

---

# Papers Documentation

*Release 1.0*

**Noel O'Boyle**

March 22, 2012



# CONTENTS

<b>1 CurlySMILES: a chemical language to customize and annotate encodings of molecular and nanodevice structures</b>	<b>1</b>
1.1 Abstract . . . . .	1
1.2 Background . . . . .	1
1.3 Results . . . . .	2
1.4 Discussion . . . . .	7
1.5 Conclusions . . . . .	8
1.6 Abbreviations . . . . .	8
1.7 Competing interests . . . . .	8
1.8 Authors' contributions . . . . .	8
<b>2 Prediction of cyclin-dependent kinase 2 inhibitor potency using the fragment molecular orbital method</b>	<b>9</b>
2.1 Abstract . . . . .	9
2.2 Background . . . . .	10
2.3 Computational and Experimental Details . . . . .	12
2.4 Results and Discussion . . . . .	16
2.5 Conclusions . . . . .	19
2.6 Competing interests . . . . .	19
2.7 Authors' contributions . . . . .	19
2.8 Acknowledgements . . . . .	19
<b>3 jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints</b>	<b>21</b>
3.1 Abstract . . . . .	21
3.2 Background . . . . .	22
3.3 Methods . . . . .	23
3.4 Implementation . . . . .	29
3.5 Results and Discussion . . . . .	30
3.6 Conclusions . . . . .	33
3.7 Availability . . . . .	33
3.8 Competing interests . . . . .	33
3.9 Authors' contributions . . . . .	33
3.10 Appendix . . . . .	34
<b>4 PubChem3D: Conformer generation</b>	<b>35</b>
4.1 Abstract . . . . .	35
4.2 Background . . . . .	36
4.3 Results and Discussion . . . . .	37
4.4 Conclusion . . . . .	49
4.5 Materials and Methods . . . . .	51

4.6	Competing interests . . . . .	52
4.7	Authors' contributions . . . . .	52
4.8	Acknowledgements . . . . .	52
<b>5</b>	<b>Modular Chemical Descriptor Language (MCDL): Stereochemical modules</b>	<b>53</b>
5.1	Abstract . . . . .	53
5.2	Background . . . . .	53
5.3	Results and Discussion . . . . .	54
5.4	Conclusions . . . . .	63
5.5	Competing interests . . . . .	64
5.6	Authors' contributions . . . . .	65
5.7	Appendices . . . . .	65
5.8	Acknowledgements . . . . .	65
<b>6</b>	<b>FlaME: Flash Molecular Editor - a 2D structure input tool for the web</b>	<b>67</b>
6.1	Abstract . . . . .	67
6.2	Background . . . . .	68
6.3	Results and discussion . . . . .	68
6.4	Implementation . . . . .	72
6.5	Conclusions and Outlook . . . . .	76
6.6	Availability and requirements . . . . .	78
6.7	Competing interests . . . . .	78
6.8	Authors' contributions . . . . .	78
6.9	Acknowledgements . . . . .	78
<b>7</b>	<b>Use of structure-activity landscape index curves and curve integrals to evaluate the performance of multiple machine learning prediction models</b>	<b>79</b>
7.1	Abstract . . . . .	79
7.2	1. Background . . . . .	80
7.3	2. Experimental . . . . .	81
7.4	3. Methods . . . . .	83
7.5	4. Results and Discussion . . . . .	84
7.6	5. Conclusions . . . . .	91
7.7	6. Abbreviations . . . . .	92
7.8	7. Competing interests . . . . .	92
7.9	8. Authors' contributions . . . . .	92
7.10	9. Acknowledgements . . . . .	92
<b>8</b>	<b>Confab - Systematic generation of diverse low-energy conformers</b>	<b>93</b>
8.1	Abstract . . . . .	93
8.2	Introduction . . . . .	93
8.3	Methods . . . . .	94
8.4	Coverage of Conformational Space . . . . .	98
8.5	Distance Distribution in Conformations of a Phenyl Sulfone . . . . .	103
8.6	Conclusion . . . . .	103
8.7	Availability and Requirements . . . . .	106
8.8	Authors' contributions . . . . .	106
8.9	Acknowledgements and Funding . . . . .	106
<b>9</b>	<b>PubChem3D: Diversity of shape</b>	<b>107</b>
9.1	Abstract . . . . .	107
9.2	Background . . . . .	107
9.3	Results and Discussion . . . . .	109
9.4	Conclusion . . . . .	119
9.5	Materials and methods . . . . .	120

9.6	Competing interests . . . . .	122
9.7	Authors' contributions . . . . .	123
9.8	Acknowledgements . . . . .	123
<b>10</b>	<b>Interpreting linear support vector machine models with heat map molecule coloring</b>	<b>125</b>
10.1	Abstract . . . . .	125
10.2	Background . . . . .	126
10.3	Methods . . . . .	127
10.4	Experimental . . . . .	130
10.5	Results and Discussion . . . . .	131
10.6	Conclusions . . . . .	135
10.7	Availability . . . . .	136
10.8	Competing interests . . . . .	136
10.9	Authors' contributions . . . . .	136
<b>11</b>	<b>Multilevel Parallelization of AutoDock 4.2</b>	<b>137</b>
11.1	Abstract . . . . .	137
11.2	Background . . . . .	137
11.3	Implementation . . . . .	139
11.4	Results and Discussion . . . . .	141
11.5	Conclusions . . . . .	144
11.6	Availability and Requirements . . . . .	144
11.7	Competing interests . . . . .	145
11.8	Authors' contributions . . . . .	145
11.9	Acknowledgements . . . . .	145
<b>12</b>	<b>PubChem3D: Similar conformers</b>	<b>147</b>
12.1	Abstract . . . . .	147
12.2	Background . . . . .	148
12.3	Results and discussion . . . . .	150
12.4	Conclusion . . . . .	164
12.5	Materials and methods . . . . .	164
12.6	Competing interests . . . . .	168
12.7	Authors' contributions . . . . .	168
12.8	Acknowledgements . . . . .	168
<b>13</b>	<b>Analysis of <i>in vitro</i> bioactivity data extracted from drug discovery literature and patents: Ranking human protein targets by assayed compounds and molecular scaffolds</b>	<b>1654</b>
13.1	Abstract . . . . .	169
13.2	Introduction . . . . .	170
13.3	Databases and Processing . . . . .	170
13.4	Results and Discussion . . . . .	171
13.5	Conclusions . . . . .	176
13.6	Endnotes . . . . .	177
13.7	Competing interests . . . . .	177
13.8	Authors' contributions . . . . .	177
13.9	Acknowledgements . . . . .	177
<b>14</b>	<b>Resource description framework technologies in chemistry</b>	<b>179</b>
14.1	Editorial . . . . .	179
14.2	1 Concepts . . . . .	181
14.3	2 Formats . . . . .	182
14.4	3 Querying the World Wide Web . . . . .	182
14.5	4 Ontologies . . . . .	182
14.6	5 Discussion . . . . .	184

14.7	6 Outlook . . . . .	185
<b>15</b>	<b>Semantic Web integration of Cheminformatics resources with the SADI framework</b>	<b>187</b>
15.1	Abstract . . . . .	187
15.2	Background . . . . .	188
15.3	Results and Discussion . . . . .	191
15.4	Conclusions . . . . .	197
15.5	Methods . . . . .	198
15.6	Authors' contributions . . . . .	198
15.7	Acknowledgements . . . . .	199
<b>16</b>	<b>ChemicalTagger: A tool for semantic text-mining in chemistry</b>	<b>201</b>
16.1	Abstract . . . . .	201
16.2	Background . . . . .	201
16.3	Methods . . . . .	203
16.4	Results and discussion . . . . .	211
16.5	Conclusions . . . . .	213
16.6	Competing interests . . . . .	214
16.7	Authors' contributions . . . . .	214
16.8	Acknowledgements . . . . .	214
<b>17</b>	<b>AMBIT RESTful web services: an implementation of the OpenTox application programming interface</b>	<b>215</b>
17.1	Abstract . . . . .	215
17.2	Background . . . . .	216
17.3	Implementation . . . . .	218
17.4	Results and Discussion . . . . .	224
17.5	Conclusions . . . . .	234
17.6	Availability and requirements . . . . .	235
17.7	Abbreviations . . . . .	235
17.8	Competing interests . . . . .	236
17.9	Authors' contributions . . . . .	236
17.10	Authors' information . . . . .	236
17.11	Acknowledgements and Funding . . . . .	236
<b>18</b>	<b>Linked open drug data for pharmaceutical research and development</b>	<b>239</b>
18.1	Abstract . . . . .	239
18.2	Findings . . . . .	239
18.3	Competing interests . . . . .	243
18.4	Authors' contributions . . . . .	244
18.5	Acknowledgements . . . . .	244
<b>19</b>	<b>Chemical Entity Semantic Specification: Knowledge representation for efficient semantic cheminformatics and facile data integration</b>	<b>245</b>
19.1	Abstract . . . . .	245
19.2	Background . . . . .	246
19.3	Results and Discussion . . . . .	249
19.4	Conclusions . . . . .	264
19.5	Methods . . . . .	265
19.6	Competing interests . . . . .	265
19.7	Authors' contributions . . . . .	265
19.8	Acknowledgements . . . . .	265
<b>20</b>	<b>Consistent two-dimensional visualization of protein-ligand complex series</b>	<b>267</b>
20.1	Abstract . . . . .	267
20.2	Background . . . . .	268

20.3 Methods . . . . .	268
20.4 Results . . . . .	274
20.5 Discussion . . . . .	276
20.6 Competing interests . . . . .	277
20.7 Authors' contributions . . . . .	277
20.8 Acknowledgements . . . . .	277
<b>21 Predicting a small molecule-kinase interaction map: A machine learning approach</b>	<b>279</b>
21.1 Abstract . . . . .	279
21.2 Background . . . . .	280
21.3 Materials and methods . . . . .	280
21.4 Results . . . . .	285
21.5 Related Work . . . . .	292
21.6 Conclusion . . . . .	293
21.7 Competing interests . . . . .	294
21.8 Authors' contributions . . . . .	294
21.9 Authors' information . . . . .	294
21.10 Acknowledgements . . . . .	294
<b>22 4D Flexible Atom-Pairs: An efficient probabilistic conformational space comparison for ligand-based virtual screening</b>	<b>295</b>
22.1 Abstract . . . . .	295
22.2 Background . . . . .	296
22.3 Methods . . . . .	297
22.4 Experimental . . . . .	302
22.5 Results and Discussion . . . . .	305
22.6 Conclusions . . . . .	310
22.7 List of abbreviations . . . . .	310
22.8 Competing interests . . . . .	311
22.9 Authors' contributions . . . . .	311
<b>23 Data governance in predictive toxicology: A review</b>	<b>313</b>
23.1 Abstract . . . . .	313
23.2 Introduction . . . . .	314
23.3 Data Governance: Main Decision Domains . . . . .	315
23.4 Review of Public Data Sources Supporting Predictive Toxicology . . . . .	317
23.5 Summary . . . . .	327
23.6 Conclusions . . . . .	327
23.7 Competing interests . . . . .	328
23.8 Authors' contributions . . . . .	328
23.9 Acknowledgements and funding . . . . .	328
<b>24 PubChem3D: Shape compatibility filtering using molecular shape quadrupoles</b>	<b>329</b>
24.1 Abstract . . . . .	329
24.2 Background . . . . .	330
24.3 Results and Discussion . . . . .	331
24.4 Conclusion . . . . .	337
24.5 Materials and methods . . . . .	337
24.6 Competing interests . . . . .	339
24.7 Authors' contributions . . . . .	339
24.8 Acknowledgements . . . . .	339
<b>25 PubChem3D: Biologically relevant 3-D similarity</b>	<b>341</b>
25.1 Abstract . . . . .	341
25.2 Background . . . . .	342

25.3	Results and Discussion . . . . .	343
25.4	Conclusion . . . . .	354
25.5	Materials and methods . . . . .	356
25.6	Competing interests . . . . .	356
25.7	Authors' contributions . . . . .	357
25.8	Acknowledgements . . . . .	357
<b>26</b>	<b>Theoretical NMR correlations based Structure Discussion</b>	<b>359</b>
26.1	Abstract . . . . .	359
26.2	Findings . . . . .	359
26.3	Availability . . . . .	362
26.4	Competing interests . . . . .	362
26.5	Authors' contributions . . . . .	362
26.6	Acknowledgements . . . . .	362
<b>27</b>	<b>AZOrange - High performance open source machine learning for QSAR modeling in a graphical programming environment</b>	<b>363</b>
27.1	Abstract . . . . .	363
27.2	Background . . . . .	364
27.3	Implementation . . . . .	365
27.4	Results and Discussion . . . . .	369
27.5	Conclusions . . . . .	373
27.6	Availability and Requirements . . . . .	373
27.7	Competing interests . . . . .	373
27.8	Authors' contributions . . . . .	373
27.9	Acknowledgements . . . . .	374
<b>28</b>	<b>Multiple search methods for similarity-based virtual screening: analysis of search overlap and precision</b>	<b>375</b>
28.1	Abstract . . . . .	375
28.2	Background . . . . .	376
28.3	Results and Discussion . . . . .	376
28.4	Conclusions . . . . .	384
28.5	Experimental Methods . . . . .	384
28.6	Competing interests . . . . .	385
28.7	Authors' contributions . . . . .	385
28.8	Acknowledgements . . . . .	385
<b>29</b>	<b>Structural diversity of biologically interesting datasets: a scaffold analysis approach</b>	<b>387</b>
29.1	Abstract . . . . .	387
29.2	Background . . . . .	388
29.3	Results and discussion . . . . .	389
29.4	Conclusions . . . . .	394
29.5	Methods . . . . .	395
29.6	Competing interests . . . . .	397
29.7	Authors' contributions . . . . .	397
29.8	Acknowledgements . . . . .	397
<b>30</b>	<b>Statistical filtering for NMR based structure generation</b>	<b>399</b>
30.1	Abstract . . . . .	399
30.2	Findings . . . . .	399
30.3	Availability . . . . .	401
30.4	Competing interests . . . . .	403
30.5	Authors' contributions . . . . .	403
30.6	Acknowledgements . . . . .	403

<b>31 PubChem3D: a new resource for scientists</b>	<b>405</b>
31.1 Abstract . . . . .	405
31.2 Background . . . . .	406
31.3 Construction and Content . . . . .	407
31.4 Utility . . . . .	411
31.5 Examples of use . . . . .	415
31.6 Conclusions . . . . .	416
31.7 Abbreviations . . . . .	418
31.8 Competing interests . . . . .	418
31.9 Authors' contributions . . . . .	418
31.10 Acknowledgements . . . . .	418
<b>32 Open Babel: An open chemical toolbox</b>	<b>419</b>
32.1 Abstract . . . . .	419
32.2 Introduction . . . . .	419
32.3 Features . . . . .	420
32.4 Implementation . . . . .	424
32.5 Using Open Babel . . . . .	429
32.6 Conclusions . . . . .	434
32.7 Availability and Requirements . . . . .	434
32.8 Competing interests . . . . .	434
32.9 Authors' contributions . . . . .	434
32.10 Acknowledgements and Funding . . . . .	435
<b>33 Adventures in public data</b>	<b>437</b>
33.1 Abstract . . . . .	437
33.2 Introduction . . . . .	437
33.3 Discussion . . . . .	438
33.4 Endnotes . . . . .	461
<b>34 Three stories about the conduct of science: Past, future, and present</b>	<b>463</b>
34.1 Abstract . . . . .	463
34.2 A story of the past . . . . .	463
34.3 A story of the future . . . . .	465
34.4 A story of the present . . . . .	467
<b>35 Openness as infrastructure</b>	<b>471</b>
35.1 Abstract . . . . .	471
35.2 Commentary . . . . .	471
35.3 Endnotes . . . . .	474
<b>36 Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on</b>	<b>477</b>
36.1 Abstract . . . . .	477
36.2 Background . . . . .	477
36.3 Scope . . . . .	478
36.4 Open Source . . . . .	479
36.5 Open Standards . . . . .	488
36.6 Open Data . . . . .	490
36.7 Other areas of activity . . . . .	492
36.8 Conclusions . . . . .	493
36.9 Competing interests . . . . .	494
36.10 Authors' contributions . . . . .	494
36.11 Acknowledgements . . . . .	494
<b>37 The Quixote project: Collaborative and Open Quantum Chemistry data management in the Internet</b>	

<b>age</b>	<b>495</b>
37.1 Abstract . . . . .	495
37.2 Background . . . . .	496
37.3 Methods . . . . .	504
37.4 Results and Discussion . . . . .	509
37.5 Conclusions . . . . .	510
37.6 Competing interests . . . . .	511
37.7 Authors' contributions . . . . .	511
37.8 Appendixes . . . . .	512
37.9 Acknowledgements . . . . .	513
<b>38 CMLLite: a design philosophy for CML</b>	<b>515</b>
38.1 Abstract . . . . .	515
38.2 Introduction . . . . .	515
38.3 Methodology of Validation . . . . .	522
38.4 Interaction and extension of conventions . . . . .	530
38.5 Conclusions . . . . .	534
38.6 Availability of Code . . . . .	537
38.7 Competing interests . . . . .	537
38.8 Authors' contributions . . . . .	537
38.9 Appendix A . . . . .	537
38.10 Appendix B . . . . .	537
38.11 Acknowledgements . . . . .	545
<b>39 Mining chemical information from open patents</b>	<b>547</b>
39.1 Abstract . . . . .	547
39.2 Background . . . . .	547
39.3 Current systems for automatic analysis of patents . . . . .	549
39.4 PatentEye . . . . .	550
39.5 The Green Chain Reaction: are chemical reactions in the literature getting greener? . . . . .	565
39.6 Competing interests . . . . .	567
39.7 Authors' contributions . . . . .	567
39.8 Acknowledgements and funding . . . . .	567
<b>40 OSCAR4: a flexible architecture for chemical text-mining</b>	<b>569</b>
40.1 Abstract . . . . .	569
40.2 Introduction . . . . .	569
40.3 Competing interests . . . . .	582
40.4 Authors' contributions . . . . .	582
40.5 Appendixes . . . . .	582
40.6 Acknowledgements . . . . .	583
<b>41 The semantic architecture of the World-Wide Molecular Matrix (WWMM)</b>	<b>585</b>
41.1 Abstract . . . . .	585
41.2 Origins/history/vision . . . . .	585
41.3 Semantics and Ontologies in Molecular Sciences . . . . .	587
41.4 Design and evolution: technologies . . . . .	591
41.5 Software development environment . . . . .	596
41.6 Virtual communities . . . . .	596
41.7 Future Development of the WWMM . . . . .	598
41.8 Competing interests . . . . .	598
41.9 Authors' contributions . . . . .	598
41.10 Acknowledgements . . . . .	599
<b>42 The semantics of Chemical Markup Language (CML): dictionaries and conventions</b>	<b>601</b>

42.1	Abstract . . . . .	601
42.2	Introduction . . . . .	601
42.3	Semantic Elements of CML . . . . .	604
42.4	Creating dictionaries . . . . .	609
42.5	Detailed use cases of dictionary construction . . . . .	610
42.6	Software support for dictionaries and units . . . . .	612
42.7	Conclusion . . . . .	612
42.8	Competing interests . . . . .	612
42.9	Authors' contributions . . . . .	612
<b>43</b>	<b>CML: Evolution and design</b>	<b>613</b>
43.1	Abstract . . . . .	613
43.2	The genesis of CML . . . . .	613
43.3	The philosophy of CML . . . . .	618
43.4	The evolution of CML . . . . .	620
43.5	JUMBO . . . . .	621
43.6	Code-driven CML Design . . . . .	622
43.7	Validation . . . . .	623
43.8	Community-driven CML Design . . . . .	624
43.9	Foreseeable evolution of CML . . . . .	625
43.10	Sustainability . . . . .	626
43.11	Competing interests . . . . .	626
43.12	Authors' contributions . . . . .	626
43.13	Endnotes . . . . .	627
43.14	Acknowledgements . . . . .	627
<b>44</b>	<b>Ami - The chemist's amanuensis</b>	<b>629</b>
44.1	Abstract . . . . .	629
44.2	Background . . . . .	629
44.3	Outcomes & Conclusions . . . . .	642
44.4	Competing interests . . . . .	642
44.5	Authors' contributions . . . . .	642
44.6	Appendix 1: Links to documentation, code resources, etc . . . . .	643
44.7	Acknowledgements . . . . .	644
<b>45</b>	<b>The past, present and future of Scientific discourse</b>	<b>645</b>
45.1	Abstract . . . . .	645
45.2	Introduction . . . . .	645
45.3	The relationship between a journal article and data . . . . .	646
45.4	The crystal structure of 1,3-dimethylcyclobutadiene . . . . .	653
45.5	Conclusions . . . . .	655
45.6	Competing interests . . . . .	656
45.7	Acknowledgements . . . . .	656
<b>46</b>	<b>Open Bibliography for Science, Technology, and Medicine</b>	<b>657</b>
46.1	Abstract . . . . .	657
46.2	Technical note . . . . .	657
46.3	Competing interests . . . . .	658
46.4	Authors' contributions . . . . .	658
<b>47</b>	<b>Semantic science and its communication - a personal view</b>	<b>659</b>
47.1	Abstract . . . . .	659
47.2	Overview . . . . .	659
47.3	Openness and the choice of BMC as publisher . . . . .	660
47.4	Open Data . . . . .	661

47.5	The semantic vision . . . . .	661
47.6	Semantic reality . . . . .	661
47.7	Chemistry as a community . . . . .	662
47.8	The value of informatics . . . . .	663
47.9	Publishing . . . . .	663
47.10	The need to change publication processes . . . . .	664
47.11	The content of the issue . . . . .	665
47.12	The future . . . . .	666
<b>48</b>	<b>Molecular dynamics simulations and in silico peptide ligand screening of the Elk-1 ETS domain</b>	<b>667</b>
48.1	Abstract . . . . .	667
48.2	Background . . . . .	668
48.3	Computational Methods . . . . .	670
48.4	Results and Discussion . . . . .	672
48.5	Conclusions . . . . .	676
48.6	Competing interests . . . . .	677
48.7	Authors' contributions . . . . .	677
48.8	Acknowledgements . . . . .	677
<b>49</b>	<b>2D-Qsar for 450 types of amino acid induction peptides with a novel substructure pair descriptor having wider scope</b>	<b>679</b>
49.1	Abstract . . . . .	679
49.2	Background . . . . .	680
49.3	Methods . . . . .	681
49.4	Results and Discussion . . . . .	684
49.5	Conclusions . . . . .	688
49.6	Competing interests . . . . .	688
49.7	Authors' contributions . . . . .	688
<b>50</b>	<b>An investigation into pharmaceutically relevant mutagenicity data and the influence on Ames predictive potential</b>	<b>689</b>
50.1	Abstract . . . . .	689
50.2	1. Introduction . . . . .	690
50.3	2. Methods . . . . .	692
50.4	3. Results and Discussion . . . . .	694
50.5	4. Conclusions . . . . .	704
50.6	Abbreviations . . . . .	705
50.7	Competing interests . . . . .	705
50.8	Authors' contributions . . . . .	705
50.9	Acknowledgements . . . . .	705
<b>51</b>	<b>Automated annotation of chemical names in the literature with tunable accuracy</b>	<b>707</b>
51.1	Abstract . . . . .	707
51.2	Background . . . . .	708
51.3	Methods . . . . .	709
51.4	Results and Discussion . . . . .	710
51.5	Conclusion and Future Application . . . . .	715
51.6	Competing interests . . . . .	716
51.7	Authors' contributions . . . . .	716
51.8	Acknowledgements . . . . .	716
<b>52</b>	<b>MyChemise: A 2D drawing program that uses morphing for visualisation purposes</b>	<b>717</b>
52.1	Abstract . . . . .	717
52.2	Introduction . . . . .	717
52.3	Implementation . . . . .	718

52.4	Results and discussion . . . . .	718
52.5	Conclusions . . . . .	726
52.6	Competing interests . . . . .	726
52.7	Acknowledgements . . . . .	728
<b>53</b>	<b>New developments on the cheminformatics open workflow environment CDK-Taverna</b>	<b>731</b>
53.1	Abstract . . . . .	731
53.2	Background . . . . .	732
53.3	Results and Discussion . . . . .	733
53.4	Conclusions . . . . .	743
53.5	Competing interests . . . . .	743
53.6	Authors' contributions . . . . .	743
53.7	Acknowledgements . . . . .	746



# CURLYSMILES: A CHEMICAL LANGUAGE TO CUSTOMIZE AND ANNOTATE ENCODINGS OF MOLECULAR AND NANODEVICE STRUCTURES

## 1.1 Abstract

CurlySMILES is a chemical line notation which extends SMILES with annotations for storage, retrieval and modeling of interlinked, coordinated, assembled and adsorbed molecules in supramolecular structures and nanodevices. Annotations are enclosed in curly braces and anchored to an atomic node or at the end of the molecular graph depending on the annotation type. CurlySMILES includes predefined annotations for stereogenicity, electron delocalization charges, extra-molecular interactions and connectivity, surface attachment, solutions, and crystal structures and allows extensions for domain-specific annotations. CurlySMILES provides a shorthand format to encode molecules with repetitive substructural parts or motifs such as monomer units in macromolecules and amino acids in peptide chains. CurlySMILES further accommodates special formats for non-molecular materials that are commonly denoted by composition of atoms or substructures rather than complete atom connectivity.

## 1.2 Background

CurlySMILES (Curly-braces enhanced Smart Material Input Line Entry Specification) is introduced as a chemical language for the specification of chemical materials and supramolecular structures. The CurlySMILES approach provides a flexible format to encode patterns in materials and molecule-based architectures. CurlySMILES includes its own set of symbols, descriptors and rules to denote respective entities and also modifies the well-established SMILES language.

SMILES is based on a set of rules that allow the representation of a molecular structure as a sequence of atom and bond symbols in a single word or string<sup>123</sup>. Unique SMILES strings are suitable as database keys while storing the structural information within the key itself<sup>4</sup>. Since SMILES notations are constructable from molecular principle, namely the molecular graph, notations can be derived for virtual, not-yet-synthesized chemical species. The flexibility

---

<sup>1</sup> SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules

<sup>2</sup> SMILES line notation

<sup>3</sup> OpenSMILES Specification

<sup>4</sup> SMILES. 2. Algorithm for Generation of Unique SMILES Notation

and portability of SMILES has been demonstrated by its use in modeling and property estimation software<sup>56789</sup> and combinatorial libraries<sup>1011</sup>.

SMILES is bridging the opposite ends of human-friendly molecular drawings, achieved with molecule editors<sup>1213</sup>, and computer-friendly connection tables (matrices), both used in representing molecular structures. Molecular information collapsed into a compact SMILES string can efficiently be managed by computer programs and stored in markup language fields, like the<sup>14</sup>.

SMILES comes in various dialects with modifications or minor extensions of the originally published language. Implementations of SMILES parsers differ with respect to the treatment and acceptance of additionally introduced symbols and syntax<sup>2</sup>. SMILES also has been extended to encode a peptide or peptoid sequence on monomer level<sup>15</sup> and in template format<sup>16</sup>.

Recently, the IUPAC International Chemical Identifier (InChI) has been designed as a string-based identifier for chemical substances<sup>17</sup>. Like a SMILES notation, an InChI string is derived from a molecular structure representations. However, InChI is intended for “behind-the-scenes” use by computers. It is typically derived from structure representations by software, whereas SMILES handily supports molecular communication between humans and computers.

The user-friendliness and popularity of SMILES encouraged us to modify this chemical language for communication of molecular architectures that can not adequately be encoded with the current SMILES language and its derivatives. CurlySMILES modifies the SMILES language by including a novel format to encode molecular details and extra-molecular features such as non-covalent interactions and attachment to a biomolecule as well as the surface of a substrate material or nanoparticle. CurlySMILES is designed with an open format to provide users with choices of integrating shorthands such as aliases or compaction of repetitive structural units. In the following, formats and rules for constructing CurlySMILES notations are described. Applications of CurlySMILES for document annotation and chemical search are then discussed.

## 1.3 Results

### 1.3.1 CurlySMILES notation

A CurlySmiles notation is a string of dot-separated component notations. A component notation can either be a plain SMILES, an annotated SMILES, or a special format notation. A plain notation maintains the grammar and rules of the known SMILES language. A plain notation is modified by introducing attributes, such as structural variations, details and decorations, enclosed in curly braces. An annotation can be anchored to a particular atomic node or placed at the end of a SMILES component. A special format notation begins with an opening and ends with a closing curly brace and includes an alias or a notation for a structure that defies molecular-graph encoding. A string with exactly one component notation is referred to as a unary CurlySMILES notation.

A CurlySMILES notation is typically not unique since the SMILES language allows for alternate notations by selecting a starting atom arbitrarily. Further, CurlySMILES provides flexible annotations formats that leaves it to a user or

<sup>5</sup> Handbook for Estimating Physicochemical Properties of Organic Compounds

<sup>6</sup> SmilogP: A Program for a Fast Evaluation of Theoretical Log P from the SMILES Code of a Molecule

<sup>7</sup> QSAR Modeling of Peripheral Versus Central Benzodiazepine Receptor Binding Affinity of 2-Phenylimidazol[1,2-a]pyridineacetamides using Optimal Descriptors Calculated with SMILES

<sup>8</sup> Artificial Neural Networks in ADMET Modeling: Prediction of Blood-Brain Barrier Permeation

<sup>9</sup> Chemical Descriptors Library (CDL): A Generic, Open Source Software Library for Chemical Informatics

<sup>10</sup> SMILIB: Rapid Assembly of Combinatorial Libraries in SMILES Notation

<sup>11</sup> SmiLib v2.0: A Java-Based Tool for Rapid Combinatorial Library Enumeration

<sup>12</sup> Molecular structure input on the web

<sup>13</sup> The PubChem chemical structure sketcher

<sup>14</sup> XML-Based IUPAC Standard for Experimental, Predicted, and Critically Evaluated Thermodynamic Property Data Storage and Capture (ThermoML)

<sup>15</sup> CHUCKLES: A Method for Representing and Searching Peptide and Peptoid Sequences on Both Monomer and Atomic Levels

<sup>16</sup> CHORTLES: A Method for Representing Oligomeric and Template-Based Mixtures

<sup>17</sup> The IUPAC International Chemical Identifier (InChI)

application software to add and granulate details. The clear separation of attributes from the molecular-graph encoding, however, provides applications with options to match and screen notations in large data sets with precedence to attributes, while deferring molecular-connectivity processing to a later stage, at which only a selected set of candidates will be considered.

Here, we focus on the core format of CurlySMILES, outlining the basic syntax that can be extended into different domains of future interest. Example notations are supplied for selected molecules and materials, demonstrating how to represent a targeted structure or a generically defined class of structures. More examples are available in Additional file <sup>1819</sup>.

Additional file 1

**CurlySMILES encoding examples.** The encoding examples illustrate the application of CurlySMILES formats and rules to derive linear notations for selected structures including stereoisomers, fragments, ring molecules with delocalized charge, coordination compounds, macromolecules, nanostructures as well as doped and surface-functionalized materials.

[Click here for file](#)

Example notations are displayed in monospace font. Parts of a CurlySMILES notation, which are given in *italics*, present descriptive metalanguage text meant to be replaced by code in CurlySMILES format.

### 1.3.2 Multiplier

A shorthand format is introduced to encode multiple occurrences of the same component notation. The multiplier is an integer greater than one and enclosed in curly braces. It is appended to a component notation as illustrated for cobalt(II) nitrate hexahydrate (Co(NO3)2{2})

A multiplier is not considered an annotation. If annotations occur at the end of a component notation, they have to precede the multiplier.

### 1.3.3 Alias

An alias is a short form for a component notation. An alias is enclosed in curly braces. CurlySMILES distinguishes between predefined and customer-defined aliases.

This distinction is critical for the implementation of a CurlySMILES parser. Look-up of the replacement notation for an alias is internal for a predefined alias, whereas customer-defined aliases require submission of a look-up dictionary by the customer.

An alias begins with a letter or a dollar sign followed by zero or more alphanumerical, hyphen, underscore, plus sign and round bracket characters. For example, we use commonly applied short notations for cations and anions of ionic liquids and solids, such as

A customer-defined alias has to be indicated by a preceding dollar sign, allowing for notations like the following:

`myCation}.{$*myAnion*}.{$*mySolvate*}{4}`

### 1.3.4 Stoichiometric formula notation

A stoichiometric formula notation (SFN) is defined to encode materials with known atomic or substructural stoichiometry, but without a discrete pattern of finite atom connectivity (molecular structure) that could be captured in a SMILES notation. SFNs are particularly useful for encoding a broad range of solids. Further, many homo- and hetero-polyatomic clusters with complex atom connectivity can be encoded as SFNs in a compact, yet distinctive

<sup>18</sup> CurlySMILES encoding examples

<sup>19</sup> Encoding examples for CurlySMILES notations

manner. CurlySMILES applies an SFN format that resembles the nomenclature typically used to name compounds by their stoichiometric composition<sup>20</sup>, but eliminates the use of sub- and superscript markup. Multiple entries of the same atomic symbol are allowed and the symbols may occur in any order. A stoichiometric integer directly follows a symbol, whereas an isotope label precedes a symbol and is marked with the ^ character; for example  $^{13}\text{C}$ . Further, selected atomic symbols may be grouped by enclosing them within round braces. If a stoichiometric integer applies to a group, it immediately follows the closing round brace. Finally, a charge notation

To distinguish an SFN-encoded component from an alias or a composite notation, an asterisk (*SFN*)

$_{23}\text{C}_6$

$_5^{4+}$

$_3(\text{CO}_3)_2(\text{OH})_2$  (carbonate mineral)

Groups in an SFN notation can be nested to any depth level:

$_2)]$

Notice that the SFN format also accommodates structures with molecular connectivity. It is the user's choice to encode such structures as SFN or SMILES notations. When in doubt about the topological description of a structure or when isomeric forms should intentionally be included, SFN is the notation of choice.

An SFN can appear within a special format notation for a component, within a composite notation to encode a constituent (see next section) and within a SMILES annotation.

### 1.3.5 Composite notation

Composite or hybrid materials are made from constituents that remain separate and distinct, even on an above-nanoscale level (mesoscale). The constituents should not be encoded by dot-separated notations, a format which should be reserved for compounds in which the different species interact with each other on a molecular/ionic level rather than on an interface or grain-boundary level. CurlySMILES defines a special format for encoding composites and other materials built from interface-connected phases:

$^*\text{constituent}^{\#1}/^*\text{constituent}^{\#2}/.../^*\text{constituent}^{\#n}\}$

Herein,  $^*\text{constituent}^{\#i}$ <sup>21</sup> is encoded as

$_2$

The first constituent is presented as an annotated SMILES notations, using the macromolecule syntax introduced below, and the second constituent is presented as SFN.

### 1.3.6 Plain SMILES notation

A component notation that does not begin with an opening curly brace is a plain SMILES or an annotated SMILES notation. A plain SMILES notation is encoded with the grammar and rules of the SMILES language<sup>1</sup>. A plain notation contains atomic node code (ANC) and may contain bond symbols and special characters to denote branching and ring formation. An ANC is either a bare atomic symbol or a square bracket atomic code (SQC) which includes an atomic symbol and, depending on the targeted structure, additional characters to denote the number of adjacent hydrogen atoms, a charge value, and an isotope label. CurlySMILES includes the symbols

CurlySMILES requires an atomic wildcard always to be encoded as a SQC. For example, the notation<sup>22</sup> as a wildcard-like any-bond symbol. Unless the bound atoms are encoded by wildcards, a bond is predetermined by the element type of the adjacent atoms and their orbital interactions. Whereas the atomic wildcard is a placeholder for atomic symbols, the symbol for an unspecified bond has not primarily the role of a bond placeholder; rather, it indicates bond-type

---

<sup>20</sup> Nomenclature of Inorganic Chemistry

<sup>21</sup> Poly(organophosphazene)s and the Sol-gel Technique

<sup>22</sup> SMARTS - A Language for Describing Molecular Patterns

ambiguity with respect to the limited classification scheme of single, double, triple, quadruple and aromatic bonds. CurlySMILES treats the character ~ as a bond symbol that encodes a bond, which cannot adequately be encoded as a covalent bond with symbols

### 1.3.7 Annotated SMILES notation

A plain notation is annotated by inserting one or more curly-enclosed annotations into the notation. An annotation has to be either anchored at a particular atomic node or appended to the end. An atom-anchored annotation (AAA) directly follows the ANC. The only characters allowed to occur between an AAA and ANC are digits that designate ring-closing. A component-anchored annotation (CAA) follows the last ANC, including its AAAs, but precedes the multiplier, if any is present in the component notation. AAA types include stereodescriptive and structural unit annotations as well as group environment, molecular detail and operational annotations. CAA types include state and shape annotations and miscellaneous interest annotations. In the following text we use the term annotation to refer to the content between the curly braces.

A stereodescriptive annotation consists of one of the upper-case letters *cis/trans* isomers by using the *E/Z* convention. Examples E1 and E2 in Additional file

A structural unit annotation defines a boundary of a structural unit. This boundary is an open or dangling bond. The annotation consists of a one-character boundary descriptor, which is a CurlySMILES bond symbol (

Stereodescriptive and structural unit annotations consist of exactly one character, while all other annotation types require a two-character annotation marker (AM), which is optionally followed by an annotation dictionary to specify attributes. The general format for a dictionary-containing annotation,

$AMk_1=^*v*_1;^*k*_2=^*v*_2;...;^*k*_n=^*v*_n\}$ ,

employs a semicolon-separated list of dictionary entries. An entry is a key/value pair,  $k_i/v_i$ <sup>23</sup>.

A group environment annotation marker (GEAM) starts with a bond symbol (

A molecular detail annotation starts with an exclamation mark. The second character of a molecular detail annotation marker (MDAM) is

The first character of an operational annotation marker (OPAM) is a plus sign. The following character is a letter. An upper-case letter indicates formal addition or substitution of a structural part. A lower-case letter indicated formal repetition of an annotated unit. OPAMs

A state and shape annotation is denoted by a state and shape annotation marker (SSAM) consisting of two lower-case letters. SSAM annotations qualitatively describe the physicochemical state, phase structure and/or the nano- or mesoscale characteristics of a material <sup>24</sup>. Examples E15 to E18 in Additional file

A miscellaneous interest annotation is denoted by a two-character miscellaneous annotation marker (MIAM). Examples E19 and E20 in Additional file

The following two examples demonstrate the combined use of annotations to encode molecular arrangements.

The imidazolium functionalized SiO<sub>2</sub> surface <sup>25</sup> in Figure 1 is encoded in CurlySMILES by applying the group environment annotation format. The two O atoms, which attach the molecular species to the material surface, are annotated with the

The material is SFN-encoded in a dictionary entry with key

The functionalized calix[4]arene of the complex shown in Figure 2 is encoded by combining structural unit, group environment and operational annotations. The molecular ring is encoded as a fragment with two open bonds. The first is encoded as a structural unit annotation and the second as an operational annotation. Corresponding atomic nodes and annotations are given in boldface:

<sup>23</sup> CurlySMILES: annotation dictionary keys

<sup>24</sup> CurlySMILES: state and shape annotation

<sup>25</sup> Supported ionic liquids: ordered mesoporous silicas containing covalently linked ionic species

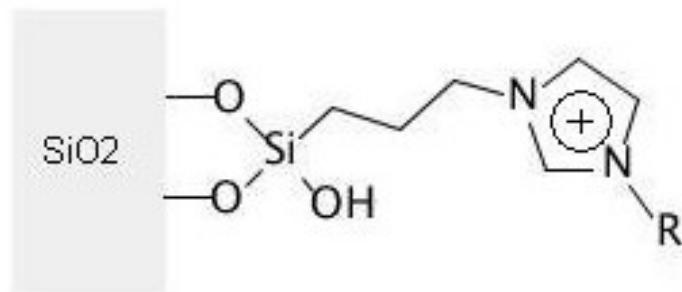


Figure 1.1: Figure 1. Alkylimidazolium ionic species immobilized on silica surface  
**Alkylimidazolium ionic species immobilized on silica surface.**

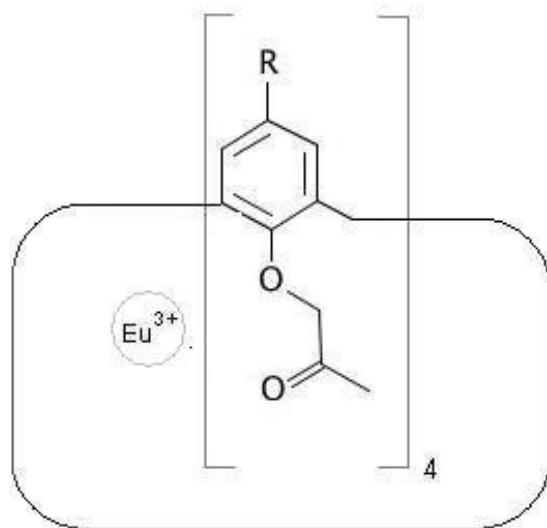


Figure 1.2: Figure 2. Eu<sup>3+</sup> cation coordinated by a cryptand  
**:sup:'3+' cation coordinated by a cryptandEu.**

`c{-}**cc{-R}cc1**C{+rn=4}`

The  $^{3+}$ /cryptand complex is given by encoding the rare-earth cation in SQC, annotated with the ligand notation:

The dictionary key

### 1.3.8 Software

A suite of Python modules have been implemented that perform parsing and molecular descriptor generation for CurlySMILES notations. These modules have been wrapped as software package <http://www.axeleratio.com/csm/py/code/downloads.htm>.

## 1.4 Discussion

### 1.4.1 Comparison of CurlySMILES with other SMILES modifications

Versions of SMILES use the symbols *Cis* and *trans* isomers are denoted via directional bond symbols

CurlySMILES introduces a format to mark a fragment or chemical group by using structural unit annotations. This allows distinction between a radical and a group. The SMILES notation [CH3] encodes a methyl radical. CurlySMILES uses the same notation to encode the radical, but encodes a methyl group as

In a CurlySMILES notation, SMILES code is strictly separated and distinguishable from other parts in the notation by using curly braces. CurlySMILES uses marked annotations to define structures in generic terms or to construct molecular patterns with a tunable depth of information granularity. CurlySMILES provides completely new, extra-molecular patterns, such as

CurlySMILES is designed for applications in specialized domains and for clients with particular tasks, including repetitive processing of certain structural entities. For this purpose, CurlySMILES includes various shorthand approaches, especially the alias format. Domain-specific abbreviations and codes for structures and materials are frequently used within chemical communities to replace long names and complex structural concepts. The alias format makes it possible to integrate those terms into notations and replace them when needed.

As a machine-readable code, a CurlySMILES notation (like SMILES or InChI) is a document-neutral representation (ASCII string) of a chemical structure that, supported by the methods in the supplied software package, can automatically be converted into a document specific format. Formula-based names of coordination compounds are a case in point.

### 1.4.2 Data mining and semantic search

A special feature of CurlySMILES is that it integrates textual parts, acronyms, other encoding schemes and client-defined aliases. Thus, a CurlySMILES notation can be used on various levels in search and data mining. CurlySMILES allows formulation of complex search pattern by using atomic-symbol placeholders and annotations that denote generic functional groups and compound classes. The annotation syntax of CurlySMILES makes it possible to associate a structural part with a specific role such as a structural repeat unit in a polymer, a substituent, ligand, cryptand, dopant, adsorbate or dissolved species. By implementing search and matching algorithm that include the information contained in CurlySMILES annotations, a plethora of search strategies can be envisioned, including precise, needle-in-the-haystack search and customer-focused report-and-review style extraction of chemical data.

## 1.5 Conclusions

CurlySMILES is a chemical language for the communication of chemical information related to molecular structure and complex nanoscale architectures. This language offers a versatile approach in encoding material composition and structure by supplementing attention to extra-molecular features. Symbols and grammar of this language allow users to encode structures and to formulate context-annotated queries with variable granularity of molecular or supramolecular details. The open format makes it easy to extend the current version to application-specific tasks.

## 1.6 Abbreviations

AAA: Atom-Anchored Annotation; AM: Annotation Marker; ANC: Atomic Node Code; CAA: Component-Anchored Annotation; CurlySMILES: Curly-braces enhanced Smart Material Input Line Entry Specification; GEAM: Group Environment Annotation Marker; InChI: IUPAC International Chemical Identifier; MDAM: Molecular Detail Annotation Marker; MIAM: Miscellaneous Interest Annotation Marker; OPAM: Operational Annotation Marker; SFN: Stoichiometric Formula Notation; SMILES: Simplified Molecular Input Line Entry System; SQC: SQuare bracket atomic Code; SSAM: State and Shape Annotation Marker

## 1.7 Competing interests

The author declares that they have no competing interests.

## 1.8 Authors' contributions

AD designed the CurlySMILES language. AD has implemented software to parse and test CurlySMILES notation and to use them in CurlySMILES-annotated archives of chemical property data and bibliographic collections.

# PREDICTION OF CYCLIN-DEPENDENT KINASE 2 INHIBITOR POTENCY USING THE FRAGMENT MOLECULAR ORBITAL METHOD

## 2.1 Abstract

### 2.1.1 Background

The reliable and robust estimation of ligand binding affinity continues to be a challenge in drug design. Many current methods rely on molecular mechanics (MM) calculations which do not fully explain complex molecular interactions. Full quantum mechanical (QM) computation of the electronic state of protein-ligand complexes has recently become possible by the latest advances in the development of linear-scaling QM methods such as the *ab initio* fragment molecular orbital (FMO) method. This approximate molecular orbital method is sufficiently fast that it can be incorporated into the development cycle during structure-based drug design for the reliable estimation of ligand binding affinity. Additionally, the FMO method can be combined with approximations for entropy and solvation to make it applicable for binding affinity prediction for a broad range of target and chemotypes.

### 2.1.2 Results

We applied this method to examine the binding affinity for a series of published cyclin-dependent kinase 2 (CDK2) inhibitors. We calculated the binding affinity for 28 CDK2 inhibitors using the *ab initio* FMO method based on a number of X-ray crystal structures. The sum of the pair interaction energies (PIE) was calculated and used to explain the gas-phase enthalpic contribution to binding. The correlation of the ligand potencies to the protein-ligand interaction energies gained from FMO was examined and was seen to give a good correlation which outperformed three MM force field based scoring functions used to approximate the free energy of binding. Although the FMO calculation allows for the enthalpic component of binding interactions to be understood at the quantum level, as it is an *in vacuo* single point calculation, the entropic component and solvation terms are neglected. For this reason a more accurate and predictive estimate for binding free energy was desired. Therefore, additional terms used to describe the protein-ligand interactions were then calculated to improve the correlation of the FMO derived values to experimental free energies of binding. These terms were used to account for the polar and non-polar solvation of the molecule estimated by the Poisson-Boltzmann equation and the solvent accessible surface area (SASA), respectively, as well as a correction term for ligand entropy. A quantitative structure-activity relationship (QSAR) model obtained by Partial Least Squares projection to latent structures (PLS) analysis of the ligand potencies and the calculated terms showed a strong correlation ( $r^2 = 0.939$ ,  $q^2 = 0.896$ ) for the 14 molecule test set which had a Pearson rank order correlation of

0.97. A training set of a further 14 molecules was well predicted ( $r^2 = 0.842$ ), and could be used to obtain meaningful estimations of the binding free energy.

### 2.1.3 Conclusions

Our results show that binding energies calculated with the FMO method correlate well with published data. Analysis of the terms used to derive the FMO energies adds greater understanding to the binding interactions than can be gained by MM methods. Combining this information with additional terms and creating a scaled model to describe the data results in more accurate predictions of ligand potencies than the absolute values obtained by FMO alone.

## 2.2 Background

A major goal in computational structure-based drug design and virtual screening protocols is to accurately predict the free energy of ligand binding to a receptor in a timescale that is amenable to drug discovery<sup>1</sup>. This is attractive for reducing costs in the discovery process by replacing wet-lab experiments with computer simulation, accelerating the discovery process and assisting in lead optimisation<sup>2</sup>. A popular procedure to identify possible lead compounds is to run a virtual screening campaign by docking a large number of diverse compounds to a receptor binding site<sup>3</sup>. A score is then given to the docked pose based on a potential function which relates the spatial orientation of a ligand in a binding site to the free energy of binding. The scoring functions are generally used in a qualitative manner to rank ligand binding poses and in doing so estimate the free energy of binding. Docking programmes are generally recognised for making reasonably successful predictions of binding modes, however, the scoring functions used to predict the binding affinity are less reliable<sup>456</sup>. There must be a balance between the attempted accuracy of the scoring function and the computational time required to perform that calculation. A compromise for improved accuracy at greater computational expense can result in overly complicated and slower functions ill-suited for the turn-around times required within a medicinal chemistry program. Methods to develop a physically satisfying model to estimate the free energy of ligand binding to a receptor accurately enough to be predictive and useful, in a reasonable amount of time, has proven challenging<sup>78</sup>.

The most rigorous theoretical methods that have been developed to estimate the free energy of binding from a thermodynamic standpoint are based on free-energy perturbation (FEP), thermodynamic integration (TI) and similar methodologies<sup>9</sup>. These methods are still limited by their use of MM force fields, and are further limited by high computational expense and are best suited to examining relative binding affinities of a small number of similar ligands. A number of approximate methods based on structural sampling, have been developed, to find appropriate stable structures and to cover enough conformational space for entropy estimations to be possible. These methods include linear-response approximation (LRA), the semi-macroscopic version of the protein-dipole Langevin-dipole approach (PDLD/S-LRA), the linear interaction energy (LIE) and molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) approaches<sup>101112131415</sup>. Also, there has been some validation for the use of a single molecular conformation, where the estimation

<sup>1</sup> The Many Roles of Computation in Drug Discovery

<sup>2</sup> Lead discovery using molecular docking

<sup>3</sup> Structure-based virtual screening protocols

<sup>4</sup> Evaluation of Docking Performance: Comparative Data on Docking Algorithms

<sup>5</sup> Detailed Analysis of Scoring Functions for Virtual Screening

<sup>6</sup> A Critical Assessment of Docking Programs and Scoring Functions

<sup>7</sup> AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading

<sup>8</sup> Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors

<sup>9</sup> Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems

<sup>10</sup> Ligand binding affinity prediction by linear interaction energy methods

<sup>11</sup> Examining Methods for Calculations of Binding Free Energies: LRA, LIE, PDLD-LRA, and PDLD/S-LRA Calculations of Ligands Binding to an HIV Protease

<sup>12</sup> Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of free energies of binding of phosphorylcholine analogs to McPC603

<sup>13</sup> Modeling electrostatic effects in proteins

<sup>14</sup> Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models

<sup>15</sup> Protein-protein interactions from linear-scaling first-principles quantum-mechanical calculations

of binding affinities is based on either physical or statistical measures<sup>81617</sup>.

The physical methods mentioned are based on calculations with a MM force field, enabling fast energy determination through the utilisation of extensive phase space sampling<sup>1819</sup>. Additionally, the system can be parameterised to account for solvation effects. However, the accuracy of the underlying force field underpins any estimation of binding free energies<sup>20</sup>. Conventional force fields are limited in that electronic effects are not accounted for adequately. It is becoming increasingly apparent that there are numerous kinds of non-classical intermolecular forces, such as cation- $\pi$ <sup>2122</sup>, dipole- $\pi$ <sup>23</sup>, halogen- $\pi$ <sup>24</sup>, carbonyl n- $\pi^*$ <sup>25</sup>, and so-called “non-conventional hydrogen bonds”, are playing an important role in inter- and intra-molecular interactions. Implementation of QM chemical calculations can significantly improve the accuracy of conventional force fields by accounting for charge transfer, polarisation effects, dispersion and other bonding interactions with greater rigor<sup>262728</sup>. QM chemical calculations explicitly describe these non-classical interactions whereas they are not accounted for by MM force fields. Such QM methods are typically based on either semi-empirical calculations<sup>29</sup> or *ab initio* methods using fractional approaches, *e.g.*, the fragment molecular orbital (FMO) method or the molecular fractionation with conjugate caps (MFCC) and related methods<sup>3031</sup>. The QM/MM method is another method that attempts to overcome the system size and sampling limitations of QM methods. In QM/MM simulations, a region that requires accurate analysis is studied quantum-mechanically, and other regions are studied by classical force field calculations.

Binding interaction energies can be studied in a new light using QM methods. The charge transfer and polarisation effects are particularly important when studying hydrogen bonding<sup>32</sup>. Many force fields treat hydrogen bond effects through their van der Waals (vdW) and fixed electrostatic contributions, however, hydrogen bonding interactions are complex. Hydrogen bonds are highly directional. There are however varying amounts of charge transfer and polarisation energy components that contribute to hydrogen bonding<sup>333435</sup>. QM methods account for dispersion forces more adequately than MM force fields because the electronic correlation effects are taken into account appropriately<sup>36</sup>. Only one of these previous studies has been performed at a level (MP2/6-311(+G(2 d,p)) for which there is hope that dispersion and polarisation effects are treated in a balanced and satisfactory way<sup>37</sup>.

QM methods have begun to demonstrate their usefulness as scoring functions for calculating ligand binding free energies. Semi-empirical methods have been used to build PLS models to describe protein-ligand interactions<sup>3839</sup>, and as computer power has increased *ab initio* QM methods have also been used<sup>304041</sup>. Historically QM methods were primarily limited to smaller systems because of the computational expense, but these methods are now tenable

<sup>16</sup> Validation and Use of the MM-PBSA Approach for Drug Discovery

<sup>17</sup> Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA

<sup>18</sup> A Multistep Approach to Structure-Based Drug Design?: Studying Ligand Binding at the Human Neutrophil Elastase

<sup>19</sup> Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular Mechanics and Continuum Solvent Models

<sup>20</sup> Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System

<sup>21</sup> Cation-pi interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide

<sup>22</sup> Cation- $\pi$  interactions in structural biology

<sup>23</sup> Origin of the Attraction and Directionality of the NH/p Interaction?: Comparison with OH/p and CH/p Interactions

<sup>24</sup> Cl- $\pi$  Interactions in Protein-Ligand Complexes

<sup>25</sup> On the Nature of Bonding in Lone Pair-Electron Complexes: CCSD(T)/Complete Basis Set Limit Calculations

<sup>26</sup> Quantum Chemical Benchmark Energy and Geometry Database for Molecular Clusters and Complex Molecular Systems: A Users Manual and Examples

<sup>27</sup> The role of quantum mechanics in structure-based drug design

<sup>28</sup> The Role of Polarization and Charge Transfer in the Solvation of Biomolecules

<sup>29</sup> Large-Scale Validation of a Quantum Mechanics Based Scoring Function?: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein-Ligand Complexes

<sup>30</sup> Fragment molecular orbital method: an approximate computational method for large molecules

<sup>31</sup> Quantum computational analysis for drug resistance of HIV-1 reverse transcriptase to nevirapine through point mutations

<sup>32</sup> The Origin of Hydrogen Bonding. An Energy Decomposition Study

<sup>33</sup> Computation of charge-transfer energies by perturbation theory

<sup>34</sup> A general treatment of solvent effects based on screened Coulomb potentials

<sup>35</sup> Divide and conquer interaction energy decomposition

<sup>36</sup> Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs

<sup>37</sup> On the accurate reproduction of *ab initio* interaction energies between an enzyme and substrate

<sup>38</sup> QM/QSAR: Utilization of a semiempirical probe potential in a field-based QSAR method

<sup>39</sup> Semiempirical Comparative Binding Energy Analysis (SE-COMBINE) of a Series of Trypsin Inhibitors

<sup>40</sup> Ab initio fragment molecular orbital (FMO) method applied to analysis of the ligand-protein interaction in a pheromone-binding protein

<sup>41</sup> Quantum mechanical map for protein-ligand binding with application to  $^{35}\beta$ -trypsin/benzamidine complex

for larger systems because of the advent of fractional QM methods. However, in order to provide reliable ligand-binding energies, additional terms accounting for solvent effects, entropy, and sampling need to be considered. Only recently has an estimate of ligand-binding energies with realistic QM methods, which considered these factors, been published<sup>42</sup>.

The FMO method is an attractive method for dealing with large biomolecular systems quantum mechanically. In the FMO method, a large molecular system is divided into smaller fragments, and the conventional molecular orbital calculations are performed for each fragment and fragment pair. This QM method is gaining attention as an accurate and fast method to correlate binding affinity to calculated values<sup>43444546</sup>. We compare our results to MM-based scoring functions and show the importance of high level QM methods to obtain reasonable binding energy predictions. Using only the FMO method resulted in values for the gas phase binding interactions, however protein-ligand interactions are more complex than this, as illustrated in the thermodynamic cycle shown in Figure 1. An alternative approach was then taken to account for all aspects of the binding phenomenon at various levels of approximation. In an effort to account for solvation and entropic binding events further terms were included, together with the enthalpic contribution of ligand binding calculated from the FMO method, to form a scoring function. The electrostatic interactions between the ligand and the protein and between the solvent and the protein-ligand complex are determined by solving the Poisson-Boltzmann equation. An entropic term was derived from the number of rotatable bonds present in the ligand. These terms were then used to build a PLS model as a scoring function to estimate the free energy of binding. The results show that consideration of other contributing terms pertaining to the thermodynamic cycle greatly enhances the predictability of free energy binding models. This was validated using a series of CDK2 inhibitors.

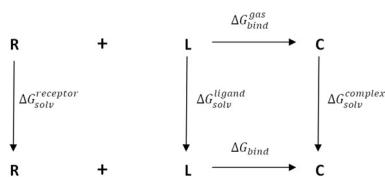


Figure 2.1: Figure 1. Schematic view of the thermodynamic cycle used to in the derivation of the binding affinity  
**Schematic view of the thermodynamic cycle used to in the derivation of the binding affinity.** The cycle calculates the receptor (R), ligand (L), and complex (C) in vacuum and then transfers them to solvent to find the solvation free energy.

## 2.3 Computational and Experimental Details

### 2.3.1 Data Set Preparation

A database of 28 CDK2 inhibitors with experimental binding affinity available in the literature was compiled<sup>4748</sup>. The reported IC<sub>50</sub> ( $\mu\text{M}$ ) values were converted to -ln(IC<sub>50</sub>) values and the free energy of binding ( $\Delta_{\text{bind}}G$ ) was calculated according to the Eq. (1) at 310 K.

The compounds with known X-ray structures were selected as the training set to compare the various methods used to predict the free energy of binding. In order to effectively validate the PLS model, compounds that were not included

<sup>42</sup> Ligand Affinities Estimated by Quantum Chemical Calculations

<sup>43</sup> Comparison of binding affinity evaluations for FKBP ligands with state-of-the-art computational methods: FMO, QM/MM, MM-PB/SA and MP-CAFE approaches

<sup>44</sup> QSAR Study of Cyclic Urea Type HIV-1 PR Inhibitors Using Ab Initio MO Calculation of Their Complex Structures with HIV-1 PR

<sup>45</sup> Novel Quantitative Structure-Activity Studies of HIV-1 Protease Inhibitors of the Cyclic Urea Type Using Descriptors Derived from Molecular Dynamics and Molecular Orbital Calculations

<sup>46</sup> Correlation analyses on binding affinity of Substituted benzenesulfonamides with carbonic anhydrase using ab initio MO calculations on their Complex structures

<sup>47</sup> Recent Developments in Fragment-Based Drug Discovery

<sup>48</sup> Identification of N-(4-Piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide (AT7519), a Novel Cyclin Dependent Kinase Inhibitor Using Fragment-Based X-Ray Crystallography and Structure Based Drug Design†

in the data set to obtain the model were placed into a separate test set to assess the predictive potential of the model. The distribution of the data set into training and test sets is shown in Table 1.

### 2.3.2 Structure Preparation

The 14 X-ray structures, corresponding to the 14 ligands in the training set were obtained from the PDB (Table 1). The remaining 14 ligands for which the X-ray structure data was not available were modelled into one of the 14 reported PDB structures based on ligand structural similarity (Table 1). The protein-ligand complexes were aligned in PyMOL<sup>49</sup>. Hydrogen atoms were added and the protonation state of the acidic and basic amino acid residues were adjusted at pH 7 using the Protonate3 D tool within MOE<sup>50</sup>. An inclusion sphere with a 4.5 Å radius was projected around the bound ligands. This area defined the residues which were to be included in the QM and MM calculations. All water molecules were removed. The N-terminals of the residues were capped with acetyl groups and the C-terminal ends were N-methyl capped using the geometry of the cleaved neighbouring residue as a vector to place the capping group. Partial charges were initially calculated to optimise the system using MM. The partial charges for the ligand binding site were calculated using the MMFF94x force field and the ligand partial charges were calculated with AM1BCC charges<sup>5152</sup>. The system was geometry optimized using MMFF94x force field in the presence of the Born continuous implicit water model, with an internal dielectric constant of 3 and an external dielectric constant of 80. The coordinates of the heavy atoms of the protein and the ligand were held fixed and the protons were energy minimised using the other default settings in MOE. The 14 modelled ligands were built by manually modifying the reference ligand and then energy minimising the ligand whilst keeping the reference heavy atoms fixed according to the method detailed above. Where appropriate, charged amino acid residues were neutralised with either a chloride anion or a lithium cation.

The FMO input files for GAMESS were prepared using Facio (version 14.2.4)<sup>5354</sup> following preprocessing of the structure in MOE. An example of the protein-ligand system is shown in Figure 2.

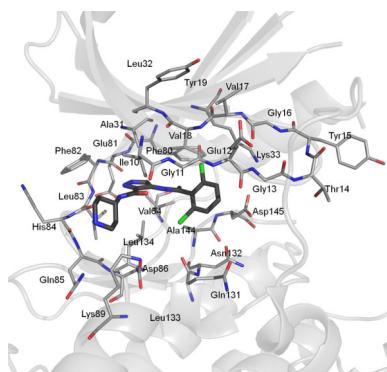


Figure 2.2: Figure 2. Orientation of the CDK2 active site in the PDB structure

**Orientation of the CDK2 active site in the PDB structure** \*\* showing the amino acid residues (grey lines) used for the QM and MM calculations\*\*. Ligand 33 is shown as grey sticks.

<sup>49</sup> The PyMOL Molecular Graphics System. (v0.99rc6)

<sup>50</sup> MOE (The Molecular Operating Environment). (2009.10)

<sup>51</sup> Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method

<sup>52</sup> Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation

<sup>53</sup> Facio: New Computational Chemistry Environment for PC GAMESS

<sup>54</sup> Development of GUI for GAMESS/FMO Calculation

### 2.3.3 The FMO Method

The FMO method has previously been thoroughly described<sup>30555657</sup>. Therefore, we provide only a short summary of the method. The FMO calculations were used to perform high-level *ab initio* quantum chemical calculations at MP2/6-31G\* (6-31G 3df for Cl and S) theory level using the one-residue-per-fragment fragmentation scheme. All calculations were run using the GAMESS implementation (either April 2008 or January 2009 version)<sup>585960</sup>.

The principle behind the FMO method is to divide a large biomolecular system into a collection of small FMO fragments and then perform molecular orbital calculations for each fragment (called monomer) and fragment pair (called dimer). Generally the system is fragmented into amino acids residues and ligands. It should be noted that FMO fragments differ from the standard assignment for amino acids residues. Here, amino acids are fragmented along the sp<sup>3</sup> bond joining the C<sub>lnonasci\_8i</sub> carbon to the peptide-bond carbonyl carbon. This simple calculation scheme significantly reduces computational time. It may be possible that this method impairs the computational accuracy because the covalent bonds are detached. However, in the FMO method, the accuracy is kept by employing projection operators made from the sp<sup>3</sup> hybrid orbital.

One of the great advantages of the FMO method is that it can be combined with a number of current quantum chemical techniques. Thus, an appropriate method for each system can be chosen. It was noted that incorporation of vdW interactions is an important consideration in studying protein-ligand interactions. Here, the FMO method provides a useful calculation scheme for dealing with these effects. Dispersion energy, which can dominant in vdW interactions, is not considered by Hartree-Fock (HF) and poorly modelled by Density Functional Theory (DFT). To achieve accurate dispersion energies correlated *ab initio* methods are required. Here, the second-order Møller-Plesset (MP2) perturbation method was used as it is the least expensive non-empirical approach. The MP2 method has been implemented in the FMO method (FMO-MP2).

The molecular system is divided into a number of monomer fragments and the *ab initio* molecular orbital calculations on the monomers are solved repeatedly at the HF level until all monomer densities become self-consistent. Then the FMO-MP2 method begins with MP2 calculations of monomers, followed by MP2 calculations of dimers. The results are used to calculate the total energy of a system following the formula:

where  $I$  and  $J$  run over all the of the fragments,  $N$ . The term  ${}_I E$  is the self-consistent field (SCF) energy of the  $I^{*th}$  fragment in the external Coulomb field of the other  $*N - 1$  fragments. The  ${}_{IJ} E$  term is the SCF energy of the  $I + J$  dimer in the external Coulomb field of the other  $N - 2$  fragments. The FMO calculations can provide PIEs, also known as inter-fragment interaction energies (IFIE), between fragments. The PIEs are used during the analysis of interaction between protein residues and the bound ligand and is derived from the FMO calculation:

where  $\Delta^{**}D^{***}IJ^*$  :sub:\ and  $**V***IJ*$  \ :sub: are the difference density matrix and the environmental electrostatic potential for dimer  $IJ$  from other fragments,<sup>40616263</sup>.

### 2.3.4 Scoring Function Used to Estimate Free Energy of Binding

The values of the free energy of binding in solvent ( $\Delta_{\text{bind}} G$ ) of each inhibitor were calculated according to Eq. (4) following thermodynamic cycle shown in Figure 1.

The solution-phase

The enthalpic binding energy of interaction term is derived from the FMO method at the MP2/6-31G\* level. The breakdown of this interaction energy can be expressed as relating to electrostatic interactions (ES), exchange repulsion

<sup>55</sup> Pair interaction molecular orbital method: an approximate computational method for molecular interactions

<sup>56</sup> Fragment molecular orbital method: application to polypeptides

<sup>57</sup> Fragment molecular orbital method: use of approximate electrostatic potential

<sup>58</sup> General Atomic and Molecular Electronic Structure System

<sup>59</sup> General Atomic and Molecular Electronic Structure System

<sup>60</sup> Advances in electronic structure theory: GAMESS a decade later

<sup>61</sup> Visualization analysis of inter-fragment interaction energies of CRP-cAMP-DNA complex based on the fragment molecular orbital method

<sup>62</sup> VISCANA: Visualized Cluster Analysis of Protein-Ligand Interaction Based on the ab Initio Fragment Molecular Orbital Method for Virtual Ligand Screening

<sup>63</sup> Intra- and intermolecular interactions between cyclic-AMP receptor protein and DNA: Ab initio fragment molecular orbital study

(EX), dispersion contributions (DI) and charge transfer (CT) with higher order mixed terms, Eq. (6)<sup>6465</sup>. Evaluation of the enthalpic ligand binding energy is commonly performed by the supermolecule method. Here, the difference between the energy of the receptor-ligand complex and the sum of the energies of the apo-receptor and the isolated ligand is considered, Eq. (10).

Thus three separate calculations are required to obtain the total ligand binding energy. However, in the FMO calculation, as all the PIEs between fragments are calculated by default, the ligand binding energy can be conveniently estimated by simply taking the sum of all the PIEs between the ligand and receptor fragments. Although the sum of PIE does not include the effect of electron redistribution in the complex as a result of ligand-protein binding, it is known that there is good qualitative agreement between the binding energy calculated by the supermolecule method and the sum of PIE<sup>44</sup>.

Entropy plays an important role in binding. During receptor-ligand complex formation, there are changes in the degrees of rotational, translational and conformational freedom, which make the process entropically unfavourable<sup>666768</sup>. The number of rotatable bonds in a ligand<sup>697071</sup> or the receptor and ligand together<sup>6772</sup> has previously been used as a measure for conformational entropy. More complex measures of determining entropy using vibrational frequencies of a ligand when complexed to a receptor have been shown to correlate well to the number of rotatable bonds<sup>73</sup>. However, vibrational entropy is not a component of conformational entropy, and do not make significant contributions to the overall entropy of the system<sup>74</sup>. The calculation of the number of rotatable bonds is also an attractive estimation for conformational ligand entropy as it is significantly less computationally demanding than other methods. Therefore we chose this method and assigned a conformational penalty of 1 kcal/mol for each rotatable bond in the ligand according to published work<sup>73</sup>.

The other term of the binding free energy is the solvation free energy<sup>66</sup>. The solvation free energy was described by a polar solvation term

The polar solvation term estimated by solving the Poisson-Boltzmann (PB) equation using MOE<sup>50</sup>. The system was parameterised as described in the Structure Preparation section. The nonpolar solvation term was estimated from the solvent-accessible surface area (SASA) of the molecule, Eq. (9). This was computed in MOE with a solvent probe radius of 1.4 Å. The values taken for `|nonascii_13|`\* and \*b were 5.0 cal/mol·Å<sup>2</sup> and 0.86 kcal/mol, respectively, as described in the literature<sup>7576</sup>. In order to speed up the calculation of the free energy of solvation we chose to use a single energy-minimised structure which has been reported in the literature to be a reasonable estimation to molecular dynamics simulations<sup>1617</sup>.

### 2.3.5 Multivariate Analysis

The statistical program SIMCAP, version 11.0.0.0, from Umetrics was used to build a PLS model<sup>7778</sup>. The X-variables originate from the components used to derive the free energy of binding in solvent, see above. The dependent Y-variable was the experimental binding affinity in -ln(IC<sub>50</sub>), Eq. (1). The variables were mean-centred and scaled to unit variance. The non-cross-validated variance coefficient ( $r^2$ ) and the cross-validated variance coefficient ( $q^2$ ) were

<sup>64</sup> Pair interaction energy decomposition analysis

<sup>65</sup> A new energy decomposition scheme for molecular interactions within the Hartree-Fock approximation

<sup>66</sup> Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction?

<sup>67</sup> The consequences of translational and rotational entropy lost by small molecules on binding to proteins

<sup>68</sup> Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein-Ligand Binding

<sup>69</sup> Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes

<sup>70</sup> The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure

<sup>71</sup> SMall Molecule Growth 2001 (SMoG2001): An Improved Knowledge-Based Scoring Function for Protein-Ligand Interactions

<sup>72</sup> A Quantum Mechanics-Based Scoring Function: Study of Zinc Ion-Mediated Ligand Binding

<sup>73</sup> Mixed Quantum Mechanics/Molecular Mechanics Scoring Function To Predict Protein-Ligand Binding Affinity

<sup>74</sup> Using a Convenient, Quantitative Model for Torsional Entropy To Establish Qualitative Trends for Molecular Processes That Restrict Conformational Freedom

<sup>75</sup> Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area

<sup>76</sup> Macroscopic models of aqueous solutions: biological and chemical applications

<sup>77</sup> SIMCAP (11.0.0.0)

<sup>78</sup> PLS-regression: a basic tool of chemometrics

used to describe how well a model can reproduce the data under analysis and the predictive ability of the model. Cross-validation was performed by dividing the training sets into 7 groups and developing a number of parallel models for the data devoid of one group. The omitted group then became the test set for the reduced model and residuals for the test set were calculated. A measure of the predictivity of the models, termed predictive residual sum of squares was derived from the sum of squares of these differences for all parallel models. The  $q^2$  value that resulted in the optimum number of components and lowest predictive residual sum of squares was used. The root mean square error of estimation (RMSEE) of the fit for observations in the model and the root mean square error of prediction (RMSEP) were also calculated.

## 2.4 Results and Discussion

### 2.4.1 Ligand and Protein Preparation

A series of 14 X-ray crystal structures of CDK2-ligand complexes with known experimental binding affinities and with resolutions of better than 2.3 Å were downloaded from the PDB<sup>48</sup>. A further 14 ligands from the same chemical series were manually docked to either one of 4 known ligand X-ray structures which had the closest chemical similarity as indicated in Table 1. Details of the proteins and the ligand structures, data set clustering into training and test sets, the resolution of the PDB structures and the experimental binding affinities are detailed in Table 1. The well resolved X-ray crystal structures meant that we were confident of the initial conformation of the complexes. Our experiments focussed primarily in the binding pocket, which for CDK2 is well resolved, particularly the residues which constitute the gate keeper and the Hinge region (residues Phe80 - Gln85 Figure 2). Our rationale for only using minimised X-ray structures as a single protein-ligand structure in preference to the averaging over a number of molecular dynamics snapshots is that this conformation can be considered to contribute significantly to and thus dominate the Boltzmann-averaged potentials for the free energy estimation. This is particularly true when the bound conformation of the ligand corresponds to a particular stable conformation of the unbound ligand. Also, a good single point calculation is more likely to be a good representation of the system than one from which the phase space is poorly sampled. Other studies have used MM/MD simulations ascertain an optimal system conformation before further analysis using FMO<sup>79</sup>. For the 14 X-ray structures examined the average local strain energy (the potential energy of the X-ray structure minus the value of the energy at a near local minimum) was 6 kcal/mol, which is an acceptably reasonable energy<sup>80</sup>. Therefore, it can be argued that the energetic penalties coming from ligand conformational strain are minimal as the ligand is already in a good binding conformation<sup>81</sup>. Using a single-point calculation is also more amenable to virtual screening. The method is not only a comparably accurate alternative to averaged snap shots over a molecular dynamics simulation, but is less time-consuming to setup and compute. It follows then that the remaining 14 ligands can be modelled by slight modifications to the X-ray solved ligands whilst maintaining the geometry of the common chemical scaffold, followed by a minimisation step (see Methods) would be a reasonable approximation of actual binding pose.

Although the structural sampling was not performed in the FMO calculation of the enthalpic contributions to binding free energy, a good correlation ( $r^2$  of 0.68) was obtained to the experimental free energy of binding, Figure 3. The binding pocket in CDK2 is not exposed to solvent and important hydrogen bonding interactions within the active site are of limited flexibility<sup>82</sup>. Under these conditions, enthalpic binding contributes significantly to the free energy of binding, and this therefore accounts for the good correlation to the FMO sum of PIE. For this target it appears that structural sampling is not crucial in order to obtain good correlations. However, the appropriate selection of atomic coordinates is an important factor to obtain well correlated data. A consideration of the optimal binding pose for the modelled ligands was out of the scope of this work, and further validations regarding conformational refinement of docked or aligned poses using the FMO method are in progress.

<sup>79</sup> Correlation Analyses on Binding Affinity of Sialic Acid Analogues with Influenza Virus Neuraminidase-1 Using ab Initio MO Calculations on Their Complex Structures

<sup>80</sup> Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding

<sup>81</sup> Calculation of Protein-Ligand Binding Affinities

<sup>82</sup> A Study of CDK2 Inhibitors Using a Novel 3D-QSAR Method Exploiting Receptor Flexibility

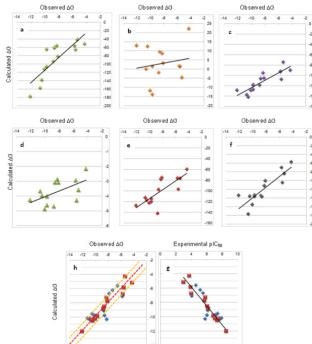


Figure 2.3: Figure 3. Calculated versus observed free energy of binding for 14 CDK2 inhibitors assessed using seven different methods

**Calculated versus observed free energy of binding for 14 CDK2 inhibitors assessed using seven different methods.** Methods include a) FMO (green diamonds),  $r^2 = 0.68$ ; b) GBVI (orange diamonds),  $r^2 = 0.03$ ; c) London dG (purple diamonds),  $r^2 = 0.73$ ; d) Affinity dG (green triangles),  $r^2 = 0.31$ ; e) Alpha HB (red diamonds),  $r^2 = 0.61$ ; f) ASE (black diamonds)  $r^2 = 0.75$ ; and g) QM-based scoring function (red squares) together with the 14 compound test set (blue diamonds) for the QM-scoring function, and h) the calculated versus the experimental  $pIC_{50}$  values for the QM-scoring function,  $r^2 = 0.94$ . For graphs a-f, and g, the line of best fit is shown in black. Graph h shows the line of best fit as a dotted red line and the two dotted yellow lines correspond to 1 log unit boundaries.

#### 2.4.2 Correlation Between MM-Based Scoring Functions and Biochemical Activity

The performance of the FMO method was compared with that of several MM scoring functions implemented in MOE, including the Generalized Born solvation model VI, London dG, Affinity dG, Alpha HB, and ASE scoring functions (Figure 3). In each of these MM methods, the protein was parameterised using the MMFF94x force field and ligand charges calculated using AM1-BCC. The 14 X-ray structures used to build the PLS model were used to compare the 6 scoring functions. The FMO method clearly outperformed three of the scoring functions and was similar to the London dG and the Alpha HB score. A good correlation was observed ( $r^2$  of 0.68) for the FMO sum of PIE and the best performing MM scoring function was the ASE score ( $r^2$  of 0.75). The ASE score has terms for the overlap of the ligand pose with alpha spheres and the overlap between ligand and receptor atom volumes approximated by Gaussians, and therefore can be thought of as mimicking dispersion interactions of ligand binding. As the CDK2 binding pocket is very hydrophobic this generalisation may be sufficient to get a good correlation to experimental binding energy. The Generalized Born solvation model VI failed to correlate the data ( $r^2$  of 0.03). The Affinity dG scoring function only considers enthalpy of ligand binding in a simplistic fashion ( $r^2$  of 0.31). This function is improved by terms to account for hydrogen bonding in the Alpha HB function ( $r^2$  of 0.61). The London dG scoring function has further improvements, adding rotational and translational entropy and a desolvation term which resulted in a good estimation of binding free energy ( $r^2$  of 0.73). The two methods that yield free energy binding predictions close to the actual values are the ASE and the London dG scoring functions.

The main purpose of a scoring function though is to rank binding poses, and here the MOE scoring functions are effective. A Pearson rank order analysis for London dG, Alpha HB and ASE score all gave a value of 0.76, the FMO method performing less well with a Pearson value of 0.64. However, an important consideration in drug development is the identification of active compounds, thus good correlations to experimental binding free energy is of more value than the rank ordering of compounds. To effectively account for other components pertaining to binding additional terms were introduced, the results of which are detailed below.

#### 2.4.3 Data Preparation

The four X-variables used to build and test the PLS model were derived from the sum of the enthalpic contributions  $G^{**}psolv:sub:\backslash$ , the nonpolar solvation term (\|nonascii\_17|\| \*G\*\*npsolv\*\| :sub:), and the entropic term<sup>48</sup>. The PLS model was tested against the 14 modelled complexes.

## 2.4.4 PLS Analysis Results

The optimum number of components in the PLS model was two which gave a very high  $q^2$  of 0.896, and the RMSEE of the fit for observations was 0.632. The  $r^2$  value was 0.939 for this model. The 4 X-variables contributed similarly to the model, and there were no outliers in the observations used to build the model. The model rank orders the compounds extremely well, with a Pearson correlation of 0.97. This robust model predicted the test set well, the  $r^2$  of the test set was 0.824, and the RMSEP was 1.005. The plot of computed and experimentally determined binding free energies is shown in Figure 3 and together with the residual differences between these values for each of the ligands is shown in Table 2. The majority of the data (Figure 3) lies within the two yellow-dashed lines, indicating errors of less than 1 order of magnitude. There are two ligands, **12** and **15** that fall well outside this boundary. An examination of the 4 components which make up the free energy term does not reveal any strong trend resulting in the large residual values. All the ligands used to train and test the model were well within the 95% confidence intervals for the predicted Y-values and there were no observations that deviated significantly from the model in X-space.

## 2.4.5 QMbased Scoring Function

FMO has been used previously to generate a charge transfer term for a quantitative structure-activity relationship (QSAR) model<sup>44</sup>. Here, we aimed at producing a QM-based scoring function which would take into consideration complex binding interactions, solvation effects and ligand binding entropy on a timescale amenable to drug discovery. The FMO methods allows for accurate treatment of charge transfer and polarisation effects. It has been noted previously that the majority of polarisation energy is within 5 Å of a ligand<sup>83</sup>. This observation justifies the 4.5 Å residue inclusion radius used to describe the binding pocket and allows for this polarisation to be incorporated into the enthalpy of binding energy term. The contribution of charge transfer effects on ligand binding have already been mentioned, and represent an important addition to a scoring function particularly when examining particular ligand-residue interactions<sup>44</sup>. However, the contribution of charge transfer on the enthalpic binding term is dependent on the wave function used. The FMO contribution to the binding free energy has a very broad range (-28 to -178), this may be a result of using the MP2 method which is known to overestimate charge transfer interactions<sup>33</sup>. Energy decomposition analysis from the FMO calculation reveals that the majority of the energy comes from the charge transfer contribution of charged atoms. The approximations for other terms in the scoring function make this overestimation less significant compared to the absolute binding energy determined by the FMO method when used in isolation.

The binding free energy is a combination of enthalpic and entropic terms. Indeed, a thorough understanding of enthalpy/entropy compensation is needed to accurately predict binding energies<sup>8485</sup>. Ligand conformational entropy contributions are also significant, and neglecting this will adversely affect binding energy predictions<sup>86</sup>. As a very simplistic method to account for this we chose to examine how the number of rotational bonds in the ligand would influence the predicted binding free energy. The good correlation obtained with our data, in this test case, indicates that this extremely fast method is adequate for this purpose. More detailed studies of entropy could be performed by normal mode analysis of molecular dynamics simulations. The lack of an adequate protein entropy term can result in an overestimation of binding free energy, and more work is needed to examine the effect of this on such calculations.

The solvation free energy is divided into polar and nonpolar terms. The nonpolar term is dependent upon the size of the ligand, which is scaled by the two constants  $\gamma$  and  $b$ . This scaling makes the nonpolar term small and negative, allowing the polar terms to dominate the solvation free energy of binding. Recently, the polarisable continuum model (PCM) implemented in the GAMESS program was used to calculate solvation energies and were compared to those obtained with PB+SASA<sup>42</sup>. It was found that PCM exaggerated the nonpolar contribution substantially, and therefore a QM treatment of solvation was not advantageous. Solvation calculations with the PCM have been implemented with the FMO method<sup>87</sup> and although this does increase computational time, more accurate treatment of solvation from a single point calculation would be possible. Solvation effects have also been developed for FMO using the PB equation

<sup>83</sup> The effect of MM polarization on the QM/MM transition state stabilization: application to chorismate mutase

<sup>84</sup> Compensating Enthalpic and Entropic Changes Hinder Binding Affinity Optimization

<sup>85</sup> Do enthalpy and entropy distinguish first in class from best in class?

<sup>86</sup> Ligand configurational entropy and protein binding

<sup>87</sup> The polarizable continuum model (PCM) interfaced with the fragment molecular orbital method (FMO)

to account for ion concentrations<sup>88</sup>. Further studies need to be performed to assess the effects these advances in treating solvation effects brings to the determination of binding free energy and the computational cost of the method.

## 2.5 Conclusions

The results show that a single point QM calculation using the FMO method gives a good correlation to experimentally determined free energies of ligand binding calculated from ligand potencies. The FMO method outperformed 3 other methods used to estimate the free energy of binding by MM-based methods. We conclude that the additional terms which treat charge transfer, polarisation and dispersion effects during ligand binding in this QM method significantly improves the estimation of ligand potency compared to MM-based procedures. Methods were then introduced to further improve upon the initial estimates. This paper presents the first attempt to calculate ligand-binding free energies using a combination of high-level *ab initio* FMO methods together with PBSA techniques to derive reasonable estimations of enthalpy, entropy and solvation energies. We used a PLS QSAR model to correlate the 4 components of our scoring function to build a model which was very robust and highly predictive. The data set was composed of ligands from a lead development program that resulted in a clinical candidate against CDK2, thus testing the QSAR model against a range of ligand potencies. The need to run a PLS model stems from the poor absolute prediction of free binding energies and therefore the need to adjust the data. This QM-based scoring function represents a new protocol to estimate ligand potencies in a congeneric series of compounds whereby single point changes can be performed on a known X-ray crystal structure to guide medicinal chemistry.

## 2.6 Competing interests

The authors declare that they have no competing interests.

## 2.7 Authors' contributions

MPM and OI ran the FMO calculations and MPM wrote the scripts for FMO analysis and developed the QSAR method. MPM and OI tested the presented methods and prepared the manuscript for this publication. RJL supervised the project. All authors have read and approved of the final manuscript.

## 2.8 Acknowledgements

The authors acknowledge Dmitri Fedorov for his support in implementing the FMO method.

<sup>88</sup> Incorporation of solvation effects into the fragment molecular orbital calculations with the Poisson-Boltzmann equation



# JCOMPOUNDMAPPER: AN OPEN SOURCE JAVA LIBRARY AND COMMAND-LINE TOOL FOR CHEMICAL FINGERPRINTS

## 3.1 Abstract

### 3.1.1 Background

The decomposition of a chemical graph is a convenient approach to encode information of the corresponding organic compound. While several commercial toolkits exist to encode molecules as so-called fingerprints, only a few open source implementations are available. The aim of this work is to introduce a library for exactly defined molecular decompositions, with a strong focus on the application of these features in machine learning and data mining. It provides several options such as search depth, distance cut-offs, atom- and pharmacophore typing. Furthermore, it provides the functionality to combine, to compare, or to export the fingerprints into several formats.

### 3.1.2 Results

We provide a Java 1.6 library for the decomposition of chemical graphs based on the open source Chemistry Development Kit toolkit. We reimplemented popular fingerprinting algorithms such as depth-first search fingerprints, extended connectivity fingerprints, autocorrelation fingerprints (e.g. CATS2D), radial fingerprints (e.g. Molprint2D), geometrical Molprint, atom pairs, and pharmacophore fingerprints. We also implemented custom fingerprints such as the all-shortest path fingerprint that only includes the subset of shortest paths from the full set of paths of the depth-first search fingerprint. As an application of jCompoundMapper, we provide a command-line executable binary. We measured the conversion speed and number of features for each encoding and described the composition of the features in detail. The quality of the encodings was tested using the default parametrizations in combination with a support vector machine on the Sutherland QSAR data sets. Additionally, we benchmarked the fingerprint encodings on the large-scale Ames toxicity benchmark using a large-scale linear support vector machine. The results were promising and could often compete with literature results. On the large Ames benchmark, for example, we obtained an AUC ROC performance of 0.87 with a reimplementation of the extended connectivity fingerprint. This result is comparable to the performance achieved by a non-linear support vector machine using state-of-the-art descriptors. On the Sutherland QSAR data set, the best fingerprint encodings showed a comparable or better performance on 5 of the 8 benchmarks when compared against the results of the best descriptors published in the paper of Sutherland et al.

### 3.1.3 Conclusions

jCompoundMapper is a library for chemical graph fingerprints with several tweaking possibilities and exporting options for open source data mining toolkits. The quality of the data mining results, the conversion speed, the LPGL software license, the command-line interface, and the exporters should be useful for many applications in cheminformatics like benchmarks against literature methods, comparison of data mining algorithms, similarity searching, and similarity-based data mining.

## 3.2 Background

The decomposition of a chemical graph into a list of features is a convenient way to assess the similarity between chemical compounds by comparing the resulting lists of features. Such representations are also called chemical fingerprints<sup>1</sup>. These encodings are important for data mining applications like similarity-based machine learning approaches or similarity searches<sup>2</sup>.

The goal of this work is to introduce an open source molecular fingerprinting library for data mining purposes which provides exact definitions of its fingerprinting algorithms. The algorithms can be parametrized with various options to adapt the encodings, for example, by applying a custom labeling function or by altering the search depth parameter. Additionally, the library can be used as a basis for new implementations. It is based on the Chemistry Development Kit<sup>3</sup>, which also provides several fingerprints in its API. However, there are several differences. The first aim of jCompoundMapper is to focus on the exact definition of its encodings, which is crucial to describe the features in data mining experiments. The second aim is to provide the functionality to export the fingerprints or pairwise similarity matrices to formats of popular machine learning toolboxes. A label or property of an input compound to be trained by a machine learning algorithm can be included.

Most fingerprint algorithms rely on either the geometrical or the topological distance between the atoms of a structure. The topological information is stored in the all-shortest path matrix, which encodes the minimum topological distance between two atoms (vertices) by the shortest path using the bonds (edges). Organic compounds are usually weakly connected because the number of covalent bonds (vertex degree) of an organic molecule is limited. In contrast, the geometry of a structure can be interpreted as a fully connected graph. The complexity of both approaches can be reduced by limiting the search depth for topological fingerprints or by introducing a distance cut-off for geometrical fingerprints.

jCompoundMapper offers a variety of topological (e.g. radial atom environments<sup>4</sup>, extended connectivity fingerprints<sup>5</sup>, depth-first search fingerprints<sup>6</sup>, or auto-correlation vectors<sup>7</sup>) and geometrical (e.g. two-point and three-point encodings<sup>8,9</sup> or geometrical atom environments<sup>10</sup>) fingerprints. If applicable, it allows for a parameterization of an encoding, such as the search depth, the distance cut-off, the geometrical scaling factor, the atom typing scheme, or the hash space.

After the feature generation step, the list of features can be mapped to a vectorial format. One possibility is to encode a set of features as a hashed fingerprint. Here, a unique identifier of a feature is used to initialize a pseudo random number generator which produces numbers in  $h$  is the maximum size of the hash space. Thus, the dimensionality of the original feature space can be considerably reduced. For example, the Fingal fingerprint<sup>11</sup>, uses the cyclic redundancy check algorithm to generate seeds for the hashing of chemical graph patterns. For an introduction into hashed fingerprints, please refer to the review by Brown<sup>1</sup>. Another strategy reserves fixed bit positions in a vector for

<sup>1</sup> Chemoinformatics - An Introduction for Computer Scientists

<sup>2</sup> Chemical Similarity Searching

<sup>3</sup> The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics

<sup>4</sup> Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance

<sup>5</sup> Extended-connectivity fingerprints

<sup>6</sup> Graph kernels for chemical informatics

<sup>7</sup> Alignment-free Pharmacophore Patterns - A Correlation Vector Approach

<sup>8</sup> Atom Pairs as Features in Structure-Activity Studies: Definition and Applications

<sup>9</sup> The Pharmacophore Kernel for Virtual Screening with Support Vector Machines

<sup>10</sup> Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D)

<sup>11</sup> Fingal: A Novel Approach to Geometric Fingerprinting and a Comparative Study of Its Application to 3D-QSAR Modelling

specific feature types, like patterns obtained at a certain parameter (such as depth or distance) with a limited number of possible combinations. The definition of the CATS2D<sup>7</sup> vector is an example for this approach.

jCompoundMapper supports native formats of common open source machine learning libraries. The exporters can be used to write feature maps to comma-separated format, LIBSVM<sup>12</sup> format (sparse and matrix), and WEKA ARFF<sup>13</sup>. Therefore, various data mining libraries can be directly applied on the output files. Furthermore, the library provides efficient data structures to compare sets of features in the case that the computation of a similarity matrix is required.

The quality of the encodings was compared on QSAR and toxicity benchmark problems in the results section. First, we conducted experiments using the support vector regression of LIBSVM on the well-known Sutherland QSAR benchmark set<sup>14</sup>. Second, we used a large Ames toxicity classification benchmark<sup>15</sup> and LIBLINEAR<sup>16</sup> to evaluate the performance using binary hashed sparse fingerprints. On the Sutherland data sets, the averaged squared correlation of the all-shortest-path and the atom triplet fingerprint was at least 5% better on ACHE than the best encoding given by Sutherland et al. On BZR and DHFR, the all-shortest path fingerprint achieved a squared correlation of 0.57 and 0.76 respectively. The performance was comparable on two data sets. On the remaining three data sets, the best encoding was more than 5% worse than the results of the best encoding published by Sutherland et al. On the Ames toxicity data set, the implementation of the extended connectivity fingerprint achieved an AUC ROC performance of 0.87, which is comparable to the performance by a non-linear support vector machine trained on state-of-the-art descriptors. Nevertheless, the goal was not an exhaustive comparison but to show that the implementations are able to obtain similar results when compared against literature results. jCompoundMapper features a command-line interface but can also be used as a Java API. It depends solely on open source libraries and is licensed under the LPGL. The source code and an executable is available at Sourceforge.

The library originated from various implementations of literature fingerprints and descriptors used in comparison studies. The encodings were employed either as part of a new approach or as a reference method<sup>1718192021</sup>.

To sum up, jCompoundMapper is an open source library for the encoding of chemical graphs as fingerprints. It can be used from the command-line interface or as a Java API. Hence, a further use in applications, like in KNIME <http://www.knime.org> nodes, is possible. The overall performance of the fingerprints in machine learning experiments indicates that structured-based models of reasonable quality can be obtained.

## 3.3 Methods

### 3.3.1 Prerequisites

#### Notation

The binned geometrical distance matrix  $g_{ij}$  Gencodes the spatial distances between two heavy atoms. The topological distance matrix  $t_{ij}$  Tencodes the shortest topological distance between atoms  $i$  and  $j$ . The labeling function  $d$  defines a distance cut-off for features, all features with  $g_{ij} > *d*$  or  $t_{ij} > d$  are omitted. A labeled path  $p$  is a sequence of atoms connected by bonds  $b$ ; connects  $p**ij :sub:\$  denotes a path connecting the  $*i*$ th atom with the  $*j*$ th atom. The depth  $*d*$  for topological patterns is the maximum number of bonds allowed for connecting the first atom with the last atom. Analogously to the definition of topological paths, a geometrical pattern must consist of different atoms, i.e. for two atoms  $*a**i*\ :sub:, a**j :sub:$  “it holds that  $i \sim j$ . Finally,  $a \sim b$

<sup>12</sup> NOTITLE!

<sup>13</sup> The WEKA Data Mining Software: An Update

<sup>14</sup> A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships

<sup>15</sup> Benchmark Data Set for in Silico Prediction of Ames Mutagenicity

<sup>16</sup> LIBLINEAR: A Library for Large Linear Classification

<sup>17</sup> Estimation of the applicability domain of kernel-based machine learning models for virtual screening

<sup>18</sup> Graph kernels for chemical compounds using topological and three-dimensional local atom pair environments

<sup>19</sup> Chronic Rat Toxicity Prediction of Chemical Compounds Using Kernel Machines

<sup>20</sup> Optimal Assignment Methods for Ligand-Based Virtual Screening

<sup>21</sup> Probabilistic Modeling of Conformational Space for 3D Machine Learning Approaches

is defined as the concatenation of alphanumerical string symbols separated by an unique delimiter. In the following, we assume a hydrogen-depleted molecular graph  $C$  with  $n$  atoms.

An encoding algorithm  $F$  has the form

where an encoding algorithm  $F$  maps some compound  $C$  to a set of features  $X$ .  $m$  depends, with the exception of fixed-vector fingerprints, on  $C$ . A feature  $f$  has an unique  $f.nom$ .  $f.id$  does not necessarily depend on  $f.nom$ . However, in most cases it is convenient to use a hash code of the string representation of a feature.

## Fundamental Matrices

The geometrical distance matrix  $g_{ij}$  is computed as the matrix of binned Euclidean distances in Ångstrom between the three-dimensional coordinates of all atom coordinates  $a \cdot c^3$ , multiplied by a scaling factor  $s^+$ . The scaling factor  $s$  influences the resolution of the geometry and should be chosen according to the size of the compounds. The entry  $i, j$  ( $g_{ij}$ ) in the matrix is calculated as follows

The computation time for the geometrical distance matrix is quadratic. The binning of the real-valued geometrical distance is important to produce discrete features  $x_f, y_f$ , which can be compared by the Dirac function

The topological distance matrix is defined as  $T_{ij}$ . The element  $i, j$  contains the shortest path between the  $i^{th}$  and the  $j^{th}$  atom ( $:sub: 'ij't$ ).  $:sub: 'ij'$  is computed by the Floyd-Warshall algorithm. Therefore, the computation time for the matrix is  $*O*(n^3)$ .

## Feature Extraction

The molecular similarity is based on the numerical identifiers  $x.id$  for a feature  $x$ . Two features are regarded as equal if  $x.id = y.id$ . In all implementations the features of a compound  $C$  are distinguishable by recurrence, which means that we include a feature if the id of a feature is different from the previously extracted features. If a feature with the same id is generated again, the count for the feature is incremented. All atom pair encodings are extracted by regarding the upper half of the distance matrix only. For each atom pair, the string representation is generated in both reading directions. Only the version with the greater hash code is included in the final set of descriptors.

The modified depth-first search applied in this library generates all possible paths originating from a root atom. Therefore, the feature space can be approximated by an  $m$ -ary tree and is therefore  $O*(:sup:'d'nm)$ , where  $n$  is the number of heavy atoms and  $m$  the number of children in an  $m$ -ary tree,  $d$  is the depth of the tree. In organic compounds, every atom has at most 4 neighbors ( $m = 4 - 1$  because one of the neighboring bonds has already been visited). Thus, the hypothetical worst case has a complexity of  $O*(n^3:sup:'d')$  at a search depth of  $d$ . If we assume an average branching factor  $|nonascii\_8|$ , which is slightly above 1 for organic compounds [#B6], the depth-first search has a complexity of  $*O*(:sup:'d'n|nonascii_9|)$ . The average branching factor depends on the average degree of a vertex, which is about 2 in organic molecules. We define  $*DFS*(:sub: 'i'a, *d)$  as the set of all possible paths originating from a root atom  $i$  with a depth up to  $d$ .

For some of the definitions, we defined a *can* function that maps a set of features to a single canonical pattern. In an implementation this function can be realized by first sorting the patterns, which is possible if a natural order can be defined on the features. Then, the list of sorted patterns can be merged to a single canonical representation.

## Atom Types and Pharmacophore Types

jCompoundMapper applies the standard atom types and ring detection algorithms implemented in the CDK. There are various typing schemes for small drug-like compounds described in the literature. In the current version, jCompoundMapper features the following typing schemes as labeling function

1. Element symbol (e.g. C, O, N, ...)
2. CDK atom types (e.g. C.sp2, O.minus, N.amine, ...)

3. Element plus the number of neighboring heavy atoms (e.g. C.2, O.1, N.2, ...)
4. Element plus ring type plus the number of neighboring heavy atoms (e.g. C.r.2, C.a.2, O.1, N.2, ...) where  $r$  is an arbitrary ring, and  $a$  is an aromatic system. If  $i$  is not contained in a ring, no ring type is set. The precedence is  $a > *r*$ .
5. Daylight-Invariants (plus optional ring flag) have the following properties, separated by a dot: Atomic number, number of heavy atom neighbors, valency minus the number of connected hydrogens, atomic mass, atomic charge, number of connected hydrogens, and a flag if the atom is member of at least one ring. (e.g. 6.2.3.12.0.1.1 for a carbon in a benzole ring)

The following listing of potential pharmacophore points (PPPs) was published by Renner et al.<sup>7</sup> for the CATS autocorrelation descriptors. If PPP atom types are needed, this list is parsed and matched with the structure using the CDK SMARTS matcher or specially implemented graph searches.

1. Hydrogen-bond donor (D): [#6H] oxygen atom of an OH-group; [#7 H,#7H2] nitrogen atom of an NH or NH<sub>2</sub> group
2. Hydrogen-bond acceptor (A): oxygen atom [#6]; [#7H0] nitrogen atom not adjacent to a hydrogen atom
3. Positive (P): [\*+] atom with a positive charge; [#7H2] nitrogen atom of an NH<sub>2</sub> group
4. Negative (N): [\*-] atom with a negative charge; [C&\$((C(=O)#8H1), P&\$((P(=O)O), S&\$((S(=O)O)) carbon, sulfur or phosphorus atom of a COOH, SOOH, or POOH group (SMARTS replaced by a direct graph search)
5. Lipophilic (L): [Cl, Br, I] chlorine, bromine, or iodine atom; [S;D2;\$(S(C)(C))] sulfur; atom adjacent to exactly two carbon atoms; sulfur atom adjacent to only carbon atoms (SMARTS replaced by a direct graph search)

### 3.3.2 Encodings

#### Topological Fingerprints

All encodings described in the following section rely on the  $d$  parameter which constrains the maximum topological distance allowed between two atoms  $i, j$  in a feature.

#### All-Path Encoding (DFS)

All-path encodings are paths generated by a graph traversal with a modified depth-first search as proposed by Ralaivola et al.<sup>6</sup>. The linear fragments are obtained by iterating over all atoms in a molecular graph and performing an exhaustive search up to a predefined depth  $d$ . To generate an unique representation for each path, a temporary path object is generated and mapped to two alphanumerical string representations by generating the original and reverse string representation of the corresponding path object. The version with the higher lexicographical order is stored.

#### All-Shortest Path Encoding (ASP)

The ASP encoding equals the DFS encoding with the exception that only the paths from an atom are stored that have shortest distances from the root atom to the last atom contained in the path, which leads to a sparser representation. During the depth-first search, all paths are removed from the temporary set of depth-first search paths that do not fulfill this constraint. To incorporate this information the  $T_{ij}$  matrix is required. Let  $\text{len}(\text{path})$  be the length (number of bonds between the  $i^{th}$  and the  $j^{th}$  atom in a path) of a path between atoms  $i, j$ , then the set of features  $F$  encoding a compound  $C$  is

Thus, the all-shortest path encoding is a subset of the paths contained in the DFS fingerprint. It is similar to topological atom pair approaches<sup>8</sup> with the exception that all-shortest paths between two atoms are explicitly stored. Borgwardt

et al. proposed a graph kernel based on the set of all-shortest paths<sup>22</sup>, however, only the vertex pairs and their shortest-path distances were included in this work. The explicit generation of paths is necessary because the Floyd-Warshall algorithm computes only the shortest distances between two vertices.

### Topological Atom Pairs (AP2D)

This encoding contains atom types and the shortest path distance information between all pairs of atoms. It can be directly extracted from  $t_{ij}$  by converting the pattern  $i \ j$  to a string feature. The features are canonicalized by generating the patterns from both reading directions and storing the version with the higher lexicographical order.

For 2-point patterns this can be easily conducted by regarding the upper half of the distance matrix  $t_{ij}$  only (i.e.  $i > *j*$ ).  $O(n^2)$  is needed for the generation of features and  $O(n^3)$  for the computation of  $t_{ij}$ .

Thus, the total computation time is cubic.

### Topological Atom Triplets (AT2D)

This encoding extends the AP2D encoding by a further atom. The set of patterns consists of atom triplets and the topological distance to the next atom in the feature  $i \ j \ k$  and  $t_{ij}$ ,  $t_{jk}$ ,  $t_{ki}$  and  $d$ . The total computation time for the features is cubic because all possible combinations of heavy atoms  $a**i:sub:\backslash , *a**j*\backslash :sub:, a**k :sub:\textcolor{red}{“} have to be considered in the worst case.$

### Topological Autocorrelation Keys (CATS2D)

The CATS2D descriptors encode the pairwise topological relationships of PPP patterns in a molecular graph by a vector of fixed size. The approach was described by Schneider et al.<sup>23</sup>, a list of PPP patterns was presented by Renner et al.<sup>7</sup>. The PPPs are defined in the subsection (“Atom Types and Pharmacophore Types”). The combination of all points leads to 15 possible pairs. The pairs are mapped to a key with a fixed dimensionality. The position of a feature in the key is determined by the index for the corresponding pattern pair of PPPs plus the topological distance. For example, this means the bit 76 in the CATS vector with  $d = 9$  belongs to the PPP pair DN, and contains the number of pairs with distance 6. In the original publication of the topological CATS descriptors,  $d = 9$  was used as the distance cut-off for the topological search, resulting in a vector with 150 dimensions. In our implementation, the search depth can be adjusted by altering the  $d$  parameter. The resulting vector has  $(d + 1) \cdot 15$  dimensions. The complete list of possible PPP pairs in the vector (block index in parentheses) is: AA (0), AD (1), AL (2), AN (3), AP (4); DD (5); DL (6); DN (7); DP (8); LL (9), LN (10), LP (11), NN (12), NP (13), PP (14). Let  $F(C) = X$  be a decomposition into all valid PPP pairs. Then, the CATS2D vector is

where  $\text{offset}(p)$  returns the predefined start index for the pattern  $p$ ,  $d**D_2(p)$  returns the topological distance between two atoms in the PPP pair, and  $\text{nonascii\_15}:sub:\text{‘}(*X)$  counts the number of occurrences of a pattern  $p$  in  $X$ .

### Pharmacophore Pair and Triplet Encodings (PHAP2PT2D, PHAP3PT2D)

The PHAP2PT2D encoding is computed similarly to the AP2D. However, instead of atom types, the information of all PPPs of an atom is used to generate the fingerprint. Thus, we have to iterate over all PPPs of an atom. Analogously, the PHAP3PT2D encoding is computed, which uses three points. To keep the notation simple, let  $\text{P}_i$  denote the set of valid PPPs for the  $*i$ th atom. Then, the set of valid 2-point pharmacophores is

and the set of valid 3-point pharmacophores is defined as

where  $t_{ij}$ ,  $t_{jk}$ ,  $t_{ki}$  and  $d$ . Actually, there are three additional inner loops over all valid pharmacophore points at atoms  $i, j, k$ . The complexity of these inner loops is theoretically  $5^3$  because the cardinality of the set of PPPs is 5. However, this

<sup>22</sup> Protein function prediction via graph kernels

<sup>23</sup> Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening

complexity is further reduced because the meaning of some PPP definitions is contradicting for some combinations, such as “atom is positively charged” and “atom is negatively charged”. The overall complexity is  $O^*(n^3)$  because of the constant computation time of the inner loops.

### SHED Key (SHED)

The SHED Keys are closely related to the pharmacophore atom pair based encodings, with the following major differences: First, the number of dimensions is fixed, second the entries do not describe a count but the entropy of the respective atom pair descriptor<sup>24</sup>. The implementation differs slightly from the original implementation because it utilizes the PPP definitions as described by Renner et al.<sup>7</sup>. The distribution is analyzed for all possible combinations of PPPs. From that distribution of pairwise features the Shannon entropy is calculated as the descriptor in the corresponding PPP pair dimension of the SHED Key. The Shannon entropy of a PPP pair  $PPP_i$  is defined as

where  $PPP_i(l)$  denotes the  $i^{th}$  PPP pair that is separated by a topological distance  $*l$ . If pattern  $i$  was not found, the value of the  $i^{th}$  dimension is set to 0. The distribution of a PPP pair is calculated by regarding the different distances  $l_1, l_2, \dots, *l, \dots, d$ . The resulting vector has 15 real-valued entries.

### Extended Connectivity Fingerprints (ECFP)

We implemented a variant of the ECFP as described by Rogers and Hahn<sup>5</sup>. Each ECFP feature represents a circular substructure around a center atom. The algorithm starts with the initial atom identifier of the center atom and grows a circular substructure around this atom throughout a defined number of iterations (search depth). For each round, the current extended version of the feature is added to the final set of features. In contrast to other radial fingerprints, the bonding information is included. Therefore, a feature can be extracted, for example, as canonical SMILES.

The current implementation of the ECFP in jCompoundMapper differs slightly from the original implementation. In the original algorithm, the identifiers of the alpha atoms of a center atom are used to calculate an updated identifier for the center atom. The algorithm only includes the alpha atoms of a center atom in each iteration and thus the connectivity information is completely discarded between the layers. However, the identifier of a center atom implicitly contains information from further and further away of the center atom in each iteration because the atom identifiers of the previous iteration are used. We explicitly model the growing substructure by using the initial atom identifiers in each iteration and keeping the connectivity information between the layers. After an iteration, new possible attachment points for a specific circular substructure are kept in memory and those attachment points are extended in the next iteration.

### Topological Molprint-like fingerprints (RAD2D)

This encoding was proposed by Bender et al.<sup>425</sup> and describes the radial environment by the atoms with the topological distance  $1, 2, \dots, l, \dots, d$  rather than the full paths containing bonds. A shell  $s*(:sub:'i'a):sub:'l'$  in our implementation contains the canonically sorted set of topological neighbors of atom  $:sub:'i'$  at a distance  $:sub:'ij'$   $t = *l$ . Additionally, we include the concatenation of all shells  $1, 2, \dots, l, \dots, d$  as additional features. Therefore, a resulting set of features contains  $n \lnonascii{18} d$  features.

### Local Path Environments (LSTAR)

This fingerprint is a radial fingerprint similar to RAD2D. The major difference is that all paths up to depth  $d$  are stored in a shell. First, the tree of all paths originating from an atom  $i$  is generated. Then, all paths of a certain length are assigned to a shell  $s*(:sub:'i'a):sub:'d'$  containing the paths originating from root atom  $:sub:'i'$  of length  $*d$ . This

<sup>24</sup> SHED: Shannon Entropy Descriptors from Topological Feature Distributions

<sup>25</sup> Screening for Dihydrofolate Reductase Inhibitors Using MOLPRINT 2D, a Fast Fragment-Based Method Employing the Naive Bayesian Classifier: Limitations of the Descriptor and the Importance of Balanced Chemistry in Training and Test Sets

is equal to a canonical representation of  $DFS^*(\ast a \ast \ast i, :sub: \ast d)$  in a single canonical feature. The paths in a shell are sorted in lexicographical order to be comparable. The resulting fingerprint contains all shells  $1, 2, \dots, l, \dots, d$ . The major difference to the Molprint-like fingerprints is that the bond information is still included.

## Geometrical Fingerprints

All geometrical encodings support the  $d$  parameter which defines the distance cut-off between two atoms. Another important parameter is the scaling factor  $s$ , as described at the beginning of this section.

### Geometrical Atom Pairs and Atom Triplets (AP3D, AT3D)

These encodings are implemented similarly as their topological pendants AP2D and AT2D. The only difference is that  $g_{ij}$  is used for the distance information. Thus, the geometrical two-point atom pair encoding (AP2D) is defined as where  $i \neq j$  and  $g_{ij} \leq d$ .

For the three-point relationships AT3D, we have

where  $i \neq j \neq k$  and  $g_{ij}, g_{jk}, g_{ki} \leq d$ .

This is a standard encoding implemented in several toolkits; a kernel based on such patterns was published by Mahé et al<sup>9</sup>.

### Geometrical CATS fingerprints (CATS3D)

Our implementation differs from the description of the original CATS3D<sup>7</sup>, which uses the Molecular Operating Environment (MOE, Chemical Computing Group, <http://www.chemcomp.com/>) patterns to depict surface features of a molecule. The version implemented in jCompoundMapper uses the PPP definitions which were also used in the implementation of the CATS2D vector. Again, let  $F^*(\ast C) = X$  be the set of all valid PPP pairs and  $\text{lenonascii}_26|\ast :sub: 'p'(\ast X)$  a function which counts a pattern  $p$  in  $X$ . Then the CATS3D vector is

where  $d \ast D_3(p)$  returns the geometrical distance of the two atoms, which equals  $g_{ij}$  between any atoms  $i, j$  contained in a feature.

### Geometrical pharmacophore fingerprints (PHAP2PT3D, PHAP3PT3D)

These fingerprints are derived from their topological variants PHAP2PT2D and PHAP3PT2D by replacing the  $T_{ij}$  matrix by  $g_{ij}$ . Let  $P_i$  denote the set of valid PPP for the  $i^{th}$  atom then  $:sub: 'i' P \text{lenonascii}_28| :sub: 'ij' g \text{lenonascii}_29| :sub: 'j' P$  is a valid two-point pharmacophores and  $:sub: 'i' P \text{lenonascii}_30| :sub: 'ij' g \text{lenonascii}_31| :sub: 'j' P \text{lenonascii}_32| :sub: 'jk' g \text{lenonascii}_33| :sub: 'k' P \text{lenonascii}_34| :sub: 'ki' g$  is a valid three-point pharmacophore, where  $:sub: 'ij' g, :sub: 'jk' g, :sub: 'ki' g \text{lenonascii}_35| *d$ .

The set of valid 2-point pharmacophores is defined as

and the set of valid 3-point pharmacophores is defined as

Again,  $P_i$  denotes the set of valid PPPs for the  $i^{th}$  atom.

### Geometrical Molprint-like fingerprints (RAD3D)

These encodings are the geometrical pendant of the topological RAD2D encoding. Similar to the RAD2D encoding, the atoms with  $g_{ij} \leq l$  at a binned geometrical distance are added to a shell descriptor. For each value in  $1, 2, \dots, l, \dots, d$  a pattern containing all shells up to distance  $l$  in a canonical order is created. Therefore, the encoding contains  $n \cdot d$  entries.

## Example of encodings

Comparison Table 1 gives a direct tabular comparison of the features extracted by the encodings together with their count or value. The features are generated from the compound presented in Figure 1.

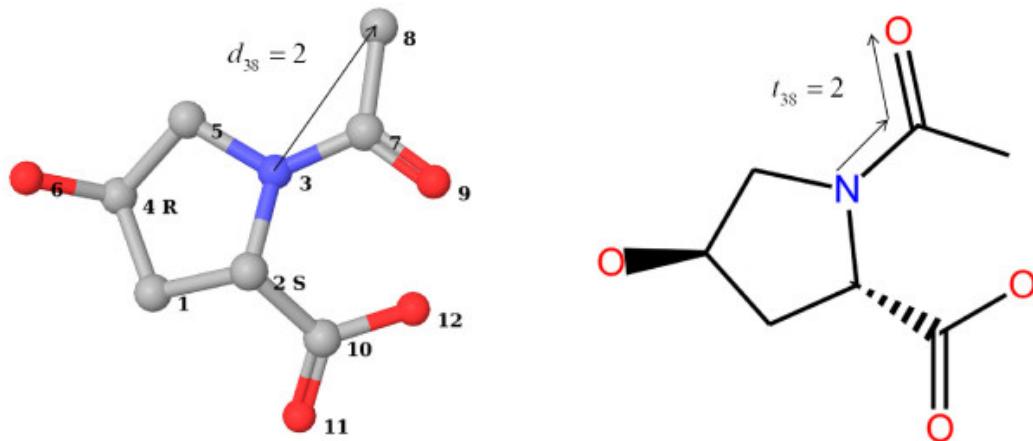


Figure 3.1: Figure 1. Topology and Geometry of Oxaceprol

**Topology and Geometry of Oxaceprol.** The geometry and topology of Oxaceprol. Pharmacophore types shown in the 3 D structure are 1 = [L], 3 = [A - [#7H0]], 6 = [D - [OH], A - [O]], 8 = [L], 9 = [A - [O]], 10 = [N], 11 = [A - [O]], 12 = [D - [OH], A - [O]]. The geometry and topology of this compound is the basis for the exemplary fingerprints shown in Table 1. Note that multiple PPPs can be assigned to an atom: In this case atoms 6 and 12 have two valid PPPs.

## 3.4 Implementation

### 3.4.1 Third-party libraries

The underlying chemical expert system is the Chemistry Development Kit (CDK)<sup>326</sup> in its current development version 1.35. It provides the basic functionality for parsing, typing, and graph algorithms for molecular data. For the command-line interface we employed the Apache Commons command-line parser 1.2 <http://commons.apache.org/cli/>. The access via the API or the binary using the command-line interface enables the user to utilize the library for batch processing. The language level is Java 1.6.

### 3.4.2 Additional functionality

#### Import and Export of Data

The valid input format is MDL SD format with attached hydrogens for the command-line tool. The CDK molecule objects can be processed using the API.

There are exporters for various formats of popular machine learning toolboxes. The ARFF format is the native WEKA<sup>13</sup> format, the support vector machine libraries LIBSVM and LIBLINEAR are supported by their sparse hashed format and precomputed matrix format. Alternatively, there are several comma-separated formats which support hashed or string features, which can be imported into toolboxes like R or MATLAB.

jCompoundMapper includes a buffered random access reader for parsing the input files. Thus, it can read files of the maximum size supported by the Java runtime environment. The memory requirements are low if the encodings are

<sup>26</sup> Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics

exported sequentially (such as the sparse LIBSVM format) because only a single encoding has to be stored at a time. If the computation of a similarity matrix is required, all encodings are kept in memory to ensure a fast computation of similarities. For this reason, a matrix computation requires additional memory for large data sets.

The label or class for learning tasks is read from the SD property and is integrated into the specific output format. As for the ARFF format, a nominal or numeric class label is created, depending on the distribution of labels in the input format. The user may overwrite the default threshold for the number of classes (currently, this is set to five).

## Hashing

All decomposition algorithms  $F^*(^*C)$  return the full list of features (the encoding). A feature  $f$  has an integer identifier  $f.^*id^* \{0, 1, \dots, 2^{32}\}$  which allows the efficient use in hash based collections. Therefore, it is possible to operate on the full set of descriptors and to generate hashed fingerprints. Hashing is useful to generate binary vectors of a fixed size. The hash function  $H$

is used to project the set of features to a binary vector of the dimension  $h$ . The size depends on the expected number of features (see Table 1).  $H$  generates the hashed bit of a pattern depending on the numerical seed  $f.id$  assigned to each feature. In most cases, the seed equals a hash code of the string representation. The hashing step is also useful to obtain nominal features for fast comparisons. A nominal feature is a feature  $f$  with a finite set of possible values like  $f \{\text{red, green, blue}\}$  or, convenient for chemical graphs,  $f \{\text{pattern included, pattern not included}\}$ .

## Similarity Matrices

jCompoundMapper offers a<sup>27</sup>. Thus, it is possible to compute distance matrices within seconds on an average desktop computer. Now, we assume two mappings  $F(A_C) = A$  and  $F(B_C) = B$ . Further, let  $p\phi(X)$  count the number of occurrences of pattern  $p$   $X$ .

The MinMax similarity is defined as

The Tanimoto similarity can be used instead if only the occurrence of a pattern is taken into account

The feature maps permit to compute similarity matrices with jCompoundMapper on the full set of features of a compound  $C$ , without introducing noise by hashing. Nevertheless, it is also possible to generate hashed binary fingerprint objects of any of the encodings.

## 3.5 Results and Discussion

### 3.5.1 Computation Times Benchmarks

Table 2 presents the performance of the different encodings as implemented in jCompoundMapper. The computation time for the atom-based approaches varies from the atom pair encodings which can be computed with 332-339 molecules per second to the depth- first search based encodings which have a performance of 68-136 molecules per second. The encodings relying on the typing using the PPP SMARTS definitions are significantly slower with about 7-8 processed molecules per second. The conversion time includes reading, typing, and feature map creation. As benchmark data set, we chose the publicly available ChemDB random background data set published in the virtual screening study by Nasr et al.<sup>28</sup>. This data set comprises 175,000 random compounds from ChemDB<sup>29</sup>.

<sup>27</sup> Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity

<sup>28</sup> Large scale study of multiple-molecule queries

<sup>29</sup> ChemDB: a public database of small molecules and related chemoinformatics resources

### 3.5.2 Machine Learning Performance

A major application of molecular encodings are structure-based machine learning and data mining methods. The aim of such applications is either the prediction of molecular properties or the ranking of compounds according to a trained model. In the following experiments, we wanted to assess the quality of the encodings implemented in jCompoundMapper for several established regression and classification benchmark problems. The encodings were used with the default parameters as given in Table 2. The compounds were prepared using CORINA<sup>30</sup> for initial coordinates and were refined using Schrödinger MacroModel<sup>31</sup> with the OPLS 2005 force field.

#### QSAR Regression Problems with LIBSVM

LIBSVM<sup>12</sup> is a library for support vector machines. For the experiments on the regression benchmarks, we decided to train \*\*-support vector regression on the benchmark compilation of eight pIC50 QSAR problems published by Sutherland et al.<sup>14</sup>. The Gram matrices were precomputed by the MinMax similarity, which is also a valid kernel function. We conducted these experiments to find out whether there are significant differences between the performances of the different encodings.

We evaluated the nested cross-validation performance of the different encodings and compared the outcome of the experiments against results from the literature. The parameters  $C$  and  $\text{nonascii\_441}$ \* of the support vector machine were selected in a nested cross-validation. In the experiments, a 10-fold cross-validation was repeated 20 times using an initial seed value. Therefore, the values represent the mean and the standard deviation, computed over 200 models. Based on these statistics, the corrected resampled \*t-test of Bouckaert and Frank<sup>32</sup> can be applied. The results are summarized in Table 3. With the exception of THERM and THR, the performance of the best encodings was at least comparable to the mean squared error values for a sophisticated graph kernel given by Fechner et al.<sup>33</sup> on the same benchmarks.

An analogue setup was used to compute nested leave-one-out cross-validation results to compare the predictive performance of the support vector machine in combination with the jCompoundMapper encodings to literature results. Again, we optimized the parameters  $C$  and  $\text{nonascii\_451}$ \* (we used a 10-fold cross-validation repeated 2 times to select the best parameter combination in the inner loop) and trained a model for each of the \*n leave-one-out sets and predicted the external sample for each model.

Table 4 and 5 summarize the results of the nested leave-one out cross-validation according to the mean squared error and Pearson's correlation coefficient. Sutherland et al. used several descriptor-based approaches to model the activity of the benchmark set (presented in Table 4 and 5) using partial least squares (PLS)<sup>14</sup>. Compared against the results (squared correlation coefficient) given for the best descriptor approach presented in this study, the jCompoundMapper encodings are competitive. The findings are summarized in Table 6 which shows a similar performance in two cases, a better performance in three cases, and a worse performance in three cases.

#### Classification of Toxic Compounds with LIBLINEAR

Another increasingly important task is to build models on large data sets of chemicals. A machine that can cope with such a setup is LIBLINEAR<sup>16</sup>, a linear large-scale support vector machine. The large Ames data set was published by Hansen et al.<sup>15</sup> and contains 6512 compounds and their measured toxicity in an Ames test. We skipped the encodings based on the PPP typer because several compounds do not have any pharmacophore point according to the PPP definition. The results were obtained by tuning the  $C$  parameter in  $\log_2 \{-8, -7, \dots, 2\}$  within a 2-fold cross-validation on the training set and evaluating the model performance on five defined splits as described in<sup>15</sup>.

Table 7 shows the AUC ROC results for the large Ames toxicity benchmark set. The ECFP and LSTAR encodings achieved the best results, comparable to the results of several supervised classifiers presented by Hansen et al.<sup>15</sup>

<sup>30</sup> Automatic Generation of 3D-Atomic Coordinates for Organic Molecules

<sup>31</sup> NOTITLE!

<sup>32</sup> Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms

<sup>33</sup> Atomic Local Neighborhood Flexibility Incorporation into a Structured Similarity Measure for QSAR

using dragonX<sup>34</sup> descriptors. Hansen et al. applied  $k$ -nearest neighbor, support vector machines with a radial basis function, Gaussian processes, and random decision forests to build models on dragonX descriptors for this problem. The approaches were evaluated on the same defined splits. The performance of LIBLINEAR with the best encodings is comparable to the best approaches Gaussian processes and support vector machines. Even the worst performing encodings (AP2D, AP3D, and DFS) were competitive to the  $k$ -nearest neighbor classifier on dragonX descriptors.

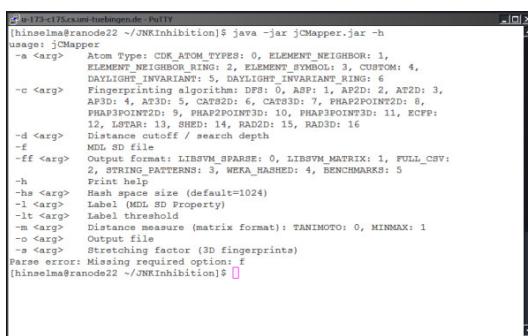
### 3.5.3 Java API and Command-line Interface

#### Java API Usage Example

The core of the library is a Java API. The API enables to process chemical information from an abstract level, similar to a workflow tool. The example given in Appendix 1 reads an MDL SD file, converts the compounds to feature maps and calculates all pairwise similarities.

#### Command-Line Interface Example

The following section gives an example of using the binary executable version of jCompoundMapper. As a case in point, this version can be used in shell scripts. Calling the command-line tool using -h gives an overview of possible parameters (see Figure 2).



```

$ java -jar jCompoundMapper.jar -h
usage: jMapper
-a <arg> Atom Type: CDK_ATOM_TYPES: 0, ELEMENT_NEIGHBOR: 1,
ELEMENT_NEIGHBOR_RING: 2, ELEMENT_SYMBOL: 3, CUSTOM: 4,
DAYLIGHT_INARIANT: 5, DAYLIGHT_INARIANT_RING: 6
-c <arg> Fingerprinting algorithm: DFS: 0, ASPI: 1, AP2D: 2, AT2D: 3,
AP3D: 4, AT3D: 5, CAT3D: 6, CATS3D: 7, PHAPPOINT2D: 8,
PHAPPOINT3D: 9, PHAPPOINT20POINT: 10, PHAPPOINT3D: 11, ECFP:
12, LSTAR: 13, SHED: 14, RAND2D: 15, RAND3D: 16
-d <arg> Distance cutoff / search depth
-f MDL SD file
-ff <arg> Output format: LIBSVM_SPARSE: 0, LIBSVM_MATRIX: 1, FULL_CSV:
2, LIBSVM_PATTERNS: 3, WEKA_HASHED: 4, BENCHMARKS: 5
-h Print help
-hs <arg> Hash space size (default=1024)
-l <arg> Label (MDL SD Property)
-lt <arg> Label threshold
-m <arg> Distance measure (matrix format): TANIMOTO: 0, MINMAX: 1
-o <arg> Output file
-r <arg> Stretching factor (3D fingerprints)
Parse error: Missing required option: f

```

Figure 3.2: Figure 2. Command-line Interface

**Command-line Interface.** The binary can be accessed via a command-line interface, which allows for scripting.

Using the defaults (or via -ff 0), jCompoundMapper generates a hashed LIBSVM output format using the depth-first search encoding with element plus neighbor count atom types.

In the following, we process the training and the known test set from the environmental toxicity challenge <http://www.cadaster.eu/node/65> which were converted to MDL SD format. The label (MDL property) to be learned is log(IGC50-1). Using these settings, the structures of the training set were mapped to hashed fingerprints with the default settings.

After the computation, an overall statistic is plotted showing e.g. the average number of features in the fingerprints. In the next step, we map the test file to the same representation. Bits in the test file were set in exactly the same positions in the vector because the random numbers are generated by using the seed value defined by the features.

In the next step, a cross-validation is conducted by using the precompiled binary distribution of LIBSVM that can be downloaded from the LIBSVM homepage. The parameters are set as follows: -t 0 sets the linear kernel (dot product), -s 3 sets regression, and -c 2 sets the error weight to 2. The file for training was produced in the previous step.

LIBSVM produces no model in cross-validation mode. However, the LIBSVM cross-validations statistics showed that the model has an  $MSE$  of 0.32 and an  $Q^2$  of 0.71, indicating a reasonable parametrization.

<sup>34</sup> NOTITLE!

Finally, the model is trained by omitting the cross-validation flag `-v`.

This step produces a separate model file, which can be used to predict the external test set that was generated during the second step. This is conducted by calling `svmpredict`.

The results are printed by LIBSVM highlighting that the performance on the external test set is  $MSE = 0.29$  and  $R^2 = 0.74$ . The result on the known test of the environmental toxicity prediction challenge would be in the top ranks of the competition. The prediction values can be obtained by opening the LIBSVM

## 3.6 Conclusions

jCompoundMapper is an open source library for molecular fingerprinting with a focus on machine learning and data mining applications. A command-line interface exists for the user who is not familiar with programming, which allows a simple usage from the shell or the application in scripts. The architecture provides the functionality to derive fingerprints from existing ones or to integrate own encodings. In contrast to closed source fingerprinting toolkits, a scientist knows exactly how the fingerprint is computed (like the labeling function, distance cut-offs) and can even inspect the source code of the generation routine. We compared the performance using linear and non-linear support vector machines on standard machine learning benchmarks in the research field. The results show that the machine learning performance using the encodings with default parameters is already close to more sophisticated state-of-the-art descriptors. The binary version provides a command-line interface allowing for the generation of models from the shell with open source software such as LIBSVM or WEKA in reasonable time on average desktop computers. The library itself uses only functionality of open source software licensed under the LGPL. Therefore, the library can be used in any project compatible with the CDK. Further projects with the library, such as a KNIME node wrapping jCompoundMapper, are planned.

## 3.7 Availability

The following files are available for download from <http://jcompoundmapper.sourceforge.net/>

1. External library, which can be integrated as Java jar library file
2. External library, including sources
3. Binary command-line tool (requires a Java runtime environment) and a short tutorial with a prepared data set

## 3.8 Competing interests

The authors declare that they have no competing interests.

## 3.9 Authors' contributions

GH wrote most of the code and the manuscript. LR implemented the Molprint-like fingerprints and the extended connectivity fingerprint, helped to design the library, and participated in writing the manuscript. AJ implemented an initial version of the pharmacophore typer and the CATS2D descriptors. NF tested some of the encodings in experiments and helped to develop the atom typing schemes. AZ supervised the study and participated in the discussion of the results. All authors read and approved the final manuscript.

## 3.10 Appendix

### 3.10.1 Appendix 1 - Usage of the API

Example of using the API: Read molecules, map the compounds to encodings, and compute a similarity matrix.

```
new RandomAccessMDLReader(new File (" molecules . sdf " ));  
  
new ArrayList <FeatureMap>();  
  
new Encoding2DAllShortestPaths ();  
  
for (int i = 0; i < reader. getSize (); i++) {  
  
new FeatureMap (rawFeatures );  
  
new double [ dim ] [ dim ];  
  
new DistanceTanimoto ();  
  
for ( int i = 0; i < dim; i++) {  
  
for ( int j = i ; j < dim ; j++) {
```

# PUBCHEM3D: CONFORMER GENERATION

## 4.1 Abstract

### 4.1.1 Background

PubChem, an open archive for the biological activities of small molecules, provides search and analysis tools to assist users in locating desired information. Many of these tools focus on the notion of chemical structure similarity at some level. PubChem3D enables similarity of chemical structure 3-D conformers to augment the existing similarity of 2-D chemical structure graphs. It is also desirable to relate theoretical 3-D descriptions of chemical structures to experimental biological activity. As such, it is important to be assured that the theoretical conformer models can reproduce experimentally determined bioactive conformations. In the present study, we investigate the effects of three primary conformer generation parameters (the fragment sampling rate, the energy window size, and force field variant) upon the accuracy of theoretical conformer models, and determined optimal settings for PubChem3D conformer model generation and conformer sampling.

### 4.1.2 Results

Using the software package OMEGA from OpenEye Scientific Software, Inc., theoretical 3-D conformer models were generated for 25,972 small-molecule ligands, whose 3-D structures were experimentally determined. Different values for primary conformer generation parameters were systematically tested to find optimal settings. Employing a greater fragment sampling rate than the default did not improve the accuracy of the theoretical conformer model ensembles. An ever increasing energy window did increase the overall average accuracy, with rapid convergence observed at 10 kcal/mol and 15 kcal/mol for model building and torsion search, respectively; however, subsequent study showed that an energy threshold of 25 kcal/mol for torsion search resulted in slightly improved results for larger and more flexible structures. Exclusion of coulomb terms from the 94s variant of the Merck molecular force field (MMFF94s) in the torsion search stage gave more accurate conformer models at lower energy windows. Overall average accuracy of reproduction of bioactive conformations was remarkably linear with respect to both non-hydrogen atom count (“size”) and effective rotor count (“flexibility”). Using these as independent variables, a regression equation was developed to predict the RMSD accuracy of a theoretical ensemble to reproduce bioactive conformations. The equation was modified to give a minimum RMSD conformer sampling value to help ensure that 90% of the sampled theoretical models should contain at least one conformer within the RMSD sampling value to a “bioactive” conformation.

### 4.1.3 Conclusion

Optimal parameters for conformer generation using OMEGA were explored and determined. An equation was developed that provides an RMSD sampling value to use that is based on the relative accuracy to reproduce bioactive

conformations. The optimal conformer generation parameters and RMSD sampling values determined are used by the PubChem3D project to generate theoretical conformer models.

## 4.2 Background

PubChem<sup>1234</sup> is an open archive for the biological activities of small molecules. It consists of three primary databases: Substance, Compound, and BioAssay. The PubChem Compound database contains the unique chemical structure content found in the PubChem Substance database. When possible, a theoretical 3-D conformer model description is generated for each and every record in the PubChem Compound database. This 3-D layer is the basis of the PubChem3D project.

PubChem provides search and analysis tools to assist users in locating desired information in the archive. The importance of this cannot be understated with more than 70 million substance descriptions, 28 million unique small molecules, 480,000 biological assays, and 110 million biological assay outcomes (results from a substance tested in an assay is considered an outcome). Nearly all of these tools focus on the notion of chemical structure similarity at some level. PubChem3D enables similarity of chemical structure 3-D conformers to augment the existing similarity of 2-D chemical structure graphs.

With the goal in mind to use theoretical 3-D descriptions of chemical structures to relate experimental biological activity, there must be an appropriate determination whether these constructs have any relevance to reality. Presumably, if the 3-D conformer model can readily reproduce a reputed “experimental bioactive ligand” conformation with sufficient regularity, one tends to feel (more) confident that the theoretical methodology may be producing biologically meaningful results. There is currently no way to prove with any absolute degree of certainty that all theoretical conformers produced will be biologically relevant; however, one can check if all known experimental “bioactive” conformers of a chemical structure can be found in its theoretical model.

The largest publicly available source of “experimental” 3-D coordinates of chemical structures is the Protein Data Bank (PDB)<sup>5</sup>. This data is not without its considerable issues<sup>678910</sup>. Most “experimental” 3-D coordinates for small molecules provided by the PDB are, in essence, theoretical models derived from fitting electron density produced by X-ray diffraction experiments to the presumed location of atoms that are part of a protein, a ligand (typically bound to the protein), and other moieties (ions, water molecules, etc.). At times, electron density is lacking or there is some degree of uncertainty as to the precise location of the small molecule atoms. In this context, the ligand location or protein binding geometry cannot be considered to be well understood, with many possible conformations of the same ligand plausible<sup>611</sup>. These concerns will be largely ignored here and all the PDB ligands will be treated as experimental fact for the purposes of this study.

There are a number of established theoretical conformer generator packages available<sup>111213141516</sup>. Many of these

---

<sup>1</sup> PubChem: a public information system for analyzing bioactivities of small molecules

<sup>2</sup> An overview of the PubChem BioAssay resource

<sup>3</sup> Database resources of the National Center for Biotechnology Information

<sup>4</sup> PubChem: integrated platform of small molecules and biological activities

<sup>5</sup> The Protein Data Bank

<sup>6</sup> Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools

<sup>7</sup> Conformational changes of small molecules binding to proteins

<sup>8</sup> A new test set for validating predictions of protein-ligand interaction

<sup>9</sup> Validation of protein structures for Protein Data Bank

<sup>10</sup> The advantages and limitations of protein crystal structures

<sup>11</sup> Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database

<sup>12</sup> Conformational freedom in 3-D databases. 1. Techniques

<sup>13</sup> Flexible 3D searching: the directed tweak technique

<sup>14</sup> A fast and efficient method to generate biologically relevant conformations

<sup>15</sup> Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors

<sup>16</sup> Chemical function queries for 3D database search

perform reasonably well<sup>6<sup>17</sup><sup>18</sup></sup>, being both fast and accurate. The specific requirements of the PubChem3D project (such as a multiplatform programmatic interface) made the choice of one of these (OMEGA<sup>19</sup>) very easy. Considering the size of PubChem, the need to relate similar conformers, and the desire to allow users to analyze biological activity patterns in real-time, it requires all 3-D information to be pre-computed and stored. This requirement is a primary limiting factor.

Conformer generator packages are very capable to produce many conformers per chemical structure. This is important as a small molecule of reasonable flexibility at room temperature can access many potential conformational shapes; however, it is impractical to store all produced theoretical conformations per chemical structure, especially when you need to consider the storage requirements for millions of compounds. A common practice is to limit the conformer count using some mix of energy-based filtering, minimum conformer root-mean-squared distance (RMSD) of pairwise atoms, and random sampling. This leaves one to determine how best to minimize the count of conformations stored while not sacrificing coverage or resolution of biologically meaningful conformer space.

In this work, one of a series covering the PubChem3D project, we attempt to answer questions regarding conformer model construction relative to the ability to reproduce PDB ligand 3-D coordinates. For example, what is the baseline conformer generation software accuracy as a function of molecular size and flexibility? Given that conformer models are produced in vacuum, is it beneficial to remove bias towards conformers with intra-molecular interaction to improve accuracy? Is energy filtering useful? What are some practical limitations when generating conformers of flexible molecules? Can one predict average theoretical conformer model accuracy? How do you minimize the count of conformers without significantly impacting accuracy? Using PDB ligand 3-D coordinates, key parameters of conformer model creation are explored to answer these questions. In the process of doing so, a useful relationship is developed relating the size and flexibility of a molecule to the accuracy of reproduction. Further examination is given to accuracy as it relates to limiting the total count of conformers considered in such a model.

## 4.3 Results and Discussion

### 4.3.1 1. Molecular size and flexibility of the PDB ligands

The size and flexibility of a molecule are important factors affecting the conformer model for a molecule. While the molecular size is approximated by the number of non-hydrogen atoms in the molecule, the molecular flexibility can be expressed in terms of the number of effective rotors<sup>20</sup>, which is given as the following:

where  $N^{**ER}$  :sub:\ is the number of effective rotors,  $*N^{**R*}$  \ :sub:\ is the number of rotatable bonds, and  $N^{**NARA}$  :sub:\ is the number of "non-aromatic" \*sp\*\ :sup:\ '3 '\ -hybridized ring atoms. Note that the value of  $*N^{**ER*}$  \ :sub:\ is not necessarily an integer, but, in this study, is frequently rounded to the nearest whole number. Effective rotors take into account the flexibility of rings as well as rotatable bonds. Figure 1 shows the distributions of the non-hydrogen atom counts and the effective rotor counts (binned by whole numbers) for the experimental structures in the Molecular Modeling Database (MMDB)<sup>21<sup>22</sup></sup> ligand dataset as downloaded from the PubChem Substance database (the MMDB contains only experimentally determined data found in the PDB). On average, the molecules in the MMDB ligand set have 17.1 non-hydrogen atoms and 4.9 effective rotors. Although molecules with up to 50 non-hydrogen atoms are considered in the present study, ~90% of them have 30 or less non-hydrogen atoms. In addition, as shown in panel (b) of Figure 1, molecules with greater than 16 effective rotors rarely occur, due to limiting the MMDB dataset to 15 rotatable bonds.

<sup>17</sup> Comparative performance assessment of the conformational model generators omega and catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations

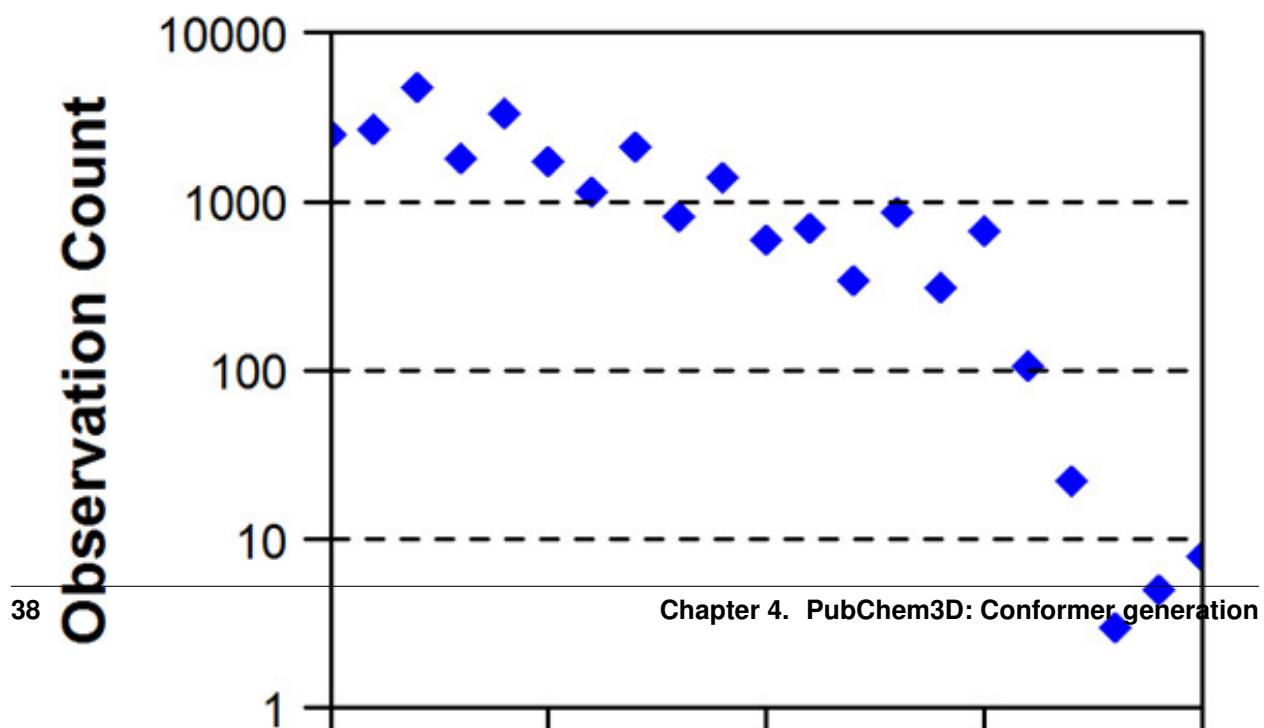
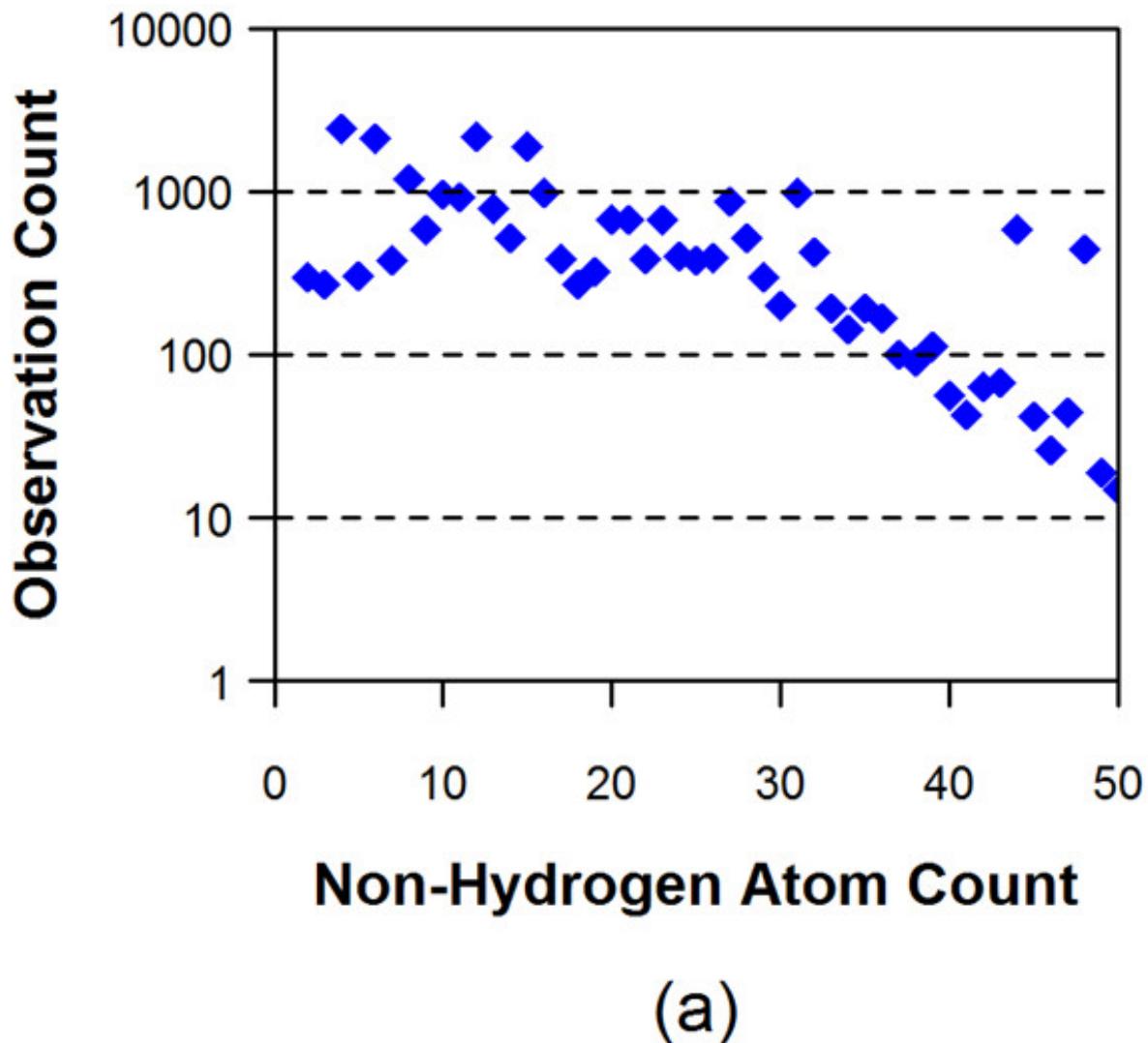
<sup>18</sup> Assessing the performance of OMEGA with respect to retrieving bioactive conformations

<sup>19</sup> OMEGA

<sup>20</sup> Assessment of conformational ensemble sizes necessary for specific resolutions of coverage of conformational space

<sup>21</sup> MMDB: Entrez's 3D-structure database

<sup>22</sup> MMDB: annotating protein sequences with Entrez's 3D-structure database



### 4.3.2 2. Parameter validation for conformer generation

OMEGA<sup>19</sup> was used in the present study. It is known to be among the fastest and most accurate conformer generation programs<sup>17</sup> available. In addition to high quality, it was the only commercially available program that had a non-windows-only C++ application programming interface (API) at the time of project initiation, a critical consideration given the computing environments at the National Center for Biotechnology Information (NCBI). A brief overview of the conformer generation algorithm of OMEGA is given in the *Materials and Methods* section. A more detailed explanation can be found elsewhere<sup>1123</sup>.

OMEGA has many adjustable parameters to generate 3-D conformations with particular attributes. Some important ones are listed in Table 1. To find an optimal set of values, the effects of primary parameters upon conformer generation were tested: (1) the fragment sampling rate for determination of fragment conformation, (2) the type of molecular force field used for the model building and torsion search, and (3) the size of the energy window that determines the energy range of conformers generated. As detailed in the *Materials and Methods* section, only a maximum of 100,000 conformations were considered for a given molecule, meaning that the conformer space of some chemical structures was not fully explored due to this “100-k limit.” Therefore, the occurrence of such cases was considered while testing optimal values of parameters.

In addition to the non-hydrogen atom pair-wise root-mean-square distance (henceforth termed simply RMSD), the Shape-Tanimoto (ST) value was also used as measure of the accuracy of the conformer models. The ST value between any two molecules A and B is given by the following equation:

where  $V^{**AA} :sub:\backslash$  and  $*V^{**BB*} \backslash :sub:$  are the self-overlap volume of A and B, respectively, and  $V^{**AB} :sub:\backslash$  is the overlap volume between A and B<sup>242526</sup>. Note that the ST score is a molecular similarity measure ranging from 0 (for no similarity) to 1 (for identical molecules), whereas the RMSD value is a molecular dissimilarity measure ranging from 0 (for identical molecules) to infinity. Among all theoretical 3-D conformers generated for a given molecule, the one with the least RMSD and the one with the greatest ST to the experimental 3-D coordinates were considered to be the most similar to the “bioactive” conformation. Therefore, the accuracy of a conformer model generated by OMEGA was evaluated using the minimum RMSD and the maximum ST values between the experimental conformation and a single theoretical 3-D conformation in the conformer model. In further discussion, the RMSD and ST values for a conformer model indicates the RMSD and ST values between the corresponding experimental structure and the most similar theoretical conformer in the conformer model, respectively.

#### 2.1. Effects of fragment sampling rate

Fragment sampling in the OMEGA model building stage helps to ensure that flexible ring systems are appropriately examined to find all unique ring conformations. In OMEGA, a float-point value for the “<sup>23</sup>:

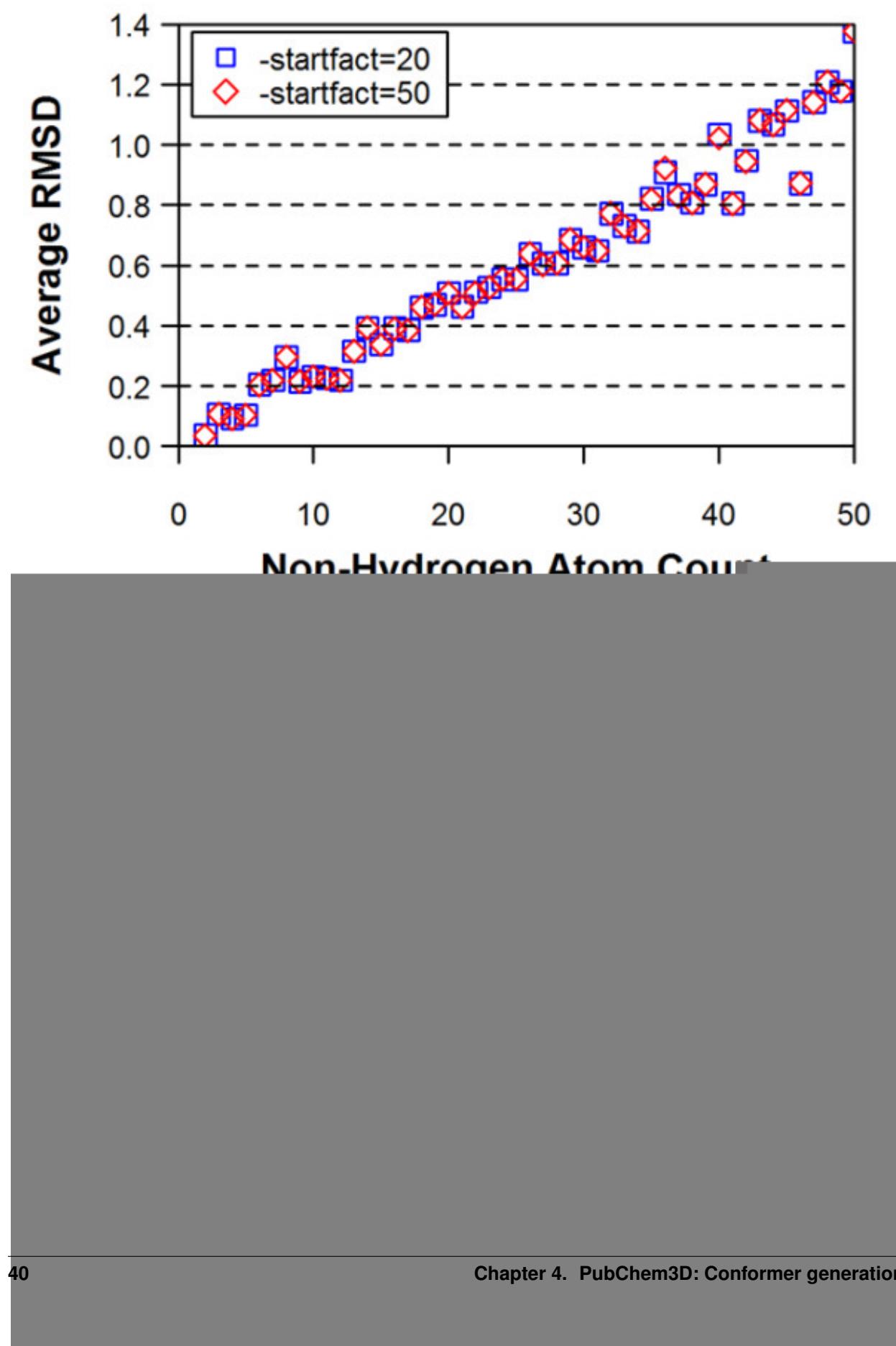
where  $N^{**samples} :sub:\backslash$  is the number of samples, and  $\Sigma(nrb = 3)$  and  $\Sigma(nrb = 4)$  are the sums of the number of atoms in the molecule that have three and four ring bonds, respectively. To investigate effects of the fragment sampling rate upon conformer generation, the conformer models generated using the default value of 20.0 for the “2. In general, using the value of 50.0 rather than the default (= 20.0) was found not to have any significant effect on the overall average RMSD and ST values between the computationally generated conformers and the experimentally determined structures. There were also no significant effects of the increased sampling rate upon the number of conformers generated and the 100-k limit counts. A similar insensitiveness to the fragment sampling rate was observed in Figure 2, which shows the average RMSD as a function of the non-hydrogen atom count and the binned effective rotor count. Thus, the default fragment sampling rate was deemed to be sufficient and the default value of 20.0 was used in the remainder of this study.

<sup>23</sup> OpenEye Omega Toolkit - C++

<sup>24</sup> ROCS - Rapid Overlay of Chemical Structures

<sup>25</sup> A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape

<sup>26</sup> A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction



## 2.2. Effects of force field choice

OMEGA has several pre-defined molecular force fields and it is possible to choose different force fields for model building and torsion search, using the <sup>[27282930313233](#)</sup>, (2) MMFF94s without coulombic interaction terms (*intra-molecular* interactions, which are assumed to not be significant when making *inter-molecular* interactions. This consideration is critical as conformer generation is performed in vacuum, which is a very different environment than a protein binding pocket. Varying the force-field terms allows this hypothesis to be tested as improved accuracy should be found if intra-molecular interactions are removed.

Table 3 shows the effects of the force field employed upon the overall average RMSD and ST of the conformer models at increasing energy window values. The type of fragment force field used during the model building stage caused minor variations (less than 0.01) in the average RMSD of conformer models for energy windows 5 and 10 kcal/mol and these disappear entirely at 15 kcal/mol, indicating an overall insensitivity to the type of fragment force field used. Relative to the earlier hypothesis, this suggests that the fragments produced during model building phase were too small to have intra-molecular interactions. On the contrary, at the energy window of 5, 10, and 15 kcal/mol, the overall average RMSD of the conformer model generated using MMFF94s\_NoEstat for the torsion search force-field was smaller by 0.21, 0.10, and 0.06, respectively, than those that used MMFF94s\_Full, indicating that exclusion of electrostatic terms from the MMFF94s\_Full increased the overall average accuracy of the conformer models significantly, but less so as energy window increased. However, almost no perceptible changes in RMSD or ST averages were found upon the additional removal of van der Waals attractive terms, as shown by nearly imperceptible changes in MMFF94s\_NoEstat and MMFF94s\_Trunc results.

One potential explanation for the higher accuracy of the conformer models generated using the MMFF94s\_NoEstat and MMFF94s\_Trunc force fields arises from their ability to provide more conformations than MMFF94s in the same energy window threshold. The MMFF94s\_Full force field includes additional terms that can lower the energy of conformations with intra-molecular interactions. Removal of such force field terms can increase the energy of conformers with intra-molecular interactions, allowing conformers without these interactions to be represented in a conformer model. This explanation is consistent with the data in Table 4, which shows that MMFF94s\_NoEstat and MMFF94s\_Trunc produced significantly more conformers per compound on average than MMFF94s\_Full. The increased number of conformers per molecule conceptually gives a better chance to have a conformer close to an experimentally determined structure, resulting in the smaller RMSD and greater ST values in Table 3. Because of this, however, the MMFF94s\_NoEstat and MMFF94\_Trunc were also found to more frequently result in 100-k limit cases. Considering that each 100-k limit case suggests a truncation of energetically possible conformations, their substantial frequency increase (by about a factor of five) as a function of increasing energy window may play a role in the reduction of RMSD differences between MMFF94s\_Full force field and MMFF94s\_NoEstat variants.

Another potential explanation for the superiority of MMFF94s\_NoEstat and MMFF94\_Trunc over the MMFF94s\_Full force field is that the lack of intra-molecular interaction is an important characteristic of a biologically relevant conformation of a small-molecule ligand found in its complex with the protein. When a small-molecule ligand binds to its target protein, the intra-molecular interaction in the ligand molecule that exists in its unbound state is likely to disappear because of a conformational change which enhances inter-molecular interaction between the molecule and the target protein. Regardless of the exact reason why, employing the MMFF94s\_NoEstat for conformer model generation appears to be a more sensible choice than the MMFF94s\_Full due to the accuracy improvement.

Overall, it appears that removal of electrostatic terms has a rather favorable effect on improving overall accuracy of reproduction of experimental bound ligand geometries. Removal of attractive van der Waals terms does not appear to have any significant effect. As such, the remainder of this study will only consider the MMFF94s\_NoEstat force field variant.

<sup>27</sup> Merck molecular force field. 1. Basis, form, scope, parameterization, and performance of MMFF94

<sup>28</sup> Merck molecular force field. 2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions

<sup>29</sup> Merck molecular force field. 3. Molecular geometries and vibrational frequencies for MMFF94

<sup>30</sup> Merck molecular force field. 4. Conformational energies and geometries for MMFF94

<sup>31</sup> Merck molecular force field. 5. Extension of MMFF94 using experimental data, additional computational data, and empirical rules

<sup>32</sup> MMFF VI. MMFF94s option for energy minimization studies

<sup>33</sup> MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries

### 2.3. Effects of energy window

In the model building step, the energy window, specified with the “[3](#) shows the effect of energy windows on the overall average accuracy of the conformer models generated. When both the model building and torsion search energy windows were 1 kcal/mol, the overall average RMSD and ST of the conformer model generated were 0.69 and 0.885, respectively. On the other hand, the use of an energy window for both stages of 30 kcal/mol resulted in a substantially improved overall average RMSD of 0.39 and an ST of 0.945. This indicates that a larger energy window gives more accurate conformer models, consistent with the data in Table 3. The increased energy window allows more conformational diversity of a molecule to be considered, but also results in more conformers per molecule (and more 100-k limit cases), as shown in Table 4. Note in Figure 3 that a rapid near-convergence in the overall average RMSD and ST occurs at the energy windows of 10 and 15 kcal/mol for model building and torsion search, respectively. The overall average RMSD and ST at these energy windows were 0.40 and 0.944, respectively. Employing bigger energy windows provided only small improvements to the overall accuracy of the conformer models. Therefore, when looking at overall average results, it initially appears reasonable to use an energy window of 10 kcal/mol for model building and 15 kcal/mol for torsion search without significant reduction in overall accuracy.

#### 4.3.3 3. Accuracy of conformer models and 100-k limit cases

Figures 4 and 5 show the average RMSD and the average ST values of the conformer models, respectively, as a function of the non-hydrogen atom count and the effective rotor count for different energy window values. An increase in the non-hydrogen atom count and the effective rotor count causes a linear increase in the RMSD and a linear decrease in the ST values, indicating that the accuracy of conformer model decreases as a function of both the molecular size and flexibility.

The conformer models that reach the 100-k limit cases may exclude important conformational diversity due to the truncated description of the molecule. Indeed, as Figures 4 and 5 show better accuracy of reproduction when truncated conformer models are excluded. Figure 6 shows effects of the 100-k limit cases upon the distributions of the non-hydrogen atom count and the effective rotor counts, and Figure 7 displays effects of the 100-k limit cases upon the average number of conformers for a molecule as a function of energy window. As one can see in Figure 6, the 100-k limit cases begin appearing for molecules with moderate size and flexibility (e.g., with ~15 non-hydrogen atoms and ~7 effective rotors). As a molecule becomes bigger and more flexible, the OMEGA conformer generation hits the 100-k limit more frequently. Therefore, removal of these cases from the dataset of 25,972 3-D reference structures leaves only a relatively small number of “non-100-k cases” for >30 non-hydrogen atoms and >10 effective rotors, causing a greater variability in the average RMSD and ST in these regions of panels (b) and (d) of Figures 4 and 5. Despite the increased variability, one can still see a slightly noticeable trend of improved accuracy in panel (d) of Figures 4 and 5. One can also see in all panels of Figures 4 and 5 that the energy window for torsion search of 25 kcal/mol gives a clear improvement over 15 kcal/mol for larger values of non-hydrogen atoms and, to a lesser extent, effective rotors. As such, for the remainder of the study 25 kcal/mol will be used for both fragment model and torsion search.

#### 4.3.4 4. Prediction of RMSD accuracy of the conformer ensemble

So far, we have found ways to get the best average RMSD and ST accuracy using OMEGA as the conformer generator. Now we are faced with the problem of data reduction, as it is not practical to store or use all conformers produced when considering millions of chemical structures. Naturally, one would like to maximally reduce the conformer count without significantly sacrificing accuracy, e.g., by a minimum RMSD separation between conformers. The lower the separation RMSD used, the greater the count of conformers that must be kept. Conversely, too large of an RMSD separation may reduce the ensemble accuracy. In an ideal world, one would know *a priori* precisely which conformers are needed and discard the rest. In the real world, this is not possible to know; however, if one can reliably predict the RMSD accuracy of a conformer model, then one can devise an RMSD separation value that could be used as a sampling threshold with some statistical assurances that the sampled conformer model should have at least one conformer within the sampling RMSD some significantly large percentage of the time.

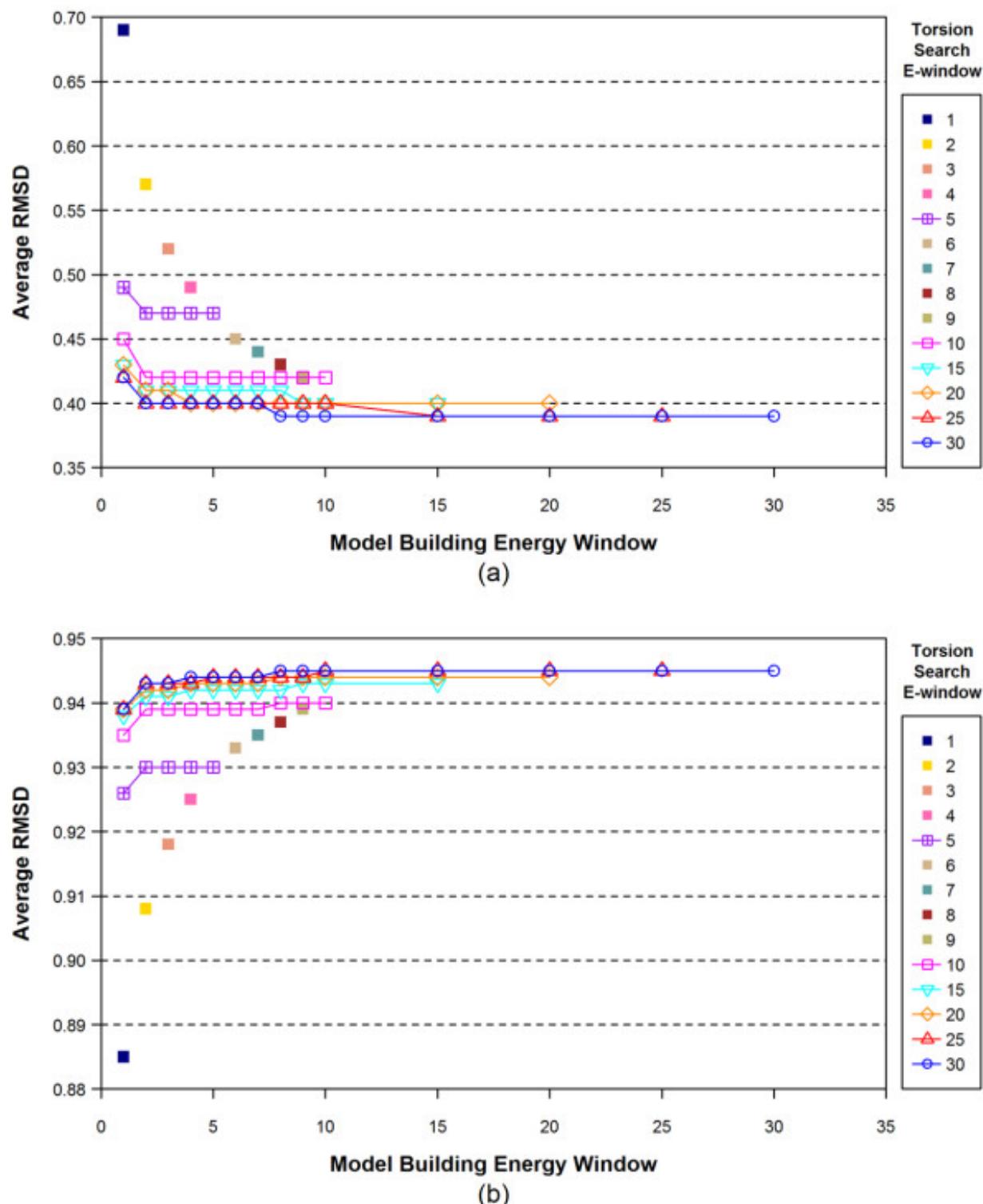


Figure 4.3: Figure 3. The overall average RMSD and Shape-Tanimoto (ST) values of the conformer models for all 25,972 structures as a function of the model building energy window and the torsion search energy window (in kcal/mol)

The overall average RMSD and Shape-Tanimoto (ST) values of the conformer models for all 25,972 structures as a function of the model building energy window and the torsion search energy window (in kcal/mol).

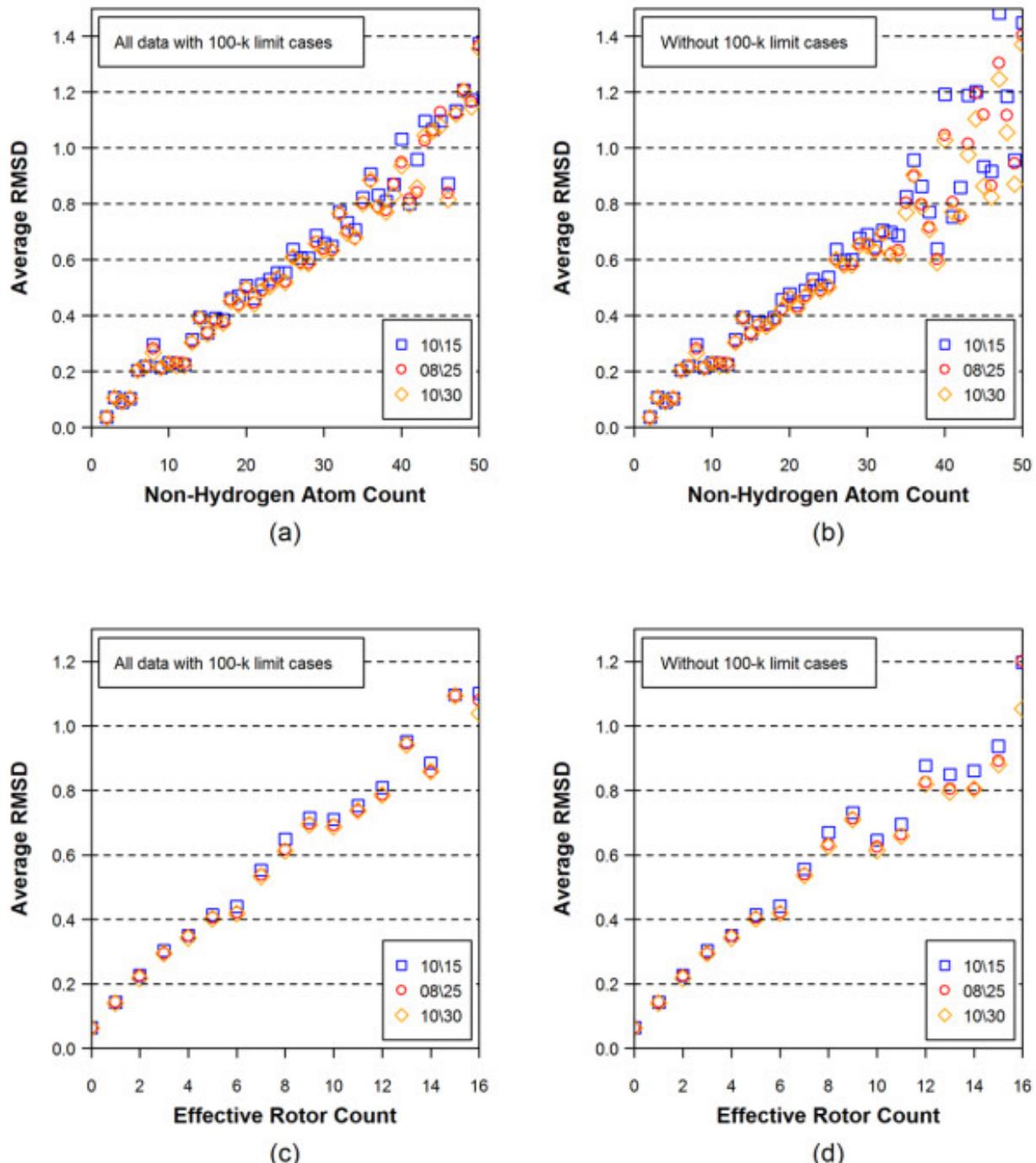


Figure 4.4: Figure 4. Average RMSD values for all 25,972 3D reference structures as a function of non-hydrogen atom count [(a) and (b)] and effective rotor count [(c) and (d)].

**Average RMSD values for all 25,972 3D reference structures as a function of non-hydrogen atom count [(a) and (b)] and effective rotor count [(c) and (d)].** In Panels (b) and (d), the 100-k limit cases, in which the number of conformers reached the maximum number allowed, were removed. The numbers in the legend boxes indicates the energy-window values used in the model building (first) and torsion search (second) stages.

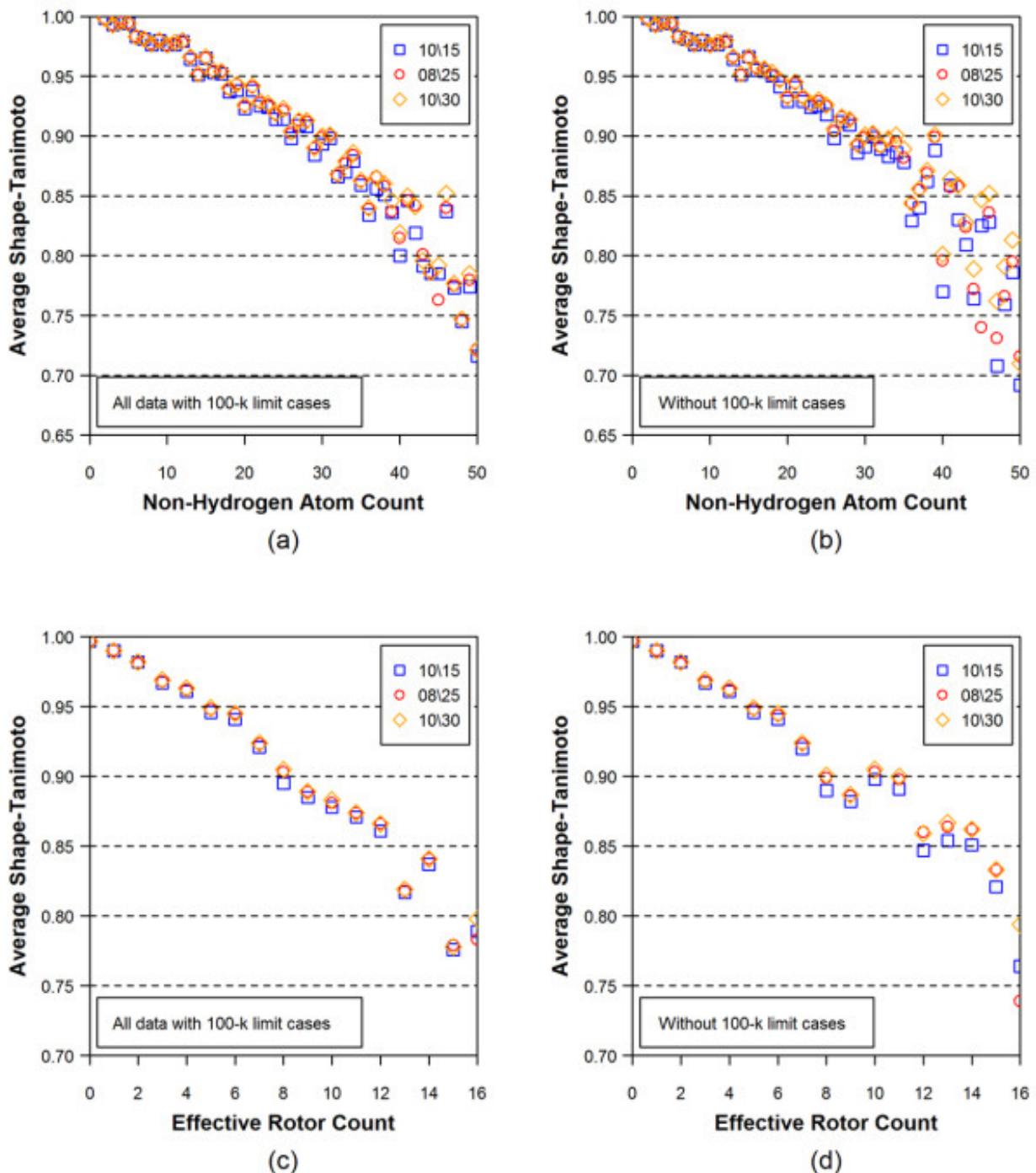


Figure 4.5: Figure 5. Average Shape-Tanimoto (ST) values for all 25,972 3D reference structures as a function of non-hydrogen atom count [(a) and (b)] and effective rotor count [(c) and (d)]

**Average Shape-Tanimoto (ST) values for all 25,972 3D reference structures as a function of non-hydrogen atom count [(a) and (b)] and effective rotor count [(c) and (d)].** In Panels (b) and (d), the 100-k limit cases, in which the number of conformers reached the maximum number allowed, were removed. The numbers in the legend boxes indicates energy-window values used in the model building (first) and torsion search (second) stages.

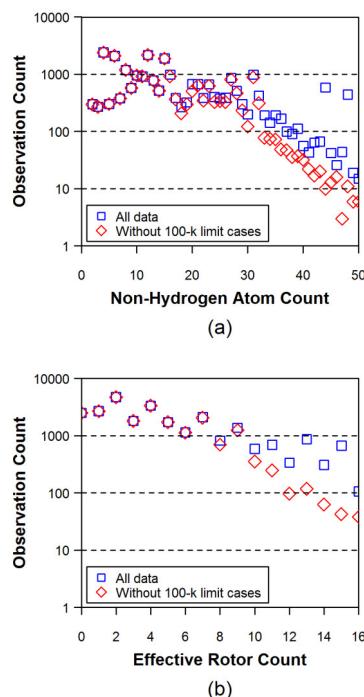


Figure 4.6: Figure 6. The effect of the 100-k limit cases upon the distributions of (1) the non-hydrogen atom count and (2) the effective rotor counts

**The effect of the 100-k limit cases upon the distributions of (1) the non-hydrogen atom count and (2) the effective rotor counts.** The MMFF94s\_NoEstat force field and the energy window of 5 kcal/mol were used for both the model building and torsion search stages.

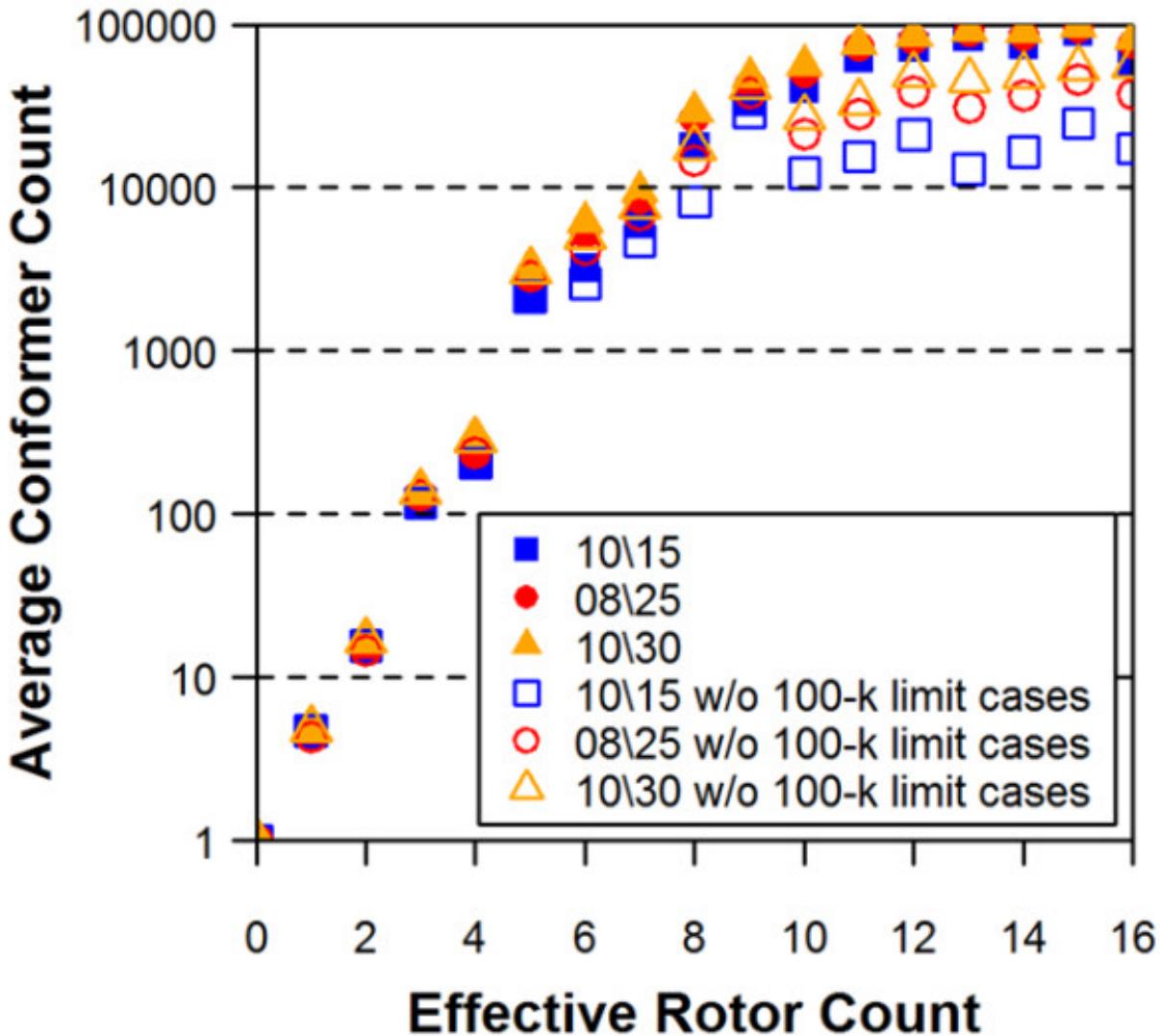


Figure 4.7: Figure 7. The average number of conformers for a molecule as a function of the effective rotor count  
**The average number of conformers for a molecule as a function of the effective rotor count.** The first two numbers in the legend box correspond to the energy-window sizes used in the model building (first) and torsion search (second). While the solid data symbols represent all 25,972 molecules, the open data symbols denote cases in which the 100-k limit cases were removed.

With the aim to find a way to derive an RMSD sampling threshold for a given chemical structure, Figures 4 and 5 show that the average RMSD accuracy of the theoretical conformer ensembles has a very linear correlation with both the non-hydrogen atom count and the effective rotor count. Therefore, a linear regression analysis was performed to derive an equation that predicts the RMSD accuracy of the conformer models using just the  $N^{**}NHA :sub:\backslash$  and the  $*N**ER*\backslash :sub:\backslash$ , yielding Eq. (4).

where  $RMSD^{**pred} :sub:\backslash$  is the predicted RMSD of a theoretical conformer ensemble for a molecule, and  $*N^{**}NHA*\backslash :sub:\backslash$  and  $N^{**}ER :sub:\backslash$  are its non-hydrogen atom and effective rotor counts, respectively. The  $R\backslash :sup:2\backslash$  value of the regression equation for all 25,972 chemical structures was 0.65 and the standard deviation of  $*RMSD^{**pred}*\backslash :sub:\backslash$  was 0.19. The correlation between the  $RMSD^{**pred} :sub:\backslash$  and the actual RMSD ( $*RMSD^{**actual}*\backslash :sub:\backslash$ ) is shown in Figure 8.

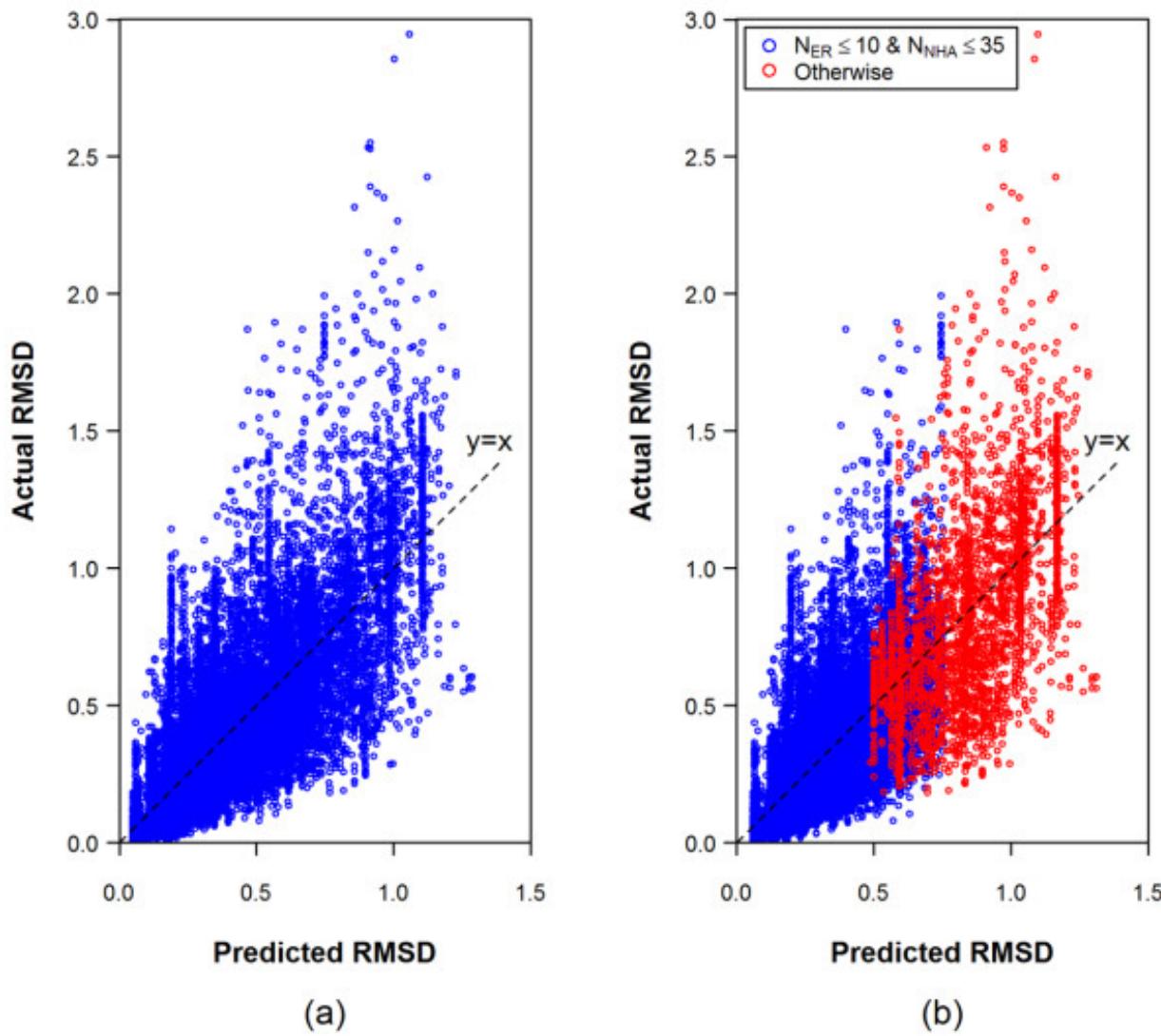


Figure 4.8: Figure 8. Comparison of the predicted and actual RMSD of the theoretical conformer models for the 25,972 experimentally determined structures

**Comparison of the predicted and actual RMSD of the theoretical conformer models for the 25,972 experimentally determined structures.** While the predicted RMSDs in panel (a) were computed using a single equation [Eq. (4)] for all 25,972 structures, those in panel (b) were computed using two different equations: Eq. (6) for the structures with  $N_{ER} \leq 10$  and  $N_{NHA} \leq 35$  (in blue), and Eq. (7) for otherwise (in red).

By design, Eq. (4) overestimates the RMSD value for half of the experimental structures and underestimates the other half. We consider it acceptable to use an RMSD sampling value where 90% of conformer ensembles have the same or lesser RMSD accuracy value on average. This is simply Eq. (4) to which we add the first standard deviation value of 0.19 to yield Eq. (5).

To highlight this, we plotted in Figure 9 the distribution of the difference between  $RMSD^{**actual}$  and  $*RMSD^{**pred}$  using Eq. (5), binned in 0.1 increments. Figure 9 shows that Eq. (5) yields an  $RMSD^{**pred}$  that is greater than or equal to  $*RMSD^{**actual}$  more than 90% of the time.

As shown in Figures 4 and 5, the average RMSD increases and the average ST decreases as a function of both  $N^{**ER}$  and  $*N^{**NHA}$ , regardless of the inclusion of 100-K limit cases. Given how the variability increases as chemical structures become larger and more flexible, potentially due, in part, to their lower populations, it may be helpful to partition data into separate groups according to their  $N^{**ER}$  and  $*N^{**NHA}$  values and perform separate regression analyses. A regression analysis of the 22,587 structures with  $N^{**ER} < 10$  and  $*N^{**NHA} < 35$  yields Eq. (6), while a regression analysis of the 3,385 structures with  $N^{**ER} > 10$  or  $*N^{**NHA} > 35$  yields Eq. (7):

The  $R^2$  values of the regression formula Eqs. (6) and (7) were 0.52 and 0.33, respectively, and the standard deviation of  $RMSD^{**pred}$  was 0.17 and 0.29 for Eqs. (6) and (7), respectively. (Attempts to partition  $*N^{**ER}$  and  $N^{**NHA}$  values using different partitioning schemes had similar  $R^2$  results.) In the same manner as used to derive Eq. (5), one can add these first standard deviations to Eqs. (6) and (7) to get RMSD sampling formulas Eqs. (8) and (9), respectively, for the two individual groups. As shown in Figure 8, the  $*RMSD^{**pred}$  values from Eqs. (6) and (7) are comparable with those from Eq. (4), despite the poor  $R^2$  values for Eqs. (6) and (7). As shown in Figure 9, where the RMSD sampling values from Eq. (5) are compared with those from Eqs. (8) and (9), the frequency distributions of the difference between  $RMSD^{**pred}$  and  $*RMSD^{**actual}$  are similar to each other. The  $RMSD^{**pred}$  value is greater than or equal to  $*RMSD^{**actual}$  for ~91.0% of the time when Eq. (5) is used, and for ~91.6% of the time when Eqs. (8) and (9) are used.

In the recent study of Hawkins *et al.*<sup>11</sup>, the quality of theoretical conformations from OMEGA was evaluated by comparing a set of 197 high-quality PDB ligand structures with corresponding OMEGA-generated theoretical conformers. They used the MMFF94\_Trunc force field to generate conformers and then reduced their count to a maximum of 200 by sampling. The mean RMSD between the 197 ligand set and their corresponding theoretical conformers was 0.67 Å. According to their bootstrapping tests, the OMEGA-generated conformers are expected 90% of the time to have an RMSD value between 0.647-0.688 Å for chemical structures with similar properties to those tested. The average rotatable bond count and non-hydrogen atom count of this 197 ligand set were 6.3 and 24.4, respectively, indicating that they are, on average, slightly bigger and flexible than the 25,972 set used in our study (4.9 rotatable bonds and 17.1 non-hydrogen atoms on average). With these average counts as the  $N^{**NHA}$  and  $*N^{**ER}$  values in Eq. (5),  $RMSD_{pred}$  is predicted to be 0.71 Å, which is comparable to 0.67 Å, the mean RMSD between the 197 ligand set and the corresponding theoretical conformers. Considering that the OMEGA parameters used in both studies were not exactly identical and that we used their reported rotatable bond count rather than effective rotor count, this may suggest that Eq (5) may have applicability beyond the parameter sets used in this study.

## 4.4 Conclusion

In the present study, theoretical conformer ensembles for 25,972 experimental MMDB ligand molecules were generated using the OMEGA software and the accuracy of the conformer models were analyzed in terms of the non-hydrogen atom pair-wise RMSD and shape similarity ST values between the theoretical conformer models and the corresponding experimental structures.

Effects of different settings for three important parameters (the fragment sampling rate, the type of the molecular force field, and the size of the energy window) that OMEGA uses for conformer generation were investigated to

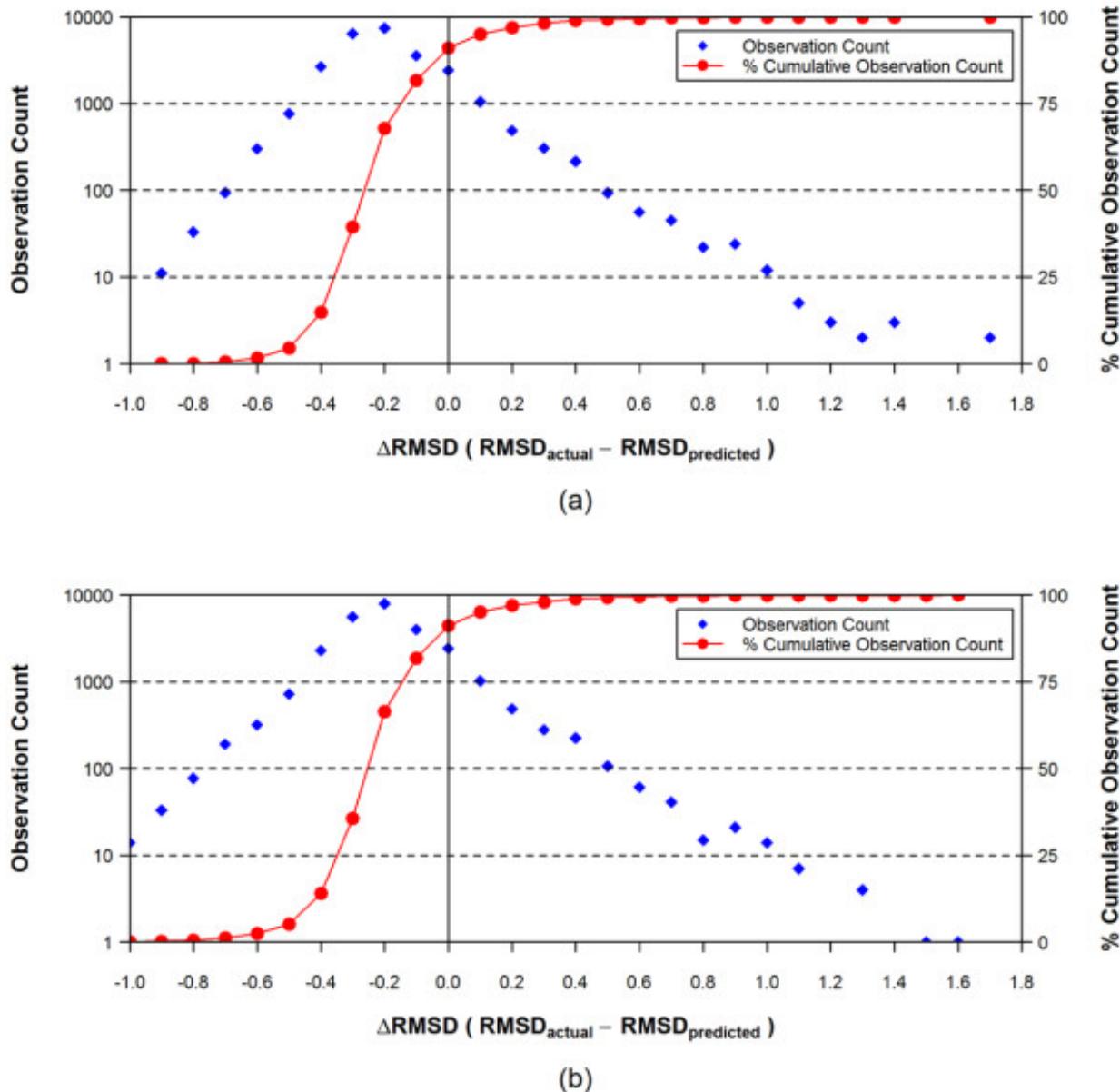


Figure 4.9: Frequency distribution of the RMSD differences between the actual RMSD and the predicted RMSD values for the conformer models of the 25,972 experimental structures, binned in 0.1 increments using Eq. (5) for panel (a) and Eqs. (8) and (9) for panel (b). (See text)

**Frequency distribution of the RMSD differences between the actual RMSD and the predicted RMSD values for the conformer models of the 25,972 experimental structures, binned in 0.1 increments using Eq. (5) for panel (a) and Eqs. (8) and (9) for panel (b). (See text).**

find their optimal settings. The use of a fragment sampling rate greater than the default (= 20.0) did not make a statistically significant change, indicating the default fragment sampling rate is already sufficient. Variation in the fragment force field type was found to provide little benefit. However, the accuracy of the theoretical conformer models was sensitive to the force field type used in the torsion search stage. When used as a torsion search force field, the MMFF94s\_NoEstat and MMFF94s\_Trunc force fields were found to generate more conformers for a given molecule and also improve the overall accuracy of the theoretical conformer models, compared to the MMFF94s force field but less so as the energy window was increased. In general, using a greater energy window in the model building and torsion search stages resulted in more accurate conformer models. However, a rapid convergence in the overall average RMSD and ST of the conformer models was observed when the energy windows were bigger than 10 kcal/mol for model building and 15 kcal/mol for torsion search. However, for larger and more flexible structures, an energy window of 25 kcal/mol for torsion search gave some noticeable improvement to the overall accuracy for larger and more flexible structures and may be better threshold for general purpose use. The average accuracy as a function of non-hydrogen atom count (size) and effective rotor count (flexibility) was very linear in the ranges considered. A regression equation was developed using these two variables to predict the accuracy of a theoretical ensemble to reproduce the experimental geometry. This equation was subsequently used to provide a RMSD sampling rate to filter conformer models such that 90% of conformer ensembles should have the same or lesser RMSD accuracy value, thus, allowing one to maximize the accuracy of a conformer model while minimizing the count of retained conformers.

## 4.5 Materials and Methods

### 4.5.1 1. Datasets

The “experimental” 3-D coordinate data set of small molecules used in the present study was downloaded from the Molecular Modeling DataBase (MMDB)<sup>2122</sup> ligand dataset as available from the PubChem Substance database at NCBI on October 20, 2006. The data set was used to calibrate the parameters used when operating the software generation package OMEGA<sup>19</sup>. Ligands that were too small or too big were discarded by limiting the non-hydrogen atom count to 2 - 50. Ligands too flexible (with a rotatable bond count greater than 15) were also eliminated. This filtering stage resulted in an initial dataset that contained 25,972 non-unique organic 3-D experimental reference structures where a 3-D conformer model could be generated.

### 4.5.2 2. Conformer generation using OMEGA

Essentially, the OMEGA application performs conformer generation in two primary stages: model building and torsion search. In the model building stage, initial molecular structures are constructed by assembling fragment templates, which are generated from fragmentation of the input molecular graph along sigma bonds. In the torsion search stage, OMEGA generates a conformer ensemble using particular rule-based torsion angles that depend on the molecular environment between connecting fragments.

There are a number of adjustable parameters available when performing conformer generation. The effects of three primary parameters upon the accuracy of conformer models generated were evaluated:

- 1.
- 2.
3. •

See Table 1 for all non-default parameters used. It is important to note that a maximum of 100,000 (100-k) conformers were generated per chemical structure. This limit is more than adequate for small or inflexible structures but for flexible compounds this limitation is of concern. For example, imagine a chemical structure with nine rotatable bonds where one systematically samples each rotatable bond four times; one would generate  $4^9$  (= 262,144) conformations. Therefore, the effects of these 100-k limit cases upon the overall accuracy of the conformer models is analyzed as a function of the non-hydrogen atom counts and the effective rotor counts. While increasing this 100-k threshold

to a larger value was not possible with earlier versions of OMEGA, one can see that increasing the total count of conformers considered by five or ten times would still not be sufficient for many flexible molecules.

To assess the accuracy of reproduction of experimental coordinates as a function of conformer generation parameters modification, two metrics were used: the RMSD of non-hydrogen atoms using the OEChem OERMSD function (with “automorph” detection turned on to allow proper treatment of symmetrically equivalent atoms and “overlay” turned on to allow rotation/translation to yield the lowest possible RMSD value) and the shape-optimized ST using the value reported by ROCS. For each conformer produced for a structure, an RMSD and ST determination were made. The lowest RMSD and greatest ST values per conformer model were used to assess “accuracy” of reproduction for the parameters used.

## 4.6 Competing interests

The authors declare that they have no competing interests.

## 4.7 Authors' contributions

EEB performed most of the research and SK wrote the first draft. SHB reviewed the final manuscript. All authors read and approved the final manuscript.

## 4.8 Acknowledgements

We are grateful to the NCBI Systems staff, especially Ron Patterson, Charlie Cook, and Don Preuss, whose efforts helped make the PubChem3D project possible. This research was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, U.S. Department of Health and Human Services. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD. (<http://biowulf.nih.gov>).

# MODULAR CHEMICAL DESCRIPTOR LANGUAGE (MCDL): STEREOCHEMICAL MODULES

## 5.1 Abstract

### 5.1.1 Background

In our previous papers we introduced the Modular Chemical Descriptor Language (MCDL) for providing a linear representation of chemical information. A subsequent development was the MCDL Java Chemical Structure Editor which is capable of drawing chemical structures from linear representations and generating MCDL descriptors from structures.

### 5.1.2 Results

In this paper we present MCDL modules and accompanying software that incorporate unique representation of molecular stereochemistry based on Cahn-Ingold-Prelog and Fischer ideas in constructing stereoisomer descriptors. The paper also contains additional discussions regarding canonical representation of stereochemical isomers, and brief algorithm descriptions of the open source LINDES, Java applet, and Open Babel MCDL processing module software packages.

### 5.1.3 Conclusions

Testing of the upgraded MCDL Java Chemical Structure Editor on compounds taken from several large and diverse chemical databases demonstrated satisfactory performance for storage and processing of stereochemical information in MCDL format.

## 5.2 Background

In our previous paper we introduced the Modular Chemical Descriptor Language (MCDL) for providing a linear representation of structural and other chemical information<sup>1</sup>. All MCDL descriptors have two unique modules that describe the composition and the connectivity of a molecule. Optional supplementary modules, which may or may not be unique, contain additional information about a compound (such as spectra, physical-chemical data, atomic

---

<sup>1</sup> Modular Chemical Descriptor Language (MCDL): Composition, Connectivity, and Supplementary Modules

coordinates, references, *etc.*). The MCDL rules were implemented in the LINDES computer program, which was designed to generate MCDL linear descriptors from files containing molecular structural information in the form of a connectivity matrix.

A subsequent development was the MCDL Java Chemical Structure Editor which is capable of drawing chemical structures from linear representations and generating MCDL descriptors from structures<sup>2</sup>. Since the module containing atomic coordinates is an optional feature of MCDL descriptors, it was necessary for the Java applet to be capable of restoring these coordinates to draw a structure. As a result, the current applet algorithm that was developed to process MCDL descriptors can also be used for processing of other coordinate-less structure representations, such as SMILES<sup>34</sup> and InChI<sup>5</sup>.

The initial MCDL concept<sup>1</sup> had one serious drawback - it did not support stereochemistry. In this paper we present optional MCDL modules and accompanying software that incorporate unique representation of molecular stereochemistry. The paper also contains some additional discussions regarding canonical representation of stereochemical isomers in MCDL format, and brief algorithm descriptions of the open source LINDES, Java applet, and Open Babel MCDL processing module software packages. The results of software testing are presented in the last section of the paper.

## 5.3 Results and Discussion

### 5.3.1 MCDL stereochemistry descriptors - theory

As noted previously<sup>1</sup>, all MCDL linear descriptors include two primary modules that uniquely describe the basic molecular structure: the composition and the connectivity modules. The connectivity module is based on molecular topology, which adequately describes the sequence of bonds that connect atoms in the molecule. However, the topology-based connectivity module is inadequate for describing the three-dimensional arrangement of those atoms, which is the distinguishing characteristic in the structures of stereoisomers [see Appendix 1]. Refinement in the molecular structure representation in the MCDL can be achieved by employing a set of supplemental stereochemistry descriptors. The task is complicated by the existence of many types of stereoisomers<sup>67</sup>. The simplest are the common “optical” isomers of compounds with asymmetric atoms and the *cis-trans* isomers of compounds with double bonds. Less-common types of stereoisomers with more complex stereogenic units include “phase” isomers found in gear-like molecules<sup>89</sup> and chiral molecular knots<sup>1011</sup>. In addition, a combination of different stereochemical types within a molecule makes comprehensive stereochemical analysis convoluted and often ambiguous. As a result, complete and unique representation of molecular stereochemistry is a compelling challenge.

Due to the complexity of underlying phenomena, specification of stereochemical information in the MCDL is currently limited to the two most common types - stereochemistry of a chiral atom in the {SA:} module and stereochemistry of a double bond in the {SB:} module. Within each type, a canonicalization procedure (described below) has been developed to generate unique stereodescriptors.

Systematic nomenclature of stereoisomers dates back to the pioneering works of Fischer<sup>1213</sup> and Cahn-Ingold-Prelog (CIP)<sup>1415</sup>, all utilizing various schemes for a unique (canonical) prioritization of substituents attached to either an

<sup>2</sup> A Java Chemical Structure Editor Supporting the Modular Chemical Descriptor Language (MCDL)

<sup>3</sup> SMILES a Chemical language and Information System. 1. Introduction to Methodology and Encoding Rules

<sup>4</sup> SMILES 2. Algorithm for Generation of Unique SMILES Notation

<sup>5</sup> An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier

<sup>6</sup> International Union of Pure and Applied Chemistry (IUPAC). Nomenclature of Organic Chemistry

<sup>7</sup> Basic Terminology of Stereochemistry

<sup>8</sup> Molecular Gearing Systems

<sup>9</sup> Stereochemical Consequences of Dynamic Gearing

<sup>10</sup> Resolution of Topologically Chiral Molecular Objects

<sup>11</sup> Topological Features of Protein Structures: Knots and Links

<sup>12</sup> Ueber die Configuration des Traubenzuckers und seiner Isomeren

<sup>13</sup> Ueber die Configuration des Traubenzuckers und seiner Isomeren. II

<sup>14</sup> Specification of Molecular Chirality

<sup>15</sup> Basic principals of the CIP-System and Proposals for a Revision

atomic center or to the two atoms connected by a double bond. The latter, and the most developed, CIP scheme uses atomic numbers as the basis for substituent priority ranking and requires sophisticated multi-level priority algorithms in the case of substituents with identical atomic numbers. CIP rules are known to produce ambiguous results due to the non-unique ranking of substituents in some complicated cases and have been under development during the last decades<sup>1617</sup>. Nevertheless, the rules work relatively well for the majority of simple organic molecules.

The MCDL employs both CIP and Fischer ideas in constructing stereoisomer descriptors. Similar to CIP, the MCDL stereochemistry descriptors are based on prioritization of substituents, but unlike CIP, the MCDL algorithm uses planes, not axes, to specify the configuration of an atomic center (Fischer's approach). Although the algorithm rules are close to the CIP rules (priority ranking)<sup>1415</sup>, the resulting MCDL descriptors are not identical to the *R*-*S* and *E*-*Z* naming conventions due in large part to the differences in the underlying prioritization approaches.

Stereoisomer descriptors are expected to be unique in all cases where canonical numbering can be implemented. It is important to note that in cases where two or more constitutionally equivalent numbering schemes can be derived, all must be taken into consideration for the selection of the unique stereochemistry descriptor. This approach is currently the only reliable method for establishing the unique (canonical) descriptors and is very similar to one that has been developed previously for the unique MCDL connectivity modules<sup>1</sup>.

### 5.3.2 Atom stereochemistry (chiral centers)

Atom stereochemistry takes into consideration the three-dimensional arrangement of substituents around an atomic chiral center. In the majority of cases it is a four-coordinated atom (such as a carbon atom), but there is a substantial number of stereoisomers having chiral centers with three substituents. Notable examples include chiral sulfoxides.

The priorities of the substituents attached to a chiral center are the MCDL priorities (based on ASCII codes) of the attached fragments and terminal atoms, if any (see below). Once these are known, a Fischer projection of the configuration at the chiral atom is drawn. A Fischer projection is a planar representation of a molecule that preserves information about chirality. With the chiral atom at its center, a horizontal line represents two bonds bending forward toward the viewer, and a vertical line represents two bonds bending back away from the viewer. In the projection, the highest MCDL priority substituent on the chiral atom is placed at the top and the second highest at the bottom. The other two substituents appear at the left and right and are positioned to preserve the configuration of the chiral atom. Once oriented in this way, the atom stereochemistry is specified in the MCDL linear descriptor as {SA:chiral fragment,top,bottom,left,right} where the four positions refer to the positions in the Fischer projection.

#### One chiral center

As an example, consider 2-hydroxy-2-methylbutanoic acid shown in Figure 1. The  $\alpha$ -carbon is chiral, and thus this molecule has two mirror-image structures. The MCDL composition and connectivity modules for both of these are C;CHH;2CHHH;CO;2OH[2,3,5,6;4;;;7]. The chiral atom is fragment 1 in the descriptor. Its substituents and their priorities are CHH (2), CHHH (3), CO (5), and OH (6). Three-dimensional representations of this compound's enantiomers are shown in Figure 2 with the priorities of the substituents indicated.



Figure 5.1: Figure 1. Structural formula of 2-hydroxy-2-methylbutanoic acid  
**Structural formula of 2-hydroxy-2-methylbutanoic acid.**

The two structures are redrawn as Fischer projection formulas in Figure 3. The four substituents have been placed on the projection so the fragment having the highest priority is at the top and the next highest at the bottom. Thus, the

<sup>16</sup> A New Effective Algorithm for the Unambiguous Identification of the Stereochemical Characteristics of Compounds During Their Registration in Databases

<sup>17</sup> A Computer-Oriented Linear Canonical Notational System for the Representation of Organic Structures with Stereochemistry

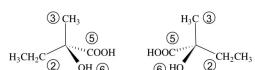


Figure 5.2: Figure 2. Three-dimensional representations of 2-hydroxy-2-methylbutanoic acid's enantiomers with the MCDL priorities of the substituents on the chiral atom indicated

**Three-dimensional representations of 2-hydroxy-2-methylbutanoic acid's enantiomers with the MCDL priorities of the substituents on the chiral atom indicated.**

stereochemistry for the left structure is specified in the MCDL as {SA:1,2,3,5,6} and as {SA:1,2,3,6,5} for the right structure.

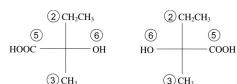


Figure 5.3: Figure 3. Fischer projections of 2-hydroxy-2-methylbutanoic acid's enantiomers  
**Fischer projections of 2-hydroxy-2-methylbutanoic acid's enantiomers.**

The example of D-lactic acid (*R*-lactic acid), shown in Figure 4, brings up a new issue to consider. The MCDL composition module of lactic acid is CH;CHHH;CO;2OH, and the chiral carbon is part of the first fragment. The attached fragments and their MCDL priorities are CHHH (2), CO (3), and OH (4). The fourth substituent on the chiral carbon, H, is actually part of fragment 1, but it must be considered on its own in order to specify the stereochemistry in the linear descriptor. To handle this situation, a new rule is added to the MCDL: Structural fragments have higher priorities than terminal atoms (numbers have lower ASCII codes than letters). Thus, of the four substituents attached to the chiral atom in lactic acid, the terminal atom H has the lowest priority. Orienting the *R* configuration as the Fischer projection in Figure 4 gives the MCDL descriptor {SA:1,2,3,4,H}.

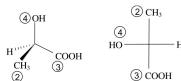


Figure 5.4: Figure 4. Three-dimensional representation and Fischer projection of D-(*R*)-lactic acid  
**RThree-dimensional representation and Fischer projection of D-.**

## Multiple chiral centers

The MCDL representation of multiple chiral centers in a molecule consists of a sequence of atomic configuration descriptors (one for each of the chiral centers), listed in descending priority order of the chiral fragments (smallest ASCII value first) and separated by semicolons. While treatment of structures with multiple chiral centers in the MCDL is straightforward in many cases, the presence of multiple chiral centers in certain quasi-symmetrical structures may lead to complications due to the fact that the unique part of the MCDL linear descriptor is generated without consideration of stereochemistry. As a result, constitutionally, but not stereochemically, equivalent chiral centers may receive arbitrarily selected fragment numbers.

Figure 5 shows the 3-dimensional structure of *meso*-tartaric acid having two constitutionally identical, but stereochemically different, chiral centers (opposite configurations). Because of the topological symmetry of the molecule, two MCDL numbering schemes are possible. (If the chiral centers had the same configuration, the two numbering schemes would be identical.) The difference can be seen in Figure 5 in which either structural fragment 1 has the *R* configuration (left structure) or fragment 2 is *R* (right structure).

Figure 6 shows the Fischer projections centered at fragments 1 and 2, respectively, of the left structure in Figure 5. The MCDL stereochemistry descriptor for this numbering scheme is {SA:1,2,3,5,H;2,1,4,H,6}.

Figure 7 corresponds to the right structure in Figure 5, giving the MCDL descriptor {SA:1,2,3,H,5;2,1,4,6,H}.

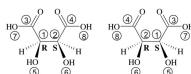


Figure 5.5: Figure 5. Two possible MCDL numbering schemes for *meso*-tartaric acid  
mesoTwo possible MCDL numbering schemes for .

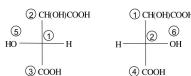


Figure 5.6: Figure 6. The Fischer projections centered at fragments 1 and 2, respectively, of the left structure in Figure 5.

**The Fischer projections centered at fragments 1 and 2, respectively, of the left structure in Figure 5.**

To choose the correct and unique stereodescriptor in cases like this, the two possible descriptors are compared position-by-position starting at the left. In this instance (<{SA:1,2,3,5,H;2,1,4,H,6} vs. <{SA:1,2,3,H,5;2,1,4,6,H}}), the first difference occurs in the fourth position where one has a 5 and the other has an H. Since 5 is a higher priority than H, this difference allows us to choose <{SA:1,2,3,5,H;2,1,4,H,6} as the MCDL stereochemistry descriptor for *meso*-tartaric acid. As a general rule, for quasi-symmetrical structures where multiple equivalent numbering schemes are possible, all must be explored for selection of the canonical (the lowest ASCII code sequence) stereochemistry descriptor. An alternative approach entails the use of hash variables for constitutionally equivalent stereogenic fragments (see *Generation of atomic MCDL stereodescriptors from structure diagrams* section below).

### Three-substituent chiral centers

The MCDL can also specify chirality of atomic centers with only three substituents in those cases where the fourth substituent position is occupied by an electron pair. In this case, the electron pair is treated as a “dummy” substituent positioned at the point of maximum distance from the three “real” substituents on a sphere of unit radius centered at the chiral atom. The priority of the electron pair is defined in the MCDL as 0 (zero), giving it, maybe surprisingly, higher priority than structural fragments and terminal atoms. For example, the chirality descriptors of ethyl(fluoromethyl)sulfoxide (CFHH;CHH;CHHH;SO[4;3,4]), shown in Figure 8, are <{SA:4,,1,2,O} for the left structure and <{SA:4,,1,O,2} for the right structure. Note that the “dummy” 0-numbered substituent is not included in the descriptor.

#### 5.3.3 Double bond stereochemistry

The configuration of a double bond can be specified after the priorities of the structural fragments making up the molecule are known. Four items of information are needed:

- the priority numbers of the two fragments containing the double-bonded atoms - (x\*\*\*\*1:sub:\ , \* \* x \* \* \* 2 \* \*\ :sub:, x\*\*I :sub:\ <\*x\*\*2\*\ :sub:)
- the higher priority connection to the higher priority fragment of the double bond, x\*\*I :sub:\ - (\* \* n \* \* \* 1 \* \*\ :sub:)

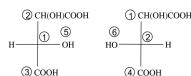
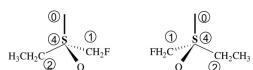


Figure 5.7: Figure 7. The Fischer projections centered at fragments 1 and 2, respectively, of the right structure in Figure 5.

**The Fischer projections centered at fragments 1 and 2, respectively, of the right structure in Figure 5.**



**Figure 5.8:** Figure 8. Three-dimensional representations of ethyl(fluoromethyl)sulfoxide’s enantiomers  
**Three-dimensional representations of ethyl(fluoromethyl)sulfoxide’s enantiomers.** The “dummy” substituent is numbered as 0.

- the connection to the lower priority fragment of the double bond,  $x^{**2}:sub:\backslash$ , that lies on the \*same side\* of the double bond as  $*n^{**1*} \backslash :sub:- (n^{***2}:sub:““)$

The stereochemistry is specified as {SB: $x^{**1}:sub:\backslash d*x^{**2*} \backslash :sub:, n^{**1*}:sub:\backslash, n^{**2*} \backslash :sub:, n^{**3*}:sub:\backslash, n^{**4*} \backslash :sub:},$  where  $n^{**3} :sub:\backslash$  and  $n^{**4*} \backslash :sub:$  are the remaining two fragments attached to  $x^{**2} :sub:\backslash$  and  $*x^{**1*} \backslash :sub:,$  respectively. Multiple double bond configurations are separated by semicolons and are listed with increasing values of  $x^{**1}:sub:““$ .

### One double bond

As an example, consider 3,4-dimethyl-3-heptene with the configuration shown in Figure 9. The priorities of the fragments containing the double-bonded atoms and their immediate connections are shown. In this figure the priority numbers of the two fragments containing the double-bonded atoms are 1 and 2. Of the two connections (3 and 6) to the higher priority fragment of the double bond (1), fragment 3 has the higher priority. The connection to fragment 2 that lies on the same side of the double bond as fragment 3 is fragment 7. Thus, the stereochemistry of the compound in Figure 9 is specified as {SB:1d2,3,7,4,6}. Including the composition and connectivity modules, the MCDL linear descriptor is 2C;3CHH;4CHHH[2,3,6;4,7;5;8;9]{SB:1d2,3,7,4,6}.



**Figure 5.9:** Figure 9. MCDL numbering of 3,4-dimethyl-3-heptene for stereochemistry descriptor generation  
**MCDL numbering of 3,4-dimethyl-3-heptene for stereochemistry descriptor generation.**

In the specification, the values of  $n^{**1}:sub:\backslash$ ,  $*n^{**2*} \backslash :sub:, n^{**3*}:sub:\backslash$  and  $*n^{**4*} \backslash :sub:$  are not necessarily structural fragment numbers. This can occur when connections to the double-bonded atoms are terminal atoms rather than structural fragments. Consider 1,2-dibromopropene in the configuration shown in Figure 10.



**Figure 5.10:** Figure 10. MCDL numbering of 1,2-dibromopropene for stereochemistry descriptor generation  
**MCDL numbering of 1,2-dibromopropene for stereochemistry descriptor generation.**

This compound consists of three structural fragments: CBr, CBrH, and CHHH. These have the MCDL priorities shown in the Figure 10. The priority numbers of the two fragments containing the double-bonded atoms are 1 and 2. Of the two connections (3 and Br) to the higher priority fragment of the double bond (1), fragment 3 has the higher priority. The connection to fragment 2 that lies on the same side of the double bond as fragment 3 is terminal atom H. The stereochemistry is thus specified {SB:1d2,3,H,Br,Br}. The complete MCLD descriptor is CBr;CBrH;CHHH<sup>2</sup>[#B3]\_{SB:1d2,3,H,Br,Br}.

## Multiple double bonds

The MCDL representation of multiple double bonds in a molecule consists of a sequence of double bond configuration descriptors (one for each of the double bonds), listed in descending priority order of the higher priority member of the double bond (smallest ASCII value first) and separated by semicolons. As in the case of compounds with multiple chiral centers, the generation of a unique double bond descriptor for compounds with multiple double bonds having certain symmetry elements can be complicated. For example, consider hexa-2,4-diene. The unique portion of this diene's linear descriptor is 4CH;2CHHH[2,3;4;5;6] regardless of the configurations of the double bonds. The *trans,\*trans\** diene (Figure 11) has only one possible numbering scheme, and its stereochemistry module is {SB:1d3,2,H,5,H;2d4,1,H,6,H}.

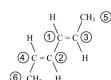


Figure 5.11: Figure 11. *Trans,\*trans\*-hexa-2,4-diene Transtrans.*

In contrast, the fragment priorities of *cis,\*trans\*-hexa-2,4-diene* can be assigned in two ways as shown in Figure 12. In the left drawing, the *trans* configuration occurs on the double bond between fragments 1 and 3 while in the right drawing the *trans* double bond is between fragments 2 and 4.

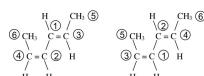


Figure 5.12: Figure 12. Two alternative MCDL numbering schemes for *cis,\*trans\*-hexa-2,4-diene cistransTwo alternative MCDL numbering schemes for .*

The two different numbering schemes give rise to stereochemistry specifications of {SB:1d3,2,H,5,H;2d4,1,6,H,H} and {SB:1d3,2,5,H,H;2d4,1,H,6,H}, respectively. To determine the correct descriptor, the *x* and *n* values of the two candidates are compared to each other beginning at the left. At the first point of difference, the descriptor having the higher priority *x* or *n* value is the correct one. The first difference in the two descriptors occurs at the third position where one has a structural fragment with priority value 5 and the other has the terminal atom H. As stated earlier, the structural fragment has a higher priority than the terminal atom. Thus, the correct stereochemistry descriptor is {SB:1d3,2,5,H,H;2d4,1,H,6,H}.

## Two- or three-substituent double bonds

Similar to atomic stereochemistry, MCDL double bond stereochemistry descriptors can be used for compounds containing three or two substituents attached to a double bond. In this case, absent substituents are replaced by “dummy” substituents having highest priority (0). Figure 13 provides the examples of *cis*- and *trans*-diaza-2-butene. The stereochemistry descriptor for *cis*-diaza-2-butene is {SB:3d4,,2,1}, and for *trans*-diaza-2-butene it is {SB:3d4,,2,,1}. Again note the absence of the digits for “dummy” substituents.



Figure 5.13: Figure 13. MCDL numbering of *cis*- and *trans*-diaza-2-butene for stereochemistry descriptor generation. “Dummy” substituents have the highest priority (0)

**cistransMCDL numbering of .**

### 5.3.4 Mixed stereogenic units (atoms and bonds)

Molecules containing both stereogenic atoms and stereogenic bonds represent a special case for the MCDL. To ensure the uniqueness of the stereochemical descriptor in this case, the following rule has been added: atoms have higher priority than bonds. A similar approach has been implemented in the recently introduced InChI chemical descriptors<sup>18</sup>. Therefore, when multiple MCDL numbering schemes are possible for these compounds, first consideration should be given to schemes that provide the highest priorities to atomic chiral centers after which the priorities of atoms in double bonds are considered. All possible numbering schemes should be considered during this iterative process (see above).

In general, the same principle applies for molecules containing more complex stereogenic units (e.g., a double bond stereogenic unit with two atoms is higher priority than an allene-type stereogenic unit with three atoms - see additional file

Additional file 1

**mcdl\_allene.** Stereochemical modules for allenes (2 pages).

[Click here for file](#)

### 5.3.5 Software implementation

#### LINDES 2.8

##### Stereochemistry algorithm

The major addition in the new release of the MCDL accompanying software, LINDES (version 2.8, see additional file

Additional file 2

**lindes28.** The source code of the C program “LINDES” (version 2.8, 47 pages).

[Click here for file](#)

The software can determine the stereochemistry of a chiral atom in a molecule if the input consists of a MOLFILE having 3D coordinates or else 2D coordinates with *up* and *down* bond designations for the chiral atom(s)<sup>19</sup>. The algorithm first orients the molecule so the substituents on the stereogenic atom are positioned as they would appear in a Fischer projection. This is accomplished by rotating the molecule (fragment) until one substituent on the chiral atom is aligned with the *+\*y\** axis and a second lies in the *yz* plane and has a negative *z* coordinate. The positions of the remaining two substituents are then identified by the signs of their *x* coordinates. This process is demonstrated in Figure 14, where the stereoisomer of CHBrClF shown at the left is rotated twice to give the orientation from which a Fischer projection can be drawn.

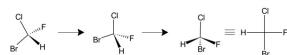


Figure 5.14: Figure 14. Rotation of CHBrClF stereoisomer to give Fischer projection orientation  
**Rotation of CHBrClF stereoisomer to give Fischer projection orientation.**

To determine the MCDL stereochemistry, the highest priority atom or fragment must be up, and the second highest must be down. In the present case, this can be accomplished by switching the positions of the Cl and Br. However, to maintain the same configuration as in the original Fischer projection, an EVEN number of switches is required. Thus, the H and F must also be switched (Figure 15). The final descriptor for this stereoisomer is {SA:1,Br,Cl,F,H}.

<sup>18</sup> The IUPAC International Chemical Identifier (InChI™)

<sup>19</sup> Description of several chemical structure file formats used by computer program developed at Molecular Design Limited

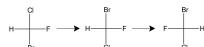


Figure 5.15: Figure 15. Atom switches in Fischer projection of CHBrClF stereoisomer to maintain configuration  
**Atom switches in Fischer projection of CHBrClF stereoisomer to maintain configuration.**

The software algorithm for determination of double bond stereochemistry works similarly to that described for chiral centers. The coordinates of the two double bonded atoms are needed along with the coordinates of one substituent on each of these. The double bond is oriented along the  $y$  axis and rotated to bring one of the substituents into the  $xy$  plane. The relative positions of the two substituents are identified by the signs of their  $x$  coordinates.

### Software limitations

Version 2.8 of LINDES is designed to handle molecules with “simple” stereochemistry: one or more stereocenters, one or more double bonds, or a combination of these. To avoid problems associated with molecular symmetry, the current implementation requires the four atoms and/or MCDL fragments **directly** attached to the chiral atom or double bond to differ (i.e., the difference can not occur in a substituent at a point removed from the stereogenic center). For example, the substituents CH<sub>2</sub>Br (MCDL fragment CBrHH) and CH<sub>2</sub>Cl (MCDL fragment CClHH) are recognized as different while CH<sub>2</sub>CH<sub>3</sub> and CH<sub>2</sub>CH<sub>2</sub>CH<sub>3</sub> are not since the MCDL fragment at the point of attachment is CHH in both cases. An exception is made for molecules containing chiral CH fragments (such as sugars). The stereochemistry of double bonds within rings is ignored. LINDES 2.8 does not determine the stereochemistry of chiral centers or double bonds with fewer than four substituents. In specific cases, stereochemistry descriptors can be generated with LINDES by adding “dummy” substituents (see discussion above).

As stated earlier, for symmetrical and quasi-symmetrical structures where multiple constitutionally equivalent MCDL numbering schemes are possible, all must be explored for selection of the unique (the lowest ASCII code sequence) stereochemistry descriptor. LINDES 2.8 does not perform this exhaustive search.

LINDES 2.8 is not designed to handle the disjointed structures of mixtures (e.g., salts). To create the MCDL descriptors of salts and other mixtures, each component must be drawn and processed independently [see Appendix 2], and the resulting component descriptors have to be combined according to MCDL rules. Also, LINDES software does not process MOLFILEs that have no bond block.

### 5.3.6 MCDL Java Chemical Structure Editor

#### Generation of structure diagrams from atomic MCDL stereodescriptors

If the stereoconfiguration of an atom is present in an MCDL descriptor, then it is reflected in the structure diagram using *up* and *down* bonds. The placement of these bonds is based on the criteria of simplicity, ease of understanding, and neat appearance of the resulting structure diagrams. The following rules have been developed to facilitate this task:

1. The maximum number of stereobonds attached to any particular atom should not exceed two. Stereobonds take a lot of space on the structure diagram, and it can become “overloaded” if more than two bonds are used to describe a chiral center (Figure 16, A).

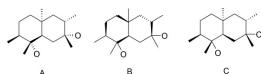


Figure 5.16: Figure 16. Different depictions of the stereochemistry in an alicyclic molecule with chiral centers  
**Different depictions of the stereochemistry in an alicyclic molecule with chiral centers.**

2. If the chiral atom is a part of a ring, a bond outside the ring is chosen to receive the *up* or *down* attribute, if possible (Figure 16, B and 16C). If there are no exocyclic bonds, then a bond within the ring is chosen. It is also desirable to choose a cyclic bond that has the minimal number of substituents (Figure 16, C).

Once a bond is selected for stereo representation, it is then necessary to determine whether the *up* or *down* attribute reflects the proper stereoconfiguration provided in the MCDL stereodescriptor. To solve this problem, the algorithm initially selects *up* for the attribute and then re-creates the MCDL stereodescriptor of the particular chiral center. If the re-created descriptor is different from the original, the attribute is changed to *down*.

The MCDL Java applet does not explicitly draw the positions of hydrogen atoms in its structure diagrams. However, in many cases *up* and *down* stereobonds terminated by hydrogen atoms are used to display the stereoconfiguration of a chiral atom. These situations require transfer of the stereochemical information from the bond terminating in hydrogen to another bond on the chiral atom. First, the applet algorithm performs a search for *up* and *down* bonds terminated by hydrogen. If any are found, the algorithm then performs a search for another single bond to this stereocenter with minimal angle to the stereobond terminated by the hydrogen atom. This new bond replaces the hydrogen-terminated bond as the stereobond.

### Generation of structure diagrams from double bond MCDL stereodescriptors

Similar to the atomic MCDL stereodescriptors, if the stereoconfiguration of a double bond is present in an MCDL descriptor, the specified stereoisomer is reflected in the structure diagram. In the original algorithm for structure diagram generation<sup>2</sup>, after generation of the initial 2D atomic coordinates, bond rotations were performed to remove overlapping. In the new algorithm, prior to any bond rotation, the 2D coordinates are used to calculate MCDL double bonds stereodescriptors, which are compared with the initial ones. If the descriptors are different, a 180-degree rotation around the double bond is performed. After the correct configurations of all the stereo double bonds are established, their relative coordinates are held fixed while other bonds are allowed to rotate to remove overlap of atoms and bonds. The lack of hydrogen atoms in the structure diagrams does not cause any problems since stereoconfiguration of a double can be determined using the positions of other substituents.

### Generation of atomic MCDL stereodescriptors from structure diagrams

The Java applet algorithm employs a fast, simplified procedure for identification of stereogenic atoms. The procedure is based on hash 8-byte variables<sup>20</sup> for stereogenic fragments, and is designed to generate canonical descriptors except for very rare complex cases involving multiple constitutionally equivalent fragments. This is demonstrated for the two structures shown in Figure 17. For structure A, in which all the chiral centers are part of symmetrical rings, the applet gives the canonical stereodescriptor {SA:2,7,10,12,H;3,8,10,H,13;4,11,14,16,H;5,11,15,H,17}. In contrast, for the acyclic structure of B, the applet gives a valid, but non-canonical stereodescriptor {SA:1,2,3,H,6;2,1,4,H,7;3,1,5,8,H}, where the canonical descriptor is {SA:1,2,3,6,H;2,1,4,7,H;3,1,5,H,8}. Identification of the specific cases where the simplified procedures produce non-canonical stereodescriptors is a difficult task beyond the scope of this article.

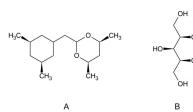


Figure 5.17: Figure 17. The test structures containing multiple stereocenters  
The test structures containing multiple stereocenters.

### Generation of double bond MCDL stereodescriptors from structure diagrams

To define the stereoconfiguration of a double bond, it is necessary to analyze the spatial arrangement of its four substituents (atoms or fragments). Stereodescriptors are not generated for cyclic double bonds. For acyclic double

<sup>20</sup> WinDat: An NMR Database Compilation Tool, User Interface and Spectrum Libraries for Personal Computers

bonds, the algorithm checks for two identical terminal groups or atoms connected to either end of the double bond using atom-centered topological indexes<sup>20</sup>. If these indexes are the same, the stereodescriptor is not generated. If the indexes are different on both ends of the double bond and a specific configuration is present in the diagram, the algorithm generates the stereodescriptor.

### 5.3.7 New MCDL file support in the Open Babel software package

Open Babel is a popular software package for conversion of chemical structure files from one format into others and as a C++ chemical toolkit<sup>21</sup>. The current version supports over 80 different chemical structure formats. Open Babel uses the SMARTS<sup>22</sup> language (SMILES<sup>3</sup> extension) for search and filtration of molecular structures. There are interfaces to other programming languages such as Perl and Python, which expand the applicability of Open Babel to other software development projects. Open Babel libraries are currently being used in more than 30 associated projects<sup>23</sup>. Therefore, support for the MCDL format in Open Babel provides a valuable opportunity to expand the usage of the MCDL.

Addition of the MCDL to Open Babel required the creation of new software modules. For example, chemical bond orders and atomic coordinates are not stored in many MCDL descriptors since this information is considered to be supplemental<sup>1</sup>. These structural parameters must be calculated during the format conversion process and molecular image generation since the majority of other chemical formats require them. Also, the existing C++ libraries of the Open Babel project did not contain modules for acyclic bond order reconstruction (*kekule.cpp* module is designed to handle aromatic bonds) and structure image generation. The required basic algorithms for bond order reconstruction and chemical structure image generation were taken from our previous effort<sup>2</sup> with appropriate modifications to fit the Open Babel specifications. The conversion capabilities to and from MCDL appear in Open Babel v2.3.0.

Methods for the generation of 2D coordinates derived from the Structure Editor significantly expand the utility of the Open Babel package. For example, structure image generation is now possible from other coordinate-less chemical structure formats, such as SMILES<sup>34</sup> and InChI<sup>5</sup>. In addition, new methods have been created (1) to check for overlapped atoms and bonds in a molecule; and, if found, to rotate the affected fragments 180 degrees around an acyclic bond or to increase the length of this acyclic bond in cases where the rotation does not work; (2) to generate the list of topologically equivalent atoms necessary to accelerate the overlapped fragment adjustment process; (3) to create the simplest image of chain structures, cycles, and condensed cycles; and (4) to calculate chiral characteristics of an atom<sup>24</sup>.

All the new classes and methods developed for MCDL inclusion in Open Babel have been written to comply with the Open Babel documentation<sup>25</sup> and are compiled in a separate plugin module to facilitate their use in Open Babel applications. The LINDES<sup>1</sup> program code (with the minor modifications such as using object-oriented methods, bond order reconstruction, and structure diagram generation procedures) is used in this module to execute the required MCDL format support functions.

## 5.4 Conclusions

### 5.4.1 Software testing

To facilitate testing of the MCDL Java Chemical Structure Editor using large databases, we developed a standalone utility *JAmodule* that can be used to process the data in batch mode. This module allows conversion of an SDF batch file into an MCDL batch file and vice versa. The original non-stereo Java algorithm<sup>2</sup> was tested by performing conversions of structure files from MOLFILE format into MCDL format and back. If everything worked correctly,

<sup>21</sup> Open Babel: The Open Source Chemistry Toolbox

<sup>22</sup> SMARTS - A Language for Describing Molecular Patterns

<sup>23</sup> Related Projects - Open Babel

<sup>24</sup> Atomic Chirality, a Quantitative Measure of the Chirality of the Environment of an Atom

<sup>25</sup> Open Babel

the final output should match the original input. The same procedure was used for testing the stereo Java algorithms presented in this paper.

The following differences were found in non-stereo testing of 20,000-records in the ChemDiv database<sup>26</sup>:

1. Porphyrines. Differences were found for 2 compounds from the set of 3 in the database. It was mentioned previously<sup>2</sup> that adequate representation of cyclooctatetraene-type structures capable of valence isomerism and valence tautomerism requires additional MCDL descriptors with information regarding specific bond order distribution in a molecule. The porphyrines belong to this class of compounds.
2. Differences for two other compounds were caused by erroneous structures. Both molecules contained atoms with illegal valences.

The structure diagram quality tests were performed using the Knovel database<sup>27</sup>, which contains a diverse set of organic compounds with applications in many different areas. To improve the graphic performance of the Java applet, we expanded the database of pre-defined templates<sup>2</sup> from 105 to 145. There was no overlapping in any of the structure diagrams. Visual observation showed that more than 90% of the structure drawings were of typographic quality.

Finally, we used the public domain NSC database<sup>28</sup> to test the accuracy of the Java MCDL stereodescriptor algorithm. Of the 42,247 records in the database, 5,188 structures contain *up* and *down* bonds and were initially selected for testing. However, manual checks found that many of these structures did not actually contain stereo elements, i.e., structures containing *up* and *down* bonds attached to non-chiral centers. These structures were removed to yield the final data set comprised of 2,418 records.

After repeated conversions of MOLFILE formats into MCDL formats and back, 16 structures had differences in stereo configurations. In most of these, the major problem was the poor quality of the initial drawing. For example, an almost linear configuration of bonds around  $sp^2$  atoms leads to ambiguity in determination of the Z/E configuration (Figure 18, 1). Sometimes the stereo notations *up* and *down* were used for illustrative purposes in symmetrical, non-chiral structures (Figure 18, 2). The Java algorithm recognition settings were developed with the assumption that the angles between the bonds of  $sp^2$  atoms would be approximately 120 degrees, and the bond distances would be nearly equivalent. It should be noted that the comparison of initial and final structures was examined by a modified CheD program<sup>29</sup> that uses somewhat different algorithms of structure drawing analysis compared to the Java applet.

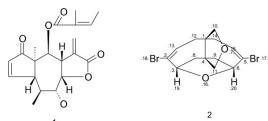


Figure 5.18: Figure 18. The graphical representation of chemical structures in the NSC database<sup>28</sup> for which determinations of stereo configurations of double bonds (1) and atoms (2) were difficult to accomplish

**The graphical representation of chemical structures in the NSC database** [#B28]\_\*\*for which determinations of stereo configurations of double bonds (1) and atoms (2) were difficult to accomplish\*\*.

The latest versions of MCDL Java Chemical Structure Editor public domain source codes and executables are deposited on SourceForge<sup>30</sup>.

## 5.5 Competing interests

The authors declare that they have no competing interests.

<sup>26</sup> ChemDiv

<sup>27</sup> Knovel - Technical Engineering Reference Information

<sup>28</sup> DTP - Human Tumor Cell Line Screen

<sup>29</sup> CheD: Chemical Database Compilation Tool, Internet Server, and Client for SQL Servers

<sup>30</sup> MCDL

## 5.6 Authors' contributions

AAG participated in design and coordination of the study including development of MCDL concept, and edited the manuscript for publication. MNB developed MCDL concept, designed LINDES and PREPROCESS software, and took part in preparation of the manuscript. SVT and AVY developed and tested MCDL Java Chemical Structure Editor and MCDL Open Babel software, and took part in preparation of the manuscript. All authors have read and approved the final manuscript.

## 5.7 Appendixes

### 5.7.1 Appendix 1

Basic molecular topology does not represent 3D features of molecular objects, so distinctively different 3D molecular objects may have identical MCDL composition and connectivity modules (e.g., conformers and stereoisomers). Unlike conformers, stereoisomers do not undergo inter-conversion in normal conditions due to restricted internal motion. For example, the high-energy barrier of rotation around double bonds leads to “cis/trans (E/Z)” isomerism. Similarly, the high-energy barrier of inversion of four-coordinated carbon atom leads to “L/D (R/S)” isomerism. Typically, this inter-conversion barrier should be at least 20-30 Kcal/mole for stereoisomers to exist as separable compounds at room temperatures. There are many examples where these energy barriers may be lower or higher depending on environment variables (solvents, pH). In addition, compounds within the same structural group may have a wide range of inter-conversion barriers so some of them can be considered as stereoisomers, and others - as conformers (see Table one in reference <sup>8</sup> as an example).

### 5.7.2 Appendix 2

An auxiliary program PREPROCESS (see additional file

Additional file 3

**preprocess.** The source code of the C program “PREPROCESS” (version 1.0, 6 pages).

[Click here for file](#)

## 5.8 Acknowledgements

This research was sponsored by the IPP program. Oak Ridge National Laboratory is managed and operated by UT-Battelle, LLC, under contract DE-AC05-00OR22725. The research at the Institute of Physiologically Active Compounds was performed under master contract DE-AC01-00N40184 with Kurchatov Institute for the U.S. Department of Energy. The authors gratefully acknowledge the efforts that contributed to the preparation of this paper, especially the valuable comments of Chris Morley and other members of Open Babel team. This paper is a contribution from the Discovery Chemistry Project.



# FLAME: FLASH MOLECULAR EDITOR - A 2D STRUCTURE INPUT TOOL FOR THE WEB

## 6.1 Abstract

### 6.1.1 Background

So far, there have been no Flash-based web tools available for chemical structure input. The authors herein present a feasibility study, aiming at the development of a compact and easy-to-use 2D structure editor, using Adobe's Flash technology and its programming language, ActionScript. As a reference model application from the Java world, we selected the Java Molecular Editor (JME). In this feasibility study, we made an attempt to realize a subset of JME's functionality in the Flash Molecular Editor (FlaME) utility. These basic capabilities are: structure input, editing and depiction of single molecules, data import and export in molfile format.

### 6.1.2 Implementation

The result of molecular diagram sketching in FlaME is accessible in V2000 molfile format. By integrating the molecular editor into a web page, its communication with the HTML elements on this page is established using the two JavaScript functions, *getMol()* and *setMol()*. In addition, structures can be copied to the *system clipboard*.

### 6.1.3 Conclusion

A first attempt was made to create a compact single-file application for 2D molecular structure input/editing on the web, based on Flash technology. With the application examples presented in this article, it could be demonstrated that the Flash methods are principally well-suited to provide the requisite communication between the Flash object (application) and the HTML elements on a web page, using JavaScript functions.

## 6.2 Background

### 6.2.1 About the Project

At present, there are various tools available for structure input on web pages, for example Marvin Sketch<sup>1</sup>, PubChem Sketcher<sup>2</sup>, NIST Molecule Editor<sup>3</sup> or JME<sup>4</sup>. The latter one is very popular because of its compact footprint and intuitive use. It has been utilized by many web developers (including the authors of this article) for sites with chemical content. Recently, a comprehensive review article<sup>5</sup> summarized the developments in this field. Prompted by the statement that so far no Flash-based web tool for structure input is available, we decided to undertake a feasibility study, aiming at the development of a compact and easy-to-use 2D structure editor, using Adobe's Flash technology and its programming language, ActionScript. As a reference model application from the Java world, we selected the Java Molecular Editor (JME). In this feasibility study, we made an attempt to realize a subset of JME's functionality in the FlaME utility (*part I, this article*). These basic capabilities are: structure input, editing and depiction of single molecules, data import and export in molfile format. As this is just the beginning of the project, we intend to continue and extend this work with additional features like reaction editing and the implementation of other data exchange formats (e.g., SMILES) (*part II, in planning*).

### 6.2.2 Objectives and expectations

The primary objective of this work has been to demonstrate a usable Flash application (written in Flash/ActionScript) for drawing and editing molecules directly within a HTML/JavaScript web page. The goals are as follows: a) we want to achieve the convenience of common structure editors such as ISIS (Symyx) Draw<sup>6</sup> or Marvin Sketch<sup>1</sup>, which are easy and intuitive to use; b) at the same time the application should be compact - in other words: small and smart. As the data exchange format (*result or output format*), the well-documented and proven MDL MOL format V2000<sup>7</sup> was chosen. This format is still very popular for structural data exchange and it is used by many databases and cheminformatics applications, despite its inherent limitations, e.g. with respect to the maximum number of atoms. The latter aspect, however, does not affect its suitability for FlaME, as this editor is primarily designed to handle "small molecules".

## 6.3 Results and discussion

### 6.3.1 The FlaME Molecular Editor: features - graphical user interface and layout

As mentioned above, the authors were inspired by the JME applet written by P. Ertl<sup>4</sup> when designing the layout of the graphical user interface (GUI). Nevertheless, we tried to create a typical and unique look-and-feel for our new tool (Figure 1). Actually, there should be no explanation needed for a novice user of FlaME. Users should immediately feel familiar with the buttons and functions in FlaME. Furthermore, tooltips are displayed in the right bottom area of the GUI which are activated by slowly moving the mouse pointer over the various buttons, thus facilitating use of the applet.

---

<sup>1</sup> Marvin Sketch

<sup>2</sup> The PubChem Project, PubChem Sketcher

<sup>3</sup> NIST Chemistry Webbook. NIST Molecule Editor

<sup>4</sup> JME Molecular Editor

<sup>5</sup> Molecular structure input on the web

<sup>6</sup> ISIS (Symyx) Draw/Accelrys Draw

<sup>7</sup> CT file formats (MDL MOL file)

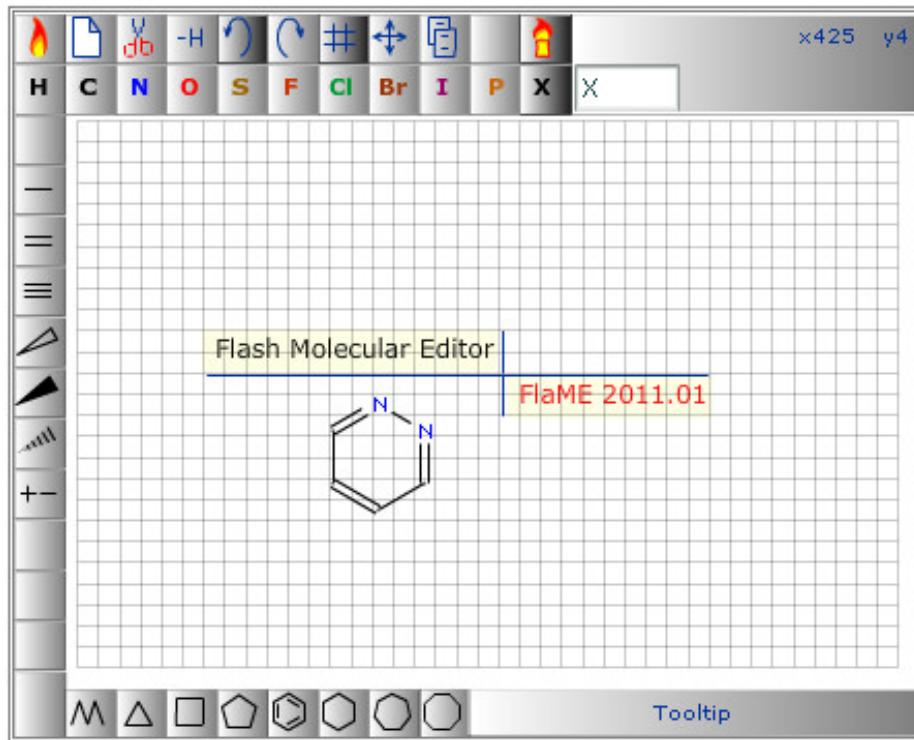


Figure 6.1: Figure 1. FlaME graphical user interface  
FlaME graphical user interface.

### 6.3.2 GUI buttons

The FlaME GUI is simple and all icons on the buttons should be intuitively understood. Figure 2 shows the icons and a description of the individual buttons.

### 6.3.3 Subjects and actions

Based on the definition of operations (drawing primitives) for molecular structure diagram sketching as discussed by A. M. Clark<sup>8</sup>, we can define the subjects in FlaME as atoms and bonds. The actions implemented so far are as follows:

- add or delete atom
- add or delete bond
- add structure fragment (as pre-defined *template structure*)
- set or change the atom type (as a pre-defined element or as input via X-button)
- set or change the bond order
- set or change the stereo style (*stereo marking*)
- set or change the ionic charge

The element or group label will change automatically depending on element and current environment of the particular structure node (*node environment*).

<sup>8</sup> Basic primitives for molecular diagram sketching



Figure 6.2: Figure 2. Description of the GUI buttons

### 6.3.4 Molecular structure output

The result of molecular diagram sketching in FlaME is accessible in V2000 molfile format<sup>7</sup>. This result can be copied to the *system clipboard* (by clicking the corresponding button). On the other hand, by integrating the molecular editor into a web page, its communication with the HTML elements on this page must be established. For this purpose, two JavaScript functions are implemented which are described below.

### 6.3.5 Flash Methods - JavaScript getMol() and setMol() functions

The Flash documentation states: “A Flash method is a JavaScript function that is specific to Flash movies. Use Flash methods to send JavaScript calls to Flash movies from a scripting environment.”<sup>9</sup>

Using the Flash methods, two functions (*getMol* and *setMol*) were written: the function *getMol()* ensures that the structure drawn in the molecular editor will be copied as a molfile to a *textarea* field (*id* = “mol”) within the web page for further use, e.g. as a query structure for a database search. The second JavaScript function *setMol()* sends the molfile from the web page to the Flash object, in our case to the FlaME molecular editor. The second line in this function just ensures compatibility with all common web browsers, such as Internet Explorer, Firefox, Opera, Chrome, Safari etc.

While the purpose of the *getMol()* function could be described as “*export as MOL file*”, the *setMol()* function could be called “*import MOL file*” and it can be used for structure input in textual form for further structural modification using FlaME (see example 1).

```
function getMol() {
    var m = thisMovie("flame").GetVariable("mol");
    document.getElementById("mol").value = m;
}

function setMol() {
    var m = document.getElementById("mol").value;
    m = m.replace(/\n/g, "r");
    thisMovie("flame").send2Flame(m);
}
```

### 6.3.6 Drawing operations and template structures

The actions of FlaME implemented so far do not reflect all of the basic primitives for molecular diagram sketching, described previously<sup>8</sup>. It was not the purpose of this tool to cover all these possible functionalities, but to provide users (and developers) with an opportunity to use a rather compact input tool for chemistry-related websites. The standard template structures included in FlaME are: cyclopropane, cyclobutane, cyclopentane, cyclohexane, benzene, cycloheptane, and cyclooctane. Once placed in the drawing area, these structures can be edited by changing the bond order (single, double and triple), by changing the atom type (element), or by changing the stereo style (stereo marking).

### 6.3.7 Use of a special key

The *Shift* key on the computer keyboard can be used to modify some of the drawing operations of FlaME: whereas ring fusion (annulation) by clicking with a ring template on an existing ring bond normally produces a fully expanded

---

<sup>9</sup> Macromedia Flash Support Center, Publishing and Exporting - Flash Methods

condensed system, pressing the *Shift* key while clicking will cause the new ring to be flipped to the opposite direction (Figure 3, structure a). Similarly, adding a ring template to an existing ring atom normally produces two rings connected by a single bond, whereas the *Shift* key modifies this behavior and leads to a spiro ring system (Figure 3, structure b). Furthermore, by pressing the *Shift* key the user can prevent the delete operation from removing terminal (“orphan”) atoms when a bond is deleted (Figure 3, structure c). The latter option can be used to create disconnected structures (e.g., salts) from a single molecule, thus providing a work-around for the current limitation of FlaME to handle only a single connection table.

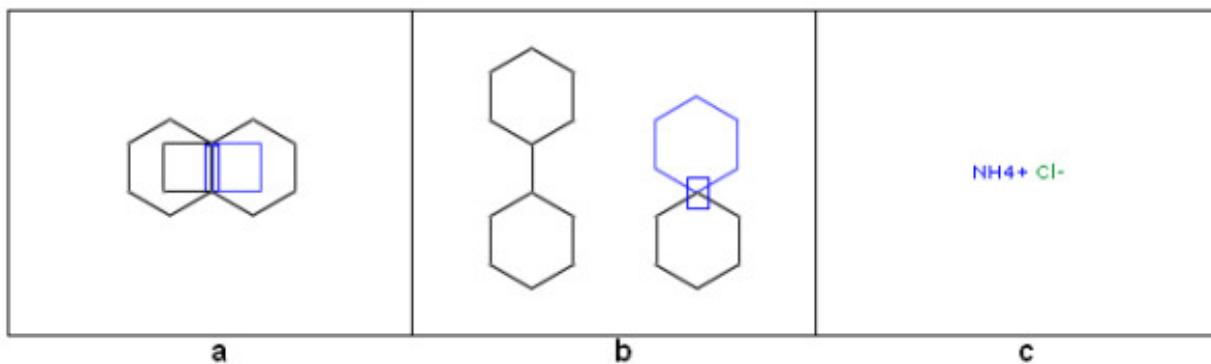


Figure 6.3: Figure 3. Using the Shift key while editing a molecular structure  
Using the Shift key while editing a molecular structure.

### 6.3.8 Features which are not yet implemented

Compared to the JME applet, FlaME still lacks a number of features which may be implemented in future versions. For instance, there is currently support for only one connection table at a time. Thus, there is no mode available for reaction input and depiction. Special definitions for query atoms are limited to A, Q, and X (see below), while R groups are not yet supported. Moreover, at present the only data exchange format is the MDL molfile format (V2000), whereas SMILES export is still on the to-do list.

## 6.4 Implementation

### 6.4.1 Development environment

Flash applications are frequently used for advertisements and as games. Very useful applications have been also developed and widely employed for e-learning purposes. Flash as well as Java and JavaScript can be used to add interactivity to web pages using the programming language ActionScript, which has a similar syntax and semantics as JavaScript (or, more generally, is a dialect of ECMAScript). The FlaME application itself was developed using Adobe Flash Professional CS4, which provides an integrated development environment. It is equipped with a compiler which can produce either compact SWF files suitable for use on web pages, or so-called Windows-projectors or Macintosh-projectors which are platform-specific executables.

FlaME is a single-file Flash application. It was written in ActionScript (AS2) and compiled as a Flash 8 SWF file. Thus, this application requires a common web browser (Firefox, IE, Opera, Chrome, Safari etc.) and the Flash Player plug-in at least in version 8. According to Adobe: “Today, over 86% of Internet-connected computers have adopted the Flash Player version 9”<sup>10</sup>.

<sup>10</sup> Adobe Flash website

Whereas there should be no principal (technical) problem to make use of an Adobe Flash Player also on Apple's mobile devices (iPhone, iPad and iPod touch), Apple has decided not to support Flash on these devices and therefore, applications such as FlaME cannot be used on them. However, essentially the same restriction applies to Java on the iPhone/iPad which also prevents Java-based structure editors from being employed on these mobile devices. Although there are some "inofficial" ways to install unsupported software on Apple's mobile devices, only a change in Apple's policy would lead to a general solution of this Flash/Java lack-of-support issue. It should be noted, however, that FlaME was developed for use on typical desktop computers rather than on mobile gadgets.

#### 6.4.2 Limitations

Because of the inherent limitations of the V2000 MDL molfile format (see the CTfile formats description<sup>7</sup>) which is used by FlaME for data exchange, the maximum number of atoms is limited to 999.

#### 6.4.3 FlaME application examples

In this section, three application examples are presented which have been built around FlaME and which are freely accessible on the web. A common web development practice is to use both an `<object>` tag and an `<embed>` tag to display Flash (SWF) content within an HTML page<sup>11</sup>. The nested-objects method requires a double *object* definition (the `<object>` tag targeting older versions of Internet Explorer and the `<embed>` tag targeting all other browsers), so it is necessary to define the object attributes and nested *param* elements twice<sup>12</sup>. The required attributes are: *classid* (the value is always as shown below), *type*, *width*, and *height*. The required *param* element is *movie*, which defines once more the URL of a SWF file. For up-to-date web browsers, one specifies the URL of the Flash application with the *src* attribute.

#### 6.4.4 Embedding SWF content (i.e., the FlaME application) in HTML

```
<object classid='clsid:d27cdb6e-ae6d-11cf-96b8-444553540000'
id='flame' height='438' width='350' align='middle'>
<param name='allowScriptAccess' value='sameDomain'/>
<param name='movie' value='flame.swf'/>
<param name='quality' value='high'/>
<param name='bgcolor' value='white'/>
<embed src='flame.swf' quality='high' bgcolor='#ffffff' width='438' height='350'
name='flame' align='middle' allowScriptAccess='sameDomain'
type='application/x-shockwave-flash'
pluginspage='<a href="http://www.macromedia.com/go/getflashplayer">_</a>'>
</object>
```

<sup>11</sup> Adobe Flash Platform - Embedding SWF content in HTML

<sup>12</sup> Embedding Adobe Flash Player content using SWFObject

## 6.4.5 Application example 1: Structure Search

This demonstration page<sup>13</sup> uses FlaME as a replacement for JME in a modified version of the open-source structure database system MolDB5R<sup>141516</sup>. The latter software employs the *checkmol/matchmol* program<sup>17</sup> as the search engine. Figure 4 shows a run in “similarity search” mode (using the Tanimoto index<sup>18</sup> derived from binary fingerprints as similarity criterion), Figure 5 shows another example in “substructure search” mode. Like with JME, special symbols (to be entered via the X-button) can be used for the definition of query atoms:

The screenshot shows the MolDB5R demo interface for structure search. At the top, there's a toolbar with various icons for file operations and a molecular editor. Below the toolbar is a legend for special symbols: H, C, N, O, S, F, Cl, Br, I, P, and X. The main workspace displays a complex organic molecule with a hydroxyl group and a branched chain. To the right of the workspace is a panel titled "MolDB5R demo: structure search". It includes a "Structure input on web:" section with a link to "FlaME: Flash Molecular Editor". Below this is a "Special symbols (to be entered via X-button):" section with definitions for A, Q, and X. Further down is a "Structure Search - Query - checkmol" section with a "Mode" dropdown set to "similarity search" (indicated by a green dot). There are also "Options" checkboxes for "strict atom/bond type comparison" and "check configuration (E/Z and R/S)". A "Search [S]" button is at the bottom of this panel. At the bottom left, there's a "Text input form (MDL molfile format)" with buttons for "Submit [T]", "Get MOL [G]", and "Show MOL [M]". The bottom right shows a scrollable list of "Hit structure(s)" with the first entry being "2:2447 6,6,9-trimethyl-3-pentylbenzo[c]chromen-1-ol (1.0000)". Below this list is a detailed view of the same molecule as in the workspace.

Figure 6.4: Figure 4. MolDB5R similarity search using FlaME for query input  
**MolDB5R similarity search using FlaME for query input.**

<sup>13</sup> FlaME: Flash Molecular Editor, application examples

<sup>14</sup> MolDB5 homepage

<sup>15</sup> Creating a Web-based, Searchable Molecular Structure Database Using Free Software

<sup>16</sup> Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: an Open-Source Approach

<sup>17</sup> The checkmol/matchmol homepage

<sup>18</sup> IBM Internal Report 17th Nov 1957, *taken from* Daylight Theory: Fingerprints - Screening and Similarity

**MolDB5R demo: structure search**

Structure input on web:

[FlaME: Flash Molecular Editor](#)

Special symbols (to be entered via X-button):

A: any atom except H  
 Q: any atom except H and C  
 X: any halogen atom

**Structure Search - Query - [checkmol](#)**

Mode:

exact search  
 substructure search  
 similarity search

Options:

strict atom/bond type comparison  
 check configuration (E/Z and R/S)

**Search [S]**

Text input form (MDL molfile format)

Submit [T]   Get MOL [G]   Show MOL [M]

**Hit structure(s)**

2:2170 [5-[(3-azaniumyloxycarbonyl-4-hydroxyphenyl)-(3-azaniumyloxycarbonyl-4-oxocyclohexa-2,5-dien-1-ylidene)methyl]-2-hydroxybenzoyl]oxyazanium

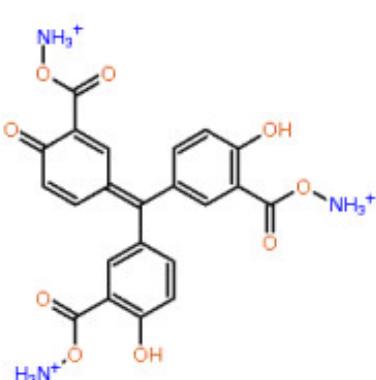


Figure 6.5: Figure 5. MolDB5R substructure search using FlaME for query input  
**MolDB5R substructure search using FlaME for query input.**

A for any atom except H

Q for any atom except H and C

X for any halogen atom

Explicit \*\*H\*\*ydrogens are entered via a separate H button.

#### 6.4.6 Application example 2: Presentation of structure collections in “slide show” mode

This web page (Figure 6) uses the JavaScript function *setMol()* to send the molfile of the selected structure to FlaME. The user can either send any of the listed structures to FlaME by clicking on the respective chemical name in the right area (printed in arbitrary color) or he/she can start a “slide show” by selecting the “play” button below the list. In the latter case, the data transfer process runs automatically with a pre-defined time interval.

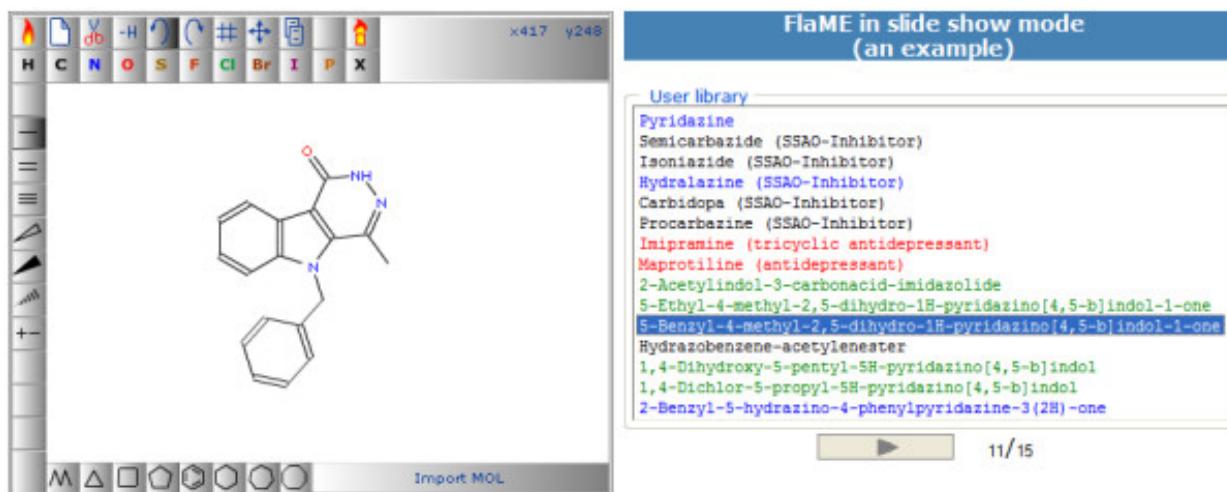


Figure 6.6: Figure 6. FlaME demo page in “slide show” mode  
FlaME demo page in “slide show” mode.

#### 6.4.7 Application example 3: FlaME editor vs. depiction

This example (Figure 7) shows the FlaME editor (left side) in combination with a depiction-only version of the software (right side). In the latter incarnation, the “move structure” mode is activated by default, as well as “zoom in” and “zoom out” functions which are accessible on a web page via keyboard shortcuts or via HTML elements (buttons) connected to appropriate JavaScript functions.

This depiction-only version of FlaME was compiled separately without any GUI buttons and layout options. By doing so, the size of the SWF file could be reduced to approximately one third (25 kB) of the original file size of the editor version.

### 6.5 Conclusions and Outlook

A first attempt was made to create a compact single-file application for 2D molecular structure input/editing on the web, based on Flash technology. At present, the project should be regarded as a feasibility study and any feedback (comments and suggestions for further development and extensions) will be welcome.

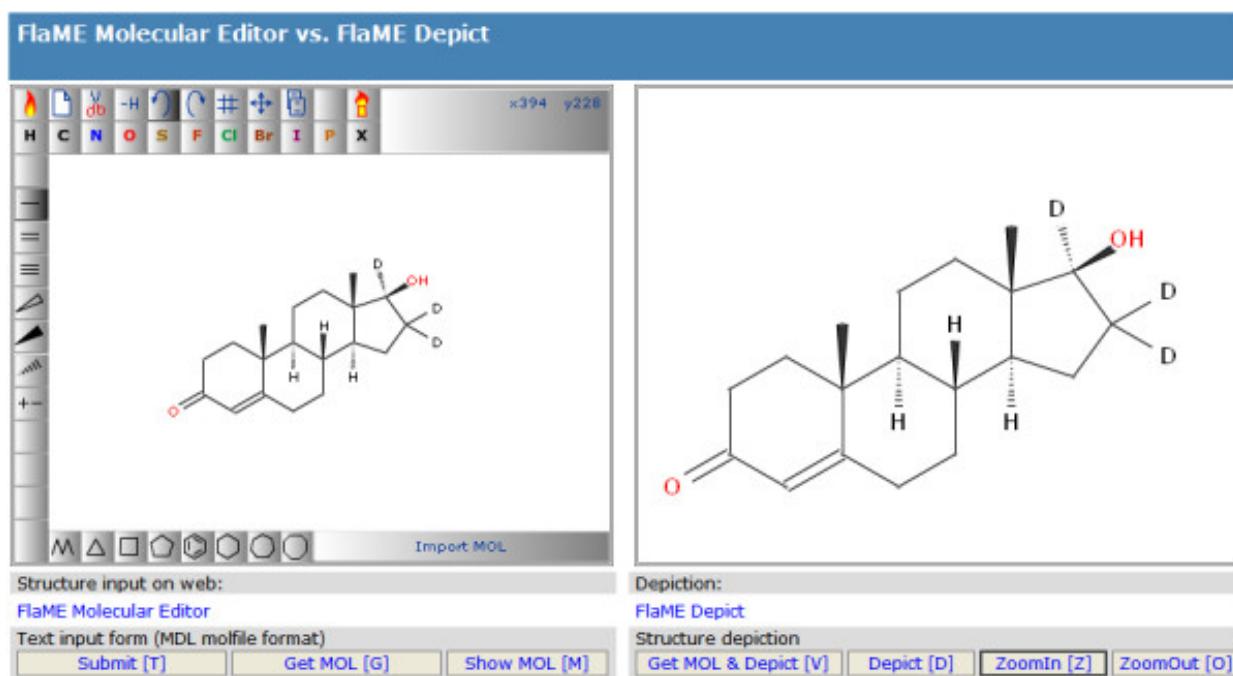


Figure 6.7: Figure 7. FlaME: Molecular Editor vs. Depiction  
**FlaME: Molecular Editor vs. Depiction.**

With the introduction of the Flash Molecular Editor (FlaME) as an alternative to similar input tools based on Java technology, we would like to start a discussion about the suitability (advantages and limitations) of the Flash environment for interactive chemistry-related websites. With the application examples presented in this article, it could be demonstrated that the Flash methods are principally well-suited to provide the requisite communication between the Flash object (application) and the HTML elements on the web page, using JavaScript functions.

The next version of FlaME should include a reaction editor and some other extensions, for instance generation of SMILES<sup>19</sup> output. For future developments, we envisage a separate Flash application for handling spectral data, e.g. for JCAMP-DX spectra viewing/editing. In this field, there are already some nice plug-ins and Java applets available (e.g., MDL Chime<sup>20</sup>, JSpecView Applet<sup>21</sup>, or the NIST JCAMP-DX Viewer<sup>22</sup>), which could be taken as models for writing a comparable Flash-based tool.

### 6.5.1 Use and Reference

Developers of chemistry-enabled websites are invited to try FlaME on their pages. The utility [Additional file<sup>23</sup>] would be desirable. However, at present the latter product is still far away from its goal to be a complete web-based animation suite.

Additional file 1

**FlaME - Flash Molecular Editor.** Platform independent SWF application file

Click here for file

<sup>19</sup> SMILES. 2. Algorithm for Generation of Unique SMILES Notation

<sup>20</sup> MDL Chime

<sup>21</sup> The JSpecView Project: an Open Source Java viewer and converter for JCAMP-DX, and XML spectral data files

<sup>22</sup> NIST Chemistry Webbook. NIST JCAMP-DX Viewer

<sup>23</sup> AJAX Animator

## 6.6 Availability and requirements

Project name: FlaME: Flash Molecular Editor - a 2D structure input tool for the web

Project homepage and application examples URL: <http://synthon.pch.univie.ac.at/flame/>

Application download URL: <http://synthon.pch.univie.ac.at/flame/flame.swf>

Operating system(s): Platform independent

Programming language: Flash, ActionScript 2

Other requirements: Adobe Flash Player 8 or higher

License: individual license on request (free of charge).

Any restrictions to use by non-academics: none

## 6.7 Competing interests

The authors declare that they have no competing interests.

## 6.8 Authors' contributions

PD was involved in the design and programming of the FlaME software, in the implementation of the software and preparation of the application examples, and the manuscript preparation. NH was involved in testing and intellectual guidance during FlaME software development, in preparation of application examples, and manuscript preparation. All authors read and approved the final manuscript.

## 6.9 Acknowledgements

We thank all interested colleagues at our department for helpful discussions during the development and suggestions for further extension of FlaME.

# USE OF STRUCTURE-ACTIVITY LANDSCAPE INDEX CURVES AND CURVE INTEGRALS TO EVALUATE THE PERFORMANCE OF MULTIPLE MACHINE LEARNING PREDICTION MODELS

## 7.1 Abstract

### 7.1.1 Background

Standard approaches to address the performance of predictive models that used common statistical measurements for the entire data set provide an overview of the average performance of the models across the entire predictive space, but give little insight into applicability of the model across the prediction space. Guha and Van Drie recently proposed the use of structure-activity landscape index (SALI) curves via the SALI curve integral (SCI) as a means to map the predictive power of computational models within the predictive space. This approach evaluates model performance by assessing the accuracy of pairwise predictions, comparing compound pairs in a manner similar to that done by medicinal chemists.

### 7.1.2 Results

The SALI approach was used to evaluate the performance of continuous prediction models for MDR1-MDCK *in vitro* efflux potential. Efflux models were built with ADMET Predictor neural net, support vector machine, kernel partial least squares, and multiple linear regression engines, as well as SIMCA-P+ partial least squares, and random forest from Pipeline Pilot as implemented by AstraZeneca, using molecular descriptors from *SimulationsPlus* and AstraZeneca.

### 7.1.3 Conclusion

The results indicate that the choice of training sets used to build the prediction models is of great importance in the resulting model quality and that the SCI values calculated for these models were very similar to their Kendall  $\tau$  values, leading to our suggestion of an approach to use this SALI/SCI paradigm to evaluate predictive model performance that

will allow more informed decisions regarding model utility. The use of SALI graphs and curves provides an additional level of quality assessment for predictive models.

## 7.2 1. Background

The use of biological property predictions has increased in recent years, due to improvements in computer technology, the rising costs of drug discovery, and a desire by regulatory agencies to better understand, predict and improve drug safety <sup>1234567</sup>. The increase in computer processing speed has allowed very complex and computationally intensive models to be developed and run on desktop computers. Model building is becoming more of a routine practice, and will likely continue to grow in importance. However, the pace of proliferation of models and model building tools has not been matched by development of approaches and tools to rigorously assess their performance.

In a recent set of papers, Guha and Van Drie have proposed the Structure Activity Landscape Index (SALI) <sup>89</sup> as an approach to better assess biochemical structure-activity relationship (SAR) model performance. The concept is derived from the observation that activities based on specific interactions (as in receptor binding) do not change linearly with linear changes in properties. For example, while building an SAR, increasing the length of an alkyl substituent may result in a 0.3 log increase in potency per carbon, when one to three carbons are added; however, addition of the fourth carbon may increase the potency by 1 log unit, constituting what Guha and Van Drie have referred to as an “activity cliff”. By identifying these activity cliffs within an SAR set the SALI procedure can improve the understanding of where the model is more or less accurate.

Quantification of the structure activity cliffs is carried out by pairwise comparisons of compounds and their related measured and predicted activities, an approach similar to that routinely taken by medicinal chemists in the generation of SAR. In practice, the activity differentials are normalized by their structural similarity measures (for example, Tanimoto similarities). A small SALI value is indicative of a smooth activity transition, whereas a large SALI value indicates the presence of an activity cliff. The SALI graph of the dataset is a representation of its SAR as a connected graph, with molecules as the nodes and the  $SALI_{ij}$  values as edges. Plotting the sum of the nodes normalized by the edges versus the normalized threshold for edge detection generates a SALI curve. The value of the curve at  $X = 0$  ( $S(0)$ ) provides the ability of the model to capture all of the edges while the value at  $X = 1$  ( $S(1)$ ) is the ability of the model to correctly identify the most significant activity cliffs.

Because of the recognized importance of transporters in drug absorption, distribution and elimination <sup>10</sup>[#B11]\_\**in vitro*\* and *in silico* models to predict transporter involvement have been established. The MDCK-MDR1 *in vitro* model uses an immortalized mammalian cell line that stably expresses the transfected human MDR1 gene product (P-glycoprotein (P-gp)) to assess the potential for P-gp mediated efflux of compounds.

MDR1 efflux is dependent on specific structural features of the transported molecule, a key requirement for the application of the SALI/SCI approach. The general features of this interaction have been described by Anna Seelig <sup>11</sup>. Seelig determined that the likelihood of a molecule to bind to and to be an efflux substrate is based on the presence and the intermolecular distance between 2 hydrogen bond acceptors. Theoretically, incorporation of these 3D descriptors should increase the accuracy of efflux predictions.

*SimulationsPlus* ADMET Predictor software enables models to be built using neural networks (ANNE), support vector machine (SVM), multiple linear regression (MLR) and Kernel Partial Least Squares (KPLS) approaches. In addition,

---

<sup>1</sup> Computer-aided drug discovery and development (CADD): *in silico*-chemical-biological approach

<sup>2</sup> *In Silico* tools streamline drug design

<sup>3</sup> Recent advances in computational prediciton of drug absorption and permeability in drug discovery

<sup>4</sup> ADMET *in silico* modeling: Towards prediction paradise?

<sup>5</sup> Building predictive ADMET models for early decisions in drug discovery

<sup>6</sup> Predicting human pharmacokinetics from preclinical data

<sup>7</sup> *In silico* prediction of drug safety: Despite progress there is abundant room for improvement

<sup>8</sup> Assessing how well a modeling protocol captures a Structure-Activity Landscape

<sup>9</sup> The Structure-Activity Landscape Index: Identifying and quantifying activity cliffs

<sup>10</sup> Regulation of drug-metabolizing enzymes and transporters in infection, inflammation, and cancer

<sup>11</sup> A general pattern for substrate recognition by P-glycoprotein

*Umetrics* SIMCA-P+ provides a partial least squares (PLS) prediction using principal component analysis (PCA). Finally, AstraZeneca has implemented the *Accelrys* Pipeline Pilot random forest (RF) model builder.

The results of SALI, S(0), S(1), Kendall  $\tau$ , and MAE assessments of various model building approaches for their utility in predicting MDR1 mediated efflux as measured in an *in vitro* assay using a training set derived from the same *in vitro* data set are compared here.

## 7.3 2. Experimental

### 7.3.1 Datasets

MDR1-MDCK efflux data were generated from compounds synthesized and tested in support of drug discovery projects at AstraZeneca (Wilmington DE) using an MDCK-MDR1 transwell assay. All chemicals used in the assay were of at least reagent grade. The assay was conducted using MDR1-MDCK cells seeded at a density of 60,000 cells/well in DMEM medium with Glutamax into Millipore 96 well plates, to final volumes of 100  $\mu$ L on the apical side and 310  $\mu$ L on the basolateral side. Cells were grown for 3 to 5 days at 37°C in a humidified 5% CO<sub>2</sub> atmosphere, with daily medium changes including a final medium change two hours prior to running the assay. At the initiation of the efflux assay, the media on both the apical and the basolateral sides was replaced with the same volume of warmed Hank's Balanced Salts Solution with or without 1  $\mu$ M test compound. The cells were incubated at 37°C for 2.5 h, and then the concentrations of compound in the apical and basolateral side were quantitated by LC/MS/MS using standard analytical methods. Cell layers were tested for integrity by addition of 100  $\mu$ M Lucifer Yellow in Hank's Balanced Salts Solution to the transwell chamber corresponding to the apical side of the cell monolayer. The fluorescence in the basolateral chamber was quantified after incubation for 1 hour at 37°C. Only data from wells with Lucifer Yellow readings demonstrating less than 0.5% cell leakage were reported. Experimental runs were accepted based on the performance of standards with known efflux ratios.

The following criteria were used to select the compounds in the training and test set: successful generation of SMILES by the SMILES generation algorithm; absence of non-organic elements (a prerequisite of the ADMET Predictor); efflux ratio 0.7; coefficient of variation for replicates 50% and elimination of censored data. Prediction models were initially generated using ANNE and RF. The structures of the compounds and measured efflux values of outliers from both models were subsequently examined for potential inconsistencies; those compounds with suspect data based on assay irregularities were then removed from the data set. The outcome of this process was a set of 818 compounds with efflux ratios between 0.7 and 119. The distribution of the efflux values in the final dataset is shown in Figure 1.

The dataset was separated into training and test sets using two different approaches. The *ADMET Predictor* software uses Kohonen mapping<sup>12</sup><sup>13</sup> to assign compounds to training and test sets. *ADMET Predictor* uses the same descriptors for generation of the Kohonen map and for the model building. Alternatively, compounds were assigned random numbers between 0 and 1 using the Excel "RAND" function. The data were sorted descending on the random numbers; the top thirty percent (245) of compounds were selected as test compounds, while the remaining 513 compounds were assigned as the training set. In this case, the maximum and minimum measured efflux values were included in the training set to eliminate prediction model extrapolation.

To test the different modeling approaches, models were built using the same training/test set data. A Kohonen map was constructed and used for all 2D or 3D models. The file listing the Kohonen map of the compounds was opened using Excel, and the listing of compounds along with their status as training/verify/test was used to define the training (Kohonen training/verify) and test (Kohonen test) values for the RF and PLS models.

<sup>12</sup> Analysis of a simple self-organizing process

<sup>13</sup> Self-organizing formation of topologically correct feature maps

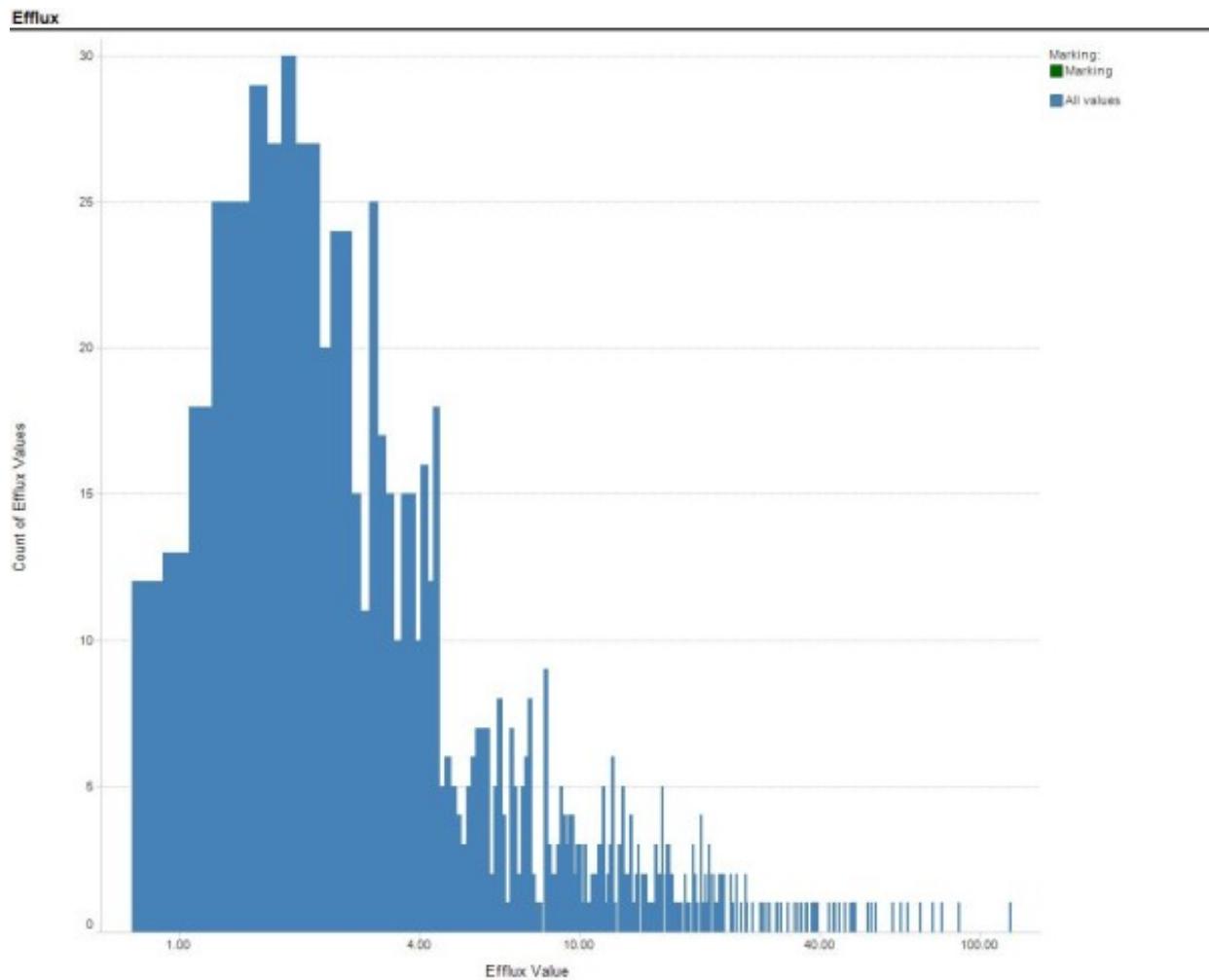


Figure 7.1: Figure 1. Distribution of Efflux Values used for Training Prediction Models  
**Distribution of Efflux Values used for Training Prediction Models.** Count of SALI values per bin versus SALI value

### 7.3.2 Model Building

#### 2D descriptors

Using ADMET Predictor descriptors and Kohonen map, ANNE, SVM, MLR, KPLS, RF and PLS models were built. In all cases, the same compounds (based on the Kohonen map) were used for test and training/(verify) sets. The ADMET descriptors (325) consisted of molecular weight, number of rings (total, aromatic, aliphatic), numbers of specific functional groups, geometric descriptors (moments of inertia, radii of gyration, surface areas), atomic partial charges, number of heteroatoms, fraction of bonds (single, double, triple), charge, hydrogen bonding descriptors, molecular ionization. The ADMET descriptors did not contain Formula, N\_Nonorgn, N\_Metal, N\_Kekule, S\_unknown, **Un-known**, AcidAtoms, or BaseAtoms, and therefore these descriptors were not included in the descriptors used for RF and PLS models. The *ADMET Predictor* application automatically removed highly-correlated descriptors, resulting in a final descriptor set of 134 descriptors actually used in ADMET model building. Feature selection was manually performed by PCA analysis for SIMCA P+ PLS models.

Alternatively, a set of 196 molecular and electronic descriptors utilized within AstraZeneca (AZ descriptors, ANNE AZ) were used. These descriptors consisted of lipophilicity, hydrogen bonding, size and shape, charge, polarity, atom counts, topology and druggability. The ADMET Predictor model building software automatically performed feature selection, resulting in a final descriptor set of 76 descriptors. Feature selection was manually performed in by PCA analysis for SIMCA P+ PLS models.

#### 3D descriptors

Using 3D sd files generated by VIDA3 (OpenEye), 3D descriptors were calculated using *ADMET Predictor*. The notable additional 3D descriptors were principal moments of inertia, second order static moments, solvent accessible areas, and those described by Seelig for identification of compounds that interact with the MDR1 transporter. Seelig's descriptors are: Type I units containing two electron donor groups with a spatial separation of  $2.5 \pm 0.3$  Å or Type II units consisting of two electron donor groups with a spatial separation of  $4.5 \pm 0.6$  Å or three electron donor groups with a spatial separation of  $4.5 \pm 0.6$  Å between the outer groups. After feature selection, a descriptor set consisting of 149 descriptors was used to build models in ADMET Predictor; all descriptors were used for RF and PLS. Feature selection was manually performed in by PCA analysis for SIMCA P+ PLS models.

### 7.3.3 Prospective model verification

To more fully evaluate the performance of the models, a set of 96 compounds was tested prospectively in the MDR1-MDCK efflux assay. This set consisted of AZ compounds that were not in the projects used for the training/test set and which met the same selection criteria as described above for the test/training compounds.

The 96 compounds were evaluated using the both 2D and the 3D descriptor sets. All models, with the exception of the random forest model, provided an assessment as to when compounds did not fall within the prediction space of the model. Only those compounds that fell within the prediction space of the respective prediction model were used for the final evaluations of model performance. The final number of compounds varied from 80 to 93, depending on the prediction model. The Pipeline Pilot RF model as implemented at AstraZeneca does not provide a flag for results not within the prediction space of the model; therefore, the RF model was not tabulated in the final analysis of the prospective data set.

After predictions, the parameters MAE, Kendall  $\tau$ , SCI, S(0) and S(1) were calculated as described below.

## 7.4 3. Methods

Mean Absolute Error (MAE) and Kendall  $\tau$  were calculated using JMP 7. SALI curves and SCI values were calculated as described in Guha and Van Drie<sup>89</sup>, using Daylight Type Fingerprints to generate Tanimoto similarity scores, and

Excel VBA programming to generate the SALI graphs, SCI and S(X) values and SALI curves.

## 7.5 4. Results and Discussion

### 7.5.1 Comparison of ADMET Predictor, in house RF and SIMCA PLUS PLS using 2D descriptors

#### Training sets

The distribution of the calculated SALI values are shown in Figure 2. All models with the exception of RF had similar MAE, Kendall  $\tau$  and S(0) values (Table 1). The results from the RF model suggested that both its accuracy (MAE) and rank ordering (Kendall  $\tau$ , S(0), SCI) properties were better than the other models tested. Using SCI as an indicator of model performance revealed marked differences between the models. MLR and RF had high SCI values approaching 1, suggesting that they could accurately rank order compounds across most of the edges of the SALI graph. ANNE, SVM, KPLS, PLS and ANNE AZ had SCI values between 0.1 and 0.5, indicating that these models would predict the edges accurately more often than inaccurately. Notably, ANNE with the random training/test set had a negative SCI value suggesting that it mispredicted more frequently than it predicted correctly. All S(1) values, except that for ANNE Random, were 1, indicating that all of the models were able to correctly rank order compounds with large SALI values (that is large activity changes with small structural changes) except ANNE Random.

Examination of the SALI curves clarified the reason for these results (Figure 3a). All models performed similarly at  $X < \sim 0.1$ . Thus, in those instances where changes in structure resulted in activity changes of a proportional magnitude, all of the models performed equally well. At  $X$  values greater than 0.3, MLR and RF could accurately order compound pairs across edges. The other models performed less well. Importantly, at  $X = 0.65$ , the S(X) for ANNE Random became negative, indicating that this model performed poorly at larger  $X$  values, completely mispredicting the nodes.

#### Test sets

The test set results, based on compounds within the same chemical space as the training set data, suggested that the model performances were roughly similar. The MAE, Kendall  $\tau$  and S(0) values were comparable in all models, including the three ANNE models. Examination of the SCI revealed three groups of values: RF, ANNE AZ, SVM, MLR and KPLS all had SCI values greater than 0.9; PLS and ANNE had SCI values between 0.7 and 0.9 while ANNE Random had a negative SCI value. Again, these results were explained by examination of the SALI curves (Figure 3b). It is clear that the different models predicted the order across the SALI edges with varying degrees of accuracy, depending on the  $X$  value. Of particular note is the observation that ANNE Random again had a negative S(X) value, crossing S(X) = 0 at an  $X$  of 0.2; thus, over most of the prediction space this model mispredicted the rank ordering.

Based on these results, there was no *a priori* link between the training and test set values. This may be somewhat expected, since the training sets predicted themselves, whereas the test sets represented unknowns. Thus, the training set performance would present a biased view of the model performance.

Also note that the S(0) and Kendall  $\tau$  values presented essentially the same measure of the model's ability to properly rank order compounds (S(0) vs. Kendall  $\tau$ , slope = 0.98, R<sup>2</sup> = 0.97, data not shown). Kendall  $\tau$ , by definition, measures the strength of the relationship between two variables <sup>14</sup>. In addition, since the interpretation of Kendall's  $\tau$  in terms of the probabilities of observing the agreeable (concordant) and non agreeable (discordant) pairs is very direct, S(0) would yield similar results if it also has that property.

---

<sup>14</sup> Practical Non-Parametric Statistics

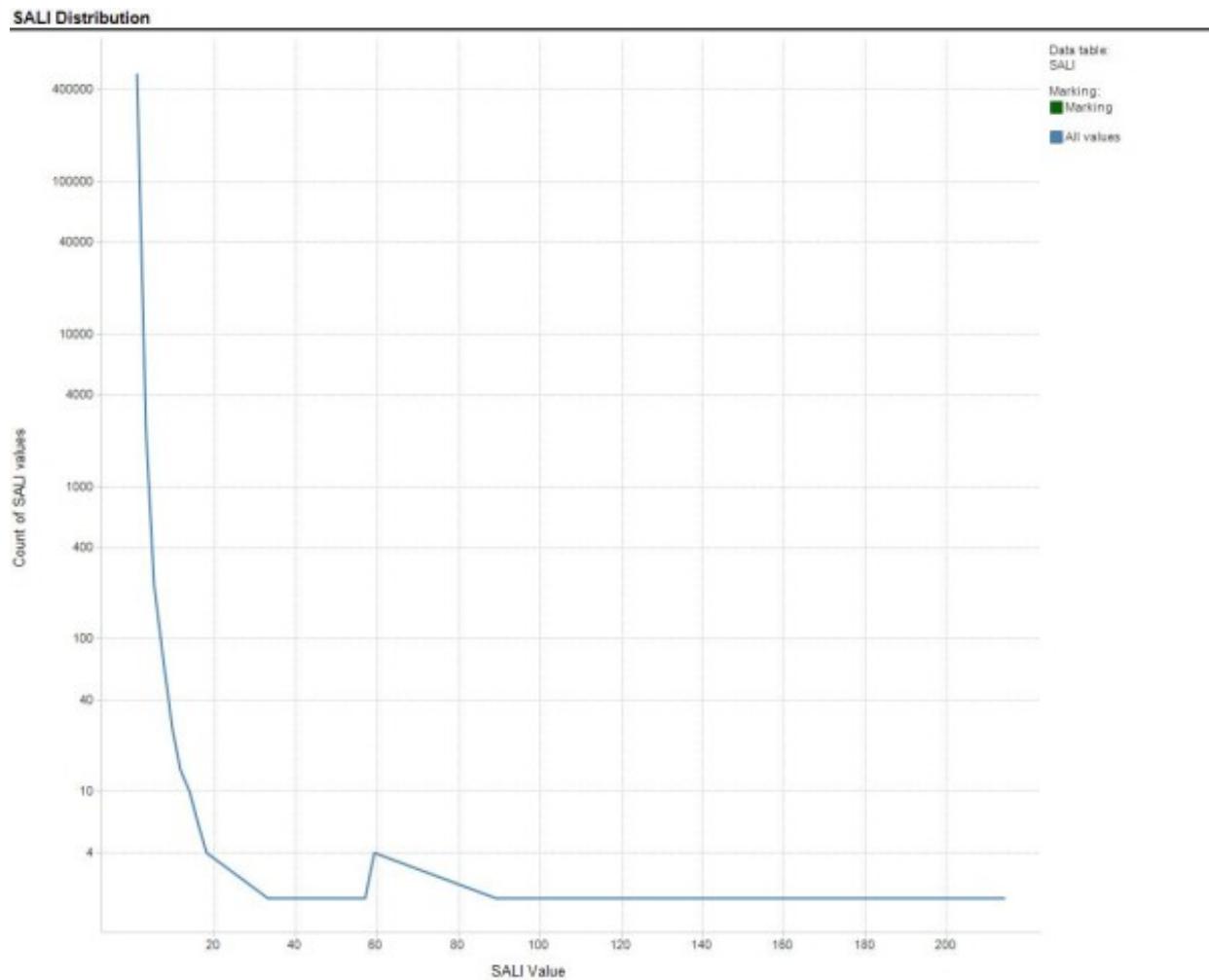


Figure 7.2: Figure 2. Distribution of SALI Values Calculated for 2D Training Sets  
**Distribution of SALI Values Calculated for 2D Training Sets.** Count of efflux value per bin versus efflux value

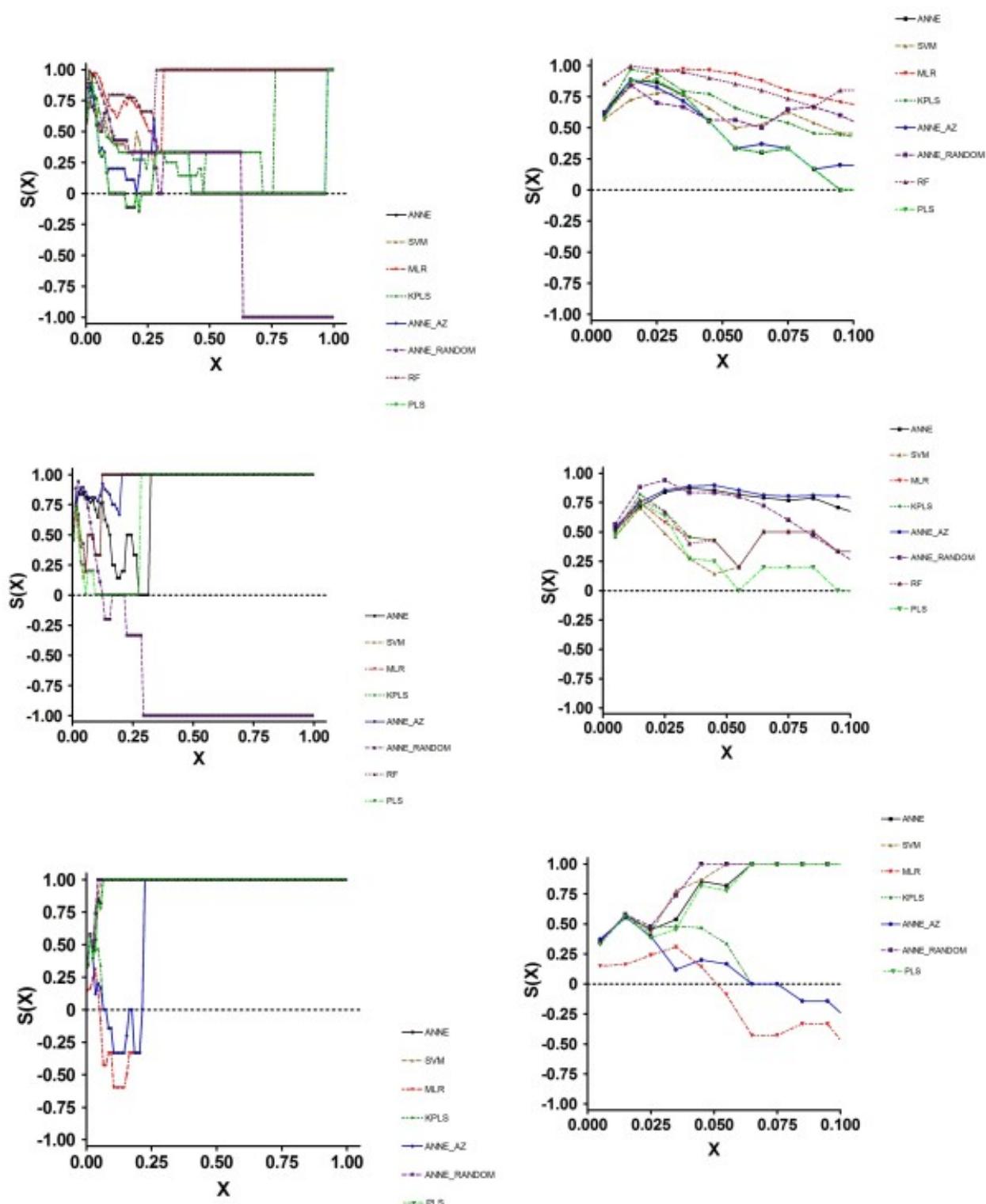


Figure 7.3: Figure 3. SALI Curves for Efflux Prediction Data Generated using 2D Descriptors  
**SALI Curves for Efflux Prediction Data Generated using 2D Descriptors.** 1a, Training Set; 1b, Test Set; 1c, Prospective Set.

## Prospective sets

In order to better assess the model quality, a prospective set of compounds which were not in the chemical space (that is, they were marketed drugs and drugs being developed within AstraZeneca for different targets and based on different scaffolds than the compounds used in the training set) of the training/test set compounds was tested in the MDR1-MDCK assay, and was also evaluated in each of the predictive models. Though these compounds were not in the same chemical space as those used in the training and test sets, their Tanimoto distances from the training set were similar to the Tanimoto distances of the test set compounds (Figure 4). Because the RF model as implemented with AstraZeneca does not identify prediction outliers, it was not used in this evaluation. With the prospective set, the Kendall  $\tau$  and S(0) values were generally around 0.35 with the exception of MLR, where the values were approximately 0.15. Thus, the models' abilities to rank order compounds overall were moderate, with the exception of MLR which was poor. The MAE values, on the other hand, showed larger differences. KPLS had the lowest MAE value of 0.19, suggesting that on average the KPLS will accurately predict the efflux value within a factor of 5% for this data set. MLR mean prediction error (0.52) was within a factor of 3, while the remaining models had mean errors (~0.3) of approximately a factor of 2. Efforts to correlate training set SCI values with any of the prospective set quality measurements were unsuccessful (data not shown).

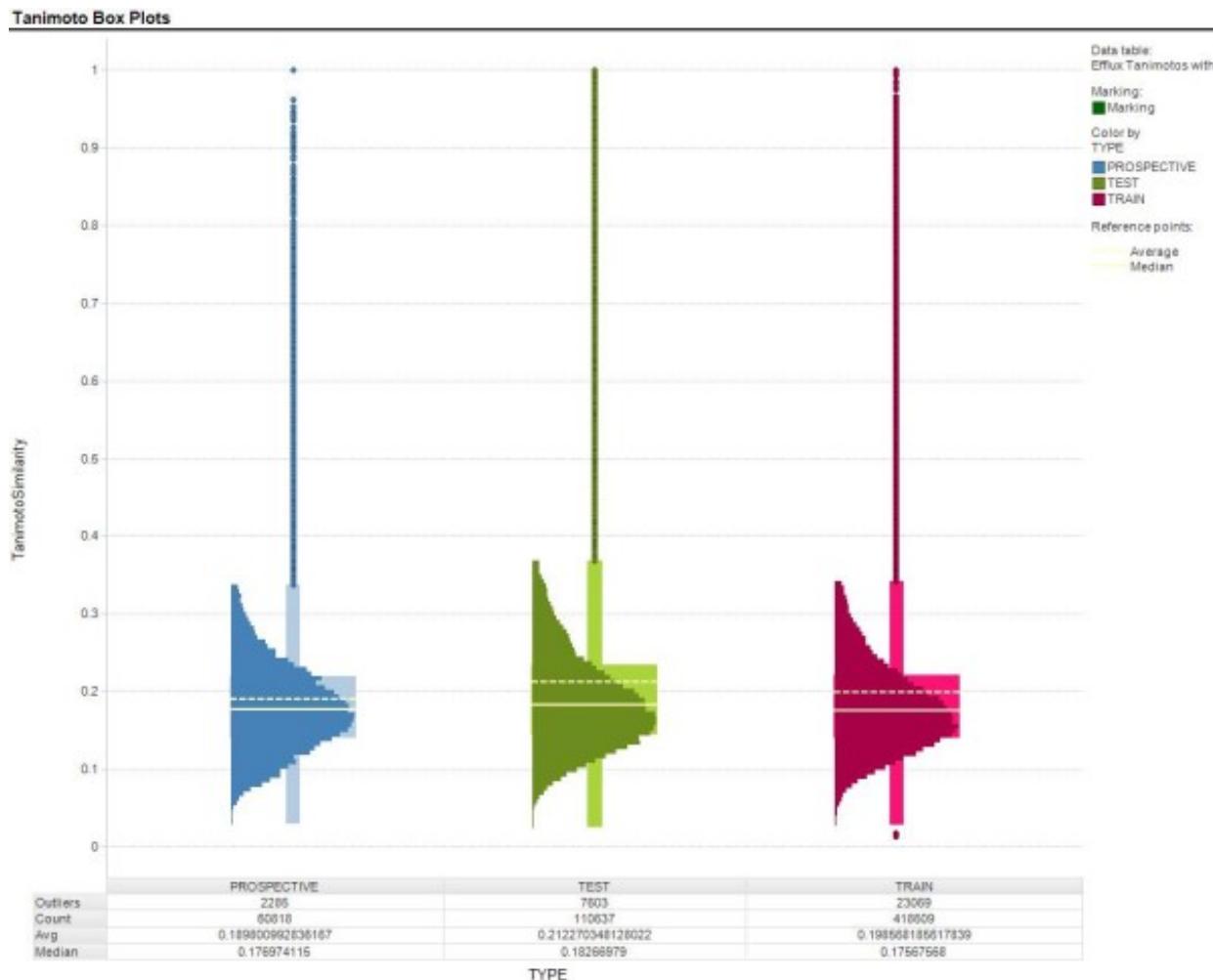


Figure 7.4: Figure 4. Box Plots of Daylight FP Scores of Training, Test and Prospective Sets  
**Box Plots of Daylight FP Scores of Training, Test and Prospective Sets.** Solid line, median Tanimoto score; dashed line, average Tanimoto score. Box shows Q1-Q3. Dots are outliers (greater than upper quartile + 1.5 times interquartile range)

As with the training and test set results, the SCI values told a different story (Figure 3c). Here, ANNE, SVM, PLS and ANNE Random were very good at accurately ordering compound pairs across nodes. MLR, KPLS and ANNE AZ had similar albeit somewhat lower ability to correctly order compounds. Again, examination of the SALI curves explained the SCI values. At all X values larger than approximately 0.1, the four best models had a prediction ordering accuracy of 100%. MLR, KPLS and ANNE AZ had S(X) values < 0 between X = 0.1 and 0.25. These data suggested that, for accurately predicting efflux values, KPLS would be the model of choice. On the other hand, to more accurately rank order compounds, ANNE, SVM, PLS and ANNE Random would be the better choices.

### 7.5.2 Comparison of ADMET Predictor, in house Random Forest and SIMCA PLUS PLS using 3D descriptors

#### Training sets

The results of these studies are summarized in Table 2 All models with the exception of RF had MAE values of approximately 0.2, and Kendall  $\tau$  and S(0) values approximating 0.6. Similar to the results with 2D descriptors, the RF model had a lower MAE and higher Kendall  $\tau$  and S(0) values.

Examination of the SCI values shows that RF had a value of 0.99, suggesting that this model was excellent at predicting edges. All other models had SCI values between 0.65 and 0.9 with the exception of PLS with a value of 0.13. Examination of the SALI curves (Figure 5a) reveals a qualitative difference in the curves models tested 2D vs. 3D descriptors. Whereas the curves from models with 2D descriptors differed across the entire range of X values, the models based on 3D descriptors showed large differences at X values between 0 and 0.4, then reached S(X) = 1 at larger values of X. This suggests that the ADMET Predictor 3D descriptors significantly improve the predictability of models across the entire range of X, particularly at large SALI values and in the case of efflux, should be included when possible.

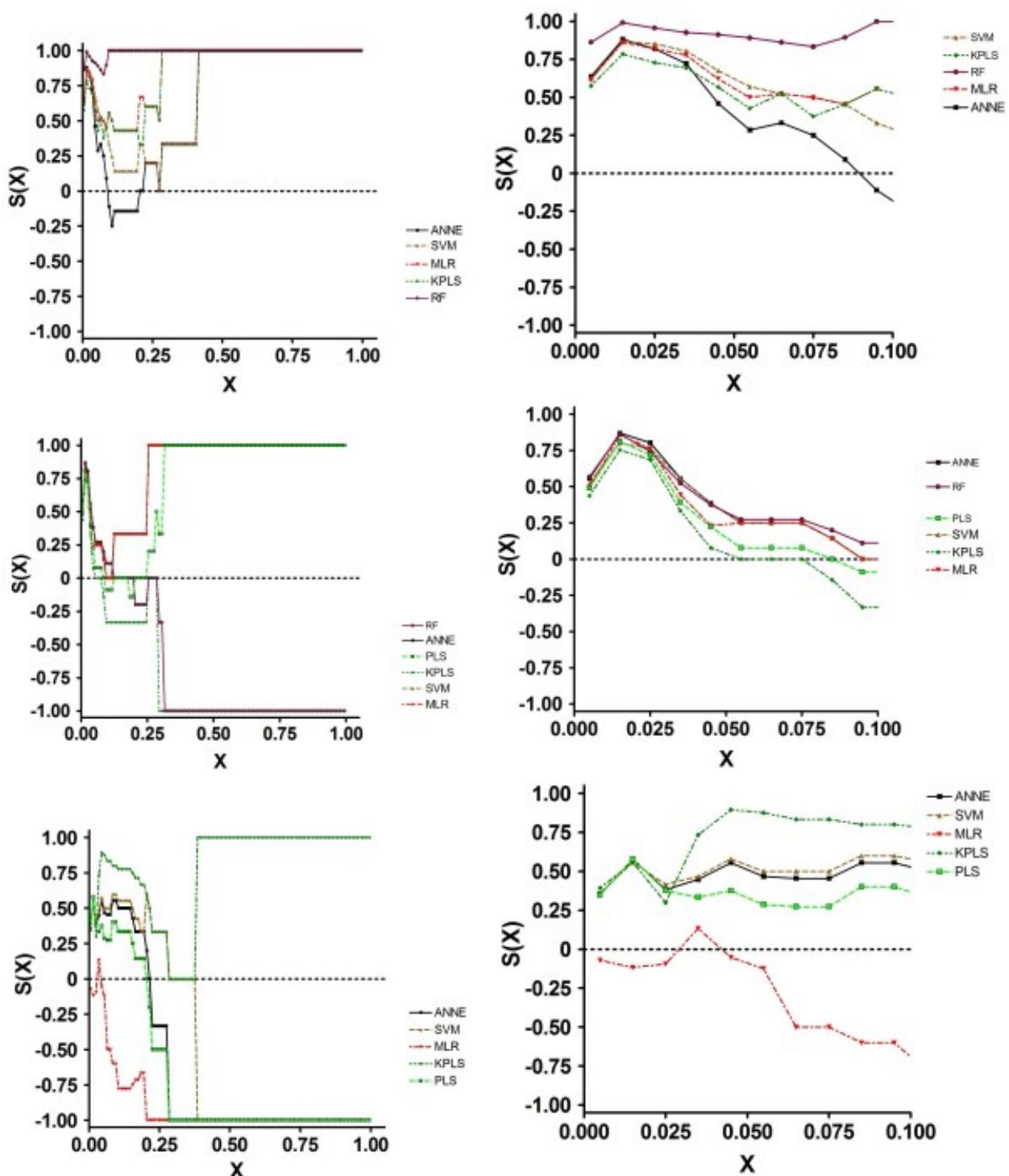
#### Test sets

Based on MAE, all models were generally similar, with KPLS showing a slightly greater inaccuracy. Kendall  $\tau$  values suggested that ANNE and RF had slightly better performance, while KPLS performance was somewhat less optimal. As with all of the data sets, the S(0) values agreed with the Kendall  $\tau$  values. In contrast to the statistical and S(0) values, the SCI results showed a wide range of values. Thus, ANNE, SVM and MLR appear to be superior to all of the other models; PLS had a somewhat lower quality, while RF and KPLS were poor in their quality. Examination of the SALI curves (Figure 5b) demonstrated the reason for the differences in model quality. All models reached S(X) = 0 at an X of approximately 0.1. ANNE, SVM and MLR increased to S(X) = 1 as X increased to 0.25, while PLS S(X) reached 1 at X = 0.35. However, at an X of 0.15, both RF and KPLS were at or below 0 (random ordering of compound pairs across the SALI edges) and attained -1 by X = 0.3. Thus, over most of the SALI space, both RF and KPLS mispredicted the order across the edges.

#### Prospective sets

The same prospective compound set used to evaluate the 2D model performance was used as a benchmark for the 3D models. These results indicate that all models with the exception of MLR had comparable performance based on MAE, Kendall  $\tau$  and S(0) values. In fact, these values were almost identical across the models. In contrast, MLR had markedly worse performance for these parameters. The MAE value was about twice that of the other models. The SCI curves (Figure 5c) indicated that all models reached an S(X) = 0 at some point; only KPLS returned to positive S(X) values and reached 1 at X = 0.4. All other models had negative S(X) values across most of the X range. These results indicate that only KPLS was capable of correctly rank ordering compounds in this prospective data set; the other models mispredicted the order.

Standard approaches to the evaluation of prediction model quality entail the examination of overall statistical properties, such as RMSE, MAE and Kendall  $\tau$ . While these approaches will provide an overall view of how the model may



**Figure 7.5: Figure 5. SALI Curves for Efflux Prediction Data Generated using 3D Descriptors**  
**SALI Curves for Efflux Prediction Data Generated using 3D Descriptors.** 1a, Training Set; 1b, Test Set; 1c, Prospective Set.

perform across the entire span of chemical space, they do little to define the quality of the model within the predictive space being examined. As a consequence, one's decision regarding the best model to use will be based on incomplete information about how and where best to use the model. In particular, one will have no knowledge as to whether the model works only in the "linear" range of the prediction space or will be able to effectively extrapolate out to more "nonlinear" regions when activity cliffs are present. Using the SALI approach, one obtains both a qualitative (graphical) and quantitative (SCI, S(0), S(1)) assessment of the model performance enabling better utilization and confidence in the predictive power of the models.

Based on standard statistical measures (MAE, Kendall  $\tau$ ), the performance of various prediction approaches (with the exception of RF) on the training sets with 2D descriptors was largely similar. Thus, to predict exact values, any of the models would suffice. However, The SCI revealed marked differences in performance when the goal is rank ordering of compound pairs, a very useful comparison when predicting in novel chemical space to instruct drug discovery. Because MLR and RF showed very high SCI values, either of these two models would be expected to be useful for predicting efflux. At the other extreme, the neural net model built using a random selection method for the training and test sets would not be expected to be useful, since this model was unable to accurately rank order compounds across the SALI edges. In addition, these results suggest that the selection of the training and test sets significantly impacts the quality of the resultant prediction model as would be anticipated from general predictive modeling experience.

When evaluating prediction models, test set data are typically used to assess model performance with unknowns. The test set standard statistical results had similar though slightly lower quality results than those observed with the training sets. Based on these results, choosing which model performed better would be difficult, since all seem to have similar prediction accuracy. However, the SCI clearly differentiates the model performance. ANNE Random is expected to be significantly inferior to all other models, in agreement with the results seen with the training sets. ANNE and PLS, while acceptable, are of lower quality than the remaining models. Based on the combined training and test set results, it can be concluded that MLR and RF would be expected to perform best within this prediction space.

One weakness in the particular data used for model building was that the training set and test set compounds generally came from the same chemical series. Ideally one would want compounds from outside of the chemical space than that used to build the model to better assess model performance. In an effort to address this concern, compounds from different chemical series were chosen as a prospective test set. In this case, exclusion of compounds which were identified by the prediction software as outside of the prediction space was found to result in an improvement in the calculated model performance values, indicating that the prediction software accurately identified the limits of its predictive ranges. With the 2D descriptors, the standard statistical measures did demonstrate differences in model performance. KPLS had the best overall accuracy while MLR had the worst with all others being similar. The Kendall  $\tau$  and S(0) values were approximately 0.35 for all models except MLR, where they were about 0.15. However, this was not the case with rank ordering. Neither KPLS, MLR nor neural nets with AZ descriptors was effective at rank ordering compounds while neural nets with ADMET descriptors, SVM and PLS were effective.

In agreement with previous studies, the selection of training set and test set compounds used for the model building is critical<sup>15</sup>. *SimulationsPlus* used Kohonen mapping to select the training and test sets, a methodology which seeks to effectively map the property space delineated by the proscribed descriptors. This approach was found to be more effective than randomly selecting the training and test sets. This was evident from the difference in SCI and S(1) values observed with the neural net models built using *ADMET Predictor* using the Kohonen map vs. random selection. The model with randomly selected sets generated the lowest values in both the training set, where model performance should be "the best" and in the test set.

Generally, there was not a large difference in model performance for prediction of efflux in the prospective data set when using 2D vs. 3D descriptors as determined by MAE, Kendall  $\tau$ , or S(0). There was, however, a large difference in the SCI values in the training sets. The training set results, though they predict themselves, provide a picture of the best possible outcomes. The observation that the SCI values in the training sets using 3D descriptors for efflux when building prediction models are larger than those built using 2D descriptors suggests that the 3 D descriptors will allow more accurate mapping of the SALI landscape, which should translate into higher quality models.

It is important to point out that the SALI range for the data sets varied according to the particular data set. The maximum SALI value for the training, test, and prospective data sets were approximately 200, 200 and 75, respectively,

---

<sup>15</sup> Generation of QSAR sets with a self-organizing map

even though the prospective set had a slightly greater spread of Tanimoto values (Figure 4) Thus, the SALI summary values (SCI, S(0), S(1)) represent the SALI landscape covered by that data set only. This is very important in model evaluation especially in the context of one project or chemical series. However, it would be less useful for generalizing the model quality over the entire SALI landscape covered by the training set since one is only looking at a portion of the entire prediction range and may be presenting a biased view of the model quality. A better representation of the model performance overall is provided by an appropriately chosen test set which more accurately reflects the SALI landscape.

Calculation and evaluation of SALI and SCI values could be incorporated into any model building paradigm. One such approach to the use of these parameters would be as follows. Various prediction models would be built using the desired descriptors, training sets, and calculation engines. The SALI values for the resultant test and training sets should be calculated, and the models with the largest SCI, S(0), and S(1) values would be chosen for continued evaluation and implementation. As one generates measured and predicted data on compounds within a project, one would use the SALI approach to continuously monitor the performance of the model for predicting the desired property. When the SCI, S(0), and/or S(1) values fall below some critical threshold, the model should be updated with the new data to improve the model performance. This will allow one to continually have knowledge of the applicability of the model in the desired chemical space.

There is still limited knowledge of the applicability of SCI to prediction model evaluation. The limitations of this approach have been summarized previously (Guha and Van Drie<sup>89</sup>). First and foremost is the requirement for a property driven by specific molecular interactions. Also, standard physical chemical properties (solubility, logD) are not readily amenable to SCI evaluation as they are not likely to have activity cliffs. Finally, more work needs to be done to identify the optimal comparison paradigm for assessment of SALI. In the present report, the SALI and SCI values were all internal to the individual data set used, resulting in different maximal SALI values. The predictive power of this approach may be improved by comparing the performance of the test and prospective sets with those of the training sets. This would likely result in the same maximum SALI value for all data sets allowing more direct comparisons of model performance with the various data sets.

## 7.6 5. Conclusions

Regulatory and competitive changes in drug discovery are driving an increase in the use of predictive sciences to speed up the development process, reduce costs, and improve safety. Though hardware and software improvements have facilitated and spread the use of prediction models, methodologies for evaluation of the model performance have not kept pace. The recent publications by Guha and Van Drie<sup>89</sup> have presented a novel approach to model performance evaluation in the context of SAR.

We have applied the SALI approach to evaluate several models for predicting efflux in MDR1-MDCK cells. The results presented here support the utility of this approach in the evaluation of model performance. Several observations here were identified as being important. First, use of SALI identified models which were better at predicting the relative order of compounds across SALI edges for small and large SALI values. This information, coupled with the SALI curves, allows evaluation of the utility of the model for correctly identifying large activity cliffs, a common occurrence in biochemical SARs. This contrasts to standard statistical approaches which will merely produce one overall number that does not discriminate large and small activity cliffs and, therefore, provides no guidance in model performance for ‘nonlinear’ processes.

Second, the aggregate SCI value was observed to be different from and complementary to accuracy measures such as MAE. Thus, one could potentially end up with models that would more accurately predict values, but would not necessarily do as good a job at identifying activity cliffs or at correctly ordering compounds across SALI edges. While the former property is certainly valuable, the latter property would be more generally utile when expanding into unknown chemical space where knowledge of the absolute value of a property may be less important than knowledge of “is this better or worse”.

The use of structure-activity landscape indices (SALI) and the SALI curve integral (SCI) was found to be very powerful in the evaluation of performance models, particularly with respect to rank ordering of compound pairs. In particular, this approach allows one to evaluate a model’s utility in the detection of large activity cliffs, a common occurrence

within drug discovery. It is the recommendation of these authors that this approach be incorporated as a quantitative and qualitative step in the evaluation of prediction models.

## 7.7 6. Abbreviations

SALI: Structure-Activity Landscape Index; SCI: SALI Curve Integral; ANNE: neural net; SVM: support vector machine; KPLS: kernel partial least squares; PLS: partial least squares; RF: random forest; MDR1-MDCK: multi-drug resistance gene 1 transfected Madine-Darby canine kidney cells; MAE: mean absolute error.

## 7.8 7. Competing interests

The authors have no competing interests.

## 7.9 8. Authors' contributions

NCLJr collated the data, performed all calculations and results interpretations, guided the design of the VBA calculation engine; KR designed and built the VBA calculation engine; JB reviewed all efflux experimental data to verify experimental integrity; LT designed, wrote and validated the robotics procedures which enabled the execution of the efflux experiments. All authors have read and approved the final manuscript.

## 7.10 9. Acknowledgements

The authors would like to acknowledge Mr. Peter Prokopiw and Ms. Cindy Shen for running the MDR1-MDCK efflux experiments and Dr. Doug Burdette for reviewing the manuscript and providing valuable comments.

# CONFAB - SYSTEMATIC GENERATION OF DIVERSE LOW-ENERGY CONFORMERS

## 8.1 Abstract

### 8.1.1 Background

Many computational chemistry analyses require the generation of conformers, either on-the-fly, or in advance. We present Confab, an open source command-line application for the systematic generation of low-energy conformers according to a diversity criterion.

### 8.1.2 Results

Confab generates conformations using the ‘torsion driving approach’ which involves iterating systematically through a set of allowed torsion angles for each rotatable bond. Energy is assessed using the MMFF94 forcefield. Diversity is measured using the heavy-atom root-mean-square deviation (RMSD) relative to conformers already stored. We investigated the recovery of crystal structures for a dataset of 1000 ligands from the Protein Data Bank with fewer than 1 million conformations. Confab can recover 97% of the molecules to within 1.5 Å at a diversity level of 1.5 Å and an energy cutoff of 50 kcal/mol.

### 8.1.3 Conclusions

Confab is available from <http://confab.googlecode.com>.

## 8.2 Introduction

The generation of molecular conformations is an essential part of many computational analyses in chemistry, particularly in the field of computational drug design. Methods such as 3D QSAR, protein-ligand docking and pharmacophore generation and searching<sup>1</sup> all require the generation of conformers, whether on-the-fly (as part of the method) or pre-generated by a stand-alone conformer generator. In contrast to 3D structure generators (such as CORINA<sup>2</sup>,

---

<sup>1</sup> Conformations and 3D pharmacophore searching

<sup>2</sup> Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures

DG-AMMOS<sup>3</sup> and smi23d<sup>4</sup>), which focus on the generation of a single low-energy conformation, conformation generators create an ensemble of conformers that cover the entire space of low-energy conformations or that part of conformational space occupied by biologically-relevant conformers.

Several proprietary conformation generators are currently available (including OMEGA<sup>5</sup>, ROTATE<sup>6</sup>, Catalyst<sup>7</sup>, Confort<sup>8</sup>, ConfGen<sup>9</sup>, Balloon<sup>10</sup> and MED-3DMC<sup>11</sup> among others) but only recently have open source conformation generators appeared: Frog2<sup>12</sup> generates conformers using a Monte Carlo approach, while Multiconf-DOCK<sup>13</sup> adapts the systematic search code from DOCK5<sup>14</sup> to generate diverse conformers via a torsion-driving approach.

Confab 1.0 is the first release of Confab, an open source conformation generator whose goal is the systematic coverage of conformational space. Accuracy has been favoured over the introduction of approximations to improve performance. The algorithm starts with an input 3D structure which, after some initialisation steps, is used to generate multiple conformers which are filtered on-the-fly to identify diverse low energy conformers. Conformations are generated using the torsion-driving approach from a set of predefined allowed torsion angles. Ring conformations are not currently sampled.

The first section of the paper describes the algorithm used by the software and some implementation details. After this, two applications of the software are described: an analysis of the conformational space of a dataset of 1000 molecules (which includes a comparison to Multiconf-DOCK), and an investigation of the conformational preferences of a particular phenyl sulfone.

## 8.3 Methods

### 8.3.1 Algorithm

The Confab algorithm is outlined in Figure 1. The input required is a 3D structure with reasonable bond lengths and angles. Since the algorithm does not currently explore ring conformations, any rings present should be in reasonable conformations.

The first step of the algorithm is the identification of rotatable bonds. These are defined as all acyclic single bonds where both atoms of the bond are connected to at least two non-hydrogen atoms, but neither atom of the bond is sp-hybridised. Note that this definition excludes rotation around bonds that interchange hydrogens (for example, the rotation of the hydrogens of a methyl group), but this does not imply any loss of accuracy as it is usual practice to exclude hydrogens when calculating the RMSD (see below).

The method used by Confab to generate conformations is known as the torsion-driving approach. A set of allowed torsion angles for each rotatable bond is assigned to each bond by searching for a match to predefined SMARTS strings in a user-configurable file (torlib.txt) included in the Confab distribution. This file is part of the Open Babel project and it assigns values to particular rotatable bonds using data from Huang et al.<sup>15</sup>.

<sup>3</sup> DG-AMMOS: A New tool to generate 3D conformation of small molecules using Distance Geometry and Automated Molecular Mechanics Optimization for in silico Screening

<sup>4</sup> smi23d

<sup>5</sup> Conformer Generation with OMEGA: Algorithm and Validation using High Quality Structures from the Protein Databank and Cambridge Structural Database

<sup>6</sup> Impact of Conformational Flexibility on Three-Dimensional Similarity Searching Using Correlation Vectors

<sup>7</sup> Catalyst

<sup>8</sup> Confort

<sup>9</sup> ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers

<sup>10</sup> Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm

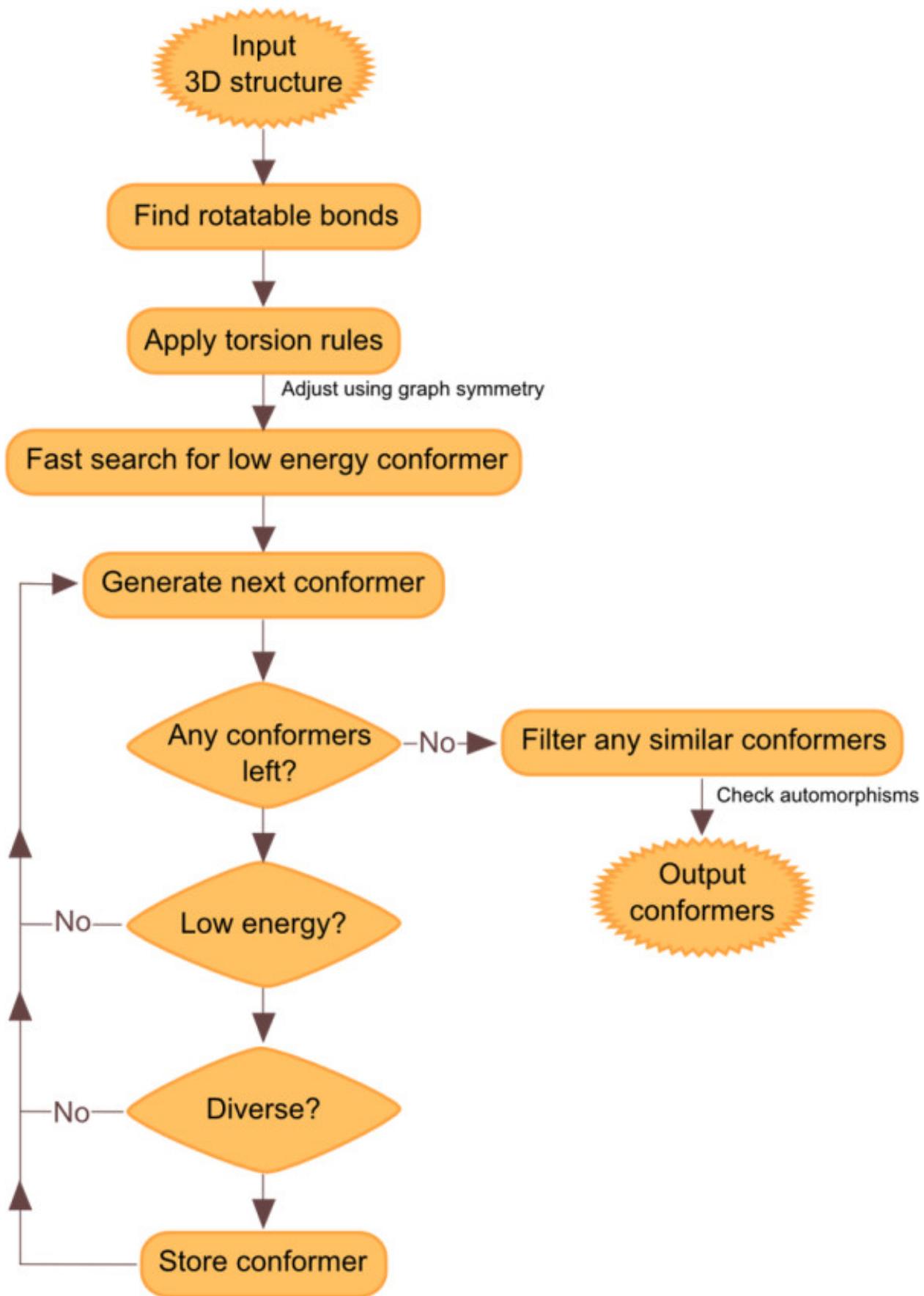
<sup>11</sup> MED-3DMC: A new tool to generate 3D conformation ensembles of small molecules with a Monte Carlo sampling of the conformational space

<sup>12</sup> Frog2: Efficient 3D conformation ensemble generator for small compounds

<sup>13</sup> MS-DOCK: Accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening

<sup>14</sup> Automated flexible ligand docking method and its application for database search

<sup>15</sup> Benchmarking Sets for Molecular Docking



Once the allowed torsion angles are assigned, they are corrected for topological (that is, graph) symmetry. The presence of such symmetry allows performance to be improved by eliminating redundant evaluations, thus reducing the number of conformations that will be tested. 2-fold symmetry is identified when a rotatable bond involves an  $sp^2$  hybridised carbon atom where the neighbouring two atoms affected by the rotation are both of the same symmetry class. When this occurs the allowed values of that torsion are halved by restricting them to those less than  $180^\circ$ . The same is done for the case of 3-fold symmetry at an  $sp^3$  hybridised carbon where the three neighbours are of the same symmetry class; in this case the torsion angles are restricted to those less than  $120^\circ$ . If graph symmetry is identified at both ends of a rotatable bond, the result is multiplicative; a 2-fold and a 3-fold symmetry combine to restrict allowed values of the torsion angles to  $360/6 = 60^\circ$ .

The next step is to obtain an estimate of the energy of the most stable conformer. Throughout Confab, energies are calculated using the MMFF94 forcefield<sup>16</sup>. The values of the bond stretching, angle bend, stretch bend and out-of-plane bending terms are constant for all conformers of the same molecule; only the torsion, Van der Waals and electrostatic terms were repeatedly evaluated. A low energy conformer is found using a simple greedy algorithm. Each torsion angle is optimised starting with the most central torsion and proceeding outwards. As this procedure is relatively fast (compared to the combinatorial problem of searching for the global optimum) it is repeated up to 16 times by testing the four most central torsions in different orders. The lowest energy conformer found is used as a reference point for applying an energy cutoff during the conformer search. If, during the actual conformer generation a lower energy conformer is found, this lower energy is used instead for the reference from that point on.

The main part of the algorithm is the systematic generation and assessment of all conformers described by the allowed torsion angles. Confab generates each of these in turn up to a user-specified cutoff (the default is  $10^6$ ) and determines its energy relative to the lowest energy conformer found so far. If this is within a user-specified energy cutoff (50 kcal/mol by default), it is assessed for diversity to the conformers already stored (see below). If it is found to be diverse, it is itself stored otherwise it is discarded. The algorithm then moves onto the next conformer.

Rather than iterate in a ‘depth-first’ manner over the torsions and their allowed angles, Confab uses a Linear Feedback Shift Register (LFSR) to iterate in a random order over all of the conformers. A LFSR allows the generation of all integers from 1 to N pseudorandomly without repetition and without any memory overhead (which is important for large values of N). By iterating randomly, Confab avoids biasing generated conformers towards a particular region of conformational space, for example towards the input conformation. It also helps increase diversity if the number of possible conformations is greater than the cutoff for the number tested.

Diversity is ensured by calculating the heavy-atom RMSD (after least-squares alignment) of the newly generated conformation to those previously stored. The alignment is carried out using the QCP algorithm of Theobald<sup>17</sup> (which we found to be about twice as fast as the popular Kabsch alignment method<sup>18</sup>). Despite this, when a molecule has many conformers and a large number of conformers have been stored, full pairwise RMSD calculations take an excessive amount of time. To minimise the number of RMSD evaluations required to discard a conformer, chosen conformers are stored in a tree structure that effectively clusters conformers on-the-fly by RMSD. Figure 2(a) shows a typical ‘diversity tree’ where each level of the tree is associated with a smaller RMSD diversity from  $3.0 \text{ \AA}$  down to the cutoff specified by the user ( $1.6 \text{ \AA}$  in the figure). Each node of the tree represents a stored conformation. Sibling nodes (that is, nodes at the same level that share the same parent node) differ by at least the RMSD diversity associated with that level. Note that sibling nodes are ordered and that the first child node of each parent is the same as the parent itself.

To illustrate the algorithm, let us imagine adding a new conformation H to the tree depicted in Figure 2(a). The algorithm starts at the top of the tree and determines which of the two branches (A or B) to take at the  $3.0 \text{ \AA}$  diversity level. To do so it checks whether H is within  $3.0 \text{ \AA}$  RMSD of A. If so, it follows the tree down to the next level, and checks to see whether it is within  $2.0 \text{ \AA}$  RMSD of A (note that it does not need to recalculate the RMSD to do this). If this is not true, then it checks for  $2.0 \text{ \AA}$  similarity to C. If so, it follows C down to the next level; otherwise it checks against D. If it is not similar to D, H is stored in the tree as the next sibling at that level of the tree (this is depicted in Figure 2(b)). When adding a new node for a conformation at a particular level, if the level is not at the bottom then child nodes containing that conformation are added at successively lower levels until the bottom level is reached.

<sup>16</sup> Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94

<sup>17</sup> Rapid calculation of RMSDs using a quaternion-based characteristic polynomial

<sup>18</sup> A solution for the best rotation to relate two sets of vectors

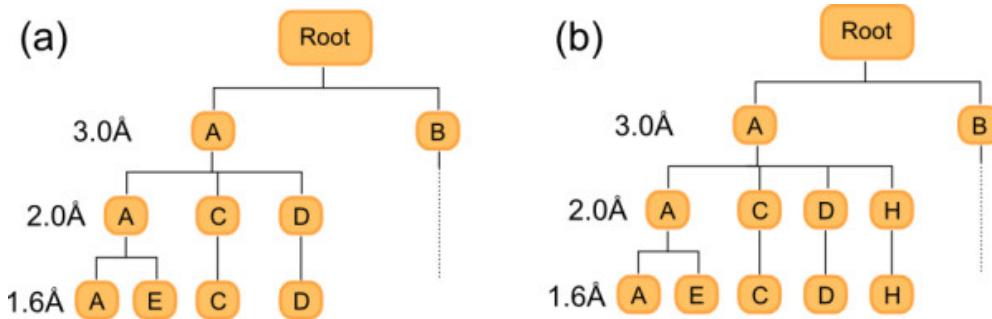


Figure 8.2: Figure 2. An example diversity tree used to filter conformations on-the-fly

**An example diversity tree used to filter conformations on-the-fly.** (a) A diversity tree containing five conformations (A to E) used to filter conformations with an RMSD of less than 1.6 Å to one of the stored conformations. (b) The same diversity tree after addition of conformer H, where H is within 3.0 Å of A but not within 2.0 Å of A, C or D.

Overall, there are two possibilities; either the algorithm reaches the bottom level and finds that the new conformation is within the RMSD cutoff of an existing conformer, in which case it is discarded, or else it is of sufficient diversity to be stored at some level of the tree.

This algorithm greatly reduces the number of RMSD evaluations during the conformer generation loop. However it does not eliminate all conformations that are similar to those already stored; conformations may be retained that differ by less than the RMSD cutoff if they end up in different branches. To prune the set of retained conformations, while still avoiding a computationally expensive pairwise RMSD calculation, all of the retained conformations are added one-by-one to a new tree in order of increasing energy. This time the algorithm used for adding conformations to the diversity tree is more robust: all sibling conformations are tested for similarity, even after finding one that is similar. The result is that the same conformation may be added at several different points in the tree. This makes the tree more effective at eliminating similar conformations at the expense of a greater number of RMSD calculations.

Calculation of an RMSD can be overestimated when a molecule's structure has automorphisms (a permutation of the atoms of a molecule that preserves the bond connections). For example, if you consider a para-substituted phenyl ring where two conformations differ by a rotation of 180° around the substituted carbons, it is clear that the calculated RMSD between the conformations should be 0. However, if the symmetry of the phenyl ring is not taken into account this will not be the case and the RMSD will be overestimated as the corresponding atoms of the two structures have moved. The symmetry-corrected RMSD is obtained by iterating over the automorphisms of the molecule and taking the minimum value of the resulting RMSDs. For performance reasons, the calculation of the RMSD is not symmetry-corrected during the main conformation generation loop. However it is used afterwards when building the final diversity tree, thereby eliminating any conformations that were retained in error.

### 8.3.2 Implementation

Confab is essentially a modified version of Open Babel<sup>19</sup>, a widely-used cheminformatics toolkit written in C++ and available under the open source GPL v2 licence<sup>20</sup>. In fact, some of the code written for Confab has been merged into the main Open Babel distribution (such as the original Kabsch alignment code) but due to an additional dependency (on tree.hh, see below) the core code has not been included in Open Babel v2.3.

The MMFF94 forcefield, the conformer generation framework and the automorphism detection are all provided by Open Babel. QCP alignment was implemented using Theobald's public domain code<sup>21</sup> in combination with the Eigen2 high performance linear algebra library<sup>22</sup>. The diversity analysis code relies on a tree data structure provided

<sup>19</sup> Open Babel, v2.3

<sup>20</sup> GNU General Public License, v2

<sup>21</sup> QCProt, v1.1

<sup>22</sup> Eigen, v2.0.15

by the Open Source tree.hh library<sup>23</sup>. The code used to implement the Linear Feedback Shift Register (LFSR) was adapted from its corresponding Wikipedia article<sup>24</sup>. Tap values for the register were taken from Alfke's Xilinx application note<sup>25</sup>.

The Confab distribution contains two command-line applications: *confab* and *calcrmsd*. The former implements the Confab algorithm to generate conformers given an input 3D structure, while the latter may be used to assess the performance of *confab* by comparing the generated conformers to a file containing crystal structures. Full details of these applications are available on the Confab website.

## 8.4 Coverage of Conformational Space

### 8.4.1 Dataset

To illustrate the performance of Confab, we used a dataset of 1000 small molecule crystal structures derived from that of Borodina et al.<sup>26</sup>. The original source is the PDB; thus this dataset represents bioactive conformations of molecules. The 3D structures of the 14504 ligands in the Borodina dataset were obtained using the PubChem Download Service (using the PubChem Substance IDs from Borodina et al.). Of these, 16 could not be handled by the MMFF94 forcefield, 5202 had no rotatable bonds (this fraction included a large number of trivial salts) and 2348 had more than 1 million conformers (according to Confab's torsion rules). 1000 structures were randomly chosen from the 6938 remaining. See Additional file

Additional file 1

**Crystal structures used to test conformational coverage.** This is a text file in SDF format containing biological conformations (as downloaded from PubChem) of 1000 molecules. This is a subset of the data used in the study by Borodina et al.

Click here for file

To avoid bias towards the crystal structures, the input conformations for Confab were generated by building the 3D structure using Open Babel. After the initial structure generation, the structures were optimised using the MMFF94 forcefield (200 steps steepest descent). Since Confab does not explore ring conformations, ring conformations were taken from the crystal structure for the initial structure generation. See Additional file

Additional file 2

**Generated 3D structures used to test conformational coverage.** This is a text file in SDF format containing 3D structures of the 1000 molecules in the dataset generated using Open Babel. These were used as the input to Confab.

Click here for file

### 8.4.2 Results

Figure 3(b) shows an overview of the dataset of 1000 structures in terms of the number of rotatable bonds in each molecule. Although the dataset contains molecules with up to 12 rotatable bonds, it is clear by comparison with the full dataset of Borodina et al. in Figure 3(a) that the reduced dataset is only a representative sample for molecules having up to 7 rotatable bonds. Beyond this, the restriction that the molecule must have fewer than 1 million conformers leads to the elimination of most of the molecules. For this reason, to avoid erroneous conclusions some of the following analyses (where stated) will not consider molecules having 8 or more rotatable bonds.

Confab was used to exhaustively generate all low energy conformers for each molecule in the dataset for diversity values ranging from 0.4 Å to 3.0 Å RMSD. The default setting of 50 kcal/mol was used as an energy cutoff. The default

---

<sup>23</sup> tree.hh, v2.65

<sup>24</sup> Linear feedback shift register

<sup>25</sup> Efficient Shift Registers, LFSR Counters, and Long Pseudo-Random Sequence Generators

<sup>26</sup> Assessment of Conformational Ensemble Sizes Necessary for Specific Resolutions of Coverage of Conformational Space

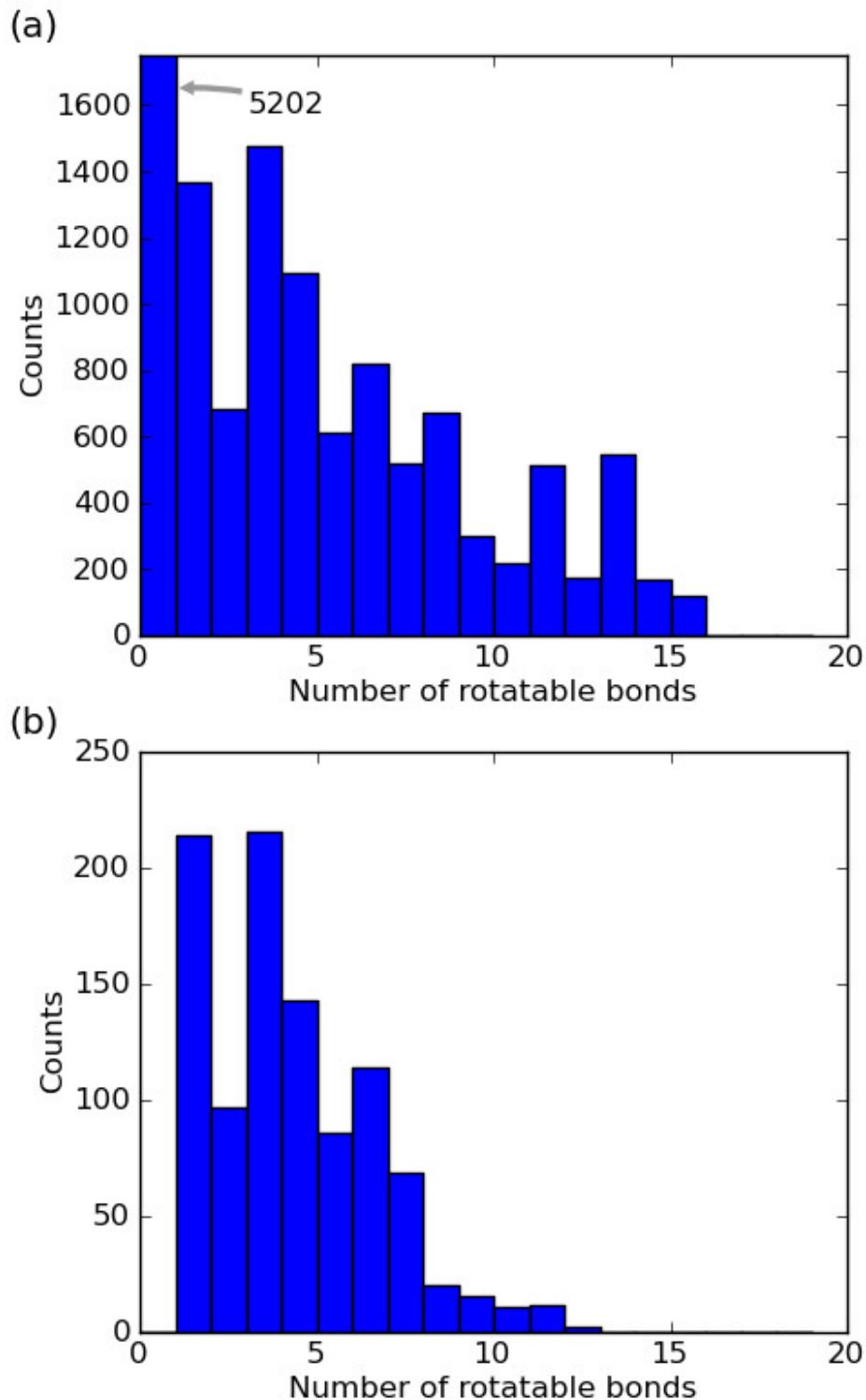


Figure 8.3: Figure 3. The distribution of molecules in terms of the number of rotatable bonds in (a) the dataset of Borodina et al., and (b) our dataset of 1000 molecules

**The distribution of molecules in terms of the number of rotatable bonds in (a) the dataset of Borodina et al., and (b) our dataset of 1000 molecules.**

value of 1 million conformers was used as the conformer cutoff; this ensured exhaustive coverage of conformational space (as defined by Confab's torsion rules) as structures with more conformers were not included in the dataset (see above). Figure 4 shows the mean time for conformer generation per molecule. This is largely independent of the diversity level for diversity levels greater than or equal to 1.0 Å. For values less than this, an increasing amount of time is spent performing the pairwise RMSD calculations against stored conformations.

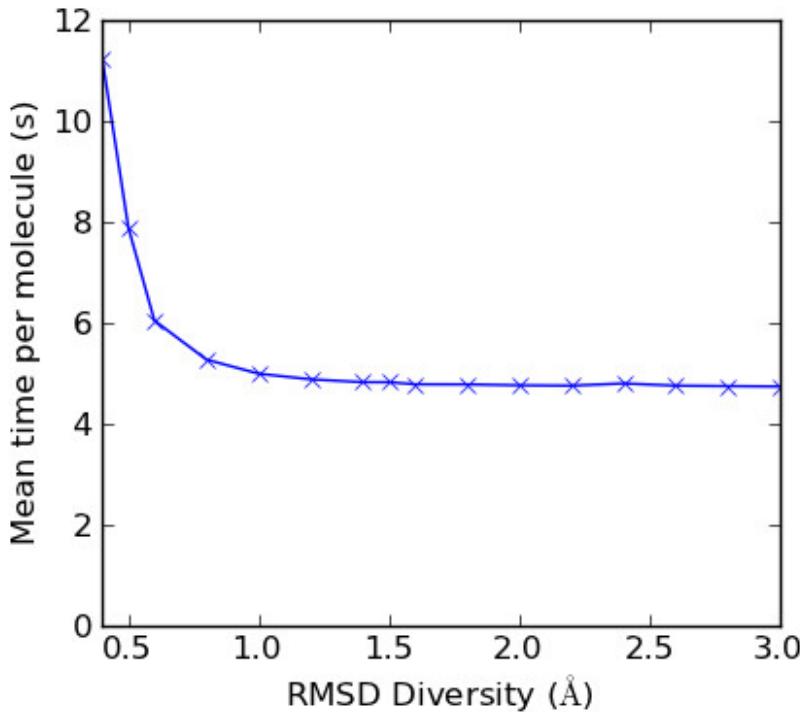


Figure 8.4: Figure 4. Effect of diversity level on speed of conformer generation

**Effect of diversity level on speed of conformer generation.** Times were measured on an Intel Xeon E5620 Processor (2.4GHz, 4C) with 32GB RAM.

Performance of conformer generators is typically measured by the percent recovery of crystal structures with respect to a particular RMSD cutoff (see for example Ref<sup>9</sup>). This is simply the percentage of molecules which have a generated conformer within a particular RMSD of the crystal structure. Commonly used values for this RMSD cutoff are 2.0, 1.5 and 1.0 Å.

Figure 5(a) shows the percent recovery at these cutoffs for different values of the RMSD diversity. At 2.0 Å RMSD diversity, 99% are within 2.0 Å RMSD of the crystal (83% within 1.5, 41% within 1.0); at 1.5 Å RMSD diversity, 99% are within 2.0 Å (97% within 1.5, 50% within 1.0); at 1.0 Å RMSD diversity, 99% are within 2.0 Å RMSD (98% within 1.5, 89% within 1.0). As expected, the percentage of crystal structures that are found decreases as the RMSD diversity increases. In particular, the curves fall off steeply once the RMSD diversity is greater than the required cutoff.

An interesting question to ask is what RMSD diversity is required to recover X% of crystal structures with respect to a particular RMSD cutoff? Figure 5(b) shows the answer to this where X is 90%, 95% or 98%. For example to find 95% of the crystal structures within a 2.0 Å cutoff an RMSD diversity of 2.4 Å (or smaller) is required, but to find the same percentage to within 1.5 Å an RMSD diversity of 1.6 Å is needed. However, even an RMSD diversity of 0.4 Å will not recover 98% of the structures to within 1.0 Å (it only recovers 96%), an indication of the inherent diversity of the generated conformers as discussed further below.

As pointed out by Borodina et al.<sup>26</sup>, if the conformational space is perfectly covered and lacks any ‘holes’ then the RMSD diversity is an upper bound of the minimum RMSD to the crystal structure. In other words, at an RMSD

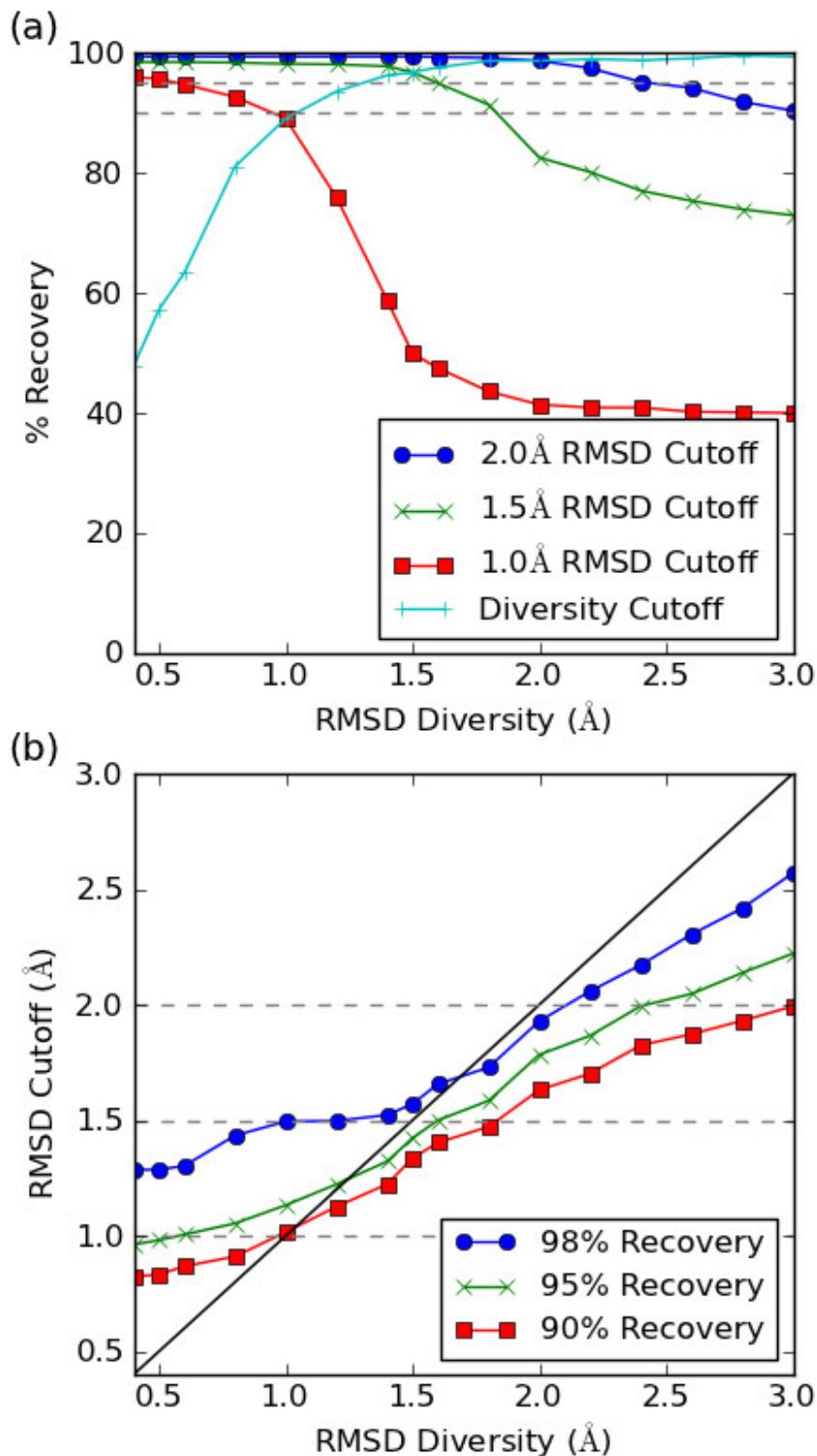


Figure 8.5: Figure 5. Performance measured as % recovery of crystal structures

**Performance measured as % recovery of crystal structures.** (a) Performance for different RMSD cutoffs. The diversity cutoff is where the value of the RMSD diversity is used as the RMSD cutoff. (b) The RMSD cutoff required to achieve a particular level of % recovery. The diagonal line indicates the maximum RMSD cutoff expected when there is complete coverage.

diversity of 1.5 Å for example, all crystal structures should be found to within 1.5 Å. The diagonal line in Figure 5(b) indicates the maximum RMSD cutoff expected if this ideal behaviour is observed. It is clear from the figure that at low RMSD diversity the actual performance is poorer than this.

There are two main problems that give rise to gaps in conformational coverage. The first is that the allowed torsion values may not encompass the specific torsion angle observed in the crystal structure. For this dataset, there are 7 molecules for which the crystal structure could not be found within 2.0 Å even at 0.4 Å RMSD diversity. These molecules (PubChem substance IDs of 584680, 823881, 825747, 826196, 828032, 830919 and 834618), of which two represent different conformations of the same molecule, all involve sugar moieties and it may be that the allowed torsion angles of the glycosidic bond are too conservative.

The second is that the granularity of the allowed torsion settings may not be sufficiently fine to allow solutions to be found to within a low RMSD cutoff. For example, a carbon-carbon single bond has 12 allowed torsion values from 0 to 360° in increments of 30°. If such a bond is centrally located in a large molecule, even if the crystal structure has similar torsion angles to one of these conformers the RMSD may differ significantly.

Based on this dataset, the inherent granularity of the Confab generated conformers is around 1.4 Å, as indicated by the “Diversity Cutoff” line in Figure 5(a) which falls off sharply as the RMSD diversity decreases below 1.4 Å. This line indicates the percent recovery at different levels of RMSD diversity when the RMSD cutoff used is the same as the diversity level. The sharp fall off below 1.4 Å is a deviation from the ideal behaviour described by Borodina et al.

Table 1 shows the median number of generated conformers tested for molecules with different numbers of rotatable bonds. Broadly speaking, about one third of the conformers pass the energy cutoff applied. Although the size of each individual subset is not very large, and the values for 6 rotatable bonds seem to be biased towards a larger number of conformers, some general points can still be made.

The number of diverse conformers is much reduced by a higher diversity level. For example, for those molecules with 7 rotatable bonds there are approximately 11000 low energy conformers of which about 13% are diverse at 0.5 Å RMSD, only 1.3% are diverse at 1.0 Å RMSD, and only 0.16% are diverse at 1.5 Å RMSD.

The values in Table 1 are in broad agreement with those reported by Smellie et al.<sup>27</sup> for a representative subset of their dataset (see table three therein). They make the point that the number of conformers required to cover conformational space is really surprisingly low. For a molecule with 7 rotatable bonds in our dataset, conformational space can be covered to within 1.0 Å with merely hundreds of conformations while just tens of conformations will achieve a coverage of 1.5 Å. Of course, these figures are expected to increase with each additional rotatable bond.

For completeness, Table 1 also reports median values for the minimum RMSD to the crystal structure. However, as a metric for coverage these values give a misleadingly positive picture compared to the percent recovery values discussed above.

#### 8.4.3 Comparison with Multiconf-DOCK

Multiconf-DOCK<sup>13</sup> is another open source conformer generator that uses a torsion driving approach to implement a systematic search to identify diverse low energy conformers. This software differs in that it uses the AMBER force field<sup>28,29</sup> (as implemented in DOCK5) instead of MMFF94. In addition, it implements performance improvements such as search tree pruning by partial energy estimation<sup>14</sup>. Like Confab, the software requires a 3D structure as input.

Multiconf-DOCK was used to generate conformations for the 1000 structures in the dataset using the same input as for Confab but converted to MOL2 using Open Babel v2.3.0. It should be noted that the specified Sybyl atom types in the input MOL2 file have an effect on the conformations generated by Multiconf-DOCK. The parameters used were taken from the example provided with the Multiconf-DOCK distribution, except that no restriction was placed on the number of generated conformations and the energy cutoff was set to 50 kcal/mol (as used for Confab). Three different RMSD diversity levels were investigated: 2.0 Å, 1.5 Å and 1.0 Å. For all three diversity levels, the mean time spent per molecule was 6.3 s (measured on the same machine used for Figure 4).

<sup>27</sup> Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage

<sup>28</sup> A new force field for molecular mechanical simulation of nucleic acids and proteins

<sup>29</sup> An all atom force field for simulations of proteins and nucleic acids

The performance in terms of percent recovery is as follows: at 2.0 Å RMSD diversity, 99% are within 2.0 Å RMSD of the crystal structure (89% within 1.5, 55% within 1.0); at 1.5 Å RMSD diversity, 99% are within 2.0 Å (97% within 1.5, 64% within 1.0); at 1.0 Å RMSD diversity, 99% are within 2.0 Å (98% within 1.5, 80% within 1.0). These values are broadly similar to those for Confab (see above). The most noticeable differences occur for the percentage of structures found to within 1.0 Å RMSD; assuming that both programs successfully remove conformations that are within the diversity cutoff, Multiconf-DOCK outperforms Confab at the 2.0 Å and 1.5 Å RMSD diversity levels but Confab performs better at 1.0 Å RMSD diversity.

Table 2 shows the median number of conformers generated by Multiconf-DOCK, along with the minimum RMSD to the crystal structure, broken down by the number of rotatable bonds. Compared to Confab the number of conformers generated is far fewer. It is difficult to say whether this represents a less comprehensive coverage of conformational space or whether this is due to the use of different forcefields. In terms of the minimum RMSD to the crystal structure, once again we see that Multiconf-DOCK performs better than Confab at the 2.0 Å and 1.5 Å RMSD diversity levels but Confab is better at 1.0 Å RMSD diversity.

## 8.5 Distance Distribution in Conformations of a Phenyl Sulfone

Many conformer generators are focused on reproducing bioactive conformations. However it is worth remembering that the generation of conformers may also be useful in other contexts. Here we use Confab to as an aid to interpret the NMR spectra for the phenyl sulfone shown in Figure 6. The peak for the methylene carbon of the ethyl ester was split unexpectedly (compared to an analogous sulfone where the phenyl group was replaced by tert-butyl), and our hypothesis was that this was due to the close approach of the methylene carbon to one of the sulfonyl oxygens in solution. Confab was used to investigate whether low energy conformations existed where the methylene group was in close proximity to a sulfonyl oxygen.

Confab was used to generate a set of conformations of the molecule with a diversity of 0.2 Å and no energy cutoff. The resulting 2014 conformations were optimised using a MMFF94 forcefield (200 steps steepest descent; implemented using Pybel<sup>30</sup>) and the final energy recorded. For each of the conformations the minimum distance between a sulfonyl oxygen and the methylene carbon was measured.

Figure 7 shows a plot of these distances versus the relative energies of the conformers with marginal histograms showing the distribution of values. The methylene carbon does not approach the sulfonyl group very closely. For low energy conformers, the distances are clustered around 4.0 Å and 5.4 Å with the former more frequent. Taking 5 kcal/mol as a cutoff, the distance can be as low as 3.7 Å but shorter distances (down to 3.0 Å) are only possible with an associated energy penalty. Figure 6 shows one of the low energy conformations (relative energy of 4.6 kcal/mol) which has a distance of 3.7 Å between the groups of interest.

## 8.6 Conclusion

The goal of this first release of Confab is to ensure complete coverage of all of the low energy conformers of a molecule. While every effort is made to maximise performance, accuracy has been the main goal. Approximations that reduce the search space on the basis of heuristics have been avoided for this reason.

Using the results from Confab 1.0 as a comparison, future work will investigate strategies to overcome the combinatorial explosion associated with large numbers of rotatable bonds<sup>31</sup> including the trade-off between speed and accuracy.

<sup>30</sup> Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit

<sup>31</sup> Systematic search in conformational analysis

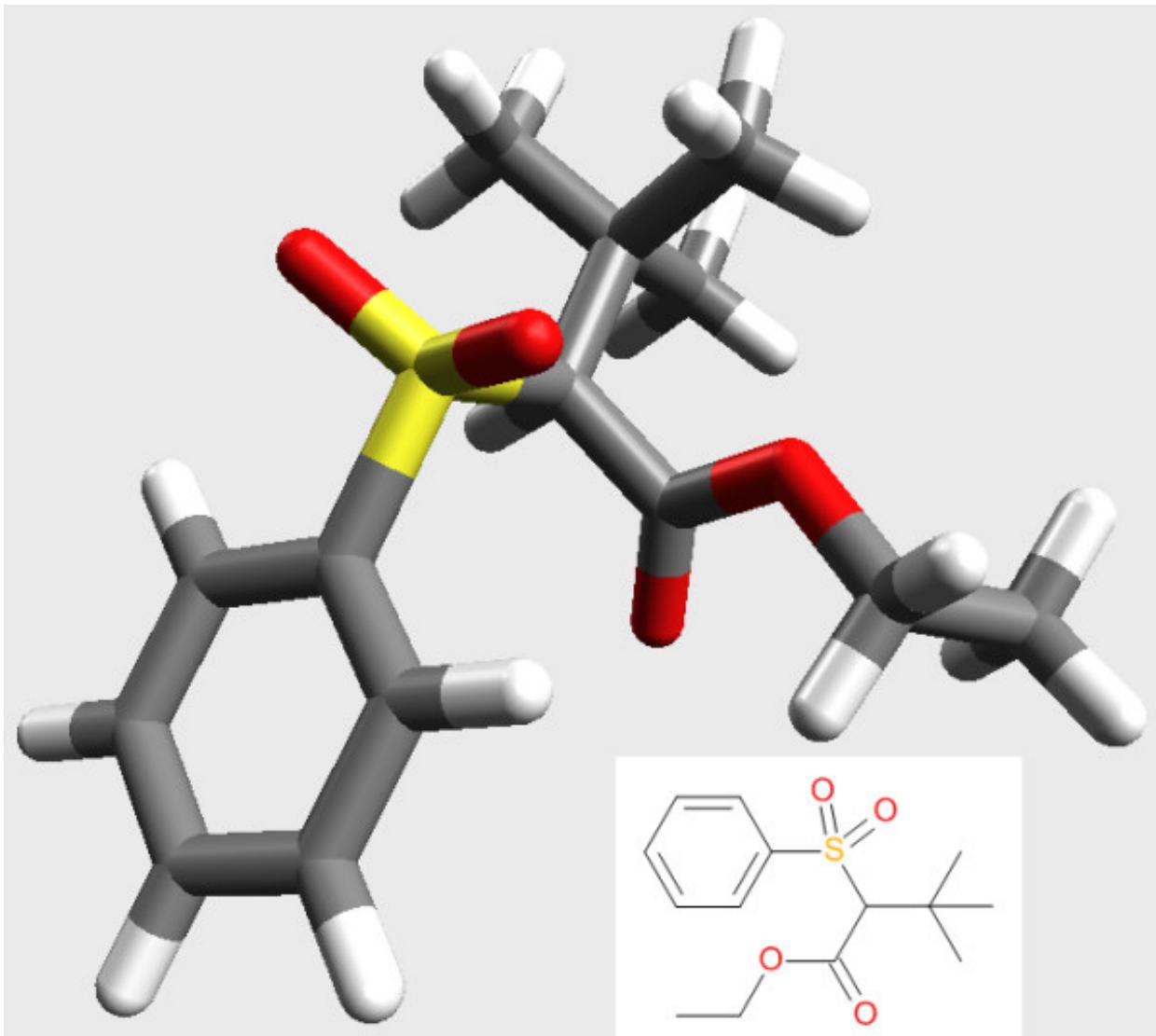


Figure 8.6: Figure 6. Structure of the phenyl sulfone studied  
**Structure of the phenyl sulfone studied.**

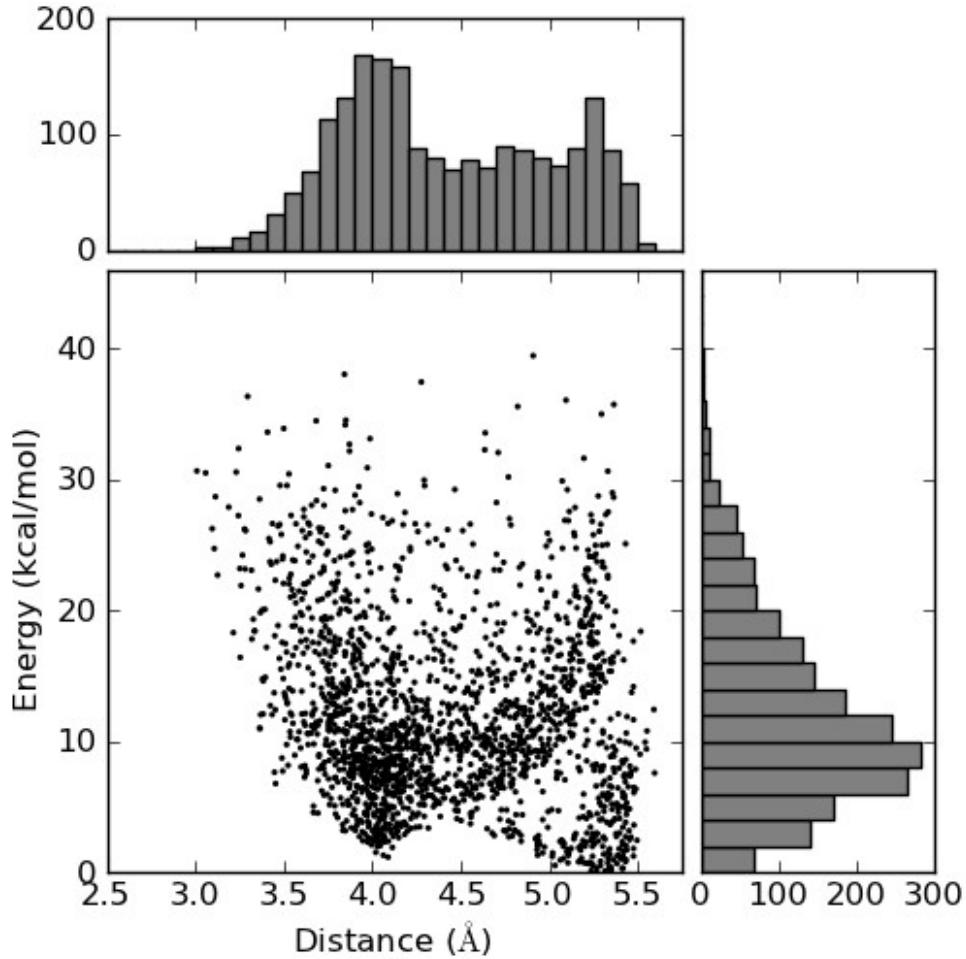


Figure 8.7: Figure 7. Scatterplot with marginal histograms of distance versus energy for the set of conformations of the phenyl sulfone in Figure 6

Scatterplot with marginal histograms of distance versus energy for the set of conformations of the phenyl sulfone in Figure 6.

## 8.7 Availability and Requirements

**Project name:** Confab

**Project home page:** <http://confab.googlecode.com>

**Operating system(s):** Cross-platform

**Programming language:** C++

**Other requirements (if compiling):** CMake 2.4+, Eigen2

**Licence:** GPL v2

**Any restrictions to use by non-academics:** None

## 8.8 Authors' contributions

NMOB devised and implemented Confab, and carried out the coverage analysis. GRH implemented the conformer generation framework in Open Babel and contributed to the forcefield code. TV implemented the automorphism code in Open Babel and contributed to the forcefield code. NMOB collaborated with CJF and ARM on the sulfone investigation. All authors read and approved the final manuscript.

## 8.9 Acknowledgements and Funding

NMOB is supported by a Health Research Board Career Development Fellowship, PD/2009/13. We thank several beta testers for their valuable feedback, and the anonymous reviewers for their constructive comments.

# PUBCHEM3D: DIVERSITY OF SHAPE

## 9.1 Abstract

### 9.1.1 Background

The shape diversity of 16.4 million biologically relevant molecules from the PubChem Compound database and their 1.46 billion diverse conformers was explored as a function of molecular volume.

### 9.1.2 Results

The diversity of shape space was investigated by determining the shape similarity threshold to achieve a maximum on the count of reference shapes per unit of conformer volume. The rate of growth in shape space, as represented by a decreasing shape similarity threshold, was found to be remarkably smooth as a function of volume. There was no apparent correlation between the count of conformers per unit volume and their diversity, meaning that a single reference shape can describe the shape space of many chemical structures. The ability of a volume to describe the shape space of lesser volumes was also examined. It was shown that a given volume was able to describe 40-70% of the shape diversity of lesser volumes, for the majority of the volume range considered in this study.

### 9.1.3 Conclusion

The relative growth of shape diversity as a function of volume and shape similarity is surprisingly uniform. Given the distribution of chemicals in PubChem versus what is theoretically synthetically possible, the results from this analysis should be considered a conservative estimate to the true diversity of shape space.

## 9.2 Background

Virtual screening of large chemical databases is now a routine practice in modern drug discovery <sup>12345678</sup>. One successful virtual screening approach is to compare the 3-D shape similarity of chemical structures using atom-centered

<sup>1</sup> Challenges and advances in computational docking: 2009 in review

<sup>2</sup> Development of anti-viral agents using molecular modeling and virtual screening techniques

<sup>3</sup> Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods

<sup>4</sup> Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results

<sup>5</sup> Virtual screening: an endless staircase?

<sup>6</sup> Molecular shape and medicinal chemistry: a perspective

<sup>7</sup> Comparison of topological, shape, and docking methods in virtual screening

<sup>8</sup> Comparison of shape-matching and docking as virtual screening tools

Gaussian functions<sup>9</sup><sup>10</sup><sup>11</sup>, e.g., as implemented in ROCS<sup>12</sup>. While this Gaussian-based approach to shape can perform hundreds or even thousands of chemical structure 3-D shape superposition computations per second per Central Processing Unit (CPU) core, even faster approaches with similar efficacy would be welcomed when searching a database of millions of chemical structures and (potentially) billions of conformers.

Attempts<sup>13</sup><sup>14</sup> have been made to use ROCS to identify reference shapes, which are then used to compute 3-D shape similarities at dramatically enhanced rates. One approach<sup>13</sup> created a binary “shape fingerprint” used much like traditional 2-D molecular connectivity fingerprints, where individual bits are “turned on” whenever the reference shape has sufficient shape similarity, as defined by the Shape Tanimoto (ST) in **Equation 1**, to the conformer being considered. Binary shape fingerprints, as an approach, were shown as a promising technique to encode the shape of a chemical structure conformer and achieve very fast 3-D similarity computation, but with the potential downside of not providing an actual 3-D conformer superposition and with no guarantee that the shape similarity values or result lists have any correlation with those provided by ROCS.

where  $V_{AA}$  and  $V_{BB}$  are the self-overlap volume of molecules A and B and  $V_{AB}$  is the overlap volume between them and the ST score ranges from 0 (for no shape similarity) to 1 (for identical shapes).

A second 3-D similarity approach using reference shapes<sup>14</sup> attempted to improve upon the first method by giving both a shape superposition and some assurance that the shape similarity ST is similar to that provided by ROCS. This was achieved by recognizing that two chemical structure conformers with similar 3-D shape align to a common reference shape in a similar fashion. By utilizing the  $3 \times 3$  rotational matrix and XYZ translational vector that align a 3-D chemical structure conformer to a common reference shape (retained after shape fingerprint generation), one could generate a superposition between conformers for each common reference shape. Given that two similar conformers may have multiple common reference shapes, one may “replay” all the alignments to common reference shapes and pick one that yields the best shape superposition. This approach achieved a 100× fold performance improvement by avoiding any shape similarity computation when shapes were too dissimilar (*i.e.*, there were no common reference shapes) and by avoiding any volume overlap maximization optimization computations. However, this methodology has its downsides. It only considered relatively small (<28 non-hydrogen atoms) and inflexible (<6 rotatable bonds) chemical structures and would not compute any shape similarity value when there was no common reference shape. Yet, in both studies<sup>13</sup><sup>14</sup>, it was shown that the use of reference shapes may provide promise to dramatically improve the throughput of shape-based alignment methodologies.

The first work described above<sup>13</sup> considered data sets of “drug-like” molecules with 12–32 non-hydrogen atoms and conformer counts between 50,000 and 500,000 to examine the growth of shape space as a function of ST value. This growth was linear when considering the logarithm of the count of reference shape and chemical structures, whether using a single conformer or multiple conformers per structure. The second work<sup>14</sup> also considered reference shapes of “drug-like” molecules of similar size, using a much larger dataset of one million chemical structures and fifteen million conformers, but only at a single ST value, as opposed to a range of ST values. Still, both studies gave valuable insight into how shape space grows with “drug-like” molecules.

In this work, we seek to expand upon these earlier two efforts by exploring in more depth the rate of growth of shape space as a function of reference shape count, conformer volume, and ST value with a much larger data set of 16.4 million biologically relevant small molecules and their 1.46 billion diverse conformers. By improving upon the understanding of the relative growth of shape space of biologically relevant molecules, new or improved “shape fingerprint”-based methodologies might be developed.

<sup>9</sup> A gaussian description of molecular shape

<sup>10</sup> A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape

<sup>11</sup> Gaussian shape methods

<sup>12</sup> ShapeTK-C++, Version 1.8.0, OpenEye Scientific Software, Inc.: Santa Fe, NM

<sup>13</sup> Small molecule shape-fingerprints

<sup>14</sup> Fast 3D shape screening of large chemical databases through alignment-recycling

## 9.3 Results and Discussion

### 9.3.1 1. Conformer generation

Conformers were generated for chemical structures in the PubChem Compound database<sup>15</sup><sup>16</sup><sup>17</sup> [#B18]\_\*Materials and Methods\* section. This resulted in 16,482,382 3-D conformer ensemble models (as of February 2008) and 1,465,813,269 diverse conformers (an average of 89 conformers per compound). The distribution of the non-hydrogen atom count, rotatable bond count, sampling RMSD, and conformer volumes (rounded to nearest integers) for these are shown in Figure 1. The average count and standard deviation of non-hydrogen atoms was 24.5 +/- 6.8 with a mode of 26 (with 1,033,645 compounds). The average count and standard deviation of rotatable bonds was 5.5 +/- 2.6 with a mode of 6 (with 2,432,059 compounds). The average and standard deviation of the sampling RMSD for the conformer ensembles was 0.82 +/- 0.20 Å with a mode of 0.8 Å (for 6,939,072 conformer ensembles). The average and standard deviation of the conformer volume was 297 +/- 64 Å<sup>3</sup>. The most common volume among the conformers was 307 Å<sup>3</sup> (for 10,920,699 conformers) and 99% of the conformers have a volume between 130 and 487 Å<sup>3</sup>. In further analyses, we focused on the conformers whose volumes were between 75 and 575 Å<sup>3</sup>, corresponding to 99.99% of all conformers.

### 9.3.2 2. Generation of reference shapes per volume

The shape diversity of a particular conformer volume may be ascertained by clustering conformers of that volume with a certain shape diversity threshold ( $\text{thresh}^{\text{ST}}$ ), which controls the “minimum” distance between any two clusters, and then by counting the number of reference shapes, each of which represents a cluster centroid and all conformers within  $\text{thresh}^{\text{ST}}$  to the reference shape. [Note that the  $\text{thresh}^{\text{ST}}$  is the “maximum” ST value between clusters since the ST score is a *similarity* measure, not a *dissimilarity* measure.] If the clustering is performed using the same  $\text{thresh}^{\text{ST}}$  value for a volume range, the shape diversity as a function of each molecular volume size may be evaluated by the growth of the number of reference shapes. However, when a constant  $\text{thresh}^{\text{ST}}$  value is used across a range of volumes, each increase in the molecular volume may result in a very rapid growth of the shape space, and hence, the number of reference shapes per volume. This is not completely desirable as the computational cost of clustering effectively increases as the square (or worse) of the total count of reference shapes (especially when this count is large), when considering  $N$  reference shapes must be compared against  $K$  conformers and  $N < K$ , compelling one to keep the count of reference shapes to a manageable size for tractability purposes.

To avoid excessive computational expense, we took an alternative approach (as described in Figure 2), in which the clustering for a given volume was performed with a dynamic  $\text{thresh}^{\text{ST}}$  value such that the resulting reference shape count became less than or equal to a certain number (chosen to be 200). In this manner, the number of reference shapes per volume was kept effectively constant (as an increase of ST by 0.01 would result in reference shape count above 200), while the growth of shape space as a function of volume is manifest by a decrease in  $\text{thresh}^{\text{ST}}$ . The detailed procedure for clustering is explained in the *Material and Methods* section and the PubChem Compound ID of the resulting reference shapes can be found on the PubChem FTP site [ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound\\_3D/ReferenceShapes/](ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound_3D/ReferenceShapes/).

Figure 3 shows the  $\text{thresh}^{\text{ST}}$  value and the reference shape counts as a function of the conformer volume. The  $\text{thresh}^{\text{ST}}$  score decreases gradually and uniformly in the 75-575 Å<sup>3</sup> range from 0.92 (for  $V = 75 \text{ Å}^3$ ) to 0.47 (for  $V = 558 \text{ Å}^3$ ). In fact, this decrease is so smooth that one can predict the ST value in the volume range 75-575 Å<sup>3</sup> using only the conformer volume (Equation 2) with an  $R^2$  value of 0.997.

where  $V$  is the conformer volume and  $\text{thresh}^{\text{ST}}$  is the shape Tanimoto for the given volume to achieve 200 or fewer reference conformers. The slope of the  $\text{thresh}^{\text{ST}}$  curve shows that the increase in the cluster distances becomes slower as the conformer volume increases; however, this reduction may be an artifact of the input. The reason for this is relatively simple. This study only considered chemical structures found in PubChem and was restricted to 50 or

<sup>15</sup> PubChem: integrated platform of small molecules and biological activities

<sup>16</sup> Database resources of the National Center for Biotechnology Information

<sup>17</sup> PubChem: a public information system for analyzing bioactivities of small molecules

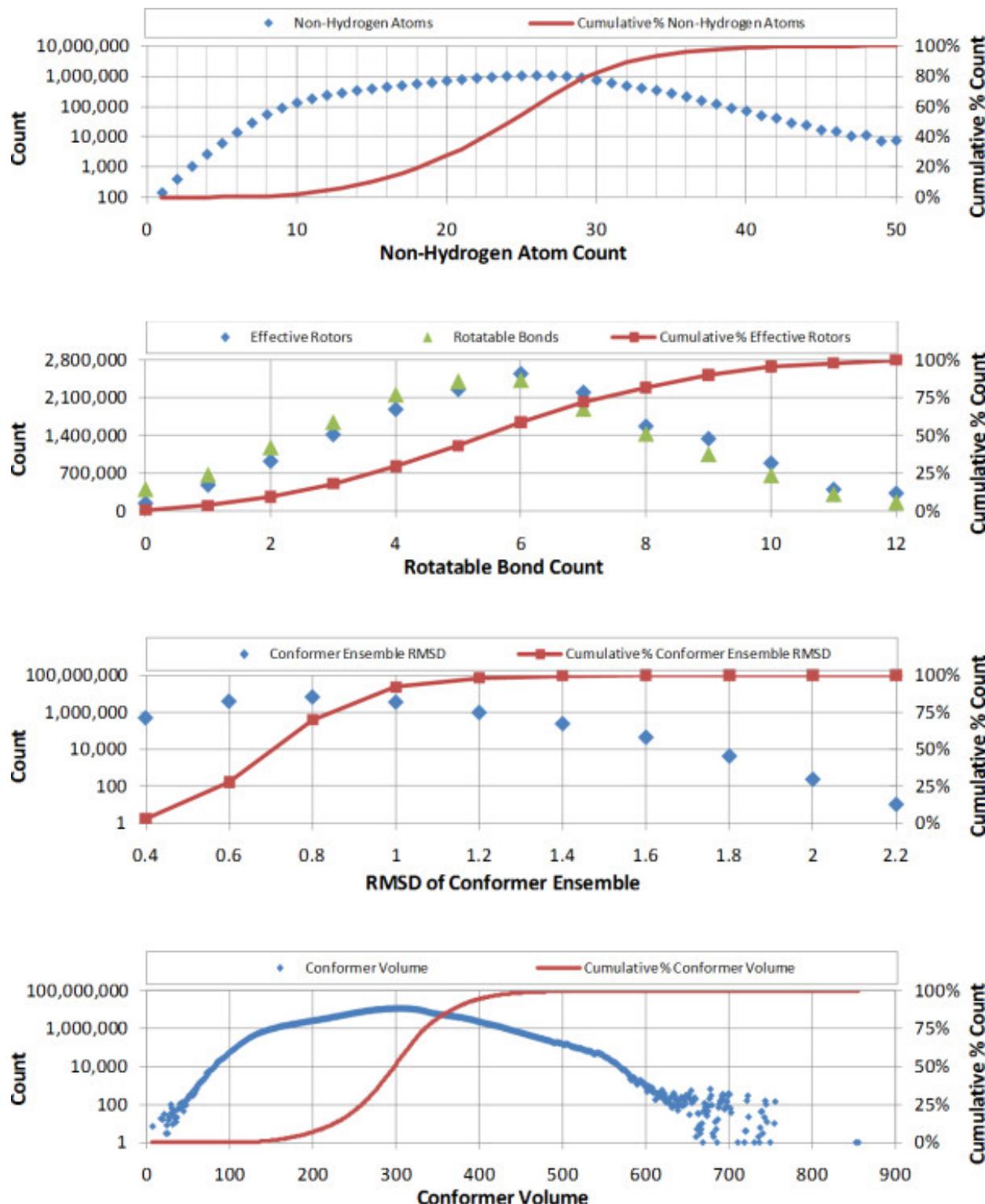


Figure 9.1: Figure 1. The distribution of non-hydrogen atom count, rotatable bond count, conformer ensemble sampling RMSD, and conformer volumes (rounded to the nearest integer) of 1,465,813,269 conformers generated from 16,482,382 molecules in the PubChem Compound database

**The distribution of non-hydrogen atom count, rotatable bond count, conformer ensemble sampling RMSD, and conformer volumes (rounded to the nearest integer) of 1,465,813,269 conformers generated from 16,482,382 molecules in the PubChem Compound database.**

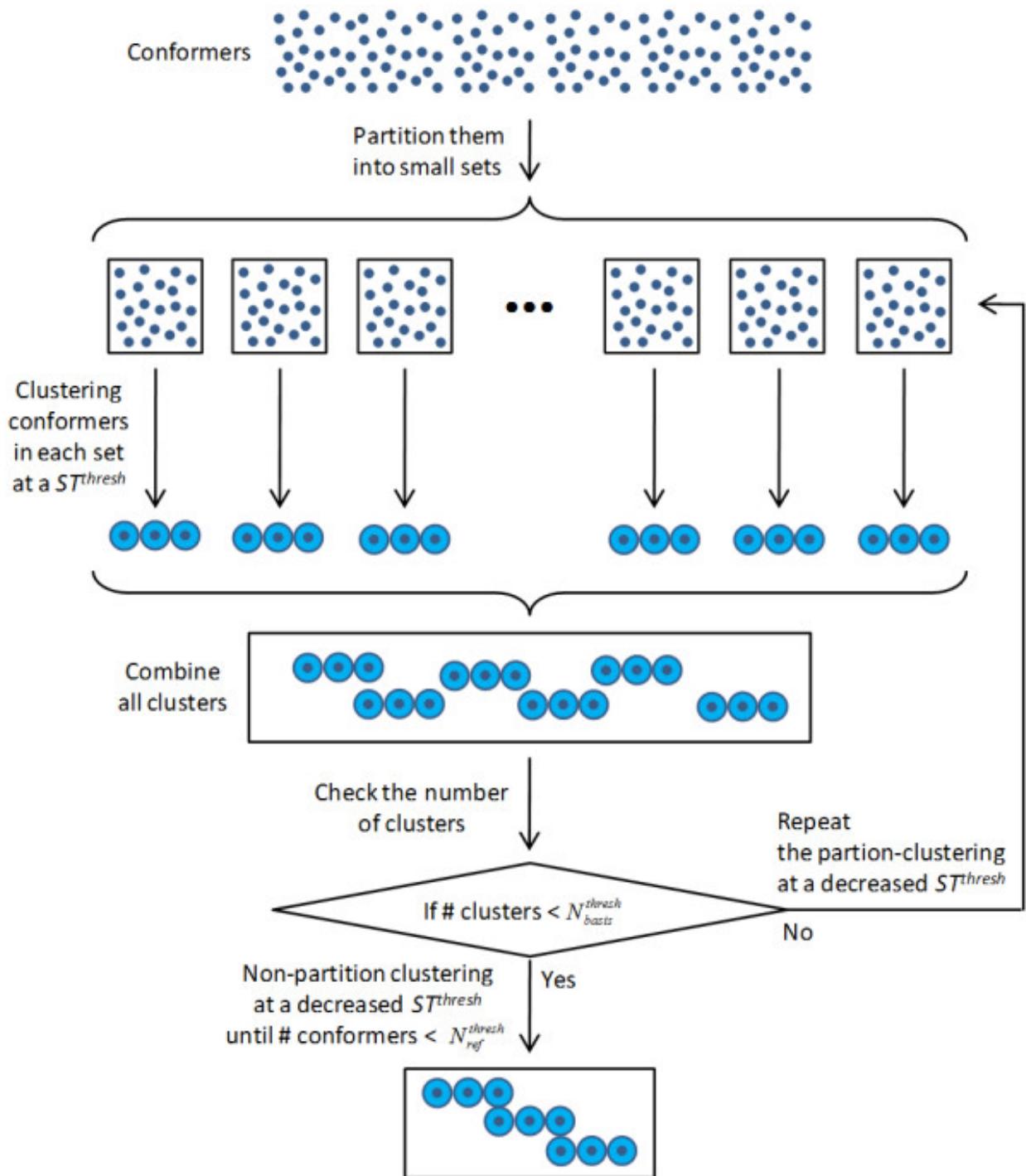


Figure 9.2: Figure 2. Partition-clustering scheme used for generating the reference shapes for a given volume  
**Partition-clustering scheme used for generating the reference shapes for a given volume.**

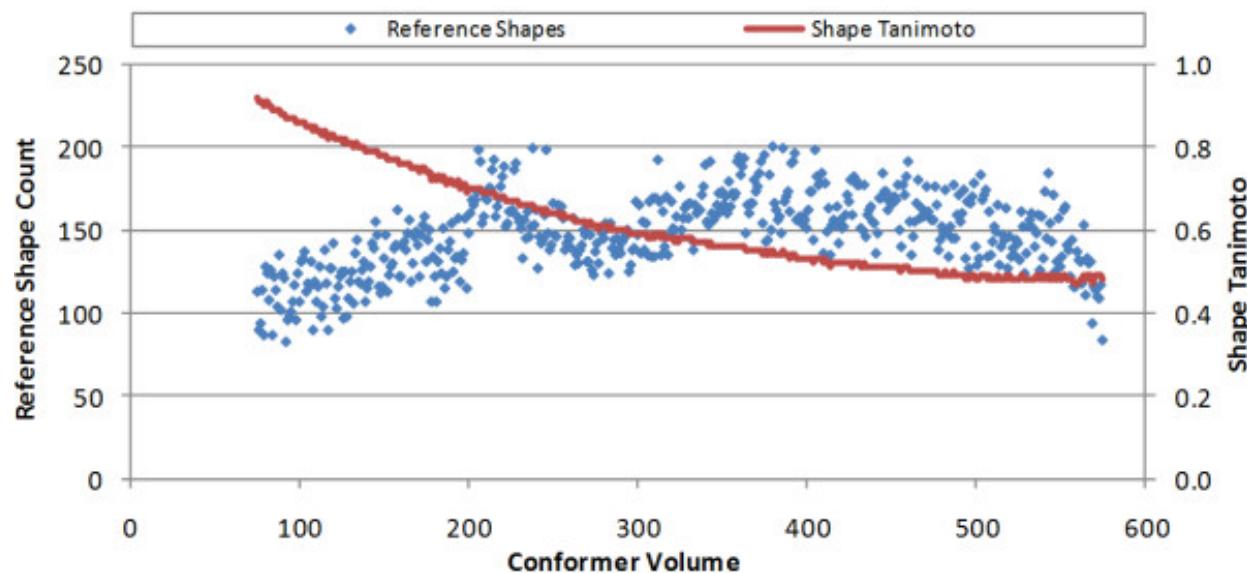


Figure 9.3: Figure 3. The Shape Tanimoto value used as a shape diversity threshold ( ${}^{\text{thresh}}\text{ST}$ ) and the resulting reference shape counts as a function of volume

**:sup:`thresh`STThe Shape Tanimoto value used as a shape diversity threshold (.**

less non-hydrogen atoms. Furthermore, the distribution of this non-hydrogen atom count had a maximum of 26. Conceivably,  ${}^{\text{thresh}}\text{ST}$  may decrease at a more rapid rate if the count of chemical structures in PubChem continued to increase as a function of non-hydrogen atom count across the entire range of non-hydrogen atom count, rather than hitting a maximum of 26. The net effect of this input artifact is that the  ${}^{\text{thresh}}\text{ST}$  curve in Figure 3 may be more linear than actually shown. We expect the entire curve as shown may shift and appear more linear as more theoretically possible and diverse chemical structures are considered; however, we believe the trends detailed in this work should still hold true, unless noted otherwise. Irrespective of the explanation provided, one should consider the curve shown in Figure 3 a conservative estimate of the absolute growth of shape space.

The reference shape count per volume was found to range from 83 (for  $V = 92 \text{ \AA}^3$ ) to the maximum allowed of 200 (for  $V = 380 \text{ \AA}^3$ ), and its average was 147.9. Interestingly, the  ${}^{\text{thresh}}\text{ST}$  curve does not reflect the maximum found in Figure 1 for conformer volume. In fact, the decrease in  ${}^{\text{thresh}}\text{ST}$  as a function of volume is very smooth, suggesting that the actual conformer count per volume, as shown in Figure 1, has little bearing on shape diversity, as shown in Figure 3. Or, put another way, the shape space of known chemicals is not near as diverse as chemical space, with a relatively small amount of reference shapes able to represent a large number of chemical structure conformers.

Another interesting observation is that a small change in  ${}^{\text{thresh}}\text{ST}$  has a large effect on reference count, as reflected in the somewhat periodic growth in shape references until the maximum value of 200 reference shapes is reached, cutting the reference shape count nearly in half. This can be roughly seen in the volume range  $75\text{--}210 \text{ \AA}^3$  and then again between  $275\text{--}375 \text{ \AA}^3$ . This reflects the use of 0.01 decrements in  ${}^{\text{thresh}}\text{ST}$  but also reflects anecdotal evidence seen when exploring the reference shapes, where each change in  ${}^{\text{thresh}}\text{ST}$  by 0.01 appeared to change the reference count by about a factor of two, much as observed by Haigh, *et al.*<sup>13</sup> This is only roughly seen in the reference shape counts as two things are changing, the volume and the  ${}^{\text{thresh}}\text{ST}$  value, and volume change involves a potentially variable change in shape space.

### 9.3.3 3. Generation of unique shapes for each volume

Reference shapes generated for a given volume are guaranteed to not be closer than the corresponding  ${}^{\text{thresh}}\text{ST}$  value so that the ST similarity between any two reference shapes for that volume cannot be greater than  ${}^{\text{thresh}}\text{ST}$ . However, it is still possible that two reference shapes of different volumes may be closer than  ${}^{\text{thresh}}\text{ST}$ , implying that some portion of the shape space covered by reference shapes for  $V = V_1$  can also be shared by reference shapes for  $V \neq V_1$ . For

this reason, we introduced the concept of the “unique shapes” for a given volume, defined as a non-overlapping set of conformers that cover the shape space spanned by the conformers whose volume is *smaller than or equal to* that volume (that is,  $V < V_1$ ). As illustrated in Figure 4, the unique shapes were classified into three groups according to the shape space they cover: (1) the “large unique shapes”, which cover the shape space spanned only by the conformers of  $V = V_1$ , (2) the “small unique shapes”, which cover the shape space spanned only by the conformers of  $V < V_1$ , and (3) the “shared unique shapes”, which cover the shape space spanned by the conformers of  $V = V_1$  *and* those of  $V < V_1$ . When the conformer volume increases from  $V < V_1$  to  $V = V_1$ , the “large unique shapes” for  $V = V_1$  explain newly added shape space whereas the “small unique shapes” for  $V = V_1$  represent the shape space not present for that volume. The unchanged portion of the shape space is explained by the “shared unique shapes” for  $V = V_1$ . Figure 5 schematically illustrates the shape space expansion upon a successive increase in the conformer volume. Note that smaller <sup>thresh</sup> STvalues were used for clustering as the volume increases (as represented by larger circles) to maintain the number of unique shapes to a manageable size and to reflect the <sup>thresh</sup> STvalue used in Figure 3 for  $V_1$ .

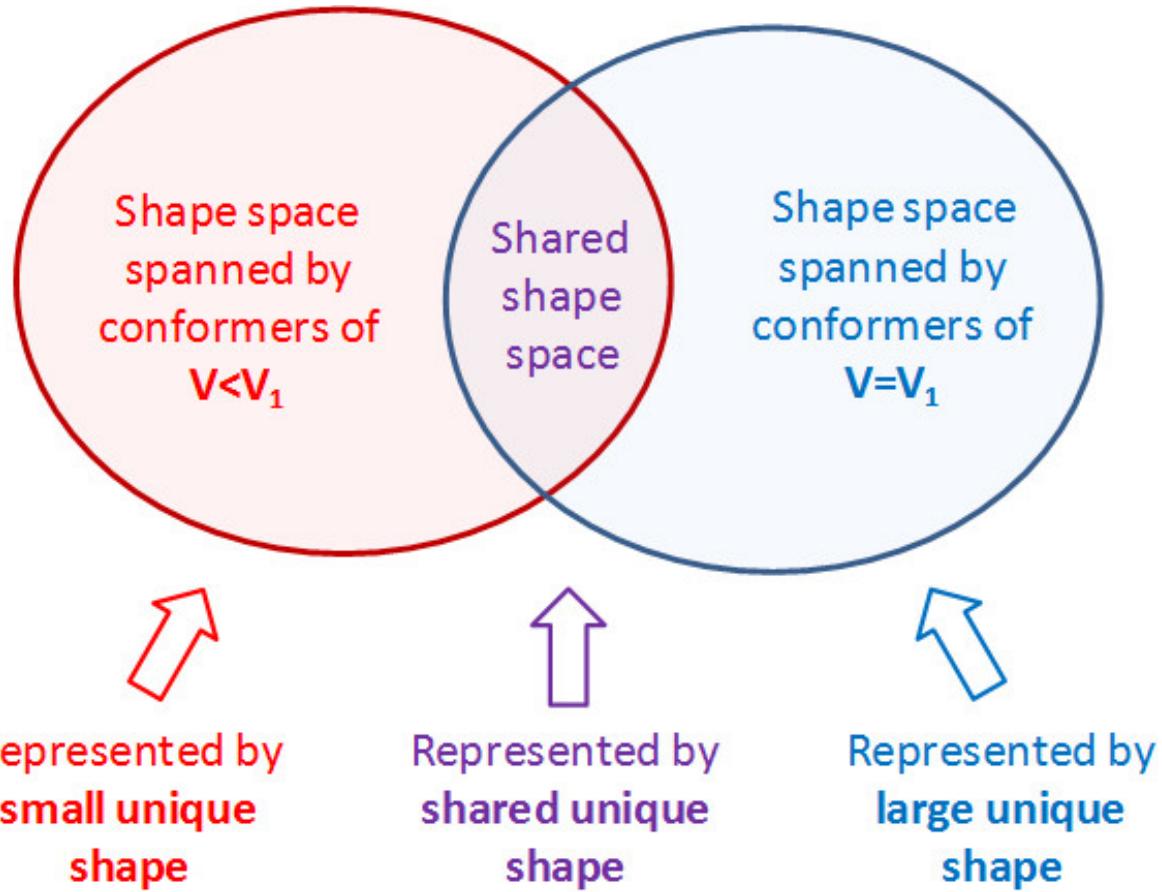


Figure 9.4: Figure 4. The concept of unique shapes for  $V = V_1$ , which cover the shape space spanned by the conformers whose volumes are less than or equal to  $V_1$

:sub:`1`<sup>1</sup>, which cover the shape space spanned by the conformers whose volumes are less than or equal to  $V$ :sub:`1`<sup>1</sup> The concept of unique shapes for  $V = V$ .

The “unique shapes” for each volume were computed using two different clustering strategies, the “small-then-large” approach and the “large-then-small” approach, as depicted in Figure 6, and detailed procedures are described in the *Materials and Methods* section. In the “small-then-large” approach [Figure 6(a)], the shape space of the conformers of  $V < V_1$  was first explored at the <sup>thresh</sup> STvalue for  $V_1$  to look for newly added shape space when the conformer volume increases to  $V_1$ . That is, the small and shared unique shapes for  $V = V_1$ , which cover the shape space spanned by conformers of  $V < V_1$ , were first generated by clustering all reference and basis shapes for  $V < V_1$ , and then the identified unique shapes were re-clustered with the reference and basis shapes for  $V = V_1$  to find the large unique

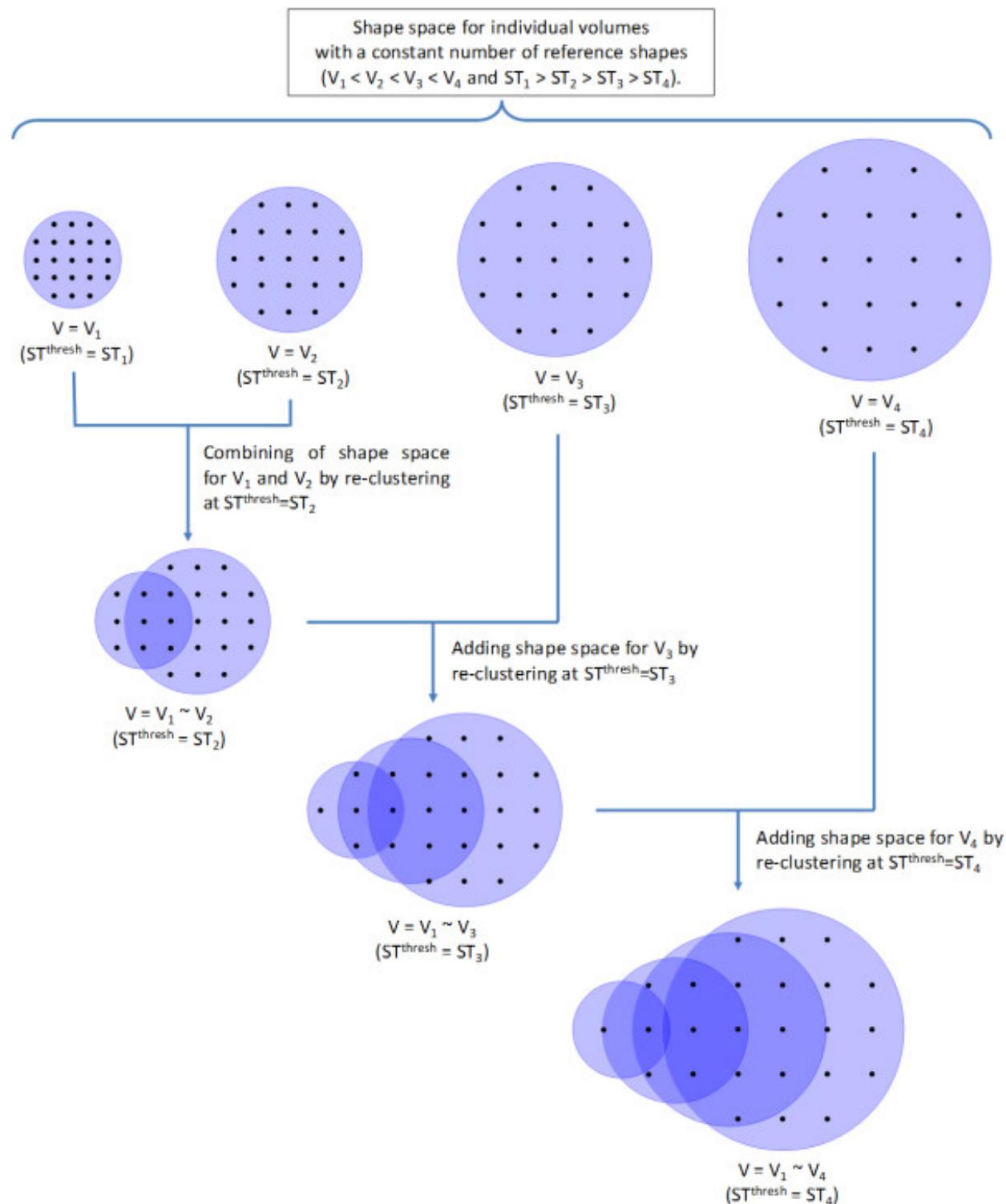


Figure 9.5: Figure 5. Schematic illustration of the shape space expansion upon a conformer volume increase  
**Schematic illustration of the shape space expansion upon a conformer volume increase.** Blue circles represent the shape space spanned by conformers of a particular volume ( $V$ ), and black dots represent reference shapes (for the individual shape spaces) or unique shapes (for combined shape spaces).  $ST^{\text{thresh}}$  indicates a Shape-Tanimoto (ST) value used as a shape diversity threshold.

shapes. On the contrary, in the “large-then-small” approach [Figure 6(b)], the large and shared unique shapes for  $V = V_1$  were determined first, by using the previously determined reference shapes for  $V = V_1$ , and then the reference and basis shapes for  $V < V_1$  were used to re-cluster to identify the small unique shapes.

The two methods resulted in two different sets of the unique shapes for each volume. The unique shape counts for both sets and the ratio between them are plotted in Figure 7(a), as a function of the conformer volume. Because both methods deal with the identical shape space, they are expected to give a number of unique shapes similar to each other; however, since reference shapes were selected randomly without any attempt to optimally minimize or maximize their count, these counts cannot be expected to be the same. As shown in Figure 7, the unique shape counts for the two sets tended to differ by 0-10%, although their ratio varied from 0.7 to 1.3 (especially for  $V > 500$ , where the conformer populations were not as numerous). This tendency may be explained by the fact that lesser volumes consider reference and basis shapes that may be considerably closer together due to larger  $^{thresh}$  STvalues. This suggests that using the larger volume reference shapes first resulted in a more efficient shape space description (*i.e.*, fewer reference shapes), when considering the union of the collective shape space for the volume range. Nonetheless, Figure 7 shows, as expected, that the total number of unique shapes gradually increases as a function of the conformer volume and its  $^{thresh}$  STvalue, indicating an overall expansion of shape space across the volume range irrespective of the change in ST value used (*i.e.*, shape space is growing faster than the decrease in ST value as a function of volume to achieve a maximum of 200 reference shapes).

Figure 8 displays the number of large unique shapes, small unique shapes, and shared unique shapes for each volume, while Figure 9 shows their proportions of the total unique shapes, which were estimated using the following equations:

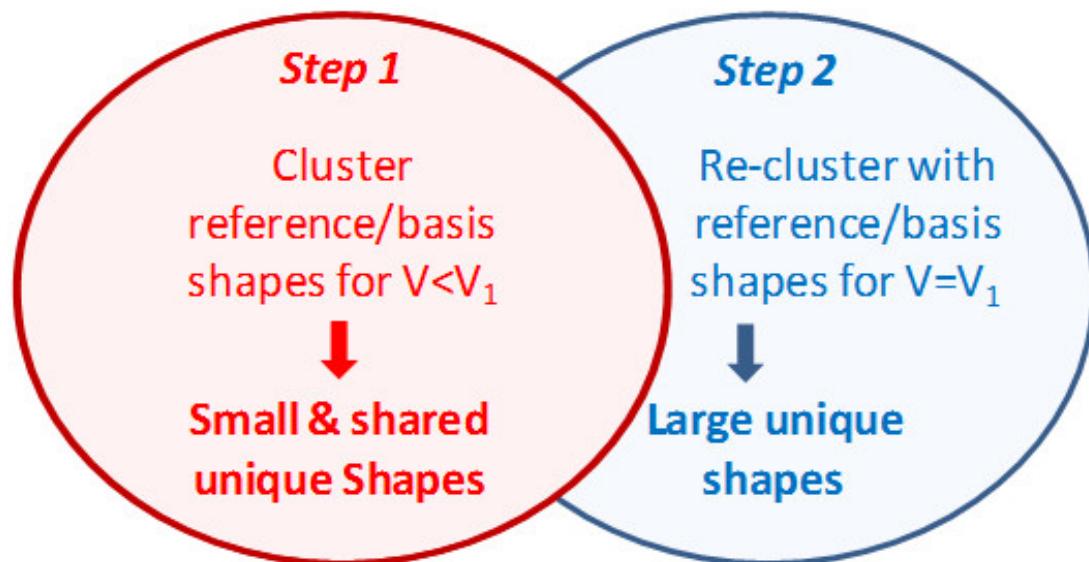
Note that the value of  $^{thresh}$  STaffects the counts of reference, basis, and unique shapes, because it determines the distance between clusters. However, the percentages of these counts plotted in Figure 9 are essentially equivalent to the fractions of the shape space that the individual counts represent, and hence, they may be considered to be independent of the  $^{thresh}$  STvalue.

There are a number of interesting observations one can make from these graphs. In Figure 7 and Figure 8 there is a banded behavior, indicated previously in Figure 3, which looks like a series of lines spaced further apart as the volume increases. This is due to the steady growth in shape space as volume increases and the use of 0.01 decrements of  $^{thresh}$  ST. Whenever the  $^{thresh}$  STdecreases by 0.01, a corresponding significant decrease in counts occurs. When the  $^{thresh}$  STvalue changes less, or does not change at all, the lines appear to be wider apart, reflecting just the growth in shape space due to volume.

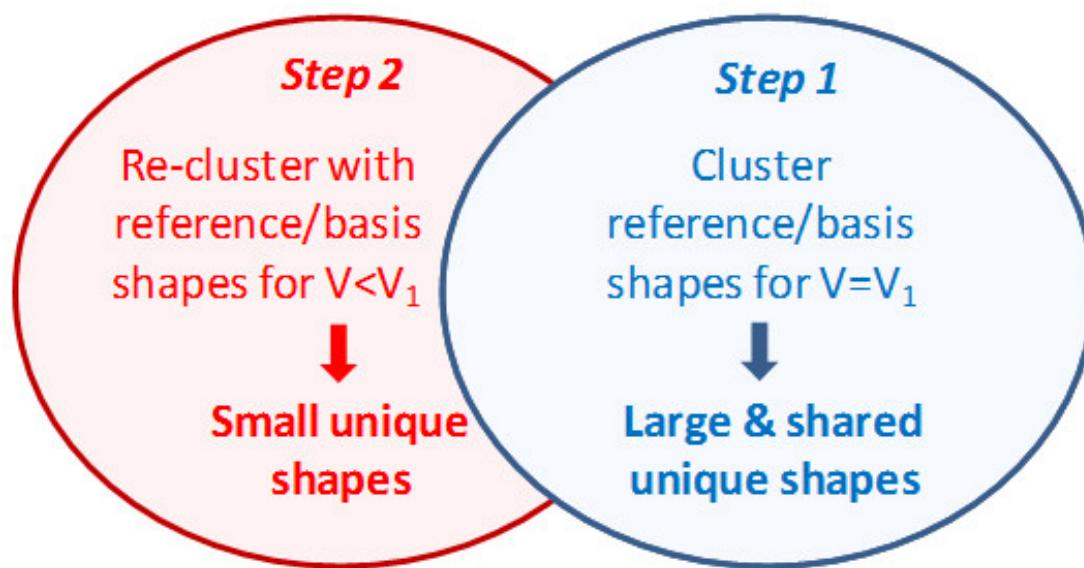
Another interesting observation in Figure 8(a), one can see that the absolute count of large unique shapes stays relatively constant in the volume range, with an average count and standard deviation of  $22.2 \pm 7.8$  and a mode of 24. There is a shallow maximum at volume  $145 \text{ \AA}^3$  followed by a relatively slow overall decline over the rest of the volume range. This decline appears most evident when the volume is beyond volume  $305 \text{ \AA}^3$ , perhaps due to the truncation of shape space considered as represented by the rapid reduction in conformer count at larger volumes and the fact that a maximum of non-hydrogen atom count occurs at 26.

Similar to the large unique shapes in Figure 8(a), the large and shared unique shapes in Figure 8(b) show a similar banded behaviour across most of the volume range, with a reference count mean and standard deviation of  $144.4 \pm 23.7$  and a mode of 140. There is a barely evident maximum volume at volume  $228 \text{ \AA}^3$  and a slightly noticeable dip at volume  $261 \text{ \AA}^3$ , prior to resuming the similar narrow band of large and shared unique shapes. This may suggest that the growth of large and shared shape space is relatively constant as a function of PubChem contents.

The small and shared unique shapes completely dominate in Figure 8(a), being nearly the same as the total count of unique shapes across the entire volume; however, the small unique shapes in Figure 8(b) show a very shallow minimum at about volume  $200 \text{ \AA}^3$  prior to significantly increasing as a function of volume. This may suggest that the overall size of PubChem shape space slows (as a function of the rate of changing ST) after a point, with large unique shapes contributing less and less to the overall shape diversity across the full volume range as the total shape space that can be represented by larger shapes diminishes. One can see this to some extent in Figure 9, where the percentage of shared shape space is “ $\Lambda$ “-shaped, reaching a maximum of 73% at volume  $217 \text{ \AA}^3$  and then steadily diminishes as a function of volume as the percentage of shape space of smaller shapes dominates. Again, it is reasonable to suggest that this observation is an artifact of the PubChem contents and not representative of what one might find if significantly more larger chemical structures were considered in the range of 30-50 non-hydrogen atoms. (*i.e.*, if the non-hydrogen atom



(a) “Small-then-large” approach



(b) “Large-then-small” approach

Figure 9.6: Figure 6. Two different approaches used to generate the unique shapes between  $V = V_1$  and  $V < V_1$ , depending on which shape space is clustered first

:sub:'1' and V<V:sub:'1', depending on which shape space is clustered firstTwo different approaches used to generate the unique shapes between  $V = V_1$ .

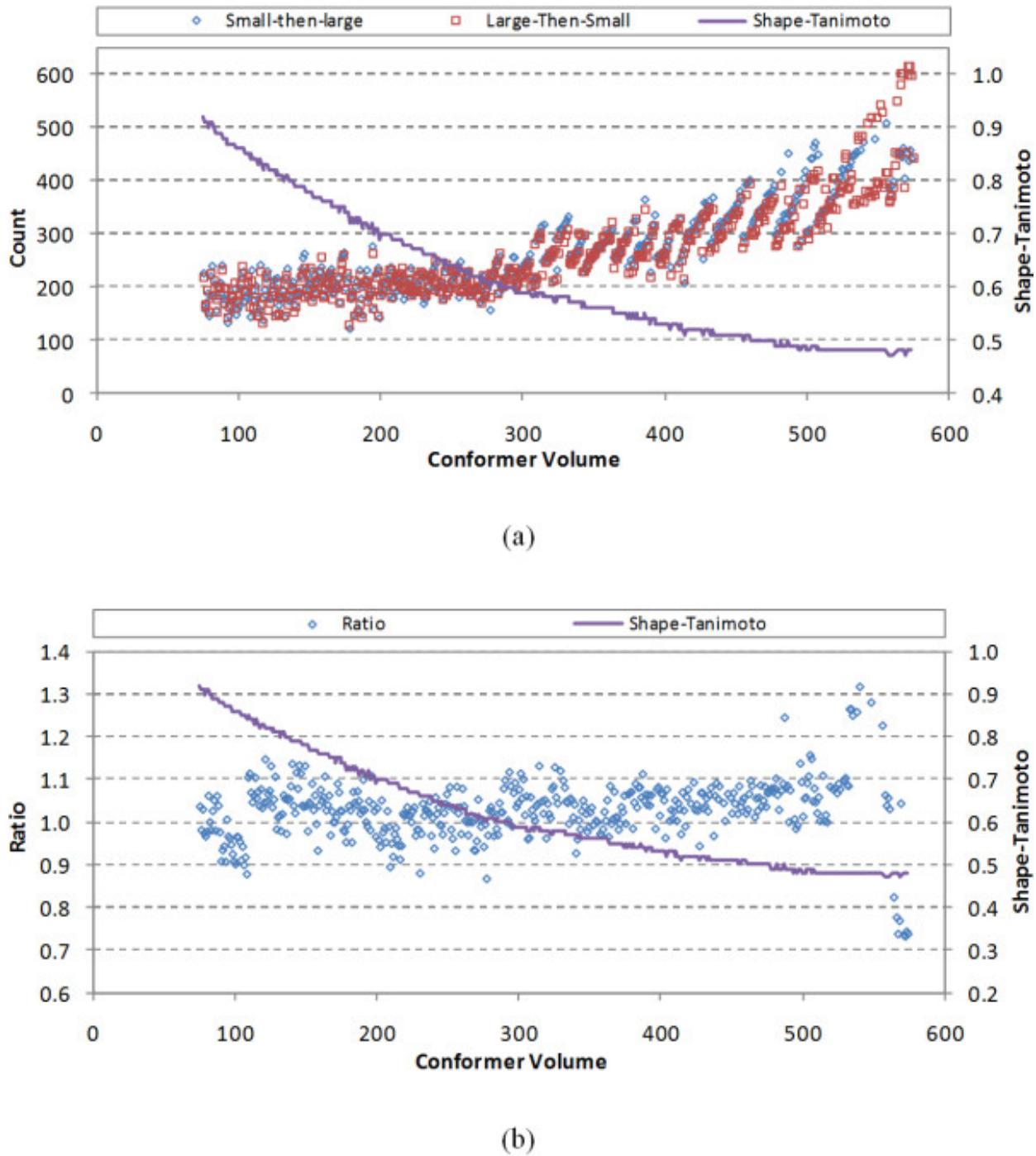


Figure 9.7: Figure 7. Unique shape counts

**Unique shape counts.** (a) The number of unique shapes generated by the “small-then-large” method and the “large-then-small” method, and (b) the ratio of “small-then-large” to “large-then-small” unique shapes as a function of conformer volume.

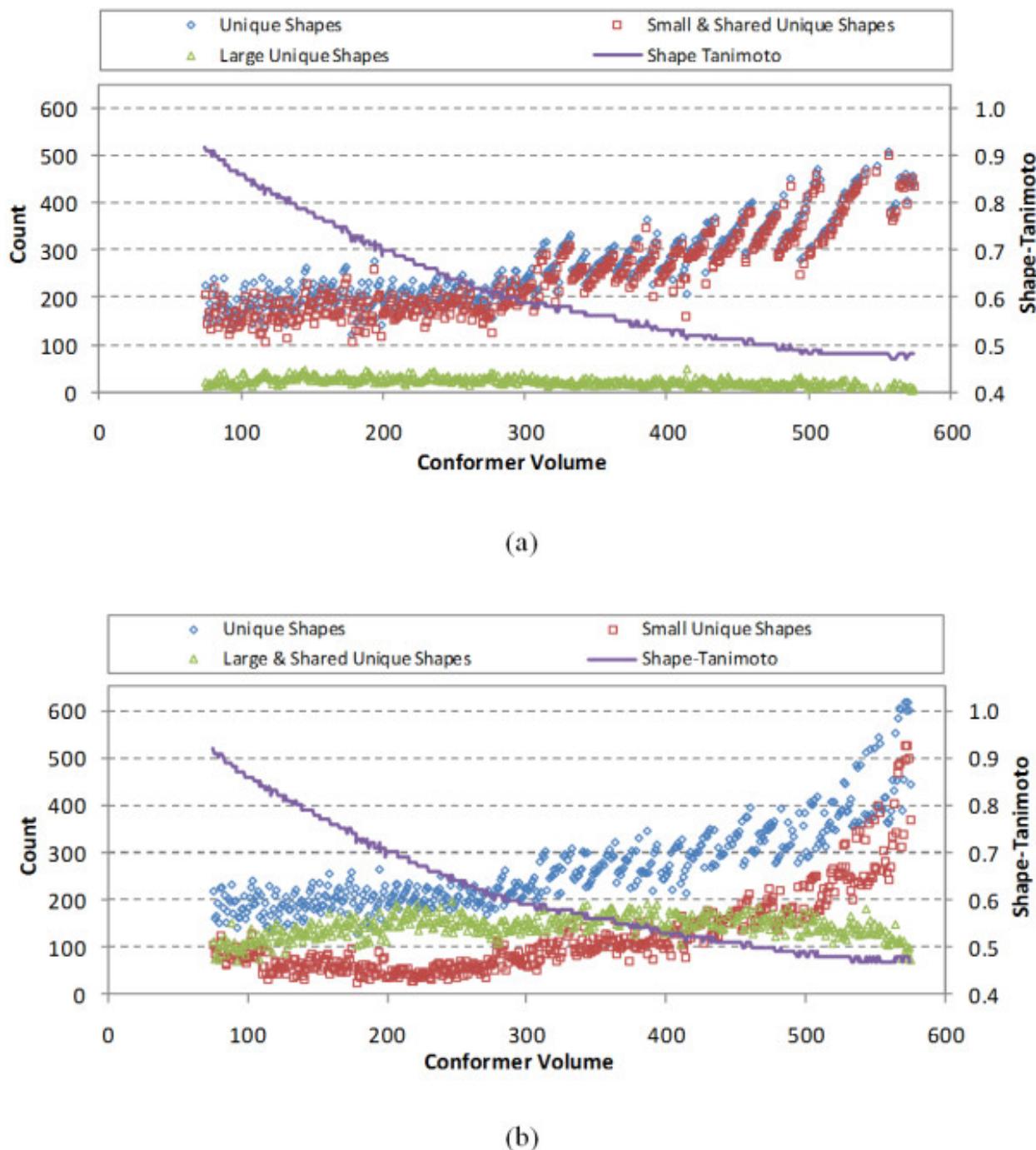


Figure 9.8: Figure 8. The number of unique shapes, small unique shapes, and large unique shapes generated using (a) the small-then-large method and (b) the large-then-small method

**The number of unique shapes, small unique shapes, and large unique shapes generated using (a) the small-then-large method and (b) the large-then-small method.**

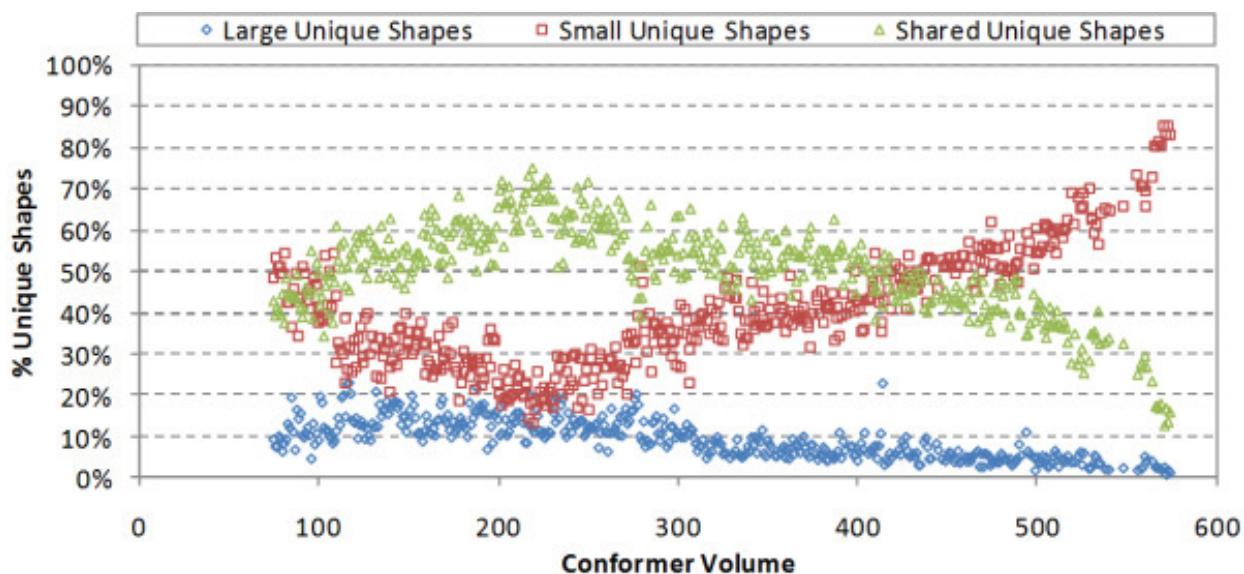


Figure 9.9: Figure 9. The percentages of the large unique shapes, small unique shapes, and shared unique shapes, being the percentage of space not covered by either large or small unique shapes [*i.e.*, shared = 1.0 - (large + small)], as a function of the conformer volume

*i.e.* The percentages of the large unique shapes, small unique shapes, and shared unique shapes, being the percentage of space not covered by either large or small unique shapes [.

count maximum was not at 26, but continued to grow until the maximum considered of 50.)

To further demonstrate this, Figure 10 shows the ratio of the fraction of the large unique shapes to the sum of the fractions of the large and shared unique shapes, which is a measure of how much of the shape space spanned by the conformers of a particular volume is not shared by the conformers smaller than that volume. For  $75 \text{ \AA}^3 \text{ V } 100 \text{ \AA}^3$ , the mean value of the ratio was 0.19, indicating that ~20% of the shape space spanned by the conformers of a particular volume in this range is unique to that particular volume, and that the other 80% is shared by the conformers smaller than that volume. The ratio decreases with the conformer volume, and the mean value for  $550 \text{ \AA}^3 \text{ V } 575 \text{ \AA}^3$  was 0.11, indicating the rate of the shape space growth decreases as the conformer volume increases, relative to the PubChem chemical structure contents.

## 9.4 Conclusion

The shape diversity of the biologically relevant conformer space of molecules and conformers was investigated using 16.4 million molecules in the PubChem Compound database (as of January 2008), covering non-hydrogen atom counts up to 50 and effective rotors up to 15, as represented by 1.46 billion diverse conformers. After binning the conformers according to their volume, cluster analysis was performed to get a maximum count of non-redundant reference shapes, representing the shape space spanned by the conformers for a particular unit volume. The  $\text{STvalue}^{\text{thresh}}$ , which defines the maximum shape similarity between any two reference shapes for that volume, gradually decreased as the conformer volume increased. There was no apparent correlation between the count of conformers clustered and the shape diversity found. Furthermore, an analysis was performed to examine the rate of increase of new reference shapes as a function of volume and the percentage of shape space unique to a particular volume. Generally speaking, the rate of addition of new reference shapes as a function of increasing volume was relatively constant across the range of volumes considered; however, the ability of a particular volume to explain the shape diversity spanned by lesser volumes increased up to a point and then decreased, ranging between 40-70% of all unique shapes for most of the considered volume range (Figure 9).

Some of the results from this analysis should be considered an artifact of the contents of PubChem in that the popula-

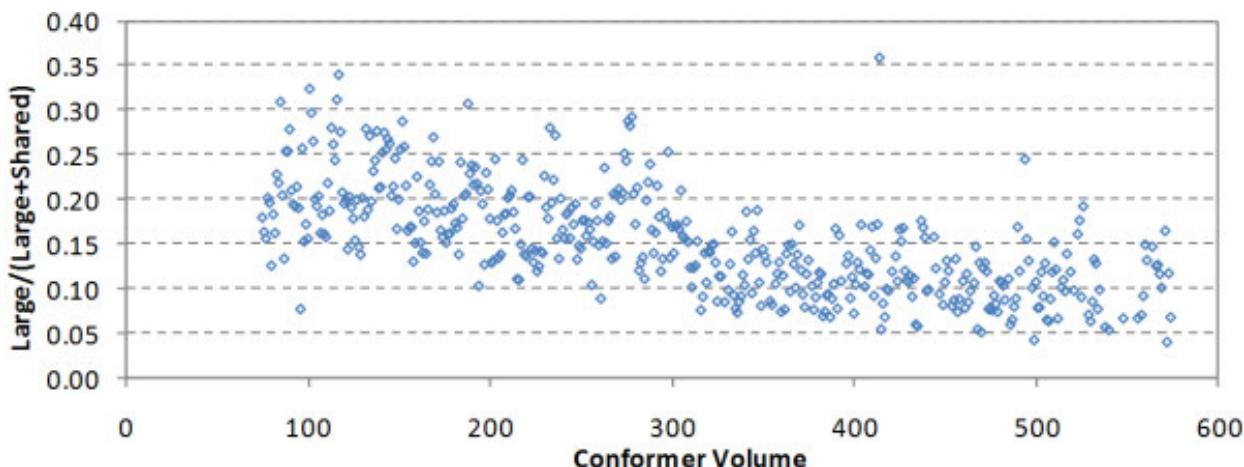


Figure 9.10: Figure 10. The ratio of the percent large unique shapes to the sum of the the percent large and shared unique shapes

**The ratio of the percent large unique shapes to the sum of the the percent large and shared unique shapes.**

tion as a function of molecular size peaks at 26 non-hydrogen atoms and then rapidly declines. An exhaustive analysis of all “reasonable” theoretically possible molecules resulting from larger molecules may provide a different trend. As such, the results of this analysis should be considered a conservative estimate.

While it is unfortunate that the PubChem shape space is truncated based on what is possible (due to the diminishing count of chemical structures with non-hydrogen atom counts greater than 26), one does see substantial evidence that shape space grows uniformly with a smoothly decreasing  $\text{thresh}^{\text{ST}}$  and increasing molecular volume. One also sees that keeping the count of reference shapes at a maximum for a given volume as an approach and analysis can allow one to achieve an understanding as to how diverse shape space is as a function of shape similarity. The apparent lack of dependence of the reference shape count with respect to the count of conformers represented by a given volume demonstrates how redundant shape space is across the volume range; however, we believe that the  $\text{thresh}^{\text{ST}}$  curve in Figure 3 may actually be linear or approach it, if provided an exhaustive set of theoretically possible but reasonable chemical structures, as the chemical structure shape possibilities surely are more diverse than the limited population of chemical structures available in PubChem for non-hydrogen atom counts greater than thirty.

## 9.5 Materials and methods

### 9.5.1 1. Biologically relevant molecules in the PubChem Compound database

From the PubChem Compound database<sup>15161718</sup>

1. Molecules only with a single covalent component were considered, since each component of mixtures and complexes has a unique Compound ID.
2. Salts were not included because their parent molecules are also in the PubChem database.
3. Molecules with a non-organic element were not included because they are not compliant with the 94 s variant of the Merck molecular force-field (MMFF94s), which was used for conformer generation (without coulomb interaction terms). For the same reason, molecules with an MMFF94 s unparameterized element type (*e.g.*, hyper-valent species) were removed.
4. Molecules that are too big or too flexible cannot have their conformational space properly sampled. Therefore, the non-hydrogen atom count was limited to a maximum of 50 and the effective rotor count was limited to 15.

<sup>18</sup> An overview of the PubChem BioAssay resource

The effective rotor count, given by the following equation, takes into account the additional flexibility due to non-aromatic rings in a molecule,

where  $n_{\text{effect}}^{\text{e}}$  is the number of effective rotors,  $nr$  is the number of rotatable bonds, and  $nn_{\text{ara}}$  is the number of “non-aromatic”  $^3\text{sp}$ -hybridized ring atoms.

5. Molecules with more than 6 undefined stereocenters were also removed because they need substantial computational resources to consider.

## 9.5.2 2. Conformer generation

The OMEGA C++ application programming interface (API)<sup>19</sup> was used to generate conformers for the molecules in PubChem and the Shape C++ API<sup>12</sup> was used to compute conformer analytic volumes. In a recent study<sup>20</sup>, a set of optimal values for some important OMEGA parameters was determined for PubChem 3-D conformer generation. This parameter set was employed for conformer generation in the present study. The MMFF94 s force-field without the coulomb interaction terms were used with the energy window limited to 25 kcal/mol. The number of conformers generated in the torsion search step was limited to 100,000 conformers. When undefined stereocenters were present, each stereo isomer was independently considered (maximum of 100,000 conformers each for up to 32 different SP3/SP2 stereo isomers) and all produced conformers combined. Conformers were then clustered using the root-mean-square distance [rounded to the nearest 0.2 increment (from 0.4 to 2.4)] given by the following equation<sup>20</sup>:

where  $n_{\text{effect}}^{\text{e}}$  is the number of effective rotors and  $n_{\text{ha}}$  is the number of non-hydrogen atoms in a molecule. The maximum number of conformers in a conformer model for each molecule was limited to 500. If clustering resulted in more than 500 conformers, the clustering RMSD was incremented by 0.2 and the conformers re-clustered, repeating until 500 or fewer conformers were achieved. Post processing of the conformer models was performed. This included full energy minimization of all hydrogen atom locations (all non-hydrogen atoms were kept frozen). Subsequent analysis removed any conformers with atom-atom “bumps”, being cases where the steric van der Waals interaction energy was greater than 25 kcal/mol.

## 9.5.3 3. General descriptions of the partition-clustering algorithm

Due to the rather large number of conformers involved, a “divide and conquer” approach with a multistage partition-based clustering algorithm (as shown in Figure 2) was employed. In the first phase of the partition-clustering algorithm, conformers were split into manageable sets (or partitions), each containing a certain number of conformers ( $N^{**}\text{setsize}^{\text{sup}}=50,000$ ). Conformers in each set were randomly sampled such that no two selected conformers had a ST distance closer than the shape diversity threshold ( $\text{thresh}^{\text{ST}}$ ). The selected conformers were retained for future analysis, as cluster representatives, while the others were considered redundant and discarded. If the count of selected conformers in a given partition was greater than  $\text{thresh}^{\text{ST}}$ . After all conformer sets were sampled using  $\text{thresh}^{\text{ST}}$ , all the conformers from each set were then combined and re-sampled as described above (e.g., divided up into partitions and sampled). When the total number of clusters became smaller than  $\text{thresh}^{\text{ST}}$ . In the end, the ST scores between any two conformers in the final reference set cannot be closer than the  $\text{thresh}^{\text{ST}}$ . A final step involved comparing the reference set with all conformers represented by the reference set.

This procedure achieves several things. Firstly, it breaks up many millions of conformers into manageable sets. Secondly, it allows the shape diversity threshold to be dynamically decreased for individual conformer sets. Thirdly, it reduces a very large number of conformers to a manageable set of conformers that represent all possible shapes present.

<sup>19</sup> OMEGA, Version 2.1, OpenEye Scientific Software, Inc.: Santa Fe, NM

<sup>20</sup> PubChem3D: conformer generation

### 3.1. Partition-clustering of conformers of a given volume

To study the shape diversity for a given volume, the conformers of the same volumes were partition-clustered, based on the procedures outlined in the previous section.

1. The 1.46 billion conformers were grouped according to their volumes rounded to the nearest integers.
2. The conformers for a given volume were partition-clustered until the total number of clusters became less than
3. The non-partition clustering was performed with the basis shapes, decreasing  $\text{thresh STvalue}$  0.01 at a time, until the number of cluster representatives became less than

### 3.2. Generation of unique shapes

To investigate the shape space redundancy between different volumes, the unique shapes (Figures 4 and 5) for each volume were generated using two different clustering schemes: (1) the “small-then-large” method and (2) the “large-then-small” method (Figure 6). In the small-then-large method, the unique shapes for  $V = V_1$  were generated from clustering of the reference and basis shapes for  $V < V_1$ , and re-clustering with the reference and basis shapes for  $V = V_1$ , to locate those shapes unique only to the current volume. On the contrary, in the large-then-small method, the unique shapes were generated by pooling the reference shapes for  $V = V_1$ , and re-clustering with the reference and basis shapes for  $V < V_1$ , to locate only those shapes that are unique to lesser volumes.

#### 1. The “small-then-large” approach

1. Pool all reference shapes of  $V < V_1$  and partition-cluster them at
2. Cluster the partition-clustered reference shapes [from step (1)] at
3. Pool all basis shapes for  $V < V_1$  and partition-cluster them at
4. Fill cluster holes in the clustered reference shapes [from step (2)], by re-clustering them with the partition-clustered basis shapes [from step (3)] at
5. Fill cluster holes in the clusters from step (4) with the reference shapes for  $V = V_1$ .
6. Fill cluster holes in the clusters from step (5) with the basis shapes for  $V = V_1$ .

#### 2. The large-then-small approach

1. Pool all reference shapes of  $V = V_1$ .
2. Pool all reference shapes of  $V < V_1$  and partition-cluster them at
3. Cluster the partition-clustered reference shapes [from step (2)] at
4. Pool all basis shapes for  $V < V_1$  and partition-cluster them at
5. Fill cluster holes in the clustered reference shapes [from step (3)], by re-clustering them with the partition-clustered basis shapes [from step (4)] at
6. Fill cluster holes in the clustered reference shapes for  $V = V_1$  [from step (1)] with the clusters from step (5).

## 9.6 Competing interests

The authors declare that they have no competing interests.

## 9.7 Authors' contributions

EEB performed most of the research. SK wrote the first draft and introduced the concept of small and large unique shapes to analyze the data. SHB reviewed the final manuscript. All authors read and approved the final manuscript.

## 9.8 Acknowledgements

We are grateful to the NCBI Systems staff, especially Ron Patterson, Charlie Cook, and Don Preuss, whose efforts helped make the PubChem3D project possible. This research was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, U.S. Department of Health and Human Services. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD. <http://biowulf.nih.gov>.



# INTERPRETING LINEAR SUPPORT VECTOR MACHINE MODELS WITH HEAT MAP MOLECULE COLORING

## 10.1 Abstract

### 10.1.1 Background

Model-based virtual screening plays an important role in the early drug discovery stage. The outcomes of high-throughput screenings are a valuable source for machine learning algorithms to infer such models. Besides a strong performance, the interpretability of a machine learning model is a desired property to guide the optimization of a compound in later drug discovery stages. Linear support vector machines showed to have a convincing performance on large-scale data sets. The goal of this study is to present a heat map molecule coloring technique to interpret linear support vector machine models. Based on the weights of a linear model, the visualization approach colors each atom and bond of a compound according to its importance for activity.

### 10.1.2 Results

We evaluated our approach on a toxicity data set, a chromosome aberration data set, and the maximum unbiased validation data sets. The experiments show that our method sensibly visualizes structure-property and structure-activity relationships of a linear support vector machine model. The coloring of ligands in the binding pocket of several crystal structures of a maximum unbiased validation data set target indicates that our approach assists to determine the correct ligand orientation in the binding pocket. Additionally, the heat map coloring enables the identification of substructures important for the binding of an inhibitor.

### 10.1.3 Conclusions

In combination with heat map coloring, linear support vector machine models can help to guide the modification of a compound in later stages of drug discovery. Particularly substructures identified as important by our method might be a starting point for optimization of a lead compound. The heat map coloring should be considered as complementary to structure based modeling approaches. As such, it helps to get a better understanding of the binding mode of an inhibitor.

## 10.2 Background

High-throughput screenings (HTS) play an important role in the early drug discovery stage. The data of these HTS are a valuable, but challenging resource for machine learning algorithms to infer predictive structure-activity relationship models for virtual screening<sup>1</sup>. In later stages of drug discovery, a lead compound is optimized for desired biophysical properties. However, as a lead compound becomes increasingly tailored to a target, there is generally less tolerance for introducing changes without an intrinsic affinity penalty<sup>2</sup>. Thus, besides a strong performance, the reasons that lead to a prediction of a compound as active or inactive is important for a medicinal chemist in lead optimization.

Recent examples of interpretable methods applied to cheminformatic problems include Naïve Bayes, decision trees, and k-nearest neighbor approaches. Bender et al.<sup>3</sup> applied Bayesian learning to radial atom environments and used the information gain to assess the significance of a substructure. Han et al.<sup>4</sup> trained decision trees on several PubChem HTS data sets. Swamidass et al.<sup>5</sup> introduced the Influence Relevance Voter, an interpretable method based on a supervised artificial neural network in combination with a k-nearest neighbor approach. Recently, Mohr et al.<sup>6</sup> employed a potential support vector machine (SVM) in combination with a maximum-common subgraph kernel to predict the genotoxicity of a compound.

Using a two step procedure, Mohr et al. labeled the atoms of a compound as important or unimportant for genotoxicity. First, the design of the potential SVM allows for the assignment of weights to atoms. Second, based on the weights, an atom is classified as important for genotoxicity if a predefined threshold is exceeded.

Linear SVMs in combination with sparse molecular fingerprints showed a convincing performance on several large-scale data sets<sup>7</sup>. In contrast to their nonlinear counterpart, linear SVMs are no black box concerning interpretability because they do not perform a nonlinear mapping from the input space to a high-dimensional feature space. Linear SVMs learn a linear discriminant function, which assigns a weight to each fingerprint feature of the input space. Recent studies<sup>8,9</sup> indicate that the interpretation of linear SVM models is possible for small regression data sets. Both approaches exploited the weights of a linear support vector regression model to extract patterns which are important for activity or selectivity against a certain protein target.

The aim of this study is to present a visualization method that allows for the interpretation of linear SVM models of large-scale data sets. We use the weights of the linear discriminant function to assign a score to each atom or bond of a compound. Based on these scores, a color is assigned to each atom or bond of a compound. We tested the visualization approach on the Kazius Ames toxicity data set<sup>10</sup>, the chromosome aberration data set compiled by Mohr et al.<sup>6</sup>, and the maximum unbiased validation data sets<sup>11</sup>.

The results show that our method is able to sensibly visualize the structure-property and structure-activity information of a linear SVM model. The heat map visualization can be combined with structure based modeling approaches to gain a better understanding of the binding mode of a compound and therefore help medicinal chemists in lead optimization.

---

<sup>1</sup> Integration of virtual and high-throughput screening

<sup>2</sup> Hit and lead generation: beyond high-throughput screening

<sup>3</sup> Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier

<sup>4</sup> Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem

<sup>5</sup> Influence relevance voting: an accurate and interpretable virtual high throughput screening method

<sup>6</sup> A maximum common subgraph kernel method for predicting the chromosome aberration test

<sup>7</sup> Large-Scale Learning of Structure-Activity Relationships Using a Linear Support Vector Machine and Problem-specific Metrics

<sup>8</sup> A Free-Wilson-like Approach to Analyze QSAR Models Based on Graph Decomposition Kernels

<sup>9</sup> Visualization of Molecular Selectivity and Structure Generation for Selective Dopamine Inhibitors

<sup>10</sup> Derivation and validation of toxicophores for mutagenicity prediction

<sup>11</sup> Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data

## 10.3 Methods

### 10.3.1 Nonlinear vs. linear SVM models

A virtual screening data set of  $l$  compounds can be represented as a set of  $l$  labeled fingerprints of compounds ( $\mathbf{x}_i, y_i$ ),  $i = 1, \dots, l$ ,  $\mathbf{x}_i = (x^{**i} : sub: 'im' x, : sub: 'ij' x \nonumber_2 \nonumber_3, : sub: 'i' y \nonumber_4 \{-1, +1\})$ . In the case of binary substructure fingerprints, each  $: sub: 'ij' x \nonumber_5 \{0, 1\}$  is an indicator for the presence or absence of a pattern (or fingerprint feature)  $*j$  in compound  $i$ . Every compound is labeled either as active ( $_i y = +1$ ) or inactive ( $_i y = -1$ ). SVMs learn a discriminant function of the form

where  $SV = \{\mathbf{x}_i | : sub: 'i' \nonumber_6 > 0\}$  is the set of support vectors,  $k$  is the kernel function and  $\nonumber_7$  is a mapping from the input space to a high-dimensional feature space. In general, the actual mapping  $\nonumber_8$  is unknown. The kernel  $*k$  implicitly performs the mapping and the calculation of the dot product  $\nonumber_9 (**x_i)^T \nonumber_{10} (**x)$ . Thus, in case of nonlinear SVMs, it is impossible to assess how a certain training set feature  $ij$   $x$  contributes to the kernel similarity  $k(**x_i, x)$ . Hence, nonlinear SVMs are a black box concerning interpretability.

In case of linear SVMs, the mapping  $**$  is the identity, which results in a linear discriminant function of the form

where the weight vector  $\mathbf{w} = (w_1, \dots, w_m)$  is optimized such that the separating hyperplane defined by  $f(**x)$  has maximum margin. The discriminant function is employed to predict the class  $sign(f(**x))$  of an unknown sample. The value of  $f(**x)$  is called prediction value. Compounds with a prediction value close to zero are close to the separating hyperplane. Consequently, the classification can be interpreted as less certain.

The weight vector  $\mathbf{w}$  can be expressed by

Hence, the weight vector  $\mathbf{w}$  contains the weighted features of the support vectors. The SVM assigns an  $\nonumber_{12} > 0$  if the compound is necessary for class separation. Thus, a compound  $*i$  which contains no information for class separation will have  $_i \alpha = 0$ . Its features  $ij$   $x$  will not have a weight  $j$   $w$  unless another compound with an  $\nonumber_{15} > 0$  also contains the pattern. Additionally, the weight  $: sub: 'ij' w = : sub: 'i' y : sub: 'i' x : sub: 'ij' \nonumber_{16}$  of a feature of a compound with  $: sub: 'i' \nonumber_{17} > 0$  is positive if the class of a compound  $**x_i$  is labeled active ( $_i y = +1$ ) and negative otherwise.

The linear discriminant function  $f(**x)$  in combination with binary substructure fingerprints is equivalent to the Free-Wilson formulation in chemometrics<sup>12</sup><sup>13</sup>. The Free-Wilson formulation assigns each pattern  $j$  a weight  $j$   $w$  according to its contribution to activity. In contrast to the Free-Wilson formulation, a linear SVM model is a classification model and not a regression model. Thus, the weight of a feature does not represent the actual contribution to binding affinity, but the relative importance of a feature. Still, a linear SVM model can in principle be interpreted in the same way as a Free-Wilson model. A pattern with large positive weight is expected to be important for activity of a compound while a pattern with large negative weight should represent inactive or non-relevant parts of a molecule. A pattern with a weight close to zero should be unimportant for class separation.

A more detailed description of maximum-margin based classifiers can be found in Schölkopf and Smola<sup>14</sup> for further reading.

### 10.3.2 Molecular Encodings

For structure-based classifiers, it is common to encode the molecular graph of a compound with binary fingerprints for large-scale learning tasks in cheminformatics. Each bit of a fingerprint indicates the presence or absence of a fingerprint feature. The specific choice of molecular encoding is crucial to obtain an interpretable linear model. The employed encoding must ensure that the set of atoms or bonds which a fingerprint feature represents is available while calculating the molecular encoding. Consequently, the weight of a fingerprint feature can be mapped back to those atoms or bonds. The mapping from sets of atoms or bonds to fingerprint features needs not to be injective. If a collision

<sup>12</sup> A mathematical contribution to structure-activity studies

<sup>13</sup> Free Wilson Analysis. Theory, Applications and its Relationship to Hansch Analysis

<sup>14</sup> NOTITLE!

occurs, the weight of a feature is mapped to all sets of atoms or bonds that caused the collision. Common molecular encodings that generate interpretable features are radial atom environments<sup>3</sup>, depth first search paths<sup>15</sup>, or extended connectivity fingerprints (ECFP)<sup>16</sup>.

We employed a variant of the ECFP to encode the molecules because ECFP features are intuitively interpretable and can be mapped back to the topology of a chemical graph. Each ECFP feature of a fingerprinted compound represents a circular substructure around a center atom. The algorithm starts with the initial atom identifier (in our case Daylight invariants<sup>17</sup>) of the center atom and grows a circular substructure around this atom. This growing can be done implicitly, like in the original algorithm, or explicitly, like in our variant. In the original algorithm the identifiers of the alpha atoms of a center atom are used to calculate an updated identifier for the center atom. In each iteration, the identifiers of the previous iteration are used as atom identifiers. This iterative procedure implicitly grows a circular substructure around the center atom because with an increasing number of iterations the updated identifier of a center atom contains information from further and further away. Our variant does the growing of a substructure explicitly by keeping the circular substructures of the previous iteration and their possible attachment points in memory. In each iteration the circular substructures are extended at their possible attachment points using the initial atom identifiers (Figure 1). Both ECFP variants generate an ECFP feature for each possible center atom and iteration. We evaluated the performance of both ECFP variants and could not observe a significant difference. We employed our variant because the information contained in a feature is defined more precisely.

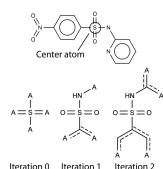


Figure 10.1: Figure 1. Illustration of the ECFP

**Illustration of the ECFP.** Each circular substructure around a center atom represents an ECFP feature. The circular substructure is grown in each iteration at the attachment points A. Any atom can be matched on an attachment point. Aromatic bonds are marked by a dashed line.

To be able to look up the substructure information of an ECFP feature, we saved the mapping from fingerprint feature identifiers to circular substructures while calculating the fingerprints of a data set. Due to the hashing, which is conducted to assign a fingerprint feature identifier, it is possible that a features identifier maps to several different substructures. However, collisions can be minimized by choosing a sufficiently large hash space.

### 10.3.3 Heat map molecule coloring

To allow a medicinal chemist to interpret a linear model, it is intuitive to color each atom or bond of a molecule according to its importance for activity. This coloring is achieved by the heat map molecule coloring.

Another intuitive approach to interpret a linear model is to select all patterns that exceed a certain weight threshold, as done in a study by Fechner et al.<sup>8</sup>. Their pattern selection approach proved to be useful on small data sets which share a common scaffold. However, preliminary experiments showed that the approach does not lead to interpretable results on large, diverse chemical data sets of assay outcomes, especially for external predictions. The top 5 ranked patterns (examples for several employed data sets can be found in Additional file *AUC > 0.9*), the chance to find a singular predictive pattern in the set of selected patterns (weight  $3 \times \ln(\text{nonascii\_211}^*)$ ) for the training set is small ( $*p < 0.3$ ). This probability drops considerably for the test set. A reason for this low probability might be that several patterns that by themselves cannot separate the two classes well might separate the classes better if combined<sup>18</sup>. For example, two patterns can perform perfectly random on their own, but can perfectly separate both classes if combined. However, combining patterns can not be accomplished by a method that inspects single patterns separately. Another reason can

<sup>15</sup> Graph kernels for chemical informatics

<sup>16</sup> Extended-Connectivity Fingerprints

<sup>17</sup> SMILES. 2. Algorithm for generation of unique SMILES notation

<sup>18</sup> An introduction to variable and feature selection

be the redundancy of the ECFP. The SVM might split the importance (weight) of a substructure among the redundant patterns that represent it.

#### Additional file 1

**Number of support vectors and top weighted fragments.** For each data set with an *AUC* 0.7 the number of support vectors and the top five weighted fragments are listed in the file.

[Click here for file](#)

Consequently, the patterns appear less important than they actually are. This problem is circumvented by the heat map coloring technique because it integrates the information of all training patterns. Hence, the importance of all redundant patterns that represent a certain substructure is fused in the coloring of a molecule. Additionally, in a chemical compound, patterns with different weights might overlap, making the interpretation unintuitive. In contrast, the information of overlapping patterns is integrated and intuitively visualized by the heat map coloring approach.

The inputs of our method are a previously trained linear SVM model, a list of compounds of interest and possibly the whole training data set depending on the employed normalization. Our method can be divided into two separate steps. First, our algorithm assigns a score to each atom and bond of a molecule based on the weights of the linear SVM model. Second, the scores are transformed to a color on a color gradient.

In this manuscript, we only performed bond coloring. Thus, we only explain the calculation of bond scores. The calculation of scores is analogous for atoms and bonds. Hence, the presented calculation of bond scores can be easily transferred to atoms if needed. We had two reasons for solely using bond coloring. Firstly, the ECFP focuses on connectivity information which can be better visualized using bond coloring. Secondly, we wanted to allow for easy element identification by using element type atom coloring. For a different fingerprinting algorithm, like radial atom environments, atom colorings might be more useful. To assign a score to each bond, our algorithm fingerprints the compounds of interest or the whole training data set again. Throughout the fingerprinting process the information, which bonds a fingerprint feature represents in a certain compound, is stored. Based on the weights of the fingerprint features, a score is assigned to each bond of a compound. The score of a bond  $b$  is equal to the sum of weights of the fingerprint features that contain the bond (Figure 2). Thus, the score  $s_b$  of a bond  $b$  is

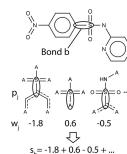


Figure 10.2: Figure 2. Illustration of pattern to bond weight mapping

**Illustration of pattern to bond weight mapping.** The weight  $w_j$  of a pattern  $j$  is added to the score  $s_b$  of a bond  $b$  if the bond is contained in the pattern  $j$ . Attachment points A can be mapped on any atom. Aromatic bonds are marked by a dashed line.

where  $B^*(f)$  is the set of bonds a feature  $f$  represents and  $w^*(f)$  is the weight of a feature  $f$  as found in the linear SVM model. In the case of ECFP features, the bonds to attachment points are included in the set of bonds a feature represents. Now, each bond of a processed compound has a score according to the weights of a trained linear SVM model. Therefore, the SVM model is responsible for assigning sensible weights if a smaller sub-fragment occurs in larger activating and non-activating fragments.

The score of a bond can not be transformed to a color directly. The score needs to be normalized to [0,1], which is achieved by

where  $\max s$  and  $\min s$  are the maximum and minimum score found, respectively. Two different normalizations are possible depending on how  $\max s$  and  $\min s$  are chosen. The first method (full set normalization) chooses  $\max s$  and  $\min s$  to be the maximum and minimum score found throughout the fingerprint calculation on the whole training data set and the compounds of interest. The second approach (single molecule normalization) sets  $\max s$  and  $\min s$  to the maximum and minimum score found in the current compound.

Both normalizations have advantages and disadvantages. The full set normalization keeps the information of the prediction value differences between compounds. Generally, a compound with a larger prediction value has larger atom

and bond scores. The relative differences of scores between compounds are not changed by the full set normalization. Thus, the relative differences in their prediction value are kept. A disadvantage of the full set normalization is that small differences between scores of a compound might not be visible. If the difference between two scores of a compound is small compared to the maximum score of the whole data set, then the difference is even smaller after normalization. Furthermore, the whole training data set needs to be fingerprinted again, which can take a considerable amount of computation time for large data sets. In contrast to the full set normalization, the single molecule normalization visualizes small differences in the scores of a compound better. Furthermore, the computation time is fast because only the compounds of interest need to be processed. The main disadvantage of the single molecule normalization is that the coloring does not contain any information about the overall activity because only differences within the molecule are taken into account.

Another normalization aspect concerns the calculation of the bond score  $b_s$ . In our implementation we do not take the fragment size of a fingerprint feature into account. To take the fragment size into account, the contribution of a fragment to a bond score  $b_s$  is divided by the number of bonds in the fragment. Thus, the weight of a feature is equally distributed among the bonds associated with the feature. This weight distribution lowers the influence of large fragments on the bond score  $b_s$ . However, putting the focus on smaller fragments might be less suited for large, diverse HTS data sets, where the scaffold is most important. Additionally, experiments (not shown) indicated that in case of the ECFP, the colorings are smoother without equal distribution of weights. Thus, we decided not to use equal weight distribution for our visualization.

After normalization to [0,1], a score can be transformed to a color on a color gradient. We use a color gradient from red over orange to green, where red represents the negative class and green represents the positive class. The whole gradient can be divided into two sub-gradients, one from red to orange and another one from orange to green. The first sub-gradient is used if

where  $rR$ ,  $oR$ , and  $gR$  are the values of the R color channel of the respective gradient colors red, orange, and green. The colors are mixed for each color channel separately and the resulting color  $(_{mix}, G_{mix}, B_{mix} R)$  is assigned to the respective bond.

## 10.4 Experimental

### 10.4.1 Virtual screening data sets

We conducted evaluation experiments on 17 maximum unbiased validation (MUV) data sets, an Ames toxicity data set (Kazius), and a chromosome aberration (CA) data set. A detailed analysis was conducted for the Kazius data set and two of the MUV data sets.

First, we employed the 17 MUV data sets compiled by Rohrer et al.<sup>11</sup> with their corresponding background data sets. Each of these data sets comprises 30 dissimilar active compounds together with 15,000 inactive compounds, which are similar to the actives with respect to several simple descriptors like volume, solubility, or mass. The MUV data sets are designed to avoid artificially high screening performance by inappropriate decoys. Additionally, the common spread of actives can have a positive impact on the interpretability of a model because the chance of overfitting a model on a small cluster of similar actives is minimized. Two of these data sets, MUV548 and MUV846, were subjected to a detailed analysis. MUV548 contains inhibitors of protein kinase A as actives and MUV846 contains inhibitors against factor XIa.

Second, we used the mutagenicity data set composed by Kazius et al.<sup>10</sup>. It comprises 2401 mutagenic and 1936 non-mutagenic compounds based on the Ames toxicity test. Kazius et al. derived 29 toxicophores from the data. Using these toxicophores, they could predict the toxicity of an external test set with an accuracy of 85% which is close to the theoretical limit of the Ames toxicity test. The authors provide a list of well defined toxicophores and demonstrate their predictive power on several compounds.

Third, we employed the CA data set compiled by Mohr et al.<sup>6</sup>. This data set consists of 351 positive and 589 negative compounds with respect to the chromosome aberration test. Mohr et al. achieved an accuracy of 89.5% on

10 predefined cross-validation folds. The data set was included because the authors' method provides visual structure-activity information on several compounds of the data set.

## 10.4.2 Experimental setup

All data sets were prepared according to the guidelines by Fourches et al.<sup>19</sup>. The structures were canonicalized and transformed with JChem Standardizer<sup>20</sup>. The options of Standardizer were set to neutralize, tautomerize, aromatize, calculate clean 2 D coordinates, and add explicit hydrogens. Explicit hydrogens were added because CDK<sup>21</sup>, the core library of the employed fingerprinting algorithms, requires correctly attached hydrogens bonded to an atom. Then, all employed data sets were checked for duplicates and fingerprinted using the ECFP variant with a depth of four and a hash space size of 2<sup>22</sup> to minimize collisions.

To evaluate the performance on each data set, we used a 5-fold two-deep cross-validation<sup>23</sup> which was repeated two times. We employed the large-scale linear SVM LIBLINEAR<sup>23</sup> to train a linear SVM model. On the CA data set we also performed an evaluation on the 10 defined splits of Mohr et al. The SVM parameter  $C$  and the weight of the negative class  $W_{-1}$  were searched using a 2-fold cross-validation.

For  $C$  we chose the grid  $\log_2(C)$  {-5, -4, ..., 7, 8} and for  $W_{-1}$  we used the grid  $\log_2(W_{-1})$  {-4, -2, 0}. We also discarded uncommon features from the data before building the model. A feature had to occur at least 3 times to be included in the training. For the detailed heat map coloring analysis we used the whole data set to train a model and only left the compound of interest for analysis out. The same  $C$  and  $W_{-1}$  grids were used.

We employed two different measures to evaluate the performance on the data sets. First, the accuracy (ACC) was computed, which is the number of correctly predicted compounds divided by the total number of compounds. The accuracy is only applicable for balanced data sets like the Kazius and CA data set. Second, we employed the area under the ROC curve (AUC). The ROC curve plots the fraction of correctly predicted actives (true positive rate) against the fraction of inactives incorrectly predicted as actives (false positive rate) for every possible threshold. The AUC is applicable for all used data sets. The higher the value of both measures, the better is the performance.

All employed PDB structures were prepared with the protein preparation wizard of Schrödinger 2010<sup>24</sup>. The settings of the preparation wizard were set to the default settings.

## 10.5 Results and Discussion

The results of the analysis of the 19 employed data sets are organized as follows. First, we present the performance on the employed data sets. Then, we briefly explain, why we selected MUV548, MUV846, and the Kazius data set for a detailed analysis and visualization. Finally, we demonstrate and discuss the heat map molecule coloring method on those three selected data sets.

### 10.5.1 Selection of data sets for visualization

We selected the three data sets of the detailed analysis by two criteria. First, the performance of a linear model trained on a data set must be reasonably good because the predictive performance of a model should be crucial to obtain sensible structure-activity relationships. Second, to be able to validate the results of a visualization, literature information on structure-activity relationships of the target of a data set must be available.

<sup>19</sup> Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research

<sup>20</sup> JChem 5.3.8

<sup>21</sup> The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics

<sup>22</sup> On the use of cross-validation to assess performance in multivariate prediction

<sup>23</sup> LIBLINEAR: A Library for Large Linear Classification

<sup>24</sup> Schrödinger 2010

The linear SVM could predict the CA data set on the 10 predefined splits with an accuracy of 72% (Table 1), which is comparable to the nonlinear SVM performance reported by Mohr et al. Although the performance of the linear SVM is considerably worse than the method of Mohr et al. (89,5%), we chose to further analyze this data set because Mohr et al. provide visualizations of several compounds of the CA data set. We predicted all compounds that were illustrated in their study externally. All of them were either predicted wrong or had a prediction value close to zero and thus were not convincing predictions. The heat map coloring showed few overlap with the substructures identified by their visualization method, which is presumably due to the 17% lower accuracy and therefore more inaccurate model. Hence, the predictive performance of a model seems to be crucial to obtain sensible structure-activity relationships with our method.

On the Kazius data set, the linear SVM achieved an accuracy of 84%, which is close to the theoretical limit of the Ames toxicity test. The convincing performance and the availability of defined toxicophores from Kazius et al. make this data set an ideal choice for a more detailed analysis.

The performance on the different MUV data sets varied between an AUC of 0.58-0.96. All data sets with kinase, protease and chaperone targets showed a promising AUC performance (0.82-0.96), whereas the protein-protein interaction and reporter gene dependent assays had a considerably worse AUC performance (0.58-0.78). It would have been interesting to analyze the data sets with a GPCR target because it is hard to get crystal structures for those targets. Therefore, information on structure-activity relationships can not be obtained from structure based modeling for GPCR targets, which would make information from our visualization valuable. However, the performance on these data sets is close to random, and thus, a visualization would not be sensible. We chose the MUV548 which can be predicted with an AUC of 0.90 and the MUV846 with an AUC of 0.958 for a detailed analysis and visualization with the heat map coloring method. The linear models of both data sets exhibit a top ranked performance compared to all data sets and plenty of literature is available for the protein targets. While the MUV832 has the best performance (0.96), there are no crystal structures, which contain a ligand similar to the data set, available. Therefore, the MUV832 was not subjected to a detailed analysis.

### 10.5.2 Visualization of Kazius Ames toxicity data set

Using a linear model of the Kazius data set, we externally predicted compounds 1028-11-1 ( $C_A$ ) and 146795-38-2 ( $C_B$ ), and applied our heat map atom coloring method with both normalization variants. Figure 3(A,B) shows the heat map coloring of the correctly predicted non-toxic compound  $C_A$ . The toxic and the detoxifying substructures described by Kazius et al. could be identified with our method. However, in addition to the detoxifying sulfonamide our method also colored parts of the aromatic ring structure red (non-toxic), which is probably caused by the fact that the sulfonamide is often attached to an aromatic ring in the data set. Compound  $C_B$  (Figure 3C,D) was correctly predicted as toxic. The compound contains the same aromatic nitro toxicophore as compound  $C_A$ , which our method identified together with parts of the attached aromatic ring. In contrast to  $C_A$ , compound  $C_B$  is toxic because it does not contain a detoxifying sulfonamide. However, the compound has a red colored chlorobenzene substructure, which is non-toxic and not detoxifying in case of  $C_B$ . The overall toxicity of both compounds is visualized by the full set normalization. Compound  $C_A$  (Figure 3B) is more reddish compared to compound  $C_B$  (Figure 3D) and therefore has a lower prediction value.

The coloring of the compounds reveals weaknesses of both normalization methods. When the single molecule normalization is applied, one can not distinguish between a non-toxic and a detoxifying substructure because the normalization can only visualize differences within the structure. Thus, it is impossible to decide if a substructure is detoxifying or non-toxic without additional information on the toxicity. Given a compound that only contains toxicophores, the most toxic substructure would be colored green and the least toxic weighted substructure would be colored red. However, the information on the toxicity of a compound is available in form of the prediction value. Thus, this weakness can be compensated. The visualization of compound  $C_A$  (Figure 3B) indicates the drawbacks of the full set normalization method: While it captures the overall toxicity of the compound, the aromatic nitro toxicophore and the detoxifying sulfonamide are less distinguishable.

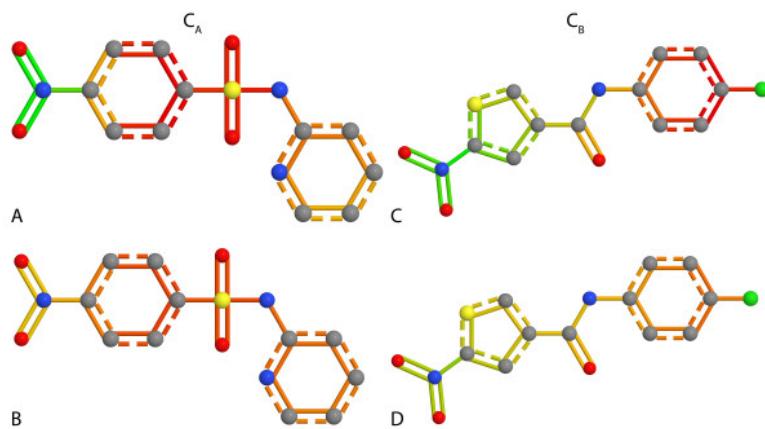


Figure 10.3: Figure 3. Kazius data set example compounds

**Kazius data set example compounds.** A heat map coloring of the non-toxic compound 1028-11-1 (C<sub>A</sub>) and the toxic compound 146795-38-2 (C<sub>B</sub>). Both compounds were predicted correctly. The color gradient ranges from green (toxic) to red (non-toxic). Both, the single molecule normalization (A,C) and the full data set normalization were applied (B,D). Compound C<sub>A</sub> contains a correctly identified aromatic nitro toxicophore. However, the compound has a detoxifying sulfonamide as well, rendering the compound non-toxic. The sulfonamide and parts of the aromatic ring were identified as non-toxic. In compound C<sub>B</sub> the aromatic nitro toxicophore was also identified as toxicophore. Compound C<sub>B</sub> is toxic because the red chlorobenzene substructure is not a detoxifying substructure.

### 10.5.3 Visualization of MUV548 protein kinase A data set

We conducted external predictions for the ligands of the PDB entries A, B, and C using a model trained on MUV548. Then, we applied our heat map coloring to the ligands. The employed PDB entries were the most suitable ones of several crystal structures available for protein kinase A because of their similarity to the compounds contained in MUV548. The ligands of other PDB structures are more dissimilar and thus presumably not in the applicability domain of a model trained on MUV548. The crystal structures of L<sub>B</sub> and L<sub>C</sub> originate from a study by Orts et al.<sup>25</sup>. The authors used an NMR based method to determine the binding orientation of low-affinity inhibitors. The method allowed for selecting the correct binding orientation of both ligands from four different orientations gained by rotation of the ligands in the binding pocket. To elucidate if our coloring method can also assist to select the correct binding orientation, we applied our method to the ligands presented in the study of Orts et al. and additionally ligand L<sub>A</sub>. The crystal structure of L<sub>A</sub> stems from an analysis of selective inhibitors against Akt1 (PKB)<sup>26</sup>. The ligand inhibits PKA with an IC<sub>50</sub> of 3.2  $\mu$ mol/L.

The coloring of the substructures of all three ligands correlates with the substructure position in the binding pocket (Figure 4). Especially the position of the green colored basic aromatic ring deep in the binding pocket is conserved for the three structures. If the ligands were rotated by 180° around the y axis, the red colored (unimportant) substructures would be located deep in the binding pocket. Hence, our approach assists to find the correct orientation of the ligands if rotated around the y axis. This rotation excludes two of the possible four orientations described by Orts et al. However, in case of the presented ligands our approach can not help to discriminate between rotations around the z axis. Yet, this limitation is not a real drawback for our method because it can be applied on the data of an HTS without performing additional NMR experiments.

The external prediction value of ligand L<sub>C</sub>, as indicated by the full set coloring (Figure 4), is lower than the prediction value of the other two ligands. On the ranked list of prediction values the ligands L<sub>A</sub> and L<sub>B</sub> are under the top 1% while the ligand L<sub>C</sub> is at position 4, 206 of 15,003 compounds. Hence, the coloring of L<sub>C</sub> could be inaccurate and changed if the ligands are included in the training set. Thus, we added all three ligands to the training set and applied the heat map coloring again. As expected, the training prediction values of all three ligands then were under the top 1% of the ranked list of training prediction values. The new position of the ligand L<sub>C</sub> in the ranked list was reflected by the full

<sup>25</sup> Crystallography-independent determination of ligand binding modes

<sup>26</sup> Design of selective, ATP-competitive inhibitors of Akt

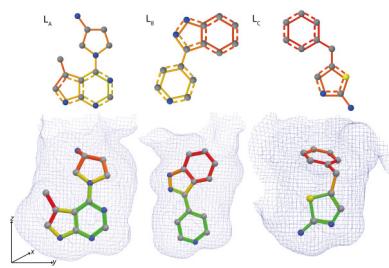


Figure 10.4: Figure 4. Orientation of different protein kinase A ligands

**Orientation of different protein kinase A ligands.** Binding orientation of the ligands of PDB entries A), B), and C). Compounds within the binding pocket were colored with the single molecule normalization, the compounds above with the full set normalization. The color gradient ranges from green (important for activity) to red (unimportant or even decreasing for activity). The binding pocket is indicated as an exclusion surface. Substructures, which are located at similar positions in the binding pocket, were colored similarly by the heat map coloring approach.

set normalized coloring. However, the single molecule normalized coloring did not change considerably for any of the ligands. The change in the full set normalized coloring was caused by a positive weighting of a large substructure of ligand L<sub>C</sub>. Re-weighting large substructures does not substantially influence weight differences within the molecule. Hence, in the case of L<sub>C</sub> the single molecule normalization might be more robust than the full set normalization.

The approach to compare the colorings of an external prediction and an inclusion in the training set might be a way to estimate the robustness of a coloring. While test compounds should never be included in model training when building predictive models, our intent is to build a descriptive model to identify features crucial for a molecule's molecular behavior. In the later case, inclusion of a compound for model building might be beneficial because additional information for finding important features is available. However, if the coloring of a compound changes drastically after inclusion in the model training, the descriptive model might not be sensible or structural aspects of a certain scaffold were not included. A robust model should not swap from completely meaningless features to sensible features by inclusion of just one compound in model training.

To evaluate if our visualization colors those substructures important for the interaction between a ligand and the target, we aligned the binding pockets of the crystal structures of L<sub>A</sub> and L<sub>B</sub> using Schrödinger 2010<sup>24</sup>. We chose L<sub>A</sub> and L<sub>B</sub> because the external prediction values of the ligands were within the top 1% of the ranked prediction value list. The important interactions of the ligands can be illustrated in comparison to the binding of ATP. The purine base of ATP is anchored in the binding pocket by hydrogen bonds to three protein residues: Glu121, Val123 and Thr183<sup>2728</sup>. In both compounds a basic aromatic ring substructure is marked as important for activity by the heat map coloring method (Figure 5). According to Schrödinger 2010, Val123 establishes an H-bond with the N1 of the pyridine of L<sub>B</sub> and the N1 of the pyrrolopyrimidine of L<sub>A</sub>. In the structure of L<sub>B</sub>, an additional H-bond connects Thr183 and the N1 of the indazole ring. In the structure of L<sub>A</sub>, Glu121 establishes an H-bond with the N7 of the pyrrolopyrimidine substructure. All H-bonds are reflected by the heat map coloring of the ligands. Additionally, the N1 of the pyrrolidine is colored as important for activity. While no interaction was detected for nitrogen N1, all active compounds of MUV548 that are based on a pyrrolopyrimidine scaffold also contain this nitrogen. Furthermore, parts of the pyrrolopyrimidine and the C5 attached methyl group of the ligand L<sub>A</sub> were marked as unimportant suggesting that the protein might be more flexible in this region. This flexibility assumption is supported by the form of the binding pocket of L<sub>C</sub> (Figure 4) which is not closed in the corresponding region. Consequently, the most important substructure, according to the heat map coloring approach, might be a basic aromatic ring substructure which is able to interact with Val123.

<sup>27</sup> Phosphotransferase and substrate binding mechanism of the cAMP-dependent protein kinase catalytic subunit from porcine heart as deduced from the 2.0 Å structure of the complex with Mn<sup>2+</sup> adenylyl imidodiphosphate and inhibitor peptide PKI(5-24)

<sup>28</sup> Staurosporine-induced conformational changes of cAMP-dependent protein kinase catalytic subunit explain inhibitory potential

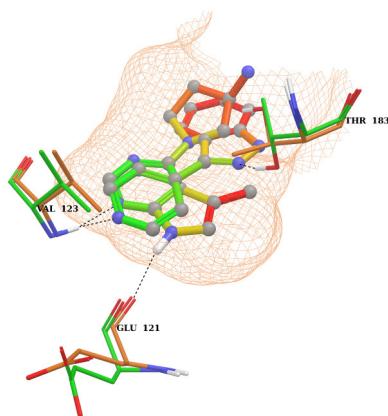


Figure 10.5: Figure 5. Aligned binding pockets of  $L_A$  and  $L_B$

**:sub:<sup>A</sup>** :sub:<sup>B</sup>Aligned binding pockets of **L**. The binding pockets of  $L_A$  and  $L_B$  were aligned and the ligands were colored with the single molecule normalization. The color gradient ranges from green (important for activity) to red (unimportant for activity or even decreasing). The green protein residues belong to  $L_A$  and the orange ones to  $L_B$ . The binding pocket is indicated as an exclusion surface. H-bonds detected by Schrödinger are indicated by a dashed line. Two similar basic aromatic rings located deep in the binding pocket are identified as important for activity.

#### 10.5.4 Visualization of MUV846 factor Xla data set

As with protein kinase A (MUV548), a plethora of crystal structures with small compound ligands are available for factor XIa. We tested our approach on the 6 ligands of the PDB entries C on MUV548. All compounds, except the ligand of PDB entry

PDB entry  $IC_{50}$  of 1.3 l<sub>nonascii\_31|M</sub><sup>29</sup>. Clavatadine A (Figure 6) is cleaved by a nucleophilic serine at the carbamate bond leaving only the carbamate side chain in the protein. Although the external prediction value of Clavatadine A is not convincing (position 3127 in the ranked list), the single molecule normalized heat map coloring of Clavatadine A identifies the carbamate bond as an important substructure for activity. A closer look at the active structures of MUV846 reveals that several active compounds also contain a carbamate bond and thus might exhibit the same binding mode. As with compound C<sub>A</sub> of the Kazius data set the full set normalization does not yield a useful coloring for Clavatadine A. The whole compound is colored reddish, which obscures the slight differences in coloring that are visible with the single molecule normalization.

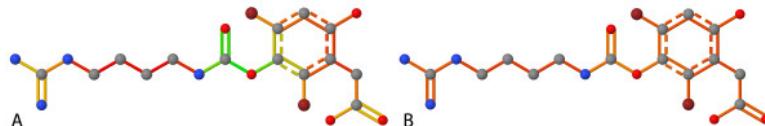


Figure 10.6: Figure 6. Clavatadine A

**Clavatadine A.** Clavatadine A colored according to a model trained on MUV846. The color gradient ranges from green (important for activity) to red (unimportant for activity). The current molecule normalization (A) and the full data set normalization (B) were both applied. The carbamate substructure is marked as important for activity.

## 10.6 Conclusions

We presented a method to visualize structure-activity and structure-property information of a linear SVM model. The heat map coloring approach assigns a color to each atom or bond of a certain molecule according to the weights of

<sup>29</sup> Clavatadine A, a natural product with selective recognition and irreversible inhibition of factor XIa

a linear SVM model. The visualization combined with linear SVMs provide an information gain compared to black box machine learning approaches like nonlinear SVMs. The method does not only provide a prediction value to label a compound as active or inactive, but also provides reasons for the labeling. Although we only tested the visualization with linear SVMs, it should in principle not be limited to linear SVMs. The visualization only requires a machine learning algorithm to assign weights to molecular fingerprinting features. The benefit of combining the visualization with linear SVMs is their promising performance and fast computation time on large-scale data sets.

We introduced two different normalization schemes. The experiments revealed advantages and disadvantages of both normalizations. However, the single molecule normalization in combination with the prediction value might be the most valuable representation for the visualization of a compound.

We evaluated our approach on a toxicity data set, a chromosome aberration data set, and the MUV data sets. Overall, the experiments show that our method sensibly visualizes structure-property and structure-activity relationships of a linear SVM model. Thus, we conclude that our method can help to guide the modification of a compound in later stages of drug discovery.

On the Kazius data set, our method allowed for identification of the toxicophores of two example compounds and therefore might help in lead optimization to obtain a less toxic compound.

The results on the MUV data sets demonstrate that our method is able to determine the correct orientation of a compound in the binding pocket. Additionally, the heat map coloring allows for the identification of important substructures for ligand protein interactions or binding mechanisms without having protein structure information. Yet, it is impossible to elucidate the exact binding mechanism or interactions without structure based approaches. Thus, the heat map coloring should be considered as complementary to structure based approaches and as such help to get a better understanding of the binding mode of an inhibitor.

The approach is not suited for identifying important side groups of a common scaffold. This deficit is mainly caused by the diversity of the employed large-scale data sets. To allow the machine learning algorithm to focus on side groups, it is necessary to employ data sets in which all compounds share a common scaffold. However, those data sets are not in the scope of a classifier, but require regression techniques.

A focus in future studies might be the combination of heat map coloring with linear support vector regression in order to elucidate the contribution of side groups to activity or selectivity.

## 10.7 Availability

All employed programs are available free of charge as executable jar and source code at <http://www.ra.cs.uni-tuebingen.de/software/ChemHeatmap/>. This includes the employed ECFP fingerprints, a modified Java version of LIBLINEAR and a graphical user interface to perform a heat map coloring of a compound. A short tutorial showing the workflow to obtain the colorings of the two compounds of the Kazius data set is also available.

## 10.8 Competing interests

The authors declare that they have no competing interests.

## 10.9 Authors' contributions

LR wrote the manuscript and implemented most of the code. GH and AJ assisted in the design and implementation of the fingerprinting algorithm. AZ supervised the study and participated in the discussion of the results. All authors read and approved the final manuscript.

# MULTILEVEL PARALLELIZATION OF AUTODOCK 4.2

## 11.1 Abstract

### 11.1.1 Background

Virtual (computational) screening is an increasingly important tool for drug discovery. AutoDock is a popular open-source application for performing molecular docking, the prediction of ligand-receptor interactions. AutoDock is a serial application, though several previous efforts have parallelized various aspects of the program. In this paper, we report on a multi-level parallelization of AutoDock 4.2 (mpAD4).

### 11.1.2 Results

Using MPI and OpenMP, AutoDock 4.2 was parallelized for use on MPI-enabled systems and to multithread the execution of individual docking jobs. In addition, code was implemented to reduce input/output (I/O) traffic by reusing grid maps at each node from docking to docking. Performance of mpAD4 was examined on two multiprocessor computers.

### 11.1.3 Conclusions

Using MPI with OpenMP multithreading, mpAD4 scales with near linearity on the multiprocessor systems tested. In situations where I/O is limiting, reuse of grid maps reduces both system I/O and overall screening time. Multithreading of AutoDock's Lamarkian Genetic Algorithm with OpenMP increases the speed of execution of individual docking jobs, and when combined with MPI parallelization can significantly reduce the execution time of virtual screens. This work is significant in that mpAD4 speeds the execution of certain molecular docking workloads and allows the user to optimize the degree of system-level (MPI) and node-level (OpenMP) parallelization to best fit both workloads and computational resources.

## 11.2 Background

Virtual screening, the use of computers to predict the binding of libraries of small molecules to known target structures, is an increasingly important component of the drug discovery process<sup>12</sup>. Although high-throughput biochemical

---

<sup>1</sup> Virtual screening of chemical libraries

<sup>2</sup> Virtual screening - what does it give us?

screening is still the predominant technique for lead compound discovery, the success of *in silico* screening in identifying drug leads has led to the growing use of virtual screening as a complement to traditional empirical methods<sup>34</sup>. There are a large number of software packages for conducting the molecular docking simulations used in virtual screening, with the open-source packages AutoDock and DOCK, and the commercial packages GOLD, FlexX and ICM, among the most popular<sup>5</sup>. Of those five packages the most widely cited is AutoDock, which has been successfully used in a number of virtual screens and in the development of the HIV integrase inhibitor raltegravir<sup>567</sup>. This work is focused on AutoDock's most recent major version, AutoDock 4.2<sup>8</sup>.

In its current iteration, AutoDock 4.2's (AD4) default search function is a Lamarkian Genetic Algorithm (LGA), a hybrid genetic algorithm with local optimization that uses a parameterized free-energy scoring function to estimate binding energy<sup>89</sup>. To perform a ligand-receptor docking experiment, the software accepts as inputs ligand and macromolecule coordinates, and then utilizes the LGA to generate ligand positions and minimize binding energies using precalculated pairwise potential grid maps<sup>10</sup>. Each *docking* is comprised of multiple independent executions of the LGA, limited to a user specified number of energy evaluations (ga\_evals) or generations (ga\_num\_generations). The individual LGA executions (ga\_runs) are clustered and ranked to generate the final docking result.

While AD4 has been widely used for virtual screening, one limitation to its usefulness is its docking speed<sup>1112</sup>. A potential way to increase AD4 performance is to parallelize aspects of its execution. Trends in processor architecture (multicore and multithreaded), and the increasing importance of highly parallel hardware such as graphics cards in scientific computation, underscore the importance of optimizing applications for parallel workloads. AD4 is a serial application not originally designed for computational clusters or to take advantage of parallel processing. There have been several previous efforts to parallelize aspects of AD4 and enable its use on high performance clusters, including: DOVIS and DOVIS 2.0 (Linux/UNIX clusters), Dockres (Linux/UNIX clusters), VS Docker (Windows clusters), and recently Autodock4.lga.MPI (an MPI implementation of Autodock4)<sup>1314151617</sup>. In general, these programs either encapsulate AutoDock in code wrappers or supply scripts that automate aspects of the preparation, distribution, execution and load balancing of AutoDock on clusters. DOVIS 2.0 uses multithreading or SSH for cluster execution, while VS Docker utilizes MPICH2 or MSMPI for cluster communication<sup>1416</sup>. Dockres runs in conjunction with several different cluster queuing systems, as does DOVIS 2.0<sup>1516</sup>. One challenge in parallelizing AutoDock for a cluster environment is that the program can generate significant network I/O during the loading of grid maps at the beginning of each docking, and when writing log files as dockings finish. Though log file writing can not easily be avoided, reuse of grid maps is a possibility as the majority of grid maps will be the same in each docking. One potential solution, if sufficient RAM is available, is to keep the grid maps in memory. This approach was used in both DOVIS and Autodock4.lga.MPI (with maps repackaged into an efficient binary format), with significant decreases in I/O observed when grid maps are loaded only once for each node<sup>1317</sup>.

In addition to optimizing AutoDock's execution on clusters, several previous efforts parallelized individual dockings. In a standard docking, the most time intensive task is the repeated execution of AutoDock's LGA, which is run tens or hundreds of times with identical structure files, grid maps and parameters. The LGA was the focus of parallelization efforts by Thormann and Pons, who parallelized the LGA of AutoDock 3.0 using OpenMP, and Khodade *et al.*, who parallelized AutoDock 3.0 and a beta version of AutoDock 4.0 using MPI<sup>1819</sup>. These approaches both resulted in a

<sup>3</sup> Virtual screening strategies in drug discovery

<sup>4</sup> Docking and chemoinformatic screens for new ligands and targets

<sup>5</sup> Protein-ligand docking: current status and future challenges

<sup>6</sup> Discovery of a novel binding trench in HIV integrase

<sup>7</sup> Virtual screening with AutoDock: theory and practice

<sup>8</sup> AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility

<sup>9</sup> Automated docking of flexible ligands: applications of AutoDock

<sup>10</sup> Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function

<sup>11</sup> Critical assessment of the automated AutoDock as a new docking tool for virtual screening

<sup>12</sup> Virtual screening for HIV protease inhibitors: a comparison of AutoDock 4 and Vina

<sup>13</sup> DOVIS: an implementation for high-throughput virtual screening using AutoDock

<sup>14</sup> VS Docker: a tool for parallel high-throughput virtual screening using AutoDock on Windows-based computer clusters

<sup>15</sup> Dockres: a computer program that analyzes the output of virtual screening of small molecules

<sup>16</sup> DOVIS 2.0: an efficient and easy to use parallel virtual screening tool based on AutoDock 4.0

<sup>17</sup> Task-parallel message passing interface implementation of Autodock4 for docking of very large databases of compounds using high-performance super-computers

<sup>18</sup> Massive docking of flexible ligands using environmental niches in parallelized genetic algorithms

<sup>19</sup> Parallel implementation of AutoDock

significant increase in AD4 execution speed, with Thormann and Pons reporting an approximately  $95\% \times N^*$  (where  $N = 8$ ) speedup, and Khodade \*et al. observing near linear speed increases on a 96-core POWER5 system<sup>1819</sup>.

Extending on these previous approaches, we had three goals for parallelization of AD4: 1) enable parallel execution of AD4 across multiple HPC architectures, 2) reduce I/O, and 3) parallelize the execution of individual docking jobs. Accordingly, we parallelized AD4 at multiple levels by: 1) utilizing MPI to distribute AD4 docking jobs across a system, 2) developing a grid map reuse scheme (conceptually similar to that implemented in DOVIS) to reduce I/O, and 3) implementing OpenMP parallelization of the LGA to achieve node-level parallelization. This standards-based parallelization scheme is significant in that it results in a highly portable parallel implementation of AD4 with user customizability in the balance between system-level and node-level parallel execution.

## 11.3 Implementation

AutoDock 4.2 (AD4) was parallelized at multiple levels using the MPICH2 implementation of the MPI standard and OpenMP application programming interface, resulting in the parallel code mpAutoDock 4.2 (mpAD4). The implementation of MPI and OpenMP in mpAD4 is standards compliant and portable to any architecture with a suitable compiler. MPI was used to parallelize the main() function of AD4 to facilitate virtual screening on MPI-enabled clusters, while OpenMP was used to implement multi-threading of the AD4 LGA. Scaling of the mpAD4 code in multithreaded and serial operation was evaluated using an IBM BlueGene/P system and a 32-core IBM POWER7 server.

### 11.3.1 MPI Parallelization

To facilitate system-level parallelization, the mpAD4 main() function was rewritten as a function call from the MPI driver. In this context, mpAD4 is executed within a master-slave scheme in which node-0 is the master node and all other nodes are slave nodes. The master node coordinates all docking activities by reading a list of docking directories from an ASCII file and then assigns individual dockings to specific slave nodes via MPI\_Send(). Once the docking assignment has been received via MPI\_Recv(), the slave nodes perform the docking work by loading necessary files, calling the mpAD4 main() function to dock the ligand that the master node has assigned to it, and writing the docking log file. To allow the user to monitor progress, the master writes three log files to track submitted dockings (MPI\_Send() call from the master), successful dockings (MPI Send() call from a slave with data indicating docking success received by the master via MPI\_Recv()) and failed dockings (MPI Send() call from a slave with data indicating docking failure received by the master via MPI\_Recv()).

### 11.3.2 I/O Optimization

AD4 requires the precalculation of one electrostatic map, one desolvation map, and individual atomic affinity grid maps for each AD4 atom type found in the ligand(s). The default AD4 behavior is to load all grid maps required for a specific docking into memory from the file system and to release that memory at the end of the docking. Thus, when the next docking begins many of the same grid maps are reloaded. In addition to the time required to load the grid maps, this behavior generates significant I/O that is unnecessary given that different dockings utilize the same electrostatic and desolvation maps, and often atomic affinity maps. Therefore, a parameter has been added to the mpAD4 executable to control grid maps persistence from one docking to another on the slave nodes. With mpAD4, as the main() function begins execution on the slave the default behavior is to load all grid maps required to dock the first ligand into compute node memory. Any remaining atomic affinity maps are loaded for subsequent dockings at the node only when required by a ligand with a previously unused atom type. Once loaded, a map persists in node memory until program termination (Figure 1). To accomplish this, the scope of the multi-dimensional array holding the grid map data changed from local (in the main() function) to global, allowing the grid map data to persist from docking to docking on a slave node. In addition, the code that manages and references this grid map array was modified to initially load only atomic affinity maps required for the first docking, and then subsequently load appropriate atomic affinity maps when required. This approach minimizes startup I/O by loading the smallest possible amount of initial

data onto the compute nodes. The user can specify grid map persistence at runtime using the flags (reload\_maps or reuse\_maps). Except where otherwise indicated, benchmarks were run with grid map reuse (gm = reuse).

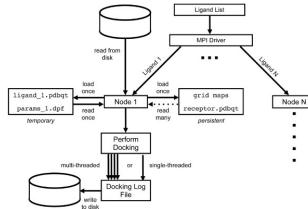


Figure 11.1: Figure 1. Workflow of mpAutoDock

**Workflow of mpAutoDock.** mpAD4 uses an MPI-driver to distribute work to individual nodes. With grid map reuse enabled, the precomputed grid maps, receptor and docking parameter files are loaded at the node level and reused for each additional docking. An individual docking on each node can be parallelized by running multiple instances of Lamarkian Genetic Algorithm in parallel using OpenMP threads.

### 11.3.3 OpenMP Parallelization

The majority of an AD4 docking is spent within the search and scoring routines, making them appealing targets for parallelization. AD4 includes several search functions, including simulated annealing (SA), genetic algorithm (GA), local search (LS) and a hybrid GA/LS (LGA). The LGA was chosen for parallelization as it was previously demonstrated to outperform either the SA or GA alone, and the LS is useful primarily for minimizing already docked structures<sup>1018</sup>. To parallelize the LGA with OpenMP, modifications to the input seed value generation and docking output handling code were required. The AD4 random number generator (RNG) utilizes a deterministic IGNLGI algorithm to generate a time-based random number seed for each LGA run. Thus, when OpenMP threads were created simultaneously with an unmodified RNG, each thread would receive an identical seed value. Therefore, the mpAD4 RNG was changed to include thread ID in the time-based seed passed to the RNG to generate unique seeds for each thread. The other code change required was related to how log information about each iteration of the LGA is written to the docking log. In the AD4, the LGA writes information to the docking log piecewise for each iteration. When the code was multithreaded, log information appeared scrambled as different threads simultaneously wrote LGA outputs. To resolve this issue in mpAD4, LGA outputs are buffered and then written en bloc after thread completion, thereby keeping the output of each ga\_run contiguous within the docking log.

### 11.3.4 Performance Profiling

In addition to the parallelization code, performance profiling has been added to mpAD4. Profiling can be turned on or off at compile time with a compiler directive. When profiling is enabled, a .csv file is updated as each docking finishes, so a user can monitor the progress of individual docking jobs and be made aware of any performance issues while the program is running. The profiling records calculation and communication start/stop times and durations from the moment the master sends the MPI message to the slave to the moment the master receives the return message from the slave with the docking status, and writes the values to a single comma delimited entry in the profiling log. Profiling outputs may be of interest to users of mpAD4 for characterizing performance bottlenecks on their system and for future developers of mpAD4. When not otherwise indicated, benchmarks were run with profiling enabled.

### 11.3.5 Blue Gene/P and POWER7 architectures

In this study two architectures were used to test mpAD4 performance, an IBM Blue Gene/P (BG/P) system and a shared-memory 32-core POWER7 p755 server<sup>2021</sup>. The BG/P system is composed of dense racks of IBM PowerPC

<sup>20</sup> IBM Power 750 and 755 Technical Overview and Introduction

<sup>21</sup> IBM System Blue Gene Solution: Blue Gene/P Application Development

450 processors running at 850 MHz with 4 cores and 4 GB RAM per compute node connected by a high performance interconnect to a storage array running the General Parallel File System (GPFS). BG/P can be configured in several different modes, including symmetric multi-processing (SMP) and virtual node (VN)<sup>21</sup>. In SMP mode, each compute node executes a single task with a maximum of four threads, with node resources including memory and network bandwidth shared by all processes. In VN mode, four single-threaded tasks are run on each node, one task per core, with each task having access to 1/4 of the total node memory. Thus, in comparing VN and SMP mode, VN mode will run four times the number of simultaneous independent MPI tasks as SMP mode, but the same number of total CPU cores will be utilized in each mode. The SMP and VN modes were used to examine differences in mpAD4 scaling and performance using MPI with multithreading SMP(OMP = 4) or MPI alone VN(OMP = 1). BG/P compute nodes do not have local disk storage, and I/O requests to the storage array are handled by dedicated I/O nodes that communicate with the network file system. Compute nodes connect to I/O nodes via a high-bandwidth “global collective network” that moves process and application data to and from the I/O nodes<sup>21</sup>. Each compute and I/O Node has three bidirectional links to the global collective network at 850 MBps per link, for a total of 5.1 GBps bandwidth per node. I/O nodes, in turn, are connected to the external file filesystem by a 10 Gb ethernet link. A BG/P system can be configured to run with a variable number of I/O nodes to model I/O replete or I/O poor systems.

In this study we tested two configurations, I/O poor (1 I/O node per 512 CPU cores) and I/O replete (1 I/O node per 64 CPU cores). When not otherwise indicated, an I/O replete configuration was used. The p755 system is a POWER7 3.3 GHz server with 32 cores and 128 GB of RAM, running the AIX 6.1 operating system. Multiparallel AD4 was compiled for BG/P with the XL C++ thread-safe cross-compiler v9.0 (bgxlC\_r) and for POWER7 using the AIX XL C++ thread-safe v11.1 (xlc\_r). For both POWER7 and BG/P, compiler optimization flag -O3 was used and the -qsmp = omp OpenMP option was specified, unless otherwise indicated. For POWER7 the -q64 flag was also used.

### 11.3.6 Ligand Libraries and Parameters

The receptor-ligand complex 1HPV (indinavir and HIV protease), and subsets of a diverse set of 34,841 compounds from the ZINC8 drug-like subset, were used to evaluate mpAD4 performance<sup>22</sup>. For the 1HPV, AutoDockTools (from MGLTools) was used to prepare the receptor and ligand<sup>23</sup>. Polar hydrogen atoms were added to the ligand and receptor .pdb files, and Gasteiger charges assigned. Indinavir libraries were then created with 4,000 copies (4 k indinavir), 8,000 copies (8 k indinavir), and 32,000 copies (32 k indinavir). The ZINC8 library ligands were prepared using the python scripts included in MGLTools package. To generate the 34,841 compound ZINC library (34 k ZINC), the 70% diversity subset of the *drug-like* subset was downloaded and compounds that failed any preparation step were discarded. A 9,000 compound subset of this library (9 k ZINC) was generated from the first 9,000 members of the 34 k ZINC library. To generate grid maps, grid box centers were defined as the center of the bound indinavir (1HPV), extending 60 grid points (0.375 Å per point) on each side. Unless otherwise specified, LGA runs were set at 20 (ga\_runs), with population size (ga\_popszie) of 150, energy evaluations (ga\_num\_evals) 250,000 and maximum number of generations (ga\_num\_generations) 27,000. All other parameter values were default for AutoDock 4.2. Except where indicated, the reuse\_maps (gm = reuse) option was used in all benchmarks.

## 11.4 Results and Discussion

To assess the performance characteristics of the hybrid parallelization and grid map reuse code, the 32 k indinavir library was docked on a 2,048(8,192) node(core) Blue Gene/P system with intermediate I/O settings (1 I/O node per 128 cores). Figure 2a shows the relative docking time in 4 different modes: VN(OMP = 1, gm = reload), VN(OMP = 1, gm = reuse), SMP(OMP = 4, gm = reload) and SMP(OMP = 4, gm = reuse). Grid map reuse reduced single-threaded execution time by approximately 17.5% due to reductions in I/O (Figure 2a). Multithreaded execution in SMP mode further reduced docking time by 10%, for an overall improvement of 25% over VN(OMP = 1, gm = reload) (Figure 2a). The improvement in docking speed observed with multithreading was due to system I/O bottlenecks experienced with single-threaded execution, as in VN mode each compute node CPU core receives an independent MPI task (8,192

<sup>22</sup> ZINC-a free database of commercially available compounds for virtual screening

<sup>23</sup> Python: a programming language for software integration and development

in this instance), while in SMP mode only physical compute nodes (2,048 in this instance) receive a task, so that one fourth as many tasks run concurrently.

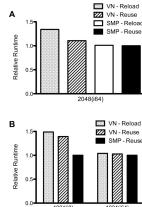


Figure 11.2: Figure 2. Impact of grid map reuse, OpenMP multithreading, and I/O on mpAD4 execution speed

**Impact of grid map reuse, OpenMP multithreading, and I/O on mpAD4 execution speed.** (A) A 32,000 copy indinavir library was docked with mpAD4 on a 2,048(8,192) node/core BG/P system in VN node mode (MPI + OpenMP with 1 OpenMP thread/node, 4x virtual nodes) with either grid map reloading or reuse and SMP mode (MPI + OpenMP with 4 OpenMP threads/node) with grid map reloading or reuse. The system was configured with 64 I/O nodes (1 I/O node per 128 cores).

Runtimes were normalized to the fastest case (SMP, gm = reuse). (B) An 8,000 copy indinavir library was docked on a 1,024(4,096) node/core BG/P system in VN mode (gm = reload or gm = reuse) and SMP mode (gm = reuse) configured with either 8 I/O nodes (I/O poor) or 64 I/O nodes (I/O replete). Runtimes were normalized to the fastest case (SMP, gm = reuse, I/O = 64).

We examined the impact of system I/O on docking using I/O times recorded by the profiling code. The major sources of I/O when running mpAD4 were the loading of grid maps and the writing of log files, while MPI communication was a negligible percentage of network traffic. I/O saturation was most apparent at the end of the initial wave of docking jobs when multiple log files are simultaneously written to disk and grid maps for the next docking jobs are loaded. In our testing with the 32 k indinavir library, instituting grid map reuse diminished VN mode average file loading time by 77%, while SMP mode (OMP = 4, gm = reuse) file loading time was 1% that of VN mode (OMP = 1, gm = reload) (Table 1). Similarly, average log file writing time was reduced by 92% with grid map reuse, and over 99% in SMP mode (Table 1). Though grid map reuse significantly reduced I/O times in our tests, the impact on overall docking times was variable.

### 11.4.1 I/O and Performance

To further examine the contribution of I/O to the performance we observed, we docked an 8 k indinavir library on 1,024(4,096) node/core BG/P system configured to be I/O poor (8 I/O nodes) or I/O replete (64 I/O nodes). In the I/O poor setting, VN mode with grid map reuse resulted in only a small increase in execution speed over VN(OMP = 1, gm = reload), while SMP mode (OMP = 4, gm = reuse) execution time was decreased by 33% (Figure 2b). Such differences were largely unapparent in an I/O replete configuration, where grid map reuse showed no benefit in overall docking speed, and SMP-mode gains were only 3% (Figure 2b). When I/O was sufficiently limited (e.g., 1,024(i8)), grid map reuse had limited impact on overall performance, likely because the I/O generated by the slave nodes writing log files was still sufficient to saturate the I/O poor system. Similarly, grid map reuse did not greatly improve performance in an I/O replete setting, as sufficient I/O capacity was available for simultaneous grid map loading and log file writing. Thus, grid map reuse greatly reduces I/O activity and can significantly improve docking performance in some settings, allowing larger systems to be effectively used for a given I/O capacity. Similarly, 4-way OpenMP multithreading reduces I/O by 75% for a given system size and I/O times by 90%, again allowing larger systems to be employed than with MPI alone.

### 11.4.2 Hybrid Scalability

To test mpAD4 scalability, small molecule libraries including a 34 k ZINC, 9 k ZINC and 4 k indinavir were run on 512(2,048), 1,024(4,096), 2,048(8,192), and 4,096(16,384) node/core BG/P systems. Figure 3a shows the speedup observed with the 34 k ZINC library in SMP and VN modes (gm = reuse). SMP mode scaling was nearly linear at

approximately 92% ideal speed on the 16,384 core system. VN mode deviated to a greater degree at 72% ideal on the 16,384 core system, a 22% decrease from SMP-mode performance (Figure 3a). Interestingly, this performance decrease was not due to I/O differences, but instead reflected improved system utilization efficiency in the multithreaded execution mode. For both SMP and VN mode, deviations from ideal occur on larger systems as a virtual screen comes to the end and fewer ligands remain to be docked than capacity of the system, resulting in portions of the system remaining idle while the remaining active jobs finish. Multithreaded execution helps to alleviate this inefficiency in two ways: 1) when using multithreading there are fewer MPI nodes in the system and a virtual screen proceeds closer to completion before nodes become idle, and 2) individual dockings are executed more quickly when multithreaded, reducing time spent with idle nodes. For very large screening libraries, node utilization efficiency differences at the end of screening are unlikely to contribute to significant difference overall docking time. However, the opposite is true as library size shrinks in comparison to system size, as demonstrated in the scaling of the 9 K ZINC library where node utilization inefficiencies are apparent in both VN and SMP mode, though SMP mode is less effected (Figure 3b).

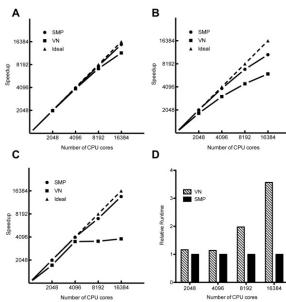


Figure 11.3: Figure 3. Scaling of MPI alone versus MPI with OpenMP multithreading

**Scaling of MPI alone versus MPI with OpenMP multithreading.** Virtual compound libraries were docked in VN and SMP modes on BG/P systems of 512(2,048), 1,024(4,096), 2,048(8,192) and 4,096(16,384) nodes(cores). (A) A 34,841 compound ZINC library was docked on variable BG/P systems sizes. Ideal speedup was calculated from the fastest 512 node result (SMP), and the relative speedup SMP and VN mode results were plotted. (B) A 9,000 compound subset ZINC library was docked on variable BG/P systems sizes, and ideal speedup was calculated from the fastest 512 node result (SMP). (C) A 4,000 copy indinavir library was docked on variable BG/P systems sizes, and an ideal speedup line was calculated from the fastest 512 node result (SMP). (D) Relative execution times for the 4,000 copy indinavir library were calculated for SMP and VN modes at each system size.

In addition to multithreading, node utilization can be improved by pre-ordering ligands to be docked by complexity (descending number of torsional angles). For a sorted 9 k indinavir library docked in VN mode on a 2,048 core system, sorting improved docking speed by 10%, though it was still 9% slower than SMP mode (data not shown). In cases where the availability of CPUs greatly exceeds the number of molecules to be screened, multithreading is particularly useful for increasing the usefully employable system size. For example, a 4 k indinavir library in single-threaded execution (VN, OMP = 1) is unable to take advantage of more than 4,000 cores (Figure 3c). In contrast, multithreading (SMP, OMP = 4) allows up to 16,000 cores to be employed, decreasing the docking time by over 70% (Figure 3c and Figure 3d). For larger systems, combining OpenMP multithreading with MPI allows for more efficient utilization of system resources at the end of screens. For smaller screens, multithreading has a clear advantages over serial execution when the number of available cores exceeds the number of ligand-receptor complexes to be docked.

### 11.4.3 OpenMP Scalability

To test the scalability of the OpenMP implementation, a 1HPV complex was docked with mpAD4 on POWER7 p755 32-core server. Docking parameters were modified as follows: ga\_runs = 128, ga\_num\_evals = 5,000,000. Table 2 shows apparent speedups and percent of ideal runtimes for a single docking of the 1HPV complex run as *serial* mpAD4 code (compiled without OpenMP) or utilizing from 1 to 32 OpenMP threads. On the POWER7 system, single-threaded OpenMP incurred a 12% overhead versus serial code. Overhead increased with thread number; with an apparent speedup of 1.5 $\times$  for 2 threads to 22.3 $\times$  for 32 threads (Table 2). Our testing on BG/P showed an overhead

of approximately 10% for either OMP = 1 or OMP = 4, with little or no additional cost for 4 threads over 1 (data not shown). Due to the overhead incurred with OMP = 1 vs. serial, a user intending to use mpAD4 only in single-threaded applications may benefit from compiling mpAD4 without OpenMP. We anticipate the specific OpenMP overhead will vary with both system characteristics and compiler options. Though here we have only demonstrated multithreading up to 32 cores, the code is currently implemented to allow up to 128 simultaneous threads, which we expect will allow further improvements total in docking speed.

#### 11.4.4 Output Comparison

The binding modes generated by single or multithreaded execution of mpAD4 were determined for a set of 76 crystallographically determined ligand-protein complexes using a BG/P system size of 512 cores in SMP(OMP = 4) and VN(OMP = 1 or serial) modes<sup>24</sup>. For each docked complex, pairwise RMSDs were calculated for the overall lowest energy ligand and lowest energy member of largest ligand cluster. When the lowest energy ligand was not also a member of the largest ligand cluster, the lesser pairwise RMSD value was used. The RMSDs from VN(serial) to VN(OMP = 1) or SMP(OMP = 4), and between VN(OMP = 1) and SMP(OMP = 4) were calculated (Table 3). The mean RMSD values in all three comparisons were less than 1.0, with median values less than 0.2 (Table 3). We therefore consider the outputs to be substantially similar.

#### 11.4.5 Performance Expectations

The implementation of MPI and OpenMP in mpAD4 is portable to systems with a suitable compiler and the required libraries. In the case of distributed-memory architectures using either Intel or AMD x86 microprocessors, we expect similar trends in terms of performance. Environmental factors that may have a large impact on performance are network bandwidth, compute node microprocessor speed, memory and the availability of node local disk storage (potentially ameliorating I/O issues associated with writing log files). Multiparallel AD4 generates little MPI communication, and we therefore anticipate that it will scale well even on clusters with limited I/O bandwidth if they possess node local disk storage and sufficient RAM to store grid maps in memory. Similarly, we would predict that the OpenMP multithreading will generate performance gains on any modern multicore microprocessor, though overhead and absolute scalability may vary with compilers, compiler options and microprocessor architecture.

### 11.5 Conclusions

We have parallelized AutoDock 4.2 using MPI and OpenMP to create mpAD4, a standards compliant and portable parallel implementation of AutoDock, with user customizability in the balance between serial and parallel execution, a capability to reuse grid maps, and extensive profiling features for performance monitoring. In our tests, grid maps reuse drastically reduced system I/O, allowing for nearly linear scaling of mpAD4 on system sizes of up to 16,384 CPU cores. OpenMP multithreading scaled up to 32 threads, resulting in a maximum speedup of 22× over single-threaded execution. We propose three potential use cases for mpAD4: 1) combining MPI and OpenMP parallelization on large systems to balance system-level and node-level parallelization to manage I/O and achieve the best possible throughput, 2) enabling larger systems to be used for screening small libraries, and to improve system utilization at all library sizes, 3) facilitating the rapid docking of one or a small number of ligand-receptor complexes on shared memory systems.

### 11.6 Availability and Requirements

Project name: mpAutoDock 4.2

Project home page: <http://autodock.scripps.edu/downloads/multilevel-parallel-autodock4.2>

---

<sup>24</sup> Development and validation of a modular, extensible docking program: DOCK 5

Operating system(s): Platform independent  
Programming language: C++  
Other requirements: MPI (MPICH2), OpenMP  
License: GNU GPL v3

## 11.7 Competing interests

The authors declare that they have no competing interests.

## 11.8 Authors' contributions

APN participated in the design of this work, performed validation and benchmarking of the parallel code, and wrote this manuscript. PKC parallelized the AutoDock code, and assisted in drafting this manuscript. JPK participated in the design of this work, and in revising this manuscript for publication. DJK assisted in the analysis of the data, and in revising this manuscript for publication. CPS conceived this study, participated in the design of this work, coordinated its execution, and helped to revise this manuscript for publication. All authors read and approved the final manuscript.

## 11.9 Acknowledgements

We thank Cindy Mestad and Steven Westerbeck at IBM Rochester, David Singer and Fred Mintzer at IBM Watson and Sharon Selzo at IBM Poughkeepsie for technical assistance, and IBM corporation for providing access to the Blue Gene/P and POWER7 systems used in this study. We acknowledge the Minnesota Supercomputing Institute for providing technical support and computational resources for this study. We are grateful to Michael Pique for thoughtful discussions and reviewing this manuscript. This work was supported by an American Heart Association Predoctoral Fellowship 09PRE2220147 (APN), NIH Predoctoral Fellowship F30DA26762 (APN), and University of Minnesota-Rochester, Bioinformatics and Computational Biology (BICB) Program Seed Grant (DJK, CPS, JPK). The distribution of the mpAD4 software is supported by NIH Grant R01 GM069832 (A. Olson, The Scripps Research Institute).



# PUBCHEM3D: SIMILAR CONFORMERS

## 12.1 Abstract

### 12.1.1 Background

PubChem is a free and open public resource for the biological activities of small molecules. With many tens of millions of both chemical structures and biological test results, PubChem is a sizeable system with an uneven degree of available information. Some chemical structures in PubChem include a great deal of biological annotation, while others have little to none. To help users, PubChem pre-computes “neighboring” relationships to relate similar chemical structures, which may have similar biological function. In this work, we introduce a “Similar Conformers” neighboring relationship to identify compounds with similar 3-D shape and similar 3-D orientation of functional groups typically used to define pharmacophore features.

### 12.1.2 Results

The first two diverse 3-D conformers of 26.1 million PubChem Compound records were compared to each other, using a shape Tanimoto (ST) of 0.8 or greater and a color Tanimoto (CT) of 0.5 or greater, yielding 8.16 billion conformer neighbor pairs and 6.62 billion compound neighbor pairs, with an average of 253 “Similar Conformers” compound neighbors per compound. Comparing the 3-D neighboring relationship to the corresponding 2-D neighboring relationship (“Similar Compounds”) for molecules such as caffeine, aspirin, and morphine, one finds unique sets of related chemical structures, providing additional significant biological annotation. The PubChem 3-D neighboring relationship is also shown to be able to group a set of non-steroidal anti-inflammatory drugs (NSAIDs), despite limited PubChem 2-D similarity.

In a study of 4,218 chemical structures of biomedical interest, consisting of many known drugs, using more diverse conformers per compound results in more 3-D compound neighbors per compound; however, the overlap of the compound neighbor lists per conformer also increasingly resemble each other, being 38% identical at three conformers and 68% at ten conformers. Perhaps surprising is that the average count of conformer neighbors per conformer increases rather slowly as a function of diverse conformers considered, with only a 70% increase for a ten times growth in conformers per compound (a 68-fold increase in the conformer pairs considered).

Neighboring 3-D conformers on the scale performed, if implemented naively, is an intractable problem using a modest sized compute cluster. Methodology developed in this work relies on a series of filters to prevent performing 3-D superposition optimization, when it can be determined that two conformers cannot possibly be a neighbor. Most filters are based on Tanimoto equation volume constraints, avoiding incompatible conformers; however, others consider preliminary superposition between conformers using reference shapes.

### 12.1.3 Conclusion

The “Similar Conformers” 3-D neighboring relationship locates similar small molecules of biological interest that may go unnoticed when using traditional 2-D chemical structure graph-based methods, making it complementary to such methodologies. The computational cost of 3-D similarity methodology on a wide scale, such as PubChem contents, is a considerable issue to overcome. Using a series of efficient filters, an effective throughput rate of more than 150,000 conformers per second per processor core was achieved, more than two orders of magnitude faster than without filtering.

## 12.2 Background

PubChem<sup>1234</sup> is a free and open public resource for the biological activities of small molecules. With more than 30 million unique chemical structures and 120 million biological test results, it is a sizeable system with an uneven degree of available information. Some chemical structures in PubChem have a great deal of biological annotation and literature associated, while many others (*e.g.*, synthesized for high-throughput screening purposes) have little to nothing known about them other than the chemical structure. To help overcome this disparity, PubChem helps users to locate or relate data in the archive by pre-computing “neighboring” relationships. One of these, known as “Similar Compounds”, associates a pair of chemical structures if they have a Tanimoto<sup>567</sup> similarity of 0.9 or greater when using the PubChem subgraph binary fingerprint<sup>8</sup> and Eq. (1).

where  $A$  and  $B$  are the respective counts of fingerprint set bits in the compound pair and  $AB$  is the count of bits in common.

The “Similar Compounds” relationship is useful to relate analogues that may have similar biological activity or function and additional biological annotation; however, “Similar Compounds” is not particularly good at finding chemical structures that can adopt similar 3-D shape and similar 3-D orientation of functional groups typically used to define pharmacophore features (henceforth, these pharmacophore feature functional groups will be referred to as “pharmacophore features” or simply as “features”), which could indicate, for example, that the molecules bind to a protein in a similar fashion. It may be useful, therefore, to provide a “Similar Conformers” relationship in PubChem to help relate relevant conformers of chemical structures.

Wanting to compute a 3-D neighboring relationship with modest computational capacity on a very large scale and actually being able to do it are two very different things. For 30 million compounds, a neighboring relationship requires a minimum of  $10^{14}$  pair-wise similarity computations. The 2-D similarity of chemical structures with binary fingerprints is relatively fast, with rates of  $10^6$  compound pair similarities per second per processor core achievable. Computing the analogous 3-D pair-wise similarity of conformers is much slower, with rates of  $10^2$  to  $10^3$  per second per processor core (depending on the degree of accuracy versus performance tradeoffs one is willing to make), when using atom-centered Gaussians<sup>9101112</sup> for the shape description. This difference in 2-D versus 3-D pair-wise similarity overlap computation rate is made yet worse by another factor of  $10^1$  (or more), when considering that 3-D methods actually need to consider multiple diverse conformers per chemical structure, since a small molecule can typically adopt multiple distinct shapes or orientations of pharmacophore features at room temperature. This puts the comparable rate of computation of 3-D chemical structure pair-wise similarity overlap at least  $10^4$  to  $10^5$

<sup>1</sup> PubChem: integrated platform of small molecules and biological activities

<sup>2</sup> PubChem: a public information system for analyzing bioactivities of small molecules

<sup>3</sup> An overview of the PubChem BioAssay resource

<sup>4</sup> Database resources of the National Center for Biotechnology Information

<sup>5</sup> Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings

<sup>6</sup> Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients

<sup>7</sup> Analysis and display of the size dependence of chemical similarity coefficients

<sup>8</sup> PubChem substructure fingerprint description

<sup>9</sup> A gaussian description of molecular shape

<sup>10</sup> ROCS - Rapid Overlay of Chemical Structures

<sup>11</sup> PAPER—accelerating parallel evaluations of ROCS

<sup>12</sup> ShapeTK-C++

slower than that for 2-D. This performance gap has led some to search for alternative approaches for determining 3-D similarity between small molecules.

In one such approach<sup>13</sup>, 3-D similarity is recast to use a binary fingerprint to achieve a conformer pair-wise similarity overlap computation speed similar to that of 2-D similarity computation. This scheme determines a set of representative 3-D reference shapes, each corresponding to a binary bit in a fingerprint. When generating the fingerprint for a 3-D chemical structure conformer, a traditional 3-D shape superposition to all reference shapes is performed. If there is sufficient similarity to a reference shape, the corresponding binary bit is set. Besides the pre-computation expense to determine the reference shapes to use and to generate the 3-D fingerprint for all conformers to be searched, this method has an important drawback. Unlike 2-D binary fingerprint methods, when two 3-D chemical structure conformers are deemed to be similar by this approach, it might not be immediately obvious as to why. The reason is simple. The common binary bit values simply identify that the two conformers share a region of shape-space, without the additional requirement that they actually share a sufficient degree of shape similarity.

An attempt<sup>14</sup> was made to improve upon the 3-D binary fingerprint approach. In effect, the method was very similar; with a predetermination of reference shapes followed by 3-D shape fingerprint computation. Yet, there were a couple of important distinctions that, in essence, allowed the method to yield conformer superposition results much like “traditional” 3-D similarity superposition methods<sup>101112</sup>. First, the alignment of the conformer to the reference shape was retained during shape fingerprint generation. Second, when a fingerprint “bit” was in common between two conformers, the retained alignments to the reference shape were used to yield an (approximate) alignment between the conformers. Dubbed “alignment recycling”, this approach recognized that conformers with similar shape align to a reference shape in a similar way. By “replaying” the alignment to common reference shapes, the best superposition between the conformer pair is the result of the similarity computation. This approach, while not as fast as the method that used only a binary fingerprint, was 10<sup>2</sup> times faster than “traditional” 3-D similarity superposition methodology. A major downside to “alignment recycling” was that it was only parameterized for relatively small and inflexible chemical structures. It means that additional work is necessary to extend this approach to larger and more flexible structures. In all, the above two 3-D fingerprint approaches showed great promise to dramatically improve the throughput of 3-D similarity computation.

To harness a 3-D fingerprint to speed 3-D similarity throughput, one must first determine the reference shapes to use. Recent efforts<sup>13</sup> to describe the shape space of biologically relevant small molecules showed exponential behavior in reference shape count resulting from changes to the minimum shape Tanimoto (ST) distance between reference shapes. However, when examining the growth of shape space per unit volume for a maximum count of reference shapes<sup>15</sup>, shape space was shown to grow gradually and smoothly as a function of ST. In addition, and generally speaking, it was shown that the shape space of a given unit volume describes 40-70% of the shape space of all chemical structures with a lesser volume. This would suggest that one could group together regions of shape space and describe it with a relatively small number of reference shapes, while avoiding the problem of having too many reference shapes. Reformulating the fingerprint definition with multiple tiers of fingerprints with different minimum ST distances between reference shapes may allow “alignment recycling” to be effective for larger and more flexible chemical structures, thus, providing a means to speed computation of a 3-D neighboring relationship on a very large scale.

In this work, we describe the multi-conformer PubChem “Similar Conformers” 3-D neighboring relationship and explain various strategies and approaches that made it a tractable problem, including extending the “alignment recycling” methodology to cover the full range of chemical structures considered in the PubChem3D project.

<sup>13</sup> Small molecule shape-fingerprints

<sup>14</sup> Fast 3D shape screening of large chemical databases through alignment-recycling

<sup>15</sup> PubChem3D: diversity of shape

## 12.3 Results and discussion

### 12.3.1 1. Description of “Similar Conformers” neighboring relationship

PubChem uses two 3-D similarity measures to determine whether two molecules are “Similar Conformers”. One of these is the shape Tanimoto (ST) for shape similarity<sup>1011121617</sup>, given by Eq. (2). The second similarity measure, defined by Eq. (3), is the color Tanimoto (CT)<sup>1012</sup>, which quantifies the 3-D shape similarity of fictitious “color” atoms, each representing the 3-D location of a particular pharmacophore feature functional group type: hydrogen-bond donor, hydrogen-bond acceptor, cation, anion, hydrophobe, or ring. The ST and CT values range between 0 (for no similarity) and 1 (for identical).

where  $V_{AA}$  and  $V_{BB}$  are the respective self-overlap volume and  $V_{AB}$  is the overlap volume of conformers A and B.

where, for each of the six independent fictitious feature atom types,  $V_{AA}$  and  $V_{BB}$  are the respective self-overlap volumes and  $V_{AB}$  is the overlap volume of conformers A and B.

Pair-wise shape and feature comparison of conformers takes two basic steps: (1) optimization of the shape superposition between two 3-D chemical structures, to find their maximum shape overlap in terms of ST, and (2) a single-point CT computation at that maximum shape overlay. PubChem 3-D “Similar Conformers” neighbors are identified as any pair-wise conformer superposition with ST and CT values of 0.8 and 0.5 (actually 0.795 and 0.495, after floating point number rounding is considered), respectively.

An important issue with 3-D neighboring is the number of conformers considered. Although PubChem generates a conformer ensemble for each molecule, consisting of up to 500 sampled conformations, it is not practical to consider all of these for 3-D neighboring. Therefore, a selection of diverse conformers for each compound is considered for the purposes of 3-D neighboring. A detailed description of how the diverse conformer set is derived can be found in the **Materials and Methods** section (See “Diverse conformer concept”).

It is important to note that 3-D neighboring using a single conformer per compound has a one-to-one correspondence between compound pairs and conformer pairs. When using multiple conformers per compound, it is possible that only a subset of possible conformer pairs per compound pair may satisfy the 3-D neighboring criteria. For clarification, a 3-D *conformer neighbor* pair is defined as any conformer pair with ST 0.8 and CT 0.5. If there is at least one *conformer neighbor* pair among all possible conformer pairs from a given compound pair, a *compound neighbor* pair results. In this work, a 3-D neighbor implies a 3-D *compound neighbor*. If further clarification is necessary, the terms 3-D compound neighbors and 3-D conformer neighbors are used.

### 12.3.2 2. The distribution of 3-D neighbors

At the time of writing, 26,153,061 PubChem Compound records (CIDs) have a “Similar Conformers” neighboring relationship using the first two diverse conformers per compound. These identified 6.62 billion unique compound neighbor pairs and 8.16 billion unique conformer neighbor pairs. The average compound neighbor count per compound, after exclusion of self-neighbor pairs, is 253. Figure 1 shows the frequency of neighbor count per compound, cumulative % CID count, and cumulative % 3-D neighbor count. Although some CIDs have more than 30,000 neighbors, 21.9 million CIDs (87.5%) have less than 1,000 neighbors, and 1.12 million CIDs (4.27%) do not even have a neighbor beyond self. This rather skewed population of the neighbor count per CID is reflected in the plot of % cumulative neighbor count versus % cumulative CID count (Figure 2). One can see that 20% of the chemical structures have more than 80% of the “Similar Conformer” neighbor pairs.

The chemical structures on the extreme end, with more than 30,000 neighbors each, have a common motif of two substituted aromatic ring systems separated by different linkers. Figure 3 depicts a single-linkage clustering of all 324 chemical structures with more than 30,000 3-D neighbors performed with the PubChem Structure Clustering tool using the PubChem 2-D dictionary-based binary fingerprint and Eq. (1) to help highlight the different chemical series represented. The most prevalent of these are based on N-phenylbenzamide (CID 7168). Neighboring reflects

<sup>16</sup> A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction

<sup>17</sup> A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape

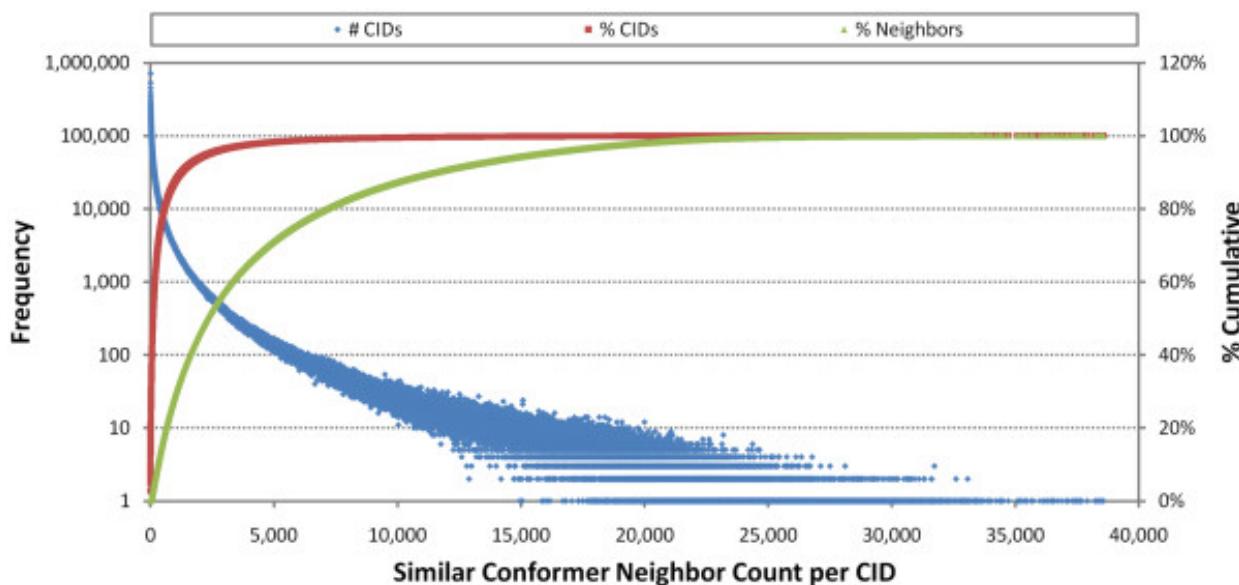


Figure 12.1: Figure 1. Count of “Similar Conformers” per compound

**Count of “Similar Conformers” per compound.** The frequency of unique 3-D compound neighbors counts per PubChem Compound record (CID) [blue diamond], percent cumulative CID count [red square], and percent cumulative 3-D neighbor count [green triangle].

the contents of PubChem. If there is a large subpopulation of chemical structures very similar to each other, those chemical structures will interrelate; however, one advantage of 3-D “Similar Conformers” neighboring is that it relates chemical structures that have similar shape and features, which can be somewhat orthogonal to a chemical series identified by 2-D “Similar Compound” neighboring (to be discussed in more detail in the next section).

Of the 1.12 million CIDs without a neighbor pair, except for self, these include a large and significant percentage of the total cases where the count of atoms or features is high, as depicted in Figure 4. The lack of 3-D neighbor means that these larger compounds lack a 3-D complement, which is not surprising given that shape space grows exponentially and PubChem3D limits consideration to chemical structures with fifty or fewer non-hydrogen atoms, making it increasingly less likely that a suitable neighbor can be found as a function of volume. Otherwise, the profile of chemical structures without neighbors is much like that for a set of 26,157,365 CIDs that represent the entire “live” PubChem3D contents as of October 2010 (designated as the *Search* set), representing a small minority of chemical structures with unique shape and feature profiles. For the first and second diverse conformers per compound, respectively, there are 1.31 million and 4.77 million cases where only the self neighbor is found. Employing a second diverse conformer allows 0.19 million additional CIDs to have a compound neighbor beyond self. The big increase in self-only neighbor pairs for the second diverse conformer, which represents the conformer most dissimilar to the first in a conformer ensemble, is notable; however, it is too early to say definitively whether these counts of no-neighbor per conformer will remain high, as more diverse conformers per compound are considered.

### 12.3.3 3. Comparison of 2-D and 3-D similarity neighbors

For a given molecule, PubChem provides a “Similar Compounds” 2-D neighboring relationship, computed using a 2-D binary fingerprint and a threshold of 0.9 Tanimoto similarity using Eq. (1). It is interesting to see how one can find related biological annotation information using the 3-D “Similar Conformers” neighboring relationship as opposed to the 2-D “Similar Compounds”. To demonstrate this, three well known molecules of biomedical interest are selected: caffeine (CID 2519), aspirin (CID 2244), and morphine (CID 5288826). The overlap of three primary types of annotation is examined. The metrics used are unique and common count of neighbors with links to: Medical

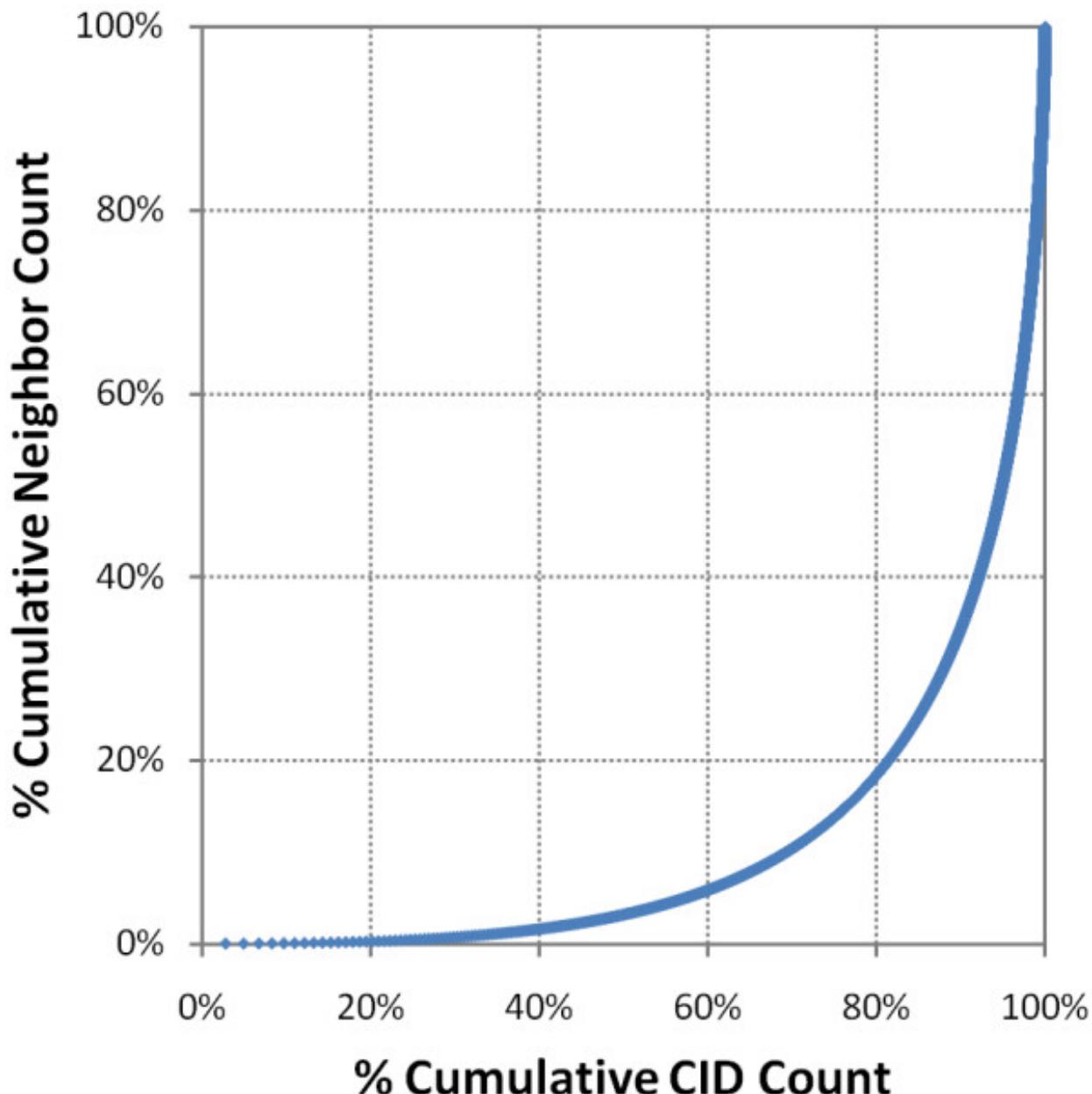


Figure 12.2: Figure 2. Most compounds have few 3-D neighbors

**Most compounds have few 3-D neighbors.** More than 80% of all CIDs have only 20% of 3-D neighbors.

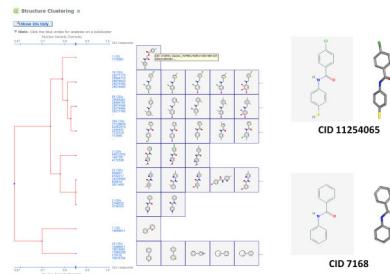
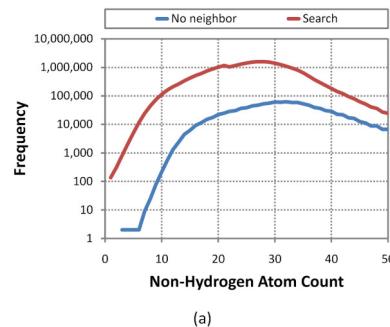
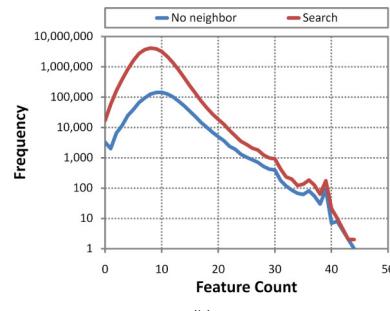


Figure 12.3: Figure 3. Compounds with the most 3-D neighbors

**Compounds with the most 3-D neighbors.** The PubChem Structure Clustering analysis of the 324 PubChem Compound records with more than 30,000 neighbors shows a common structural motif of two (aromatic) rings separated by a linker. N-phenylbenzamide (CID 7168) scaffold is present in the majority of these, with CID 11254065 having the most 3-D neighbors in all of PubChem.



(a)



(b)

Figure 12.4: Figure 4. Molecules without neighbors

**Molecules without neighbors.** Non-hydrogen atom count and feature atom count profiles for the 1.12 million CIDs without a neighbor pair (other than the self-neighbor pair) compared to those for all 26.1 million neighbored CIDs (*Search* set), showing “no neighbor” cases are found across the entire range but accounting for much of the larger count cases.

Subject Heading (MeSH)<sup>18</sup>, through which one can locate scientific literature about a similar chemical structure in PubMed<sup>19</sup>; PubChem BioAssay database<sup>3</sup>, where one can find biological and experimental data, including protein binding inhibition values; and protein 3-D structures<sup>20</sup>, representing 3-D structures of a discrete protein with a bound ligand, determined by X-ray crystallography or NMR spectroscopy. Figure 5 gives the overlaps found between 2-D and 3-D neighboring relationships. As one can see, caffeine has 1,231 2-D neighbors, but only 302 of these are in common with its 2,298 3-D neighbors. The non-overlapping parts between the 2-D and 3-D neighboring show how similar, yet unique, chemical space is located. Of the unique 3-D neighbors, they expand, beyond its 2-D counterpart, the available biomedical annotation that may be related and relevant, with an additional 23 MeSH links, 274 biological experiments, and a doubling of the protein 3-D structures to consider. A similar result is found in the case of aspirin and morphine. It appears clear that in these cases 3-D similarity complements 2-D similarity with a mostly unique set of chemical structures that help one to discover connections between small molecules that might otherwise be missed. While this near orthogonality of neighbor sets won't be true for all chemical structures, it can be helpful to locate and relate available information in a vast data system such as PubChem.

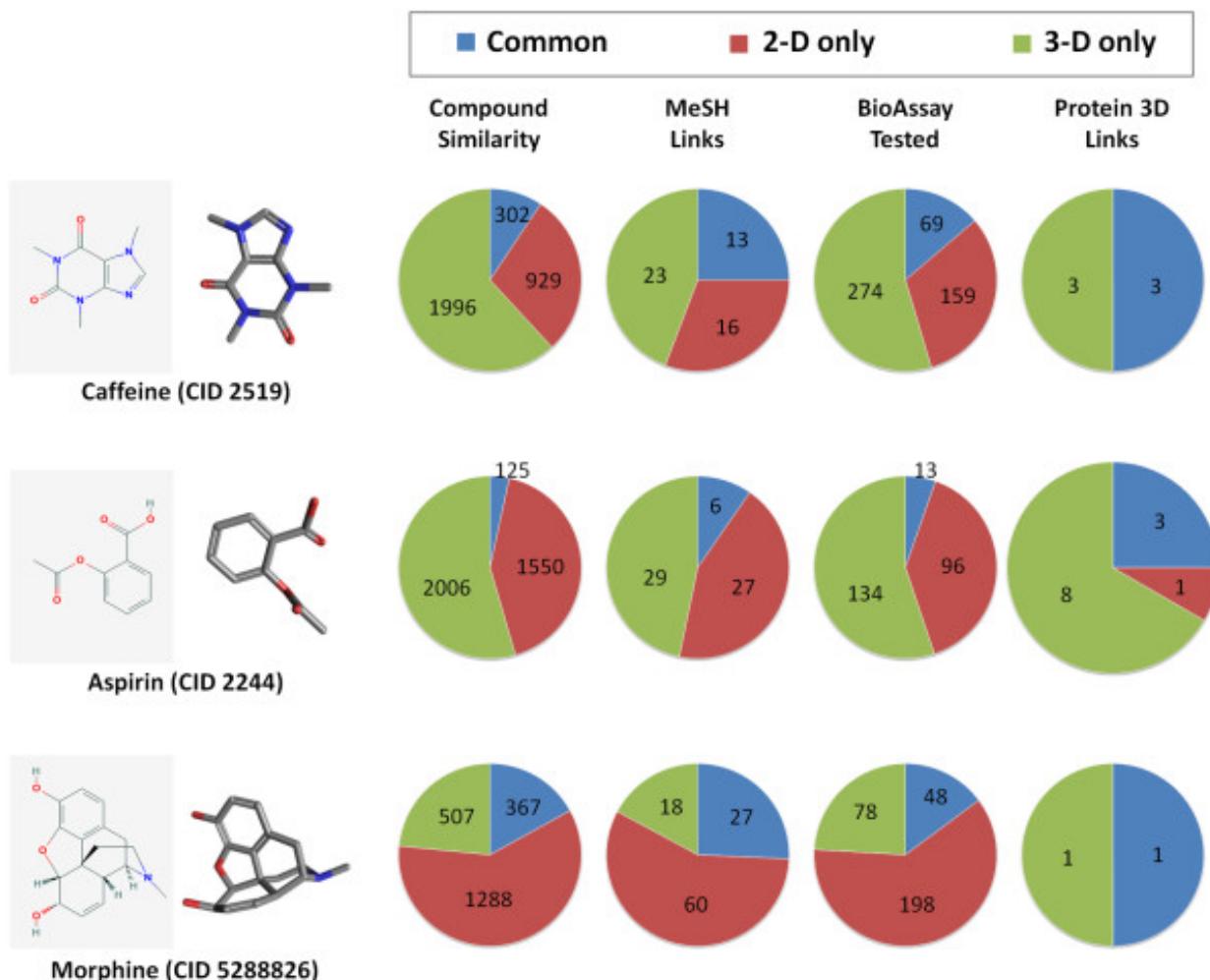


Figure 12.5: Figure 5. 2-D neighbors versus 3-D neighbors

**2-D neighbors versus 3-D neighbors.** Comparison of the 2-D “Similar Compound” and 3-D “Similar Conformer” neighboring relationships using three well known small molecules, caffeine, aspirin, and morphine, demonstrates how each neighboring relationship can help find related chemical structures with unique biological annotation.

<sup>18</sup> Medical Subject Headings<sup>19</sup> PubMed<sup>20</sup> MMDB: annotating protein sequences with Entrez’s 3D-structure database

To further emphasize how the 3-D “Similar Conformers” neighboring relationship may complement the 2-D “Similar Compounds” neighboring relationship, the 2-D and 3-D similarity scores of eight drug molecules with the same mechanism of action are compared in Figure 6, and the 3-D alignment for particular compound pairs, whose 2-D and 3-D similarity difference are relatively large, are depicted in Figure 7. All eight drugs are known inhibitors of prostaglandin synthase<sup>2122232425</sup> and were carefully selected for illustrative purposes from the PubChem Compound database via the MeSH pharmacological action of “anti-inflammatory agents, non-steroidal” (MeSH ID 68000894), also known as NSAIDs. While the 2-D similarity between drug molecules is calculated using the PubChem subgraph fingerprint<sup>8</sup>, the 3-D similarity scores represent the best ST and CT similarity values from all possible combinations of the first ten diverse conformers of each compound pair. Although all eight molecules inhibit the same target, only one molecule pair (CIDs 3332 and 3394) is identified as a 2-D neighbor, as shown in the lower triangle of the similarity score matrix. The 3-D similarity approach, however, identified 11 molecule pairs as 3-D neighbors. For example, although the 2-D similarity score between CIDs 1302 and 2581 is 0.43, there are significant 3-D shape and feature overlaps (ST = 0.92 and CT = 0.55) between them (Figure 7). If fewer conformers are used, the number of resulting 3-D “Similar Conformers” neighbor pairs will be reduced. When using 2, 3, 5, 7, and 10 diverse conformers, a total of 2, 3, 9, 11, and 11 compound pairs and 2, 3, 14, 22, and 27 conformer pairs, respectively, met the 3-D neighboring criteria for the eight drug molecules.

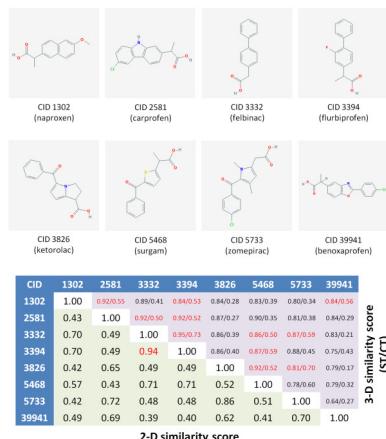


Figure 12.6: Figure 6. Similarity score matrix for selected non-steroidal anti-inflammatory drugs

**Similarity score matrix for selected non-steroidal anti-inflammatory drugs.** The lower triangle of the score matrix corresponds to the 2-D similarity scores computed using the PubChem fingerprint, and the upper triangle corresponds to the 3-D similarity ST/CT scores. The matrix elements in red indicate the 2-D “Similar Compounds” (with a 2-D score of 0.9) or 3-D “Similar Conformers” (with a 3-D score of ST 0.8 and CT 0.5). The first ten diverse conformers were used for each molecule.

While not all eight selected NSAID drug molecules are 3-D neighbors of each other, examining the 3-D neighbors of the 3-D neighbors shows that each of the eight drug molecules is related to one or more of the eight drug molecules, effectively forming a cluster of related drugs that are highly similar in terms of shape and pharmacophore features but rather dissimilar in terms of 2-D graph similarity. Actually, this “cluster” of NSAID drugs presented in Figure 6 is part of a larger 3-D cluster, with only eight of thirteen members being selected for clarity and demonstrative purposes. In addition, this is only one of several NSAID drug “clusters” that one can find using 3-D similarity. For the purposes of brevity and focus, only the drug class NSAIDs is explored, but suffice it to say that there are other examples one can find with other drug target classes that are similarly demonstrative.

<sup>21</sup> DrugBank 3.0: a comprehensive resource for ‘Omics’ research on drugs

<sup>22</sup> Zomepirac

<sup>23</sup> Pharmacologic properties of fenbufen

<sup>24</sup> Inhibition of prostaglandin activity and synthesis by fenbufen (a new nonsteroidal antiinflammatory agent) and one of its metabolites

<sup>25</sup> Pharmacology of benoxaprofen (2-[4-chlorophenyl]- $\alpha$ -methyl-5-benzoxazole acetic acid), LRCL 3794, a new compound with anti-inflammatory activity apparently unrelated to inhibition of prostaglandin synthesis

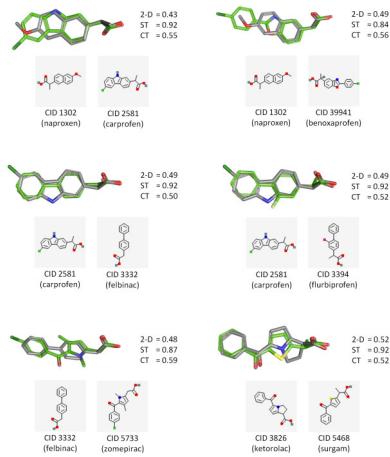


Figure 12.7: Figure 7. 3-D superposition of selected 3-D “Similar Conformers” pairs

**3-D superposition of selected 3-D “Similar Conformers” pairs.** Although there is little 2-D similarity, using the PubChem fingerprint, significant 3-D similarity are found between selected non-steroidal anti-inflammatory drugs.

If a molecule has known bioactivity, there is a reasonable expectation<sup>2627</sup> that its similarity neighbors may also be similarly bioactive. As demonstrated in Figure 6 and 7, the 3-D “Similar Conformers” relationship can be useful to identify structurally similar molecules that may be completely missed when only the 2-D “Similar Compounds” relationship is exploited. Therefore, one might consider to use PubChem’s precomputed 2-D and 3-D neighboring relationships as complementary virtual screening tools or to help understand how chemical structures relate to each other relative to their biological efficacy.

#### 12.3.4 4. Effect of using multiple conformers

Taking into account all conformers of each CID for 3-D neighboring using the current methodology is simply not practical. The PubChem “Similar Conformers” neighboring relationship described here considers (at the time of writing) only two diverse conformers per compound (with a third conformer per compound soon to be released). One may wonder, as more conformers are considered, does one locate more chemical structures and, if so, to what extent? Is there a point of “diminishing returns”, where a plateau forms in the curve of unique neighbor count as a function of diverse conformer count? Indirect evidence addressing aspects of these questions can be found in the 3-D neighboring data PubChem provides.

PubChem assigns different unique compound identifiers (CIDs) for different isotopomers of the same chemical structure. For example, CID 2244 and CID 450661 are both aspirin (Figure 8), but they differ from each other in the mass of one of the carbonyl carbon atoms. Although they are effectively identical for 3-D neighboring purposes, the conformer generation processing employed in PubChem3D resulted in different “default” conformers that are effectively mirror images of each other, with an insignificant energy difference of less than 0.5 kcal/mol. Superposition of the default conformers of these two CIDs yields a ST of 0.83, meeting the ST neighboring threshold; however, the CT at this superposition is only 0.27, which is not similar enough to satisfy the “Similar Conformers” 3-D neighboring threshold. As shown in Figure 8 and Table 1, the neighbors for the first three diverse conformers of CID 2244 and CID 450661 each have some degree of overlap, and, in some cases, this overlap is significant. For example, 62% (775 of 1,251) of the 3-D neighbors for the first diverse conformer of CID 2244 are identical to the 3-D neighbors found for the second diverse conformer of CID 450661. Similarly, 63% (812 out of 1,296) of the 3-D neighbors of the second diverse conformer of CID 2244 overlap with those of the first diverse conformer of CID 450661, while the third diverse conformer of CID 2244 shares 60% (730 out of 1,214) of its neighbors with the third conformer of CID 450661. Although there is a great deal of similarity between different chosen conformers of aspirin, they still identify a sizeable population of unique 3-D neighbors between CID 2244 and CID 450661, and, thus, unique shape/feature space. This demonstrates

<sup>26</sup> Molecular shape and medicinal chemistry: a perspective

<sup>27</sup> Application of belief theory to similarity data fusion for use in analog searching and lead hopping

the sensitivity of the conformers used during neighboring processing, even for simple chemical structures like aspirin; however, considering PubChem is using a diverse conformer scheme, as more conformers are used in neighboring, the coverage of the conformational variation improves. This leaves one to wonder, how many diverse conformers per compound might be necessary to saturate this coverage and moderate the effects of this sensitivity?

To help address this question more directly, 4,218 compounds were 3-D neighbored against all of PubChem3D. This set of 4,218 compounds were selected using a query of the PubChem Compound database (“*has pharm*”[Filter] AND “*has 3d conformer*”[Filter] AND 0[AtomChiralUndeDefCount] AND 0[BondChiralUndeDefCount]). This query means that the queried chemical structures have known pharmacological action as annotated by MeSH<sup>18</sup>, have a conformer model in PubChem3D, and have zero undefined SP2/SP3 stereo centers. (The last criterion is utilized solely to limit the count of chemical structures considered and should have no bearing on the results of this test.) The PubChem CIDs for the selected chemical structures are available in Additional file

Additional file 1

**List of PubChem Compound identifiers (CIDs) that were neighbored using different counts of diverse conformers per compound.**

[Click here for file](#)

These molecules were selected as they are among the most biologically relevant small molecule chemical structures known, being heavily studied in the biomedical literature and consisting, in large part, of most known drugs. Of the very broad range of 367 pharmacological actions defined for the 4,218 small molecules, the three with greatest compound count were enzyme inhibitors (336), anti-bacterial agents (237), and antineoplastic agents (230). These small molecules with known biological action (*Query set*) were neighbored against 26,157,365 compound records (*Search set*), representing the entire “live” PubChem3D contents as of Oct. 2010, using up to 1, 3, 5, 7, and 10 diverse conformers per compound for both compound sets. As shown in Table 2, the average conformer counts between the *Query* set and *Search* set are similar, with the query set being slightly less flexible. The non-hydrogen atom count and feature count profiles depicted in Figure 9 for the *Query* set are also comparable to those found for the *Search* set.

Looking at Table 2, one can see that the average counts of neighbors per conformer and those per compound increase as a function of diverse conformer count. Interestingly, as shown in Figure 10, the average count of compound neighbors per compound appears highly correlated with the logarithm of total conformer pairs considered by neighboring. This suggests one must exponentially increase the count of conformer pairs to achieve a complementary linear increase in unique compound pairs.

It is not completely clear why this should be so, but one consideration comes to mind. It may be an artefact of the nature of the diverse conformer relationship, whereby a default conformer is chosen as the first, the most diverse conformer to the default conformer is the selected as the second, and each subsequent diverse conformer must be furthest away from the previously selected diverse conformers. This means that the most diverse conformers for a chemical structure are always considered first. Subsequently, each additional diverse conformer will increasingly resemble the previous diverse conformers, potentially yielding compound neighbors found previously by the other conformers for the same chemical structure. This is reflected by the ratio of conformer and compound 3-D neighbors. At three, five, and seven diverse conformers, 38%, 53%, and 61%, respectively, of the conformer neighbors point to the same compound neighbors. By ten diverse conformers, 68% of the conformer neighbors point to the same compound neighbors. With this said, one thing is clear. Neighboring more diverse conformers per compound will result in more compound neighbors per compound; however, the computation effort expended to do this grows exponentially as an increasing ratio of conformer neighbors show you more ways two compounds are interrelated.

One interesting aspect of Table 2 and Figure 10 is that the average conformer neighbor count per conformer grows very slowly. A ten times growth in conformers, corresponding to a 68 times increase in conformer pairs considered, results in only a 70% increase in the average conformer neighbor count. This is somewhat surprising given the argument above. It appears to suggest that each added diverse conformer of a chemical structure is also adding a significant portion of unique shape/feature space. This is seen in Table 1, whereby the conformer neighbors of each of the first three diverse conformers of aspirin (CID 2244 or CID 450661) mostly had very little overlap, typically less than 20%, of similar conformer neighbors with other diverse conformers of the same chemical structure. While the degree of unique shape/feature space being added may diminish as more diverse conformers are added, it would still appear to be rather substantial even at ten diverse conformers per compound. Eventually, one may expect, as even more diverse

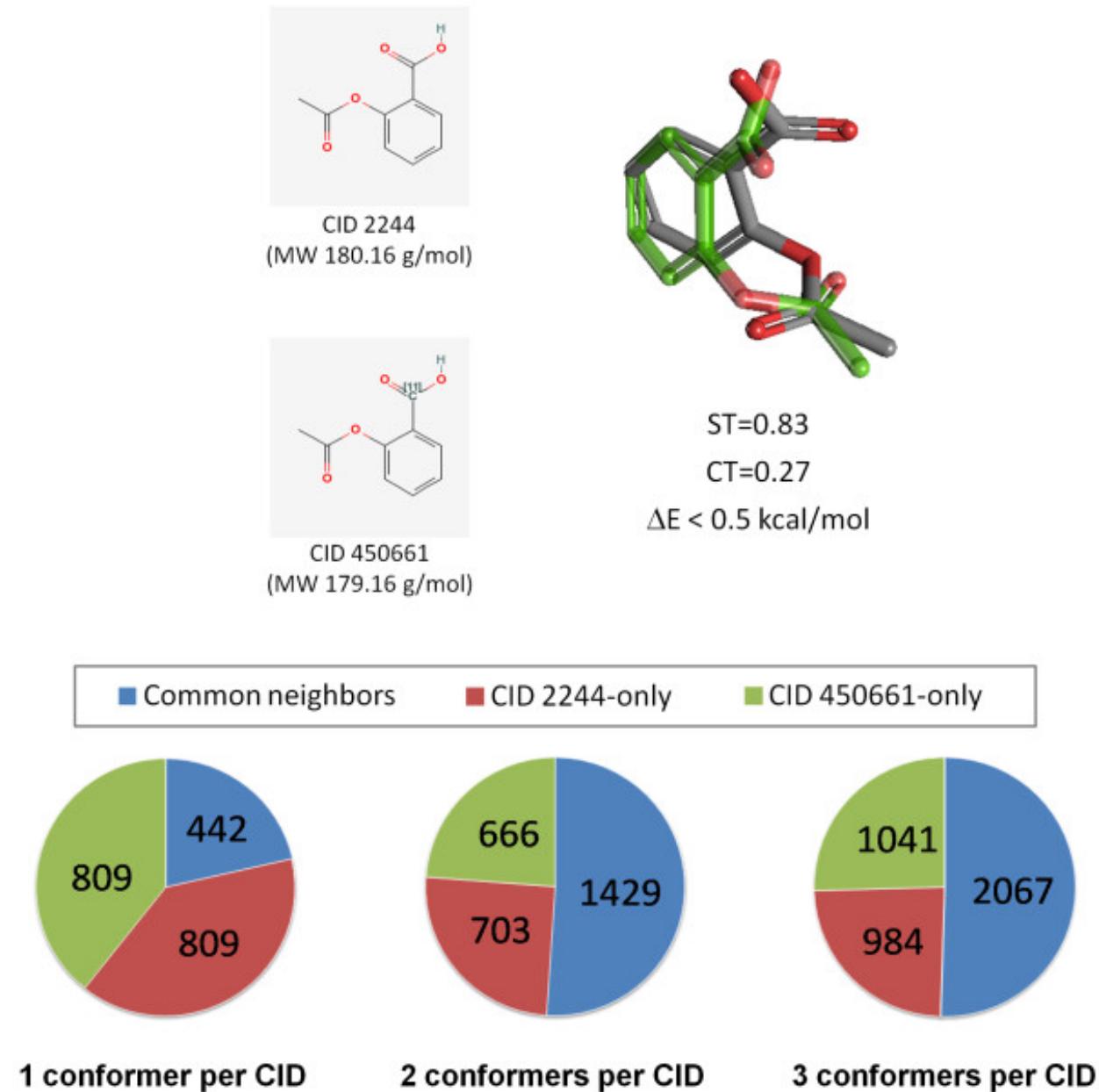
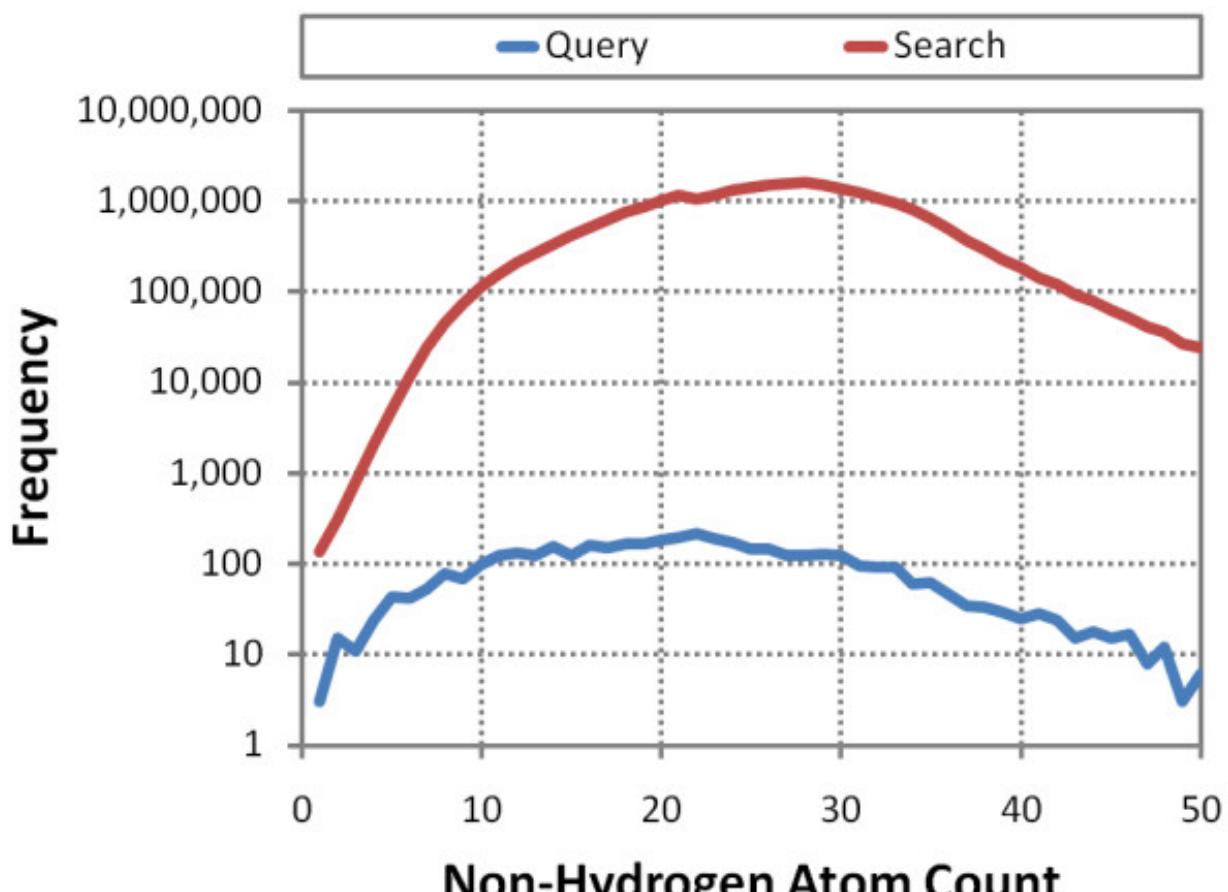
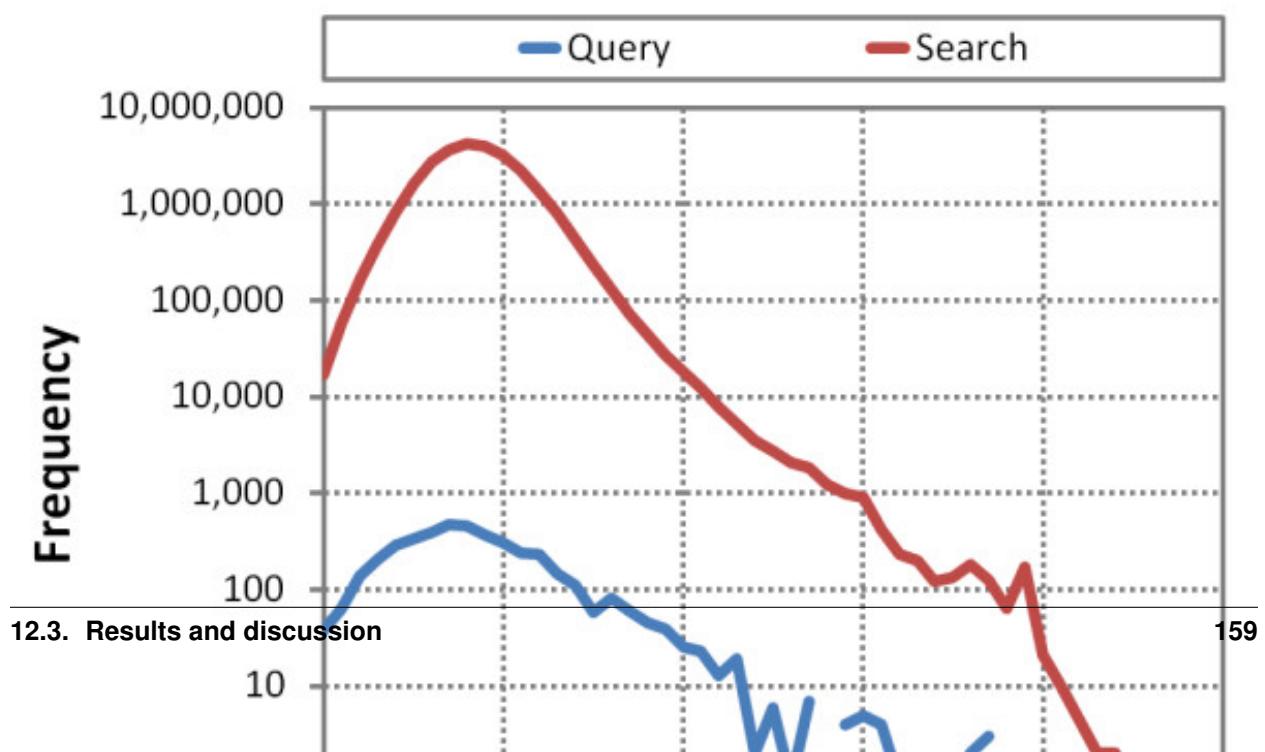


Figure 12.8: Figure 8. Sensitivity of conformer choice in 3-D neighboring

**Sensitivity of conformer choice in 3-D neighboring.** Independent conformer processing for CID 2244 and CID 450661, which differ by a single isotope, resulted in default conformers that are effectively mirror images. The 3-D neighbors are different, but less so as more diverse conformers are used, illustrating the sensitivity of the input conformers to 3-D neighboring, but also how using multiple conformers during neighboring can help mitigate such effects.



(a)



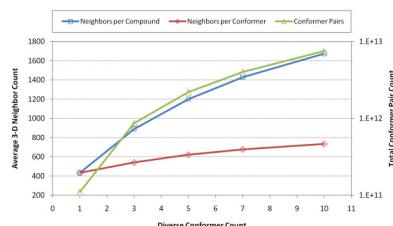


Figure 12.10: Figure 10. Compound 3-D neighbor count correlated to Log(Conformer pair count)

**Compound 3-D neighbor count correlated to Log(Conformer pair count).** Plot of count of 3-D neighbors per compound and neighbors per conformer [left Y-axis] and a plot of Log<sub>10</sub>(total conformer pairs) [right Y-axis] as a function of diverse conformer count considered during neighboring.

conformers are considered, that the average count of conformer neighbors per conformer may grow substantially, as conformers increasingly yield similar neighbor lists, but clearly this point is not yet reached at ten conformers per compound, as reflected by the continued growth in average count of compound neighbors per compound. Perhaps, for most chemical structures, this point may be reached by twenty diverse conformers. Using the computers and algorithms of today, and as reflected in the total search time in Table 2, twenty diverse conformers per compound is still a mountain too high to climb for a collection of the size of PubChem.

### 12.3.5 5. Efficiency of 3-D neighboring scheme

Although the overall speed of 3-D neighboring depends on various factors, such as atom count, use of a precomputed shape grid approach, etc., a modern computer processor core can process on the order of 10<sup>2</sup> to 10<sup>3</sup> 3-D conformer pair superpositions per second, when using a Gaussian-based shape definition. In theory, 26.1 million compounds with two diverse conformers per compound would require more than a quadrillion (10<sup>15</sup>) pair-wise conformer superposition determinations, corresponding to +40,000 years of processor core computation; however, PubChem 3-D neighboring processing was completed in about two months using ~2,500 computer processing cores (which represents more the throughput achieved in terms of actual time on a somewhat chaotic and somewhat unstable shared compute cluster rather than actual CPU time), meaning it took ~400 years of compute server time. How was this achieved?

To demonstrate the efficiency of the PubChem3D neighboring system, and reusing the previous example of querying 4,218 known bioactive small molecules against all of PubChem, Table 3 gives the percentage of conformer pairs excluded by filter type and the percentage of time spent in each stage of the neighboring processing. In the first stage, a series of three filters are utilized to screen out conformer pairs incapable of achieving the ST and CT thresholds of 0.8 and 0.5, respectively, required to be a neighbor. The most effective of these is the CT feature filter with 65% efficiency for this test set, which is to say more than half of all conformer pairs encountered can be effectively ignored. One nice aspect is that this CT feature filter operates on compound pairs, as opposed to conformer pairs. The other two filters at this stage check for incompatible shape or feature volume between conformer pairs. The total CPU time spent performing these three filters is less than 1%, yet they are effective, removing 68% of all conformer pairs from further consideration.

Alignment recycling is the next stage after filtering. This methodology consists of: comparing a shape fingerprint; locating common reference shapes; and then reuse of the alignment to the common reference, where the shape overlap and the feature overlap are computed at that recycled alignment to the reference shape. This is repeated for each common reference shape and only the best superposition is kept.

Alignment recycling provides two opportunities to further remove conformer pairs from consideration. If a reference shape cannot be found in common, the conformers are considered to be too different to be a neighbor. This alignment recycling fingerprint filter removes an additional 4% of all conformer pairs (14% of all conformer pairs not already filtered). If the pre-optimized best overlap from alignment recycling is not sufficiently large (yielding an ST of at least 0.735), the conformer pair is considered to be incapable of being a neighbor. This alignment recycling overlap filter removes 27% of all conformer pairs (96% of all conformer pairs not already filtered) but consumes 86% of CPU time. Together, all filtering steps remove 99.8% of conformer pairs prior to optimization of the conformer superposition at

the recycled alignment. The final shape optimization step consumes 10% of the CPU time, retaining less than 0.6% of optimized conformer pairs as neighbor pairs. About 66% of conformer pairs shape-optimized are rejected due to an insufficient ST value (<0.795) to become a neighbor and the remainder rejected due to insufficient CT value (<0.495) at the shape-optimized superposition.

The overall throughput of the 3-D neighboring methodology is consistent across the range of diverse conformers considered, at a rate of ~150,000 conformers per second. The other overhead reported in Table 3 involves mostly the billions and trillions of timing measurements but also involves some memory allocation aspects. In reality, with timing statistics turned off, there is very little other overhead to the method. While the total size of the input binary data files grows as a function diverse conformer count, ranging from 19 GB to 159 GB, the computational density is more than sufficient to avoid making input of these search files a bottleneck, provided at least four conformers are being queried simultaneously. If fewer than four conformers are queried at a time, and the input binary files are not memory resident, input can be a bottleneck.

### 12.3.6 6. Alignment recycling

The alignment recycling methodology<sup>14</sup> was extended to cover non-hydrogen atom counts from 0-50 and rotatable bond counts from 0-15. This was achieved by leveraging our recent study on the diversity of shape space<sup>15</sup>, where shape space was shown to grow gradually as a function of conformer volume and a dynamic shape similarity threshold for a relatively constant count of reference shapes. This curve (the Unique-Shape Tanimoto in Figure 11) was used to effectively partition shape space into seven regions. Each fingerprint region has a distinct shape similarity threshold (the Fingerprint Tanimoto in Figure 11) and covers the entire shape diversity of a given conformer volume range. As Table 4 shows, there are a total of 3,311 reference shapes across all seven regions, representing the entire shape diversity of 5.2 billion conformers for the entire contents (live and non-live) of the PubChem3D system (+45.9 million small molecules).

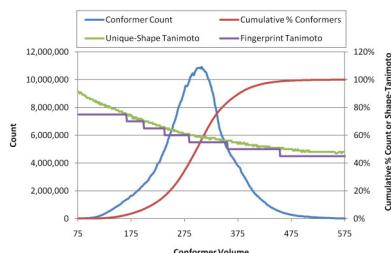


Figure 12.11: Figure 11. Shape fingerprint design

**Shape fingerprint design.** Plot of conformer count (blue line) [left Y-axis], cumulative % conformers (red line) [right Y-axis], unique-shape Tanimoto (green line) [right Y-axis], and fingerprint Tanimoto (purple line) [right Y-axis] as a function of conformer volume ( $\text{\AA}^3$ ). The fingerprint Tanimoto indicates the seven volume regions of the reference shape and its corresponding ST minimum distance between reference shapes.

When computing the shape fingerprint of a conformer, if a reference shape has a shape optimized superposition that is greater than or equal to the fingerprint shape similarity threshold (12, for the first ten diverse conformers from the 26.1 million compounds (246 million conformers) covered in the study of 4,218 small molecules of biomedical interest, there are at most a total of 129 reference shapes used per conformer, with an average and standard deviation of  $39 \pm 13$ . This sparseness is to be expected as the shape fingerprint primarily identifies a specific region of shape space. Figure 13 depicts the count of set bits per fingerprint region across the 246 million conformers. As Table 4 shows, each fingerprint area covers a specific volume range. So, one should not expect a conformer with volume  $100 \text{ \AA}^3$  to have reference shapes in the conformer volume range 433-999, and vice versa. In fact, while each conformer has at least one reference shape set, many of the 246 million conformers considered do not have any reference shapes set in one of the seven different fingerprint regions. For the fingerprint reference shape volume ( $\text{\AA}^3$ ) ranges 1-165, 166-199, 200-238, 239-285, 286-344, 345-432, and 433-999, a total of 83.2%, 62.4%, 35.1%, 11.6%, 2.4%, 2.6%, and 4.1% of

the 246 million conformers, respectively, are not using the fingerprint region. This is reflected in the relatively high counts of conformers with no reference shapes, as depicted in the magnified section of Figure 13.

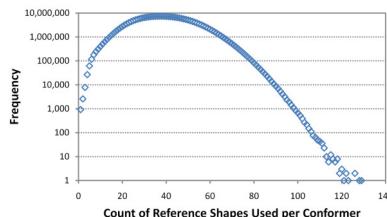


Figure 12.12: Figure 12. Shape fingerprint bits are sparsely set

**Shape fingerprint bits are sparsely set.** Frequency plot of the total count of fingerprint reference shapes set per conformer for the first ten conformers of the 26,157,365 PubChem3D Compound records in the *Search* set, corresponding to 246,874,949 conformers.

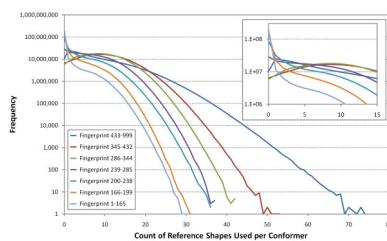


Figure 12.13: Figure 13. Some shape fingerprint volume regions are mostly unused

**Some shape fingerprint volume regions are mostly unused.** Plot of the frequency of the shape fingerprint bit counts per fingerprint volume region for the first ten conformers of the 26,157,365 PubChem3D Compound records in the *Search* set, corresponding to 246,874,949 conformers. A significant percentage of the conformers do not use fingerprint reference shapes in the volume ranges 1-165, 166-199, and 200-238.

The relative popularity of each discrete 3,311 reference shapes varies markedly. Depicted in Figure 14, one can see the frequency of use of each reference shape defined in a given fingerprint volume range for the 246 million conformers. In each fingerprint volume range, there exist a very small number of reference shapes that clearly stand out as being used most often. Afterwards, the use of individual reference shapes falls off sharply and then gradually, until only peripheral reference shapes that are rarely used are left. This motif is seen for all fingerprint volume regions and may reflect the relative uniqueness (or lack thereof) of shapes across the first ten diverse conformers in PubChem.

### 12.3.7 7. Superposition storage

Superposition of two conformers requires modification of the coordinates of one conformer relative to the other. Retention of the rotational matrix and translation vector is a practical approach to retain a superposition between conformers to avoid having to re-compute a superposition or store modified coordinates of a conformer.

Storage of superposition results in PubChem3D involves identification of: the two conformers involved, often with one of the two conformers implicitly identified (*e.g.*, by storing the superposition as a subordinate property of a conformer); the  $3 \times 3$  rotation matrix; and the  $3 \times 1$  translation vector. The PubChem3D conformer ID is often represented as either a 64-bit unsigned integer (sometimes stored in 16-character hex form), with the 32-high bits representing the PubChem Compound identifier (CID) and the 16-low bits representing the local conformer ID (LID), or two numbers ":" separated (*e.g.*, CID.LID). Storage of the rotation and translation parts represents more of a challenge, given there are twelve floating point numbers to convey. To provide for a more compact superposition representation, the ability to pack/unpack the rotation and translation into a 64-bit unsigned integer was developed. While described in more detail in the **Materials and Methods** section below, this involves transforming the rotation matrix into a quaternion

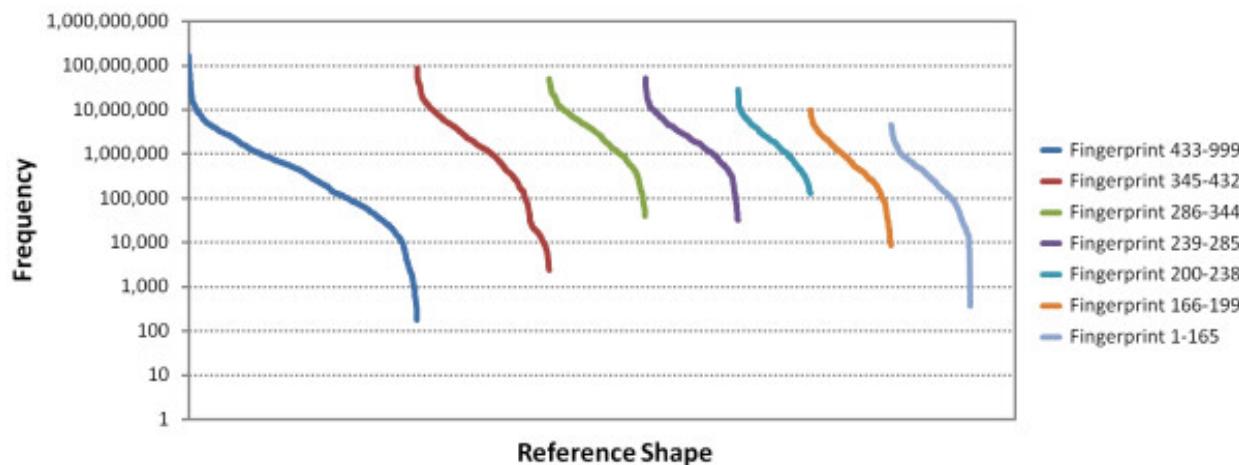


Figure 12.14: Figure 14. Frequency of fingerprint reference shape use

**Frequency of fingerprint reference shape use.** The frequency of use of the 3,311 fingerprint reference shape bits, separated by fingerprint volume region, by the first ten conformers of the 26,157,365 PubChem3D Compound records in the *Search* set, corresponding to 246,874,949 conformers. Some reference shapes are very popular while others are rarely used.

and packing each of the four ( $Q_w$ ,  $Q_x$ ,  $Q_y$ ,  $Q_z$ ) components into 32-bits, 8 bits each. The remaining 32-bits are used to encode the translation vector.

To study the loss in accuracy due to encoding/decoding the conformer superposition information into a 64-bit integer, 1.85 billion unique conformer neighbor pairs in the 0-20 million CID range involving conformers that are the first diverse conformer of a compound were used. The chemical structure and 3-D coordinates of each conformer pair were: downloaded from the PubChem3D data system database; the superposition between the conformers was optimized, yielding a before ST/CT value pair; the superposition rotation and translation was encoded, decoded, and applied to the original downloaded conformer pair coordinates; and then a single point ST and CT value was computed, yielding an after ST/CT value pair. The difference in the before/after ST and CT values were binned in 0.001 increments and the population of the occupied bins are plotted in Figure 15 and summarized in Table 5.

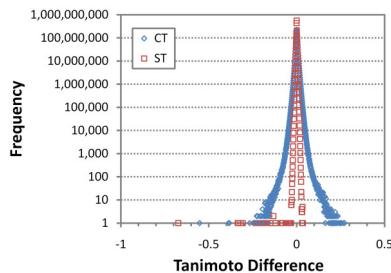


Figure 12.15: Figure 15. Effect of superposition packing on ST/CT

**Effect of superposition packing on ST/CT.** The difference in the ST/CT scores (binned in 0.001 increments) before and after packing superposition translation/rotation information into an unsigned 64-bit integer. A positive difference indicates an enhancement of the ST/CT scores and a negative difference indicates an error in the ST/CT scores.

Perhaps most remarkable, the superposition encode/decode procedure is just as likely to enhance the ST and CT values as detract from them. Also interesting is that the CT error curves are much broader, reflecting, in part, the much greater positional sensitivity of the CT measure. Small deviations in rotation have an increasing effect the further an atom is from the molecule center. Fictitious feature atoms are relatively sparse, have small atomic radii, and are often close to the periphery of the chemical structure. Shape similarity, on the other hand, is not as sensitive, as real atoms are relatively dense and most atoms in the molecule are typically near the steric center, thus, fewer atoms are affected from rotation encoding effects. As a whole, the use of a 64-bit integer to store a conformer pair superposition results

in relatively few cases where the Tanimoto difference (after-before) is less than 0.025, with the chances for this to occur for ST and CT being 1 in 14.6 million and 1 in 955, respectively. If the error from being off a small fraction of a degree from the original superposition is too much, one could simply re-optimize the conformer superposition provided by PubChem, as the benefits in terms of the ease of storage are considerable.

## 12.4 Conclusion

In the present paper, the PubChem 3-D “Similar Conformers” neighboring relationship and the methodology used in its computation are described. PubChem 3-D neighbors are defined as any two conformers with a shape-optimized superposition yielding a similar 3-D conformer shape (ST value of 0.8) and similar 3-D orientation of functional groups typically used to define pharmacophore features (CT value of 0.5). In the cases of chemical structures without features, a similar 3-D conformer shape with ST value of 0.93 is used.

To make the calculation of this 3-D neighboring relationship tractable, a series of filters were designed to avoid the time-consuming shape-superposition computation between conformer pairs that could not possibly be 3-D neighbors. This resulted in an average throughput of 150,000 conformer pairs per second per processor core, a speed sufficient to consider multiple diverse conformers per compound in the 3-D neighboring relationship.

Neighboring the first two diverse conformers of 26.1 million PubChem Compound records yielded 8.16 billion 3-D conformer neighbor pairs and 6.62 billion 3-D compound neighbor pairs, with an average of 253 “Similar Conformers” per PubChem Compound. Comparison of the PubChem 3-D “Similar Conformers” neighboring relationship with the PubChem 2-D “Similar Compounds” neighboring relationship using three well-known bioactive molecules (aspirin, caffeine, and morphine) showed a considerable degree of uniqueness between the two neighboring relationships and providing a number of related structures with significant biological annotation. This was also illustrated by the ability of the 3-D neighboring relationship to associate eight selected non-steroidal anti-inflammatory drugs (NSAIDs) to each other, despite little 2-D pair-wise similarity between most of the compound pairs. Additional study of 4,218 small molecules of biomedical interest across a range of diverse conformers shows that neighboring more conformers per compound will result in being able to associate more chemical structures to each other; however, an exponential increase in the count of conformer pairs considered results in only a linear increase in additional compound 3-D neighbor pairs.

## 12.5 Materials and methods

### 12.5.1 1. Chemical structure 3-D representation

Theoretical 3-D descriptions of the 26,157,365 chemical structures covered in this work and found in the PubChem Compound database<sup>12</sup> are generated as described in our previous studies<sup>15<sup>28</sup></sup>. It is important to note that these conformers are not energy minima on a potential energy surface, but an ensemble of energetically-accessible (at room temperature), biologically-relevant (able to reproduce most known bioactive conformations), sampled (with a minimum atom pair-wise RMSD separation) conformations that the molecule may cover. In theory, these ensembles describe all relevant molecular shapes (including all important energy minima) within the resolution of the clustering RMSD for the conformer ensemble.

### 12.5.2 2. Molecular shape and features

An atom-centered Gaussian description<sup>9<sup>10<sup>11<sup>12</sup></sup></sup></sup> using Bondi radii<sup>29</sup> is utilized to compute 3-D similarity. Fictitious “feature” atoms (also known as “color” atoms) are defined to represent the pharmacophore feature functional group

---

<sup>28</sup> PubChem3D: conformer generation

<sup>29</sup> van der Waals volumes and radii

types present in a chemical structure. The Mills/Deans implicit force-field<sup>30</sup>, as implemented in the OEShape C++ Toolkit, is employed to identify these features. The six feature types recognized are: anion, cation, hydrogen-bond donor, hydrogen-bond acceptor, hydrophobe, and ring. Feature atom 3-D coordinates are computed relative to the steric center of real “parent” atoms that comprise each feature. Post processing of feature atom assignment identifies any features of the same type within 1.0 Å of each other and merges the unique parent atoms that comprise the two features. This post processing step is performed iteratively, until no additional features are merged. The radius used for all feature atoms is 1.08265 Å.

Shape similarity computation utilizes the shape Tanimoto (ST) via Eq. (2) and only considers the non-hydrogen atoms in the molecule. Feature similarity, unlike shape similarity, involves summing the individual overlaps of the six component feature atom types when computing the  $A$ ,  $B$ , and  $AB$  found in Eq. (2); thus, yielding Eq. (3) for the feature similarity measure, color Tanimoto (CT). Otherwise, the feature similarity computation method is identical to the shape similarity computation method.

When optimizing the shape superposition between a conformer pair, the OpenEye OEShape C++ toolkit is used via the OEBestOverlay object, with the parameters OEOverlapRadii::All and OEOverlapMethod::Analytic. Any other shape or feature computation utilizes in-house implementations using the Grant and Pickup<sup>9</sup> Gaussian-based shape methodology. For all in-house shape-based approaches, an exponent lookup table of size 6,001 is used in lieu of exponent calculation for the range of (-12.0 to 0.0) in 0.002 increments. Exponent values outside of this range are zero. All other terms in the Grant and Pickup shape-based methodology are computed exactly.

A grid-based approach is used by parts of the 3-D neighboring methodology to estimate the shape overlap with  $O(N)$  computational complexity. In these cases, a 3-D lattice of points separated by 0.25 Å give the shape overlap of a carbon probe-atom at the grid point to the query conformer. A cut-off distance of 4.5 Å is used for each query conformer atom, where no additional contribution to shape overlap is considered.

### 12.5.3 3. Diverse conformer concept

Although the theoretical conformer ensemble for each molecule may have up to 500 conformers (averaging ~110), it is not practical to consider all conformers for PubChem 3-D neighboring. Therefore, a diverse conformer concept is introduced that orders the conformers in the ensemble for a chemical structure by their combined shape and feature dissimilarity, with the most dissimilar conformers first. The lowest-energy conformer in the conformer ensemble is selected as the default, first diverse conformer to seed the process. The conformer with the least combo Tanimoto (being the sum of the ST and CT similarity values for the ST-optimized superposition) to the first conformer is selected as the second most diverse conformer. The conformer with the least sum of combo Tanimoto to the first two conformers is selected as the third, and so on until all conformers are assigned a diverse conformer ordering. In the case of a tie, the conformer with the largest sum of combo Tanimoto to all unassigned conformers is selected. If a tie persists, the conformer with the least LID (local conformer identifier) is selected.

### 12.5.4 4. Shape fingerprint definition

Haigh *et al.*<sup>13</sup> applied a clustering technique to select a diverse set of reference shapes that cover the overall shape space of possible 3-D shapes, and generated 3-D molecular shape fingerprints using these reference shapes. Comparison of molecular shape fingerprints was shown to be orders of magnitude faster than shape-overlap-based approaches such as ROCS<sup>10</sup>, illustrating its potential in screening a large 3-D chemical database. Therefore, we applied the 3-D shape fingerprint technique, in conjunction with “alignment recycling”<sup>14</sup>, for use in computing a “Similar Conformer” relationship.

In our recent study<sup>15</sup>, a dynamic shape similarity threshold ( ${}^{\text{thresh}}\text{ST}$ ) was employed in clustering conformers of a particular volume such that the resulting reference shape count became less than or equal to a certain number (200). In this manner, the number of reference shapes per volume can be kept relatively constant while the growth of the shape space as a function of volume is manifest by a decrease in [11](#)). The plot of the Unique-Shape Tanimoto versus the conformer volume was used to choose appropriate [11](#)). The <sup>15</sup> were then pooled and clustered at

<sup>30</sup> Three-dimensional hydrogen-bond geometry and probability information from a crystal survey

The resulting “unique shape” count is listed in Table 4, with the conformer count that belongs to the shape space represented by the corresponding unique shapes. It indicates that the shape space spanned by 5.24 billion conformers (the entire contents of the PubChem3D archive, live and non-live, from more than 45.9 million unique chemical structures) can be represented in such a manner so as to only require 3,311 unique reference shapes (a number which may grow as a function of time). Figure 14 shows the frequency of use of the various 3-D fingerprint reference shapes, with some being heavily utilized while others are rarely used. The volume range 433-999 is the largest in both volume spanned and count of reference shapes. We anticipate that this volume range may need to be split into separate regions in the future.

## 12.5.5 5. PubChem 3-D neighbor processing

PubChem Compound (CID) records are partitioned into two sets, “live” and “non-live”. A “live” CID is one that has at least one current version PubChem Substance record pointing to it. The “non-live” partition contains all CIDs not considered to be “live”. For each record that is contained in the PubChem3D system and considered to be “live”, PubChem computes a “Similar Conformer” relationship that considers both shape similarity (ST 0.8) and feature similarity (CT 0.5). [Chemical structures without features, while rare, can have a similar conformer relationship with other featureless chemical structures provided the ST 0.93.] Essentially, this amounts to a 3-D similarity search of a given conformer across the first N-diverse conformers of “live” CIDs, where at the time of writing “N” is two, with a third in the process of being added. This processing we call “neighboring”.

This PubChem3D similarity search is a multistep process designed: to filter out conformer pairs that cannot possibly be neighbors, to generate a near-optimal superposition between conformer pairs, and to perform a final optimization of the superposition to maximize the shape volume overlap between conformer pairs. These distinct stages are described below.

### Stage 1: Filtering

There are multiple filters used in PubChem3D neighboring, each with different degrees of computational cost and efficiency. The cheapest filters utilize the ST and CT equations [Eq. (2) and Eq. (3)], the pre-computed self-overlap volumes ( $A$  and  $B$ ), the predefined Tanimoto thresholds, and a rearranged Tanimoto equation solving for  $AB$ , which now represents the minimum overlap ( $_{\min}AB$ ) necessary to meet the threshold criteria (ST or CT). If this  $_{\min}AB$  is greater than either  $A$  or  $B$ , then the conformer pair may be eliminated from further consideration as the maximum possible overlap between the conformer pair will be smaller than  $_{\min}AB$ , yielding a Tanimoto value less than the respective threshold. The resulting equation for CT, using the threshold of 0.5, is:

This approach to filtering can be used a second time (although at greater computational cost) for feature similarity by using the individual feature counts. For each feature type, one uses the feature count ( $A$  and  $B$ ) and the common count (either  $A$  or  $B$ , whichever is the lesser). One can then compute a  $\max CT$  that must be above the threshold of 0.5 to possibly be a neighbor, as shown in Eq. (5):

An enormous advantage to this filter is that it operates on compound pairs. This means it applies to all conformers being neighbored for the respective compound pair, unlike the previous CT filter which is per conformer pair. So, as the count of conformers per compound is increased, the utility of this filter is magnified.

Taking the exact same approach for ST as CT is not possible, as the shape volume  $AB$  overlap can be greater than either  $A$  or  $B$ , due to the Grant and Pickup atom-centered Gaussian-based formulation of molecular shape, the radii used, and the bond distances of atoms. The net effect is such that the  $AB$  overlap can be significantly greater than either  $A$  or  $B$ . The filter for ST (and the threshold of 0.8) becomes:

The ratios in both Eqs. (6) and (7) must be between 0.75 and 1.50 for the conformer pair to possibly be a neighbor.

## Stage 2: Shape fingerprint comparison and alignment recycling

The next stage of filtering utilizes the PubChem3D shape fingerprint computed for each conformer. For a given conformer pair to become neighbors to each other, there must be a common reference. If a common reference cannot be found, the conformer pair cannot be neighbors and no further computation is necessary. For each common reference, the  $3 \times 3$  rotation matrix and  $3 \times 1$  translation vector that aligns the conformer to the reference is utilized. The resulting  $3 \times 3$  rotation matrix and  $3 \times 1$  translation vector are applied to the coordinates of one of the conformers. This provides a superposition of one conformer to the other. A combined shape and feature overlap for the conformer pair using this alignment is then computed as following:

where  $_{\text{shape}} \text{AB}$  is the common shape volume between the conformer pair (using a precomputed shape-grid at 0.25 Å resolution) and  $_{\text{feature}} \text{AB}$  is the sum of six component feature overlaps (using a feature atom radius of 1.25 Å, to account for proximate color atoms).

This approach is repeated for each common reference shape. The reference-shape-derived conformer alignment yielding the largest  $_{\text{combo}} \text{AB}$  value is used in a final  $_{\text{shape}} \text{AB}$  overlap computation, this time not using a grid-based approach. The final  $_{\text{shape}} \text{AB}$  overlap is used along with the pre-computed self shape overlap values per conformer to compute the ST at this overlap geometry. If the computed ST is not greater than 0.735, the conformer pair is considered to not possibly be a neighbor. [The “grid”  $_{\text{shape}} \text{AB}$  can be sufficiently different than the “exact”  $_{\text{shape}} \text{AB}$  value, resulting in the loss of many neighbor relationships.]

## Stage 3: Shape overlay optimization and ST/CT score computation

Using the “alignment recycling” conformer pair superposition from the previous stage, a final superposition optimization to maximize the shape volume overlap between the conformer pair is performed using the OEShape C++ toolkit<sup>12</sup>. If the final conformer superposition yields an ST 0.8 (actually 0.795 after rounding to the nearest 0.01 is considered), the CT is also computed at the same conformer alignment. If CT 0.5 (actually 0.495 after rounding to the nearest 0.01 is considered), the conformer pair is considered to be a neighbor. As mentioned previously, if both conformers are devoid of features, alternatively, an ST 0.93 (actually 0.925 after rounding to the nearest 0.01 is considered) is sufficient to be considered a neighbor. The  $3 \times 3$  rotation matrix and  $3 \times 1$  translation vector and the ST/CT similarity values are retained for the conformer pair.

### 12.5.6 6. PubChem 3-D neighbor processing addendum

There are additional aspects to neighbor processing that are germane to their accuracy and use. To minimize input overhead and memory utilization, all input data is encoded in a highly compact 64-bit aligned binary format (gzip or bzip2 compression reduces file size by only 4%) that contains all information necessary to perform the neighboring computation. One side effect of this encoding is that conformer coordinates are transformed into an integer value with a resolution of +/- 0.0015 Å and restricts coordinates to the range (-50 Å, +50 Å), which is more than sufficient for all conformers in the PubChem3D data system. Another side effect of this encoding scheme is that, to obtain the superposition alignment between neighbored conformers, one must first transform the conformers to their steric center (*i.e.*, subtract the coordinate average per axis) prior to applying the stored  $3 \times 3$  rotation matrix and then the  $3 \times 1$  translation vector (in that order). For conformers stored in the PubChem3D system, conformers are already at the steric center (and rotated into the non-mass weighed inertial frame of reference).

Another major consideration is that the  $3 \times 3$  rotation matrix and  $3 \times 1$  translation vector for all fingerprints and neighbor pairs are encoded and stored as a 64-bit unsigned number. This involves transforming the  $3 \times 3$  rotation matrix into the quaternion and then encoding these four values ( $Q_w$ ,  $Q_x$ ,  $Q_y$ , and  $Q_z$ ) and the  $3 \times 1$  translation vector. Each quaternion value is encoded in eight bits with a range of (-1, +1) and a resolution of 0.00784. Care must be taken when either encoding or decoding these quaternion values to always normalize, as doing so reduces the encoding error. Encoding the  $3 \times 1$  translation vector uses a slightly different approach to achieve a maximum range of (-100 Å, +100 Å). Three of the remaining 32 bits are used to hold the sign of the X, Y, and Z translation. The remaining 29 bits encode the three absolute values in the following fashion. A value of 1.0 is added and the log of the resulting number

is taken and divided by 812 ( $812^3 = 535387328$ , which just fits within 29 bits) and rounded to the nearest whole number. The integer encoded X, Y, and Z are then combined as such:

The result of using a log value provides for a translation encoding that gives the best precision at 0 Å (+/- 0.0028 Å) and the worst at 100 Å (+/- 0.29 Å). Considering the requirement that a conformer pair must be at the steric center prior to encoding, the encoding error due to translation is minimal.

Validation of the encoding/decoding accuracy as a function of ST and CT values across +1.85 billion unique neighbor pairs show (Table 5) that there is nearly an equal possibility to enhance the ST or CT value as it is to detract from it. However, the CT is much more sensitive to encoding/decoding than ST, with a 1 in 38 chance of yielding a value less by 0.01 than that reported. This is to be expected, as the radius used for feature atoms in CT computations is nearly 60% smaller than for carbon atoms, meaning small changes in the alignment can have a big effect on similarity, especially for features that are furthest from the steric center (a “torque arm” effect). Additionally, the shape overlap optimization does not consider feature atoms, meaning a small change in rotation or translation may either increase or decrease the CT value, considering no attempt was made to optimize feature alignment.

## 12.6 Competing interests

The authors declare that they have no competing interests.

## 12.7 Authors' contributions

EEB performed most of the research and SK wrote the first draft. SHB reviewed the final manuscript. All authors read and approved the final manuscript.

## 12.8 Acknowledgements

Many thanks to Roger Sayle for providing key insights on ways to improve the rotational/translational matrix pack/unpack scheme. We are grateful to the NCBI Systems staff, especially Ron Patterson, Charlie Cook, and Don Preuss, whose efforts helped make the PubChem3D project possible. This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, U.S. Department of Health and Human Services. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (<http://biowulf.nih.gov>).

# **ANALYSIS OF *IN VITRO* BIOACTIVITY DATA EXTRACTED FROM DRUG DISCOVERY LITERATURE AND PATENTS: RANKING 1654 HUMAN PROTEIN TARGETS BY ASSAYED COMPOUNDS AND MOLECULAR SCAFFOLDS**

## **13.1 Abstract**

### **13.1.1 Background**

Since the classic Hopkins and Groom druggable genome review in 2002, there have been a number of publications updating both the hypothetical and successful human drug target statistics. However, listings of research targets that define the area between these two extremes are sparse because of the challenges of collating published information at the necessary scale. We have addressed this by interrogating databases, populated by expert curation, of bioactivity data extracted from patents and journal papers over the last 30 years.

### **13.1.2 Results**

From a subset of just over 27,000 documents we have extracted a set of compound-to-target relationships for biochemical *in vitro* binding-type assay data for 1,736 human proteins and 1,654 gene identifiers. These are linked to 1,671,951 compound records derived from 823,179 unique chemical structures. The distribution showed a compounds-per-target average of 964 with a maximum of 42,869 (Factor Xa). The list includes non-targets, failed targets and cross-screening targets. The top-278 most actively pursued targets cover 90% of the compounds. We further investigated target ranking by determining the number of molecular frameworks and scaffolds. These were compared to the compound counts as alternative measures of chemical diversity on a per-target basis.

### 13.1.3 Conclusions

The compounds-per-protein listing generated in this work (provided as a supplementary file) represents the major proportion of the human drug target landscape defined by published data. We supplemented the simple ranking by the number of compounds assayed with additional rankings by molecular topology. These showed significant differences and provide complementary assessments of chemical tractability.

## 13.2 Introduction

An important factor in assessing the global progress in drug research is the number of targets for which therapeutic small-molecule modulators have been, are being, or could be, generated. This question was addressed in the landmark publication in 2002 that introduced the “druggable genome” concept<sup>1</sup>.

This total of approximately 3,000 human proteins was arrived at by homologous family extrapolation from the targets of approved drugs at that time. The count of successful targets was updated in 2006 and stood then at 324, of which the subset of human proteins was 207<sup>2</sup>. Despite many publications covering this topic, the inclusion of explicit listings of target identifiers, extrinsic to the data sets from which they were derived, are rare, with the partial exception of a poster that included 185 human targets of approved oral drugs<sup>2</sup>.

Notwithstanding, there are now public databases from which it is possible to browse and extract targets with explicit links to bioactive compounds. DrugBank is one such resource<sup>3</sup>. It has a total of 6,827 drug entries including 1,431 FDA-approved small molecule drugs and 5,212 research compounds linked to 4,477 non-redundant protein sequences. These include primary targets, cross-screening targets, metabolising enzymes and associations inferred from compound name with protein name co-occurrences automatically extracted from the literature. The Therapeutic Targets Database (TTD) contains conceptually similar information to DrugBank but organised into a different data structure<sup>4</sup>. It provides sequence subsets of their total of 1,675 targets divided into 348 approved, 260 clinical trial and 1,067 research targets. The BindingDB resource also includes approved and research targets with a focus on measured small-molecule binding affinities and ligands. It currently includes 5,526 protein targets and 271,419 compounds<sup>5</sup>. The largest public resource of this type is the ChEMBL database with 8,091 targets and 658,075 compounds extracted from medicinal chemistry journal papers (N.B. a subset of ChEMBL data is now incorporated into BindingDB)<sup>6</sup>. Three of the databases above, DrugBank, TTD and ChEMBL, have recently been included in a comparative study of compounds and targets<sup>7</sup>.

## 13.3 Databases and Processing

The company GVKBIO<sup>8</sup> has developed a suite of databases over the last 9 years that are now unified under a single query interface, termed GVKBIO Online Structure Activity Relationships (GOSTAR)<sup>910</sup>. The results we present are from two of the six GOSTAR components, the Medicinal Chemistry (MCD) and Target (TGD) Databases. Their combined utility for mining drug research data has already been described<sup>11121314</sup>. In addition, the comparison of

---

<sup>1</sup> The druggable genome

<sup>2</sup> How many drug targets are there?

<sup>3</sup> DrugBank: a knowledgebase for drugs, drug actions and drug targets

<sup>4</sup> TTD: Therapeutic Target Database

<sup>5</sup> BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities

<sup>6</sup> ChEMBL

<sup>7</sup> Mapping Between Databases of Compounds and Protein Targets

<sup>8</sup> GVK BIO

<sup>9</sup> About GOSTAR

<sup>10</sup> Database Systems for Knowledge-Based Discovery

<sup>11</sup> The influence of drug-like concepts on decision-making in medicinal chemistry

<sup>12</sup> Physicochemical property profiles of marketed drugs, clinical candidates and bioactive compounds

<sup>13</sup> Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success

<sup>14</sup> Gaining Insight into Off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis

compound and target content of these with other bioactivity databases has been reported in publications that included the expansion of coverage between 2006 and 2008<sup>1516</sup>.

The data in MCD and TGD are derived from the large-scale expert extraction of structure-activity relationships (SAR) from patents and journal papers reporting the results of drug discovery research<sup>9</sup>. The basic process is familiar to scientists working in this area. By inspecting a document “D” they can identify the description of a biochemical assay “A” (e.g. for enzyme activity) with a quantitative result “R” (e.g. a Ki) for a compound “C” (e.g. a specific chemical structure) that defines it as an activity modulator (e.g. an inhibitor) of protein target “P” (e.g. a protease). An outline of these relationships is shown in Figure 1.

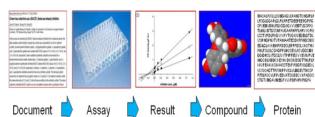


Figure 13.1: Figure 1. Depiction of the key entities and the relationships between them (D-A-R-C-P) used to populate the MCD and TGD databases

**Depiction of the key entities and the relationships between them (D-A-R-C-P) used to populate the MCD and TGD databases.**

At GVKBIO the relationships between these five entities of document, assay description, assay result, compound structure and protein target (D-A-R-C-P) are manually abstracted by a team of expert curators and transferred to document-centric relational databases. These contain data predominantly from the research phases of drug discovery but, because this extends back over 30 years, much of the primary data for approved drugs is included. The difference between them is that MCD extracts data from 120 journals selected for their high content of D-A-R-C-P relationships on a per-journal basis. TGD extracts the same relationships from patents covering the “big ten” target classes (kinases, GPCRs, proteases, nuclear hormone receptors, ion-channels, transporters, lipases, phosphatases, oxidoreductases and transferases). The process involves a triage to select a representative of the patent family for extraction. The addition of compounds to the database is limited to exemplified structures linked to quantitative or qualitative assay data. While all structures with quantitative results are extracted, where the activity data is ranged or only qualitative, the number of compounds extracted is capped at 200 or 100 examples, respectively<sup>10</sup>.

Details of these databases are described elsewhere but briefly, structures and related metadata for the GOSTAR database records are stored in an Oracle database<sup>17</sup>. The compound counts are defined by a unique structure identifier based on the Standard InChIKey<sup>18</sup>. Protein information was added using NCBI Entrez Gene as primary source for protein (gene) names and identifiers (EGID)<sup>19</sup>. Where documents specified distinct alternative splice forms in assays, the common name used by the authors for that splice form was included with the EGID.

Target classes were assigned according to an internal schema. GVKBIO internally developed tools were also used to generate frameworks, scaffolds, and graph skeletons. The data was mined by running SQL queries against MCD and TGD subsets of the GOSTAR database. Additional filters were species, targets having an Entrez Gene name and assay type. Tables and graphs were generated in Excel.

## 13.4 Results and Discussion

The content statistics of the aggregated MCD and TGD sources, with combined and separate numbers for patents and journal papers, are shown in Table 1.

The following aspects can be expanded. The average redundancy (records-per-unique structure) is 1.5 because some compounds, particularly reference reagents and established drugs, have assay data included in many documents. The

<sup>15</sup> Complementarity Between Public and Commercial Databases: New Opportunities in Medicinal Chemistry Informatics

<sup>16</sup> Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds

<sup>17</sup> Curation of inhibitor-target data: process and impact on pathway analysis

<sup>18</sup> History of InChI

<sup>19</sup> Entrez Gene: gene-centered information at NCBI

predominant assay type is termed “type-B” or binding assay because it encompasses the enzyme inhibition and receptor binding assays most commonly reported for compounds tested against molecular targets *in vitro* and, implicitly, with binding specificity. The last three rows show the stringency used to define the final target listing. The target names in row 12 encompass both defined and undefined molecular targets (e.g. protein complexes or unresolved subfamilies) that are linked to compounds via a type-B assay result. These are further restricted in row 13 to only those molecular targets mapping to a protein identifier (e.g. an Entrez Gene ID or a Swiss-Prot accession). We added a final restriction to human sequences (row 14). We made this simplification choice for two reasons. The first was to exclude the many proteins used as cross-screening targets from mouse, rat and other model organisms. The second reason is that resolving anti-infective molecular target protein IDs also comes up against the problem of orthologous redundancy due to the multiplicity of viral, as well as bacterial sub-types, strains and species.

### 13.4.1 Counting Distinct Human Protein Targets

We used the MCD and TGD databases to compile a list of all human gene identifiers that were linked with compounds via the results of type-B assays. The full list of these protein sequence identifiers, compound counts and document counts is included in Additional file 2.

Additional file 1

**Additional material.** A list of proteins with names, symbols and Entrez Gene identifiers (Microsoft EXCEL). It also includes compound and document counts and the molecular framework breakdown for the compound sets.

[Click here for file](#)

To maximise the curatorial specificity of mapping compounds to protein sequences, a number of splice form designations are included where these names have been used in assays descriptions (mainly from journal papers in MCD). These cases produced 135 entries for 48 Entrez Gene IDs (EGIDs). While, in general, only small numbers of compounds are linked to these non-canonical protein sequences (i.e. alternative splice forms of the UniProt or RefSeq sequences corresponding to the EGID), these are important to capture for pharmacological differences. The human EGID total in Additional file

The summed number of compounds is 1,673,803. However, because the same compound may be assayed against different targets in the same or different documents, the unique set is 823,179. The average is 964 and the median 41 compounds-per-target. The top-278 proteins cover 90% of the total compounds at a cut-off of just over 1000 compounds-per-target. The summed number of documents is 53,440. The unique totals, 12,764 journal articles and 15,170 patents, are lower because of those that include results for more than one target. A subset of the top-50 targets with a cut-off just below 9,000 compounds-per-target is shown in Table 3. The binned distribution for the complete Additional file 4.

Inspection of our results indicated, not unexpectedly, a correlation between the number of compounds and number of documents. However, this was a very broad distribution because the extraction averages (given in Table 1) of 14 compounds-per-paper and 44 compounds-per-patent, varied by at least one order of magnitude for the former and two orders of magnitude for the latter. In the following section target proteins will thus be referred to by their rank on the basis of compounds. Those within the top-50 are listed in Table 2 while any below these in the ranking are listed in Additional file

### 13.4.2 Content of *bona fide* Drug Targets

Detailed elaboration of what constitutes a drug target is outside the scope of this work but this has been reviewed<sup>20</sup>. We, as do most descriptions for sources of this type, use the term “target” broadly to encompass any compound-to-protein mapping in our large dataset. We consider the target figures and divisions given by TTD to be a good approximation (they include a proportion of authenticated one-to-many mappings) to a set of *bona fide* primary targets (i.e. where the interaction *in vitro* is mechanistically causative for the therapeutic effect *in vivo*). It should be noted that, without inspection of the individual documents or “prior knowledge”, it is difficult to discriminate within database

---

<sup>20</sup> Drug target central

records *per se* between a *bona fide* drug target, a protein assay included for the purpose of discerning compound selectivity, off-target effects or modulating multiple targets with the same compound (i.e. polypharmacology)<sup>21</sup>. This classification problem is encountered for any large-scale collation of compound-to-protein mappings. It cannot be discerned clearly enough to be specified in the TCD database records because, while journal authors will typically explain the context and objectives of multiple assays, patent applicants often do not.

Nevertheless, it is clear from Table 2 that many of the top-50 proteins are not (yet) successful targets of approved drugs. A formal test was applied by determining the gene symbol intersect between Table 2 and the 185 targets of approved oral drugs from 2006<sup>2</sup>. Despite there being some new targets for post-2006 approved drugs the result was only 23 in common, indicating that a high compound ranking *per se*, is not necessarily a predictor of successful approval. The targets-in-common across the entire list were 160. Inspection of the 25 targets not matched indicated that, in most cases, the primary literature either included assay data from non-human proteins (e.g. mouse or rat) or that the cell-based receptor pharmacology assays were not classed as “type B”. One interesting exception is what could be classified as orphan target, tyrosine-3-hydroxylase, TH [Swiss-Prot

### 13.4.3 Cross-screening and Para-targets

The difficulty of discriminating primary targets from cross-screening activities is illustrated at the top of Table 2 for factor X, F10 [Swiss-Prot 2 is the cannabinoid receptors 1, CNR1 [Swiss-Prot<sup>22</sup>. In addition, CNRI provides an example of screening data derived from a specific splice variant with unique pharmacological profile, designated as cannabinoid receptor 1B [Swiss-Prot<sup>23</sup>. Other para-target pairs illustrate different aspects. For the beta amyloid cleaving enzymes BACE1 [Swiss-Prot<sup>24</sup>.

### 13.4.4 Anti-targets

The first anti-target (i.e. cross-screening for potential liabilities in development) in the list, ranked at 83, is the hERG Kv11.1 potassium channel, KCNH2 [Swiss-Prot<sup>25</sup>. Another anti-target is the drug efflux pump, ATP-binding cassette, sub-family B (MDR/TAP) member 1, ABCB1 [Swiss-Prot<sup>26</sup>.

### 13.4.5 Non-targets

The first non-target (i.e. without an established therapeutic context) is Trypsin, PRSS1 [Swiss-Prot

### 13.4.6 Failed Targets

Target names can be recognised in the list where compounds in Phase III trials have been publically declared as either having safety concerns or did not show efficacy. An example of the former, the cannabinoid receptor 1, CNR1 [Swiss-Prot<sup>27</sup>. During the initial drafting of this manuscript we selected the cholesterol ester transfer protein, CETP [Swiss-Prot<sup>28</sup>. However, within months, there was a more successful phase III outcome for anacetrapib (CID 11556427) targeting the same protein<sup>29</sup>. Thus, the extent to which late-stage failures constitute de-validation remains an open question, given not only that some of those targets can still “make it” but also that efficacy in a pharmacogenetically stratified cohort or repurposing for an alternative indication might still be achievable.

<sup>21</sup> Network pharmacology: the next paradigm in drug discovery

<sup>22</sup> Cannabinoid receptors as therapeutic targets

<sup>23</sup> Identification and characterisation of a novel splice variant of the human CB1 receptor

<sup>24</sup> The role of cathepsins in osteoporosis and arthritis: Rationale for the design of new therapeutics

<sup>25</sup> hERG (KCNH2 or Kv11.1) K Channels: Screening for Cardiac Arrhythmia Risk

<sup>26</sup> Transporter-Mediated Efflux Influences CNS Side Effects: ABCB1, from Antitarget to Target

<sup>27</sup> Efficacy and safety of the weight-loss drug rimonabant: a meta-analysis of randomised trials

<sup>28</sup> The end of the road for CETP inhibitors after torcetrapib?

<sup>29</sup> Safety of Anacetrapib in Patients with or at High Risk for Coronary Heart Disease

Nevertheless, the ability to flag likely de-validation in the listing we have produced would be valuable. However, the capture of historical data has the limitation that targets can achieve a high ranking if many compounds have been generated during validation and proof-of-concept studies even where these eventually fail. In addition, negative data produced during the research phase is less likely to be published. Our data can be analysed on a per-year basis, so the observation of a sustained decline in compounds (i.e. less publications on that target) can infer that validation has stalled (data not shown).

### 13.4.7 Tractability Assessment by Molecular Frameworks Analysis

The upper part of our compounds-per-target distribution (Table 2 and Additional file *de facto* chemical tractability ranking. The term is used here as a measure of the probability that a useful level of potency for chemical modulation of the therapeutically relevant biochemical activity of a protein can be readily achieved *in vitro*. While this is likely to be related to the HTS primary hit-rate, it must be remembered that a high proportion of the compounds in MCD and TGD have gone through some hit-to-lead optimisation. We thus choose to differentiate, on a target basis, between chemical tractability and druggability. We consider the latter to be the likelihood of developing compounds with appropriate *in vivo* bioavailability, efficacy and safety profiles<sup>30</sup>. These two characteristics are usually related because high chemical tractability facilitates the generation of more compound series *in vitro* which, in turn, provide more optimisation options *in vivo*. The main caveat with ranking targets just by compound numbers (as in Table 3) is that, in order to be useful, a tractability metric needs to factor-in the chemical diversity of the compound set. For example, targets mapped to large numbers of highly similar analogues might actually be less tractable than those with smaller absolute compound numbers but covering a broader range of chemotypes.

We have consequently exploited the compound listing to produce a detailed assessment of chemical diversity by comparing molecular frameworks and scaffolds on a per-target basis. These are well-developed concepts in medicinal chemistry and there are a number of ways in which chemical structures can be abstracted. An approach, initially described by Bemis and Murcko<sup>31</sup>, considers such frameworks as a collection of ring systems connected by linkers, after removing side chains. A more detailed hierarchy was used by Xu and Johnson<sup>32</sup> to define Molecular Equivalence Indices (MEQIs) as tools for molecular similarity measures. These approaches have been used for classifying and visualising compound collections<sup>33 34</sup>, scaffold-hopping<sup>35</sup>, comparing small sets of bioactive molecules<sup>36</sup> and large vendor libraries<sup>37</sup>, target selectivity<sup>38</sup> and to differentiate between drugs, clinical candidate and bioactive molecules<sup>39</sup>.

For our analysis we generated five levels of frameworks and scaffolds using software developed at GVKBIO:

1. Molecular Framework 1 (MF1): This is generated from the normalised molecular structure by removing all terminal side chains. Exocyclic double bonds (atoms connected to ring systems through multiple bonds) and double bonds directly attached to the linker are kept.
2. Molecular Framework 2 (MF2): This is derived from MF1 by removing exocyclic double bonds and double bonds directly attached to the linker.
3. Carbon Scaffold (CS): This is derived from MF2 by ignoring all atom types other than Carbon.
4. Atom Type Scaffold (ATS): Also derived from MF2 but ignoring bond types.
5. Graph Scaffold (GS): Also derived from MF2 but ignoring bond types or atom types.

<sup>30</sup> A practical view of 'druggability'

<sup>31</sup> The Properties of Known Drugs. 1. Molecular Frameworks

<sup>32</sup> Using Molecular Equivalence Numbers To Visually Explore Structural Features that Distinguish Chemical Libraries

<sup>33</sup> The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification

<sup>34</sup> Interactive exploration of chemical space with Scaffold Hunter

<sup>35</sup> A 3D similarity method for scaffold hopping from the known drugs or natural ligands to new chemotypes

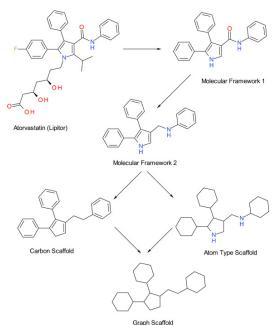
<sup>36</sup> Scaffold Distributions in Bioactive Molecules, Clinical Trials Compounds, and Drugs

<sup>37</sup> Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers

<sup>38</sup> Investigation of the Relationship between Topology and Selectivity for Druglike Molecules

<sup>39</sup> Molecular Topology Analysis of the Differences between Drugs, Clinical Candidate Compounds, and Bioactive Molecules

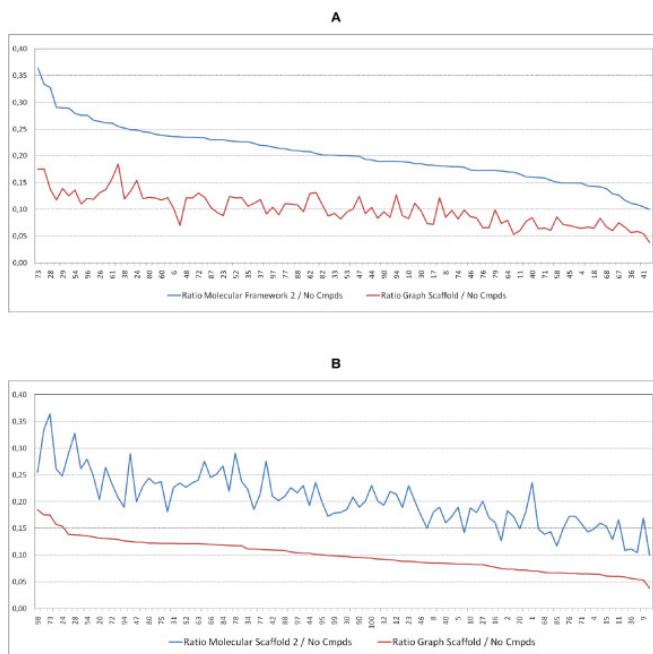
An example of the five levels of molecular topology hierarchy is shown for atorvastatin (CID 60823) in Figure 2. We applied these abstractions to the entire compound set, on a per-target basis, and the results are included in Additional file 5.



**Figure 13.2: Figure 2. The molecular topology hierarchy exemplified for Atorvastatin (Lipitor)**  
**The molecular topology hierarchy exemplified for Atorvastatin (Lipitor).**

We can see that the metalloprotease MMP1 drops from its original compound ranking at 11 down to 19 when ranked by MF2. The cathepsin CTSS moves in the opposite direction from 28 in the original ranking up to 7 by MF2. In the GS ranking we see the elastase ELNA rising from 49 to 20 but the kinase MAPK14 dropping from 4 to 14. Thus, for an individual target the tractability depends significantly on the molecular framework level used for ranking.

The MF2 level is particularly relevant for medicinal chemistry because it represents a practical scaffold level from which substituents can be permuted for the preparation of analogue series or compound libraries for SAR studies. For this reason, we have extended the analysis in Table 5 by plotting top-100 targets from Additional file 3.



**Figure 13.3: Figure 3. Sorted MF2 to number of compounds ratio (a) and Graph Scaffold to number of compounds ratio (b)**

**Sorted MF2 to number of compounds ratio (a) and Graph Scaffold to number of compounds ratio (b).** This is plotted for all targets with more than 4869 compounds from Additional file

More compounds with fewer MF2 scaffolds indicate lower tractability (e.g. an MF2: compound ratio of 0.13 for ESR2 from a total of 6,695 compounds). A larger ratio indicates higher tractability (e.g. 0.36 for HDAC1 from a total of 6,124 compounds). We suggest this complements the ranking by compounds alone and, in this case, clearly differentiates the relative rankings of 67 for ESR2 and 73 for HDAC1.

The molecular scaffold results can be conceived as collapsing the ensemble of structures mapped to a target in progressive stages of abstraction. Thus, moving from MF2 and GS we see a reduction as more compounds collapse into the latter. The target trends in Figure 3 are different for MF2 and GS. In addition, the spiked shape of the abstractions show these can be highly target-specific. As an example of utility, the visualisation of the chemotype landscape for targets with very large compound sets (e.g. over 10,000) is much easier when the GS ring-type abstractions can be displayed and browsed.

The utility of using public data for examining tractability before embarking on drug discovery project directed against targets and the correlation with ligand-based experimental assessments has recently been pointed out<sup>40</sup>.

## 13.5 Conclusions

We have triaged a commercial database to provide human target protein identifiers ranked by the numbers of compounds linked to them via direct biochemical assay data and the numbers of documents from which these associations were extracted. As far as we are aware, this is the largest published listing of this type and presents a detailed assessment of the major part of the human molecular target landscape that has been, or is, under active investigation<sup>41</sup>. The unique of scale of this is exemplified by comparing the equivalent compound-to-target count for F10 in ChEMBL of 5,871 against 42,869 in this work. This is because the process includes the extraction compounds and data not only from journal articles but also from patents.

Nevertheless, there are limitations (beyond our triage choices) that preclude this being a complete capture of the available data. The first is that in the PubChemBioAssay database, while the direct assay methods may have been published as documents, the compound structures, protein identifiers and result sets are only instantiated in *silico*<sup>4243</sup>. The second limitation is the necessity to cap the number of examples extracted from a patent. The third is that patent data extraction is currently limited to the “big ten” target classes and English language applications (but efforts are underway at GVKBIO to expand this). The fourth is journal selection as opposed to all journals. Whilst these pragmatic constraints may bias the extractions, we propose that, in SAR terms, they are selective for higher quality data.

Our complete set of results include many proteins that would not be considered *bona fide* drug target candidates, not only for the reasons already pointed out in the review of the list, but also by being in the tail of the compound distribution. However, the inclusion of even the singletons (one compound from one publication) is useful not only because they have been authenticated by expert extraction but also both the target and the compound may have a wider set of relationships using different species and/or assay type restrictions. Imposing any cut-off for “target likelihood” is clearly arbitrary but taking a lower limit of 20 compounds-per-target still covers just over 1000 proteins. This brings it into congruence with the data-supported target count of 836 human proteins for which moderately potent small-molecule chemical starting points had previously been reported<sup>44</sup>.

Our breakdown of the compound sets into molecular scaffolds provides a useful measure of target-specific chemical tractability. Nevertheless, we can point out factors that may be skewing the ranking upwards. The first is the cross-screening effect already mentioned where many compounds mapped to a target are not being optimised for that target. A second effect is that resources assigned to target projects are determined by factors such as market potential, competitive positioning and unmet clinical need. This skews the distribution away from an objectively neutral ranking of tractability *per se* towards those targets the research community is collectively “working hardest” on. This intense focus also produces patent thickets (in the sense that many of the synthetically feasible chemotypes and analogues that

<sup>40</sup> Fragment screening to predict druggability (ligandability) and lead discovery success

<sup>41</sup> Visualizing the drug target landscape

<sup>42</sup> An overview of the PubChem BioAssay resource

<sup>43</sup> PubChem as a public resource for drug discovery

<sup>44</sup> Global mapping of pharmacological space

can bind to a particular active site have already been claimed) that will also drive the expansion of chemical diversity for popular targets.

Readers are encouraged to explore their own additional analyses for Additional file

## 13.6 Endnotes

Protein designations first used in the text are given as their common name followed by the HGNC approved human gene symbol as used in the result tables. These are followed by the Swiss-Prot ID. Drug names are accompanied by their PubChem compound identifiers (CIDs).

## 13.7 Competing interests

AstraZeneca and GVKBIO have a business relationship.

## 13.8 Authors' contributions

The study was conceived by CS, SM and SARPJ. The data was generated by KB and the manuscript drafted by CS and SM. All authors read and approved the final manuscript.

## 13.9 Acknowledgements

We would like to thank Niklas Blomberg for his encouragement and perceptive reviewing of the manuscript.



# RESOURCE DESCRIPTION FRAMEWORK TECHNOLOGIES IN CHEMISTRY

## 14.1 Editorial

The Resource Description Framework (RDF) is providing the life sciences with new standards around data and knowledge management. The uptake in the life sciences is significantly higher than the uptake of the eXtensible Markup Language (XML) and even relational databases, as was recently shown by Splendiani et al.<sup>1</sup> Chemistry is adopting these methods too. For example, Murray-Rust and co-workers used RDF already in 2004 to distribute news items where chemical structures were embedded using RDF Site Summary 1.0<sup>2</sup>. Frey implemented a system which would now be referred to as an electronic lab notebook (ELN)<sup>3</sup>. The use of the SPARQL query language goes back to 2007 where it was used in a system to annotate crystal structures<sup>4</sup>.

The American Chemical Society (ACS) Division of Chemical Information (CINF) invited scientists from around the world to present their use of RDF technologies in chemistry on 22nd-23rd August 2010 at the 240th ACS National Meeting in Boston, USA. During three half-day sessions, the speakers demonstrated a mix of smaller and larger initiatives where RDF and related technologies are used in cheminformatics and bioinformatics as Open Standards for data exchange, common languages (ontologies), and problem solving. The fifteen presentations were grouped in the themes computation, ontologies, and chemical applications. Figures 1, 2 and 3 display the most important keywords reflecting the abstracts of the talks in each session as word clouds<sup>5</sup>.

The goal of the meeting was to make more chemists aware of what the RDF Open Standard has to offer to chemistry. We are delighted to continue this effort with this Thematic Series, for which the speakers (and others) were invited to present their work in more detail to a wider chemistry community. The choice of an Open Access journal follows this goal. At this place, we would like to thank Pfizer, Inc., who had partially funded the article processing charges for this Thematic Series. Pfizer, Inc. has had no input into the content of the publication or the articles themselves. All articles in the series were independently prepared by the authors and were subjected to the journal's standard peer review process.

In the remainder of this editorial, we will briefly outline the various RDF technologies and how they have been used in chemistry so far.

---

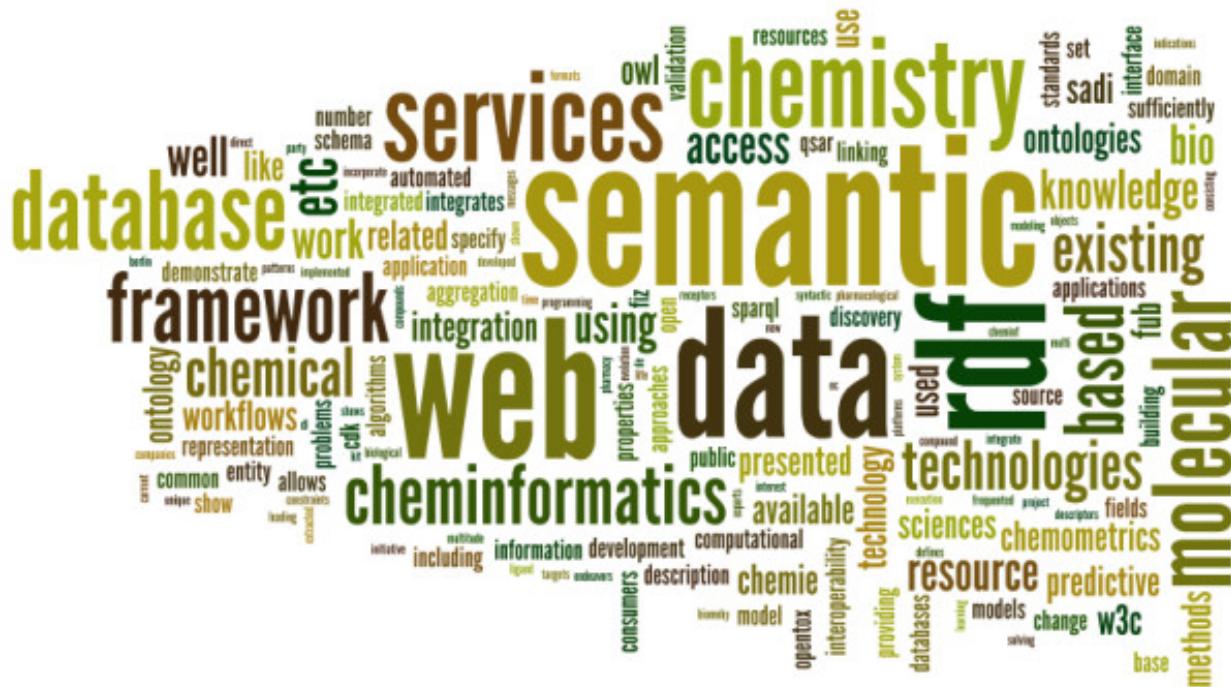
<sup>1</sup> Biomedical semantics in the Semantic Web

<sup>2</sup> Chemical markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators

<sup>3</sup> Dark Lab or Smart Lab: The Challenges for 21st Century Laboratory Software

<sup>4</sup> Collaborative Annotation of 3D Crystallographic Models

<sup>5</sup> Wordle



**Figure 14.1: Figure 1. Keyword cloud for the RDF and Computation session**



**Figure 14.2: Figure 2. Keyword cloud for the RDF and Ontologies session**



**Keyword cloud for the RDF and Chemical Applications session.**

## 14.2 1 Concepts

The core RDF specification was introduced by the World Wide Web Consortium (W3C) in 1999<sup>6</sup> and defines the foundation of the RDF technologies. It has evolved into a set of recommendations by the W3C published in 2004 (See Table 1). RDF specifies a very simple data structure linking a subject to an object or a value (literal) using a predicate. Cheminformaticians will recognize this data structure as an edge from graph theory. This structure allows us to represent facts like “vanillin dissolves in methyl alcohol”<sup>7</sup>. RDF uses Uniform Resource Identifiers (URIs) to identify things. Therefore, the RDF equivalent of the solution statement could be like this so-called triple:

<http://dbpedia.org/resource/Vanillin> <http://example.com/dissolvesIn> <http://dbpedia.org/resource/Methanol>.

Since URIs may be used to reference resources on any server worldwide, RDF triples allow to span a global graph data structure. This is not surprising, since RDF is the core technology behind the proposed Semantic Web<sup>8</sup>. In fact, the Web nature is clear here, as one can follow both the URIs for vanillin and methanol to obtain further information on those two chemicals. These molecules' URIs are said to be dereferencable, allowing agents to spider the Web for information following the hyperlinks, quite like how you follow hyperlinks on websites. Hence, the term Semantic Web.

Recent projects such as Bio2RDF<sup>9</sup>, Chem2Bio2RDF<sup>10</sup>, and OpenTox<sup>11</sup> have brought genomic, chemical and pharmaceutical knowledge to the Semantic Web by expressing it in RDF. These three projects aim at making databases with chemical knowledge available from a central access point, interlinking the individual data sets. Smaller data sets are also becoming available as RDF, such as the Open Notebook Science Solubility data<sup>12</sup>.

<sup>6</sup> Resource Description Framework(RDF) Model and Syntax Specification

<sup>7</sup> Open Notebook Science Challenge: Solubilities of Organic Compounds in Organic Solvents

8 The Semantic Web

<sup>9</sup> Bio2RDF: Towards a mashup to build bioinformatics knowledge systems

<sup>10</sup> Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data

<sup>11</sup> Collaborative development of predictive toxicology applications

12 NOTITLE!

## 14.3 2 Formats

The actual use of RDF depends on various further standards. For example, standards were required that describe how RDF statements are exchanged. Several standards serve this purpose: RDF/XML is an XML-based serialization<sup>13</sup>, while simpler formats exist with N-Triples<sup>14</sup> and Notation3<sup>15</sup>. For integration with current web practices, RDFa has been defined to allow RDF triples to be embedded in HTML pages<sup>16</sup>. Additionally, a proposal has been written that describes how RDF can be serialized as Javascript Object Notation (JSON)<sup>17</sup>, and while this is not a formal specification yet, a new RDF working group will formalize this into a new standard<sup>18</sup>. Several of these serialization standards are used in the papers in this Series.

Using these serializations, RDF can be downloaded directly from pure RDF documents (RDF/XML, Notation3), or extracted from RDFa-based web pages using online RDF extraction web services, like <http://www.w3.org/2007/08/pyRdfa/>. These approaches make it simple to aggregate chemical data from web pages.

## 14.4 3 Querying the World Wide Web

The most promising technology in the RDF family is the SPARQL Protocol and RDF Query Language (SPARQL)<sup>19</sup>, which has been applied by Chen et al. in three chemogenomics use cases<sup>10</sup>. One of the use cases shows how SPARQL queries are used to find compounds that are active in bioassay for genes related to proteins to which the chemical dexamethasone binds, using information from PubChem, Uniprot, and DrugBank, all made available as RDF in the Chem2Bio2RDF database. The other use cases in this paper use the same approach by aggregating data sources before querying them. As such, it is similar to querying data stored in a relational database. However, an important difference between SPARQL and SQL query engines is the underlying data they act on: a graph of triples for RDF data, and rectangular tables in relational databases. This difference implies that RDF resources must have at least some common elements, whereas a relational DBMS assumes an identical data structure for all records of a table.

For example, Jankowski used a public SPARQL service to extract boiling points of a series of alkanes from an XHTML webpage with the data made machine readable with RDFa, and visualized that using Javascript in another web page dynamically<sup>20</sup> (see Figure 4). A second important difference is that SPARQL queries can be *federated*<sup>21</sup>. Federated SPARQL allows one to query various RDF providers in one query. This has been used recently in the Receptor Explorer tool to help translational research by connecting basic neuroscience research with clinical trials<sup>22</sup>. Being able to query resources in this manner, brings us a step closer to systems biology approaches.

## 14.5 4 Ontologies

With RDF we have a data structure to link resources and provide details about those resources, and SPARQL provides us with the tools to query and aggregate that data. The next standard we will discuss now is the Web Ontology Language (OWL) which brought the RDF technology to the ontology community<sup>24</sup>. Ontologies are most certainly not new to chemistry<sup>25</sup> nor biology or life sciences, but the OWL standard makes it much easier to use ontologies,

<sup>13</sup> RDF/XML Syntax Specification (Revised)

<sup>14</sup> RDF Test Cases

<sup>15</sup> Notation 3 - An readable language for data on the Web

<sup>16</sup> RDFa in XHTML: Syntax and Processing

<sup>17</sup> RDF in JSON

<sup>18</sup> New RDF Working Group, RDF/JSON, RDF API...

<sup>19</sup> SPARQL Query Language for RDF

<sup>20</sup> SPARQL to Chart

<sup>21</sup> Federated SPARQL

<sup>22</sup> A journey to Semantic Web query federation in the life sciences

<sup>24</sup> OWL Web Ontology Language Overview

<sup>25</sup> Chemical inference. 3. Formalization of the language of relational chemistry: ontology and algebra

```
PREFIX cc: <http://github.com/egonw/cheminformatics.classics/1/#>
SELECT *
FROM
<http://www.w3.org/2007/08/pyRdfa/extract?
uri=http%3A%2F%2Fegonw.github.com%2Fcheminformatics.classics%2Fclassic1.html&forma
t=pretty-xml&warnings=false&parser=lax&space-preserve=true>
{
    ?mol cc:p0 ?p ;
        cc:t0 ?t .
}
```

Plot Results

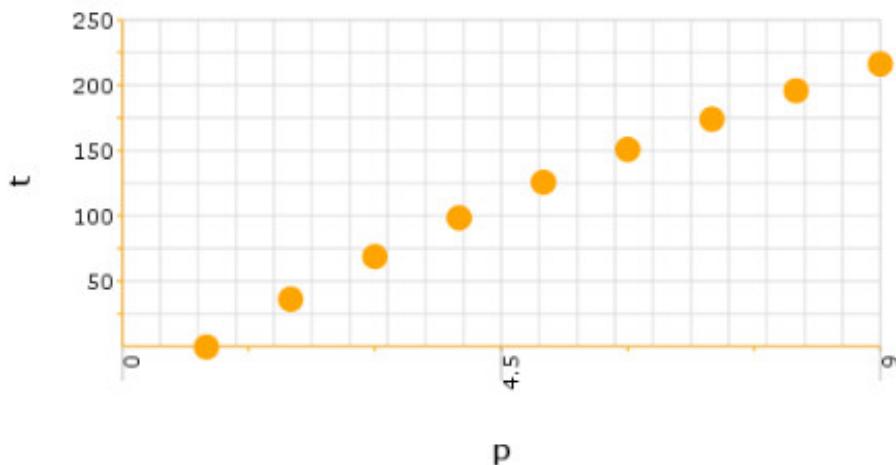


Figure 14.4: Figure 4. Web page using SPARQL to visualize alkane boiling points extracted from another web page  
**Web page using SPARQL to visualize alkane boiling points extracted from another web page.** Web page with JavaScript by Jankowski visualizing the boiling point of a series of alkanes from Wiener<sup>23</sup> extracted with SPARQL from a second, XHTML+RDFa web page at <http://egonw.github.com/cheminformatics.classics/classic1.html>

partly because they are formulated in RDF themselves. Ontologies, like controlled vocabularies and thesauri, describe what things mean, by linking terms to a human-readable definition. As such, ontologies are used for sharing knowledge in a common language, as well as to organize that knowledge. While linking resources is not new either, expressing the content of resources in explicit terms allows humans and software to reason formally on the content and to find possible sources of error. For example, Konyk et al. have used OWL to link PubChem, DrugBank, and DBpedia, noting that it offers new ways to discover knowledge<sup>26</sup>.

There are currently not many ontologies in chemistry, but many OBO Foundry-based ontologies can be reused using an OBO to OWL mapping<sup>27</sup>. This makes available chemical ontologies like the CO ontology<sup>28</sup>, the ontology of Chemical Entities of Biological Interest (ChEBI)<sup>2930</sup>, and the Chemical Information Ontology <http://code.google.com/p/semanticchemistry/>, but also other ontologies in the life sciences, such as the Gene Ontology<sup>31</sup>. This way, OWL provides a universal standard to link data sources in life sciences, transcending traditional boundaries between the various domains.

The current state is that different RDF resources are using different ontologies. This does not necessarily have to be a problem, because the ontologies can be explicitly mapped to each other. This way, equivalent terms from two ontologies can be formally defined as equivalent, using the OWL predicates owl:equivalentClass and owl:equivalentProperty for classes, and owl:sameAs for instance. Making the equivalence explicit this way helps to illustrate the provenance of data integration efforts.

## 14.6 5 Discussion

This Thematic Series shows the current state of the use of RDF in chemistry, as presented at the ACS RDF 2010 meeting in Boston, and provides an insight into the progress of these methods. Much of the research is currently explorative, rather than formative, though standards are being proposed. It may very well turn out that some aspects of chemistry will never be expressed in RDF, and some computation will be done without ontology-based reasoning. It is important to realize here where RDF is positioned, namely for linking resources.

However, the use of RDF for already well-defined data structures in chemistry is not obvious. Data types like connection tables and various matrices are possible, but the use of URIs makes such structures needlessly verbose. Moreover, there is no need to format already well-formalized data structures into RDF, such as the various uses of matrices in computational chemistry as RDF triples. In fact, several papers in this series outline how to combine knowledge expressed with RDF with computational services. This shows that RDF is not an isolated framework, but one that can be integrated into existing cheminformatics workflows.

What RDF does not solve, are the following issues that remain in cheminformatics. RDF is about knowledge representation, and while ontologies take care of meaning and provide requirements to verify formal data consistency, it does not enforce any data quality, data structure, or data availability. This is, in fact, similar to other ways of providing data. For example, a data set with boiling points may or may not include information about experimental error. Metabolomics data may name the molecules for which concentration profiles have been measured, or the original accurate masses from which the identity was deduced.

It must be clear, therefore, that the RDF technologies are not the solution to everything. Their use does not guarantee an impressive scientific scenario. Instead, it can help simplify data analysis and particularly data integration, making it easier to handle large volumes of data accurately, or at least, with an explicitly defined accuracy.

As such, the use of explicit, semantic formats can be considered a gold standard of scientific practise. It is about adding as much detail to your lab notebook as you need. But, it does not inhibit you from writing nonsense in your notebook.

---

<sup>26</sup> Chemical Knowledge for the Semantic Web

<sup>27</sup> OBO to OWL: a protege OWL tab to read/save OBO ontologies

<sup>28</sup> CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules

<sup>29</sup> ChEBI: a database and ontology for chemical entities of biological interest

<sup>30</sup> GO faster ChEBI with Reasonable Biochemistry

<sup>31</sup> Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL

## 14.7 6 Outlook

The future of the use of RDF technologies as open standards in chemistry looks bright, and fills the needs in chemistry for semantically linking chemical data to other data sources. RDF technologies provide a domain-independent way for representing knowledge and their open nature assures many alternative approaches for making data available as RDF. This Thematic Series shows a few novel and creative applications of these RDF technologies, and we hope they may serve as seminal work in cheminformatics for future years.



# SEMANTIC WEB INTEGRATION OF CHEMINFORMATICS RESOURCES WITH THE SADI FRAMEWORK

## 15.1 Abstract

### 15.1.1 Background

The diversity and the largely independent nature of chemical research efforts over the past half century are, most likely, the major contributors to the current poor state of chemical computational resource and database interoperability. While open software for chemical format interconversion and database entry cross-linking have partially addressed database interoperability, computational resource integration is hindered by the great diversity of software interfaces, languages, access methods, and platforms, among others. This has, in turn, translated into limited reproducibility of computational experiments and the need for application-specific computational workflow construction and semi-automated enactment by human experts, especially where emerging interdisciplinary fields, such as systems chemistry, are pursued. Fortunately, the advent of the Semantic Web, and the very recent introduction of RESTful Semantic Web Services (SWS) may present an opportunity to integrate all of the existing computational and database resources in chemistry into a machine-understandable, unified system that draws on the entirety of the Semantic Web.

### 15.1.2 Results

We have created a prototype framework of Semantic Automated Discovery and Integration (SADI) framework SWS that exposes the QSAR descriptor functionality of the Chemistry Development Kit. Since each of these services has formal ontology-defined input and output classes, and each service consumes and produces RDF graphs, clients can automatically reason about the services and available reference information necessary to complete a given overall computational task specified through a simple SPARQL query. We demonstrate this capability by carrying out QSAR analysis backed by a simple formal ontology to determine whether a given molecule is drug-like. Further, we discuss parameter-based control over the execution of SADI SWS. Finally, we demonstrate the value of computational resource envelopment as SADI services through service reuse and ease of integration of computational functionality into formal ontologies.

### 15.1.3 Conclusions

The work we present here may trigger a major paradigm shift in the distribution of computational resources in chemistry. We conclude that envelopment of chemical computational resources as SADI SWS facilitates interdisciplinary

research by enabling the definition of computational problems in terms of ontologies and formal logical statements instead of cumbersome and application-specific tasks and workflows.

## 15.2 Background

The introduction and subsequent widespread availability of computers in the latter half of the 20<sup>th</sup> century has had an enormous impact on chemistry and related sciences. A wide range of problems which could only be addressed by tedious manual or semi-automated computation a few decades prior suddenly became readily accessible with computers. The explosion of the diversity of the various software packages addressing every aspect of chemistry that followed can only be compared, in relative terms, to the Cambrian explosion in species diversity. Myriads of file formats, programming languages, platforms, operating systems, programming paradigms, distribution models, and access methods have been employed in hundreds of largely-independent projects, each vying for widespread adoption and often offering a unique set of functionalities and features to target a specific subdomain or application of chemistry. Consequently, computational life scientists are now obliged to spend considerable efforts on software package integration to make any progress in their daily investigations.

This problem has been especially acute for interdisciplinary studies, perhaps rising in relevance and importance with the relatively recent rise of Systems Science to prominence. For instance, to build a simple ordinary differential equation-based model of a system of partially enzyme-catalysed reactions, one may need to generate the three-dimensional structures of involved molecules, compute their energies of formation and solvation, approximate pKa values, evaluate enzyme interactions, predict kinetics, and finally solve kinetic equations, all in different software packages, which might be located on different operating systems or have unique shamanic execution procedures known only to the high priests of these packages. Even practitioners of narrower specialities are not spared the wrath of software integration, albeit on a smaller scale. Although tools to interconvert the output and input files for many of these software components have been developed<sup>1,2</sup>, and although a number of chemical packages offer access to their functionalities through programming interfaces (e.g.<sup>3,4,5</sup>), one is left wishing that researchers in chemistry-related fields could still do more science rather than pipelining.

As scientific publishing accelerates and high-throughput experimentation platforms become increasingly pervasive, the problem of integration of the disparate computational, literature, and experimental resources is transformed from that of removing a daily nuisance to that of finding a solution without which science cannot move forward effectively. With the introduction of computational web services for life sciences (e.g.<sup>6,7</sup>), a step in the direction of addressing this problem has been made. With web services, tasks can be posted directly to computational resources with job execution instructions that conform to service-specific schemas, usually defined with a standard specification, most prominently the Web Service Definition Language (WSDL)<sup>8</sup>. Given sufficient knowledge of the service schemas, it is technically possible to automate workflow construction and provide seamless integration of web service components to fulfil a greater overarching task. In practice, however, the lack of shared and consistent schema elements with *formal semantics* has severely limited this integrative potential due to difficulties of automatically integrating service schemas themselves.

The next step of the evolution of web services was reached with the adoption of the Semantic Web and the corresponding development of Semantic Web Technologies to enable not only machine-understandable knowledge representation, but also the exposition of this knowledge and underlying concepts to automated, formal logic-based reasoning. Given a collection of knowledge triples that utilize types and relations from a formal ontology, it has finally become possible to automatically classify, integrate, and interconnect entities and concepts, much like a human expert would. To truly capitalize on this potential, simple XML-based approaches in resource specification and annotation would have

---

<sup>1</sup> The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics

<sup>2</sup> The Blue Obelisk - Interoperability in Chemical Informatics

<sup>3</sup> Predictive models for carcinogenicity and mutagenicity: frameworks, state-of-the-art, and perspectives

<sup>4</sup> JOELib Java-based cheminformatics library

<sup>5</sup> RDKit Cheminformatics Package

<sup>6</sup> Semi-automatic web service composition for the life sciences using the BioMoby semantic web framework

<sup>7</sup> CDK-Taverna: an open workflow environment for cheminformatics

<sup>8</sup> Web Service Description Language Specification

to make way for Resource Description Framework (RDF)<sup>9</sup> and Web Ontology Language (OWL)<sup>10</sup> to enable automated integration of static knowledge resources with computationally generated information and provide results for cross-domain queries. This ability is indispensable in the life sciences domain to address interdisciplinary problems in toxicology or metabolism, for example. For such problems, it is not only often the case that no single database contains all the information necessary to build a working model or formulate trustworthy predictions, but it is also true that much database information is fragmented and often incomplete. Some of this missing information could be computed to fill in the gaps preventing integrative model construction, but relevant computational resources, many of which are web services, remain inaccessible to a single query method, partly due to the aforementioned integration issues. Although large collections of chemical data have recently become represented in RDF and exposed to SPARQL querying<sup>11 12 13</sup>, seamless and facile integration of computational resource output to enable query completion has been difficult to attain with currently existing technologies.

Early solutions proposed for automated service integration in life sciences often drew on elements of Semantic Web Technologies. With Semantic Annotations for WSDL and XML Schema (SAWSLD), it has become possible to annotate WSDL documents with terms from formal ontologies, in a process termed ‘lifting’ to enable a greater degree of resource integration than with simple XML service specifications<sup>14</sup>. Services thus annotated would then be more readily available to integration from a central service registry. However, tight integration into the Semantic Web that would allow natural formal reasoning over such services has not yet been adequately addressed, with SAWSLD and WSDL services often requiring adaptors for service integration. The Web Service Modelling Ontology (WSMO), on the other hand, aims to construct and support a complex framework, implemented in Web Service Modelling eXecution environment (WSMX), whereby web services and all aspects of their behaviour are formally semantically represented<sup>15</sup>.

Within the life sciences web service domain, numerous practical solutions to web service integration have been proposed, but are too numerous to completely discuss here<sup>16</sup>. Often, these solutions focussed on the construction of a common domain ontology, vocabulary, or registry to improve service annotation and discovery in a given domain (e.g.<sup>17 18</sup>). More recently, service frameworks that relied on more general ontologies for service annotation and input/output specification have also become available<sup>19 20</sup>. Further, frameworks integrating REST service discovery and ontology-assisted workflow composition have also appeared recently<sup>21</sup>. Implicit in these approaches has been the need to adopt and adhere to common service annotation, input and output specifications, or common domain ontologies. This has meant that although some of these SWS initiatives have enjoyed considerable success in enabling service interoperability and supporting facile manual workflow composition, truly seamless and automated service integration into the Semantic Web was not reached, as it was inhibited by semantic service platforms themselves.

This situation has changed with the recent introduction of the SADI framework<sup>22</sup>. Web services created with SADI consume and produce RDF graphs, operating on instances of input and output classes formally defined in supporting service OWL ontologies. The input class of a SADI service subsumes the output class, as these services are stateless, atomic, and annotative. That is, each service carries out a single primitive function and annotates an instance of the input class with information through a particular predicate. SADI services are also REST-like, in that there is only a standard basic set of HTTP verbs that they may respond to, namely GET and POST. A GET operation on a given service returns its semantic description, while a POST of a well-formed RDF graph to the service initiates service execution and returns the same RDF graph with the annotations created by the service. If a SADI service is computationally-intensive, standard asynchronous execution mechanisms are available.

<sup>9</sup> Resource Description Framework Specification

<sup>10</sup> Web Ontology Language Specification

<sup>11</sup> Bio2RDF: towards a mashup to build bioinformatics knowledge systems

<sup>12</sup> Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data

<sup>13</sup> Linking Open Drug Data Project

<sup>14</sup> Semantic Annotations for WSDL and XML Schema Specification

<sup>15</sup> Semantically-enabled Service Oriented Architecture: Concepts, Technology and Application

<sup>16</sup> Evolution of web services in bioinformatics

<sup>17</sup> TAMBIS: transparent access to multiple bioinformatics information sources

<sup>18</sup> Feta: A Light-Weight Architecture for User Oriented Semantic Service Discovery

<sup>19</sup> BioMOBY: an open source biological web services proposal

<sup>20</sup> SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services

<sup>21</sup> Data Mining Workflow Templates for Intelligent Discovery Assistance and Auto-Experimentation

<sup>22</sup> SADI SemanticWeb Services - ‘cause you can’t always GET what you want!

Unlike its aforementioned predecessors, SADI service specification is extremely simple as it neither imposes nor invents a central schema, ontology, or message structure, using standard web components instead. Because of the formal logical definition of the input class, output class, and the introduced predicate, SADI services can be tightly integrated into the Semantic Web and very naturally reasoned about by a machine client. This combination of simplicity of specification and power of formal reasoning allows SADI web services to be seamlessly integrated into SPARQL queries with simple machine reasoning clients, as if the data that they can potentially generate was already available in an RDF triple store. One such prototype client, Semantic Health And Research Environment (SHARE), operates on SPARQL queries and is capable reasoning about the desired overall query goal and chaining services and information together such as to reach this goal in the least computationally expensive way<sup>23</sup>. SHARE draws on a central freely accessible service registry that contains information about service input, output, and predicate types to carry out this automated workflow construction. Thus, in order to pose a query through SHARE, a human agent has to be aware of service specification details in the registry to be able to create a well-formed query. For this, the user needs to acquaint themselves with the input and output classes operated upon and annotations created by the collection of SADI services available in a given SADI instance by perusing the service registry<sup>24</sup>. Thus, one only needs to identify an input class that contains the information that is already available as a starting point, as well as the service-specific predicates corresponding to the annotation that is desired.

In order to maximise service interoperability, it is of course recommended to reuse concepts and classes as much as possible by adhering to upper-level domain ontologies, but this is not a requirement in SADI and concepts can be manually mapped to those appearing in supporting service ontologies, if required. By the virtue of allowing external formal ontologies to be referenced in SPARQL queries executed by SHARE, this approach provides support for discourse in science while disambiguating and explicitly highlighting the points of disagreement. For example, a small molecule according to one researcher may be one that is no heavier than 500 Daltons, while another may insist that number to be 750 Daltons. With OWL, both of these viewpoints may be represented with an explicit specification which can then be used to classify a set of molecules automatically using the same SADI services, and filter down into further assertions and reasoning seamlessly. Further, multiple researchers may construct ontologies that may model molecules, and consequently *smilesmolecule* differently. So long as a formal logical mapping can be either inferred or directly made to concepts used by the service ontologies, this difference can be accommodated and the construction of computational workflows may be initiated. In other words, SADI enables liberation from conformity in existing consortia-generated ontologies and facilitates discourse and disagreement which drive science forward, at the cost of making the end user aware of the ontologies used by the existing SADI services.

SADI, with supporting machine reasoning clients, also enables the conversion of ontologies into workflows. As we have seen, a formal definition of a small molecule as having a molecular mass descriptor within a particular range in an external ontology will trigger the execution of an appropriate service through a SPARQL query posted to SHARE, if no such data is already available, and the input and output classes the service operates upon are consistent with the external ontology. Therefore, integrative workflows (in e.g. molecular classification) can be constructed just by redefining the problem in terms of classifying a given entity into a particular formal ontology-defined class or a set of classes of interest in a given study (Figure 1). Furthermore, because computational tasks are explicitly specified and service invocation is controlled, SADI allows for a greater reproducibility and interoperability of computational analysis.

In this work, we shall describe and discuss the exposition of a range of computational chemical resources as SADI services and their resultant amenability to seamless automated semantic integration through ontologies, SPARQL queries, and with graphical interfaces.

---

<sup>23</sup> SHARE: A Semantic Web Query Engine for Bioinformatics

<sup>24</sup> SADI Service Registry

Figure 15.1: Figure 1. Seamless service integration into the Semantic Web with SPARQL queries over RDF-encoded resources in chemistry, as enabled by SADI

**Seamless service integration into the Semantic Web with SPARQL queries over RDF-encoded resources in chemistry, as enabled by SADI.**

## 15.3 Results and Discussion

### 15.3.1 Exposing CDK QSAR Functionality with SADI

The exposure of computational functionality of a particular software package or application programming interface begins with the isolation of the smallest accessible functional units of the software at hand. At the most detailed level, this process may be likened to decomposition of an API into its constituent classes, which may often become input classes in the supporting service ontology, and their corresponding methods, which may be viewed as the actual computational functionality of these services (Figure 2). This comparison is limited and highly simplified, but it captures the general essence of what we are trying to do. If one concerns themselves solely with the primary *functionalities* of a given piece of software relevant to a particular problem, individual services may envelop more than one basic method in a given API or software. One limitation that arises as a consequence of the annotative nature of SADI services is the requirement to support transformative functionalities by differentiating the input and output, even if both are of the same class in the software package. For example, a method that removes hydrogen atoms from a given input molecule specified by a SMILES string and returns a SMILES string (or void), would have to be converted into a SADI service that operates upon an input class that consists of molecules that have a SMILES string specified and produces output typed to a class that consists of molecules that are annotated with a SMILES string as well as a hydrogen-free SMILES string. It must be noted that SADI services do not *have* to return all of the information present in the specification of the annotated entity that is a member of the input class. For a member of the output class of a service whose input is a molecule that has a SMILES string, it is possible to introduce only the new service-generated annotations on the existing input entity (present at a dereferenceable URI) in the output. Thus, in our hydrogen depletion example, only a reference to the input entity and the new annotation need to be included in the output, without the need to include all of the inherited entity attributes.

In this study, we have exposed a portion of QSAR descriptor calculating *functionality*, to calculate descriptors starting from a SMILES string molecular specification, as implemented in the Chemistry Development Kit. Because we are primarily concerned with demonstrating the envelopment of *functionalities* of CDK in this work, we have chosen to distribute the calculation of every available descriptor as a separate web service, even if such services relied on multiple methods and classes in CDK (Table 1).

The second step in service creation is formal description of the input and output classes in a service ontology. For our whole set of QSAR descriptor services, we have created a single service ontology <sup>25</sup>, extended from the CHEMINF

<sup>25</sup> Lipinski Service Ontology

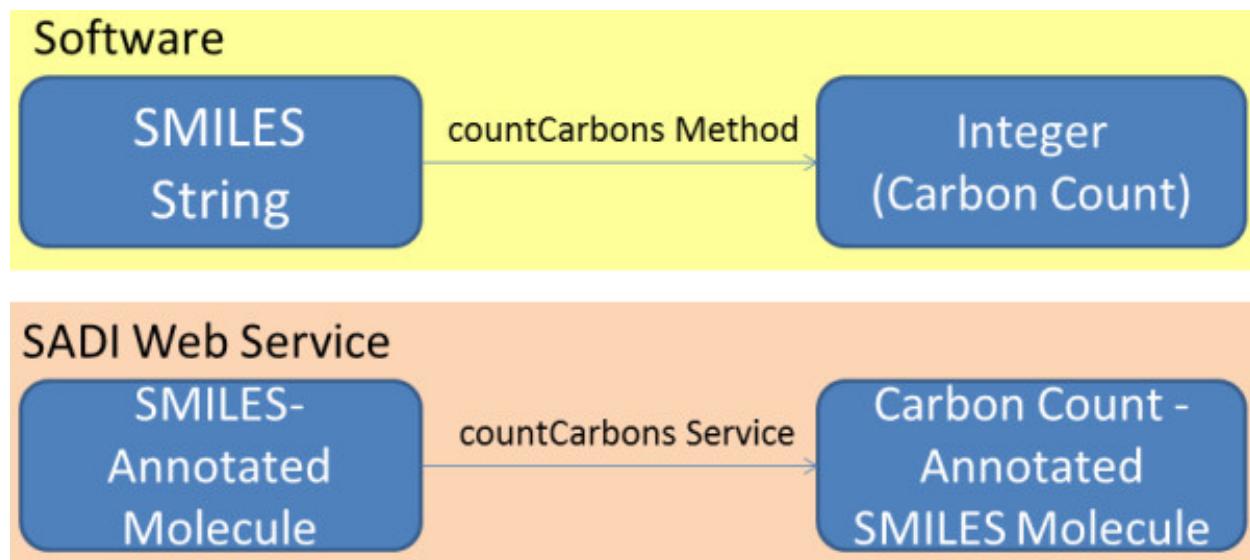


Figure 15.2: Figure 2. In principle, classes and methods in APIs can, with some adjustments, be converted to input/output classes and functionality-encapsulating services

**In principle, classes and methods in APIs can, with some adjustments, be converted to input/output classes and functionality-encapsulating services.** Note that since SADI services are annotative, the input class subsumes the output class which merely contains the extra annotation computed by the service.

ontology<sup>26</sup> that contains concepts relevant to formal specification of chemical information in general and descriptor information in particular. The reuse of concepts and relations from widely accepted higher-level ontologies in CHEMINF translates into greater integration of service input and output classes into cross-domain queries.

For each descriptor calculating service, the input class is a *smilesmolecule* which is formally defined as the following.  
*molecule and 'has attribute' some ('SMILES descriptor' and 'has value' some string)*

This input specification assures that the service will receive and operate upon an entity of the type molecule that has a SMILES descriptor and that this descriptor has a string value which the service can parse, transform into a molecular graph, and for which it can carry out descriptor calculations with a given API, in this case CDK. Note that the terms *molecule*, *has attribute*, and *has value* are reused from upper-level ontologies, meaning that concepts introduced in third-party ontologies constructed using the same upper-level ontologies will be much easier to integrate with than if we were to invent our own terms. The input is an RDF-XML graph submitted to the service through a simple HTTP POST to the service URL (Listing 1).

Listing 1. A fragment of the RDF input graph for the CDK descriptor services, in N3 form.

<http://semanticscience.org/>.

<http://semanticscience.org/resource/>.

Because the input class has to subsume the output class, the output entity has to have all the features of the input entity, but service-computed annotations should decorate the entity in the output. For instance, the definition of the output class *bondcounts smilesmolecule* for a bond count descriptor calculating service<sup>27</sup> is as follows.

*smilesmolecule and 'has attribute' some ('bond count' and 'has value' some int)*

This class definition specifies that in the output, a given *smilesmolecule* instance will be annotated with a bond count descriptor which would have an integer value (Listing 2).

<sup>26</sup> CHEMINF Ontology

<sup>27</sup> Bond Count Descriptor Service

Listing 2. A fragment of the RDF output graph produced by the CDK bond count calculator service, converted to N3 RDF form.

```
http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl#.
http://semanticscience.org/resource/.
http://www.w3.org/1999/02/22-rdf-syntax-ns#.
http://www.w3.org/2001/XMLSchema#int.
```

The service OWL ontology, containing these input and output class specifications, along with the relevant predicate (*has attribute*) specifications, has to be made distributed such as to assure that these resources have dereferenceable URIs and that the ontology itself is readily available for machine reasoning agents. In order to expose services for invocation and automated workflow composition with the SHARE client, one needs to also register the service on the central SADI registry. Descriptor information can then be obtained by submitting SPARQL queries that are no different from queries over triple stores that are already populated with RDF knowledge, to the SHARE client. In essence, we are seamlessly querying *all* the data at our disposal, even the knowledge that does not yet exist, but can be generated. For example, to determine the number of hydrogen bond donors in a given molecule, one may submit the following query to the SHARE client (Listing 3).

Listing 3. A sample SPARQL query to determine the value (specified by the ?value variable) of the hydrogen bond donor count descriptor for a molecule specified in the given (lipinski\_test) RDF graph, submitted to a SHARE client<sup>28</sup>.

```
http://www.w3.org/1999/02/22-rdf-syntax-ns#.
http://semanticscience.org/resource/.
http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl#.
http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl.
http://semanticscience.org/sadi/ontology/lipinski_test.rdf.
```

This query returns the identity of the input molecule (whose SMILES descriptor is specified in the lipinski\_test RDF graph), along with the value of its corresponding logP descriptor. In the background, SHARE reasons about the available services based on the information requested in the SPARQL query, *as well as the information already available in the input graph*, and automatically matches the appropriate service or services to the request. Thanks to the formal reasoning carried out by the SHARE client using the service-specific ontologies and any other ontologies referenced in the query, it is possible to infer the services needed to carry out a particular task even if the request does not use concepts identical to those found in the service definition. For example, if a predicate *hatChemischeDeskriptor* can be inferred to be equivalent to *hasChemicalDescriptor* through its formal axiomatic definition, the same set of services shall be called to fulfil queries using either predicate. Thus, the integration of SADI services into the Semantic Web by means of integration into SPARQL queries with SHARE is seamless and requires no additional programming on the part of the life science researcher.

### 15.3.2 SADI-Enabled Format Interconversion and Software Interfacing

Though ubiquitous in chemical databases, SMILES strings do not address every chemical entity specification need. For example, one may be interested in the three-dimensional configuration of a given molecule directly, or in a more standard and canonical way of representing chemical graph structure with InChI strings<sup>29</sup>. Conversely, the SMILES string needed for our services to operate may not be present, but an InChI descriptor may be available instead. Finally, disparate services may operate on different formats, such that one may produce a molecule specified with an InChI string, while another may need to consume a SMILES string. Clearly, the ability to interconvert a wide range of chemical formats and representations is essential for wrapping and integrating into a single workflow the functionality of

<sup>28</sup> SHARE Web Interface

<sup>29</sup> The IUPAC International Chemical Identifier: InChI - A New Standard for Molecular Informatics

entire software packages that have no exposed programming interfaces, but are accessible for command-line interface scripting.

As a means of demonstrating the format conversion capacity as well as integration of multiple disparate software packages with SADI, we have created an Open Babel (version 2.3.0) based format conversion service to convert InChI strings to SMILES strings<sup>30</sup>. The implementation of this functionality and SPARQL querying for resultant data is virtually identical to that of other descriptor computing services, and is readily accessible, either through the SHARE client or through a direct POST of an RDF graph containing an instance of the *inchimolecule* class (specified below) as input.

*molecule and ‘has attribute’ some (‘InChI descriptor’ and ‘has value’ some string)*

The resultant output, by virtue of classifying into the *smilesmolecule* class, since it now contains the SMILES string representation of the queried molecule, can subsequently be consumed by all of the QSAR descriptor computing services. A collection of services to convert file formats can therefore be envisioned in order to connect multiple chemical calculation packages together, on the fly.

### 15.3.3 Lipinski Rule of Five the Semantic Way

The simplicity of SADI architecture allows for natural computational resource integration into the Semantic Web, as demonstrated by seamless service invocation through simple SPARQL queries. The automated computational workflow construction that can be achieved thanks to this tight resource integration can be demonstrated by carrying out simple Lipinski Rule of Five analysis<sup>31</sup>. This well-known rule postulates that drug-like compounds can be most often characterized as having a molecular mass of less than 500 Daltons, fewer than 5 hydrogen bond donors, fewer than 10 hydrogen bond acceptors, and a logP value between -5 and 5. The definition of a Lipinski-consistent molecule lends itself quite easily for formal representation using concepts from the CHEMINF ontology, as follows.

*smilesmolecule*

and ‘hasChemicalDescriptor’ some (‘mass descriptor’ and ‘has value’ some double[< = 500.0])

and ‘hasChemicalDescriptor’ some (‘hydrogen bond donor count’ and ‘has value’ some int[< 5])

and ‘hasChemicalDescriptor’ some (‘hydrogen bond acceptor count’ and ‘has value’ some int[< 10])

and ‘hasChemicalDescriptor’ some (‘logP descriptor’ and ‘has value’ some double[< 5.0, > -5.0])

In this formal definition of a drug-like molecule, each statement linking the input SMILES molecule to a particular descriptor conforms to the output class and annotating predicate specification of a corresponding descriptor calculator service. This means that when a SPARQL query is posed to a SHARE client to determine whether a given instance of the *smilesmolecule* class is drug-like, the client will be capable of identifying and executing the four services necessary for the completion of this query (Listing 4).

Listing 4. A SPARQL query to determine whether a molecule (in lipinski\_test RDF graph) is drug-like. If it is the case, the URI corresponding to the matching molecule will be returned.

<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

<http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl#>.

<http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl>.

[http://semanticscience.org/sadi/ontology/lipinski\\_test.rdf](http://semanticscience.org/sadi/ontology/lipinski_test.rdf).

The overall effect of this is that in the absence of necessary existing data, SHARE creates a web service workflow to complement the information already available, based solely on the formal definition of Lipinski drug-like molecules in a reference ontology. Not only does this lead to improvements in computational workflow reproducibility and concept

---

<sup>30</sup> Service for converting InChI to SMILES

<sup>31</sup> Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings

disambiguation, but it also allows for straightforward means of concept reassessment from within a common framework during the course of scientific discourse. For example, the Lipinski Rule of Five has been extensively discussed, assessed and revised since its introduction<sup>32</sup>. By expressing their alternative definitions of drug-like compounds within the same framework as that of the original Rule of Five, it may have been possible to reduce ambiguity and inconsistencies in results stemming from inadvertent inconsistencies of data sources or precise methods of computational resource invocation. Further, the integration of service-computed data with more involved analysis (e.g. statistical regressions) as well as operations on entire data sets are possible, and demonstrations of this have been discussed at length elsewhere<sup>22</sup>.

### 15.3.4 Mechanisms for Parameter and Computational Experiment Provenance Specification

A number of algorithms and software packages, which may be wrapped as SADI services, require the specification of one or more parameters. The abundance of algorithm implementations and implementation-specific parameters, coupled with their under-reporting in scientific literature may often result in irreproducibility of computational experiments or discrepancies in research findings and conclusions. SADI services can specify parameters defined in an OWL ontology to control the execution of a specific computational algorithm, or to select a given algorithm from a set of equivalent algorithms which would otherwise be logically equivalent and called either together or at random, depending on the preferences and settings employed by the end user. The service description (using the GET) will display which type corresponds to the parameter class. A user wanting to specify the parameter must do so by adding its explicit description to the input RDF graph. Additionally, the provenance for the data item obtained by running the service is preserved by annotating the output as being the product of a parameterized data transformation. Besides the parameters used, this approach also allows us to explicitly specify the software (and its version), the agent (who executed it). For instance, using CHEMINF concepts, one may construct the following simplified generic output class.

An instance of *parameterized data transformation* may be placed, along with the input, into the input RDF graph and referred to in the SPARQL query to execute service computational functionality according to explicit, precise, and reproducible specifications. To demonstrate parametric execution capacity, we have created a prototype service to compute a scaled octanol-water partition coefficient value<sup>33</sup>. For some compounds, it may be necessary to apply correction factors to arrive at more accurate predicted logP values. Our service computes a logP value which is multiplied by the value of the scaling factor parameter specified in the input as follows.

**Listing 5.** A simplified input to the parameterized logP calculating service, converted to N3.

<http://semanticscience.org/>.

<http://semanticscience.org/resource/>.

If no parameter is specified, the service has an internally-specified default parameter to fall back on. In both cases, the value of the parameter is reported in the output, and the parameter itself is linked to the process executed in order to obtain the value of the descriptor (Listing 6). Because the output of a parameterized service preserves this provenance information explicitly on the descriptor this service generates, it is then possible to query over only descriptors generated using a particular set of parameters, or with a given software package. This is useful when addressing the construction of toxicological models using data derived from multiple disparate data sources or across chemical entity databases. Finally, this preserved provenance information makes our calculation fully and unambiguously reproducible.

**Listing 6.** A simplified output of the parameterized logP calculating service, converted to N3.

<http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl#>.

<http://semanticscience.org/resource/>.

<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

<http://www.w3.org/2001/XMLSchema#double>.

<sup>32</sup> Lead- and drug-like compounds: the rule-of-five revolution

<sup>33</sup> Parameterized LogP Calculator Service

<http://www.w3.org/2001/XMLSchema#double>.

### 15.3.5 Integration and Repurposing of Chemical Resources

Service interoperability, even within a single framework, relies on the compatibility of service inputs and outputs. With SADI, formal definition of input and output classes in supporting service ontologies, especially if these ontologies draw on common upper-level concepts, facilitates service integration by enabling class equivalence inference. However, if one service produces output in terms of a molecule that has a SMILES descriptor for example, no conceivable web service framework will magically enable that output to be directly consumed by a service that demands three-dimensional molecular structure specified. In these cases, intermediary services have to be made available to bridge the gap. For example, the existing SADI service to retrieve the KEGG pathways a given drug is involved in, based on a molecule's KEGG Drug identifier<sup>34</sup>, would have to be connected to *smilesmolecule* instance-generating services through a KEGG Drug identifier matching service.

If service input/output classes are logically equivalent or compatible, however, no such pipelining services are required to repurpose services for uses not originally anticipated. Consider, for example a functional group annotation service created by us to assist in lipid annotation and classification<sup>35</sup>. Given an instance of a *smilesmolecule*, this service enumerates functional group instances (from a predefined collection) occurring in the input molecule through an upper-level ontology *has proper part* predicate and produces a semantic equivalent of a chemical fingerprint in the *annotatedsmilesmolecule* output class. Although this information was originally used to classify molecules into various lipid classes, we may repurpose it for defining a customized class of chemical compounds: drug-like alkynes, as follows.

*lipinskismilesmolecule and hasProperPart some Alkyl\_Group*

In order to invoke service execution, one needs to submit a SPARQL query, much like that for the Lipinski Rule of Five use case, to SHARE. The SHARE client will then be capable of inferring not only the necessary services to invoke in order to classify an input molecule into the *lipinskismilesmolecule* class, but would also call on the functional group annotator service to obtain *hasProperPart* annotations and complete the reasoning. Thus, it is possible to build up increasingly complex queries *ad infinitum* and let the machine reasoning clients take care of the invocation and orchestration of the web services necessary to obtain the information needed to address the query.

It is also easy to imagine a service that enumerates pharmacologically active functional groups working in conjunction with QSAR descriptor computing services to logically select compounds that are predicted to be drug-like and non-toxic, out of a large collection of combinatorially-generated chemical entities. Furthermore, thanks to the ready integration and repurposing of SADI services, it is also possible to combine QSAR descriptors with molecular pharmacological activity data to obtain a formally defined QSAR model as an output of a model creator service that could wrap existing QSAR software or mathematical scripts. Finally, it is worth stressing that due to the simplicity of SADI services, they are not precluded from working with other services, or be described and accessed through other web service frameworks.

### 15.3.6 Exposing Chemical Database Resources as SADI Services

Web services are not solely limited to tasks involving computational capacities, but can be linked to a range of processes, including those carried out by experimental or industrial platforms. In certain cases, it is advantageous to use web services to encapsulate relational database lookup and the conversion of resultant information to RDF. Although large corpora of RDF data derived from the numerous publicly accessible chemical databases have been exposed for querying through SPARQL endpoints<sup>1113</sup>, and although this has a proven potential in facilitating cross-domain querying in chemistry, there is a number of reasons web service-based lookups may be preferable. For example, because the major data providers do not directly publish their information in RDF there may sometimes be a delay in the conversion and incorporation of new data by the RDF triple store providers. Further, not all of the desirable information may

---

<sup>34</sup> KEGG Pathway Association Retrieval Service

<sup>35</sup> Functional Group Annotation Service

be available in the RDF triple stores, or information might be available in a form that makes it difficult or awkward to map to one's own service ontologies, for example.

To preserve the atomic nature of SADI web services and allow for maximal flexibility in workflow construction, it is preferable to encapsulate the lookup of each index-value pair type as a separate web service in a manner identical to that of creating a CDK QSAR service. Here, the input class definition would have to require specification of the index used to look up the database and the service output class would contain entities annotated with the value or values retrieved from the database. Because this task has been demonstrated and implemented elsewhere, we shall limit our discussion of implementation to what is already stated. One point that we would like to observe is that encapsulation of database lookup functionality as SADI services allows seamless integration of chemical database resources into SPARQL queries even in the absence of corresponding RDF data. In the end, both computational and experimental resources will be available in addressing a given SPARQL query.

## 15.4 Conclusions

Chemistry is indeed an immense and rapidly growing discipline with a wealth of disparate computational and database resources which are currently largely isolated and inaccessible to truly integrative queries across the entirety of the chemical (deep) web. Thus, we believe that there is an urgent need of exposing chemical resources in a manner that would be conducive to supporting a more productive way of carrying out chemical research. In this work, we have attempted to address this issue by demonstrating what we believe to be the future of chemical service distribution and chemical resource integration into the rapidly expanding Semantic Web. Using our set of SADI services to envelop the CDK QSAR-relevant descriptor functionality to decide whether a molecule was drug-like, we have demonstrated a highly integrative behaviour afforded by the simplicity of the formal semantic service specification of the SADI framework. In the future, the widespread adoption of explicit formal specification of computational tasks afforded by Semantic Web technologies may lead to an improved reproducibility and reduced ambiguity of chemical research.

With the Lipinski Rule of Five example of SHARE-assisted automated workflow construction, we have demonstrated the kinds of powerful and natural queries that could be accessible in cheminformatics research if all of the functionalities of CDK were distributed as SADI services. Although the complexity of queries amenable to SHARE automated reasoning is somewhat limited to the capacity of the supporting formal reasoning software and computational resources of the host machine, we believe that with time, this limitation shall diminish to the point of vanishing, as existing reasoners are improved and new ones become available. Engineering limitations aside, provided an ontology of common tasks and a set of adequately specified services, researchers in the future would, in principle, only need to specify their end goal or the kind of information they seek, potentially with natural language queries, and obtain it without having to be well-versed with computational tools, programming, or pipelining. At the same time, parameter-based service control would enable advanced users to express service execution specifics. Intermediate users or those wishing to specify parameters manually or string together SADI services alongside the many other cheminformatics and bioinformatics services would also be able to do this through graphical programming in the Taverna web service interface, using the Taverna SADI plugin<sup>36</sup>. This approach could be applied to computational queries, both big and small, because the SADI framework specifically addresses synchronous and asynchronous service execution modes. This paves the way to integration of more than just database and computational resources into scientific queries, but also potentially to automation of experimentation platforms, similar to the platform deployed for the robot scientist.

Finally, distribution of resources with SADI may act as a form of insurance against computational resources being lost into oblivion as a result of changes in platform popularity or difficulties in porting computational resources across platforms, since SADI services expose a standard, platform-independent interface. Distributing computational capacity as SADI web services in the cloud may become an attractive possibility in the future. In our future work, we intend to significantly expand our collection of web services to envelop all of chemical functionality of CDK, as well as openly accessible cheminformatics and computational packages, potentially in the cloud.

We believe that the amount of knowledge created or creatable in chemistry and related fields on a daily basis has far exceeded the potential of a single human to analyse and integrate information efficiently. In order for chemistry to progress and in order for us to handle these massive and exponentially growing amounts of data, the greater chemistry

<sup>36</sup> SADI Taverna Plugin

and life sciences communities have to start exploiting the power of the Semantic Web and deferring some reasoning to machine agents. We believe that the SADI web services, the semantic resource envelopment, and the seamless machine reasoning they enable constitute the first step on our journey to a way of practicing science that transcends disciplines, knows no barriers, and encompasses all human knowledge without taxing the beholder with menial and irrelevant tasks: a self-aware science.

## 15.5 Methods

### 15.5.1 Supporting Service Ontologies

We have developed a formal OWL ontology, Lipinski Service Ontology (LSO) to capture the formal definition of service input and output classes, as well as the predicate with which a given service carries out annotations. LSO is a derivative of the CHEMINF ontology for representing chemical information and chemical descriptors, and relies on an upper level ontology, Semantic Science Integrated Ontology (SIO)<sup>37</sup>. Within LSO, we have defined a single input class for all the services, *smilesmolecule*, and a large and growing set of output classes to correspond to the output of each service individually.

### 15.5.2 Service Creation with CDK and OpenBabel

We implemented descriptor calculating functionality based on classes implementing the IMolecularDescriptor interface of CDK, version 1.3.0. Where a descriptor calculation returned multiple results, we created a separate service for each of the results within the descriptor vector thus returned, in order to preserve the atomic nature of SADI services. For cases where a three-dimensional molecular configuration was necessary in order to compute a particular descriptor, we employed the ModelBuilder3D class of CDK. For the InChI-to-SMILES service demonstrating the wrapping of programmatically inaccessible computational capacity distribution, we employed Java system calls to Open Babel<sup>38</sup> (version 2.3.0) from within the SADI service. Although we are well aware of the Open Babel API, we have chosen to access the compiled Babel binary from the command line as a means of demonstrating that numerous other command-line tools may be semantically exposed in a similar fashion.

### 15.5.3 SHARE and SADI Service Distribution

The SHARE client and SADI skeleton for generating services are freely available for download and development. We distributed our services as Java servlets, using the Jetty servlet container. We then registered our SADI services to the central service registry and queried them on the freely accessible public SHARE interface with the queries provided in text. Functionality of SADI web services that are registered in either the central or a local service registry can also be employed in manually created workflows in Taverna through the SADI Taverna plugin.

## 15.6 Authors' contributions

LLC wrote the paper and created the demonstration services. LLC and MD created the supporting service ontologies. MD contributed to the paper and provided guidance. Both authors have read and approved the final manuscript.

---

<sup>37</sup> Semanticscience Integrated Ontology

<sup>38</sup> Open Babel Open Source Chemistry Toolbox

## 15.7 Acknowledgements

This research was funded in part by NSERC CGS for LLC and the CANARIE NEP-2 Program for the C-BRASS project. We would like to thank Dr. Mark Wilkinson and Luke McCarthy for helpful discussions on SADI.

We acknowledge the article processing charge for this article that has been partially funded by Pfizer, Inc. Pfizer, Inc. has had no input into the content of the article. The article has been independently prepared by the authors and been subject to the journal's standard peer review process.



# CHEMICALTAGGER: A TOOL FOR SEMANTIC TEXT-MINING IN CHEMISTRY

## 16.1 Abstract

### 16.1.1 Background

The primary method for scientific communication is in the form of published scientific articles and theses which use natural language combined with domain-specific terminology. As such, they contain free flowing unstructured text. Given the usefulness of data extraction from unstructured literature, we aim to show how this can be achieved for the discipline of chemistry. The highly formulaic style of writing most chemists adopt make their contributions well suited to high-throughput Natural Language Processing (NLP) approaches.

### 16.1.2 Results

We have developed the ChemicalTagger parser as a medium-depth, phrase-based semantic NLP tool for the language of chemical experiments. Tagging is based on a modular architecture and uses a combination of OSCAR, domain-specific regex and English taggers to identify parts-of-speech. The ANTLR grammar is used to structure this into tree-based phrases. Using a metric that allows for overlapping annotations, we achieved machine-annotator agreements of 88.9% for phrase recognition and 91.9% for phrase-type identification (*Action* names).

### 16.1.3 Conclusions

It is possible parse to chemical experimental text using rule-based techniques in conjunction with a formal grammar parser. ChemicalTagger has been deployed for over 10,000 patents and has identified solvents from their linguistic context with >99.5% precision.

## 16.2 Background

In many scientific disciplines, the primary method of communicating scientific results is in the form of a scientific paper or thesis which uses free flowing natural language combined with domain-specific terminology and numeric phrases. As such, they contain unstructured data, which is not identifiable by machines and not easily re-usable. Information providers have built businesses around the manual abstraction of unstructured data from the literature

by human domain experts. Apart from the considerable labour cost and delay after the original publication, human abstraction is also a considerable source of error and data corruption.

A typical synthesis procedure taken from the organic chemistry literature, reads as follows:<sup>1</sup>

#### 5-Cyclobutyl-2,3-dihydro-[1H]-2-benzazepine 82:

Potassium carbonate (0.63 g, 4.56 mmol) and thiophenol (0.19 g, 1.69 mmol) were added to the 2-nitrobenzene sulfonamide **50** (0.50 g, 1.302 mmol) in N, N-dimethylformamide (33 cm<sup>3</sup>) at room temperature and the mixture was stirred for 16 h. Deionised water (50 cm<sup>3</sup>) was added and the aqueous phase was extracted with ethyl acetate (5 × 50 cm<sup>3</sup>). The organic extracts were dried (MgSO<sub>4</sub>) and concentrated under reduced pressure to give the title compound **82** (0.259 g, 1.302 mmol, ca. 100%) as an oil used without further purification.

The example shown here shows highly stylized and formulaic language, which occurs in many disciplines, and is not just restricted to chemistry, and consists of:

|nonascii\_2| **Semi-structured documents:** Usually delimited by typographic conventions such as newlines and bold text, rather than formal markup.

|nonascii\_3| **Domain-specific entities:** Entities and terminology from different scientific domains.

|nonascii\_4| **Stock phrases:** ‘X was added to a flask...’.

|nonascii\_5| **Data phrases:** ‘(0.259 g, 1.302 mmol, ca. 100%)’.

Therefore, scientific papers are an attractive target for the development of machine processes for automatic information extraction. Text-mining uses NLP (Natural Language Processing) tools for the automatic discovery of previously unknown information from unstructured data. The information generated through text mining can be used for:

\*\*\* The classification of documents (information retrieval).

\*\*\* The determination of occurrence and co-occurrence of specific terms (indexing).

\*\*\* The extraction of simple relationships.

\*\*\* The systematic extraction of data from related studies.

\*\*\* The generation of ‘mashups’ between different disciplines, such as the interactive Crystallography timemap developed by Ben O’Steen<sup>2</sup>. This visualisation shows authors of papers, geo-located onto a map and organised by the date of publication.

Text-mining in chemistry is not as prevalent as it is biology, and the tools are less developed. Text-mining in biology is often used for the automatic extraction of information about genes, proteins and their functional relationships from text documents<sup>3456</sup>. The NLP tools in biology are also well developed, and we aim to create the equivalent in chemistry for part-of-speech taggers such as the GeniaTagger<sup>78</sup> as well as syntactic parsers such as Enju<sup>9</sup>.

### 16.2.1 Aims and Objectives

The aim of this paper is to show how text-mining has been achieved for the discipline of chemistry using our ChemicalTagger tool. Chemists not only produce a significant amount of data-rich scholarly communication artefacts, but have also adopted the highly formulaic style of writing outlined above. Consequently, these publications are an attractive target for automated data extraction. The sample paragraph quoted above will be used as an example throughout this paper, but it is stressed that the techniques reported here can be applied to much of science.

<sup>1</sup> Synthesis of 5-hydroxy-2,3,4,5-tetrahydro-[1H]-2-benzazepin-4-ones: selective antagonists of muscarinic (M3) receptors

<sup>2</sup> IUCr Crystal publication data

<sup>3</sup> Automatic Annotation for Biological Sequences by Extraction of Keywords from MEDLINE Abstracts: Development of a Prototype System

<sup>4</sup> A survey of current work in biomedical text mining

<sup>5</sup> Mining semantically related terms from biomedical literature

<sup>6</sup> Methods in Biomedical Text Mining

<sup>7</sup> Developing a Robust Part-of-Speech Tagger for Biomedical Text

<sup>8</sup> Genia Tagger

<sup>9</sup> Evaluating impact of re-training a lexical disambiguation model on domain adaptation of an HPSG parser

Previous work has concentrated on the identification and extraction of chemical entities from scientific papers<sup>10 11 12</sup>, but did not address the extraction of the relationships linking these entities to both each other as well as to the document object from which they were extracted. The current work aims to address these issues using novel methods to extract information such as units, mixtures, amounts of substances and roles (such as solvents, reactants and products) as well as *Action* phrases using linguistic context. ChemicalTagger was initially developed in the context of physical science and has been designed to interoperate with bioscience tools and requirements as explored and presented at Dagstuhl 2008<sup>13</sup>. The chemical literature considered in this work consists of journal articles, open-access theses and reports (*e.g.* company reports). Most of these documents have the general structure of ‘Introduction’, ‘Materials and Methods’, ‘Experiments’, ‘Results’, ‘Discussion’ and ‘Summary and Conclusion’. This paper will focus on the experimental section, which usually consists of paragraphs such as the example shown above. The next section will demonstrate how relationships between entities can be extracted using our ChemicalTagger tool and stored in a machine-understandable format.

## 16.3 Methods

ChemicalTagger is an open-source tool for tagging and parsing experimental sections in the chemistry literature. It takes a string of text as input and produces a structured XML document as output. The ChemicalTagger workflow can be divided into five main steps: text normalisation, tokenisation, tagging, phrase parsing and finally *Action* phrase identification. These steps will be described further below:

### 16.3.1 Text Normalisation

Text normalisation is a preprocessing step that transforms the text into a format that is consistent for tagging. This involves removing nonprinting Unicode characters from text (these are the Unicode character set values 0 to 31, 127, 129, 141, 143, 144, and 157) and normalising the spacing between the words. To demonstrate, a sample phrase from the experimental paragraph above will be used:

*‘Potassium carbonate (0.63 g, 4.56 mmol) and thiophenol (0.19 g, 1.69 mmol) were added to the 2-nitrobenzene sulfonamide’*

The text-normaliser first cleans the text of nonprinting characters such as non-breaking spaces, tabs and carriage returns. It then proceeds to formatting the spaces between alphanumeric and non-alphanumeric characters (*i.e.* commas, brackets, full stops...) within the sentence. In the sentence above, strings such as ‘(0.63 g,’ and ‘(0.19 g,’ could cause problems for the tagger as they are composed of four separate elements that have been combined, and consequently would be mistagged. The normaliser will break such strings down into their constituent parts *i.e* ‘(0.63 g,’ and ‘(0.19 g,’ respectively. Special care is taken with decimal points within numbers as well as brackets and commas within chemical names. After normalisation, the following text is produced:

*‘Potassium carbonate (0.63 g, 4.56 mmol) and thiophenol (0.19 g, 1.69 mmol) were added to the 2-nitrobenzene sulfonamide’*

### 16.3.2 Tokenisation

Tokenisation is the process of splitting a phrase into into a sequence of meaningful elements called tokens. A token can be made up of one or more words and is not necessarily alphanumeric (*e.g.* commas, exclamation marks, full stops etc...). Many different splitting patterns are conceivable in natural language processing and hence many different tokenisers exist, with the most common one being the whitespace tokeniser. An adapted whitespace tokeniser is used by ChemicalTagger, since chemical names, in particular, are fragile to common methods of tokenisation as they

<sup>10</sup> Experimental Data Checker: Better Information for Organic Chemists

<sup>11</sup> High-Throughput Identification of Chemistry in Life Science Texts

<sup>12</sup> CHIC - Converting Hamburgers into Cows

<sup>13</sup> Ontologies and Text Mining for Life Sciences: Current Status and Future Perspectives

contain potential inter-token characters such as space, hyphens, brackets and commata. Running the tokeniser on the normalised sentence above produces the following tokens (Figure 1):

Potassium [carbonate] [1] [0.60] [2] [1] [4.96] [3] [and] [the] [isopropyl] [4] [0.19]  
[5] [1.60] [mixed] [6] [were] [added] [to] [the] [2-nitrobenzene] [7] [cinnamaldehyde]

Figure 16.1: Figure 1. Tokenisation  
Tokenisation.

### 16.3.3 Tagging

Tagging is the process of assigning grammatical roles to the tokens. ChemicalTagger uses a three-step cascading tagger. The first step involves running a chemical entity recogniser (OSCAR) on the tokens. ChemicalTagger then falls back on a customised regex tagger and then a parts-of-speech tagger for the tokens which have not been identified. The taggers will be discussed further below:

#### OSCAR-Tagger

OSCAR<sup>11</sup> is used for the recognition of chemical entities in text. OSCAR (Open Source Chemistry Analysis Routines) is an open source extensible system for the automated annotation of chemistry in scientific articles. It can be used to identify:

- Chemical names, including formulae and acronyms.
- Reaction names, such as *hydrolysis* and *Wolff-Kishner*.
- Ontology terms.
- Enzymes.
- Chemical prefixes and adjectives.

In addition, where possible, any chemical names detected will be annotated with structures derived either by lookup, or name-to-structure parsing using ‘OPSIN’<sup>14</sup> or with identifiers from the ChEBI [#B15]\_(‘Chemical Entities of Biological Interest’) ontology. The extracted information is stored in XML format. Identified chemical entities are marked up using the **ne** (named entity) tag. The tag has four attributes:

- **Id**: The id of the token within the document.
- **Surface**: The text that makes up the entity.
- **Type**: The chemical entity name, which can be either a chemical compound (CM), reaction name (RN), ontology term (ONT), chemical pre × (CPR), enzymes(ASE) or chemical adjective (CJ).
- **Confidence**: The confidence score associated with the identification of the entity, if the entity was identified using OSCAR’s MEMM machine learning algorithm<sup>15</sup>.

The XML output resulting from running the OSCAR parser on our sample text provides the following:

<sup>14</sup> Chemical Name to Structure: OPSIN, an Open Source Solution

<sup>15</sup> Cascaded classifiers for confidence-based chemical named entity recognition

At this stage, the chemical tokens have been successfully marked up (tokens denoted by single box and OSCAR-tagged tokens are shown in double boxes) (Figure 2):

Figure 16.2: Figure 2. OSCAR Tagging  
OSCAR Tagging.

### Regex-Tagger

The Regex-Tagger is used to mark-up *chemistry-related* terms that are not recognised by OSCAR. These include nouns such as *solution* and *mixture* and verbs such as *quench* and *evaporate* that are specific to the chemistry domain. The Regex-tagger uses regular expressions that are stored in a *rules* file together with the customised tags. *chemistry-related* terms can include the following:

**|nonascii\_96| Boldface Numbers:** These numbers usually refer to a chemical in the experiment, such as the number **50** in our example paragraph.

**\*\*Action\*\* :** Verbs that refer to specific *Actions* in an experiment, such as adding, removing, dissolving etc...

**|nonascii\_98| Physical States:** The different aggregation states a chemical compound may have such as liquid (including oils), solid (including crystals) or gas.

**|nonascii\_99| Units:** Standardised quantities such as mmol, g and mL.

This information is then passed to a regular expression tagger. Running this tagger on the sample phrase yields the following (tokens are denoted by single box, OSCAR-tagged tokens are in double boxes and regex-tagged tokens are underlined) (Figure 3):

Figure 16.3: Figure 3. Regex Tagging  
Regex Tagging.

## English Parts-of-Speech Tagger

The final step of tagging involves marking up the general English language tokens. English parts-of-speech (POS) taggers are widely available and for the purposes of this work, the Penn Treebank is used. A treebank is a parsed text corpus (i.e. annotated with syntactic structure) that is used in corpus linguistics. The Penn Treebank<sup>16</sup> is commonly used for English parts-of-speech tagging and is made up of 4.5 million American English words. Typical tags include:<sup>17</sup>

|nonascii\_100| NN singular or mass noun

|nonascii\_101| NNS plural noun

|nonascii\_102| VB verb, base form

|nonascii\_103| VBD verb, past tense

|nonascii\_104| CD cardinal number (one, two, 2, etc.)

|nonascii\_105| CC Conjunctions (and, or, plus etc...)

This treebank is used within a parts-of-speech tagger, provided by OpenNLP. OpenNLP<sup>18</sup> is a suite of open source Java projects, data sets and tutorials supporting research and development in natural language processing. Running this tagger against the non-tagged text gives the following (tokens denoted by single box, OSCAR-tagged tokens are in double boxes, regex-tagged tokens are underlined and English POS-tagged tokens are in italics) (Figure 4):



Figure 16.4: Figure 4. English POS Tagging  
English POS Tagging.

At this stage, the text has been tokenised and tagged, it is now ready for parsing.

### 16.3.4 Phrase Parsing

Parsers build on tagged tokens to assign syntactical structure to text. The goal of phrase parsing in ChemicalTagger is to build the chemical equivalent of a Chomsky<sup>19</sup> tree structure of a sentence. Figure 5 is a syntactic tree model of a simple sentence.

In this tree model **S** is a sentence, **D** is a determiner, **N** a noun, **V** a verb, **NP** a noun phrase and **VP** a verb phrase.

In human discourse, sentences are parsed in multiple valid ways. However, the formalistic structure of the chemical domain has a high probability for only one parse to be found. Therefore, a formal approach was decided on for phrase parsing so ChemicalTagger uses ANTLR<sup>20</sup>. ANTLR (ANother Tool for Language Recognition) is a parser generator that uses LL(\*) parsing to automate the construction of language recognisers. It was designed to generate grammars for formal programming languages, but is applicable to any domain where an underlying implicit grammar exists. We believe that the type of language in our corpus can largely be described by a formal grammar, therefore ANTLR is used as a novel method for parsing phrases in chemical language.

LL(\*) parsers are recursive descent parsers; they analyse input sequences by working their way down from the topmost non-terminal symbol until they reach a terminal node. In natural language these terminal nodes are the tokens. LL(\*) parsers also use leftmost derivations and the symbols at each step are consumed from Left-to-Right, the '\*' in LL(\*) refers to the use of arbitrary lookahead to make decisions. According to Parr<sup>21</sup>:

<sup>16</sup> The Penn Treebank Project

<sup>17</sup> Penn Treebank TagSet

<sup>18</sup> Open NLP Website

<sup>19</sup> NOTITLE!

<sup>20</sup> NOTITLE!

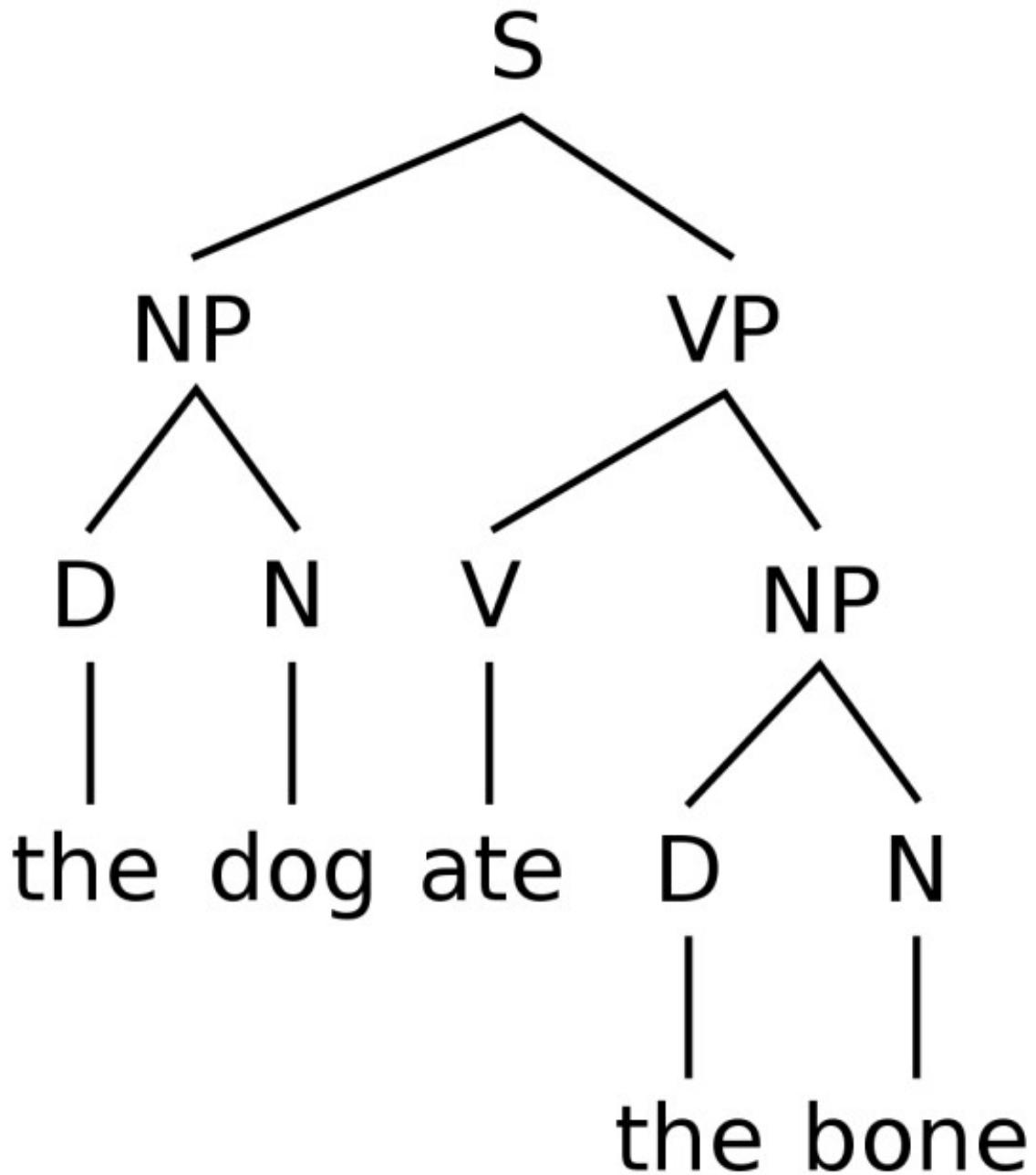


Figure 16.5: Figure 5. Basic English Syntax Tree  
Basic English Syntax Tree. [http://en.wikipedia.org/wiki/File:Basic\\_english\\_syntax\\_tree.svg](http://en.wikipedia.org/wiki/File:Basic_english_syntax_tree.svg).

LL(\*)'s arbitrary lookahead is like bringing a trained monkey along in the maze. The monkey can race ahead of you down the various paths emanating from a fork. It looks for some simple word sequences from your [...]phrase that distinguish the paths. LL(\*) represents a significant step forward in recognizer technology because it dramatically increases the number of acceptable grammars without incurring a large runtime speed penalty.

To demonstrate how ANTLR is used, a simplified version of ChemicalTagger's grammar is described below. For clarity, uppercase symbols represent terminals (symbols that can not be broken down into smaller constituents) while lowercase words represent non-terminals (symbols that can be broken down into smaller constituents).

The top-rule in our grammar is a *sentence* and it can be made up of a *nounphrase* and a *verbphrase*:

Using left derivation, the left non-terminal token *nounphrase* is selected. A *nounphrase* can be made up of a *determiner*, *adjective(s)* and *noun(s)*. A *noun* can include a *molecule* which in turn consists of an OSCAR recognised moiety followed by a *quantity*. A *quantity* consists of comma-separated numbers and units contained within brackets. This set of rules can be represented in ANTLR as follows:

Once recursion down the *nounphrase* tree is completed, *verbphrase* is next. A *verbphrase* could consist of an *adverb*, *verb(s)* followed by a *prepphrase*. A *prepphrase* is made up of a preposition followed by a *nounphrase*.

Running this grammar over the sample sentence produces the following output (Figure 6) in the form of an Abstract Syntax Tree (AST). The add-phrase shown here is only a simple example; more complex rules are defined to cover most of the grammar within the chemistry domain.



Figure 16.6: Figure 6. AST Output of ANTLR Parse  
AST Output of ANTLR Parse.

### 16.3.5 Action Phrase Identification

The text has now been tagged and parsed, the next step is to assign roles to the parsed phrases. The roles, in this instance, refer to *Actions* carried out during a chemical synthesis (*e.g.* adding, dissolving, evaporating *etc.*). After surveying the literature in collaboration with domain experts, 21 different types of *Action* phrases were defined. The complete list of phrases can be found in Table 1.

A postprocessing class was used to analyse the Abstract Syntax Tree, identify the *Action* phrases and output the tree to XML. Postprocessing our sample sentence gives the following XML output:

The following types of phrases can now be extracted from the preparation (Figure 7):

```

<Add-Phrase> [Potassium iodate (0.63 g, 4.56 mmol) and thiolein (0.19 g, 1.00 mmol) were added to the 2-aminobenzoic acid (0.50 g, 3.00 mmol) in N,N-dimethylformamide (5 mL) at room temperature]<br/>
<Dissolve-Phrase> [2-aminobenzoic acid (0.50 g, 3.00 mmol) was dissolved in N,N-dimethylformamide (5 mL)]<br/>
<Stir-Phrase> [<br/>The mixture was stirred for 30 s]<br/>
<Wash-Phrase> [<br/>The collected organic extracts were washed with water]
  
```

Figure 16.7: Figure 7. Action Phrase Markup  
Action .

It is important to note, that the parser also extracts nested noun-phrases such as the **Dissolve-Phrase** found within the **Add-Phrase** as shown above.

## Role Identification

This simple approach to *Action* phrase identification can yield good results. Other inferences can be made at this stage, such as the identification of ‘roles’. Typical roles for chemical compounds are products, reactants and solvents. Using linguistic context such as *Action* names and their position in the text we are able to detect this. For example in the following Dissolve-Phrase:

*2-nitrobenzene sulfonamide* **50** (0.50 g, 1.302 mmol) *N,N-dimethylformamide* (**33 cm<sup>3</sup>**) and Wash-Phrase:

the combined organic extracts were washed **brine**

it can be inferred that *N,N-dimethylformamide* and *brine* are solvents using cues such as their position after the preposition(underlined) and the type of Action phrase in which they are contained. The compound *2-nitrobenzene sulfonamide* may be classified as a reactant as a result of its location at the start of the text and the bold number **50** following the compound. Bold numbers are commonly used as identifiers for reactants and products in organic chemistry literature. It can also be inferred that the product of this reaction is the compound *5-Cyclobutyl-2,3-dihydro-[1H]-2-benzazepine*, because of its location in the title

**5-Cyclobutyl-2,3-dihydro-[1H]-2-benzazepine** **82**:

and the Yield-Phrase:

to give the title compound **82** (0.259 g, 1.302 mmol, ca. 100%) as an oil.

The structure provided through ChemicalTagger facilitates these inferences (see the ‘Architecture and Deployment’ section).

## 16.3.6 Output Representation

The ultimate goal of ChemicalTagger is to create machine processable structured data from natural language. The parse trees and nodes need to be preserved and labelled to identify any phrase or language component within them. Output formats for this data include CML, XML and RDF. Storing information in a structured machine-processable format makes it readily available for querying and visualisation tools. For example, a query could be run to retrieve all reactions that use *N,N-dimethylformamide* as a solvent and *2-nitrobenzene sulfonamide* as a reactant that have yields greater than 80%. This would be a useful tool for grouping together similar reactions. Structured information can also be visualised, Figure 8 shows one method of visualising extracted reaction paths.

In this figure, the numbers in the nodes refer to the products and reactants, the colours reflect the extracted information about the colour of the product and the shapes of the nodes refer to the aggregation state of the product:

- Ellipses: Unknown.
- 3D Boxes: Solid.
- Double Circles: Oil.
- Octagon: Gum.
- Triple Octagon: Foam.
- Diamond: Crystals or Needles.

As such, this graph provides a useful summary and a highly visual map of the chemistry reported in the paper -a document summary- and further analyses of this and graphs derived from other papers will open the door to the development of novel measures of document similarity (e.g. in terms of the chemical transformations reported in a corpus of synthesis papers).

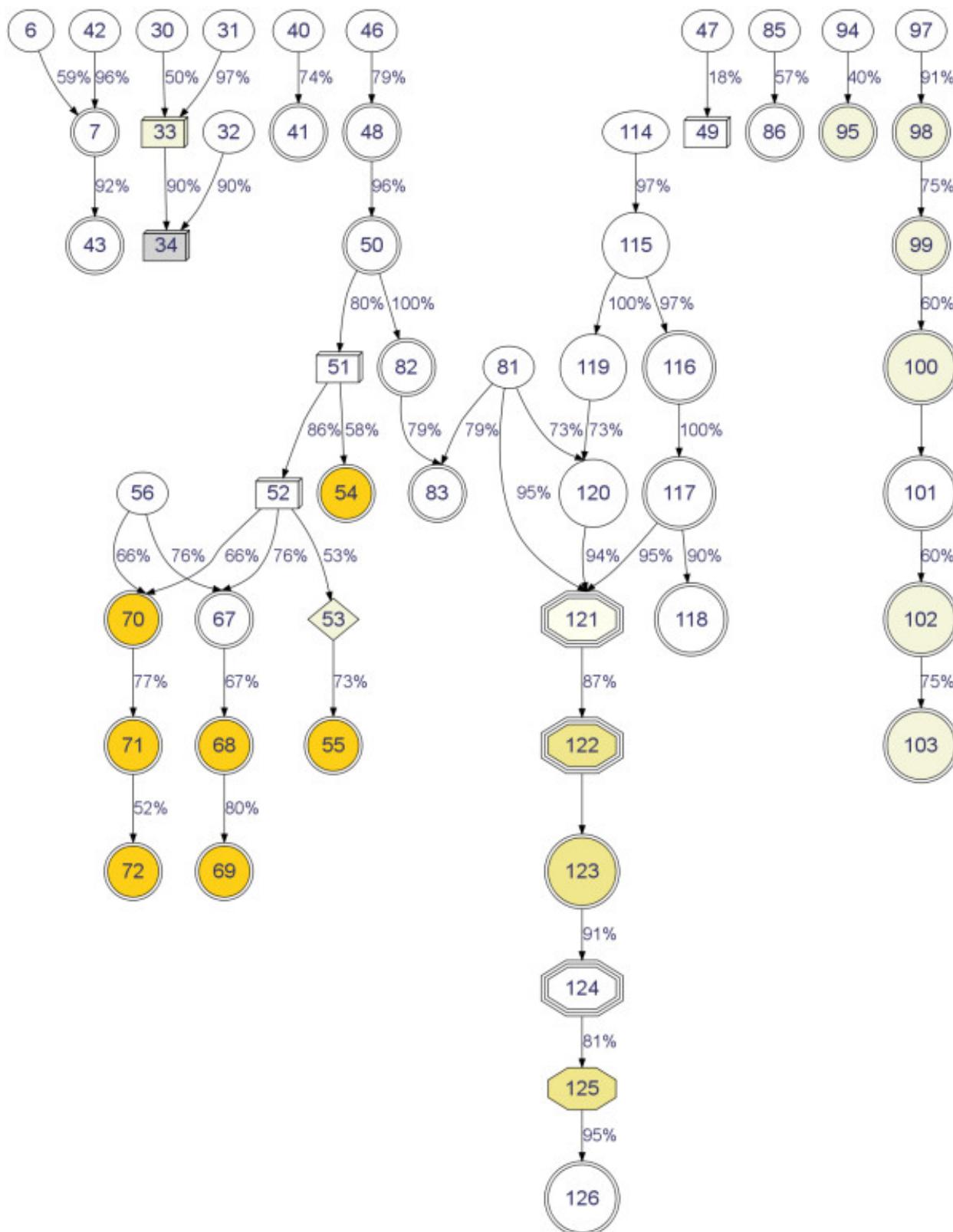


Figure 16.8: Figure 8. Graph of Reaction Paths  
Graph of Reaction Paths.

### 16.3.7 Architecture and Deployment

ChemicalTagger has been developed in a modular manner using the Java framework, making individual components such as tokenisers, vocabularies and phrase grammars easily replaceable. This facilitates the study of a wide range of chemical subdomains which vary in syntactic style, vocabulary and semantic abstraction. Moreover, it is possible to convert ChemicalTagger's output into CML<sup>21</sup> using a ChemicalTagger2CML converter. Thus, identified phrase-based chemistry such as solutions, reaction and procedures can converted into computable CML. This then allows for the construction of machine-processable synthesis information and searchable indices<sup>22</sup>.

ChemicalTagger has been used in an initial study to index large numbers (*ca.* 10,000) of patents from the European Patent Office. Preliminary results of this work were presented at the Science Online meeting<sup>23</sup> where the methodology and deployment were demonstrated. The **Dissolve** phrases were extracted to determine what solvents were used. Although precise metrics were not used, the false positive rate (i.e. identification of a compound that was not a solvent) was very low (less than 0.5%) showing that ChemicalTagger greatly enhances the precision of identification of chemical compounds as well as providing the most likely role.

The modular structure of ChemicalTagger allows for adaption to general formulaic scientific language. Thus phrases that refer to conditions such as temperature (*at a temperature of 25°C*), time (*left to equilibrate for 24 hours*), atmosphere (*under a nitrogen atmosphere*), and pressure (*caused by high pressure*) can be found in atmospheric or bioscience papers and we believe that ChemicalTagger will identify these phrases with high precision without further modification. We are intending to promote ChemicalTagger as an Open Source general scientific NLP tool. The source code is available at

<https://bitbucket.org/lh359/chemicaltagger>

and further information about ChemicalTagger can be found at

<http://www-ucc.ch.cam.ac.uk/products/software/chemicaltagger>

## 16.4 Results and discussion

Evaluation was performed by preparing a corpus of experimental paragraphs from the chemical literature and conducting an inter-annotator study. The purpose of the inter-annotator agreement study is two-fold: evaluating the agreement between human annotators agree with each other and assessing the performance of ChemicalTagger against human annotators. Although chemistry is a relatively closed domain, writing styles vary and therefore such an assessment is necessary to get a clear picture of the quality of any machine extracted information.

### 16.4.1 Corpus Assembly

The test corpus was assembled by carrying out searches for polymer synthesis related keywords in SciFinder Scholar<sup>24</sup>. The keywords were 'atom transfer radical polymerization', 'condensation polymerization' and 'anionic polymerization' and papers were chosen at random from a variety of journals. This was done in order to accommodate different writing styles and conventions used across the literature. 50 paragraphs from the experimental sections of these papers were used to create the corpus.

<sup>21</sup> A universal approach to web-based chemistry using XML and CML

<sup>22</sup> Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions

<sup>23</sup> Science Online London 2010

<sup>24</sup> Scifinder Scholar

### 16.4.2 Inter-Annotator Study

The study was carried out by four annotators, who are all trained chemists with formal backgrounds in different areas of chemistry. The annotators were provided with annotation guidelines, that can be found here<sup>25</sup>. The guidelines specify the structure of 21 different types of phrases that commonly occur in the chemistry literature and contain examples of annotated phrases from the experimental sections. The annotation process consisted of the chemists manually annotating 50 paragraphs from the test corpus and classifying the phrases according to the annotation guidelines. A point-and-click software tool was provided to facilitate annotation. After human annotation was completed, ChemicalTagger was run over the test corpus. Table 2 shows the number of *Action* phrases marked up by all four annotators alongside the number of *Action* phrases marked up by ChemicalTagger.

### 16.4.3 Evaluation

Evaluation was performed by pair-wise comparison of the annotations (*i.e.* Annotator A vs Annotator B); and the phrases as well as the *Actions* assigned to them were evaluated. Similarity of the annotations was measured using a Dice coefficient, a similarity matrix which is defined as:

$|X|$  and  $|Y|$  represent the annotations recognised by a pair of annotators.  $|X \cap Y|$  is the intersection of these annotations. The value of the similarity coefficient  $s$  therefore is twice the shared information over the combined set.

#### Identity of annotations

Previous work on annotations concentrated on named entities where strict rules for agreement between annotators. For example, in the sentence *We used sodium chloride solution* only the multiworded token *sodium chloride* would be allowed, while *sodium* and *sodium chloride solution* would both score negative. However, providing guidelines for measuring similarity between phrases is difficult. Conjunctions are problematic as are anaphora such as *Salt was dissolved in water and concentrated at 80°C*. In this example there are two phrases, but ‘and’ is not part of either and its inclusion could score negatively. Alongside the identification of the extent of the phrase (which should be exact) the annotators were also asked to identify the types of phrase (in this case *dissolve* and *concentrate*). It is possible to match the extent correctly and misidentify the type, or vice versa. These considerations are critical to interpreting the performance of ChemicalTagger.

The *Action* types and phrases in the test corpus were evaluated separately. A string match was used to evaluate the *Action* types and a machine-annotator agreement of 91.9% was achieved (See Table 3).

Evaluating phrase similarity was more challenging as annotators can often get the sense of the markup without the exact extent. Exact string match produced a low inter-annotator agreement of 55.5% and machine-annotator agreement of 48.7%. For example, *and concentrated at 80°C* and *concentrated at 80°C*. do not match identically but have sufficient overlap that it is clear that the annotators were in agreement. Therefore a set of metric techniques based on string filtration were developed. The filter removed common stock words and tokens, such as preceding adverbs and prepositions as well as ‘.', ‘;’, ‘;’, ‘and’, ‘to’, ‘the’ and ‘a’, from consideration. This improved the observed average Dice Coefficient considerably and achieved an inter-annotator agreement of 76.2% and a machine-annotator agreement of 60.4% (See Table 4).

#### Text Alignment

While filter matches improve the Dice coefficient considerably, this does not account for the ambiguity involved in defining and thus marking up the beginning and end of *Action* phrases. For example, the two *Action* phrases:

A 25 ml three-necked round-bottomed flask fitted with a dean-stark trap a condenser a nitrogen inlet/outlet

and

---

<sup>25</sup> Annotation Guidelines for Marking up Chemistry Phrases

To a 25 ml three-necked round-bottomed flask fitted with a dean-stark trap a condenser a nitrogen inlet/outlet and magnetic stirrer

would, using the above metric, be treated as two different entities although they are essentially the same *Action* phrase. As such, a disagreement between two annotators is recorded if both have marked up slightly different beginnings and endings.

To solve this problem and get a true measure of the inter-annotator agreement, we have used the Needleman-Wunsch algorithm<sup>26</sup> to align and compare annotations by different annotators. The Needleman-Wunsch algorithm is a dynamic algorithm commonly used in bioinformatics to perform the global alignment of protein sequences. The algorithm aligns sequences by matching common characters and inserting spaces in unknown or non-matching locations. The alignment is performed by assigning scores for aligned characters in the form of a similarity matrix, with gaps being heavily penalised. An optimal alignment is then found. To illustrate how the algorithm works against annotations, consider the example in Table 5. The algorithm has detected that phrases 1 to 3 highlighted by both annotators match each other, while annotator A's phrase 4 does not match anything marked up by annotator B (and therefore gives a value of -1). Annotator A's fifth sentence was identified to correspond to annotator B's fourth sentence. The following matrix is produced by the alignment algorithm:

A Dice coefficient was then calculated on the results of this alignment. Using this algorithm a machine-annotator agreement of 88.9% was achieved (see Table 6).

#### 16.4.4 Further Work

Current work investigates the use of chemical treebanks for recognising parts-of-speech tags as well as phrases. As mentioned earlier, a treebank is a parsed text corpus that is used in corpus linguistics for studying syntactic phenomena. It can also be used for training and testing parsers. Once parsed, a corpus will contain evidence of both frequency (how common different grammatical structures are in use) and coverage (the discovery of new, unanticipated, grammatical phenomena).

In life sciences, the Enju parser was adapted to biomedical domain by providing the GENIA treebank<sup>9</sup>. We aim to create an equivalent treebank for chemistry using an open-access corpus of paragraphs taken from the experimental sections of papers from the chemistry domain. This treebank will be produced semi-automatically by first running ChemicalTagger on the corpus and then manually correcting the mistagged nodes and trees. The treebank produced by this semi-automatic curation process will then be used as input for the development of a machine-learning-based parser for ChemicalTagger. An analysis of this parser's performance can then be carried out by evaluating its output against that of the ANTLR-based ChemicalTagger.

### 16.5 Conclusions

We have shown that structured scientific data can be extracted from unstructured scientific literature using ChemicalTagger. We have also demonstrated that, using text mining and natural language processing tools, we can extract both chemical entities and the relationships between those entities, and make the resulting data available in a machine-processable format. We have shown that these graphs are useful for the generation of highly informative visualisations. While machine extraction can yield good results, it nevertheless remains an act of ‘information archaeology’ and as such necessarily imperfect. We therefore strongly urge, that the scientific community move towards an ethos where scientific data is published in semantic form and where both authors and publishers feel under an obligation to make this information openly available. Were this to happen on a significant scale, it would lead to a revolution where millions of chemical syntheses every year can be automatically analysed by machine, which in turn could lead to significant improvements in our ability to do science. Opportunities generated through the large-scale availability of semantic data include:

- Formal semantic verification of published information leading to higher quality information from authors, for reviewers and for technical processing.

<sup>26</sup> A general method applicable to the search for similarities in the amino acid sequence of two proteins

- Greater understandability by readers (including machines).
- Automatic analysis of reaction conditions and results.
- Greater formal representation of chemical reactions.

We hope, however, that the extraction tools demonstrated here will have only a limited lifetime before they are replaced by semantic authoring.

### 16.5.1 Copyright Implications

It is important to note that these extraction tools are restricted to the copyright associated with the data. Patents and Open Access (CC-BY) papers explicitly allow data extraction. Theses may depend on the copyright or explicit rights within the thesis. Most publishers of chemistry are not universally Open Access and we have engaged with them over several years trying to find a straightforward answer. The authors have raised this issue with both specific publishers (*e.g.* Elsevier, who publish Tetrahedron) and the STM Publisher's Association. Elsevier have referred this to their 'Universal Access' department and currently cannot say whether or not this is permitted. It has been agreed with STM publishers that bibliographic data is Open (CC-BY or CC0). There is no agreement, at the moment, on what data can be extracted.

## 16.6 Competing interests

The authors declare that they have no competing interests.

## 16.7 Authors' contributions

LH co-authored the paper, developed ChemicalTagger and evaluated its performance. DJ co-authored the paper and developed ChemicalTagger. NA co-authored the paper, setup the test corpus and co-authored the annotation guidelines. PMR co-authored the paper and was the principle investigator on the project.

## 16.8 Acknowledgements

The authors wish to acknowledge the following for their contributions: Daniel Lowe (University of Cambridge) for providing substantial feedback on deploying ChemicalTagger and contributing to the code base, Nicholas England (University of Cambridge) and Dr Colin Batchelor (Royal Society of Chemistry) for participating in the inter-annotator agreement study and for valuable discussions in the preparation of the annotation guidelines. This research was funded by JISC and Unilever. The authors also acknowledge the article processing charge for this article that has been partially funded by Pfizer, Inc. Pfizer, Inc. has had no input into the content of the article. The article has been independently prepared by the authors and been subject to the journal's standard peer review process.

# AMBIT RESTFUL WEB SERVICES: AN IMPLEMENTATION OF THE OPENTOX APPLICATION PROGRAMMING INTERFACE

## 17.1 Abstract

The AMBIT web services package is one of the several existing independent implementations of the OpenTox Application Programming Interface and is built according to the principles of the Representational State Transfer (REST) architecture. The Open Source Predictive Toxicology Framework, developed by the partners in the EC FP7 OpenTox project, aims at providing a unified access to toxicity data and predictive models, as well as validation procedures. This is achieved by i) an information model, based on a common OWL-DL ontology ii) links to related ontologies; iii) data and algorithms, available through a standardized REST web services interface, where every compound, data set or predictive method has a unique web address, used to retrieve its Resource Description Framework (RDF) representation, or initiate the associated calculations.

The AMBIT web services package has been developed as an extension of AMBIT modules, adding the ability to create (Quantitative) Structure-Activity Relationship (QSAR) models and providing an OpenTox API compliant interface. The representation of data and processing resources in W3C Resource Description Framework facilitates integrating the resources as Linked Data. By uploading datasets with chemical structures and arbitrary set of properties, they become automatically available online in several formats. The services provide unified interfaces to several descriptor calculation, machine learning and similarity searching algorithms, as well as to applicability domain and toxicity prediction models. All Toxtree modules for predicting the toxicological hazard of chemical compounds are also integrated within this package. The complexity and diversity of the processing is reduced to the simple paradigm “read data from a web address, perform processing, write to a web address”. The online service allows to easily run predictions, without installing any software, as well to share online datasets and models. The downloadable web application allows researchers to setup an arbitrary number of service instances for specific purposes and at suitable locations. These services could be used as a distributed framework for processing of resource-intensive tasks and data sharing or in a fully independent way, according to the specific needs. The advantage of exposing the functionality via the OpenTox API is seamless interoperability, not only within a single web application, but also in a network of distributed services. Last, but not least, the services provide a basis for building web mashups, end user applications with friendly GUIs, as well as embedding the functionalities in existing workflow systems.

## 17.2 Background

Most of the common tasks in toxicity prediction consist of several typical steps, such as access to datasets, descriptor calculation and validation procedures. Usually, the components that implement these steps are developed from scratch for every new predictive application and this often leads to undesirable duplication of effort and/or lack of interoperability. The availability of a universal set of interoperable components could facilitate the implementation of new specialized applications that combine algorithms in the most appropriate way and allow fast and rigorous model development and testing.

The OpenTox framework <sup>1</sup> is being built as a collaborative effort by the partners in the OpenTox EC FP7 project, and is an attempt to design and implement a framework of web accessible components, solving common tasks in chemical properties prediction. The design objectives were to build a component based system, independent of programming languages and operating systems, where the components could interoperate between themselves and with external software packages, being able to aggregate data from different sources and staying open for future extensions. OpenTox made two major technological choices in order to keep the developments within these constraints: (i) the REpresentational State Transfer (REST) software architecture style allowing platform and programming language independence and facilitating the implementation of new data and processing components; (ii) a formally defined common information model, based on the W3C Resource Description Framework (RDF) <sup>2</sup> and communication through well-defined interfaces ensuring interoperability of the web components.

REST is a software architecture style for network based applications, defined by Roy T. Fielding by analyzing the properties of the World Wide Web and other network architectures, and deriving the architectural constraints that made the WWW successful <sup>3</sup>. There is a plethora of information on RESTful design principles <sup>4</sup>, development frameworks and examples. The REST architecture can be briefly summarized as Resource Oriented and the essential architectural constraints are as follows. Every important information entity or collection of entities is considered a resource and is named and addressable (i.e. its content can be retrieved by its address) and supports limited number of operations (e.g. read and write). Any information entity or collection of entities could be considered a resource. A resource may return its content in different formats. The content is regarded as resource “representation”. Some operations are safe (have no side effects - e.g. reading a resource) and idempotent (have same effect if executed multiple times), while others are not (e.g. creating new resources). The RESTful API design includes a specification of the allowed representation formats for each resource/operation pair. Another important design constraint is the usage of hyperlinks. It is considered good practice if all resources could be reached via a single or minimum number of entry points. Thus, the representation of a resource should return links to the related resources.

The REST style web services became a popular alternative of SOAP based services and they are considered lighter and easier to use. Contrary to the established WS-\* <sup>5</sup> standards, there are currently no standards for RESTful applications, but merely design guides. While the most widely deployed REST applications use the HTTP protocol (and therefore HTTP URIs as identifiers, HTTP methods as operations, and MIME types to specify representation formats), the architecture itself is protocol independent, therefore REST systems that do not use the HTTP protocol could exist, and vice versa. A RESTful application is characterized by complying with the architectural constraints, which are selected to ensure a set of particular properties of a distributed system. It is worthwhile to recall that the REST architecture is envisioned to be a collection of independently deployed and interacting distributed software entities, much like as there are millions of independent web servers, which constitute the WWW. Multiple independent and interacting deployments, is also the intended usage of the OpenTox REST API and AMBIT services as one of its implementations, rather than being a web application made available by a single authority or service provider.

The design of a REST system, based on the HTTP protocol, starts by identifying the domain objects, followed by mapping the objects to resources and defining identifiers (URI patterns) and operations (HTTP verbs) on each resource. In the case of OpenTox, the minimum set of domain objects, identified collaboratively by the partners <sup>1</sup>, consists of

---

<sup>1</sup> Collaborative Development of Predictive Toxicology Applications

<sup>2</sup> Resource Description Framework

<sup>3</sup>

18. Fielding, Representational State Transfer (REST)

<sup>4</sup> RESTful Web Services

<sup>5</sup> OASIS Standards

chemical compounds, properties of chemical compounds, datasets of chemical compounds and their properties (measured or calculated), algorithms (including descriptor calculation, regression, classification, structural alerts, quantum chemistry algorithms, etc.), predictive models (e.g. a model, obtained by applying a machine learning algorithm to a training dataset), validation algorithms, and reports. In addition, tasks are introduced as special resources to allow representation and handling of long running asynchronous jobs. Every resource is identified by a unique web address, following an agreed pattern, specific to the resource type (e.g./algorithm/{id} for algorithms/compound/{id} for compounds, etc.). The resources can be created (HTTP POST), updated (HTTP PUT) and deleted (HTTP DELETE), or their representations retrieved (HTTP GET). While there are diverging opinions whether POST or PUT should be used for creating resources in a REST application, our view (supported by<sup>3</sup>) is that this issue is irrelevant for the characterisation of a system as RESTful, as the design goals of the REST software architecture style (scalability, statelessness, cacheability, uniformity) are not violated by either choice. The particular choice of using POST for creating subordinate resources is a pragmatic one, as it is supported by popular REST frameworks, the HTTP protocol specification<sup>6</sup>, and the Atom Publishing Protocol<sup>7</sup>, which is often cited as a reference implementation of a REST system. Two additional features of POST's standard defined behaviour have been also accounted for as desirable properties in our design: (i) non-idempotent, meaning that multiple identical requests would probably result in the creation of separate subordinate resources with identical information<sup>4</sup>, and (ii) the URIs of the newly created resources are determined by the server, rather than specified by the client. On the other hand, updates of existing resources (PUT) require the client to specify the resource URI, again in full compliance with the HTTP protocol specification<sup>6</sup>.

The common information model of the OpenTox domain objects is based on the Resource Description Framework (RDF) and described by the OpenTox ontology<sup>8</sup>. It should be noted that the initial design of the OpenTox API (version 1.0) was based on a XML schema, but it was later decided to adopt RDF as a more powerful approach to describe objects and their relationships, as well as to facilitate the reuse of ongoing ontology developments in bioinformatics. Any entity could be described in RDF as a collection of triples (or statements), each triple consisting of a subject, a predicate, and an object. The predicate (also called a property) denotes the relationship between two objects (e.g. *Model1 has\_training\_dataset Dataset1*). The objects are modelled in RDF as *Classes* (rdf:Class), and Classes have specific *Instances*. Relationships are modelled with *Properties* (rdf:Property).

Thus, the Resource Description Framework allows defining a data model (how the data is organized), instead of specifying data format (how the data is written into a file). Once a data model is defined, it could be serialized into different formats, for example RDF/XML<sup>9</sup>, N3<sup>10</sup>, TURTLE<sup>11</sup>. The OWL Web Ontology Language<sup>12</sup> is built on top of RDF, and, compared to RDF, imposes restrictions on what is allowed to be represented. Because of such restrictions, the OWL subsets OWL-Lite and OWL-DL (Description Logic) allow performing automated machine reasoning. In OWL, there are Object properties and Data properties (owl:Property, which is a subclass of rdf:Property). An Object property specifies a relation between Instances, while a Data property specifies a relation between an Instance and a simple data value (string, integer, etc.). Properties cannot be used as Classes and vice versa.

Both REST and RDF technologies encourage data model development and consider assigning resource identifiers important. However, there are differences, as REST identifiers are used as addresses of the underlying protocol (e.g. HTTP URIs) and it is essential that URIs are dereferenceable. While the RDF representation allows HTTP URIs as resource identifiers, these are considered names, not addresses, and are not necessarily dereferenceable. HTTP URIs are hierarchical, while RDF does not exploit the hierarchy, and splits HTTP URIs into a prefix and identifier instead. REST resources define clear boundaries between information entities, while data, represented via RDF, is usually perceived as one linked graph. The common usage of RDF for data integration is to convert data coming from diverse sources into a (typically read only) single triple storage and provide a query interface (SPARQL endpoint). On the contrary, web services provide distributed and dynamically generated information. Most REST services define data formats<sup>13</sup> as a means for communication, rather than an explicit data model. The simultaneous use of RDF and REST is not yet widespread and there are ongoing debates on various related topics. Nevertheless, there is an added value

<sup>6</sup> RFC 2616 Hypertext Transfer Protocol – HTTP/1.1

<sup>7</sup> RFC 5023 The Atom Publishing Protocol

<sup>8</sup> OpenTox ontology

<sup>9</sup> RDF/XML Syntax Specification

<sup>10</sup> Notation 3 (N3) A Readable RDF Syntax

<sup>11</sup> Turtle - Terse RDF Triple Language

<sup>12</sup> OWL Web Ontology Language Guide

<sup>13</sup> Microformats

of merging both technologies for independent deployments of multiple services, able to dynamically generate linked data with dereferenceable links. This could lead to an enrichment of the information space and scalability, in a manner similar to a deployment of many web servers that provide hypertext documents.

The OpenTox framework integrates both technologies into a distributed web services framework, where both data and processing resources are described by ontologies: either existing ones, or developed within the project. The framework consists of simple modules, developed by different partners and with different programming languages, running on a set of geographically dispersed servers, and communicating via Internet. The modules can be used to build more complex use cases, embed OpenTox web services into workflows, build web mashups, consume the web services via rich client applications, etc.

This paper describes a particular implementation of a subset of OpenTox web services, based on the AMBIT<sup>14</sup><sup>15</sup> project. AMBIT is an open source software for chemoinformatics data management, which consists of a database and functional modules, allowing a variety of queries and data mining of the information stored in a MySQL<sup>16</sup> database. The modules were initially designed and developed to serve as building blocks of a desktop application (AmbitXT), as per the requirements of a CEFIC LRI<sup>17</sup> contract. The AmbitXT application features a Swing graphical user interface, and provides a set of functionalities to facilitate the evaluation and registration of chemicals according to the REACH requirements: for example workflows for analogue identification and assessment of Persistence, Bioaccumulation, and Toxicity (PBT). The downloadable installer includes a large database, covering all REACH registered chemicals, as well as several publicly available datasets featuring toxicity data. Users can also import their own sets of chemical structures and data. Downloading and running the application locally on the user machine is usually considered an advantage, especially when handling confidential data. On the other hand, with the growing popularity of the Web browser as a platform for applications, cumbersome downloads of custom desktop solutions are becoming less convenient nowadays and are even considered obsolete sometimes.

The AMBIT software was considerably enhanced within the framework of the OpenTox project, not only by providing an OpenTox API compliant REST web service interface to most of its functionalities, but also by adding the ability to describe data, algorithms, and model resources via corresponding ontologies and to build QSAR models. AMBIT REST web services are distributed as web archive (*war* file) and can be deployed in an Apache Tomcat<sup>18</sup> application server or any other compatible servlet<sup>19</sup> container. All Toxtree<sup>20</sup><sup>21</sup> modules for predicting the toxicological hazard of chemical compounds are also integrated within this package and available as REST web services via the OpenTox model API. In addition, a separate project<sup>22</sup>, implementing an OpenTox Ontology service, has been created. It consists of a simple implementation of a triple storage, exposing a SPARQL endpoint, and allowing RESTful updates via HTTP POST and DELETE commands.

## 17.3 Implementation

AMBIT is implemented in Java, uses a MySQL database as backend, and relies on The Chemistry Development Kit<sup>23</sup><sup>24</sup><sup>25</sup> for cheminformatics functionality. The OpenTox API implementation introduces two additional major dependencies, namely, the Restlet<sup>26</sup> library for implementation of REST services, and the Jena<sup>27</sup> RDF API. Apache Maven<sup>28</sup> is used for software project management (organizing dependencies and building of executables). The source

<sup>14</sup> AMBIT project

<sup>15</sup> Chapter 17. Open Source Tools for Read-Across and Category Formation

<sup>16</sup> MySQL

<sup>17</sup> CEFIC Long Range research initiative

<sup>18</sup> Apache Tomcat

<sup>19</sup> Java Servlet Technology

<sup>20</sup> Toxtree

<sup>21</sup> An evaluation of the implementation of the Cramer classification scheme in the Toxtree software

<sup>22</sup> Implementation of simple OpenTox Ontology service

<sup>23</sup> The Chemistry Development Kit

<sup>24</sup> The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics

<sup>25</sup> Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo-and Bioinformatics

<sup>26</sup> Restlet, REST framework for Java

<sup>27</sup> Jena - A Semantic Web Framework for Java

<sup>28</sup> Apache Maven

code is available in a Subversion repository at the SourceForge site <sup>29</sup>. There are two top level Maven projects, *ambit2-all* and *ambit2-apps*, consisting of several sub-modules. The first is used to organize and build modules, while *ambit2-apps* uses these modules as dependencies and builds the end user applications. The Toxtree source code <sup>30</sup> also includes dependencies on some of the *ambit-all* modules, and, on the other hand, is itself a dependency of the end user applications, in which it has been incorporated, such as AmbitXT and REST web services. The entire package currently consists of 30 Maven modules. The larger number of modules (30, compared to 21, as reported in the previous publication <sup>15</sup> that describes the standalone application), is mostly due to refactoring towards better organization of dependencies and introduction of new algorithms. The REST services implementation is organized in two modules, *ambit2-rest* and *ambit2-www*; the first one contains generic REST and RDF functionality, while the second is an implementation of the OpenTox API and builds the web application used to run AMBIT REST services.

Table 1 provides a non-exhaustive overview of the most important objects and operations of the OpenTox API, implemented by the AMBIT services. The complete description of the API <sup>1</sup> includes specifications of the input parameters and the result codes. An up-to-date version is available from the dedicated wiki at the OpenTox web site <sup>31</sup>. Currently, there is no AMBIT implementation of the OpenTox validation and reporting services; however, remote validation and reporting services are compatible, and can be used to validate models created via AMBIT services. Incorporation of the Authentication and Authorization API is under development.

The RDF representation of OpenTox objects is defined by the OpenTox ontology. The current version is available at <http://www.opentox.org/api/1.1/opentox.owl> The namespace prefix used in this paper is “ot.”, e.g. ot:Model refers to the <http://www.opentox.org/api/1.1/opentox.owl#Modelclass>. OpenTox REST resources are instances of the relevant RDF classes (e.g. <http://apps.ideaconsult.net:8080/ambit2/model/9> is an instance of the ot:Model class). Appendixes 1 and 2 provide examples how to retrieve the representations of an OpenTox model and algorithm, respectively. As a consequence of being exposed as REST web services, all OpenTox objects URIs are dereferenceable. The examples provided in the Appendixes rely on the cURL <sup>32</sup> command line tool for transferring data with URI syntax, which supports all HTTP operations (as well as other protocols). Any tool or programming language library, supporting the HTTP protocol, can be used to communicate with the OpenTox REST services. The examples use live demo instances of the AMBIT implementation of the services, but are also applicable, with minor trivial changes, to any OpenTox compliant service.

### 17.3.1 Appendix 1: An example how to retrieve the representation of an OpenTox model

```
curl -H "Accept:text/n3" http://apps.ideaconsult.net:8080/ambit2/model/9
@prefix ot: <http://www.opentox.org/api/1.1#>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
<http://apps.ideaconsult.net:8080/ambit2/model/9>
  a ot:Model ;
  dc:title "XLogP" ;
  ot:algorithm
  <http://apps.ideaconsult.net:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.XLogPDescriptor>;
  ot:predictedVariables
  <http://apps.ideaconsult.net:8080/ambit2/feature/22114>.
```

<sup>29</sup> AMBIT source code

<sup>30</sup> Toxtree - Toxic Hazard Estimation by decision tree approach, source code

<sup>31</sup> OpenTox API wiki site

<sup>32</sup> cURL tool

```
<http://apps.ideaconsult.net:8080/ambit2/feature/22114>.
```

```
  a  ot:Feature.
```

```
<http://apps.ideaconsult.net:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.XLogPDescriptor>
```

```
  a  ot:Algorithm
```

### 17.3.2 Appendix 2: An example how to retrieve the representation of an OpenTox algorithm

```
curl -H "Accept:text/n3"
```

```
http://apps.ideaconsult.net:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.XLogPDescriptor
```

```
@prefix ot: <http://www.opentox.org/api/1.1#>.
```

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
```

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
```

```
@prefix bo: <http://www.blueobelisk.org/ontologies/chemoinformatics-algorithms/#>.
```

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
```

```
@prefix ota: <http://www.opentox.org/algorithmTypes.owl#>.
```

```
<http://apps.ideaconsult.net:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.XLogPDescriptor>
```

```
  a  ot:Algorithm, ota:DescriptorCalculation ;
```

```
    dc:title "XLogP"^^xsd:string ;
```

```
    bo:instanceOf bo:xlogP.
```

The examples provided in Appendixes 3 and 4 illustrate how processing is performed via HTTP operations. The *dataset\_uri* parameter refers to the ToxCast<sup>33</sup> dataset, which consists of 320 chemicals, and is essentially an example of batch processing via the OpenTox API.

### 17.3.3 Appendix 3: An example of launching XLogP prediction for a dataset

```
curl -H "Accept:text/uri-list" -X POST -d "dataset_uri='<http://apps.ideaconsult.net:8080/ambit2/dataset/112>'_"
```

```
http://apps.ideaconsult.net:8080/ambit2/model/9 -v
```

```
< HTTP/1.1 202 Accepted
```

```
http://apps.ideaconsult.net:8080/ambit2/task/232289a2-2ce8-4f2e-9a62-8db02887577b
```

Note that both the dataset and the models are accessed indirectly via URIs, so the only data transferred on input and output are those URIs, not actual content. The result is a Task URI, and the HTTP return code 202 Accepted is an indicator that the processing has not been completed yet. In case processing was completed, the return code would have been OK 200 and the returned URI - an ot:Dataset, where results could be retrieved.

### 17.3.4 Appendix 4: An example of polling the status of asynchronous job (Task URI)

```
curl -i -H "Accept:text/uri-list"
```

<sup>33</sup> ToxCast - Predicting Hazard, Characterizing Toxicity Pathways, and Prioritizing the Toxicity Testing of Environmental Chemicals

<http://apps.ideaconsult.net:8080/ambit2/task/232289a2-2ce8-4f2e-9a62-8db02887577b>

HTTP/1.1 200 OK

[http://apps.ideaconsult.net:8080/ambit2/dataset/112?feature\\_uris\[\]](http://apps.ideaconsult.net:8080/ambit2/dataset/112?feature_uris[])=[http://apps.ideaconsult.net:8080/ambit2/dataset/112?feature\\_uris\[\]](http://apps.ideaconsult.net:8080/ambit2/dataset/112?feature_uris[])

Finally, we retrieve the prediction results from the URI shown in Appendix 4. The prediction results (Appendix 5) are represented as ot:Dataset (e.g. table with variable number of columns), which consists of data entries (ot:DataEntry) relating compounds (e.g. rows) to features (columns, ot:Feature). The table “cells” are represented as instances of the ot:FeatureValue class. A short excerpt, consisting of only two data entries (out of the total of 320 data entries included in this particular dataset), is shown in Appendix 5.

### 17.3.5 Appendix 5: An example of prediction results retrieval by HTTP GET command on URI, received as shown in Appendix 4

```
curl -H "Accept:text/n3"
"http://apps.ideaconsult.net:8080/ambit2/dataset/112?feature_uris%5B%5D=http%3A%2F%2Fapps.ideaconsult.net%3A8080%2Fambit2%2Fm
@prefix ot: <http://www.opentox.org/api/1.1#>.
@prefix dc: <http://purl.org/dc/elements/1.1#>.
@prefix: <http://apps.ideaconsult.net:8080/ambit2#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix otee: <http://www.opentox.org/echaEndpoints.owl#>.
[] a ot:Dataset ;
    ot:dataEntry
        [a ot:DataEntry ;
            ot:compound http://apps.ideaconsult.net:8080/ambit2/compound/147678/conformer/419677# ;
            ot:values
                [a ot:FeatureValue ;
                    ot:feature <http://apps.ideaconsult.net:8080/ambit2/feature/22114#> ;
                    ot:value "2.74"^^xsd:double
                ]
        ];
        ot:dataEntry
            [a ot:DataEntry ;
                ot:compound <http://apps.ideaconsult.net:8080/ambit2/compound/2146/conformer/419678#> ;
                ot:values
                    [a ot:FeatureValue ;
                        ot:feature <http://apps.ideaconsult.net:8080/ambit2/feature/22114#> ;
```

```
    ot:value "1.59"^^xsd:double
]
].
<http://apps.ideaconsult.net:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.XLogPDescriptor>
  a  ot:Algorithm.

<http://apps.ideaconsult.net:8080/ambit2/feature/22114>
  a  ot:Feature, ot:NumericFeature ;
  dc:title "XLogP" ;
  ot:hasSource
<http://apps.ideaconsult.net:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.XLogPDescriptor>
;
  =  otee:ENDPOINT_Octanol-water_partition_coefficient.
```

**The Ontology Service** is a separate project, which does not depend on AMBIT modules, and which compiles into a different web application. It currently uses the Jena TDB<sup>34</sup> persistence mechanism, and was initially designed as a proof-of-concept service to illustrate the added value of gathering RDF triples of several remote OpenTox services into the same triple storage and enabling SPARQL queries. According to our experience, the performance of the TDB storage, especially when embedded as a web service and being concurrently accessed by many users, is not optimal, and other available solutions are being evaluated. Currently, it is planned to use the ontology service as a registry of all deployed OpenTox services (both local and remote).

**AMBIT REST services** is a web application that includes all resources listed in Table 1 except the ontology service. All OpenTox objects are implemented as subclasses of *org.restlet.resource.ServerResource*<sup>35</sup>, and reside in the *ambit-www* module, which compiles into a single web archive (*ambit2.war*). The Algorithm and Task resources are implemented as in-memory objects. Compounds, Features, Datasets, and Models rely on a MySQL database backend. The common architecture is as follows.

## GET operations

The *ServerResource* receives input parameters and creates a query object, which encapsulates the database query. The query object could be as simple as the definition of a resource retrieval by its primary key or it could represent more complex queries like searching by several parameters, similarity search, or substructure search (SMARTS) pre-screening. The query object is processed by a generic “batch processing” class, which retrieves domain objects (e.g. compounds, features, datasets, dataset entries, or models) one by one, applies further processing if necessary, and serializes back from the server to the client the resource representation in the desired format by a “reporter” class. This setup allows for easy handling of new query types (by adding new query classes) and for adding many serialization formats (by writing new reporter classes). The supported MIME types for datasets (besides the mandatory *application/rdf+xml*) currently are: *chemical/x-mdl-sdfile*, *text/n3*, *application/x-turtle*, *chemical/x-mdl-molfile*, *chemical/x-cml*, *chemical/x-daylight-smiles*, *chemical/x-inchi*, *text/x-arff*, *application/pdf*, *text/uri-list*, *text/csv*, *text/plain*. Experimental support for YAML and JSON is also available. The most efficient implementation of a “reporter” class is to serialize the domain objects into the stream immediately after receiving them, without keeping the objects, or any related data, in memory. Unfortunately, when Jena is used to generate a RDF representation of a domain object, it requires building the entire RDF triple model prior to serialization. To avoid this overhead, the dataset RDF/XML serialization was re-implemented to use the Streaming API for XML (StAX)<sup>36</sup>, resulting in reduced response time of dataset retrieval (2-10 times improvement, depending on the size of the dataset).

<sup>34</sup> TDB - A SPARQL Database for Jena

<sup>35</sup> Restlet documentation on ServerResource class

<sup>36</sup> The Streaming API for XML

## POST and PUT operations

Instances of *ServerResource* receive input parameters, create a task resource, put it into an execution queue, and immediately return the task URI and representation in the requested MIME type to the client. The execution queue consists of *java.util.concurrent.Callable* objects<sup>37</sup>, while completed tasks are light objects, containing only input and output URIs. The result, as per the OpenTox REST API, is always a URI: either representing the result, or an intermediate Task object. The tasks are available via the Task service (Table 1), and are used, via GET, for accessing either the status of a unfinished task, or the URI of the results - for the completed ones. This defines a generic processing scheme where, for implementing new type of processing (e.g. integrating a new algorithm), it is sufficient to subclass the *ServerResource* and attach the specific type of *Callable* object that implements the new algorithm.

POST and PUT on datasets, compounds, and feature resources are used to create new resources or update the content of existing ones, and always return the URI of the new resources or the URI of the updated ones. POST on machine learning algorithms (e.g. regression, classification, or clustering) creates a new model resource and returns its URI. The representation of a model URI can be retrieved via GET to inspect the model details (e.g. training dataset, independent variables, specific parameters). POST on a model URI creates a new dataset, containing prediction results, and returns its URI. Returning the URI of a subordinate resource upon POST is in compliance with REST recommendations (and HTTP specifications<sup>6</sup>), as the content of the result URI could be later accessed via GET, obeying the cacheability constraint of the architecture. Neither REST nor HTTP strictly defines the meaning of “subordinate” resource; we however consider the OpenTox API interpretation compliant to the REST architecture, because in all of the cases, presented above, POST on a resource creates a new dependent resource, and is defined in a uniform manner. An important difference to remote procedure call (RPC) based architectures is that the client does not send the complete data to be processed; the processing service receives only the data URI, which it uses to retrieve the appropriate representation when it needs the data. The distinction between information resources and their representations, which is considered a key feature of REST, enables the processing resource to choose the most appropriate representation (i.e. no additional data conversion is necessary!) and keep track of the data provenance by simply referring to the data URI and its relevant metadata. This design also allows to dynamically generate predictive models, immediately making them available online, and maintaining in the underlying representation of linked resources all the information required to reproduce the model building process, which was one of the initial design goals of the OpenTox framework.

The results of applying the REST constraints to information processing elements, like data analysis algorithms, leads to a change in the way of thinking, modelling, implementing, and perceiving data processing. From a point of view of the REST architecture, a data processing algorithm is just another resource that retrieves data, given its identifier, and creates a resulting resource with another identifier. The difference between the data and processing elements vanishes.

## DELETE operations

Usually implemented by deleting objects from the database backend, the integrity is managed via a standard relational database foreign keys mechanism. Integrity between local and remote objects is not addressed. If a local object refers to a remote OpenTox object, e.g. predictions stored as an AMBIT dataset by a remote model, and the remote model service becomes unreachable, this will not be reflected in any way. This is similar to the generic problem of broken hyperlinks on the Web and might be addressed in future by some suitable keep-alive or synchronization mechanism.

## RDF input/output

Jena in-memory models are used to read incoming RDF data and to serialize domain objects into RDF formats. The lack of streaming RDF readers and writers is a major disadvantage for the use of RDF for data transfer. A possible workaround is to introduce a persistent RDF storage, but the performance gain has still to be evaluated. Another disadvantage of making domain objects available in RDF is the lack of support from most popular scripting languages, used to build web applications (e.g. JavaScript). As a workaround, JSON (Java Script Object Notation)<sup>38</sup> serialization of RDF is considered, and although many proposals and implementations exist, there is currently no standard for JSON

<sup>37</sup> JDK Documentation

<sup>38</sup> JSON (JavaScript Object Notation)

serialization. Two of the existing JSON libraries have been evaluated, with the results not encouraging - the volume of the JSON representation is comparable to that of RDF/XML, and the same is true for the corresponding memory consumption. Possible workarounds are either to build client applications in programming languages with good RDF support or to provide alternative formats with efficient streaming support. Fortunately, the REST architecture natively supports multiple representations per resource, which allows using the most appropriate format for carrying out a particular task.

A clear advantage of the availability of RDF representations for the OpenTox objects, data, algorithms, and models is that it allows to combine easily the RDF representations of remote resources into a standard triple storage, annotating and cross-linking objects with terms from existing ontologies. Publishing a dataset of chemical structures and their properties as linked data becomes as straightforward, as uploading a *sdf* file into an OpenTox dataset service, with optional subsequent annotation of property tags.

## 17.4 Results and Discussion

We have implemented a large subset of the OpenTox API in the open source AMBIT REST package, and have made it available both as live demo online services and as a downloadable package, allowing third parties to install and run separate instances of the services, either on Intranet or publicly on the Internet.

The major advantage is the ability of the framework to hide implementation details and offer diverse functionality via a uniform application programming interface, which, while generic, allows encapsulating very diverse data and predictive algorithms and allows seamless integration of remote services. Additionally, representing domain objects via the Resource Description Framework allows to explicitly assert relationships between data and data generation processes.

### 17.4.1 Uniform access to data

The OpenTox compound and dataset API provide generic means to access chemical compounds and aggregate various data. **Chemical compounds** are assigned unique URIs, and can be retrieved, created, or deleted via HTTP POST, PUT and DELETE commands, submitted to the compound service <http://host:port/{service}/compound>. The GET command returns a representation of the chemical compound in a specified MIME format (Appendix 6). Changing the MIME format in this example will return the representation of the compound in that format, making the service essentially work as a format converter.

### 17.4.2 Appendix 6: An example of retrieving a compound in a specified format (Chemical MIME for SMILES in this example)

```
curl -H "Accept:chemical/x-daylight-smiles" http://apps.ideaconsult.net:8080/ambit2/compound/1
```

O=C

The concept of a **dataset of chemical compounds** is central to the OpenTox web services functionality. Algorithm services accept a dataset URI in order to build a model or to generate descriptor values. Model services accept a dataset URI in order to apply a model and obtain predictions. Predictions are also returned as a dataset URI, whose contents could be subsequently retrieved (Appendix 5). Search results (by identifiers, similarity, or substructure), are available as datasets as well.

The OpenTox Dataset (ot:Dataset class) can be thought of as a file of chemical compounds, along with their properties, which is identified (and referred to) by a unique web address, instead of a filename, and can be read and written remotely. The dataset POST operation allows uploading datasets in RDF representation, as well as files with chemical structures with arbitrary set of fields. AMBIT services do not restrict entering and uploading data to predefined fields only. Instead, arbitrary data can be imported, and later annotated to establish the semantics of the fields. When uploading data in RDF format, the client has full control of the fields' representation. This is a substantial improvement

over most of the current practices with popular chemical formats, which usually involve describing the meaning of the fields in separate documents, targeted at human readers; sadly, this approach tends to lead to quite frequent peculiarities.

### 17.4.3 Appendix 7: A RDF representation of a single entry from the DSSTox Carcinogenic Potency Database dataset

```

@prefix ot: <http://www.opentox.org/api/1.1#>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix: <http://apps.ideaconsult.net:8080/ambit2/>.
@prefix otee: <http://www.opentox.org/echaEndpoints.owl#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix ac: <http://apps.ideaconsult.net:8080/ambit2/compound/>.
@prefix ad: <http://apps.ideaconsult.net:8080/ambit2/dataset/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix af: <http://apps.ideaconsult.net:8080/ambit2/feature/>.

af:21611
  a  ot:Feature ;
  dc:title "ActivityOutcome_CPDDBAS_Mutagenicity" ;
  ot:hasSource ad:10 ;
  =  otee:Mutagenicity.

af:21604
  a  ot:Feature ;
  dc:title "TD50_Dog_mg" ;
  ot:hasSource ad:10 ;
  ot:units "mg" ;
  =  otee:ENDPOINT_Carcinogenicity.

ac:144089
  a  ot:Compound.

ad:10
  a  ot:Dataset ;
  ot:dataEntry
    [ a  ot:DataEntry ;
      ot:compound ac:144089 ;
      ot:values
        [a  ot:FeatureValue ;

```

```
ot:feature af:21604 ;
  ot:value "blank"^^xsd:string
];
ot:values
[a  ot:FeatureValue ;
  ot:feature af:21611 ;
  ot:value "active"^^xsd:string
]
].

```

A simple example is representing carcinogenicity data from two public datasets, DSSTox CPDBAS: Carcinogenic Potency Database<sup>39</sup> (Appendix 7) and ISSCAN: Chemical Carcinogens Database<sup>40</sup>. Both datasets are available as *sdf* files, with fields described in human readable documents. The outcome of the carcinogenicity study is represented in the “ActivityOutcome” field in CPDBAS (with allowed values “active”, “unspecified”, “inactive”), while in ISSCAN, a numeric field named “Canc” is used with allowed value of 1, 2, or 3. The description of the numbers (3 = *carcinogen*; 2 = *equivocal*; 1 = *noncarcinogen*) is only available in a separate “Guidance for Use” *pdf* file. Ideally, toxicity prediction software should offer comparison between the data and models, derived from both datasets, which is currently impossible without involving human efforts to read the guides and establish the semantic correspondence between the relevant data entries if and when possible. Moreover, every toxicity prediction package has to do the same. The two files in the example are selected only because they are publicly available and widely known. This is a typical scenario of the current state of representing toxicity data. Even if the toxicity data is highly structured within a commercial or in-house database, the usual practice for exchanging it is through export into unstructured file formats. ToxML<sup>41</sup> is a notable example of an attempt of a structured file format for data exchange in toxicology, but it has not yet been adopted beyond its original authors, even though there are ongoing efforts in this direction<sup>42</sup>. There are a variety of relevant ontology development efforts<sup>4344</sup>, but these are in most cases done in a different context, and are only partially applicable to representations of toxicology studies.

Being aware of the lack of standards in this area, the authors of the OpenTox API have designed it in a way to provide a generic approach towards data representation, keeping the flexibility of importing arbitrary named fields, but still allowing assignment of computer readable annotations to the fields. This is illustrated in Appendixes 8 and 9.

#### 17.4.4 Appendix 8: A RDF representation of the “Canc” field of the ISSCAN dataset, available via AMBIT services and OpenTox API (prefixes are the same as in Appendix 7, and therefore omitted)

```
ad:9 a ot:Dataset ;
  rdfs:seeAlso "http://www.epa.gov/NCCT/dsstox/sdf_isscan_external.html" ;
  dc:source "ISSCAN_v3a_1153_19Sept08.1222179139.sdf" ;
  dc:title "ISSCAN: Istituto Superiore di Sanita, CHEMICAL CARCINOGENS: STRUCTURES AND EXPERIMENTAL DATA".
```

af:21573

a ot:Feature ;

<sup>39</sup> CPDBAS: Carcinogenic Potency Database Summary Tables

<sup>40</sup> ISSCAN: Chemical Carcinogens Database

<sup>41</sup> ToxML

<sup>42</sup> ToxML project

<sup>43</sup> The Open Biological and Biomedical Ontologies

<sup>44</sup> MIBBI: Minimum Information for Biological and Biomedical Investigations

```

dc:title "Canc" ;
ot:hasSource ad:9 ;
= otee:ENDPOINT_Carcinogenicity.

```

The fields in *sdf* files and other formats can contain arbitrary attributes, which are represented as instances of the *ot:Feature* class from the OpenTox ontology. Every feature is identified by a unique URI, which is hosted at a feature service (<http://host:port/{service}/feature>) and is dereferenceable (a representation of the feature can be retrieved through a GET command). The RDF representation includes a feature name (via *dc:title* property), units (via *ot:units* property), and a link to the resource (via *ot:hasSource*) that was used to generate this property or where it was originally read from. Currently, the range of *ot:hasSource* property is defined to be one of *ot:Algorithm*, *ot:Model*, or *ot:Dataset*. Using the *owl:sameAs* property, it is possible to assert that an instance of the *ot:Feature* class is the same as another resource, defined in some other ontology. An example is shown in Appendix 8, where the feature af:21573 is asserted to be the same as the *otee:ENDPOINT\_Carcinogenicity* individual from a simple ontology<sup>45</sup> that enables the representation of physicochemical properties and toxicology endpoints as defined in the ECHA guidance document<sup>45</sup>. The same approach, as well as using the *rdf:type* property, can be applied to assign more elaborate representations of toxicity studies to a particular feature, provided that an ontology describing the study exists. This technique is used to represent the ToxCast data in AMBIT services, and enables linking and querying related entries from the GO ontology<sup>46</sup>.

#### **17.4.5 Appendix 9: A RDF representation of a subset of fields of the CPDBAS dataset, available via AMBIT services and OpenTox API (prefixes are the same as in Appendix 7, and therefore omitted)**

af:21603

```

a ot:Feature ;
dc:title "STRUCTURE_MolecularWeight" ;
ot:hasSource ad:10 ;
= <http://example.org#an-ontology-entry-representing-molecular-weight>.

```

af:21607

```

a ot:Feature ;
dc:title "STRUCTURE_ChemicalName_IUPAC" ;
ot:hasSource ad:10 ;
= <http://example.org#an-ontology-entry-representing-IUPAC name>.

```

af:21610

```

a ot:Feature ;
dc:title "ActivityOutcome_CPDBAS_Rat" ;
ot:hasSource ad:10 ;
= otee:ENDPOINT_Carcinogenicity.

```

ad:10

```

a ot:Dataset ;
rdfs:seeAlso "http://www.epa.gov/NCCT/dsstox/sdf\_cpdbas.html" ;

```

<sup>45</sup> Guidance on information requirements and chemical safety assessment Chapter R.6: QSARs and grouping of chemicals

<sup>46</sup> GO Ontology

dc:title “CPDBAS: Carcinogenic Potency Database Summary Tables - All Species”.

Instances of the ot:Feature class (Appendix 9) are used to represent arbitrary properties, including chemical identifiers (e.g. *STRUCTURE\_ChemicalName\_IUPAC*), properties like molecular weight (e.g. *STRUCTURE\_MolecularWeight*), or calculated descriptors (Appendix 5) and model predictions (Appendix 11). If ot:hasSource points to an OpenTox algorithm or model URI, it could be directly used to launch the calculations for any new compound or dataset by simply initiating a HTTP POST to this URI, with an input parameter, pointing to the compound or dataset. This ensures keeping track of all the processing steps performed by the OpenTox services, and provides sufficient information to reproduce or repeat the calculations (Appendix 5). Features can be deleted by sending a DELETE command to the feature service, and created or updated via POST and PUT commands by providing a RDF representation as an input parameter. AMBIT services automatically create features when a dataset is being uploaded. If the uploaded dataset is not in RDF format, the features are generated with dc:title equal to the field name in the file and ot:hasSource property linking to the dataset, the combination of both properties used as a unique key. The features representation can be modified and annotated later by sending an appropriate RDF representation to the feature URI via a HTTP PUT command.

The use of dynamically generated and dereferenceable URIs for RDF resource identifiers differs from the classic recommendation of using “stable” identifiers from a predefined ontology. However, we consider the dynamically generated RDF graph an advantage of OpenTox services, and, moreover, it does not preclude linking dynamically generated resources with equivalent resources that have stable identifiers, if such exist. For example, features are expected to be associated via owl:sameAs links with stable identifiers, describing specific chemical properties. Arbitrary RDF statements, including both dynamically generated and stable resources could be added as well. The dynamically generated RDF representations allow quickly publishing information in RDF format and making it available online. Models and predictions also immediately become available as RDF resources online, and include live local and remote links, keeping track of the provenance (how predictions have been calculated and where the data came from). Given the availability of the OpenTox services as open source, anybody interested could run an instance of the services themselves, for as long as necessary. Because of the interoperable and distributed design, multiple instances of services running at multiple places could communicate and generate dynamically linked data. The URIs and addresses of networking resources generally don’t have infinite lifetime, but this is not considered disadvantage for the World Wide Web, where, if any piece of the dynamic infrastructure is perceived important - for economic or any other reasons - it will certainly remain available for longer than average. The fact that HTTP URIs are transient and dependent on the service location is a consequence of the early Internet design as a medium for host-to-host communication, rather than one for data access, and also of the lack of location independent application names in Internet protocols <sup>47</sup>. Revising the current status of network resources naming towards persistent and self-certifying names and content-oriented networking is a field of active research nowadays, and may render the disputes about dereferenceability and stability of resource identifiers irrelevant in future.

Finally, it is trivial to retrieve the RDF representations from an arbitrary set of geographically distributed services. It is equally easy to create a snapshot of the content of a given subset of services of particular interest, either for archiving purposes, or in order to import it into a RDF triple storage and expose it via a SPARQL endpoint.

We support the view <sup>4849</sup> that the current practice of aggregating data via loading RDF dumps into a single triple store is not always the best approach, but rather a temporary solution, until emerging technologies for distributed querying and reasoning become more efficient and scalable enough to eliminate the need of centralized data stores. Meanwhile, web services as the OpenTox REST ones, that provide dynamically generated RDF data via resolvable identifiers, can be crawled in a similar way as search engines crawl the web. However, there is the additional benefit of results being retrieved and reasoning performed over structured data, instead of just analysing keywords and links as popular search engines typically operate today.

---

<sup>47</sup> NOTITLE!

<sup>48</sup> An evaluation of approaches to federated query processing over linked data

<sup>49</sup> Active Knowledge: Dynamically Enriching RDF Knowledge Bases by Web Services, Proceeding

## 17.4.6 Uniform approach to data processing, model building, and predictions

The ability to represent data in a generic way, as explained above, greatly simplifies **data processing**. The latter can be described as the following three simple steps:

1. Read data from a web address, representing an ot:Compound or an ot:Dataset instance;
2. Perform processing; store results as ot:Dataset representation (e.g. ot:FeatureValue instances);
3. Write the ot:Dataset RDF representation to an OpenTox data service; return the URI of the resulting dataset.

The OpenTox API specifies two classes that perform processing - ot:Algorithm and ot:Model, supported by <http://host:port/{service}/algorithm> and <http://host:port/{service}/model> services, respectively. The lists of available algorithms can be retrieved by a GET command. The type of the algorithm is specified by sub-classing the algorithm instance from the respective class in the Algorithm types ontology<sup>1</sup>. Two major types of algorithms are data processing ones and model building algorithms.

Models are generated by the respective algorithms, given specific parameters and data. The process of **model creation** (e.g. using statistical algorithm to build a model) is initiated by sending a POST command to the algorithm service (example available in the Supporting Information [Additional file]

Additional file 1

**Supporting Information.** Examples of accessing various AMBIT REST services via the cURL tool.

[Click here for file](#)

1. Optionally read data from a web address, representing an ot:Dataset instance;
2. Create a model; describe it as an ot:Model instance; this includes specifying ot:Feature instances that contain the results, via the ot:predictedVariables property, as well as linking any independent and target variables via the ot:independentVariables and the ot:dependentVariables properties;
3. Assign a unique URI to the model, and return the URI;
4. A POST command to the model URI, with a dataset or compound URI as input parameter, could be later used to calculate predictions.

This architecture turns out to be successful in encapsulating different algorithms and models in a single API. A summary of the algorithms, included in AMBIT REST services, is shown in Table 2 and the full list can be retrieved originally from <http://apps.ideaconsult.net:8080/ambit2/algorithm> or from <http://host:port/ambit2/algorithm> in any other installation of the *ambit2.war*.

Most of the algorithms (except Weka and Toxtree) are considered data processing algorithms, and accept a dataset URI as input parameter, returning URI of the resulting dataset. The calculated values are included as feature values, as explained above. The structure optimization algorithm returns a dataset with links to the new 3D structures. SMARTCyp and SOME algorithms return their results as features as well, but the features represent calculated atomic properties. The MCSS algorithm accepts a dataset and creates a model, containing a set of maximum common substructures. The model could be further applied to new datasets or compounds. The superservice is an algorithm, which encapsulates descriptor calculation and model prediction, by automatically identifying which descriptors are required by a given model, launching the calculation, and, when results are available, applying the model itself. Toxtree algorithms are implemented as a model building algorithm, although being fixed rules and not requiring a training dataset. Thus, upon installation of the web application, the Toxtree model needs to be created by sending a HTTP POST to the corresponding algorithm. The Weka algorithms are selected to be representative of regression, classification, and clustering algorithms. They accept a dataset URI and a feature URI (referring to the target variable), and generate a model URI, as specified in the API. The implementation of Weka algorithms as OpenTox REST services is a generic one; inclusion of all algorithms, available in the Weka package, is just a matter of configuration, and the list will be extended in future releases. The RDF representation of all algorithms and models can be retrieved by submitting a GET command.

### 17.4.7 Registering data, algorithms and models; SPARQL query

The OpenTox ontology service provides a place for registering OpenTox resources, running at remote places, as well as searching capabilities via SPARQL. Registering a resource into the ontology service requires sending a POST command to the service, with a parameter, pointing to the resource being registered (see Supporting Information [Additional file]

### 17.4.8 Appendix 10: An example of retrieving information about a specific model (X and Y variables; learning algorithm; variables, containing the predictions; endpoints)

```
PREFIX ot: <http://www.opentox.org/api/1.1#>
PREFIX ota: <http://www.opentox.org/algorithms.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX otee: <http://www.opentox.org/echaEndpoints.owl#>

SELECT ?Model ?algorithm ?xvars ?descriptorAlgorithms ?yvars ?endpoints ?predicted
WHERE {
?Model rdf:type ot:Model.
OPTIONAL { ?Model dc:title ?title }.
OPTIONAL {
?Model ot:algorithm ?algorithm.
?algorithm rdf:type <http://www.opentox.org/algorithmTypes.owl/#Regression>.
}.
OPTIONAL {
?Model ot:independentVariables ?xvars.
OPTIONAL { ?xvars ot:hasSource ?descriptorAlgorithms }.
}.
OPTIONAL {
?Model ot:dependentVariables ?yvars.
OPTIONAL { ?yvars owl:sameAs ?endpoints }.
}.
OPTIONAL {
?Model ot:predictedVariables ?predicted.
OPTIONAL { ?predicted owl:sameAs ?endpoints }.
}.
}
```

Any number of ontology services can be installed, thus allowing clustering and querying resources of interest to specific applications. Policies and access rights for protecting the resources are currently under development. Alternatively, a RDF triple storage of choice could be used to aggregate resources, generated by different implementations of OpenTox services.

A RDF graph, describing two models (tumm:TUMOpenToxModel\_kNN\_92 and am:33), running on remote services and using the same training dataset (ot:trainingDataset ad:R545) and descriptors (ot:independentVariables af:22213, af:22137, af:22252, af:22127; the link to the descriptor calculation service shown only for the af:22127), hosted and calculated by AMBIT services, is provided in Appendix 11.

#### **17.4.9 Appendix 11: A RDF graph, representing two remote models, using the same training dataset (the RDF content was aggregated by retrieving the RDF representations of multiple web services, and is available as Supporting Information [Additional file]**

Additional file 2

**Supporting Information.** RDF graph, representing two remote models, using the same training dataset.

[Click here for file](#)

```
@prefix: <http://apps.ideaconsult.net:8080/ambit2/>.
@prefix ot: <http://www.opentox.org/api/1.1#>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix tuma: <http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/>.
@prefix tumm: <http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/model/>.
@prefix ota: <http://www.opentox.org/algorithmTypes.owl#>.
@prefix otee: <http://www.opentox.org/echaEndpoints.owl#>.
@prefix bo: <http://www.blueobelisk.org/ontologies/chemoinformatics-algorithms/#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix am: <http://apps.ideaconsult.net:8080/ambit2/model/>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix ac: <http://apps.ideaconsult.net:8080/ambit2/compound/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix ad: <http://apps.ideaconsult.net:8080/ambit2/dataset/>.
@prefix ag: <http://apps.ideaconsult.net:8080/ambit2/algorithm/>.
@prefix af: <http://apps.ideaconsult.net:8080/ambit2/feature/>.

tumm:TUMOpenToxModel_kNN_92
a ot:Model ;
dc:title "OpenTox model created with TUM's kNNregression model learning web service." ; ot:algorithm tuma:kNNregression ;
ot:dependentVariables
af:22200 ;
```

```
ot:independentVariables
af:22213, af:22137, af:22252, af:22127 ;
ot:predictedVariables
af:27501 ;
ot:trainingDataset ad:R545.

am:33
a ot:Model ;
dc:title "Caco-2 Cell Permeability" ;
ot:algorithm ag:LR ;
ot:dependentVariables
af:22200 ;
ot:independentVariables
af:22213, af:22137, af:22252, af:22127 ;
ot:predictedVariables
af:26182 ;
ot:trainingDataset ad:R545.

ag:LR
a ot:Algorithm, ota:Supervised, ota:EagerLearning, ota:SingleTarget, ota:Regression;
dc:title "Linear regression"^^xsd:string.

af:22127
a ot:Feature ;
dc:title "FPSA-2" ;
ot:hasSource
<http://apps.ideaconsult.net:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.CPSADescriptor>.
```

#### 17.4.10 Linked resources

Uploading data and running calculations via the OpenTox API and its implementation by AMBIT services generates a multitude of linked resources, all available via their RDF representations. The links could span many remote sites, running various implementations of OpenTox services. For example, a model, built by model services running at site A, will be accessible via its web address, but the representation could include links to the training dataset and prediction variables, hosted at OpenTox services running at site B. The features, representing predicted variables, contain links back to the remote model. An illustration of linked resources, generated by OpenTox services, is provided on Figure 1 and Additional file

#### 17.4.11 Comparison with similar systems

The design of the OpenTox REST API and its implementation started at the beginning of the OpenTox FP7 project in 2008. At that moment we were not aware of any API with comparable functionality and design. There were examples of REST services in cheminformatics, but usually designed as a monolithic system and not available for download

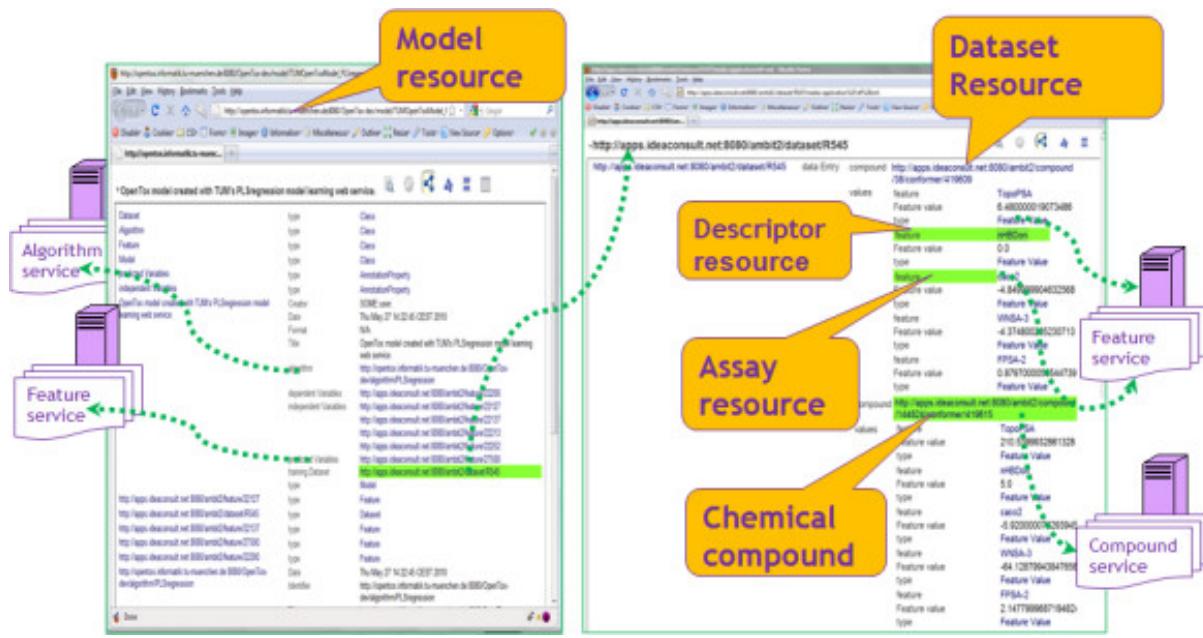


Figure 17.1: Figure 1. Illustration of linked resources, generated by OpenTox services  
Illustration of linked resources, generated by OpenTox services.

and installation elsewhere. The OpenTox framework is designed and developed collaboratively with the intention to be a modular and interoperable distributed system. The SADI framework<sup>50</sup><sup>51</sup> is the only other existing system which combines REST and RDF technologies to perform bio- and cheminformatics tasks. It should be noted, though, that these systems have been developed independently, without mutual awareness, until quite recently. While both approaches might seem similar to some extent, there are significant differences in their design and implementation.

The main goal of the OpenTox framework is to provide distributed means for building, using, and validating predictive models. We are not fully aware whether SADI services support generating and validating new predictive models via machine learning techniques or other methods. OpenTox services are independent, and can be mashed up or invoked in serial or parallel fashion by explicit invocation of command tools, existing workflow systems, or custom user interface. SADI's strong point is in the usage of implicit invocation of web services, given a SPARQL query. The SHARE engine<sup>52</sup> decides which services to invoke in order to fill in the missing data. The SADI services use HTTP, but define HTTP resources only for the processing elements, not for the data elements. The calculations are initiated by a POST command, and the data is returned in the body, resembling a typical processing by a remote procedure call, rather than a REST resource. Input data is subsumed into the output data, and neither of the data has its own dereferenceable identifier. OpenTox services work by accepting a URI of an input resource and return a URI of the resulting resource. The content of the latter could be retrieved by a subsequent GET operation if necessary - as a whole or in parts. This allows processing of datasets of arbitrary number of entries. Dataset is a central type of resource in OpenTox, while we are not aware of a corresponding concept in SADI. Implementation-wise, SADI services require a RDF triple storage as a backend, while OpenTox services do not mandate any particular backend representation; it is sufficient only to serialize resources to RDF on input/output in order to be compatible with the OpenTox API. Another difference exists due to the requirement to define a custom input/output format for each SADI processing service, while OpenTox services have a uniform interface, which resembles conceptually the standard input and output streams in UNIX operating systems, and brings proven flexibility when composing services into arbitrary workflows. Finally, SADI strives to provide a single ontology, describing all cheminformatics services. We believe that this is hardly achievable in a truly distributed system, and therefore designed OpenTox in a different way; we provide a skeleton ontology, allowing representation of a few basic classes, generate dynamic resources, and link/annotate them with all

<sup>50</sup> SADI framework

<sup>51</sup> SADI, SHARE, and the in silico scientific method

<sup>52</sup> SHARE: A Semantic Web Query Engine for Bioinformatics

relevant third party ontologies.

### 17.4.12 Applications

Although all AMBIT REST services support HTML MIME format and could be accessed through a web browser, the intended use is via custom client applications, which would consume the web services, and provide a friendly user interface, tailored to specific use cases. An example is the ToxPredict<sup>153</sup> web application, which provides a customized user interface for searching data, selecting and applying models, and displaying prediction results. Integration of REST services into workflow systems and rich client applications are other options, subject to further work.

### 17.4.13 Installation

- Download the web application archive (*war*) file from <http://ambit.sourceforge.net/>
- Deploy the *war* file into a servlet container
- Ensure MySQL is installed and running at the default port
- Create an empty database by issuing a POST request to <http://host:8080/ambit2/admin/database> URI as shown in the command below. Note: *mysqlprivuser* should be an existing MySQL user with sufficient privileges to create a database.

```
curl -X POST -d "dbname = ambit2" -d "user = mysqlprivuser" -d "pass = mysqlprivpass" http://host:8080/ambit2/admin/database
```

- On success, reading the URI <http://host:8080/ambit2/admin/database> will return the database name
- Import your own data by sending a POST command to <http://host:8080/ambit2/dataset> or using the web interface. Use the OpenTox API to run algorithms and models.

**Plans for future developments** include protecting resources via the OpenTox Authentication and Authorization API<sup>54</sup>, which relies on a customized OpenAM<sup>55</sup> service; extend dataset and feature representations to accommodate hierarchical data; develop applications with specialized user interfaces that would consume the services; improve and refactor the services' implementation in order to provide a skeleton code for easy deployment of third party algorithms and models, compliant with the OpenTox API; provide a client library for accessing the OpenTox API.

## 17.5 Conclusions

The AMBIT REST services package has been developed as an extension of AMBIT modules, wrapping their functionalities as REST web services, and adding some new ones. This implementation covers a large subset of the functionality, specified by the OpenTox API, and is available both as live demo online web services and as a downloadable web application, which can be deployed in a compatible servlet container. The services, implementing the OpenTox API for compounds, datasets, and features, enable importing arbitrary files with chemical structures and their properties, allowing linking to computer readable information about the data fields, as well as keeping provenance information. In addition, they support multiple structures of the same compound, which is useful for storing and working with multiple conformations, as well as for comparing structures, originally residing in different source databases. Uploading a file with chemical structures and data makes it automatically available in several formats, including the mandatory RDF representation, defined by the OpenTox ontology. The services, implementing the OpenTox API for algorithms and models, provide a unified web service interface to several descriptor calculation, machine learning, and similarity searching algorithms, as well as to applicability domain and toxicity prediction models. The complexity and diversity of the processing is reduced to the simple paradigm “read data from a web address, perform processing, write to a

---

<sup>53</sup> ToxPredict demo application

<sup>54</sup> OpenTox Authentication and Authorisation API

<sup>55</sup> OpenAM

web address". The online service allows running predictions without installing any software, as well sharing datasets and models between online users. The downloadable web application allows researchers to set up their own systems of chemical compounds, calculated and experimental data, and to run existing algorithms and create new models. The advantage of exposing the functionality via the OpenTox API is that all these resources could interoperate seamlessly, not only within a single web application, but also in a network of many cooperating distributed services.

Exposing functionalities through a web application programming interface allows to hide the implementation details of both data storage (different database types vs. memory vs. file system backend) and processing (descriptor calculation algorithms using CDK, OpenBabel, commercial or in-house implementations). The availability of data and processing resources as RDF facilitates integrating the resources as Linked Data<sup>56</sup>. The distributed algorithm and model resources automatically generate RDF representations, making the linked data dynamic, and not relying on a single traditional triple storage. The classes in the OpenTox ontology are designed to cover the minimum number of building blocks, necessary to create predictive toxicology applications. The OpenTox ontology relies on external ontologies to represent descriptor calculation algorithms, machine learning methods, and toxicity studies. We believe that such modularity better reflects how any particular domain is described in reality<sup>57</sup>, compared to monolithic ontologies, which could be difficult or even impossible to reach consensus on, and would be hard to maintain. RDF natively provides means to link multiple concepts to a same resource, either by multiple inheritance, or owl:sameAs links, and we intend to use these techniques, together with the current dataset representation, to describe complex toxicological studies.

The AMBIT REST services package is one of the several independent implementations of the OpenTox Application Programming Interface, being developed within the OpenTox project. While creating an ontology (even for a rather limited domain) by consensus is a challenge by itself, the value of having multiple independent implementations of services using the ontology is enormous, as it clearly highlights components that have not been explicitly described, and are thus open to diverging and possibly conflicting interpretations. This demonstrates that the OpenTox API could be implemented equally well either as a completely new project or as an extension of an existing software. It also demonstrates OpenTox API's ability to provide a unified interface to diverse algorithms and data, and to encourage defining explicit relationships between data and processing routines. Last but not least, the services provide a sound basis for building web mashups, end user applications with friendly GUIs, as well as embedding the functionalities in existing workflow systems.

## 17.6 Availability and requirements

- **Project name:** AMBIT implementation of OpenTox REST web services
- **Project home page:** <http://ambit.sourceforge.net/>
- **Operating system(s):** Platform independent
- **Programming language:** Java
- **Other requirements:** Java 1.6 or higher, Tomcat 6.0 or higher, MySQL 5.1 or higher
- **License:** GNU LGPL (ambit2-all modules), GNU GPL (web services)
- **Any restrictions to use by non-academics:** None
- **Online web services:** <http://apps.ideaconsult.net:8080/ambit2/>

## 17.7 Abbreviations

API: Application Programming Interface; CDK: The Chemistry Development Kit; HTTP: Hypertext Transfer Protocol; MIME: Multipurpose Internet Mail Extensions; (Q)SAR: (Quantitative) Structure Activity Relationship; REST: REpresentational State Transfer; RDF: Resource Description Framework; URI: Universal Resource Identifier.

<sup>56</sup> Wikipedia entry for Linked Data

<sup>57</sup> Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies

## 17.8 Competing interests

The authors declare that they have no competing interests.

## 17.9 Authors' contributions

NJ performed most of the AMBIT web services software development, coordinated the OpenTox framework implementation, and provided valuable insights on a range of important aspects of the manuscript. VJ contributed to the design and testing of the webservices, performed data gathering and curation, and helped drafting the manuscript. All authors read and approved the final manuscript.

## 17.10 Authors' information

Nina Jeliazkova (NJ): Nina received a M.Sc. in Computer Science from the Institute for Fine Mechanics and Optics, St. Petersburg, Russia in 1991, followed by a PhD in Computer Science (thesis “Novel computer methods for molecular modelling”) in 2001 in Sofia, Bulgaria, and a PostDoc at the Central Product Safety department, Procter & Gamble, Brussels, Belgium (2002 - 2003). Her professional career started as a software developer first at the oil refinery Neftochim at Burgas, Bulgaria (1991 - 1995), then at the Laboratory for Mathematical Chemistry, Burgas, Bulgaria (1996 - 2001). She joined the Bulgarian Academy of Sciences in 1996 as a researcher and network engineer at the Network Operating Centre of the Bulgarian National Research and Education Network. She is founder and co-owner of Ideaconult Ltd, and is technical manager of the company since 2009. She participated in a number of R&D projects in Belgium and Bulgaria, authored and co-authored about 40 scientific papers in Bulgarian and international journals, as well as several book chapters.

Vedrin Jeliazkov (VJ): Vedrin received a M.Sc. in Computer Science from the University Paris VII - Denis Diderot, Paris, France in 1998. From 1996 to 1998 he worked for the R&D department of Electricité de France, Clamart, France, as a software developer, and was responsible for the design of quality assurance tests. From 2001 to 2002 Vedrin had been employed by the European Commission as an advisor to the director of “Essential Technologies and Infrastructures” at the Information Society Directorate-General. From 2003 to 2007 he was researcher at the Bulgarian Academy of Sciences and network engineer at the Network Operating Centre of the Bulgarian National Research and Education Network. Vedrin is one of the founders and owners of Ideaconult Ltd, and is a full-time researcher at the company since 2007. He participated in a number of R&D projects in France, Belgium, and Bulgaria, authored ten research papers, co-authored one book and has given several talks in scientific conferences.

## 17.11 Acknowledgements and Funding

The AMBIT software was initially developed within the framework of the CEFIC LRI project EEM-9 ‘Building blocks for a future (Q)SAR decision support system: databases, applicability domain, similarity assessment and structure conversions’, [http://www.cefic-lri.org/projects/1202813618/21/EEM9-PGEU-Quantitative-Structure-Activity-Relationships-software-for-data-management/?cntnt01template=display\\_list\\_test](http://www.cefic-lri.org/projects/1202813618/21/EEM9-PGEU-Quantitative-Structure-Activity-Relationships-software-for-data-management/?cntnt01template=display_list_test), and further extended under a subsequent CEFIC LRI contract for developing AmbitXT, <http://www.cefic-lri.org/lri-toolbox/ambit>. The AMBIT web services package was developed within the OpenTox project - An Open Source Predictive Toxicology Framework, <http://www.opentox.org/> funded under the EU Seventh Framework Programme: HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs (Quantitative Structure-Activity Relationships) for toxicology, Project Reference Number Health-F5-2008-200787 (2008-2011). Support for applicability domain algorithms has been integrated into the AMBIT web services package in the framework of the CADASTER project - Case studies on the development and application of in-silico techniques for environmental hazard and risk assessment, <http://www.cadaster.eu/>, funded under the EU Seventh Framework Programme: ENV.2007.3.3.1.1 In-silico techniques for hazard-, safety-, and environmental risk-assessment, Project Reference Number 212668 (2009-2012).

We acknowledge the article processing charge for this article that has been partially funded by Pfizer, Inc. Pfizer, Inc. has had no input into the content of the article. The article has been independently prepared by the authors and been subject to the journal's standard peer review process.



# LINKED OPEN DRUG DATA FOR PHARMACEUTICAL RESEARCH AND DEVELOPMENT

## 18.1 Abstract

There is an abundance of information about drugs available on the Web. Data sources range from medicinal chemistry results, over the impact of drugs on gene expression, to the outcomes of drugs in clinical trials. These data are typically not connected together, which reduces the ease with which insights can be gained. Linking Open Drug Data (LODD) is a task force within the World Wide Web Consortium's (W3C) Health Care and Life Sciences Interest Group (HCLS IG). LODD has surveyed publicly available data about drugs, created Linked Data representations of the data sets, and identified interesting scientific and business questions that can be answered once the data sets are connected. The task force provides recommendations for the best practices of exposing data in a Linked Data representation. In this paper, we present past and ongoing work of LODD and discuss the growing importance of Linked Data as a foundation for pharmaceutical R&D data sharing.

## 18.2 Findings

Pharmaceutical research has a wealth of available data sources to help elucidate the complex biological mechanisms that lead to the development of diseases. However, the heterogeneous nature of these data and their widespread distribution over journal articles, patents and numerous databases makes searching and pattern discovery a tedious and manual task. From the perspective of a pharmaceutical research scientist, the ideal data infrastructure should make it easy to link and search across open data sources in order to identify novel and meaningful correlations and mechanisms. In this paper, we present work from the Linked Open Drug Data (LODD) task force of the World Wide Web Consortium (W3C) Health Care and Life Science Interest Group (HCLS IG) that aims to address these issues by harnessing the power of new web technologies.

The LODD task force works with a set of technologies and conventions that are now commonly referred to as *Linked Data*. The primary goal of the Linked Data movement is to make the World Wide Web not only useful for sharing and interlinking documents, but also for sharing and interlinking *data* at very detailed levels. The movement is driven by the hypothesis that these technologies could revolutionize global data sharing, integration and analysis, just like the classic Web revolutionized information sharing and communication over the last two decades.

Linked Data is based on a set of principles and standard recommendations created by the W3C. Single data points are identified with Hypertext Transfer Protocol (HTTP,<sup>1</sup>) Uniform Resource Identifiers (URIs). Similar to how a Web page can be retrieved by resolving its HTTP URI (e.g., ‘<http://en.wikipedia.org/wiki/Presenilin>’), data about a single entity

---

<sup>1</sup> HTTP Specifications and Drafts

in the Linked Data space can be retrieved by resolving its HTTP URI (e.g. ‘<http://dbpedia.org/resource/Presenilin>’). However, instead of Web pages, the primary data model of Linked Data is the Resource Description Framework (*RDF*<sup>2</sup>, *[#B2J]*). In *RDF*, entities, their relations and properties are described with simple subject-predicate-object \*triples. Out of these simple triples, sophisticated networks of interlinked data can be built, potentially spanning over several different locations on the web. Since every entity in this network can be resolved through HTTP, it is possible to navigate and aggregate the globally distributed data, enabling the important features of transparency and scalability that made the Web successful.

There is a large array of other standard recommendations based on *RDF*. Networks of *RDF* data can be queried by an intuitive and powerful query language called *SPARQL*<sup>2</sup>. The Web Ontology Language (*OWL*,<sup>3</sup>) makes it possible to do complex logical reasoning and consistency checking of *RDF/OWL* resources. These reasoning capabilities can be used to harmonize heterogeneous data structures. Another related standard is *RDFa*<sup>4</sup>, which makes it possible to embed *RDF* statements into human-readable Web pages, effectively bridging the domains of human-readable and machine-readable data. Chen et al. provide an extensive review of *RDF/OWL* - based projects relevant to drug discovery in a recent publication<sup>5</sup>.

To date, participants of the LODD project have made twelve open-access datasets relevant to pharmaceutical research and development available as Linked Data (table 1). These are DrugBank<sup>6</sup>, ClinicalTrials.gov<sup>78</sup>, DailyMed<sup>9</sup>, ChEMBL<sup>1011</sup>, Diseaseome<sup>12</sup>, TCMGeneDIT<sup>1314</sup>, SIDER<sup>15</sup>, STITCH<sup>16</sup>, the Medicare formulary and the three most recent additions, RxNorm<sup>17</sup>, Unified Medical Language System (UMLS,<sup>18</sup>) and the WHO Global Health Observatory<sup>19</sup>. To be kept up to date, the original datasets are periodically retrieved and the Linked Data representations are refreshed. The URIs for representing entities in the linked datasets are stable and are chosen by the LODD participants.

Not all of these datasets can currently be considered fully ‘open’ as outlined by the Panton Principles<sup>20</sup>. For example, some of the source have non-commercial clauses in the license agreement. The LODD project is actively exploring the exact conditions for modification and redistribution defined by the data providers, and acknowledges the limitations with respect to openness some of these datasets currently have.

The LODD datasets are linked with each other, as well as with datasets provided by other Linked Data projects, such as Bio2RDF<sup>21</sup> and Chem2Bio2RDF<sup>22</sup>, as well as primary data providers that offer their resources in *RDF*, such as UniProt<sup>2324</sup> and the Allen Brain Atlas<sup>25</sup>. The links between datasets are depicted in Figure 1. Overall, there are several dozens of biomedically relevant linked datasets available to date.

While the number of linked biomedical datasets has grown significantly over the last years, there is still a marked lack of mature applications that enable end-users to explore and query these datasets. Linked data browsers such

<sup>2</sup> SPARQL Query Language for *RDF*

<sup>3</sup> OWL Web Ontology Language Overview

<sup>4</sup> RDFa Primer

<sup>5</sup> The use of web ontology languages and other semantic web tools in drug discovery

<sup>6</sup> DrugBank: a comprehensive resource for *in silico* drug discovery and exploration

<sup>7</sup> LinkedCT: A Linked Data Space for Clinical Trials

<sup>8</sup> Home - ClinicalTrials.gov

<sup>9</sup> DailyMed: About DailyMed

<sup>10</sup> Role of open chemical data in aiding drug discovery and design

<sup>11</sup> ChEMBL

<sup>12</sup> The human disease network

<sup>13</sup> Publishing Chinese medicine knowledge as Linked Data on the Web

<sup>14</sup> TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining

<sup>15</sup> A side effect resource to capture phenotypic effects of drugs

<sup>16</sup> STITCH: interaction networks of chemicals and proteins

<sup>17</sup> RxNorm: prescription for electronic drug information exchange

<sup>18</sup> The Unified Medical Language System (UMLS): integrating biomedical terminology

<sup>19</sup> WHO | Global Health Observatory (GHO)

<sup>20</sup> Panton Principles

<sup>21</sup> Bio2RDF: Towards a mashup to build bioinformatics knowledge systems

<sup>22</sup> Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data

<sup>23</sup> UniProt

<sup>24</sup> The Universal Protein Resource (UniProt) in 2010

<sup>25</sup> Allen Brain Atlas: Home

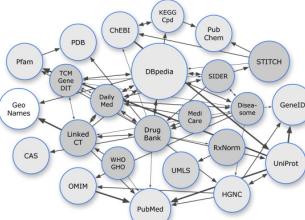


Figure 18.1: Figure 1. A graph of some of the LODD datasets (dark grey), related biomedical datasets (light grey), related general-purpose datasets (white) and their interconnections

**A graph of some of the LODD datasets (dark grey), related biomedical datasets (light grey), related general-purpose datasets (white) and their interconnections.** Line weights correspond to the number of links. The direction of an arrow indicates the dataset that contains the links, e.g., an arrow from A to B means that dataset A contains RDF triples that use identifiers from B. Bidirectional arrows usually indicate that the links are mirrored in both datasets.

as Marbles<sup>26</sup> or Sig.ma<sup>2728</sup> are currently too generic for most end-users (although they can be very helpful for developers). These shortcomings are addressed by TripleMap (Figure 2,<sup>29</sup>), a new web-based application that can be used for the navigation, visualization and analysis of the LODD resources and other RDF datasets. To illustrate the use of TripleMap and the LODD resources, the following simple scenario could be imagined: A researcher interested in Alzheimer’s Disease decides to find out everything that they can about the disease by querying an integrated version of the Linking Open Drug Data (LODD) sets. They open TripleMap and start their search by typing “Alzheimer’s” into the Diseases search box. As they type, TripleMap provides a dynamic auto-complete list of all of the disease related entities across all LODD data sets that match the search string. The researcher selects “Alzheimer’s Disease” and drags and drops it into the TripleMap workspace. Now, the researcher can view a range of information known about the properties of the disease in the right-hand “properties panel” including links out to Pubmed, Online Mendelian Inheritance in Man (OMIM,<sup>30</sup>) Uniprot<sup>24</sup> and other sources. These sources provide the user with rapid access to overview information about the disease.

The researcher is now interested in discovering entities that are associated with Alzheimer’s Disease. They select the Alzheimer’s Disease icon in the workspace and the system automatically shows them a number of associated disease genes provided by Diseaseome, compounds provided by DrugBank and DailyMed, and clinical trials provided by LinkedCT. The researcher starts to explore relationships between entities by selecting two genes, presenilin (PSEN1) and amyloid precursor protein (APP), and dragging them into the workspace. In addition to finding genes related to Alzheimer’s Disease, the user is interested in compounds known to be related to the disease. The user finds several compounds and pulls them into the workspace. The user is also interested in finding out what clinical trials are currently being run for Alzheimer’s Disease and the system shows 200 such trials. With a simple click and drag action they pull all 200 trials into the workspace. As entities are added to the workspace, if there are known associations between them, those associations are also shown to the user as semantically tagged edges. This ability to show a researcher unexpected associations between entities that are related to their field of interest is at the heart of the value of an application like TripleMap and the extensive, rich, interconnected data available in the LODD data sets.

Linked Data as an emerging technology is still not free from shortcomings. A major problem is the heterogeneity in how data is modeled. Even when the entities between datasets are mapped to each other, it can still be difficult to intuitively write queries that span datasets because of this heterogeneity. This problem is being addressed by another task force of the W3C HCLSIG, which aims to bridge the data in the growing number of LODD datasets with a well-engineered top-level ontology, the translational medicine ontology (TMO,<sup>31</sup>). Another problem is how to efficiently query RDF in distributed SPARQL databases without requiring the aggregation of RDF data at a central location.

<sup>26</sup> Marbles Linked Data Engine

<sup>27</sup> Sig.ma: Live views on the Web of Data

<sup>28</sup> sig.ma - Semantic Information MASHup

<sup>29</sup> TripleMap

<sup>30</sup> OMIM Home

<sup>31</sup> The Translational Medicine Ontology: Driving personalized medicine by bridging the gap from bedside to bench



Figure 18.2: Figure 2. TripleMap <https://www.triplemap.com> is a web-based application that provides a rich, dynamic, visual interface to integrated RDF datasets such as the LODD. On the left hand side of the application a researcher uses an icon-based menu representing biomedical entities such as compounds, diseases and assays to search for entities and view their associations

**TripleMap** \*[www.triplemap.com](http://www.triplemap.com)\* \*\* is a web-based application that provides a rich, dynamic, visual interface to integrated RDF datasets such as the LODD. On the left hand side of the application a researcher uses an icon-based menu representing biomedical entities such as compounds, diseases and assays to search for entities and view their associations\*\*. Entities can be dragged and dropped from the icon menu into the application's zoomable workspace. In the middle of the application the user navigates maps of entities and their associations in the zoomable workspace much like users of Google Maps are able to scan and zoom into and out of geographically based maps. On the right hand side of the application the user can view an integrated set of all of the available properties for a selected entity. As entities are added to the workspace the system automatically generates semantically tagged edges between associated entities.

Again, this is addressed by ongoing work on query federation by members of the W3C HCLS IG<sup>32</sup>. Finally, there has been a lack of applications with good user interfaces to make Linked Data resources accessible to end-users outside the biomedical informatics community. This is addressed by several ongoing endeavors such as the European Khresmōi project<sup>33</sup>.

A challenge to creating linked data that is specific to the domain of chemistry is the provision of chemical identifiers. It is for this reason that W3C HCLS IG supports efforts to standardize unique identifiers for chemical compounds such as the IUPAC International Chemical Identifier (InChI,<sup>34</sup>).

The pharmaceutical industry is starting to embrace Linked Data with examples of projects being presented by Eli Lilly, Johnson & Johnson and UCB Pharma. While the adoption of Linked Data is still not yet very widespread in individual companies, it is on the agenda of several large-scale cross-pharma projects. An European project, the Open Pharmacological Space (OPS) Open PHACTS (Pharmacological Concept Triple Store) project under the European Innovative Medicines Initiative (IMI,<sup>35</sup>) wants to create an open source, open standards and open access infrastructure to enable integration of chemical and biological data to support drug discovery. The project intends to reach this goal by using Linked Data and managing the data in an RDF triple store. Collaboration across several IMI projects should also encourage the coordinated use of Linked Data to enhance data sharing. On the pre-competitive data sharing side of pharmaceutical informatics, the members of the Pistoia Alliance<sup>36</sup> are developing the Semantically Enriched Scientific Literature (SESL) project. The goal of SESL is to test the feasibility of executing federated querying across full text literature and bioinformatics databases by performing SPARQL queries on a triple store of assertions from the chosen data sources. The PRISM Forum<sup>37</sup> has also issued a letter recommending the adoption of Linked Data that has been supported by its membership of 15 of the top 20 pharmaceutical companies. The European OpenTox project<sup>3839</sup> uses RDF as a standard for the exchange of predictive toxicology related data. The OpenTox framework defines algorithms, models, data sets, and chemical compounds, in a distributed data storage and computing facility.

Proprietary systems for providing integrated pharmaceutical data exist. The Accelrys/Symyx products<sup>40</sup> are popular examples, and can be both accessed online or installed locally. Accessing the data provided by these products often requires proprietary tools and internal installations also require ongoing work to be kept up-to-date. Furthermore, many of these products are based on individual databases that are not linked. Since the amount of data and the number of potential data sources is growing, it will become harder for single software vendors to create all-encompassing solutions. The nascent Linked Data infrastructure could help to make the creation of integrated solution more sustainable, easier to maintain and vendor-neutral.

Over the next years, the LODD group will continue to work jointly with both academic and industry partners. It will aim to become an umbrella for other Linked Data providers and consumers in the pharmaceutical domain, assisting with documentation, interlinking, quality management, and compliance with standard formats and vocabularies. Another strand of work will focus on how to integrate public Linked Data with non-public, in-house datasets of biomedical research institutions and pharmaceutical companies.

The LODD task force is open to new participants and interested individuals or groups are invited to get in contact with the authors of this paper.

## 18.3 Competing interests

CB declares association with Entagen, LLC, a for-profit company that is building commercial software for semantic technologies such as TripleMap. All other authors declare no competing interests.

<sup>32</sup> A journey to Semantic Web query federation in the life sciences

<sup>33</sup> Khresmōi - Medical Information Analysis and Retrieval

<sup>34</sup> International Union of Pure and Applied Chemistry

<sup>35</sup> Home | IMI - Innovative Medicines Initiative

<sup>36</sup> Pistoia Alliance | Open standards for data and technology interfaces in the life science research industry

<sup>37</sup> The PRISM Forum Association - Home

<sup>38</sup> Collaborative development of predictive toxicology applications

<sup>39</sup> Welcome to the OpenTox Community site – OpenTox

<sup>40</sup> Scientific Informatics Software for Life Sciences, Materials R&D | Accelrys

## 18.4 Authors' contributions

MS wrote major parts of the manuscript and organized the paper writing process. AJ converted several of the LODD datasets. CB developed the TripleMap software. SS organized the Linked Open Drug Data task force. All authors participated in discussions and developments of the Linked Open Drug Data task force of the W3C Health Care and Life Science Interest Group. All authors read and approved the final manuscript.

## 18.5 Acknowledgements

We acknowledge the article processing charge for this article that has been partially funded by Pfizer, Inc. Pfizer, Inc. has had no input into the content of the article. The article has been independently prepared by the authors and been subject to the journal's standard peer review process.

# CHEMICAL ENTITY SEMANTIC SPECIFICATION: KNOWLEDGE REPRESENTATION FOR EFFICIENT SEMANTIC CHEMINFORMATICS AND FACILE DATA INTEGRATION

## 19.1 Abstract

### 19.1.1 Background

Over the past several centuries, chemistry has permeated virtually every facet of human lifestyle, enriching fields as diverse as medicine, agriculture, manufacturing, warfare, and electronics, among numerous others. Unfortunately, application-specific, incompatible chemical information formats and representation strategies have emerged as a result of such diverse adoption of chemistry. Although a number of efforts have been dedicated to unifying the computational representation of chemical information, disparities between the various chemical databases still persist and stand in the way of cross-domain, interdisciplinary investigations. Through a common syntax and formal semantics, Semantic Web technology offers the ability to accurately represent, integrate, reason about and query across diverse chemical information.

### 19.1.2 Results

Here we specify and implement the Chemical Entity Semantic Specification (CHESS) for the representation of polyatomic chemical entities, their substructures, bonds, atoms, and reactions using Semantic Web technologies. CHESS provides means to capture aspects of their corresponding chemical descriptors, connectivity, functional composition, and geometric structure while specifying mechanisms for data provenance. We demonstrate that using our readily extensible specification, it is possible to efficiently integrate multiple disparate chemical data sources, while retaining appropriate correspondence of chemical descriptors, with very little additional effort. We demonstrate the impact of some of our representational decisions on the performance of chemically-aware knowledgebase searching and rudimentary reaction candidate selection. Finally, we provide access to the tools necessary to carry out chemical entity encoding in CHESS, along with a sample knowledgebase.

### 19.1.3 Conclusions

By harnessing the power of Semantic Web technologies with CHESS, it is possible to provide a means of facile cross-domain chemical knowledge integration with full preservation of data correspondence and provenance. Our representation builds on existing cheminformatics technologies and, by the virtue of RDF specification, remains flexible and amenable to application- and domain-specific annotations without compromising chemical data integration. We conclude that the adoption of a consistent and semantically-enabled chemical specification is imperative for surviving the coming chemical data deluge and supporting systems science research.

## 19.2 Background

The importance of cataloguing and adequately representing chemical information has been realized fairly early in the development of chemistry and related sciences. From the dawn of the era of organic synthesis, thousands of chemical entities, reactions, and experimental outcomes were catalogued and stored in a human-readable form, some dating to as early as the eighteenth century when the understanding of molecular reactivity and chemical structure was nowhere near its current level (preserved in e.g.<sup>1</sup>). During the relatively long history of the development of chemical information archiving technologies, a large number of persistent redundancies and factors complicating chemical knowledge federation have been introduced. It may be argued, however, that these problems may be reduced to three major categories, some of which have been only recently partially addressed: i) a lack of a consensus canonical identifiers of all chemical entities, including reactions and macromolecules, as well as and their constituents, ii) absence of a single common flexible representation to satisfy the needs of most sub-disciplines of chemistry, and iii) a lack of a consensus chemical database structure or schema. We argue that the bulk of present-day complications in integrating chemical information can be traced to these three problems and that until an information representation that addresses these issues is introduced, truly integrative chemical research shall be a complicated and costly endeavour.

### 19.2.1 Consensus Chemical Entity Identifiers

Many modern chemical databases rely on internal chemical entity identifiers which are usually created in a sequential manner, producing an index whose value increases for every new chemical entity in the database<sup>2345</sup>. Unfortunately, such indexing systems require manual or semi-automated cross-database matching, resulting in difficulties of federating chemical data. Attempts to create reproducible and canonical molecular identifiers that bear molecular graph information could be dated to the introduction of computers to the field of chemistry in the 20<sup>th</sup> century. The fact that molecules could be represented as graphs and features of these graphs could be used to arrive at a shorthand depiction of molecular structure, had swiftly led to the creation of the first fragment-based line notation in chemistry: the Wiswesser Line Notation (WLN)<sup>6</sup>. In this notation, molecular fragments were abbreviated and recorded with a limited character set to reconstitute various molecular parts and their connectivity. Unfortunately, no efficient way to create a *canonical* molecular representation for this line notation existed, meaning that a given molecule could be referred to by multiple different WLN strings in different chemical databases. This shortcoming was overcome with the introduction of the SMILES notation<sup>7</sup>, which explicitly represented chemical molecules as graphs with atoms being nodes and bonds edges, along with an efficient algorithm to create a canonical, reproducible SMILES string representation of a given molecule. Unfortunately, multiple algorithms for SMILES canonicalization have been devised over the years, leading to software-, and therefore, database-specific canonical molecular SMILES representations. Finally, The International Chemical Identifier (InChI) notation has addressed this issue by providing algorithms and software to enable consistent canonical representation of chemical structures, but has unfortunately not yet addressed

---

<sup>1</sup> Crossfire Database Suite

<sup>2</sup> Chemical Entities of Biological Interest: an update

<sup>3</sup> PubChem as a public resource for drug discovery

<sup>4</sup> ChemSpider Database of Chemical Structures and Property Predictions

<sup>5</sup> ChEMBL Database

<sup>6</sup> How the WLN began in 1949 and how it might be in 1999

<sup>7</sup> SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules

the efficient canonical representation of many other chemical entities, such as reactions and macromolecules<sup>8</sup>. More recently, due to the unwieldy nature of InChI for many larger molecules as well as web search engine complications, InChI keys have been introduced, producing a 25-character hash based on the elements of chemical graph structure<sup>8</sup>. Although InChI keys cannot be used to reconstitute chemical structure without lookup tables, their use has enabled cross-database chemical searches using common web search engines. It is not unreasonable to believe that a universal adoption of the IUPAC standard InChI keys in the role of database indexes could potentially facilitate knowledge federation immensely.

## 19.2.2 Common Chemical Information Representation

Simple line notations have been useful as chemical structure identifiers and bearers of information necessary for the vast majority of cheminformatics tasks, such as chemical database searching or basic reactive transformation outcome predictions. However, these notations could not address the needs of structural, biological, and computational chemists, among others. For this purpose, myriads of chemical file formats incorporating elements of discipline-specific controlled annotations and geometric molecular configuration, have been developed over the past half century. One of the most popular formats to address this need has been the Structure-Data File (SDF)<sup>9</sup>, which combines molecular structural and atomic connectivity information with data annotations. Unfortunately, these annotations may often be confusing or contradictory, as they commonly bear no units, data source information, or specific references to the moieties or molecular entities to which the annotations correspond. For instance, an octanol-water partition coefficient annotation may be specified as follows.

While it may be a straightforward annotation to the creators of a given database, and while it may be more or less easily interpreted by a human agent, it bears no information with respect to corresponding units, algorithms, or parameters used in generating this value. Furthermore, if two different databases containing SDF data for partition coefficients for the same molecule were to be integrated, this integration would require human interpretation, specialized parsers, and if a relational database is used, a special field to store this information in order to enable queries over it. This task is convoluted by the limited availability of annotation specification or outright lack thereof, prompting many cheminformatics applications to re-evaluate descriptor values in a given study.

Thus, an ideal representation would be able to refer to every chemical entity and its part unambiguously and to capture information in a controlled, reproducible, and machine-understandable way to enable machine reasoning and to facilitate data integration. To address this, numerous XML-based chemical representation schemes have been created (e.g. <sup>10</sup><sup>11</sup><sup>12</sup>), enabling highly detailed reaction modeling and chemical representations but unfortunately they have neither been widely embraced by the chemical community, nor have they allowed for seamless machine-mediated information integration. One XML-based representation, the Chemical Markup Language<sup>13</sup>, backed by a controlled vocabulary, has been rather successful in specifying most aspects of chemistry, from small molecules and their connectivity to polymers and crystal structures<sup>14</sup>.

Unfortunately, while most elements of this specification can be parsed out using one of the many XML libraries, certain elements do not render themselves to facile interpretation. Consider the sample CML specification of a water molecule (Figure 1). In order to identify the member atoms in a given bond, it is necessary to carry out string processing as an intermediate step. Further, while many of the elements of CML are defined in a controlled vocabulary, the lack of explicit, consistent, and formal axiomatization of the involved concepts gives rise to difficulties in inferring connections between chemical concepts where no such connections are stated explicitly, something that is possible in formal ontology-backed RDF-based information specifications. Although CML specifications have been increasingly evolving to incorporate elements of the Semantic Web, the lack of widespread adoption of the format, and the limited availability of large-scale CML-based chemical knowledge repositories, have somewhat limited CML-assisted federation of the world of chemical data. Furthermore, the implementation of coverage of additional chemical concepts in

<sup>8</sup> The IUPAC International Chemical Identifier: InChI - A New Standard for Molecular Informatics

<sup>9</sup> Connectivity Table File Formats

<sup>10</sup> XyM Markup Language (XyMML) for Electronic Communication of Chemical Documents Containing Structural Formulas and Reaction Schemes

<sup>11</sup> Ontology Aided Modeling of Organic Reaction Mechanisms with Flexible and Fragment Based XML Markup Procedures

<sup>12</sup> Model Tool to Describe Chemical Structures in XML Format Utilizing Structural Fragments and Chemical Ontology

<sup>13</sup> Chemical Markup, XML, and the World Wide Web. 1. Basic Principles

<sup>14</sup> Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language

most chemical representations require a formal, rigorous representation specification, complicating the incorporation of data represented using domain-specific representation extensions. We believe that an ideal chemical representation would require no specialized wrapper or interpreter, would be generic such as to allow for facile and conflict-free extensions, would be based on a formal ontology, and would be encoded in a machine-*understandable* (as opposed to simply machine-readable, as in CML) manner and therefore facilitates automated reasoning and data integration.

```
<?xml version="1.0"?>
<molecule xmlns="http://www.xml-cml.org/schema">
  <atomArray>
    <atom id="a1" elementType="O"/>
    <atom id="a2" elementType="H"/>
    <atom id="a3" elementType="H"/>
  </atomArray>
  <bondArray>
    <bond atomRefs2="a1 a2" order="1"/>
    <bond atomRefs2="a1 a3" order="1"/>
  </bondArray>
</molecule>
```

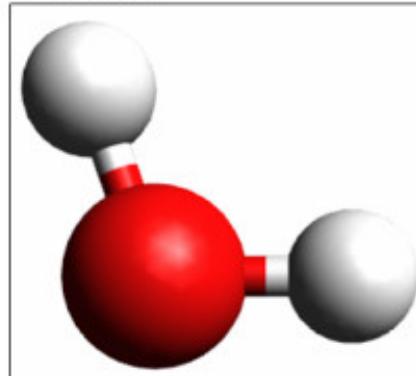


Figure 19.1: Figure 1. A simplified specification of a water molecule in CML  
**A simplified specification of a water molecule in CML.**

### 19.2.3 Chemical Knowledge Integration

The final point of contention in the world of chemical information archiving is a universal open architecture for chemical databases. Chemical data currently exists in a large collection of application- or institution-specific databases that offer little in the way of an integrative searching approach. In fact, many of these databases expect the end user to rely solely on the information that they provide in their research. In the world where cross-discipline borders are increasingly disappearing, such philosophy should have no place or foothold. It should be possible to seamlessly query for, say, the physical properties of a given chemical entity, as well as for its metabolic fates and toxicity data, and ordering information, all from a single interface. Though a number of databases, such as PubChem<sup>3</sup> and ChemSpider<sup>4</sup> currently offer database cross-links to a number of relevant information providers, the perusal and integration of this information still requires human or human-assisted procedures. Furthermore, until an explicit mapping to a given data source is introduced within a database interface, this data source is inaccessible or difficult to access, making the data practically non-existent to the users of these databases. This situation is complicated by the fact that many data repositories and web services that could potentially generate the required data require unique interfaces and access methods, leaving an immense amount of potentially useful information inaccessible.

With the advent of the Semantic Web, a number of these issues have been addressed. With the concept of linked data and the Resource Description Framework (RDF)-based knowledge representations, a new way of modeling, querying and distributing data became available<sup>15</sup>. With RDF, knowledge is represented in terms of subject-predicate-object triples, where each member of a given triple may be a dereferenceable Universal Resource Identifier (URI) for a particular concept or entity. Thus, what was traditionally referred to as a database, could now be considered a knowledge base, as the initially inert data points were given a machine-understandable meaning through reference to formally specified concepts in supporting ontologies. Thus, two entries from two different knowledge bases could be inferred to relate to the same concept even if no such inference had been explicitly stated, through machine reasoning over axioms in the supporting ontologies. Furthermore, truly integrative queries that could draw on the entirety of the linked data web have now become a reality.

<sup>15</sup> Resource Description Framework Specification

A number of efforts<sup>16</sup><sup>17</sup><sup>18</sup><sup>19</sup> have already been successful in the integration of a large portion of chemical information into the linked open data cloud, demonstrating the utility of doing so with successfully fulfilled integrative queries. Such integrative efforts address a number of issues, from the representation of small molecules and their adverse effects to explicit specification of multiple facets of macromolecular structure and interactions. In fact, the chemically-relevant data cloud constitutes a major portion of the entirety of the linked data available on the web<sup>20</sup>. Many of these efforts provide facile means and tools, such as our chemical information RDFization plugin for Open Babel, to represent and distribute any arbitrary chemical information on the Semantic Web<sup>21</sup>. Although these efforts provide a means to of integrating chemical information and of breaking data out of domain or institutional data silos and exposing it for common searches, the triplified data often bears, to a greater or a lesser extent, the same problem as those found in the original databases. For instance, multiple redundant entries for a given molecule, based on database-specific indices may exist without explicit equivalence assertions. Parthood relationships may be missing from such knowledge-bases altogether, and chemical descriptors may be assigned to entities without reference to generating software, experimental conditions, parameters, or data sources. Finally, many of these specifications implement specifications that preclude facile extension of asserted knowledge. For example, a model where an octanol-water partition coefficient is expressed through a predicate and a value, as in ‘ethanol hasPartitionCoefficient 1.5’ is not as readily amenable to specification of the conditions under which this value had been generated as its counterpart where the descriptor is given a URI and is fully annotatable with the required information.

## 19.2.4 Overview

To rectify the aforementioned chemical information integration problems, we propose CHESS, an RDF-based chemical information specification that is backed by the CHEMINF ontology<sup>22</sup>. Due to the expansive nature of the subject of chemical information representation and the limited space and time to present our work, we shall only focus on selected aspects of semantic chemical information encoding with CHESS, emphasizing principles and consequences rather than specification details. Thus, we shall explore the representation of molecules and all of their constituents with the exception of electrons, representation of chemical descriptors, the consequences of our representation in terms of efficiency of chemical database searches *without specialized cheminformatics plugins*, and finally, we shall briefly cover reaction representation and implications of our representation on reaction candidate selection. By no means do we claim that the representation specification presented here is complete, but would like to rather refer the reader elsewhere for a more detailed and rigorous explanation and implementation examples<sup>23</sup>.

## 19.3 Results and Discussion

### 19.3.1 CHESS Representation Overview

The underlying principle in CHESS is to minimize the amount of context-specific rules and regulations, while maximizing the coverage of information represented with the given set of rules. We have also followed an expanded set of principles and requirements in formulating CHESS specification in order to ensure its suitability as a universal chemical exchange language on the Semantic Web, as follows.

1. The most important requirement for CHESS as a universal chemical information framework is the ability to identify and represent chemical entities in a database-, software-, and discipline-independent fashion. For this purpose, we recruit InChI keys and canonical atom numbering arising from the InChI canonicalization algorithm. This also means that atoms, bonds, and functional groups within a molecule have unique and consistent

<sup>16</sup> The semantic smart laboratory: a system for supporting the chemical eScientist

<sup>17</sup> Bio2RDF: towards a mashup to build bioinformatics knowledge systems

<sup>18</sup> Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data

<sup>19</sup> Linking Open Drug Data Project

<sup>20</sup> Linking Open Data Initiative

<sup>21</sup> Chemical Knowledge for the Semantic Web

<sup>22</sup> CHEMINF Ontology

<sup>23</sup> Chemical Entity Semantic Specification

identifiers. Furthermore, all other (physical and informational) entities, such as descriptors, reactions, and macromolecules should also have canonical representations from which consistent identifiers could be obtained.

2. The flexibility and extensibility to support interfacing with and capture of information from a majority of the existing file formats and databases. In previous work, our group has demonstrated OWL serialization of approximately 100 different structure formats with an Open Babel plugin<sup>21</sup>. We continue this trend by providing means to triplify topological and structural information encoded by SMILES, InChI, and SDF chemical formats, as well as by providing an extensible model for specifying chemical properties and descriptors.
3. CHESS should be descriptive enough to provide means for capturing data at various levels of granularity, from atoms to substances, as well as heterogeneity of data from molecular orbitals to chemical reactions. The information thus captured should preserve explicit correspondence to the circumstances of its creation and the other related data points. That is, all the positional descriptors for the atoms in a particular molecule should preserve a correspondence to each other, reconstituting a single conformer, as well as to the parameters and conditions under which they were observed or computed.
4. CHESS should be supportive of Semantic Web Technology-based implementations of basic cheminformatics tasks pertinent to useful analysis of chemical information, such as semantic drug discovery, chemical similarity searching, or reactive pattern matching.
5. Concepts used in CHESS must be backed by a formal ontology in order to facilitate reasoner-mediated integration of chemical information. For the purposes of our specification, we have chosen the CHEMINF ontology.

Thus, the overall specification for CHESS is quite simple, and involves only three broad categories of players: i) chemical entities, comprised of reactions, complexes, molecules, functional groups, bonds, and atoms (but extensible to e.g. electrons, macromolecular assemblies or even subatomic particles), ii) chemical descriptors which could be entity variant or invariant or could be complex and contain multiple CHEMINF ontology-typed descriptors, each with appropriate value, uncertainty, unit, and chemical configuration annotations, and iii) chemical configurations themselves, which reflect upon the sum of relevant conditions under which data has been derived, as well as the sources of data (Figure 2).

Please note that the chemical configuration is a reflection of the sum of the conditions that may change the value of a given descriptor, as well as the data source.

### 19.3.2 Molecular Specification

Let us consider in greater detail the methodology of CHESS specification generation on the example of an ethanol molecule. First of all, it is necessary to decide which chemical entities are of interest in a given study. Here, we shall focus on the molecule itself as well as the connectivity of its constituents, including functional groups, bonds, and atoms. In order to respect the first principle of CHESS, it is imperative to generate unique canonical identifiers for each of these entities, according to a set of simple rules that will consistently result in reproducible, database-independent identifiers, some of which we outline here (Figure 3). Please note that although we use our own base URI in our work [http://semanticscience.org/resource/CHESS\\_](http://semanticscience.org/resource/CHESS_), we would like to invite the wider chemical community to initiate a discussion on adopting a single, standard base URI for semantic chemistry, to which all entities will be assigned.

In order to enable reasoning and inference over this chemical information, each represented entity also has to be explicitly assigned to a class that is defined within a supporting ontology. This is important to enable querying over broad general concept topics, such as the identification of all instances of oxygen atoms in a given database, or the weakest bond of a particular type in a given molecule, for example. For this purpose, we draw on the concepts present in the Chemical Entities of Biological Interest (CHEBI) ontology and the Semanticscience Integrated Ontology (SIO)<sup>24</sup> to assign general classes to the appropriate chemical entities. For instance, the ethanol molecule may be assigned to a general class of molecular entities (CHEBI:23367), or if the correspondence is present, to a more specific class of molecules, such as that of primary alcohols (CHEBI:15734). Functional groups or molecular substructures may also be assigned to a general class within the CHEBI ontology (e.g. CHEBI:33249), bonds to an appropriate SIO

---

<sup>24</sup> Semanticscience Integrated Ontology

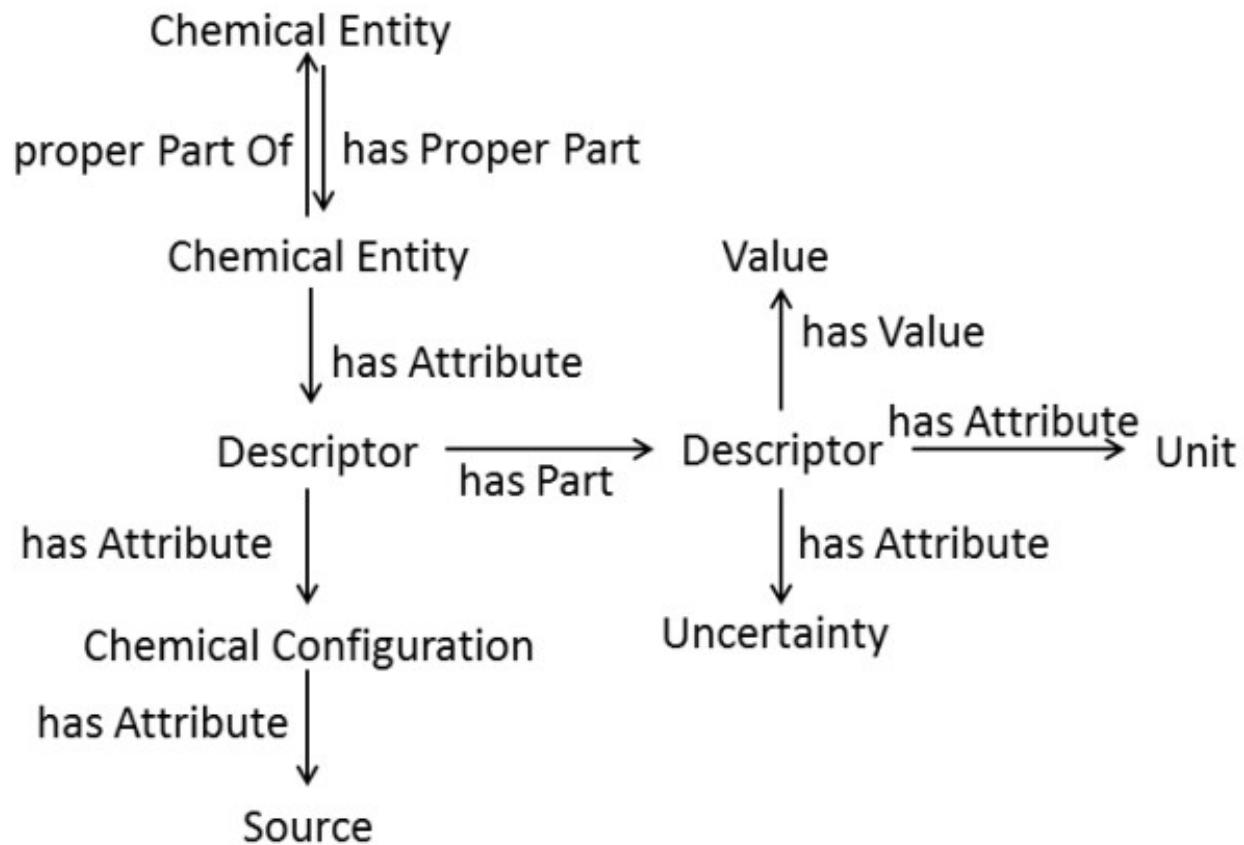


Figure 19.2: Figure 2. A simplified overview of the general features of the CHESS specification.  
A simplified overview of the general features of the CHESS specification. Please note that the chemical configuration is a reflection of the sum of the conditions that may change the value of a given descriptor, as well as the data source.

	Molecule	Standard InChI Key
	Functional Group Instance	Containing Entity InChI Key, FG, Hash: (Canonical Specification, Ordered List of Atoms)
	Bond	Containing Entity InChI Key, B, Canonical Specification, Ordered List of Atom Indices
	Atom	Containing Entity InChI Key, A, Canonical Index, Symbol
	Reaction	R, Hash: Ordered List of Reactant, Catalyst, and Product InChI Keys, Stoichiometric Coefficients

Figure 19.3: Figure 3. Principles for generating canonical identifiers for some of the many chemical entity types  
**Principles for generating canonical identifiers for some of the many chemical entity types.** Please note that these identifiers are for instances of chemical entities rather than classes of chemical entities (e.g. all oxygen atoms or all C-O bonds) and necessarily involve the canonical identifier of their containing entity, molecule in this case.

class (SIO\_011118), atoms to their appropriate types in CHEBI (e.g. CHEBI:25805 for oxygen), and reactions to the *chemical reaction* class in SIO (SIO\_010345). The end-user is not limited to the pre-defined classes in the SIO or CHEBI ontologies. Because these classes are fully extensible, it is possible to define a more specific class for each of the chemical entities presented here. For instance, one may extend SIO's *covalent chemical bond* (SIO\_011118) to create a subclass corresponding to carbon-oxygen single bonds, or extend the broad class of functional groups or molecular substructures in CHEBI to correspond to a class of substructures that satisfy or exactly match a general pattern, such as CCO, as we shall see later.

Though functional group or substructure specification is optional, as is that of any component not relevant to the chemical information represented, in order to demonstrate chemical database searching and reaction candidate matching in this study, we have automatically generated a set of unique atom-centric molecular sub-graphs consisting of heavy atoms and containing first, second, and/or third neighbours of each heavy atom in a given molecule. For example, the oxygen-centered fragmentation products of *n*-propanol of connectivity 1, 2, and 3 are the hydrogen-suppressed graphs CO, CCO, and CCCO, respectively. These fragments are given unique and reproducible identifiers, based on fragment chemical graph structure and the canonical indices of the member heavy atoms, as well as their molecule of origin. It must be noted that the CHESS specification itself is not limited to this automatically generated set of fragments, but rather we are using this set it in order to achieve further goals of enabling chemical similarity searching. Customized fragment or functional group annotations, just like annotations of any other type, may be added to the triple store containing the chemical entities under investigation at any time.

To complete the semantic description of the molecular skeleton, mereological relationships between the various sub-components of a given entity have to be asserted. These relationships are captured with *has proper part* (SIO\_000053). Based on this information, the complete chemical graph can be reconstituted, and our chemical entity under investigation is ready for further annotation or querying. Here, we provide examples of the specification for each entity discussed (Appendix 1).

Appendix 1. RDF/N3 CHESS representation of ethanol and its constituent parts.

```
http://www.w3.org/1999/02/22-rdf-syntax-ns#>
http://semanticscience.org/resource/SIO_>
http://semanticscience.org/resource/CHESS_>
http://purl.org/obo/owl/CHEBI#>
```

### 19.3.3 Semantic Web-Enabled Cheminformatics: Chemical Searching

Since the specification we have so far described can be used to reconstitute the molecular graph, we can now venture to study the most optimal approaches to enabling some of the most common tasks in cheminformatics. Here, we shall initially focus on the classical task of chemical database searching by chemical similarity. Although it is possible to invoke plugins or intermediary specialized software to enable rapid database searching, we argue that unlike many other information representations and formats, CHESS allows us to fully represent the chemical graph and should therefore be readily amenable to graph manipulation and similarity searching. Furthermore, we believe that, within the limit of providing an expressive enough specification of chemical entities under investigation, the efficiency of querying a knowledgebase created using a given knowledge representation is a good indicator of the efficiency of the representation itself.

Chemical similarity searching is a complex topic that lies at the heart of cheminformatics and can be carried out in a wide variety of ways to address a number of problems. Because our specification provides us a complete chemical graph description, we have chosen to first attempt a Semantic Web-native solution for the substructure matching problem, using description logic-safe rules<sup>25</sup> and SPARQL query language-based queries<sup>26</sup> on the molecular structure. As a benchmark, we have used representative subsections of the LIPIDMAPS database<sup>27</sup> of lipids and their structures of sizes 10, 100, and 10000 molecules in partitions DB10, DB100, and DB10000, respectively. As query graphs, we

<sup>25</sup> Semantic Web Rule Language

<sup>26</sup> SPARQL Query Language for RDF

<sup>27</sup> LIPID MAPS online tools for lipid research

have used a series of linear carbon chains, from ethyl to pentyl, cyclopentene, and a number of lipid-related functional groups, including glycerol, sterol, a fatty acyl moiety, a sphingolipid moiety, and a prenol lipid moiety.

While carrying out our tests, we have become aware of the complexity of modeling bonds as explicit entities. While this specification allowed for facile annotation of bonds with properties and descriptors, it resulted in significant search performance hits, forcing us to reconsider elements of our specification. As a result, we have created a test set where, apart from specifying explicit bonds, we also linked bonded atom instances with the appropriate bidirectional relationships that corresponded to bond type (single, double, triple, or aromatic). This improved performance considerably and allowed us to carry out our tests. Herein we find another demonstration of the versatility of semantically enabled information representations: we have been able to extend and amend the information in our knowledge repository without much additional effort or adverse effects on the knowledge repository. So long as our specification is consistent with the formal axioms underpinning the concepts in a supporting ontology, there is no barrier preventing us from extending our specification indefinitely.

One may search for molecules containing a particular sub-graph using the explicit specification of the structure of the sought molecular sub-graph as a SPARQL query (Appendix 2). It must be noted that answering this query in a typical SPARQL engine involves the exhaustive examination of all the candidate molecules that may potentially contain a collection of atoms that satisfy the laid out bonding criteria. Unfortunately, since the SPARQL query engines currently available have not been explicitly optimized chemical searching needs, they often lack many of the mechanisms developed over the past decades to accelerate the solution of this problem, and resemble the brute force approach to graph matching more closely.

Appendix 2. An automatically generated (based on graphical user input) SPARQL query to identify all molecules containing an ethyl subgraph, altered for improved readability. Note the use of ‘has single bond with’ direct atom relationship to improve query performance.

[http://www.w3.org/1999/02/22-rdf-syntax-ns#>](http://www.w3.org/1999/02/22-rdf-syntax-ns#)

[http://semanticscience.org/resource/SIO\\_>](http://semanticscience.org/resource/SIO_>)

[http://semanticscience.org/resource/CHESS\\_>](http://semanticscience.org/resource/CHESS_>)

<http://purl.org/obo/owl/CHEBI#>>

Description logic-safe rules may be used to reason about instances of an OWL-DL ontology where the description follows a graph-like pattern instead of the general ‘tree-like’ expression. In this way, one can combine classical DL-reasoning with graph-like descriptions to classify molecules as more specific kinds of compounds, provided that they satisfy certain sub-graph conditions. For instance, an ethyl compound would be a kind of a molecule that contains an ethyl group with two carbon atoms linked together by a single bond. Overlooking for the sake of simplicity of the immediate example the other necessary conditions, such as the lack of branching or membership in a ring, this may be represented as the following rule (Appendix 3).

Appendix 3. An automatically generated (based on graphical user input) dl-safe rule for identifying an ethyl group-containing molecule, altered for improved readability.

Both, the SPARQL querying, and the rule-based reasoning, completed in the allocated time of 120 seconds on the simple carbon chain-based queries. However, the iterative identification of compounds with more complex substructures, such as those relevant in the classification of lipids, fails with rule reasoning using DB10 for the majority of molecules in the database. In all cases, we observed either memory exhaustion (despite having allocated 5GB of memory), or premature termination. Note that the time limit was imposed after observing that even if Pellet (see chapter “Query Testing”) was allowed to reason over an unconstrained amount of time, this would only delay termination due to memory exhaustion. SPARQL-based query answering, on the other hand, succeeds in identifying molecules containing many fatty acyl patterns, but fails between 20-40% of the time in DB10, and between 60-70% of the time in DB100 for prenol lipid, sphingolipid and sterol lipid SPARQL queries. Generally, SPARQL query completion exhibits a dramatic drop across all test cases with increasing number of atoms in the database molecules searched at a given time (Figure 4). In addition to this discouraging result, the increase of search pattern size or complexity had an even more profound effect on query completion, from exceeding the completion time limit to SPARQL query engine-triggered query termination due to the query computational load exceeding any reasonable expectations, of 11 years, for example (results not shown).

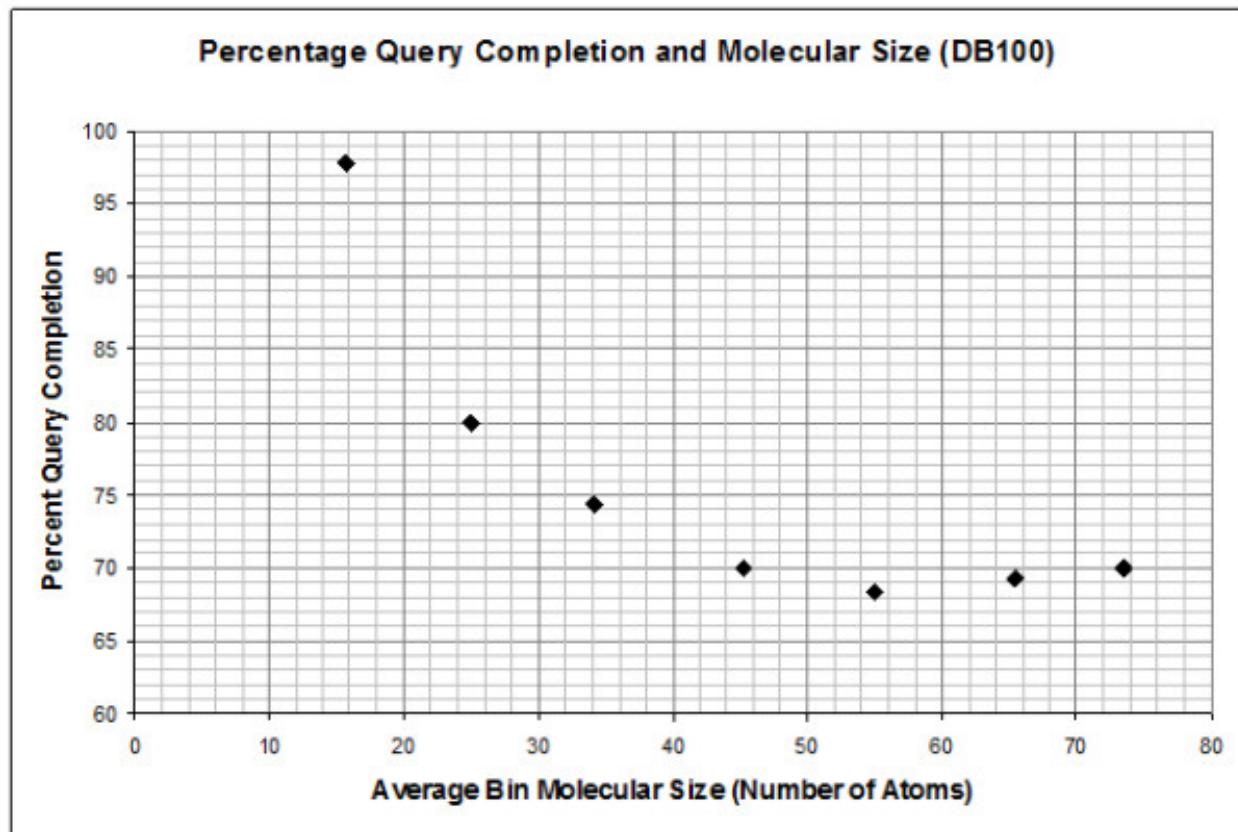


Figure 19.4: Figure 4. SPARQL query completion and average molecular size in DB100  
SPARQL query completion and average molecular size in DB100.

In terms of performance for queries that were successful, both DL-safe rule reasoning and SPARQL-based querying are 1-4 orders of magnitude slower than using Open Babel, and DL-safe rules are 1-2 orders of magnitude slower than SPARQL queries. By comparing the number of atoms in the query structure to the completion time we observe that Babel performance is linear, but the performance of searching with both, SPARQL and DL-safe rules, appears exponential or parabolic (Figure 5).

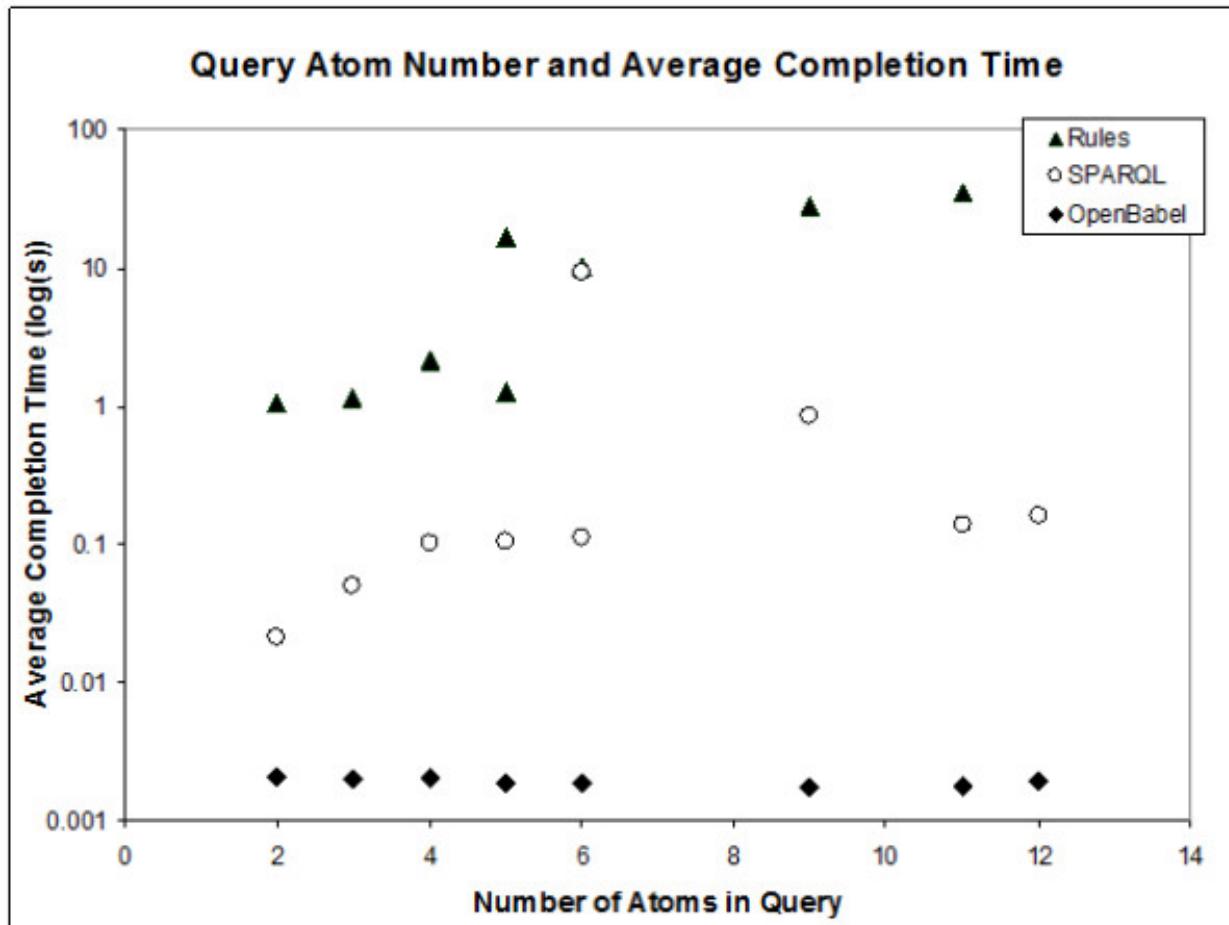


Figure 19.5: Figure 5. General performance trends for the two query modes relative to Open Babel matching  
**General performance trends for the two query modes relative to Open Babel matching.** Note that the two points at six query atoms for SPARQL queries is due to alternate structures: cyclical structures are more time-consuming than linear ones.

Given these discouraging results, even for an extremely small data set, it was clear that an alternative approach or a redefinition of the problem to a more manageable one was in order. As we have discussed, by incorporating functional group/sub-structure information at the time of creating the molecular specification, or by adding it to the existing representation stores, we can redefine our searching problem. By doing so, it becomes possible to use the same device as the one used in fingerprint-based searches: in order to identify database molecules that are similar to the query, it is simply necessary to rank them in terms of the number of substructures that belong to the same class as those in the query. The resultant query exhibits relatively rapid completion times that depend weakly on the complexity of the queried structure and linearly on database size (Appendix 4). Overall, the completion time of this query for DB10000 is well within the stipulated time limit for all the queries attempted (results not shown). Although this is encouraging, many commercially and scientifically important chemical databases enumerate several orders more molecular entities, casting a shadow over the applicability of this approach to large-scale applications, until a more detailed study of search performance demonstrates otherwise.

Appendix 4. The general form of the SPARQL query to identify molecules similar to a queried molecule, which can

be obtained from graphical user input. Query amended for readability.

[http://semanticscience.org/resource/SIO\\_>](http://semanticscience.org/resource/SIO_>)

[http://semanticscience.org/resource/CHESS\\_>](http://semanticscience.org/resource/CHESS_>)

Although we have not tested the more efficient tools currently available, and although the performance of SPARQL query engines<sup>28</sup> and machine reasoners<sup>29</sup> continue to improve, we can see that chemical similarity searches that rely solely on existing Semantic Web tools are possible, but may be problematic for very large chemical knowledge bases. Certainly, we believe that optimization of Semantic Web-based chemical searching solutions with the latest and the most efficient tools, as well as the application of existing tools to very large (over a million molecules) stores of chemical information, warrants further elaboration in an expanded, separate study. Other, significantly more rapid searching solutions that draw on both, the Semantic Web technologies, and existing methodologies, are also available. For example, it is possible to create custom SPARQL functions that encapsulate specialized code to either carry out pairwise similarity comparisons between a query and database molecules by reconstructing and comparing chemical graphs from InChI, SMILES, or SMARTS annotations. It is also possible to envision specialized functions to carry out simple Tanimoto comparison of query and database molecule fingerprint strings. Finally, semantically-enabled web services may be used to carry out such searches<sup>30</sup>. However, the purpose of this excursion has not been the immediate creation of a solution that could outperform specialized and streamlined code carrying out optimized sub-graph detection or similarity calculations on in-memory stores of molecular fingerprint strings<sup>31</sup>. Rather, the fact that it is possible to attempt parser- and specialized tool-free analysis and integration of chemical data, demonstrates the potential, power, and versatility of investigations afforded by adopting a semantic specification of chemical entities.

### 19.3.4 Chemical Descriptor Specification

Having explored in detail semantic specification of various chemical entities and their parts, let us turn our attention to their annotation. Chemical annotations may be classified into two broad types: those that are dependent solely on the composition and the nature of a given entity (e.g. standard InChI strings or heavy atom count), and those that capture empirically or theoretically derived data that varies depending on the circumstances of the chemical entity or data observation (e.g. solubility, computed logP). In CHESS, the two cases are specified through a common approach stipulated in the CHEMINF ontology with one major difference: non-constant descriptors are assigned to a *chemical configuration* that reflects upon the circumstances of descriptor creation and the circumstances of the entity to which these descriptors refer. For instance, the calculated free energy of formation of a molecule in gas phase depends not only upon the computational package employed to derive the value and parameters like temperature or the level of theory used, but also on the geometric configuration of the molecule.

Unlike chemical entities, it is not absolutely crucial that descriptors have canonical and reproducible names, as they are rarely used as focal points around which other annotations are integrated. That is, although it is important to ensure that descriptor identifiers do not clash, thus overwriting or contradicting any data previously assigned to a given descriptor, the precise form of descriptor identifier and its canonical nature are of secondary importance. In this work, invariant descriptors receive unique and canonical identifiers, based only on a hash of essential descriptor components: value, uncertainty (if any), and units. Variable descriptors, on the other hand, derive their identifiers from these parameters in addition to the canonical identifier of their corresponding chemical configuration, itself derived by hashing values of its annotations, presented in lexicographical order. Multi-component descriptors that include other descriptors, such as the non-invariant positional descriptor for atoms, receive an identifier that is a hash of the canonical identifiers of constituent descriptors in lexicographical order instead of a direct hash on the value, uncertainty, and unit annotations. In both cases, the resultant hash is appended to the identifier of the entity to which these descriptors refer in order to obtain the final identifier.

The scheme for specifying descriptors is quite simple (Figure 6). Descriptors may contain other descriptors and may have an unlimited number of annotations, such as units and uncertainty, but must have a value assigned. Furthermore,

<sup>28</sup> BigOWLIM High-Performance Semantic Repository

<sup>29</sup> FaCT++ Reasoner

<sup>30</sup> SADI SemanticWeb Services - 'cause you can't always GET what you want!

<sup>31</sup> Small Molecule Subgraph Detector (SMSD) toolkit

CHESS follows the CHEMINF approach to modeling computationally-derived descriptor provenance. Thus, a chemical descriptor is an information content entity that is a specified output of an algorithm (e.g. Mannhold logP algorithm<sup>32</sup>), and an output of a parameterized or non-parameterized software execution process. This process has to be annotated with the identity of the software agent employed, which may be further annotated with e.g. software version, and any parameters (formally defined in an ontology) used in carrying out the calculation. Experimentally-derived descriptors follow a similar scheme, but refer to experimental observations and processes.

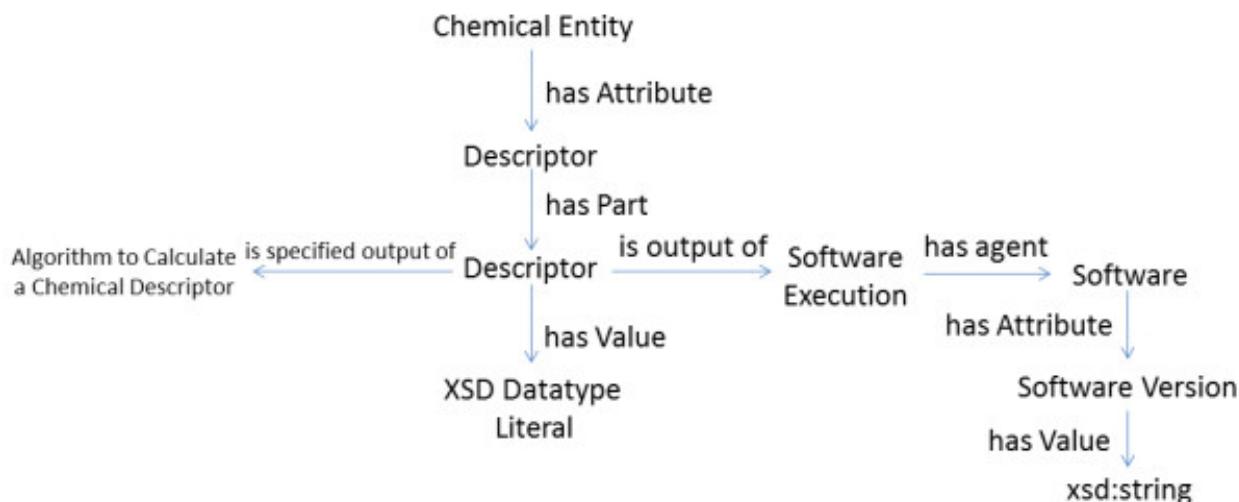


Figure 19.6: Figure 6. A simplified descriptor specification, as per the CHEMINF ontology approach  
A simplified descriptor specification, as per the CHEMINF ontology approach.

In addition to this specification, descriptors are also annotated with a chemical configuration. This concept is useful not only as a nodal point for storing source and other provenance information, but also in uniting the descriptors that have been derived under a uniform set of conditions and for the system under investigation that is in a given, well-defined state. Because chemical configurations are also, in part, specific to a geometric configuration of a given entity, it also allows a non-confounded aggregation of data, e.g. in representing atomic coordinates for multiple conformations of the same molecular entity or investigating the thermodynamic properties of molecules that may have different electronic configurations (e.g. singlet or triplet oxygen) or molecules that may be exposed to different temperatures. This approach also does away with data retrieval and integration complexities arising from heterogeneously derived information attached to a single chemical entity in certain databases, manifested in the requirement for extensive involvement of a human expert to identify a set of descriptors suitable for a particular comparison. So long as chemical information is properly annotated, descriptors specified with our approach (Appendix 5) are readily amenable to facile querying and retrieval (Appendix 6), significantly reducing the workload on an individual researcher. Alternatively, querying the knowledgebase for all descriptors from different databases that refer to the same geometric configuration or experimental conditions, is also possible.

Appendix 5. Sample specification of variable descriptors with reference to a chemical configuration, amended for readability.

```

http://www.w3.org/1999/02/22-rdf-syntax-ns#>
http://semanticscience.org/resource/SIO_>
http://semanticscience.org/resource/CHESS_>
http://purl.org/obo/owl/CHEBI#>
http://purl.org/obo/owl/UO#UO_0000019>.
http://pubchem.ncbi.nlm.nih.gov.”
  
```

<sup>32</sup> Substructure versus Whole-molecule Approaches for Calculating Log P

Appendix 6. Sample query of variable descriptors that would retrieve coordinate information for all atoms of ethanol that originate from PubChem, amended for readability.

```
http://www.w3.org/1999/02/22-rdf-syntax-ns#>
http://semanticscience.org/resource/SIO_>
http://semanticscience.org/resource/CHESS_>
http://purl.org/obo/owl/CHEBI#>
http://pubchem.ncbi.nlm.nih.gov".
```

### 19.3.5 Chemical Information Integration

Having described at length the representation of chemical information and various entities, let us consider the practical effects of decisions taken in CHESS on information integration. Although we have generated and successfully integrated moderately-sized subsections of various publically accessible chemical databases, the overall effect of our specification can be illustrated on the example of a limited set of molecules present in multiple databases or having descriptors created by different computational procedures. To demonstrate the facile cross-database and cross-study information integration afforded by consistent canonical entity identifiers, consider two instances of the same compound, antidepressant melitracene, in two different databases, PubChem and ChEMBL. Although these entries are cross-linked in their respective databases, the comparison and integration of information regarding this entity involves a procedure that requires a degree of human involvement, especially if this entry has to be further cross-linked to another entry in any number of other databases, each having unique data fields or approaches to data presentation and representation.

With CHESS, it is possible to independently encode the information in each repository in separate RDF graphs. However, because chemical entity URIs for melitracene are the same in all cases [http://semanticscience.org/resource/CHESS\\_GWWLWDURRGNSRS-UHFFFAOYSA-N](http://semanticscience.org/resource/CHESS_GWWLWDURRGNSRS-UHFFFAOYSA-N), all of these graphs effectively collapse into a single graph, with a SPARQL query to retrieve information relevant to melitracene capable of seamlessly drawing on the entirety of the chemical knowledge, without regards to originating database- or software-specific integration barriers. At the same time, this allows us to address issues relating to data correspondence, as bonds with the same URI are assured to be the same entity. This eliminates the need for substructure matching or other intermediate steps in carrying out cross-database comparisons. This ability may be especially useful in cases where, for example, multiple computational experiments are performed and the computed bond lengths need to be compared to the experimentally observed bond lengths or to the results of other computational experiments.

As a simple demonstration of this capability, we have generated a small set of 90 CHESS-encoded antidepressants containing selected information from two databases, as well as three different computational packages. The descriptors we have represented in this knowledgebase were of relevance to satisfying Lipinski's Rule of Five<sup>33</sup>, allowing us to potentially pool all of the available information in these disparate databases and computational experiments to address the question of whether the compounds in our knowledgebase were, in fact, drug-like. For each source of information, a number of descriptors were intentionally left out to demonstrate the assurance of information complementarity and preservation of information correspondence with the CHESS representation. To test this, we have created a custom class with a formal definition corresponding to Lipinski's Rule of Five and reasoned over it using the Pellet reasoner plugin (version 1.4)<sup>34</sup> in Protégé software<sup>35</sup> (version 4.0, build 115). As this class was correctly populated with 88 instances of chemical entities satisfying the Rule of Five, it became apparent that consistent molecular identifiers permitted the effective collapse of the multiple knowledge sources to provide the information necessary in order to fulfil the classification and identify drug-like chemical entities. Furthermore, the correspondence of the molecular descriptors specified in our knowledgebase was adequately preserved.

Appendix 7. The formal axiomatic definition of a class of chemical entities that satisfy Lipinski's Rule of Five, using CHEMINF concepts (specified using the Manchester OWL syntax).

<sup>33</sup> Lead- and drug-like compounds: the rule-of-five revolution

<sup>34</sup> Pellet OWL 2 Reasoner

<sup>35</sup> Protégé Ontology Editor

### 19.3.6 Variable Level Granularity Semantically Enriched Annotations and Queries

We have established that CHESS addresses the problem of seamless data integration across multiple sources of information by demonstrating a query that reproduces a common Rule of Five filter, on an integrated data set from three different sources. However, this kind of filtering can currently be readily carried out for most chemical databases through their search engine interfaces provided by the suppliers of such databases, such as PubChem. In the absence of such exposed search interfaces, however (e.g. when there is no option to restrict search results based on molecular mass), users of these chemical information repositories are faced with the task of manually parsing or calculating the chemical information that is needed for answering their research question. In addition to this, annotations of existing entities with new information in smaller studies are often lost due to the database barriers discussed at length in this work, or the practicality of publishing ‘smaller’ scientific data as an accessible database that is open to querying. Finally, the level of annotation granularity allowed in a given database may be insufficient for a particular application. That is, while data on individual atoms and bonds certainly exists in PubChem, it is impossible to refer to, retrieve, or annotate these individual entities, at least not in a manner that would immediately meaningfully connect the annotations generated as a part of a given study with a given PubChem entity and would allow these annotations to be discovered and queried.

In contrast, CHESS is flexible and extensible: annotations and information represented in CHESS is assured to be searchable, no matter what information the database vendor feels like exposing. So long as there is chemical information represented in CHESS, it is subject to logical queries and semantic integration. Furthermore, all chemical entities, such as atoms, bonds, and molecules, can be fully annotated using a range of vocabularies, and these annotations can be linked directly to the entities being annotated, even if the original entity and its annotations reside in different RDF graphs. With CHESS, it is now possible to facilitate open publishing of scientific information and assure that the precious scientific knowledge is preserved, no matter how small a study has been carried out.

To demonstrate the benefits of this approach, let us consider a practical case that is impossible to address with currently existing chemical databases. For this, let us examine phenolic antioxidants, which constitute an important class of molecules that are used in industrial processes and as nutritional supplements to help alleviate the damaging effects of free radicals, such as lipid peroxidation in oils. It has been shown that Bond Dissociation Enthalpies (BDEs) of the phenolic O-H bonds can be used as excellent predictors of the potency of a phenolic compound. When designing phenolic antioxidants for biological systems (e.g. humans), one has to be mindful that the BDE of the phenolic O-H bond has to be higher than that of the weakest O-H bond in ascorbate (67 kcal/mol) to allow for biological antioxidant recycling, but lower than that of the O-H bond of  $\alpha$ -tocopherol (78 kcal/mol) to surpass the potency of existing physiological antioxidant defences<sup>36</sup>. While the accurate annotation and searching of bond-level information is currently impossible in major chemical information repositories, we have developed a demonstrative set of chemical entities with O-H BDE information annotation (available from our companion website<sup>23</sup>), including the computational method, software, and some of the parameters used to compute this information for ethanol (Appendix 8).

Appendix 8. A representative portion of the CHESS specification of a ethanol, its constituent OH bond, and the parameters used in BDE calculation for this bond.

```
http://www.w3.org/1999/02/22-rdf-syntax-ns#>
http://semanticscience.org/resource/SIO_>
http://semanticscience.org/resource/CHESS_>
http://purl.org/obo/owl/CHEBI#>
http://semanticscience.org/resource/CHEMINF_>
http://purl.org/obo/owl/UO#UO_0000012“>
```

Please note that this representation was simplified and modified to improve readability.

A knowledgebase with such information could have been published and shared as an outcome of any of a number of studies on this subject, with annotations attached directly to existing chemical entities in one of the major chemical databases. Instead, this information was sealed away in a series of PDF and HTML documents, accessible only to

<sup>36</sup> Development of novel antioxidants: design, synthesis, and reactivity

those with the time and resources to locate and read them. In order to retrieve *phenolic* antioxidants with potential applications in biological systems, one could combine the molecular structural features (the presence of a phenol group) with thermochemical annotation information on the OH bond. Furthermore, since computationally derived thermochemical parameters are best compared when the method, software, and parameters used to derive them are uniform, we may include these requirements in our query to retrieve uniform and useable information. Such an integrative query will provide all the potentially potent novel phenolic antioxidants from a chemical knowledgebase (Appendix 9). Please note that such integrative, variable-granularity, and flexible queries are impossible for the major conventional chemical information repositories. Truly, the imagination of CHESS users is the limit for the expressivity and the semantic enrichment of the represented chemical information.

Appendix 9. Sample query that may be carried out across a single or multiple SPARQL endpoints to retrieve all potentially promising phenolic antioxidants from an annotated compound collection, amended for readability.

```
http://www.w3.org/1999/02/22-rdf-syntax-ns#>
http://semanticscience.org/resource/SIO_>
http://semanticscience.org/resource/CHESS_>
http://purl.org/obo/owl/CHEBI#>
```

### 19.3.7 Representing and Querying Reactive Transformations

As a final demonstration of usability of CHESS-encoded chemical information, let us consider reaction representation and querying. As mentioned earlier, reactions in CHESS are considered to be chemical entities, whose identifiers come from their participants and stoichiometry. Though reactions may or may not take place under a certain range of circumstances, we model them as abstract processes that exist without regards to their likelihood of occurring. Because as long as a specified reaction respects universal principles, such as the conservation of matter and energy any reaction may happen at any time (some are more likely than others), we maintain that the sole features relevant to unique reaction specification are the chemical identities and stoichiometries of the participating chemical entities (Figure 3).

In general terms, there are two types of reactive transformations specified: those involving generic chemical moieties and their transformations and those that involve precisely defined chemical entities. Both cases may be represented in CHESS in a uniform and consistent fashion. The only difference between the two cases is the specification of the entities involved, which in the former case are of the functional group/substructure type, and in the latter are instantiated molecular entities. In either case, the most important feature that decides whether a molecule will undergo reactive conversion is the presence of a characteristic functional group within the substrate molecule. That is, reactions can be viewed as transformations between the various isolated regions, or functional groups, with the rest of the molecule attenuating or increasing reactivity. The presence of the requisite functional group through explicit mapping of constituent atoms and bonds makes it possible to infer the role of a given molecule and its components in a reactive process (Figure 7).

The amount of information regarding a reactive transformation that could be inferred is entirely dependent upon the amount of information present in the reference knowledgebase and in the reaction specification RDF graph. For a demonstrative example, let us consider a transformation of primary alcohols to aldehydes, which could be catalyzed by an inorganic catalyst or by an enzyme. Since generic functional group transformations are involved, reaction specification would involve functional group types and atom types, rather than corresponding instances. Substrates are linked to the reaction using *has input* (SIO\_000230), products using *has product* (SIO\_000312) and catalysts using *has agent* (SIO\_000139). Further, to maintain a record of the correspondence of every transforming entity, the transformations would be specified with *transforms into* (SIO\_000655), which can operate upon whole molecules, functional groups, and atoms (Appendix 10).

Appendix 10. Sample reaction definition of CCO functional group transformed into CC = O functional group with CHESS in N3-turtle, amended for readability.

```
http://semanticscience.org/resource/CHESS_>
```

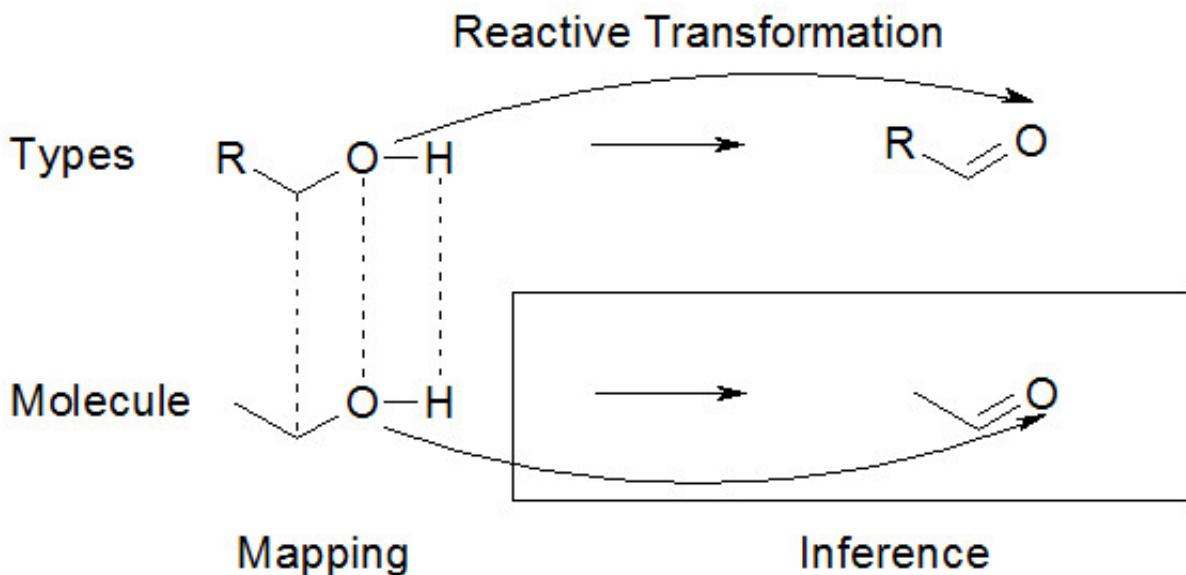


Figure 19.7: Figure 7. Mapping components of an instantiated molecule to reactive transformation definition participants can enable a range of inferences relevant to predicting reaction outcome

**Mapping components of an instantiated molecule to reactive transformation definition participants can enable a range of inferences relevant to predicting reaction outcome.**

[http://semanticscience.org/resource/SIO\\_>](http://semanticscience.org/resource/SIO_>)

<http://purl.org/obo/owl/CHEBI#>>

$\text{O}_3$

To enable reaction matching, it is necessary to also obtain the atom and functional group typing information for the chemical entities in the CHESS knowledgebase. For this purpose, one may use an extension of the described atom-centric fingerprinting procedure by typing atom and functional group instances to the generic classes of atoms and functional groups they instantiate. Thus, instead of merely asserting membership of group LFQSCWFLJHTTHZ-UHFFFAOYSA-NFGXYZ in the ethanol molecule (where to save space, XYZ is the appropriate hash), it would be typed to the FGXXX functional group. A similar procedure has to be carried out for atoms if atom-centric searches need to be enabled, e.g. for tracing the precise passing of atoms between molecules, and checking the metabolic history of every single atom (Figure 8). As mentioned earlier, this process may be carried on at any point, and the knowledgebase may be amended with the required information.

The resultant specification of reactions and chemical entities enables simple, yet powerful SPARQL-based queries, e.g. to find the potential reactions of a given molecule and identify the atoms and functional group instances involved in these reactions (Appendix 11).

Appendix 11. A query that will return all the reactions that ethanol may potentially be involved in, amended for readability.

[http://semanticscience.org/resource/CHESS\\_>](http://semanticscience.org/resource/CHESS_>)

<http://semanticscience.org/ontology/sio.owl#>>

<http://purl.org/obo/owl/CHEBI#>>

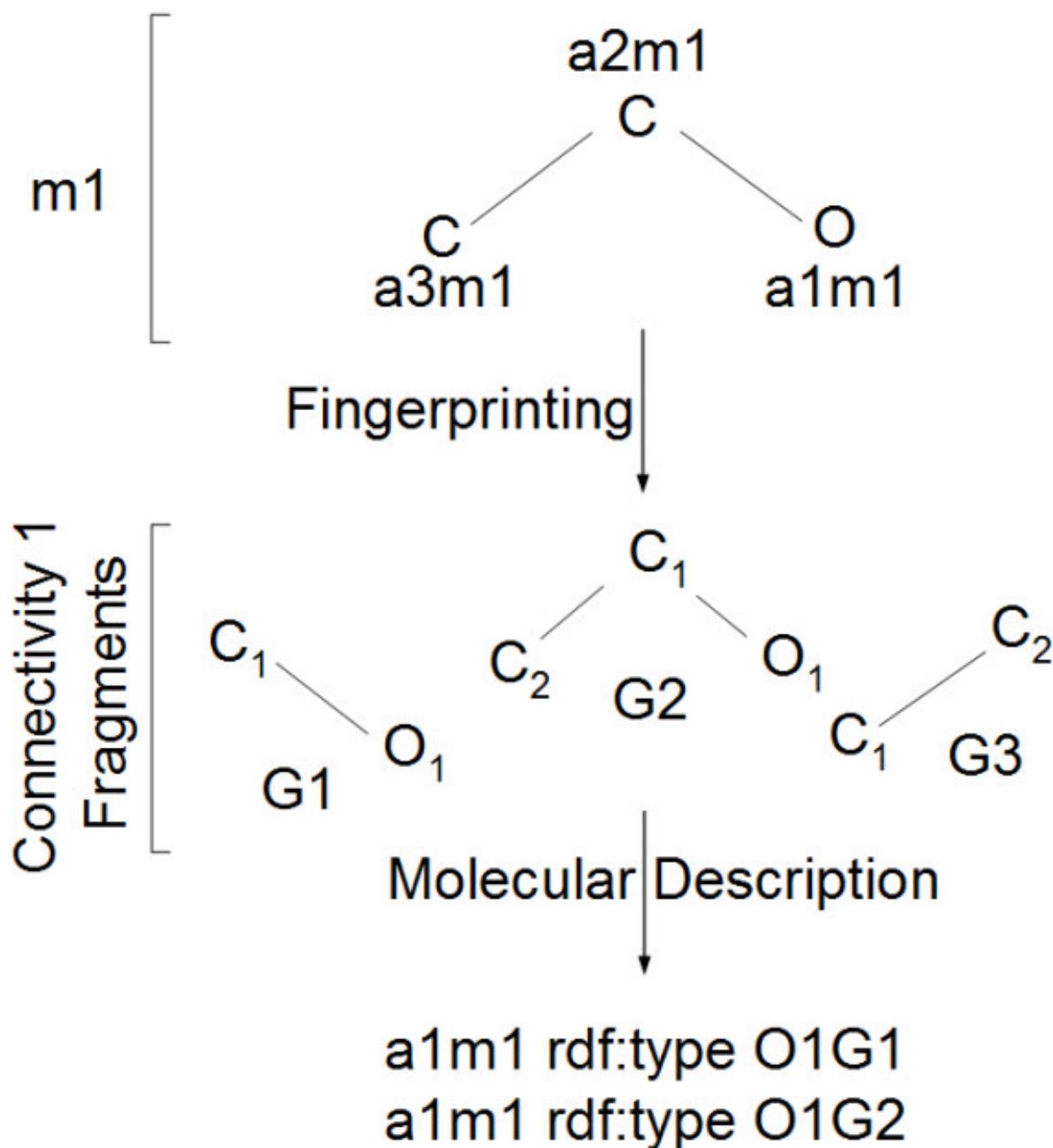


Figure 19.8: Figure 8. A modification of molecular fingerprinting used for the complete description of chemical structures in terms of an exhaustive list of functional groups

**A modification of molecular fingerprinting used for the complete description of chemical structures in terms of an exhaustive list of functional groups.** Molecular fragments resulting from fingerprinting ( $G1$ ,  $G2$ ,  $G3$ ) may be stored and treated as descriptive functional groups, along with user-submitted ones.

### 19.3.8 Supporting Tools and Interfaces

An evolving project site with supporting information, graphical user interface-backed tools, sample RDF specifications and data sets (including a sample set of ~17000 ChEBI compounds), as well as further information, is available<sup>23</sup>. The implementation of CHESS provided on the website has a scope limited to compounds that can be well-represented with current version of InChI. Therefore, we currently exclude polymers with repeating units of arbitrary length and Markush structures from the scope of CHESS, for example. On the other hand, organometallic compounds and metals that can be represented with InChI are included within our scope. While there is no theoretical limit to the size of a compound that can in principle be represented using CHESS, our sample implementation is practically limited (by the available computational resources on the server) to lower molecular weight compounds.

## 19.4 Conclusions

Unfortunately, many of the large chemical databases currently do not possess the means of chemical data integration and federation. Either for historical reasons or for efficiency improvement, a large number of these databases have been purpose-built for capturing data within a particular domain, and without much consideration of trans-domain knowledge aggregation. This further complicates the task of database integration and poses as an obstacle to productive chemical research. Fortunately however, the Semantic Web provides an excellent opportunity for significantly simplifying this problem with the appropriate data representation and sufficiently advanced data conversion tools.

In this work, our principal goal has been to present a novel chemical representation formalism that draws on the Semantic Web principles. We have attempted to make a compelling case for a universal semantic specification of chemical entities in cheminformatics by demonstrating the power, integrative capacity, and the flexibility of representation afforded by fully embracing Semantic Web technologies. By adopting consistent and canonical identifiers for every aspect of chemical entities identified here, we have demonstrated facile cross-domain chemical knowledge integration while preserving correct data correspondence and explicit data provenance information. Furthermore, we have demonstrated the power of CHESS in enabling integrative chemical research that draws on the entirety of chemical information available on the Web. While we do not believe that any specification can natively address outright errors in databases, CHESS representations allow us to explicitly and formally define the meaning of our data and enable machines agents to automatically reason over this data, checking it for consistency and completeness. This, in turn, enables a more accurate scientific discourse and a more reproducible and transparent way of doing science.

We have also demonstrated mechanisms by which chemical configuration-specific information may be encoded without loss of inter-configuration information aggregation and without introducing intra-configuration information mixing that is sometimes an unfortunate occurrence in traditional databases. For example, atomic coordinate information and heat of formation data may exist for multiple conformers of a single molecule, with every conformer annotated with the appropriate data. In principle, this chemical configuration concept allows one to aggregate information about the various electronic states of a given molecule when CHESS will be extended to include the explicit specification of electrons and related concepts. While CHESS does not aim to make statements with respect to the preferred chemical configuration of a given compound under a given set of conditions, CHESS allows the unambiguous and explicit identification of precise chemical configurations for the purpose of advancing and facilitating interdisciplinary scientific discourse.

We believe that outsourcing of chemical information integration to machine agents is of increasing importance as the rapidly growing collection of diverse chemical information already available on the web is overwhelming human integrative capacity. If no steps are taken to create, standardize, and adopt a set of consistent standard Semantic Web chemical information exchange ontologies and representation formalisms soon, we are risking missing yet another opportunity to truly federate the chemical web and trigger a transition to a new era of chemical research. Therefore, with this work, we would like to invite the broader cheminformatics community to initiate the discussion of representations, standards, and supporting ontologies. Truly, the infinite chemical space is full of mysteries, marvels, and opportunities - and we believe that it is only through the concerted and unified efforts of researchers in all fields of science, enabled by Semantic Web technologies, that we may hope to one day chart it.

## 19.5 Methods

### 19.5.1 Supporting Ontologies

CHEMINF is a collaboratively developed Web Ontology Language (OWL)<sup>37</sup> ontology for representing chemical information and descriptors, freely available to the broad cheminformatics community. SemanticScience Integrated Ontology (SIO) is a general ontology that provides over a 150 object relations and over 900 classes of various entities, including physical, processual, and informational ones. There is only one data property in the ontology, ‘has value’, with the other relations describing aspects of mereology, spatial positioning, temporal ordering, qualities, attributes, representations, participation, and agency. In addition to these ontologies, we use the CHEBI ontology for chemical entities and concepts.

### 19.5.2 Triplification of Chemical Information

The encoding of chemical information into its CHESS form was carried out with software we developed (sample source code available on the companion website<sup>23</sup>), based on the Jena API<sup>38</sup> and the Chemistry Development Kit<sup>39</sup>. A sample dataset of 90 antidepressants was obtained from a mesh-based keyword search on ‘antidepressant’ over the PubChem database. Chemical descriptors were computed using CDK and the Open Babel Java API<sup>40</sup>. Unique identifiers were obtained by hashing pertinent chemical data as a 40-letter SHA-1 hash using the Java Security API.

### 19.5.3 Query Testing

In order to test our chemical similarity queries, we used OpenLink Virtuoso version 6.01.3126 SPARQL endpoint<sup>41</sup>. To test DL Rule-based queries, we used the Pellet reasoner version 2.2<sup>42</sup>, running on a single CPU core. The machine used for these tests had dual Intel Xeon CPU at 2.9GHz, with 32GB of RAM.

## 19.6 Competing interests

The authors declare that they have no competing interests.

## 19.7 Authors' contributions

LLC wrote the paper, implemented supporting tools, generated triple stores, and ran tests. LLC and MD created the supporting ontologies. MD contributed to the paper, provided guidance, and ran tests. Both authors have read and approved the final manuscript.

## 19.8 Acknowledgements

This research was funded in part by National Science and Engineering Research Council of Canada CGS award for LLC and Discovery Grant funding to MD.

<sup>37</sup> Web Ontology Language Specification

<sup>38</sup> Jena Semantic Web Framework

<sup>39</sup> The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics

<sup>40</sup> Open Babel Chemistry Toolbox

<sup>41</sup> OpenLink Virtuoso

<sup>42</sup> Pellet: A practical OWL-DL reasoner



# CONSISTENT TWO-DIMENSIONAL VISUALIZATION OF PROTEIN-LIGAND COMPLEX SERIES

## 20.1 Abstract

### 20.1.1 Background

The comparative two-dimensional graphical representation of protein-ligand complex series featuring different ligands bound to the same active site offers a quick insight in their binding mode differences. In comparison to arbitrary orientations of the residue molecules in the individual complex depictions a consistent placement improves the legibility and comparability within the series. The automatic generation of such consistent layouts offers the possibility to apply it to large data sets originating from computer-aided drug design methods.

### 20.1.2 Results

We developed a new approach, which automatically generates a consistent layout of interacting residues for a given series of complexes. Based on the structural three-dimensional input information, a global two-dimensional layout for all residues of the complex ensemble is computed. The algorithm incorporates the three-dimensional adjacencies of the active site residues in order to find an universally valid circular arrangement of the residues around the ligand. Subsequent to a two-dimensional ligand superimposition step, a global placement for each residue is derived from the set of already placed ligands. The method generates high-quality layouts, showing mostly overlap-free solutions with molecules which are displayed as structure diagrams providing interaction information in atomic detail. Application examples document an improved legibility compared to series of diagrams whose layouts are calculated independently from each other.

### 20.1.3 Conclusions

The presented method extends the field of complex series visualizations. A series of molecules binding to the same protein active site is drawn in a graphically consistent way. Compared to existing approaches these drawings substantially simplify the visual analysis of large compound series.

## 20.2 Background

Many methods in structure-based drug design, like virtual screening, scaffold hopping, and docking, are dealing with series of protein-ligand complexes. They are all characterized by several poses or ligands bound to one active site, which is, except for potential conformational flexibility, non-varying for the whole set. The comparative visual inspection of the different binding patterns is facilitated by a depiction mode, which takes the constant part into account. While in the context of three-dimensional (3D) visualization the superimposition of ligands in one graphical active site representation is common practice<sup>1</sup>, the orientation of two-dimensional representations is often affected by the attempt to provide a planar and aesthetically ideal arrangement of all diagram elements. This leads to a heterogeneous overall picture within a complex series and makes the comparison of the binding modes difficult.

An approach for the 2D depiction of protein-ligand complex series with an automatically generated consistent layout of the residues for all diagrams was introduced in the software MOE<sup>2</sup> in 2007. The built-in 2D drawer is able to deal with single proteins, which contain multiple ligands, as well as with multiple members of one protein family. Generally speaking, the layout generation is done in two steps: First, the planar ligand diagrams are aligned in their original 3D position and this alignment is transformed to the x-y-plane. In a second step, the residues are placed based on pseudo-atom positions on a grid, which are derived from the superimposed ligands. Although the method works well in practice, the protein amino acids are represented as spherical objects only such that the individual hydrogen bonding pattern cannot be derived from the figure.

Also Ligplot<sup>3</sup> can be used to depict series of complexes with a consistent 2D layout for all drawn residues. In this case, the layout generation is a semi-automatic process: while the initial diagram layouts are generated automatically, the user has to choose one of them as template and subsequently to align the residue centroids manually by means of certain meta-files.

In this work, we will present an extension of PoseView<sup>456</sup> that generates series of complex diagrams with a consistent receptor layout. In its previous versions, PoseView generated automatically 2D layouts for single protein-ligand complexes by means of a ligand centered algorithm. The objective of the new approach is to find a global position for each residue providing an intersection-free arrangement of directed interactions for each individual complex diagram. In contrast to the algorithm for single complexes the new methods take the 3D arrangement of the residues into account assuming that some of the 3D adjacencies can be conserved. Therefore, these adjacencies are used as basis for the initial 2D residue arrangement around the ligand.

## 20.3 Methods

In this section, we will give an overview of the whole layout generation procedure and explain the algorithms, which differ from the layout generation algorithm for single complexes. A more detailed description of the underlying PoseView methodology can be found in<sup>456</sup>. Below, these methods are summarized, where necessary for the understanding of our new approach. Additionally, a graphical representation of the work flow is given in Figure 1.

The algorithm starts with the determination of the interactions between the different ligands and the receptor and the initialization of the individual molecules. For each individual complex, a drawing complexity score and an initial layout is calculated based on the structure diagrams of the ligand and residue molecules. Furthermore, a tree is derived from the complex' connectivity. Subsequently, the residues, which are part of any of the complexes in the series, are collected and stored in the global template. Then, a global layout is computed for the diagrams of the selected residues starting with the determination of the optimal global target sequence (circular order of amino acids around the ligand) derived from the original 3D residue adjacencies. The global target sequence is optimized by means of the individual trees which are derived from the complexes. Subsequently, the individual ligand layouts are modified and

<sup>1</sup> Visualization of macromolecular structures

<sup>2</sup> 2D depiction of protein-ligand complexes

<sup>3</sup> LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions

<sup>4</sup> Drawing the PDB: Protein-Ligand Complexes in Two Dimensions

<sup>5</sup> From modeling to medicinal chemistry: automatic generation of two-dimensional complex diagrams

<sup>6</sup> Molecular complexes at a glance: automated generation of two-dimensional complex diagrams

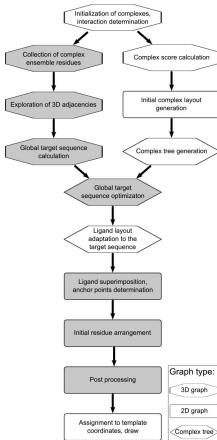


Figure 20.1: Figure 1. PoseView work flow

**PoseView work flow.** The PoseView algorithm proceeds on different graphs which is represented by the different shapes of the text boxes. Calculation steps based on the 3D complex representation are denoted by an octagon, the ones based on the topological tree representation are denoted by a hexagon and structure diagram based algorithms are denoted by rectangles. Additionally the steps are subdivided in those which address the global structure (gray background) and those which modify the individual complexes (white background).

their interaction atom arrangements are adapted to the global target sequence. The global layout is computed based on the convex hull<sup>7</sup> of all superimposed ligand diagrams and the resulting global interaction starting coordinates, called anchor points. This layout generation includes an initial placement of each residue diagram and a subsequent post optimization analog to the residue layout calculation for single complexes. In a last step, the global amino acid atom coordinates are assigned to the individual complexes and then they are drawn.

The algorithm employs three different graphs as underlying data structures:

- the local 3D graphs which are composed of the ligand and the interacting residues for each individual molecular ensemble
- the corresponding 2D graphs containing the coordinate sets of the molecular structure diagrams
- a tree for each complex, containing only topological information based on the connectivity of the individual molecular ensembles

In Figure 1 the methods are labeled according to the type of graph underlying the calculation step.

### 20.3.1 Terms and definitions from the layout generation for single complexes

Before starting the algorithm description, some of the basic terms, which are defined in previous papers and used in the following section, will be mentioned here:

- The order of ligand interaction atoms that is generated by a circular walk around the ligand is referred to as *interaction atom order*.
- A *good layout* is characterized as an arrangement of all depicted complex elements that, on the one hand, is collision free and on the other hand fulfills aesthetic and chemical structure diagram conventions. The quality of the complex diagram layout results from the combination of an intersection-free interaction atom order, the convenient geometric positioning of the single structure diagrams (SD), and the consequential arrangement of interaction lines.
- Each residue in a complex has a *main interaction direction* with a defined starting point. It is the resultant from the individual optimal directions of all directed interactions which are connecting one residue with the ligand. For each

<sup>7</sup> Introduction to algorithms

molecule of the complex ensemble, the individual directions are derived from the convex hull of the 2D atom coordinates which leads to a radial orientation and avoids collisions of the structure diagram and its interaction lines. The main direction is calculated for both the ligand and the residue. The centroid of the corresponding interacting atoms is defined as the *main interaction direction starting point*. The placement of a residue is realized by superimposition of the matching ligand and residue main interaction vectors.

### 20.3.2 Initialization of complexes and interaction determination

The input and initialization of the individual complexes is performed using the chemistry model and file handling utilities implemented in FlexX<sup>8</sup>. The interactions can originate from either the built-in geometry-based interaction model in PoseView<sup>4</sup> or calculated by any other software. In the case of external interaction calculation, the interactions have to be defined in the comment block of an input file in mol2 format. If the interactions are determined by the PoseView model and the protein is defined in a PDB file, the separation of residues from the 3D structure is done as described previously<sup>6</sup>.

### 20.3.3 Complex score calculation

Before starting the layout calculation, the complexes are scored and ordered according to their drawing complexity, which is determined mainly by the number of interacting residues and the number of residues with more than one directed interaction to the ligand. The latter ones are responsible for the need to potentially modify the initial ligand structure diagram layout in order to provide an intersection-free arrangement of interaction lines. Therefore, the complexes are ordered according to the following score  $*\sigma*$ :

where  $\#_i$  interactions is the number of directed interactions from residue  $i$  to the ligand.

In case of greedy layout decisions in the context of the sequential calculation steps like the determination of ligand anchor point coordinates, the scoring ensures that the more complicated complexes are treated first. The subsequent ligand superimposition method as well as the calculation of the initial global residue sequence take advantage of this ordering.

### 20.3.4 Initial complex layout generation

For each complex of the series, the structure diagrams of all interacting ligands and residues are initially generated as basis for the following steps. At this point, no optimization procedures are performed even though collisions may occur and the interaction atom order may be suboptimal. The resulting 2D coordinates are used as starting point for the following algorithms.

### 20.3.5 Representation of single complexes as a tree

A rooted tree, in the following referred to as complex tree, is derived from the initial complex layout whose nodes represent atoms or groups of atoms like non-interacting ring systems and whose edges represent a covalent bond or interaction. This leads to a uniform representation of all complex parts - ligand atoms and bonds, interactions, residue atoms and bonds - and permits on the one hand a condensation of parts of the complex, which are irrelevant for layout decisions, and on the other hand an ordered layout processing. Both features improve the average run time in comparison to the full enumeration of possible bond modifications on the basis of the structure diagram representation. The tree is directly derived from each individual complex, reflecting the relative 2D arrangement of structure diagram elements by its edge sorting. Subsequently, it is processed in order to simplify the following layout generation process under conservation of all chemical and topological information that is needed to generate valid 2D layouts. Initially, for each atom a node is inserted and these nodes are connected by edges according to the connectivity in the original protein-ligand complex by the covalent bonds and interactions. Unlike acyclic parts of the structure diagram, rings

---

<sup>8</sup> A fast flexible docking method using an incremental construction algorithm

are represented by a single central node such that circles are avoided. Additionally, for each ring atom that is starting point of a substituent or an interaction, an additional node is inserted and an edge, that connects this new node with the center node.

The residue part of the complex is represented only partly in the tree: In contrast to the ligand, whose atoms are all considered in the tree generation, the residue atoms are only included if they interact with any ligand atom. Hence, all residue atom nodes are leaves of the tree. For directed interactions between ligand and residue three different layout scenarios are possible: In the first case, there is only one interaction between both molecules such that one node has to be inserted in the tree to represent the residue part. In the second case, one atom of the residue forms more than one interaction to the ligand. This leads to a representation of this one atom by multiple nodes in order to avoid circles in the graph. Such nodes get an adjacency label, that is realized by the global interaction order sequence as described in the following subsections. During the subsequent interaction order optimization, the method tries to find an order, which satisfies this adjacency demand. The third case offers two distinct atoms of the same residue interacting with the ligand. This is also solved by inserting multiple nodes and setting an adjacency label. Hybrid forms are treated the same way.

The edges are labeled with modification operations depending on the related bond or interaction (for the assignment procedure of bond modifications see <sup>6</sup>). Due to the nature of 2D bond rotations, which are the equivalent of 180° flips, their effect on the interaction atom order is canceled by a second rotation of any other edge that influences the order of the same interaction atom range. Therefore, subsequent edges of the tree, whose rotations cause the same layout modification, can be condensed to single edges. The number of potential modifications for ligands with long carbon chains for example can be significantly decreased: if five adjacent edges are merged, the number is decreased from 2<sup>5</sup> to 2 potential modifications. Branches, which represent non-interacting molecule parts and cannot influence the interaction atom order, are cut. Based on this condensed tree, a root is calculated by determination of the central node: Starting at each leaf node, a depth first search is performed and the nodes are labeled according to their depth in order to find the longest path in the tree. The node that lies in the middle of the longest path is selected as root node. Now, edges can be directed in root-to-leaf order and sorted according to the polar angle between the corresponding bonds and the bond representing the unique edge to the parent node. This results in a tree, whose leaf order is equal to the order of residues around the ligand in the complex. The number of leaves is equal to the number of directed interactions. A graphical example for the tree generation and layout modification is given in Figure 2.

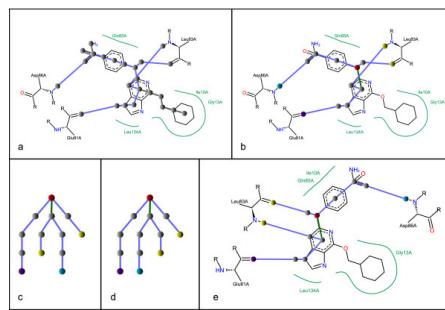


Figure 20.2: Figure 2. Complex tree generation

**Complex tree generation.** a) In the initial complex tree, a node is inserted for each acyclic atom, for each ring center, for each ring atom that is either an interacting atom or a substituent starting atom, and for each amino acid interaction atom. Subtrees which lie not on a path between interaction atoms are denoted by gray edges. b) After edge condensation and removing subtrees which are not needed the resulting tree contains only one rotatable edge which is highlighted in green. The root node is colored red and the leaves are colored accordingly to the amino acid they belong to. c) The derived complex tree has a suboptimal leaf order, which is optimized by rotation d) such that the yellow nodes are adjacent. e) The tree modifications are applied to the ligand structure diagram.

### 20.3.6 Collection of complex ensemble residues

Beyond the optimal individual layout, a good global layout that compromises with all individual optimal layouts has to be computed. The generation of such a layout starts with the collection of all different interacting residues based on the individual complexes. For all single complexes the interacting residues are enumerated and, if not already found in a previous complex, a 3D representation is added to the global template structure. The mapping of equal residues is realized by comparing their 3-letter code, their sequence number and chain ID. Subsequent to the collection, the structure diagrams of all residues are generated and also stored in the template.

### 20.3.7 Exploration of 3D adjacencies and global target sequence calculation

For many complexes more than one ligand structure diagram layout provides an intersection-free arrangement of diagram elements. As parts of the ligands in their bound conformation are planar or rigid because they consist of ring systems or non-rotatable bonds, 3D residue adjacencies are a good heuristic starting point to calculate the global initial interaction order, also called global target sequence. In this step, distances and adjacencies of the 3D residues are computed along the active site molecular surface<sup>9</sup> in order to find an adequate circular arrangement of the residues around the ligand. Walking along the surface is necessary because at narrow points in the binding pocket the direct distance between two residues on either side of the lumen is relatively small in comparison to the path length along the surface. In this case, using the surface path as distance function is more suitable, because it takes account to the fact that the ligand lies between both residues and that they are therefore not adjacent. A surface triangulation<sup>10</sup> is used as basis for the path calculation and a breadth first search<sup>11</sup> is performed starting at each residue that is member of the complex ensemble. A Hamiltonian cycle of all complex ensemble residues is calculated for the resulting complete adjacency graph by using an approximation to the minimal spanning tree<sup>12</sup>. The order of residues in the Hamiltonian cycle is used as initial global target sequence.

### 20.3.8 Global target sequence optimization

A global target sequence represents the order of all interacting residues available in the template structure whereas a local target sequence is derived from the global target sequence by deleting all residues which doesn't take part in the formation of the currently processed complex. An optimal global target sequence is characterized by an intersection-free matching of all individual residue sequences of the different complexes. The initial global target sequence is therefore subsequently optimized under consideration of the first  $n$  ligands of the complex series by checking if their interaction atom order can be modified via edge modifications such that an intersection-free matching to the global target sequence is possible. The default value of  $n$  is set to 20. In case all complexes can be drawn with an intersection-free matching, the algorithm stops. Otherwise, the sequence of residues is changed randomly and tested again. The acceptance of a new order is controlled by a Simulated Annealing method in order to avoid getting trapped in local optima in terms of increasing numbers of intersections.

### 20.3.9 Ligand layout adaptation to the target sequence

As previously described, the residue sequences in the single complexes are represented by the leaf order of the complex trees and a tree can feature more than one leaf per residue. The leaf sequences are iteratively matched to the local target sequence, which are derived from the optimized global target sequence, in the order that was calculated during the complex scoring. In contrast to the ligand-centered method, which was applied in case of single complexes<sup>6</sup>, the ligand has now to be fitted in a given residue arrangement. Therefore, the leaf order of the tree has to be modified by rotating or exchanging edges until it matches the residue order in the local target sequence. The matching is realized by inserting additional edges, one from each leaf node to the position of the corresponding residue in the local target

---

<sup>9</sup> Analytical molecular surface calculation

<sup>10</sup> The ball-pivoting algorithm for surface reconstruction

<sup>11</sup> Shortest path through a maze

<sup>12</sup> An analysis of several heuristics for the traveling salesman problem

sequence, see Figure 3. In some cases, more than one edge leads to the same position in the target sequence owing to the occurrence of multiple leaves representing the same residue. As the fitting part, intersections in the matching are tried to be solved by modifying the tree applying the available modification operations. To solve a matching intersection, the common parent edge of the two appropriate leaf nodes is selected and, if possible, modified. In case of a rotation, all sub-trees containing rotatable edges have to be rotated back in order to keep the sequence valid and to affect only the relative order of the two matching edges in question. In contrast to this, exchanges of edges, which are descending from the same node, do not invert the leaf order and the sub-trees stay untouched, see Figure 4.

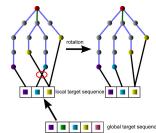


Figure 20.3: Figure 3. Target sequence alignment

**Target sequence alignment.** The global target sequence contains one entry for each residue being part of the series complex ensemble. A local target sequence is derived by deleting all entries which have no corresponding residue in the individual complex in question. After inserting the matching edges (black), they are searched for intersections. By modification of the complex tree edges (blue and green), in this case a rotation of the green edge, the intersections are removed.

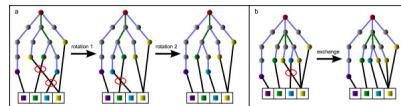


Figure 20.4: Figure 4. Tree modifications

**Tree modifications.** A rotation of an edge (a) inverts the leaf order of the whole subtree. A rotation of the rotatable edges in the relevant subtree compensates this inversion. In case of exchanges (b) the leaf order in the subtrees is not changed.

### 20.3.10 Ligand superimposition and anchor point determination

Beginning with the superimposition of ligands all following layout generation steps are based on the precalculated 2D structure diagram information. The placement starts with the ligands by determination of the anchor points for the different residues. The term anchor point is defined as the global coordinate for the starting point of the main interaction direction of a residue on the ligand side. Thus, the number of anchor points is equal to the number of residues in the global structure. Corresponding to the anchor point, each residue features a global residue coordinate, which defines the global starting point of the interaction main direction of all interactions starting at this particular amino acid in any of the complexes. The computation of the global residue coordinate will be described in the following paragraph. The ligand anchor points are calculated by iteratively superimposing the ligands of the single complexes according to the order that is defined by the complex scoring. The first ligand is translated such that its centroid lies in the origin. Then, for each residue that is interacting with this particular ligand, the main interaction direction starting point is calculated and stored as the anchor point. All other ligands are superimposed to the firstly placed ligand by minimizing the RMSD between the common subset of own and already placed template anchor points. If in the course of the superimposition new anchor points are placed, they are assigned to the appropriate residue in the global structure.

### 20.3.11 Initial global residue arrangement

Similar to the method for single complexes, the global positioning of the residue structure diagrams is based on a convex hull, but the underlying point set is, unlike in the case for single complexes, derived from the superimposed anchor points. The convex hull is represented as a circular path consisting of directed edges. Hence, each node has one incoming and one outgoing edge. To each anchor point, an edge of the convex hull is assigned: If the anchor point

is a convex hull vertex, the edge leading to this vertex is chosen; otherwise the edge with the smallest distance to the anchor point in question is selected. From all interaction main directions of the individual complexes calculated in the initial complex layout generation, the overall main direction is chosen to be the median when sorting their directions by the polar angle to the corresponding edge of the convex hull. The global residue coordinate is set to a point on this straight line with a distance of five standard bond length from the anchor point. The adjustment of the residue structure diagram is done by superimposing the global main direction of the ligand and the inverted resultant direction of all individual residue interaction directions of the individual complexes.

### 20.3.12 Global layout post optimization

Analog to the generation of single complexes, the initial placement may cause collisions. These are handled with an approach that is in principle the same as described before<sup>6</sup>. The major difference is that the collision detection is not performed on basis of atom and bond coordinates but by testing for overlaps between the convex hulls of the global residue structure diagrams and the convex hull of ligand anchor points respectively, because the atom-wise comparison would slow down the collision handling significantly. Additionally, intersections of interaction lines as well as intersections crossing convex hulls are detected.

### 20.3.13 Drawing the complexes

Subsequent to the global layout generation, the coordinates are assigned to the single complexes of the series. The ligand is drawn superimposed to the corresponding anchor points and the amino acids are drawn at their global positions. The interaction atoms of the ligand are not necessarily identical with the anchor coordinates. Thus, the interaction lines have to be adapted to the local complex coordinates. In a final step, the hydrophobic contacts are placed and drawn; they are not part of the global layout.

## 20.4 Results

The new method was applied to different test sets. In the following, three examples will be presented: two of them feature different ligands bound to the same protein (PARP and UK) while the other data (ER\*α\*) is composed of different crystal structures from the PDB<sup>13</sup> with an individual protein file for each of the complexes. Based on the presented application examples, the strength and weaknesses of the new approach will be discussed.

Before starting the layout calculation, the complexes with only one directed interaction or without directed interactions are removed from the sets as well as duplicates. Complexes are recognized as duplicate if their ligands and the interaction patterns are identical. A prerequisite for a successful layout generation process is that all ligands are bound to the same active site and protein chain; otherwise no common residues can be found by the algorithm and the layout alignment fails. In all examples the complexes are sorted according to their score, such that the ones with the highest number of interactions come first.

### 20.4.1 Poly ADP Ribose Polymerase (PARP)

The ligands for the PARP example (Figure 5) were taken from the ZINC database<sup>14</sup> and docked to a protein provided by the PDB (PDB code:

While in Figure 5b the order of interaction atoms is properly aligned to the target sequence, a collision is caused by the ligand layout. This could be avoided by an additional ligand layout post optimization step that searches for alternative ligand layouts, which improves the geometric arrangement of ligand interaction atoms without affecting their topological order. In this case, flipping the upper ring system and rotating the amide group would remove the

---

<sup>13</sup> The protein data bank

<sup>14</sup> Benchmarking sets for molecular docking

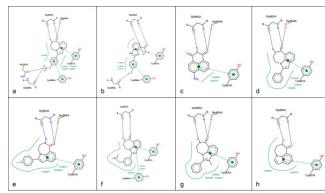


Figure 20.5: Figure 5. Aligned visualization of Poly ADP Ribose Polymerase complexes

**Aligned visualization of Poly ADP Ribose Polymerase complexes.** The ligands were taken from the ZINC database<sup>14</sup> and docked to a protein provided by the PDB (PDB code:

collision between interaction lines and the ligand structure diagram. For all other complexes, a collision-free layout could be generated.

Figure 6 shows the same series like Figure 5, and the complexes are in the same order but in contrast to the previously shown depiction the complexes are drawn independently from each other in their default orientation. In this case, the similarity is not as obvious as in the aligned visualization with a consistent residue layout.

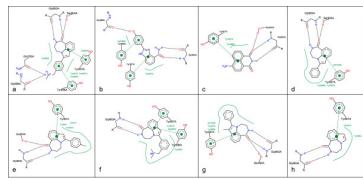


Figure 20.6: Figure 6. Visualization of Poly ADP Ribose Polymerase complexes with default orientation

**Visualization of Poly ADP Ribose Polymerase complexes with default orientation.** The depicted complexes are identical to the protein-ligand complexes in Figure 5, but their layout is calculated independently from each other.

## 20.4.2 Estrogen Receptor $\alpha$ (ER $\alpha$ )

In contrast to the former example, the ligands of ER\* $\alpha$ \* feature a more heterogeneous picture (Figure 7). The data set is an representative collection of crystal structures from the PDB (PDB codes, ordered according to the complex scoring of PoseView: 7a, b, d, e, and 7l the carboxylate of Glu 353A is hydrogen bonded to the hydroxyl group of the ligand. Due to its conformational flexibility, the interaction switches from one to the other carboxyl oxygen between the different complexes. 7a (\* $\pi$ \* stacking interaction and in the second case it is not.

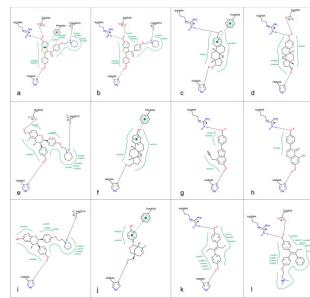


Figure 20.7: Figure 7. Aligned visualization of Estrogen Receptor \*\* $\alpha$ \*\* complexes

\*\* $\alpha$ \*\* complexes Aligned visualization of Estrogen Receptor

### 20.4.3 Urokinase (UK)

The third example consists of a randomly selected subset of the UK complex series provided by Brown and Muchmore<sup>15</sup> (different ligands bound to 1OWK), Figure 8. Here, the binding mode shows only minimal variations and is characterized by a salt bridge between Asp 191A and an amidinium group of the ligand. The remaining hydrogens of this amidinium group form a hydrogen bond to Gly 220A on the one side and Ser 192A on the other side of the Asp191A. Figure 8o shows a disadvantage of the template based orientation of the residues: contrarily to all other complexes the Gln 194A has an acceptor function but the side chain oxygen points not towards the ligand and a collision occurs. In case of flexible side chains and different possible protonation states a compromise between a consistent layout and flexibility in order to avoid collisions is necessary.

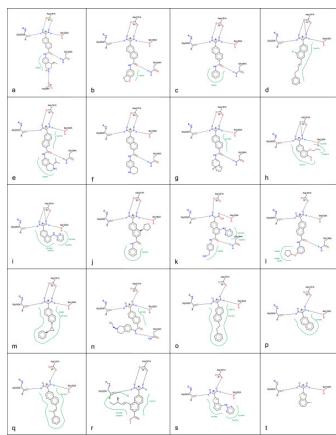


Figure 20.8: Figure 8. Aligned visualization of Urokinase complexes

**Aligned visualization of Urokinase complexes.** The ligands are taken from a dataset provided by Brown and Muchmore<sup>15</sup>. The protein file is stored in the PDB with the accession code

## 20.5 Discussion

We have implemented an extension of the PoseView algorithm that automatically generates consistent residue layouts for series of related complexes with different ligands bound to one protein. The layout generation is performed receptor-based taking into account the 3D residue adjacencies as well as the ligand topology. If not defined, the interactions and the resulting complex ensemble can be determined during run time.

All presented test sets feature a good overall layout quality that is comparable to the results of the PoseView version for single complexes. The ligand and the residues forming directed interactions are drawn in atomic detail as structure diagrams and arranged such that the visualized complexes are mainly collision free. As intended, the comparability and legibility within a complex series was considerably improved due to the consistent residue layout. While the residue orientation is fixed the ligand orientation changes over the different diagrams. An example can be found in Figure 7b and 7i. This is caused by the difference in the interaction patterns and the minimization of the deviation between the optimal interaction directions of the ligand and the real interaction directions given by the globally set amino acid positions. In contrast to known methods<sup>23</sup>, this approach combines a high degree of detail considering the IUPAC structure diagram conventions with the independence from any particular interaction model. An unsolved challenge is the handling of different protonation states and side chain orientations within one series. Also the depiction of residues which form no interactions to the particular ligand, for example colored light grey, would enhance the readability.

In summary, the presented method extends the field of complex visualization. The aligned depiction of related complexes in atomic detail offers the possibility to get a quick insight in the differences and similarities within a series.

<sup>15</sup> Large-Scale Application of High-Throughput Molecular Mechanics with Poisson-Boltzmann Surface Area for Routine Physics-Based Scoring of Protein-Ligand Complexes

## 20.6 Competing interests

The authors declare that they have no competing interests.

## 20.7 Authors' contributions

KS developed, implemented and tested the presented method. KS prepared the manuscript for this publication. MR supervised and coordinated the project. Both authors have read and approved of the final manuscript.

## 20.8 Acknowledgements

We thank Birte Seebeck and Nadine Schneider for their support in the test set preparation. The project was funded by the Klaus Tschira Stiftung gemeinnützige GmbH.



# PREDICTING A SMALL MOLECULE-KINASE INTERACTION MAP: A MACHINE LEARNING APPROACH

## 21.1 Abstract

### 21.1.1 Background

We present a machine learning approach to the problem of protein ligand interaction prediction. We focus on a set of binding data obtained from 113 different protein kinases and 20 inhibitors. It was attained through ATP site-dependent binding competition assays and constitutes the first available dataset of this kind. We extract information about the investigated molecules from various data sources to obtain an informative set of features.

### 21.1.2 Results

A Support Vector Machine (SVM) as well as a decision tree algorithm (C5/See5) is used to learn models based on the available features which in turn can be used for the classification of new kinase-inhibitor pair test instances. We evaluate our approach using different feature sets and parameter settings for the employed classifiers. Moreover, the paper introduces a new way of evaluating predictions in such a setting, where different amounts of information about the binding partners can be assumed to be available for training. Results on an external test set are also provided.

### 21.1.3 Conclusions

In most of the cases, the presented approach clearly outperforms the baseline methods used for comparison. Experimental results indicate that the applied machine learning methods are able to detect a signal in the data and predict binding affinity to some extent. For SVMs, the binding prediction can be improved significantly by using features that describe the active site of a kinase. For C5, besides diversity in the feature set, alignment scores of conserved regions turned out to be very useful.

## 21.2 Background

The question whether two molecules (a protein and a small molecule) can interact can be addressed in several ways. On the experimental side, different kinds of assays<sup>1</sup> or crystallography are applied routinely. Target-ligand interaction is an important topic in the field of biochemistry and related disciplines. However, the use of experimental methods to screen databases containing millions of small molecules<sup>2</sup> that could match with a target protein, for instance, is often very time-consuming, expensive and error-prone due to experimental errors. Computational techniques may provide a means for speeding up this process and making it more efficient. In particular in the area of kinases, however, docking methods have been shown to have difficulties so far<sup>3</sup> (Apostolakis J: Personal communication, 2008). In this paper, we address the task of interaction prediction as a data mining problem in which crucial binding properties and features responsible for interactions have to be identified. Note that this paper is written in a machine learning context, hence we use the term “prediction” instead of “retrospective prediction” that would be used in a biomedical context.

In the following, we focus on protein kinases and kinase inhibitors. Protein kinases have key functions in the metabolism, signal transmission, cell growth and differentiation. Since they are directly linked to many diseases like cancer or inflammation, they constitute a first-class subject for the research community. Inhibitors are mostly small molecules that have the potential to block or slow down enzyme reactions and can therefore act as a drug. In this study we have 20 different inhibitors with partially very heterogeneous structures (see Figure 1).

We developed a new computational approach to solve the protein-ligand binding prediction problem using machine learning and data mining methods, which are easier and faster to perform than experimental techniques from biochemistry and have proven successful for similar tasks<sup>567</sup>. In summary, the contributions of this paper are as follows: First, it uses both kinase and kinase inhibitor descriptors at the same time to address the interaction between small heterogeneous molecules and kinases from different families from a machine learning point of view. Second, it proposes a new evaluation scheme that takes into account various amounts of information known about the binding partners. Third, it provides insight into features that are particularly important to achieve a certain level of performance.

This paper is organized as follows: In the following sections, we first present the methods and datasets we used, then we give a detailed description of variants of leave-one-out cross-validation to measure the quality of predictions, present the experimental results and finally draw our conclusions.

## 21.3 Materials and methods

### 21.3.1 Data

This section introduces the Ambit Biosciences’ dataset<sup>7</sup> that provides us with class information for our classification task. From the dataset we define a two class problem by assigning to each kinase inhibitor pair “binding” or “no binding” according to the measured affinities of interaction read out by quantitative PCR. This dataset is obtained by ATP site-dependent competition binding assays and represents the first approach to mass screening of protein kinases and inhibitors. Table 1 shows overview statistics concerning the size and the class distribution of the dataset. Table S1 in Additional File<sup>8</sup>.

Additional file 1

**In this additional file we show in a table how often an inhibitor binds to a certain group of kinases (group in a phylogenetic meaning).** It can be clearly seen that nearly all inhibitors bind to several kinase groups. This means that there is generally no kinase group to that an inhibitor binds consistently.

---

<sup>1</sup> Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G

<sup>2</sup> Efficient database screening for rational drug design using pharmacophore-constrained conformational search

<sup>3</sup> Modified AutoDock for accurate docking of protein kinase inhibitors

<sup>5</sup> Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines

<sup>6</sup> Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds

<sup>7</sup> Classifying ‘drug-likeness’ with kernel-based learning methods

<sup>8</sup> Kinase, inhibitor data, features, binding matrix

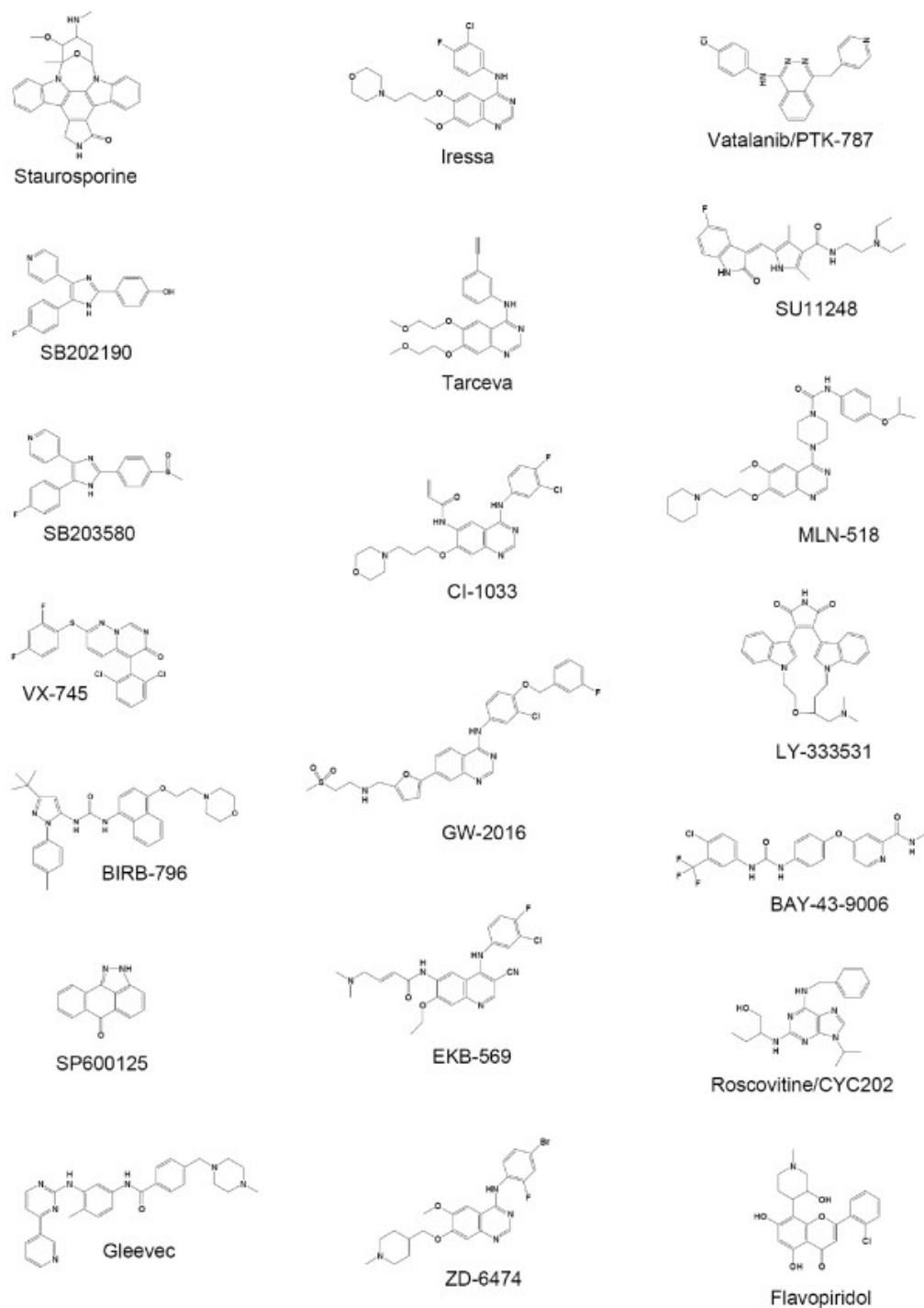


Figure 21.1: Figure 1. Training set inhibitors

**Training set inhibitors.** Structures of the 20 inhibitors that were subject of our study<sup>4</sup>.

[Click here for file](#)

For assessing the specificity of protein kinases for inhibitors, Ambit applied ATP site-dependent competition binding assays that directly and quantitatively measure the binding of inhibitors to the ATP binding site of kinases.

### 21.3.2 General approach

According to the SAR (Structure Activity Relationship) paradigm that activity is related to structure, we put the focus on features that describe the structure of molecules, that are leading to certain structures (sequence-based features) or that determine the chemical environment of the active site of a kinase or the molecule as a whole with respect to the inhibitors. Features from further categories may also give hints whether an interaction can take place, like information about the similarity of molecules, e.g., alignment scores or other phylogenetic features.

### 21.3.3 Feature generation

In the following, we will describe the features chosen to describe the kinases and inhibitors.

#### Feature generation for kinases

One class of features for the kinases represent sequence-based features. These features are derived from consecutive patterns of single amino acids. However, only frequent patterns are regarded as interesting, since only those can be informative for prediction. Active sites, for instance, are usually highly conserved and of special interest since inhibitors bind to that region. We scanned the PROSITE database (release 51.2) for PROSITE patterns that match our protein kinase sequences. These patterns are characteristic clusters of residue types occurring over a rather short section of a protein sequence. For the generation of further frequent patterns we implemented Agrawal and Srikant's APriori algorithm<sup>9</sup> with minor modifications. During the levelwise search<sup>10</sup> that enables us to find all frequent patterns, we count per example: multiple occurrences within a kinase are only counted once. As a refinement operator to generate patterns for the next level, we used pattern merging. In a pattern we allow wildcards, however, their number is restricted to two in order to reduce the search space. We excluded wildcards at the beginning and at the end of patterns, since they do not carry any significant biological information. The sequence-based features are represented in bitvectors that indicate whether a sequence is present or absent in a kinase.

Each position in a bitvector corresponds to a sequence where "1" indicates presence and "0" absence. Another class of features are phylogeny-based features. We extract available phylogenetic information about the kinases from KinBase<sup>11</sup> since a closer phylogenetic relationship implies a higher sequence similarity and thus also a higher similarity in the overall 3D structure - especially at conserved sites like the active center. We grouped the kinases into Serine/Threonine and Tyrosine kinases and made a finer division into kinase groups and kinase families. The phylogeny-based information is presented in nominal form, e.g. the phylogenetic feature "group" has several categories like AGC, CAMK or CK1.

Other phylogenetic features are directly derived from kinase sequence alignments. We implemented three different types of alignment procedures. First, a global alignment algorithm<sup>12</sup> was used that aligns two amino acid sequences over their full length. The protein kinases we investigated differ enormously in length, reaching from 275 to 1,607 amino acids, and therefore many gaps requiring to be introduced. This may obscure existing evolutionary relationships making therefore the alignment scores less useful. To overcome this problem, first we applied CLUSTALW<sup>13</sup> to get a multiple sequence alignment (MSA), from which we selected highly conserved sequence stretches (frames) where each frame must satisfy the following three criteria:

<sup>9</sup> Fast algorithms for mining association rules

<sup>10</sup> Levelwise search and borders of theories in knowledge discovery

<sup>11</sup> KinBase: The kinase database at Sugen/Salk

<sup>12</sup> A general method applicable to the search for similarities in the amino acid sequence of two proteins

<sup>13</sup> CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice

- At least one position in a frame must be highly conserved. A highly conserved position is a residue in which one amino acid occurs in more than 100 out of the 113 cases.
- The frame border is at most five amino acids away from a highly conserved position.
- A highly conserved position must be part of the active center.

For each kinase pair, each frame is matched with the corresponding frame from the second sequence and scored according to the scoring matrices (see below). The scores for the frame pairs are summed up to get an overall score for all conserved regions, whereas a higher score should indicate a more similar active center and more similar binding properties. To calculate the scores, we implemented two different techniques. First, we just cut out the amino acid sequences of the frames from the MSA and scored it without any further modification. Second, we realigned the cut out frames before calculating the total score. For all alignment procedures we used PAM120 and Blosum62 as substitution matrices with uniform costs for gap opening and gap extension. The alignment-based features are represented with a 113-dimension vector where each dimension or position represents a numeric alignment score.

Additionally, we also use single residues which contribute to the active site and inhibitor binding as position specific features. These features' values are either the respective amino acid or the physico-chemical class of the amino acid. In Table 2, the features, the number of features, and the feature type in each group used in our study for describing the kinases are summarized.

### Feature generation for inhibitors

For the description of the inhibitors we used features based on their 2D structures, preferred binding partners (primary targets) and binding patterns. We visually clustered the inhibitors by simply looking at their 2D structure so that inhibitors with similar shapes are grouped together (see Table 3). Primary targets are kinases for which an inhibitor shows a highly preferred binding compared to other kinases. In this context, "primary target" concerns kinases in general and is not restricted to the kinases under consideration in this paper. Binding patterns represent the binding behavior of an inhibitor to a set of kinases and may serve also as features since similar properties on known targets give hints to binding properties on unknown targets. Therefore, we implemented a  $k$ -nearest-neighbor method (KNN) to detect each inhibitor's  $k$  nearest neighbors. In this study we used  $k = 3$ . The calculation is based on data from Ambit's binding matrix. Note that this calculation is only possible in the "soft case" evaluation (to be presented below) since only in that case all the information of a test kinase (respectively inhibitor) relative to all training inhibitors (respectively kinases) is given. As a distance measure, we used a function counting the number of common bindings of two inhibitors  $i$  x,  $j$  x( $c$ ), and the more complex Tanimoto coefficient (1) that counts besides ( $c$ ) the number of bindings of inhibitor  $i$  x to a kinase ( $a$ ) and the number of bindings of inhibitor  $j$  x to the same kinase ( $b$ ):

To describe the inhibitor structures, we applied the graph mining tool Free Tree Miner (FTM)<sup>14</sup>. With this tool, the 2D structure representations of the inhibitors are mined for frequently occurring acyclic substructures. Such substructures can describe, for instance, a hydrophobic group in an inhibitor important for bindings or an extended region that would exclude small active sites of kinases as binding partners due to steric hindrance. To avoid the exclusion of probably important substructures right from the beginning, we set the minimum frequency threshold rather low to 10%.

Additionally, we also calculated geometric features of the inhibitors from 2D data like their diameter, length and width that might prevent kinase binding due to steric hindrance.

Besides this, various chemical features determine whether or not a binding at the active site can take place. For the calculation of such features, we applied the cheminformatics library JOELib2<sup>15</sup>. In this way, we obtained the following physicochemical features: XlogP, molecular weight, hydrogen bond acceptor/donor count, rotatable bond count, tautomer count and topological polar surface area. All these features are suitable for building basic structure-activity relationship (SAR) models<sup>16</sup>. We also described the inhibitors with pharmacophores. A pharmacophore is, in general, a 3D substructure of a molecule that is meaningful for its medical activity. It can be seen as an abstraction of the molecular structure to a usually small number of key features that contribute to the majority of the activity together with their geometric arrangement that is represented by pairwise distances. For the actual calculation of the pharmacophores we,

<sup>14</sup> Frequent free tree discovery in graph data

<sup>15</sup> JOELib2: A Java based cheminformatics (computational chemistry) library, 2008

<sup>16</sup> NOTITLE!

only for simplicity, used the 2D information of the inhibitors. We calculated so-called 3-point pharmacophores<sup>17</sup> for our set of inhibitors. Such pharmacophores consist of three essential atoms (negatively or positively charged atoms, acceptor or donor atoms) and their distances in space. We calculated all 3-point pharmacophores, sorted the atoms lexicographically in order to avoid duplicates, and used the atoms as well as their (discretized) distances as features. Table 3 summarizes the features, the number of features and the feature type in each group that we used for describing the inhibitors.

The instances consisting of kinase-inhibitor pairs are represented by concatenating kinase and inhibitor feature vectors, i.e. each kinase is concatenated with each inhibitor. Formally, this can be stated as

where  $\mathbf{k}_i$  represents the feature vector of the  $i^{th}$  kinase and  $\mathbf{l}_j$  the feature vector of the  $j^{th}$  inhibitor.

#### 21.3.4 Feature selection/reduction

Feature selection techniques attempt to determine appropriate features that can discriminate well between classes. Feature sets that are too larger may contain many uninformative features leading to overfitting or a decrease in prediction accuracy or efficiency. On the other hand, feature sets which are too small may not contain enough information to determine the target class and may cause underfitting.

The feature sets generated by APriori usually contain many solution patterns which are redundant or less useful as they are too small (i.e., strings/trees of length one). Such elements can be removed, and the size of the complete solution set can be reduced significantly, e.g. by computing so-called *border elements*<sup>18</sup>, i.e., the most specific patterns that are still solutions. We calibrated Free Tree Miner to solely output border elements. APriori was implemented to output only features that are border elements and larger than a user defined size threshold. Finally, we used in our study 14 sequence-based apriori features and 78 free trees (see Tables 2, 3).

#### 21.3.5 Classification

For classification, we used standard schemes like decision tree (C5) and large margin (SVM) learning methods. C5<sup>19</sup> is commercial improvement of C4.5<sup>20</sup> written in C and popular for its efficiency. For the SVM<sup>21</sup>, we used Weka's<sup>22</sup> implementation of Sequential Minimal Optimization (SMO)<sup>23</sup>. We tested three kernels (linear (E1), quadratic (E2) and radial basis function (RBF)) with Weka's default parameter setting including the cost factor  $C = 1.0$ . A higher  $C$  slows down the running time of the classifiers. A  $C$  of 0.1, however, renders the RBF kernel SVM to a majority class predictor. For an SVM with a linear kernel the opposite is true, but it performs in all cases on a lower level. The performance of the quadratic kernel SVM remains nearly the same on the test data, on the training data, however, a smaller  $C$  decreases the predictive power. For C5, the main task consists of finding the best pruning options to control overfitting. C5 provides the option to prune with confidence intervals and with a minimum support of training instances that must be covered by each leaf of the tree. We used C5's default settings, with a pruning confidence factor of 25% and a minimum support in each leaf of 2. Subsequently, global pruning can be used to optimize the tree's performance further.

Note that for C5, continuous or numeric features are discretized using standard procedures<sup>20</sup>. For SVMs, nominal features are transformed to “binary numeric” using Weka's standard filter *NominalToBinary*<sup>2224</sup>. All features used within SVMs are normalized by the Weka workbench by default. The kernels we applied are constructed out of all these normalized features.

---

<sup>17</sup> NOTITLE!

<sup>18</sup> NOTITLE!

<sup>19</sup> Data Mining Tools See5 and C5.0, 2008

<sup>20</sup> NOTITLE!

<sup>21</sup> A tutorial on support vector machines for pattern recognition

<sup>22</sup> The WEKA data mining software: An update

<sup>23</sup> Improvements to Platt's SMO algorithm for SVM classifier design

<sup>24</sup> NominalToBinary: A Java class for converting nominal features to “binary numeric”, 2011

## 21.4 Results

### 21.4.1 Evaluation

We use leave-one-out cross-validation (LOOCV) to evaluate our classification results. LOOCV may appear uncommon, at first sight, in this setting with 2260 instances since it is generally recommended (along with the bootstrap) for smaller datasets. This is because a smaller number of folds would result in an even larger variance. LOOCV is known to deliver estimates with a small bias, whereas the variance can be high. However, with more than 2000 instances, the training sets do not vary a lot; therefore, even the variance is low in this case. Usually, ten times ten-fold cross-validation is preferred on such datasets for practical reasons, to avoid the excessive running times of LOOCV. However, we wanted to test the “purest” setting and also obtain maximally unbiased error estimates. Finally, it should be clear that the proposed evaluation variants can easily be extended towards regular k-fold cross-validation, by leaving out pairs of sets of kinases and sets of inhibitors in turn. To evaluate the quality of a model, we used three established performance measures: In-/correctly classified instances, recall and precision:

Note that (4) is also known as Sensitivity and True Positive Rate (TPR), (5) as Selectivity and Positive Predicted Value (PPV), (6) as Specificity and True Negative Rate (TNR), and (7) as Negative Predicted Value (NPV).

In the following, we will present a new way of evaluating classifiers in the present setting, and give an overview of four different variants of LOOCV applied here. Since we aim for predictions for pairs of kinases and inhibitors, different amounts of information may be available for the two potential binding partners.

#### The hard and the soft case

Figure 2 shows the two extreme variants of our different implementations of LOOCV. The left-hand side of the figure shows the “hard case”, in which no information about the test kinase and the test inhibitor is allowed in the training dataset. This would, for instance, represent the scenario in which a binding prediction is performed for a completely unknown pair of a kinase and an inhibitor. In the “soft case” (see the right-hand side of Figure 2), however, all information about the test kinase and the test inhibitor is already known in the training set - except for the pair itself to be predicted.

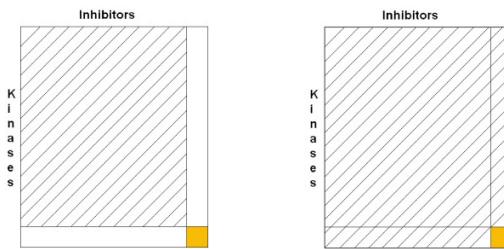


Figure 21.2: Figure 2. Hard and soft case of LOOCV

**Hard and soft case of LOOCV.** Illustration of the hard (left) and the soft (right) case of LOOCV.

#### The mixed and the mixed-mixed case

The two cases between the extreme variants of our different implementations of LOOCV are shown in Figure 3. The left side of the figure illustrates the “mixed case”, in which the equal percentage of information on the test kinase and the test inhibitor are put into the training set. This means that a certain random fraction from the test kinase and the same random fraction from the test inhibitor is put into the training set. To give an example, if we use 50% of the test inhibitor information, we put 10 kinase-inhibitor pairs in the training set where the inhibitor in the pair must be the inhibitor to be predicted. For the kinase the same holds, but 50% make up 57 pairs. On the right side, the “mixed-mixed case” is illustrated, in which the training dataset contains information on the test kinase and test inhibitor in an

unequal proportion. This represents the situation in which experimental information concerning the binding patterns of a certain test kinase to the inhibitors is partly available. For the test inhibitor, the same holds, but in a different proportion.

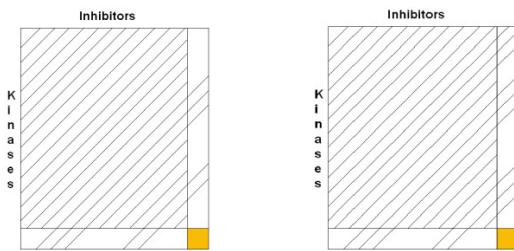


Figure 21.3: Figure 3. Mixed and mixed-mixed case of LOOCV

**Mixed and mixed-mixed case of LOOCV.** Illustration of the mixed (left) and the mixed-mixed (right) case of LOOCV.

### Results for different feature sets

We start with the results from the soft case evaluation. In the following, we show how different feature sets (see Table 4) affect the performance of the classifiers. The overall plan of the experiments was (1.) to start with a set of base features for both kinases and inhibitors, (2.) to refine the representation of kinases in the next step, and after that (3.) to refine the representation of inhibitors. For the representation of kinases, we add alignment-based features, then position-specific features, and ultimately both alignment-and position-specific features. For the representation of inhibitors, we start with the base features and then add further descriptors (CF, GF and P in the table) in a final refinement step.

In preliminary experiments, we evaluated the individual performance of feature groups (Table 5). Here the features for kinases perform very similarly, all in a range between 73% and 74% (for C5), whereas the features for inhibitors differ in their performance: the predictive accuracy of CF, GF, KNN, FTs, and P range between 79% and 80% (again for C5), with the remaining two feature groups (PT and MS) lagging behind. Results for SVMs are mostly comparable (see Table 5).

Figure 4 shows the prediction accuracies for different feature sets for both SVMs and C5 that were run with different parameter settings. In all cases C5, without the option “global pruning”, outperforms the other variant with global pruning. Compared to the SVM, it is extremely fast and handles large feature sets well concerning runtime and memory.

For C5, the usage of global alignment scores as features (FS2) reduced the prediction accuracy on test data significantly. This may be explained by the fact that these scores take into account the complete amino acid sequence, whereas only a small part constitutes the active center and is therefore important for the binding to an inhibitor. These non-informative sequences clearly outweigh the informative ones, and so they obscure the information and make the scores for global alignments less useful. Alignment scores of extracted conserved regions perform clearly better on the training and test set. In this case it is particularly remarkable that cut out and realigned conserved regions (FS3) perform 1.6% better than without realigning on the test set.

The most difficult prediction task for the applied classifiers was the correct prediction of a binding between an inhibitor and a kinase. For all different feature sets, C5 as well as SVMs have the lowest values for the recall of the positive class (Figure 4). Particularly conspicuous are the extremely low values for the recall of the positive class of SVMs with an RBF and a linear kernel. For C5, feature sets 3, 6 and 7 clearly show the best positive recall (and prediction accuracy) for the test data. Particular attention should be paid to FS3, which comprises, besides the basic feature set, only local alignment scores. This indicates that local alignment scores are very suitable for making predictions with C5. The negative recall is relatively constant for all feature sets. However, the tradeoff between positive and negative recall is visible.

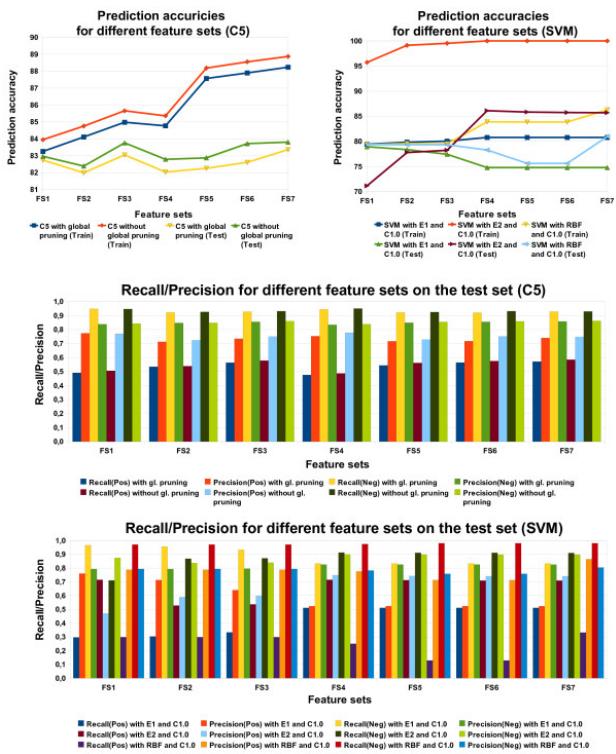


Figure 21.4: Figure 4. Performance on different feature sets (soft case)

**Performance on different feature sets (soft case).** Prediction accuracies, recall and precision for different feature sets from C5 and Support Vector Machines with different parameter settings (soft case).

Comparing FS1 with FS3, or FS4 with FS7, shows that a higher positive recall leads to a lower negative recall. A combination of global and local alignment scores, as well as a combination of position specific features with abstract position specific features, degrades the predictivity. Only when we combined alignment score features (FS4) with position specific features (FS5) to feature set 6, the prediction accuracy increases significantly on the training and test set. This combination is then further improved by adding chemical features leading to the best prediction accuracy we obtained from C5 on the test set. This success can largely be attributed to the use of chemical features and the diversity of the features.

With SVMs, we used global alignment scores as features (FS2) only once, since they slow down the computation enormously and when combined with other feature sets, do not contribute to an increase in the prediction accuracy on the test set. However, we tested the influence of using position specific features with (FS5) and without abstractions (FS4), where the use of abstract position specific features did not show an improvement of the prediction accuracy.

### Comparison of different kernels in SVMs

The differences between the kernels are clearly observable from Figure 4. The quadratic kernel performs with higher success than the linear and RBF kernel for all feature sets except for FS1 and FS2 for both kernels and FS3 for the RBF kernel on the test set. On the training set, this fact must be mainly attributed to overfitting. On the test set, the best results are obtained with feature set 4. This indicates that SVMs with a quadratic kernel work best with position specific features. This may be due to the fact that we described here for the first time the active site of a kinase with position specific features. An addition of further features does not lead to an increase in the predictive power for both training and testing.

For the linear and RBF kernel things are different. A larger amount of features increases the predictivity on the training data set but harms it on the test data set except for FS7 with the RBF kernel. From Figure 4 it is evident that the recall for the negative class normally drops from feature set 1 to feature set 7 for the linear kernel. The reason for this may be that SVMs are not able to predict the “binding” class with features that do not discriminate immediately between the classes. Hence, SVMs mostly predict the majority class “no-binding”, leading to a high negative recall. But with increasing ability to discriminate between the classes, more bindings are predicted correctly, leading to an increase in the recall for the positive class. On the test set this is accompanied by a decrease in the recall for the negative class and an increase in its corresponding precision (Figure 4). For the RBF kernel we obtain the best predictivity as well as the highest positive and negative recall with FS7. The chemical features seem to be the most decisive ones with respect to the RBF SVM. From feature set 1 to feature set 6 the negative recall and precision remain relatively constant. The positive recall and precision, however, decrease significantly.

### Performance on random feature sets

We also tested the performance on features sets with random feature values. For feature set 3, we assigned random integers (FS3\_ran) to all local alignment score values where a random integer must be in the range between the smallest and largest value of the true values. Results for C5 and SVMs with a quadratic kernel are shown in Figure 5. As expected, the usage of random features harms performance. For C5, it is visible that for feature set 3 the drop of the performance is larger than for feature sets 6 or 7. The same holds for SVMs with respect to feature set 4 and 7. This can be explained by the fact that the amount of random features for feature set 3 is much higher than for feature set 7, for instance, where random values are assigned to three rather small feature groups (chemical, geometrical and pharmacophoric features). Particularly, the worse performance of the random feature sets 6 and 7 indicates that the sheer size of a feature set does not necessarily contribute to a better performance through chance correlation or overfitting, but that the diversity of a feature set is the factor positively impacting predictivity.

### Results for the hard and the soft case evaluation strategy

All the results in this section are based on feature set 7. C5 shows only slight, nonsignificant differences between the hard and the soft case concerning the training data, however, on the test data there are large differences (Figure 6). In

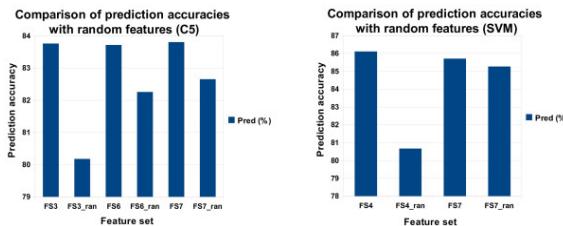


Figure 21.5: Figure 5. Comparison of prediction accuracies with random features

**Comparison of prediction accuracies with random features.** Comparison of prediction accuracy for different feature sets including random features.

the hard case, recall and precision values for the positive class are very low, which indicates that the classifier is not good at identifying kinase and inhibitor features responsible for binding. From the hard to the soft case, these two values clearly increase. In contrast, there is only a small drop in the recall and the precision for the negative class. As for C5, the prediction accuracy on the test and the training set of SVMs (with a linear, quadratic or RBF kernel and cost factor C = 1.0) always increases from the hard to the soft case. In the hard case, there are even worse results regarding the recall for the positive class (5.0% recall). The quadratic kernel clearly overfits on the training data.

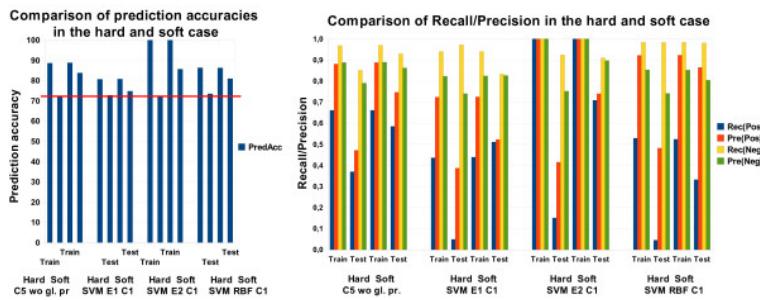


Figure 21.6: Figure 6. Performance comparison of the hard and the soft case

**Performance comparison of the hard and the soft case.** Comparison of the prediction accuracy and recall/precision in the hard and the soft case.

In some hard cases for different feature sets (not shown here), especially for small feature sets with predominantly features that are not able to discriminate between classes, C5 as well as SVMs are performing as good as a majority class classifier since they predict everything as non-binding. For more complex feature sets like in the cases shown here, C5 and the SVM classifiers with a linear or quadratic kernel are slightly worse than a majority class classifier (red line in Figure 6) that would reach 73.6%. SVMs with an RBF kernel, however, perform slightly better although the recall for the positive class is very low. However, in this case, this is compensated by a high precision for the positive class as well as a high recall for the negative class (see Figure 6).

In the soft case, we compare our prediction accuracy results with a simple baseline classifier that calculates the probabilities for a binding from the binding matrix, not taking into account any information about the molecules. This is only possible in non-hard cases, since the information how often a test kinase/inhibitor binds to a training set inhibitor/kinase is directly taken into account. More precisely, the simple baseline classifier calculates, separately, the probabilities  $\text{kinp}(b)$  and  $\text{inhp}(b)$  of a test kinase/inhibitor binding on the training set. Subsequently, these probabilities are multiplied and it is determined whether the product is greater than the threshold  $*\theta*$  that was optimized empirically:

A smaller value than  $*\theta*$  results in a “no-binding” prediction, otherwise “binding” is predicted. This simple classifier is able to reach 78.5% prediction accuracy without any knowledge about the kinases or inhibitors except their binding patterns. It is clearly better than a majority class classifier, but worse than models that consider additional information about the molecules.

The difference between the hard and the soft case are the kinase-inhibitor pairs in the training set that contain either the test kinase or the test inhibitor. The performance improvement of the classifiers must be attributed to these pairs. Figure 7 shows the results if we remove all kinase-inhibitor pairs from the training set which do not contain the test kinase or the test inhibitor. On the training set (consisting of 131 instances or kinase-inhibitor pairs), a strong performance can be observed. Compared to the soft case (see Figure 6 soft case), however, the results on the test set, disregarding SVMs with a linear kernel, show a clear performance loss if we remove kinase-inhibitor pairs from the training set which contain neither the kinase nor the inhibitor to be predicted. The usage of test kinase and test inhibitor information, solely, is not sufficient to obtain reasonable results. This means that the applied machine learning methods require the other pairs in order to generalize well.

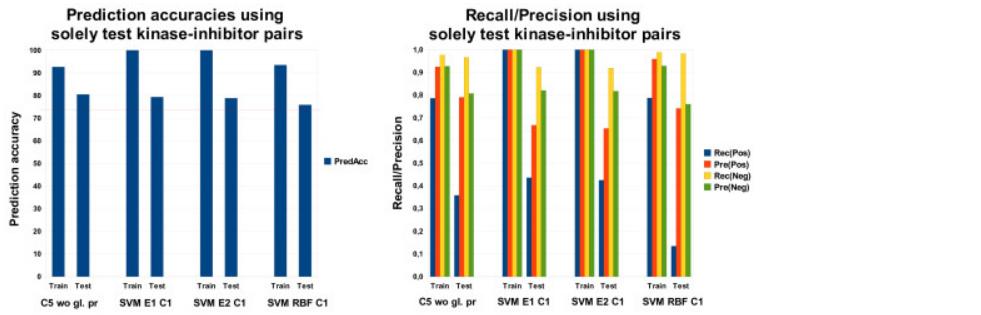


Figure 21.7: Figure 7. Performance using solely test kinase-inhibitor pairs

**Performance using solely test kinase-inhibitor pairs.** Comparison of prediction accuracy and recall/precision using solely test kinase-inhibitor pairs in the training set.

## Results for the mixed and the mixed-mixed case

All the results in this section are based on predictions of C5 that was run on feature set 7 without global pruning. Figure 8 shows the prediction accuracies of three mixed cases, the soft and the hard case, a simple baseline classifier and a majority classifier, as well as the recall and precision values for our predictions. The results for the mixed cases and the simple baseline classifier are obtained by averaging the results from ten runs of C5 with identical parameter settings. Note that we took randomly a certain fraction of test kinase/inhibitor information in the training set. This means that the results in each run can be slightly different. Hence, it is necessary to conduct several, in our case 10, experiments with the same parameter setting and average the results in order to take these variations into account.

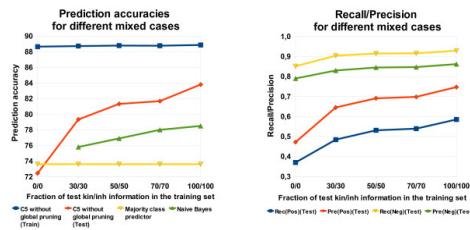


Figure 21.8: Figure 8. Performance comparison of different mixed cases (C5)

**Performance comparison of different mixed cases (C5).** Comparison of prediction accuracy and recall/precision for different mixed cases (C5 without global pruning).

The performance on the test set is strongly influenced by the usage of different fractions of the test kinases and the test inhibitors in the training set. The performance on the training data, however, is nearly independent of it (see Figure 8). The strong increase in performance from the hard case to the first mixed case (30/30) indicates the importance of the information about the test kinase and the test inhibitor. The usage of information from the test kinase and the test inhibitor leads to a substantial increase in the recall and the precision for the positive and the negative class. The

classifier learns to discriminate better between the classes and thus to predict a correct binding more often without losing performance on the negative class.

Further results (not shown here) reveal only slight differences in the training set performance for the mixed-mixed case. Here, the classifier performs clearly independent of the number of instances from the test kinase and the test inhibitor in the training set. On the test set, however, there is great variability in the prediction accuracy of different mixed-mixed cases. First, we analyze the performance if a percentual amount of information on the test kinase and the test inhibitor is added to the training set. Note that this means that actually more information on the test kinase is added since the dataset consists of more kinases than inhibitors. Second, we analyze the case in which equal information on the test kinase and the test inhibitor is added to the training set, i.e. if 10 pairs consisting of the test kinase and 10 different training set inhibitors are added, then 10 pairs of the test inhibitor and 10 different training set kinases also have to be added. Further note that the training set without test molecule information can be seen as a reference. This reference represents the hard case. For both variants, it is tested how information on test kinases and test inhibitors in the training set can improve the performance compared to the reference. From Figure 9, the worst and the best values are obtained in the hard and the soft case, respectively. The same holds for the corresponding cases in Table 6 (0/0 and 19/19). Results with percentual and absolute test molecule information are similar. For the dataset under investigation, it can clearly be seen that the information about the kinase test molecule is more important than that of the tested inhibitor. If no information from the test inhibitor and only little information from the test kinase is in the training set, the prediction accuracy increases significantly. In contrast, no information from the test kinase and little information from the test inhibitor leads to a remarkably lower increase in the prediction accuracy (see Figure 9 and Table 6). This can be clearly seen, for instance, in Table 6 for the cases 10/0 and 0/10.

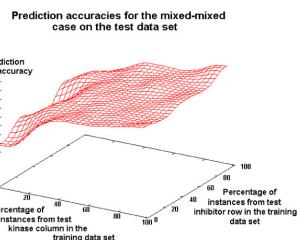


Figure 21.9: Figure 9. Performance comparison of different mixed-mixed cases (C5)

**Performance comparison of different mixed-mixed cases (C5).** Comparison of prediction accuracy for the soft, hard, mixed and mixed-mixed cases (C5 without global pruning).

We benchmarked our kinase inhibitor binding prediction approach for the mixed-mixed cases and the soft case with a majority class classifier as well as with a simple baseline classifier (Table 7). The more informed classifier, C5, performs in all cases, except one, better than both reference classifiers. Mostly, a clear performance improvement with respect to prediction accuracy can be observed. This means that the feature extraction from the kinases and the inhibitors is beneficial. The same holds for the mixed cases and the soft case.

In summary, the best model achieved 83.8% predictive accuracy with a recall of 0.59 and a precision of 0.75 for the positive class. The most frequently used features in the learned decision tree are position-specific features, local alignment features and the JOELib2 chemical features.

## Results for an external test set

In addition, the classifiers were tested on an external dataset consisting of 19 kinase inhibitors and 177 protein kinases<sup>25</sup>. It is the result of a later study from Ambit Biosciences and produced in the same way as the dataset described in the section “Data”. Note that the original dataset consists of 38 inhibitors and 317 kinases. We removed inhibitors and kinases that are contained in the training set<sup>7</sup> and those where information is missing needed for descriptor calculation. The class distribution of this compiled dataset is similar with 26.6% bindings and 73.4% non-bindings. Testing on an external dataset corresponds to the hard case since neither information about the inhibitors nor information about

<sup>25</sup> A quantitative analysis of kinase inhibitor selectivity

the kinases is available. The best results on feature set 7 we obtained with C5 without global pruning (prediction accuracy on test set: 74.1%). SVMs with an RBF kernel are also able to outperform a majority class predictor. SVMs with a linear or quadratic kernel, however, perform slightly worse than a majority class predictor (Figure 10). These results represent a clear improvement in comparison to the hard case in the LOOCV setting. Primarily, this improved performance can be explained by structurally similar inhibitors present in the training set, which are not available in LOOCV (see Figure 1).

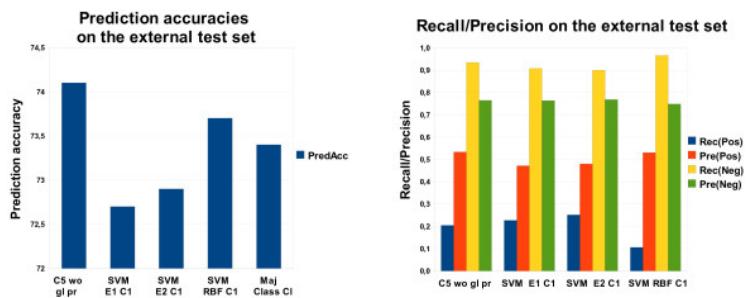


Figure 21.10: Figure 10. Performance on the external test set

**Performance on the external test set.** Prediction accuracy and recall/precision on the external test set with feature set 7, for both C5 and SVMs.

## 21.5 Related Work

Kinase inhibitor predictions have been investigated over the past few years. Basically, there exist two approaches. The simpler, and more established one, is to calculate a vectorial representation of both kinase inhibitor and non-kinase inhibitor molecules and using the result with standard machine learning algorithms to predict the probability of a molecule to be a kinase inhibitor. This approach was, for instance, taken by Briem and Günther<sup>26</sup>. In their study, they used a Schering in-house dataset of small molecules encoded by 120 fragment-based Ghose-Crippen descriptors and applied several machine learning techniques (SVMs, artificial neural networks, kNN with GA-optimized feature selection and recursive partitioning) to distinguish between kinase inhibitors and molecules with no reported activity on any protein kinase. Since the original dataset was strongly imbalanced, a ensemble-based sampling procedure was applied to ensure balanced training sets. In the end, 13 training sets were generated for model learning and applied to an independent test set. Briem and Günther analyzed to what extent machine-learning algorithms are capable of learning kinase inhibitor likeness and to compare the different classifiers. Results are reported for each of the 13 individual sample classifiers and for a consensus majority vote of all members of the ensemble. The results show that the latter generally outperforms averaging over the individual models. All four methods exhibited a reasonable discriminative power. Comparing the individual classifiers with respect to standard quality measures, SVMs seem to be the best choice. This is also true for a further compiled test set with significantly different structures.

Xia and colleagues<sup>27</sup> used a modified Naive Bayes classifier to model multifamily and single-target kinase inhibitors. In their study, they used Amgen's CORP datasets (around 200.000 molecules) composed of kinase inhibitors, potential kinase inhibitors, and random drug molecules. To describe the molecules, standard physicochemical features as well as a 2D structural fingerprint were used. To assess the performance of the Bayesian model, the positions of active compounds in ordered scoring lists of the test set were used. The approach was validated by first using an equal proportion of training and test instances (1:1) and second using a much smaller training set (1:9). The results suggest that only 10% of the data are enough to yield a performance nearly as good as if 50% were used. 85% of the active compounds occurred in the top 10% of the ordered molecules. This underlines the power of the Bayesian model which is also confirmed on 172 novel kinase inhibitor compounds from different structural classes that were classified with

<sup>26</sup> Classifying "kinase inhibitor-likeness" by using machine-learning methods

<sup>27</sup> Classification of kinase inhibitors using a Bayesian model

the 1:9 Bayesian split model. 70% of these new compounds were found in the top 10% and 85% in the top 20% rank-ordered compounds.

Compared to our study, Briem and Günther as well as Xia and colleagues used only information of small molecules (kinase inhibitors, potential kinase inhibitors and random drug molecules) for predicting the probability of a molecule being a kinase inhibitor. Information about kinases is not considered.

The second, and more difficult approach, is to use features from kinase inhibitors and protein kinases in combination. Weill and Rognan<sup>28</sup> presented a novel low-dimensional fingerprint approach encoding ligands and target properties to mine the protein-ligand chemogenomic space. Kinase inhibitors are represented by standard descriptors, while protein transmembrane cavities are encoded by a fixed length bit string describing pharmacophoric properties of a defined number of binding site residues. Due to the complexity of the cavity, this study is restricted to G protein-coupled receptors (GPCRs) with a homogeneous cavity description. Several machine learning classifiers on two training sets of roughly 200.000 receptor-ligand fingerprints with different definitions of inactive decoys are applied for model learning. Two external test sets of 60 GPCRs were used to validate the models. Experimental results suggest that SVMs with an RBF kernel perform best with respect to a balanced accuracy measure combining true positive and true negative rate. The authors demonstrate that protein-ligand fingerprints outperform the corresponding ligand fingerprints in predicting either putative ligands for a known target or putative targets for a known ligand. They conclude that, with respect to GPCRs, predicting ligands is significantly easier than predicting targets.

Our approach resembles the one of Weill and Rognan in that both kinase and inhibitor information is used for modeling. However, a key difference in Weill and Rognan's approach is the restriction to GPCRs, whereas in this paper, we take a broad spectrum of different kinase families into account and thus are able to make predictions for a larger range of kinases and inhibitors.

## 21.6 Conclusion

We tackled the prediction task whether a binding between a protein kinase and an inhibitor can take place, given a set of features describing both molecules. We applied and tested a range of data mining and classification tools. Finally, we used both C5 and Support Vector Machines together with three variants of leave-one-out cross-validation to learn and validate concepts of protein kinase inhibitor bindings and the influence of information available about the potential binding partners. The approach performs well in the soft case validation and comparable with a majority class classifier in the hard case validation. On an external test set we obtained a clearly better performance than a majority class predictor with C5 and SVMs with an RBF kernel. As expected, the performance can be improved by features describing the active site of the kinases by local alignment scores for C5 or position specific features for SVMs. These features are frequently used by C5 and increase the prediction accuracy substantially. Primary chemical features and a diversity in the feature sets had a positive influence on the performance of the classifiers. Note that once a pair of a kinase and an inhibitor is classified as binding, a regression model could be applied subsequently to predict the quantity of the binding affinity.

In summary, the contributions of the paper are as follows: First, we presented a machine learning approach to modeling the binding affinity of inhibitors to kinases. In particular, it is the first time that the complete dataset of Fabian *et al.*<sup>7</sup> with information for all pairs from a set of inhibitors and a set of kinases, is used in predictive modeling (classification). Second, we proposed novel validation schemes for this kind of problem, depending on how much information is available for the inhibitor and for the kinase. Third, our experiments showed that for the decision tree learner C5 alignment-based features are very useful, but that a combination with position-specific features and certain chemical features is necessary for obtaining the best results. For SVMs, the best results are obtained with a quadratic kernel and position specific features. The best predictive accuracy of 86.1% indicates that machine learning methods are able to detect a signal in the data and predict binding affinity to some extent. However, it is clear that there is ample room for improvement for all kinds of methods and that the prediction of kinase-inhibitor binding will remain a relevant research topic for a long time to come.

---

<sup>28</sup> Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: Application to G protein-coupled receptors and their ligands

## 21.7 Competing interests

The authors declare that they have no competing interests.

## 21.8 Authors' contributions

FB implemented the methods and conducted the experiments. FB, LR and SK made substantial contributions to the conception, design and coordination of the study. All three analyzed and interpreted the results, were involved in drafting the manuscript, revised it critically for important intellectual content, and gave the final approval of the version to be published.

## 21.9 Authors' information

FB is a PhD student at the computer science department of Technische Universität München. After receiving his diploma in bioinformatics from the Ludwigs Maximilians Universität München and Technische Universität München, he began to work on predictive toxicology in the scientific staff of Prof. Stefan Kramer (SK) in the Machine Learning and Data Mining in Bioinformatics group at Technische Universität München. His current research interests include predictive toxicology, machine learning, data mining, bioinformatics and cheminformatics.

LR is a Post-doc in the group of SK. He received a diploma and a Ph.D. in biology from Technische Universität München. After a year in a biotech-startup company he came back to Technische Universität München for a postgraduate study in computer science and joined SK's group upon completion. He is interested in data mining and integration of chemical and biological data.

SK is professor of bioinformatics at the computer science department of Technische Universität München. After receiving his doctoral degree from the Vienna University of Technology, he has spent a few years as an assistant professor in the Machine Learning lab of the University of Freiburg. He was the co-organizer of the Predictive Toxicology Challenge 2000-2001, an international competition in toxicity prediction. He has organized several conferences and workshops, edited special issues of journals, given invited talks and tutorials, and serves on the program committees of major data mining and machine learning conferences and on the editorial board of the Machine Learning journal. His current research interests include data mining, machine learning, and applications in chemistry, biology and medicine.

## 21.10 Acknowledgements

OpenTox - An Open Source Predictive Toxicology Framework, <http://www.opentox.org/> is funded under the EU Seventh Framework Program: HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs (Quantitative Structure-Activity Relationships) for toxicology, Project Reference Number Health-F5-2008-200787 (2008-2011).

# 4D FLEXIBLE ATOM-PAIRS: AN EFFICIENT PROBABILISTIC CONFORMATIONAL SPACE COMPARISON FOR LIGAND-BASED VIRTUAL SCREENING

## 22.1 Abstract

### 22.1.1 Background

The performance of 3D-based virtual screening similarity functions is affected by the applied conformations of compounds. Therefore, the results of 3D approaches are often less robust than 2D approaches. The application of 3D methods on multiple conformer data sets normally reduces this weakness, but entails a significant computational overhead. Therefore, we developed a special conformational space encoding by means of Gaussian mixture models and a similarity function that operates on these models. The application of a model-based encoding allows an efficient comparison of the conformational space of compounds.

### 22.1.2 Results

Comparisons of our 4D flexible atom-pair approach with over 15 state-of-the-art 2D- and 3D-based virtual screening similarity functions on the 40 data sets of the Directory of Useful Decoys show a robust performance of our approach. Even 3D-based approaches that operate on multiple conformers yield inferior results. The 4D flexible atom-pair method achieves an averaged AUC value of 0.78 on the filtered Directory of Useful Decoys data sets. The best 2D- and 3D-based approaches of this study yield an AUC value of 0.74 and 0.72, respectively. As a result, the 4D flexible atom-pair approach achieves an average rank of 1.25 with respect to 15 other state-of-the-art similarity functions and four different evaluation metrics.

### 22.1.3 Conclusions

Our 4D method yields a robust performance on 40 pharmaceutically relevant targets. The conformational space encoding enables an efficient comparison of the conformational space. Therefore, the weakness of the 3D-based approaches on single conformations is circumvented. With over 100,000 similarity calculations on a single desktop CPU, the utilization of the 4D flexible atom-pair in real-world applications is feasible.

## 22.2 Background

Sorting and comparing molecules from chemical databases represent two of the key tasks in cheminformatics<sup>1</sup>. The sorting of such databases, with respect to a given set of queries (molecules) and similarity functions, is known as virtual screening (VS). The goal of VS is to enrich molecules with similar properties (e.g., biological activity) to the query molecules and to discover new chemical entities in a small fraction of the database. To ensure the desired properties (e.g., biological activity) and to evaluate the success of the VS run, it is necessary to further analyze the enriched molecules by means of biological assays. The success of a VS run consists of two different aspects. First, the enriched molecules should have similar properties as the query molecules. Second, the discovery of new chemical entities that consist of different scaffolds in comparison with the query molecules, and, therefore represent an information gain. Based on the focus on a relevant subset of the database and the possible structural information gain, VS experiments represent a fundamental approach in the drug discovery pipeline<sup>23</sup>.

In the last two decades a plethora of different similarity functions were proposed<sup>45</sup>, and the development of new functions is still an open field of research. All similarity functions can be categorized by the dimension of the applied representation of molecules. 1D similarity functions are based on molecular property counts such as molecular weight or number of hydrogen bond acceptors. 2D approaches make use of the adjacency matrix of the molecular graph, and, therefore they are also called topological-based approaches. MOLPRINT2D<sup>6</sup>, substructure-based fingerprints like BCI<sup>7</sup> and DAYLIGHT<sup>8</sup> as well as the MACCS<sup>9</sup> keys are well known 2D similarity methods. Those topological or structural fingerprints yield promising results with respect to the enrichment of active molecules, but often lack the ability to discover new chemical entities<sup>10</sup>. 3D similarity functions are based on the shape<sup>11121314</sup> or geometrical distance information<sup>151617</sup> of molecules. Information of the conformational ensembles of molecules extends the 3D-based methods and can be seen as 4D approaches<sup>1819</sup>.

Based on the key-lock principle of Hermann Emil Fischer, it could be expected that the shape of molecules plays an important role for the biological activity. However, the shape of a molecule is not unique, but rather a function of internal parameters like the torsion angles. Hence, each rotatable bond represents a degree of freedom and increases the number of possible shapes (conformations) of the molecule. The resulting space, which contains all possible conformations, represents the conformational space of the molecule. Based on this increased complexity, it is not surprising that several literature studies reported a more robust VS performance of 2D methods in comparison to 3D approaches<sup>2021</sup>. Further arguments for 2D methods are their simplicity and speed<sup>22</sup>.

In a comprehensive study, Venkatraman et al.<sup>21</sup> investigated the performance of different 2D and 3D methods on a wide range of pharmaceutically relevant targets. The results of the study underpin the predominant opinion that 2D-based approaches are superior to 3D approaches with respect to the enrichment of active molecules. The performance

---

<sup>1</sup> Flexophore, a New Versatile 3D Pharmacophore Descriptor That Considers Molecular Flexibility

<sup>2</sup> Integration of virtual and high-throughput screening

<sup>3</sup> NOTITLE!

<sup>4</sup> Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation

<sup>5</sup> How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space

<sup>6</sup> Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance

<sup>7</sup> Chemical Fragment Generation and Clustering Software

<sup>8</sup> Daylight Chemical Information Systems Inc

<sup>9</sup> NOTITLE!

<sup>10</sup> Measuring CAMD Technique Performance: A Virtual Screening Case Study in the Design of Validation Experiments

<sup>11</sup> Application of 3D Zernike descriptors to shape-based ligand similarity searching

<sup>12</sup> A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape

<sup>13</sup> Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations

<sup>14</sup> ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential

<sup>15</sup> Pharmacophoric pattern matching in files of three-dimensional chemical structures: Comparison of conformational-searching algorithms for flexible searching

<sup>16</sup> Distance Profiles (DiP): A translationally and rotationally invariant 3D structure descriptor capturing steric properties of molecules

<sup>17</sup> Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology

<sup>18</sup> Treating Chemical Diversity in QSAR Analysis: Modeling Diverse HIV-1 Integrase Inhibitors Using 4D Fingerprints

<sup>19</sup> 4D-Fingerprints, Universal QSAR and QSPR Descriptors

<sup>20</sup> Unconventional 2D Shape Similarity Method Affords Comparable Enrichment as a 3D Shape Method in Virtual Screening Experiments

<sup>21</sup> Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods

<sup>22</sup> One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation

of the 2D and 3D approaches with respect to the knowledge gain by means of the discovery of new chemical entities was not evaluated by the study. A possible reason for the inferior performance of 3D methods is the geometric information that is based on one conformation of the molecule<sup>21</sup>. One opportunity to improve the performance of 3D methods is to apply the 3D methods on different conformations of the molecules and use the mean or maximum similarity value. The drawback of this workaround is the quadratic increase in computation time, which scales with the number of conformations. To address this runtime issue, it is necessary to perform the similarity calculation on the complete conformational ensemble in one step in a feasible manner. These limitations of 3D approaches also affect the performance of instance-based machine learning QSAR/QSPR models. To improve the robustness of those QSAR/QSPR models, we developed a 4D-based approach that is able to compare the conformational space of molecules within one step in feasible time<sup>23</sup>. The results showed that our approach produces robust models that are superior to similar 3D and 2D approaches. Given the fact that the reasons for the inferior performance of 3D-based methods seem to be similar in both applications (VS and QSAR/QSPR), it is possible that our 4D-based approach is also able to increase the VS performance in comparison to 2D and 3D methods.

The aim of this study is to evaluate our 4D approach as VS similarity function on a variety of literature VS benchmark data sets. Additionally, we compare the results to state-of-the-art 2D and 3D approaches to assess the performance of our method. We employed VS performance metrics that measure the chemotype enrichment performance to reduce the influence of artificial enrichment. The results show a robust performance of our approach in comparison to state-of-the-art 2D and 3D approaches. Therefore, our conformational space comparison is able to reduce the weakness of 3D-based methods without the time-demanding pair-wise comparison of individual conformations.

## 22.3 Methods

This section describes our 4D flexible atom-pair (4D FAP) similarity measure on the conformational space of molecules. To allow an efficient comparison of the conformational space of molecules, our approach needs a special encoding of the conformational ensembles, which can be seen as a preprocessing step. First, We describe our conformational space encoding. Afterwards, a modified Expectation Maximization (EM) algorithm will be presented that computes generative models, which represent the behavior of the molecules in their conformational space. Finally, the actual similarity calculation, which operates on the preprocessed molecules, will be explained.

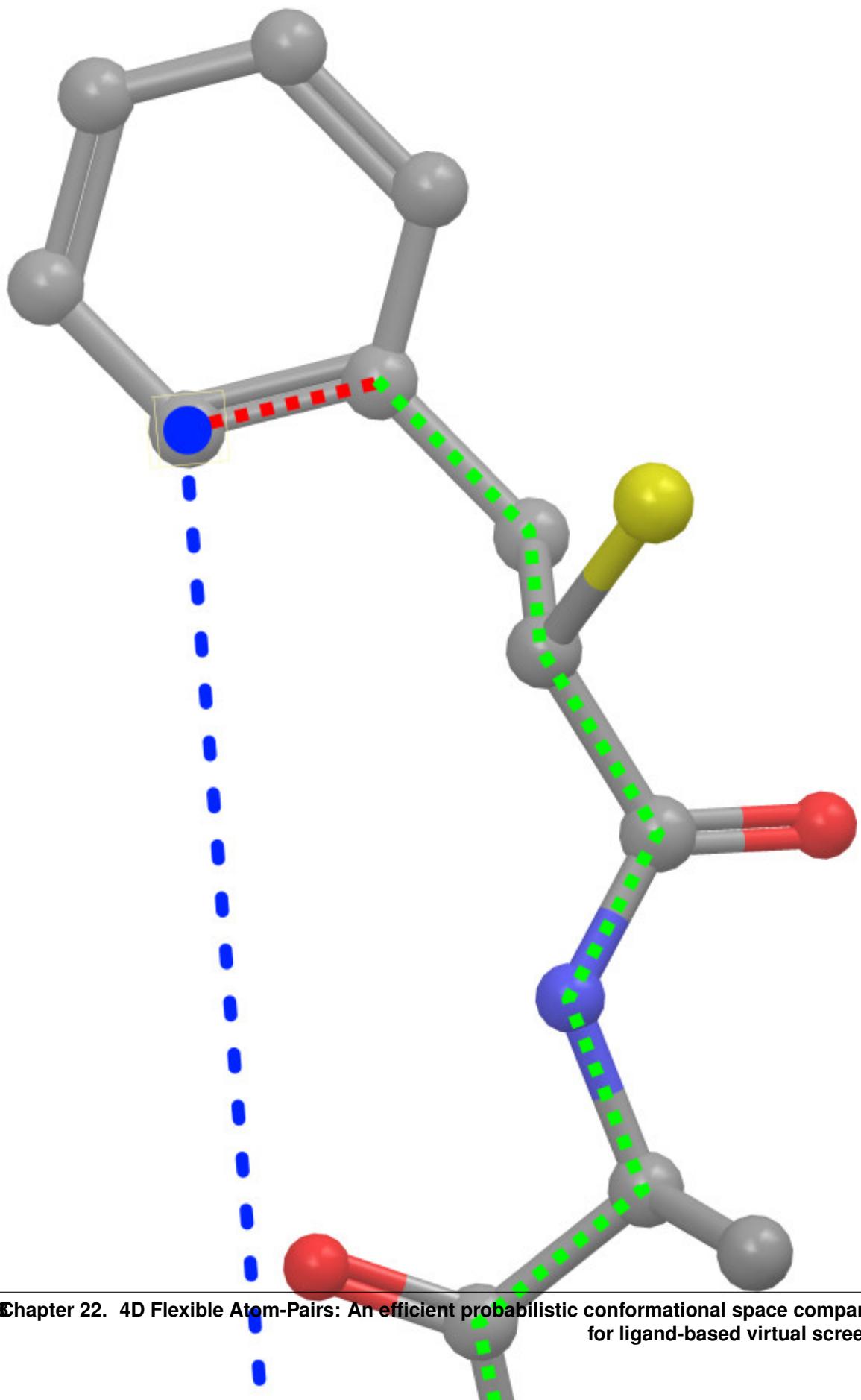
### 22.3.1 Conformational Space Encoding

To ensure a fast comparison of the conformational space of molecules, it is necessary to transform the complex information of the conformational space of molecules into a representation that is suitable for the integration into fast similarity functions. Therefore, we decompose the information of the conformational space into small portions. Given a conformational sampling  $M$  of molecule  $M$  with  $|M|$  heavy atoms, the encoding is based on the distance behavior of atom-pairs in the conformational space. Hence, the conformational space  $M$  of molecule  $M$  is segmented into the distance behavior of

Figure 1 represents exemplarily an atom-pair and the corresponding geometric distance. Not all of the  $M$  have a flexible distance behavior in the conformational space. The distance relation of neighboring atoms or atoms of a ring system only shows a small variability of the distance. Therefore, our encoding separates the atom-pairs into two disjoint classes: The flexible and the rigid atom-pair class. The separation is realized by a heuristic that employs the number of rotatable bonds in the shortest path of the corresponding molecular graph. Figure 1 visualizes the shortest path of the marked atom-pair. A bond is supposed to be rotatable if it is a single bond and not a bond of a ring system. If the number of rotatable bonds in the shortest path is 1, the atom-pair represents a flexible, otherwise a rigid atom-pair. Terminal rotatable bonds (rotatable bonds that are adjacent to one of the atoms that form the atom-pair) are not counted in the heuristic because a rotation of such a bond has no influence on the distance relation of the atoms (Figure 1).

Given the class of flexible atom-pairs from the heuristic above, our encoding measures the distance of each atom-pair and conformation of the given conformational sampling  $M$ . This results into atom-pairs that have  $\text{N} : \text{sub}: 'M' \text{ C}$

<sup>23</sup> Probabilistic Modeling of Conformational Space for 3D Machine Learning Approaches



distance values, where  $\text{N}_{\text{M}}$  represents the number of sampled conformers of molecule  $M$ . We refer to the atom-pairs containing the distance information in the conformational space as distance profiles.

These distance profiles can be visualized by means of normalized histograms, which represent the relative frequency of observing the corresponding atom-pair distance within a binned distance range. A histogram-based visualization of the distance profile from the atom-pair of Figure 1 can be seen in Figure 2. The application of histograms in a similarity function entails two major drawbacks. First, the binning size represents a parameter and has substantial impact on the resulting similarity value. Second, the storage of the information needs more space than a model-based encoding. Therefore, we decided to describe the distance behavior in the conformational space by means of Gaussian Mixture Models (GMMs). After the encoding of the distance profiles as GMMs the preprocessing of the molecules is finished.

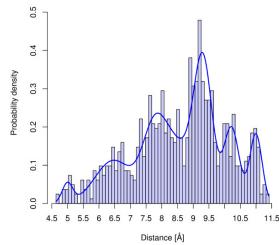


Figure 22.2: Figure 2. Atom-pair distance distribution

**Atom-pair distance distribution.** Histogram-based visualization of the distance distribution of the marked atom-pair of Figure 1. The line represents the corresponding GMM that models the distance behavior of the atom-pair in the conformational space.

### 22.3.2 Gaussian Mixture Models and Parameter Estimation on Distance Profiles

Mixture models are probabilistic models that represent a complex distribution based on a linear combination of individual sub-distributions. Applying Gaussian distributions as sub-distributions in a mixture model yields a GMM as given in Equation 1, where  $p^*(x)$  represents the probability density at the point  $x$ ,  $\pi_c$  determines the weight of the  $c$ -th Gaussian distribution, and  $\mu_c$  and covariance matrix  $\Sigma_c$ .

GMMs are generative models for real-world data and involve two advantages in our application. First, a conformational ensemble of a molecule represents only a discrete sampled approximation of the complete conformational space. Therefore, the flexible atom-pairs contain a series of sampled distance values. A generative model, fitted to the distance values, represents a continuous function, and, therefore describes a more realistic model in comparison to discrete or binned values. Second, the models can be efficiently stored because only the model parameters are necessary for a similarity calculation between such models. A drawback of the GMMs is the parameter estimation for a given data set. Given the distance values of a flexible atom-pair, it is necessary to fit the parameters  $\pi_c$ ,  $\mu_c$ ,  $\Sigma_c$ , and  $C$  (number of Gaussian components) of Equation 1 to the distance values. A popular approach to determine the parameters of a mixture model is the Expectation Maximization (EM) algorithm<sup>24</sup>. This algorithm is based on the maximum likelihood framework and optimizes the objective function given in Equation 2.

The EM algorithm represents an iterative process that consists of two steps. The first step (E-step) evaluates the responsibilities that the  $k$ -th component of the GMM was responsible for generating the  $n$ -th ( $x_n$ ) data point of the given data set  $X$  (Equation 3).

The second step (M-step) updates the parameters of the GMM on the basis of the responsibilities of the previous E-step (Equation 4-7).

These two steps are repeated until a predefined convergence criterion is reached. The EM algorithm optimizes the parameters of the GMM and guarantees a local optimum solution. Therefore, it is necessary to execute the EM algorithm with different initial parameters to avoid a model from a local optimum with an inferior likelihood value.

<sup>24</sup> Maximum Likelihood from Incomplete Data via the EM Algorithm

The EM algorithm estimates the parameters of a GMM based on a predefined number of Gaussian components. A suitable number of components is crucial for a useful model. Therefore, a model selection step that determines an optimal number of components is necessary. To reduce the risk of overfitting (high number of Gaussian components), several model selection criterions, such as the Bayesian information criterion<sup>25</sup> or the Akaike information criterion<sup>26</sup>, were proposed that penalize an increased number of components. This model selection step involves a significant runtime overhead and can be avoided if the number of sub-distributions can be estimated. In our application, a GMM has to model the distance behavior of the corresponding atom-pair in the conformational space. The distance behavior of an atom-pair can be seen as a function of the flexibility of the shortest path in the molecular graph.

Therefore, the number of flexible bonds in the shortest path (as applied to classify the atom-pairs) can also be applied as a heuristic to determine the number of Gaussian components for the GMM. In an earlier study we already presented the comparable performance of the heuristic in comparison to model selection criterions<sup>23</sup>. This heuristic avoids the model selection step and reduces the runtime of the preprocessing step. Figure 2 presents the GMM that models the distance behavior of the atom-pair in Figure 1. The presented EM algorithm assumes that all samples of the data set are equally important for the final model. Transferred to our application this means that each conformation has the same influence on the final model. Based on a thermodynamic point of view, this assumption of equal influence holds if all conformations of the ensemble have the same energy. To emphasize the influence of low energy conformations on the final model, we developed an extension of the EM algorithm that integrates the importance of each sample into the optimization process. In an earlier study, our modified EM algorithm generated improved QSAR/QSPR models in comparison to models based on equally weighted GMMs<sup>27</sup>.

### 22.3.3 Boltzmann Weighted Expectation Maximization Algorithm

To increase the importance of low energy conformations on the final GMMs of a molecule, we apply the normalized Boltzmann distribution as given in Equation 8 to determine a probability value for a given conformer.  $\Delta E_n$  represents the energy offset of the n-th conformer to the global optimum of the conformational ensemble, R presents the gas constant, and T the temperature of the canonical ensemble.

These probability values have to be integrated into the EM algorithm and modify the objective function as outlined in Equation 9, where  $E$  symbolizes a vector containing the energy values of the conformers and  $p^*(E_n)$  depicts the probability of the n-th conformation.

The E-step (computation of the responsibilities) remains unchanged, and, therefore the responsibilities are calculated as stated in the Equation 3. However, the equations of the M-step (update of the parameters) need the integration of the probability values as listed in the Equations 10-13.

Based on the described modifications, the EM algorithm computes GMMs that represent the distance behavior of atom-pairs as a function of the frequency of observing an atom-pair at a certain distance as well as the probability that the canonical ensemble will occupy these states (conformations). Figure 3 visualizes a weighted (probabilities of the conformations) histogram-based representation of the same atom-pair visualized in Figure 2. The Boltzmann weighted model of Figure 3 shows that the distances at 9.25 Å and 11 Å are energetically favorable, and, therefore the probability density is increased in comparison to the unweighted model of Figure 2. In contrast, the conformations with low range distances have higher energy values and, as a consequence, the corresponding probability densities are reduced.

### 22.3.4 4D Flexible Atom-Pair Similarity Function

After the preprocessing of the molecules (encoding the distance distributions by means of GMMs) the actual 4D similarity calculation can be conducted. The similarity function operates on the molecular graph (adjacency matrix) and the GMMs of the flexible atom-pairs. Therefore, the conformational ensemble of the molecules is not further needed.

<sup>25</sup> Estimating the dimension of a model

<sup>26</sup> A new look at the statistical model identification

<sup>27</sup> Boltzmann-Enhanced Flexible Atom-Pair Kernel with Dynamic Dimension Reduction

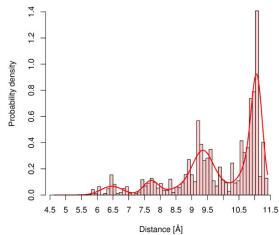


Figure 22.3: Figure 3. Boltzmann weighted atom-pair distance distribution

**Boltzmann weighted atom-pair distance distribution.** Boltzmann weighted histogram-based visualization of the distance distribution of the atom-pair of Figure 1. The line describes the probability density of the GMM that was computed by the Boltzmann weighted EM algorithm.

In a first step, the 4D FAP creates for each heavy atom of the molecule an atom-pair prefix tree. This data structure represents an efficient approach for search and comparison operations and was already applied as a data structure for an atom-pair-based similarity measure<sup>2829</sup>. Each prefix tree has one atom as root node and contains all atom-pair information to the remaining  $|M| - 1$  heavy atoms of the molecule (leaves of the tree). The preprocessing step divides the atom-pairs into two disjoint classes. Therefore, an atom-pair tree  $T$  contains the two different sub-trees  $R$  and  $F$  for the rigid and flexible atom-pair class, respectively. The rigid sub-tree  $R$  contains the information of the rigid atom-pairs that were not modeled by GMMs. To increase the information content of the rigid atom-pairs, the sub-tree additionally contains the topological distance information of each atom-pair. An example of such an atom-pair tree can be seen in Figure 4.

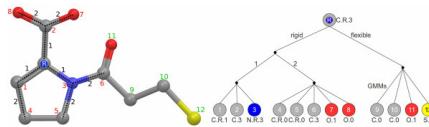


Figure 22.4: Figure 4. Flexible atom-pair tree

**Flexible atom-pair tree.** The left molecule represents the example molecule for the tree on the right side. The white ‘R’ marks the atom that serves as root atom (point of origin for the atom-pairs) for the tree. The black numbers symbolize the topological distance for the rigid atom-pairs. The red and green numbers correspond with the leaf numbers of the tree on the right side. The red or green color of these atom numbers indicates the membership of the atom to the rigid or flexible sub-tree, respectively.

The nodes in these prefix trees can be labeled by any arbitrary atom typing scheme. We applied a labeling function that consists of three different elements. The first element is the element symbol of the atom. A ring flag indicates the membership to a ring system. The final value is the result of the number of neighboring heavy atoms minus the number of neighboring hydrogen atoms.

Given the prefix trees of two molecules  $A$  and  $B$ , the 4D FAP computes a similarity matrix  $S$ , where each entry represents the similarity value between two atom-pair trees. Based on the two different sub-trees, an entry  $S_{ij}$  in the similarity matrix is the sum of two distinct similarity calculations

The rigid sub-tree contains the labels of the atoms and a topological distance value. This type of information represents nominal features and enables the use of simple similarity functions. We applied the Tanimoto similarity function as stated in Equation 14.  $R_i$  and  $R_j$  represent the rigid atom-pair sub-trees of the  $i$ -th and  $j$ -th atom, respectively.

The comparison of the flexible atom-pair sub-tree consists of two different similarity functions. Given the flexible atom-pair sub-trees  $F_i$  and  $F_j$  containing  $|F_i|$  and  $|F_j|$  flexible atom-pairs (number of leaves in the sub-tree). The first function compares the atom labels of two given flexible atom-pairs  $AP_n$  and  $AP_m$  as presented in Equation 15.  $n$  and  $m$  represent the atom labels of the atom-pairs. The function is a Dirac function on the atom labels and returns a value of 1.0 if the labels are equal and 0 otherwise.

<sup>28</sup> Optimal assignment methods for ligand-based virtual screening

<sup>29</sup> Graph kernels for chemical compounds using topological and three-dimensional local atom pair environments

The second similarity function compares the behavior of the atom-pairs in the conformational space. For this purpose, a correlation measure on GMMs is applied as denoted in Equation 16, where  $m$  and  $n$  symbolize the GMMs of the flexible atom-pair  $AP_m$  and  $AP_n$ , respectively.

The assembly of both similarity functions for flexible atom-pairs results in Equation 17 and represents the similarity function for the flexible atom-pair sub-tree.

Unlike the similarity function for the rigid atom-pair sub-tree, the similarity function for the flexible atom-pair sub-tree is not based on nominal features. Therefore, Equation 17 performs a pair-wise comparison of all atom-pairs and sums up the individual similarity scores. To avoid overestimated similarity values of sub-trees with an increased number of flexible atom-pairs, the similarity value is normalized by Equation 18

After computation of the similarity matrix  $S$ , which contains all pair-wise similarity values of the atom-pair trees, the 4D FAP computes a final similarity value based on the matrix  $S$ . The original 4D FAP, as applied in QSAR/QSPR studies<sup>2327</sup>, sums up the entries of the  $S$  matrix and normalizes the sum to obtain a value in the range [0.0, 1.0]. Another possibility to compute a final similarity value represents the optimal assignment. This approach was introduced into the field of cheminformatics by Fröhlich et al.<sup>3031</sup> and applied as a VS similarity function in a previous study<sup>28</sup>.

Preliminary experiments (not published) showed that an optimal assignment on the matrix  $S$  improves the VS performance in comparison to the normalized summation of the matrix elements. Therefore, we changed the final computation step of the 4D FAP to perform an optimal assignment on the matrix  $S$  as stated in Equation 19. Given the molecules  $A$  and  $B$  (with  $|A| < |B|$  w.l.o.g.),  $\text{Inonascii\_14|}$  represents a function that maps each value of  $i$  [1, ...,  $|A|$ ] on a value in the range [1, ...,  $|B|$ ] in such a way that the sum of the similarity entries is maximized. The final sum of the optimal assigned similarity values of the atom-pair trees is also normalized by Equation 18. We refer to the optimal assignment-based variant of the 4D FAP as 4D FAP<sub>OA</sub>.

## 22.4 Experimental

In this section we initially characterize the applied VS benchmark data sets as well as their preparation step. Afterwards, the protocol for the conformational sampling of the molecules as well as a short description of the VS evaluation metrics follow. Finally, we present a brief overview of literature VS methods that were applied to classify the results of the 4D FAP<sub>OA</sub>.

### 22.4.1 Data sets

To evaluate the 4D FAP<sub>OA</sub> on a wide range of pharmaceutically relevant targets, we employed the directory of useful decoys (DUDs) release 2<sup>32</sup>. These data sets were introduced as a benchmark data set compilation for the evaluation of docking algorithms<sup>33</sup>. Ligand-based VS, based on similarity values to a query structure, can be afflicted with an analogue enrichment bias. This bias results from the enrichment of structurally similar molecules with respect to the query structure. These similar structures represent only a limited information gain, and, therefore the results of the experiment will have an analogue enrichment bias.

To reduce this bias and to enable a fair comparison between similarity-based and docking-based algorithms, Good and Oprea applied a lead-like filter<sup>34</sup> on the data sets and clustered the actives<sup>35</sup>. These filtered and clustered data sets were already applied in a ligand-based VS study<sup>28</sup> and are publicly available<sup>36</sup>. Table 1 shows a complete overview of the 40 targets and the number of actives and decoys for the DUD release 2 and the filtered variant. For the VS experiments we applied the target ligands as query structure for the respective active and decoy data set. The data sets were not further modified to allow a fair comparison of the results.

<sup>30</sup> Optimal assignment kernels for attributed molecular graphs

<sup>31</sup> Kernel Functions for Attributed Molecular Graphs - A New Similarity-Based Approach to ADME Prediction in Classification and Regression

<sup>32</sup> DUD - A Directory of Useful Decoys

<sup>33</sup> Benchmarking Sets for Molecular Docking

<sup>34</sup> Is There a Difference between Leads and Drugs? A Historical Perspective

<sup>35</sup> Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: a Help or Hindrance in Tool Selection?

<sup>36</sup> DUD Filtered - Lead-like filtered DUD

The evaluation of similarity functions by means of the DUD data sets represents a retrospective evaluation. Analogous to the “Kubinyi paradox”<sup>37</sup> of QSAR models, the solely retrospective evaluation possibly implies the risk that the development of new methods or the improvement of existing approaches will increase their retrospective performance at the expense of the prospective performance. However, the DUD data sets contain over 100,000 molecules for 40 different targets. Consequently, the evaluation on all 40 data sets is based on an increased molecular diversity in comparison with the usually smaller and less diverse benchmark data sets of QSAR experiments. Therefore, the risk of an inferior prospective performance of VS similarity functions as a result of their optimization for the retrospective performance is reduced but still present.

## 22.4.2 Conformational sampling

To create the conformational ensembles of the molecules, we applied the ConfGen tool of Schrödinger<sup>38</sup>. Recent studies showed the ability of ConfGen to compute reasonable conformers of molecules<sup>3940</sup>. The tool provides four standard parameter schemes that sample the conformational space at different resolutions. To compute useful GMMs in the preprocessing step, it is necessary to sample the conformational space at a high resolution. Therefore, we modified the ‘comprehensive’ parameter scheme of ConfGen to further increase the resolution. We reduced the heavy atom rmsd for distinct conformers from 0.5 Å to 0.1 Å. This modification results in more conformers but does not increase the runtime of the conformational sampling. The energy values, which are necessary for the Boltzmann weighted GMMs, were computed by the OPLS 2005 force-field with standard parameters.

The applied conformational sampling algorithm as well as the force-field model have a major impact on the final results of the 4D FAP<sub>OA</sub>. Different conformational sampling algorithms compute different sets of conformers, which in turn yield different atom-pair distance profiles. The force-field computes an energy value for each conformer and determines the weight of each measured atom-pair distance. As a result, a different conformational sampling protocol will yield different GMMs of the atom-pairs. Hence, the computed similarity values differ and will probably change the results. However, the aim of this study is not the evaluation of the impact of different conformational sampling protocols on the 4D FAP<sub>OA</sub>, but the evaluation of the 4D FAP<sub>OA</sub> as a VS similarity function based on the given protocol.

## 22.4.3 Evaluation metrics

To evaluate the performance of our 4D FAP<sub>OA</sub> approach, we applied different standard evaluation metrics. The receiver operating characteristic (ROC) curve represents a function that plots the true positive rate as a function of the false positive rate. The area under the ROC curve (AUC) represents a quantification of the curve and facilitates an easier comparison of results. The AUC is calculated as given in Equation 20, where  $N_{\text{actives}}$  depicts the number of actives,  $N_{\text{decoys}}$  represents the number of decoys, and  $i$ -th active structure. The received value is in the range [0.0, 1.0], where 0.5 indicates a random performance.

The AUC metric represents a measure that evaluates the performance on the complete data. However, a major goal of VS experiments is the enrichment of active structures in a small fraction of the database. Therefore, it is necessary to apply additional metrics that focus on the early enrichment behavior. A common metric for the early enrichment problem is the enrichment factor at a predefined fraction of the data set (\*x\*) as given in Equation 21.

The enrichment factor depends on the number of actives, and, therefore it is not a robust metric. Korff et al.<sup>41</sup> proposed the relative enrichment factor (REF) as stated in Equation 22 to remove the dependency on the number of actives structures.

The enrichment of active structures that are based on different scaffolds emerged to an additional important goal of VS experiments. All metrics that evaluate the so-called chemotype enrichment are based on a clustering of the

<sup>37</sup> Pharmacophore Discovery - Lessons Learned

<sup>38</sup> NOTITLE!

<sup>39</sup> ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers

<sup>40</sup> Drug-like Bioactive Structures and Conformational Coverage with the LigPrep/ConfGen Suite: Comparison to Programs MOE and Catalyst

<sup>41</sup> Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set

active structures into different chemotypes (scaffolds). Mackey and Melville showed the integration of the scaffold information into common VS metrics<sup>42</sup>. We decided to apply the arithmetic weighting on the ROC enrichment as given in Equation 23. Based on the information of the clustering, a structure  $i$  obtains a weight that is inversely proportional to the number of structures ( $jN$ ) in the cluster  $i$ -th active structure of the  $j$ -th cluster is contained in the first  $*x\%$  of the data set.

The evaluation metrics listed above represent only a small fraction of possible metrics. Other popular metrics for the early enrichment evaluation are the BEDROC<sup>43</sup> score or the enrichment factor. To enable future comparisons with the presented results of the 4D FAP<sub>OA</sub>, we computed for each target of the filtered DUD data set a result file that contains several additional VS metrics (e.g., BEDROC score at nine predefined alpha values). Additionally, the files contain the complete ranking of the molecules that allows the computation of the VS metric of choice. The 40 result files are contained in the additional file

Additional file 1

Archive of the 4D FAP<sub>OA</sub> result files. This is a Gzip compressed Tar archive containing the result files of the 4D FAP<sub>OA</sub> on the filtered DUD data sets. The result files are tab-separated plain text files including the following information: method name, active data set with size, cluster information and the distribution of the active molecules over the clusters, decoy data set with size, ratio active:decoy, AUC, awAUC, BEDROC scores for predefined alpha values as suggested by Truchon and Bayly<sup>43</sup>, enrichment factors, relative enrichment factor<sup>41</sup>, ROC enrichments, awROC enrichments at predefined false positive fractions, chemotype enrichment, ROC and awROC data points, and the ranking of each structure to compute other VS metrics.

[Click here for file](#)

#### 22.4.4 Literature Similarity Functions

We employed a wide range of different 2D and 3D similarity approaches to assess the performance of the 4D FAP<sub>OA</sub>. Due to the fact that we compare our approach to 20 other approaches, we only mention the name of the method and the applied type of information. For a comprehensive description we refer to the original publications.

Different optimal assignment approaches were already evaluated on the filtered DUD data sets in an earlier publication<sup>28</sup>. The best approach of this study was a two-step hierarchical assignment (2SHA) that first operates on a substructure level and afterwards on the atomic level. A second approach of that study optimally assigns the atom-pair (OAAP) environment trees and represents a similar 3D concept in comparison to the 4D FAP<sub>OA</sub>. The optimal assignment kernel (OAK)<sup>3031</sup> and its flexibility extension, the OAK<sub>FLEX</sub><sup>44</sup>, were also evaluated in this earlier publication.

Cheeseright et al.<sup>45</sup> introduced FieldScreen as a multiconformer-based VS tool. FieldScreen utilizes a database that contains conformers of each molecule. Therefore, it operates on a conformational ensemble in a similar way as the 4D FAP<sub>OA</sub> and represents an interesting reference approach. FieldScreen employs four different types of locally optimized molecular field points to compute a similarity value between two given molecules.

Venkatraman et al. conducted a comparison study in which a plethora of different 2D and 3D approaches were evaluated on the original as well as the filtered DUD data sets<sup>21</sup>. We compared the performance of the 4D FAP<sub>OA</sub> to the main results of this study. The study conducted by Venkatraman et al. employed the 2D fingerprint methods: OPENBABEL<sup>46</sup>, DAYLIGHT<sup>8</sup>, BCI<sup>7</sup>, MACCS<sup>9</sup>, and MOLPRINT2D<sup>6</sup>. As 3D-based approaches they utilized ROCS<sup>12</sup> with two different scoring schemes ROCSS (shape only) and ROCS<sub>SC</sub> (shape and chemistry). The EON<sup>47</sup> approach compares the electrostatic fields computed by the Poisson-Boltzmann equation and was also evaluated using two different parameterizations. EON<sub>SE</sub> is based on the shape and the electrostatic, whereas EON<sub>SCE</sub> additionally uses chemical information.

<sup>42</sup> Better than Random? The Chemotype Enrichment Problem

<sup>43</sup> Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem

<sup>44</sup> Atomic Local Neighborhood Flexibility Incorporation into a Structured Similarity Measure for QSAR

<sup>45</sup> FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set

<sup>46</sup> The Open Babel Package

<sup>47</sup> The Use of Three-Dimensional Shape and Electrostatic Similarity Searching in the Identification of a Melanin-Concentrating Hormone Receptor 1 Antagonist

SHAEP<sup>14</sup> is based on a maximum common subgraph approach that is employed to perform a superposition of the molecules. The method operates only on the shape of molecules ( $\text{SHAPE}_S$ ) or on the shape and the electrostatic ( $\text{SHAPE}_{\text{SE}}$ ). The Ultrafast Shape Recognition (USR)<sup>17</sup> employs four distance relations of each atom and computes the first three moments of each distribution to obtain 12 descriptor values for each molecule. ESHAPE3D is based on a heavy atom distance matrix that is employed to compute fingerprints. The ESHAPE<sub>HYD</sub> alternatively uses the hydrophobic heavy atoms.

PARAFIT<sup>13</sup> computes a similarity value based on spherical harmonic expansions of molecular surfaces. Another important class of similarity functions are the pharmacophore-based approaches. These approaches operate on an abstract representation of the molecules by means of pharmacophore features. These features are divided into different classes (e.g., hydrogen bond acceptor, aromatic, or hydrophobic) and represent important interaction points of molecules. The distance relation between these pharmacophore features plays an important role and can be measured in a topological or geometrical manner. Therefore, the pharmacophore-based approaches can also be divided into 2D- and 3D-based approaches. Korff et al.<sup>41</sup> compared different structure- and ligand-based VS approaches on the DUD data sets. This study contains two different pharmacophore-based methods. The topological pharmacophore point histogram (TopPPHist) computes for each pair of pharmacophore classes a distance histogram based on the topological distances. Therefore, the TopPPHist represents a 2D-based pharmacophore approach. Finally, the distance histograms are converted into a descriptor vector. The Flexophore approach<sup>141</sup> computes geometrical and binned distance histograms for each pharmacophore point pair based on a representative set of given conformers. The final comparison between two molecules is similar to the maximum common subgraph-isomorphism because the pharmacophore points together with the distance histograms form complete graphs.

## 22.5 Results and Discussion

The results section is divided into four different subsections. The first subsection compares the results of the 4D FAP<sub>OA</sub> approach with other optimal assignment-based approaches that were already evaluated on the DUD data sets<sup>28</sup>. The second part is based on the results of Venkatraman et al.<sup>21</sup> and compares the average performance of the 4D FAP<sub>OA</sub> with 15 state-of-the-art 2D and 3D approaches. Afterwards, a comparison with the pharmacophore-based approaches of Korff et al.<sup>41</sup> follows. The final subsection focuses on the performance difference between 3D approaches on multiple conformers and our 4D FAP<sub>OA</sub> approach.

### 22.5.1 Comparison with other Optimal Assignment Methods

The comparison with other optimal assignment methods measures the influence of the applied information type on the final performance. The OAAP represents the comparable 3D approach in comparison with the 4D FAP<sub>OA</sub>, and, therefore directly measures the performance gain of the 4D extension. As an early enrichment metric we applied the awROCE<sub>5%</sub>, which also assesses the chemotype enrichment performance. To reduce the bias introduced by a low number of chemotypes, we only applied data sets that have at least 15 different chemotypes. The AUC value was applied to evaluate the performance on the complete data sets.

Table 2 shows the results of the four optimal assignment methods and the 4D FAP<sub>OA</sub>. The direct comparison of the OAAP and the 4D FAP<sub>OA</sub> indicates that the 4D FAP<sub>OA</sub> outperforms the OAAP on 10 out of 13 data sets with respect to both performance measures. The OAAP is superior to the 4D FAP<sub>OA</sub> on the COX2, HIVRT, and the PDGFrB data sets. These three data sets are more rigid data sets with respect to the number of rotatable bonds of the query compounds. The query compounds of COX2, HIVRT, and PDGFrB have 5, 9, and 7 rotatable bonds, respectively. In comparison, the most flexible data sets are the ACE and EGFr data sets with 18 and 14 rotatable bonds, respectively. The correlation between the performance gain on the AUC metric  $n$  is the number of samples (13 data sets) and  $r$  the correlation, the probability that both variables (AUC performance gain and flexibility of query compounds) result in such a correlation if there is no true correlation of the variables ( $|nonascii\_19|* = 0.0$  is 0.0277. *Therefore, the correlation is significant* ( $*p = 0.05$ ) and indicates that the performance gain of the 4D FAP<sub>OA</sub> is a function of the flexibility of the data set. In comparison to the other optimal assignment methods there is no correlation apparent. However, the OAK, OAK<sub>FLEX</sub>,

and 2SHA are based on a different type of information (local atom similarity based on atom and bond features), and therefore, a direct comparison of the correlations is not meaningful.

The comparisons of the 4D FAP<sub>OA</sub> with all other optimal assignment approaches show that the 4D FAP<sub>OA</sub> outperforms all other methods on 6 and 9 data sets with respect to the awROCE<sub>5%</sub> and AUC, respectively. These results yield a best average rank of 1.58 for the 4D FAP<sub>OA</sub> with respect to the AUC. For the awROCE<sub>5%</sub> results the 2SHA achieves the best average rank of 2.15 followed by the 4D FAP<sub>OA</sub> with an average rank of 2.31. In a direct comparison the 4D FAP<sub>OA</sub> outperforms the 2SHA approach on 7 data sets, whereas the reverse case only occurs on 5 data sets. To conclude, the 4D FAP<sub>OA</sub> shows a robust performance on 13 data sets. Considering the results of the complete data sets (AUC) the 4D FAP<sub>OA</sub> outperforms all other optimal assignment methods. The ability of 4D FAP<sub>OA</sub> to early enrich different scaffolds is comparable with the 2SHA approach.

The encoding of the conformational space should be most beneficial if the flexibility of the query structure is high. Therefore, we discuss the results on the two data sets with the most flexible query compounds, the ACE and EGFr data set, in more detail.

Figure 5 shows the ROC plot of all optimal assignment methods on the ACE data set. The curve of the 4D FAP<sub>OA</sub> passes always above the other curves with the exception of the 2SHA curve between 0.3 and 0.4 false positive rate. In the early enrichment range (0.0 - 0.1 false positive rate) the 4D FAP<sub>OA</sub> shows a strong increase of the true positive rate without any longer phases of stagnation (horizontal elements in the curve). The other optimal assignment methods show a similar behavior till 0.03 false positive rate, but they stagnate until 0.1 false positive rate. Therefore, the 4D FAP<sub>OA</sub> has an offset of the true positive rate of nearly 0.2 in comparison to the other methods. The 4D FAP<sub>OA</sub> is also the first approach that is able to retrieve all actives of the data set (0.7 false positive rate). The second approach that retrieves all actives is the 2SHA approach at a false positive rate of 0.9. The comparable 3D approach (OAAP) is always inferior in comparison to the 4D FAP<sub>OA</sub>.

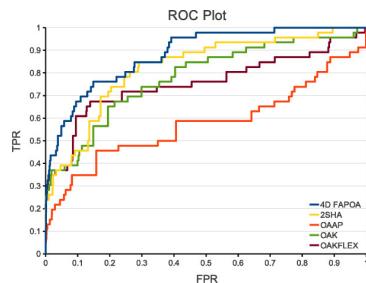


Figure 22.5: Figure 5. ROC plot on ACE

**ROC plot on ACE.** ROC plot of all optimal assignment methods on the filtered ACE data set. TPR and FPR denote the true positive rate and false positive rate, respectively.

To evaluate the chemotype discovery on the complete data set, we plotted the fraction of the discovered chemotypes as a function of the fraction of the ranked data set. A chemotype is considered as discovered if one compound of the chemotype is ranked.

Figure 6 presents the chemotype discovery of all optimal assignment approaches on the ACE data set. The curves of the 4D FAP<sub>OA</sub>, 2SHA, OAK, and OAK<sub>FLEX</sub> show a similar behavior over the complete data set. Only the OAAP has an inferior chemotype discovery rate until 40% of the data set. Therefore, the information gain by discovering new chemotypes is increased by the 4D FAP<sub>OA</sub> in comparison to its similar 3D method (OAAP). However, the other approaches that are based on a different type of information (OAK, OAK<sub>FLEX</sub>, and 2SHA) show a similar discovery rate.

The ROC plots and the chemotype discovery on the EGFr data set can be seen in the Figures 7 and 8, respectively. In both figures a considerably performance gain of the 4D FAP<sub>OA</sub> is apparent. The 4D FAP<sub>OA</sub> is able to retrieve all actives within 30% of the data set (Figure 7). All chemotypes were discovered within 23% of the data set (Figure 8). All other optimal assignment methods retrieve at least 20% of the actives and at least 15% of the chemotypes within the last percent of the data set. In comparison to the OAAP the performance gain of the 4D FAP<sub>OA</sub> is approximately

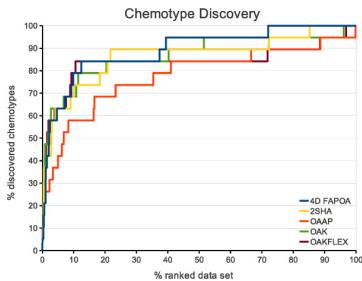


Figure 22.6: Figure 6. Chemotype discovery on ACE

**Chemotype discovery on ACE.** Chemotype discovery of all optimal assignment methods on the filtered ACE data set.

twice that of the OAK, OAK<sub>FLEX</sub>, and 2SMA. Therefore, our encoding of the conformational space entails a significant performance gain on the EGFr data set.

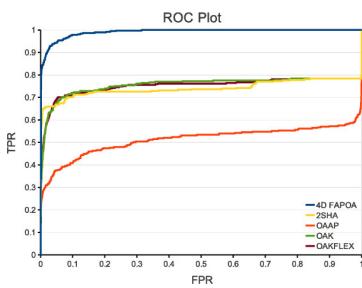


Figure 22.7: Figure 7. ROC plot on EGFr

**ROC plot on EGFr.** ROC plot of all optimal assignment methods on the filtered EGFr data set.

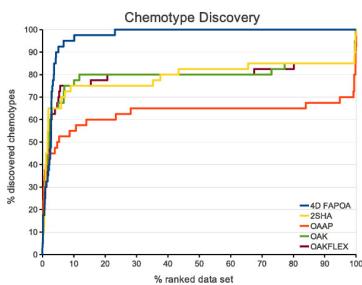


Figure 22.8: Figure 8. Chemotype discovery on EGFr

**Chemotype discovery on EGFr.** Chemotype discovery of all optimal assignment methods on the filtered EGFr data set.

Another important property of a VS similarity function is the computation performance. To enable a VS experiment on a real-world database, the VS similarity function should be able to process a reasonable number of compounds in a feasible time. All presented VS similarity functions that are based on the optimal assignment approach were developed at our department, and, therefore we are able to perform a fair comparison of the computation time. We computed the average computation time of each optimal assignment method on the 13 data sets, which were used in Table 2, to approximate a reasonable performance for drug-like compounds.

The 4D FAP<sub>OA</sub> approach has an averaged performance of  $38.8 \pm 27.56$  similarity calculations per second. This computation time is based on preprocessed molecules (GMMs already computed). The OAK yields  $27.34 \pm 3.40$  calculations per second, whereas its flexibility extension (OAK<sub>FLEX</sub>) computes  $41.03 \pm 7.32$  molecules per second. The OAAP represents the fastest approach with  $51.49 \pm 18.07$  computations per second. In contrast, the 2SMA is the slowest method with a throughput of  $14.04 \pm 1.78$  per second. All calculations were done on a Core2Duo CPU with

2 GHz using one core and 1 GB memory. As a result, the 4D FAP<sub>OA</sub> is fast enough to screen over 100,000 molecules within one hour on a desktop CPU using only one core. The similarity calculation can be easily parallelized to further increase the throughput, and, therefore the approach should be fast enough for real-world applications.

The preprocessing step (conformational sampling and GMM calculation) represents an additional computational task of our approach. However, the preprocessing step has only to be computed once for each molecule. Additionally, the computation of different conformers (conformational sampling) is often necessary for different tasks in the drug discovery pipeline. Furthermore, our encoding is a model-based encoding that reduces the memory usage in a database in comparison to the storage of multiple conformers of a molecule.

## 22.5.2 Comparison with State-of-the-Art 2D and 3D Approaches

In this subsection we compare the performance of the 4D FAP<sub>OA</sub> with different state-of-the-art 2D and 3D approaches. Venkatraman et al.<sup>21</sup> conducted a comprehensive evaluation of 15 different literature methods on the DUD data sets. The study contains the averaged (over all 40 data sets) relative enrichment factors at 1%, 5%, and 10% as well as the AUC values for each method. Unfortunately, the study lacks any evaluation metric that rates the chemotype discovery of the approaches. Therefore, the results in this section are only based on the early enrichment (relative enrichment factors) and the performance on the complete data set (AUC). All results are based on the filtered data sets and compiled in Table 3.

The results of Table 3 confirm that the 2D approaches are more robust in comparison to the 3D methods. Only the ROCS<sub>SC</sub> is able to yield comparable results in comparison to the MACCS keys and MOLPRINT2D. The 4D FAP<sub>OA</sub> is able to utilize the GMMs as a source of reasonable information, and, therefore the approach yields the best results with respect to the relative enrichment factor at 5% and 10% as well as the AUC metric. Only the BCI approach is able to marginally improve the results with respect to the relative enrichment factor at 1%. The best performance of the 4D FAP<sub>OA</sub> on three out of four metrics results in the best average rank of 1.25. The BCI and DAYLIGHT fingerprints yield an average rank of 2.75 and represent the best 2D-based approach. ROCS<sub>SC</sub> is the best 3D-based approach with an average rank of 5.0, and, therefore higher ranked as the 2D-based approaches MOLPRINT2D (6.0) and the MACCS keys (7.0). All other 3D-based methods are inferior in comparison to the 2D-based approaches. To conclude, the 4D FAP<sub>OA</sub> benefits from the conformational space information and is able to yield the best average performance of all methods.

## 22.5.3 Comparison with Pharmacophore-Based Approaches

Korff et al.<sup>41</sup> evaluated the TopPPHist and the Flexophore approach on the 40 targets of the DUD data sets. The early enrichment performance was assessed by the relative enrichment factor at 1% of the data set. To evaluate the chemotype enrichment, Korff et al. counted the discovered chemotypes within the enriched data set fraction with respect to the chemotype definition of Good and Oprea<sup>35</sup>. Table 4 lists the relative enrichment factors and the number of discovered chemotypes for each of the 40 data sets of the DUD.

With respect to the early enrichment performance the TopPPHist and the Flexophore approach achieved an average relative enrichment factor of  $37.34 \pm 31.38$  and  $43.31 \pm 33.25$ , respectively. The application of the 4D FAP<sub>OA</sub> resulted in an average relative enrichment factor of  $55.45 \pm 33.26$  and increased the performance of the Flexophore approach by over 20%. However, based on their abstract representation of molecules, one of the strengths of pharmacophore-based approaches is the ability to discover new chemical entities. This abstraction from the query scaffold can be seen in the chemotype discovery results of Table 4. The Flexophore approach needs 20% less active compounds to discover a similar amount of chemotypes (94) in comparison with the 4D FAPOA (98). The 2D-based TopPPHist discovered only 66 chemotypes over all 40 data sets and showed an inferior chemotype discovery in comparison with the 4D-based approaches (Flexophore, 4D FAP<sub>OA</sub>).

## 22.5.4 Comparison with Multiple Conformer Approaches

The results of the previous sections demonstrated the inferior performance of 3D-based approaches in comparison with 2D-based methods. A common technique to tackle this deficit of 3D approaches is to utilize multiple conformers and average or use the maximum of all pair-wise similarity values. The number of necessary similarity computations scales with  $O^*(n^2)$ , where  $n$  represents the number of conformers of the molecules. Therefore, this technique implies a significant increase in computation time. However, the averaging over multiple conformers increases the available information content of the 3D-based approaches to a level that is similar in comparison to the 4D FAP<sub>OA</sub>. The 4D FAP<sub>OA</sub> has a model-based description of the conformational space, whereas the 3D-based approaches explicitly have the conformational space. Consequently, a comparison of the 4D FAP<sub>OA</sub> with 3D-based approaches on multiple conformers represents an interesting comparison based on an equal source of information.

Venkatraman et al.<sup>21</sup> evaluated the ROCS<sub>SC</sub> (best 3D-based approach of Table 3) in three additional experiments on the unfiltered DUD data sets with different ensembles of size 10, 100, and 1000 conformers per molecule. Table 5 lists in detail the AUC performance of the ROCS<sub>SC</sub> on different ensemble sizes and the AUC results of the 4D FAP<sub>OA</sub>. The table also contains the AUC results of the ROCS<sub>SC</sub> on one given conformation as a baseline to evaluate the performance gain of the multiple conformer setup.

The average AUC of the ROCS<sub>SC</sub> increases from 0.692 (AUC(1)) over 0.703 (AUC(10)) to 0.725(AUC(100)). The results on 1000 conformers are marginally inferior (average AUC(1000) of 0.722) in comparison to the results on 100 conformers. As a result, the ROCS<sub>SC</sub> slightly benefits from the additional information content of multiple conformers. However, the average AUC of the 4D FAP<sub>OA</sub> is 0.80, and, therefore superior in comparison to all four ROCS<sub>SC</sub> setups. These results are verified by the average ranks of the approaches. The 4D FAP<sub>OA</sub> is able to achieve the best AUC value on 27 out of 40 data sets and demonstrates its robust performance on a wide range of pharmaceutically relevant targets. The best ROCS<sub>SC</sub> setup (100 conformers) yields on eight data sets the best result. Please note that the different average AUC values in Table 3 and 5 are the result of the applied data sets (filtered DUD in Table 3 and unfiltered in Table 5).

Despite the robust and superior performance of the 4D FAP<sub>OA</sub> on the majority of the 40 data sets, the weak performance on the HIVPR data set is conspicuous. The HIVPR data set has an average number of heavy atoms of 36.3 and represents the data set with the largest compounds of all 40 DUD data sets. The 4D FAP<sub>OA</sub> entails an optimal assignment step to compute a final similarity value based on the atom-pair tree similarity matrix  $S$ . If the approach computes the similarity value between the query compound and a data set compound, the  $i$ -th row of  $S$  represents the atom-pair tree with the  $i$ -th atom of the query compound as root node. Analogously, this applies to the  $j$ -th column of  $S$  and the  $j$ -th atom of the data set compound. The optimal assignment step maps each atom of the query compound onto an atom of the data set compound. With an increased size of atoms the number of possibilities (possible mappings) scales with  $O^*(n!)$ , where  $n$  is the number of heavy atoms. This increase also increments the risk of a topological error in the assignment step. Topological errors are assignments that do not preserve a substructure mapping (e.g., atoms of a ring are assigned to atoms of different rings). Figure 9 shows a mapping with several topological errors. These topological errors maximize the final similarity value, but from a chemical point of view these mappings are questionable. Therefore, these errors can negatively influence the ranking of the compounds on the HIVPR data set.

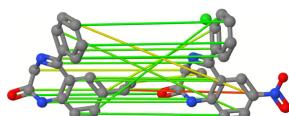


Figure 22.9: Figure 9. Optimal assignment with topological errors

**Optimal assignment with topological errors.** Example mapping with several topological errors. Figure was taken from Jahn et al.<sup>28</sup>

The FieldScreen approach by Cheeseright et al.<sup>45</sup> represents a VS similarity function that applies four different types of locally optimized field points and operates on a multiconformer database. Therefore, it also operates on a comparable information content as our 4D FAP<sub>OA</sub> approach. Cheeseright et al. evaluated the FieldScreen approach on the filtered DUD data sets and applied the chemotype information on the result metrics. The results of the FieldScreen approach as well as the 4D FAP<sub>OA</sub> are listed in Table 6.

The 4D FAP<sub>OA</sub> yields a superior early enrichment performance (awROCE<sub>5%</sub>) on 9 out of 13 data sets. Concerning the performance on the complete data set (AUC) our approach outperforms FieldScreen on 8 data sets. The 4D FAP<sub>OA</sub> is able to increase the mean early enrichment and complete data set performance by 30% and 16%, respectively. The major improvements of FieldScreen in comparison to the 4D FAP<sub>OA</sub> are on the FXa and HIVRT data sets. These data sets also consist of larger molecules, and, therefore the risk of topological errors is increased and is probably a reason for the inferior 4D FAP<sub>OA</sub> performance.

To conclude, the best 3D-based approach of Table 3 (ROCS<sub>SC</sub>) could increase the performance if it is applied on multiple conformer data sets. However, the performance gain was not strong enough to reach the results of the 4D FAP<sub>OA</sub>. The comparison with the FieldScreen approach yields similar results and underpinned the robust performance of the 4D FAP<sub>OA</sub>. The detailed evaluation of the results reveals a weakness of our approach if the compounds of a data set have an increased number of heavy atoms. This weakness is likely the result of the optimal assignment step and was already reported as a weak point of optimal assignment approaches<sup>28</sup>. Nevertheless, the 4D FAP<sub>OA</sub> represents a robust similarity measure for small and medium sized drug-like compounds.

## 22.6 Conclusions

We presented a VS similarity function that operates on GMM encoded conformational space information. Our approach is able to compare the conformational space of molecules within one step, and, therefore avoids the application of time-consuming averaging techniques. The approach was already applied in QSAR experiments and demonstrated its robust performance in comparison to similar 3D-based QSAR models<sup>2328</sup>.

The aim of this study was to evaluate our approach as VS similarity function. Therefore, we compared the results of the 4D FAP<sub>OA</sub> with 20 other 2D- and 3D-based approaches. Additionally, we applied two approaches (ROCS<sub>SC</sub> and FieldScreen) that operate on multiple conformers to provide a comparison of approaches that are based on a similar information content.

The results showed that our approach is able to achieve superior results on a wide range of pharmaceutically relevant targets. Even the best 3D approach, with respect to the results of Venkatraman et al.<sup>21</sup>, applied on multiple conformers is inferior in comparison to our approach.

The preprocessing, which is necessary to encode the conformational space information by means of GMMs, represents an additional computational step. However, all compounds have only be computed once and the encoded models need less space in comparison to the storage of conformational ensembles. The computational speed of the actual similarity function is fast enough to screen over 100,000 compounds within one hour on a standard desktop CPU with one core. Therefore, our approach should meet the requirements of real-world VS applications.

The complete source code of the preprocessing tool (computing GMMs based on conformational ensembles) as well as the 4D FAP<sub>OA</sub> similarity function are publicly available on our department website <http://www.cogsys.cs.uni-tuebingen.de/software/4DFAP>.

## 22.7 List of abbreviations

ACE: angiotensin-converting enzyme; AChE: acetylcholinesterase; ADA: adenosine deaminase; ALR2: aldose reductase; AmpC: AmpC \*β\*-lactamase; AR: androgen receptor; CDK2: cyclin-dependent kinase 2; COMT: catechol O-methyltransferase; COX-1: cyclooxygenase-1; COX-2: cyclooxygenase-2; DHFR: dihydrofolate reductase; EGFr: epidermal growth factor receptor; ER: estrogen receptor; FGFr1: fibroblast growth factor receptor kinase; FXa: factor Xa; GART: glycinamide ribonucleotide transformylase; GPB: glycogen phosphorylase \*β\*; GR: glucocorticoid receptor; HIVPR: HIV protease; HIVRT: HIV reverse transcriptase; HMGR: hydroxymethylglutaryl-CoA reductase; HSP90: human heat shock protein 90; InhA: enoyl ACP reductase; MR: mineralo-corticoid receptor; NA: neuraminidase; P38: P38 mitogen activated protein; PARP: poly(ADP-ribose) polymerase; PDE5: phosphodiesterase 5; PDGFrB: platelet derived growth factor receptor kinase; PNP: purine nucleoside phosphorylase; PPAR<sub>l<sub>nonascii\_39</sub></sub>: peroxisome proliferator activated receptor \*γ\*; PR: progesterone receptor; RXR<sub>l<sub>nonascii\_41</sub></sub>: retinoic X receptor \*α\*;

SAHH: S-adenosyl-homocysteine hydrolase; SRC: tyrosine kinase SRC; TK: thymidine kinase; VEGFr2: vascular endothelial growth factor receptor.

## 22.8 Competing interests

The authors declare that they have no competing interests.

## 22.9 Authors' contributions

AJ designed and developed the main part of the 4D FAP<sub>OA</sub>, has written the manuscript, participated in the design of the experiments and the discussion of the results. LR participated in the design of the experiments and the discussion of the results. GH contributed to the development of the 4D FAP<sub>OA</sub>, participated in the design of the experiments and the discussion. AZ participated in the design of the 4DFAP<sub>OA</sub>, the design of the experiments, and the discussion of the results.



# DATA GOVERNANCE IN PREDICTIVE TOXICOLOGY: A REVIEW

## 23.1 Abstract

### 23.1.1 Background

Due to recent advances in data storage and sharing for further data processing in predictive toxicology, there is an increasing need for flexible data representations, secure and consistent data curation and automated data quality checking. Toxicity prediction involves multidisciplinary data. There are hundreds of collections of chemical, biological and toxicological data that are widely dispersed, mostly in the open literature, professional research bodies and commercial companies. In order to better manage and make full use of such large amount of toxicity data, there is a trend to develop functionalities aiming towards data governance in predictive toxicology to formalise a set of processes to guarantee high data quality and better data management. In this paper, data quality mainly refers in a data storage sense (e.g. accuracy, completeness and integrity) and not in a toxicological sense (e.g. the quality of experimental results).

### 23.1.2 Results

This paper reviews seven widely used predictive toxicology data sources and applications, with a particular focus on their data governance aspects, including: data accuracy, data completeness, data integrity, metadata and its management, data availability and data authorisation. This review reveals the current problems (e.g. lack of systematic and standard measures of data quality) and desirable needs (e.g. better management and further use of captured metadata and the development of flexible multi-level user access authorisation schemas) of predictive toxicology data sources development. The analytical results will help to address a significant gap in toxicology data quality assessment and lead to the development of novel frameworks for predictive toxicology data and model governance.

### 23.1.3 Conclusions

While the discussed public data sources are well developed, there nevertheless remain some gaps in the development of a data governance framework to support predictive toxicology. In this paper, data governance is identified as the new challenge in predictive toxicology, and a good use of it may provide a promising framework for developing high quality and easy accessible toxicity data repositories. This paper also identifies important research directions that require further investigation in this area.

## 23.2 Introduction

With the enormous growth of organisational data and various possible ways to access such data, more and more organisations have become aware of the increasing significance of governing their data. Data governance involves a set of processes to improve data consistency and accuracy, reduce the cost of data management and increase security for the available data<sup>123</sup>. A good data governance framework ensures that organisational data is formally managed throughout the enterprise and provides efficient access to accurate data and business intelligent tools for further analysis.

It is important to note that data governance is different from data management. Data governance complements data management, but never replaces it. In general, management is about the decisions organisation make and it also involves implementing such decisions. Governance concerns what decisions need to be made by whom to ensure effective management, while aiming to provide a structure for achieving these tasks<sup>45</sup>. In other words, governance covers not only the decision domains, but also accountability for decision-making. Take organisational data quality for example, data governance provides a structure for identifying who in the organisation holds the decision making right to determine the standards for data quality, which aspects of data quality need to be included, and how to ensure such standards are attained. On the other hand, data management involves determining the actual metrics which will be employed to assess the pre-defined data quality standards. This paper is mainly concerned with the data governance aspect.

If an organisation settles on a good data governance framework, people within that organisation can more easily and effectively create a clear mission, achieve clarity, maintain scope and focus, increase confidence of using the organisational data, establish accountabilities, and define measurable successes<sup>1</sup>. However, establishing such data governance framework is not an easy task. The characteristics of the data, even the data itself, can be quite different from various organisations. This makes it very difficult to set a unique framework to assess and govern organisational data. In addition, data governance requires the bringing together of diverse expectations and expertises from different departments across the enterprise to achieve agreed, consistent, transparent and repeatable set of processes that enable better data-related decision making.

In many domains of life sciences such as pharmacology, cosmetics and food protection, toxicity assessment at the early stage in a chemical compound discovery process is receiving increasing attention. Predictive toxicology aims to address this problem. The large number of publicly available data sources, development of computational chemistry and biology, and rapidly increasing number of *in vitro* assays have contributed to the development of more accurate QSAR predictive models. The new REACH legislation<sup>6</sup> would require more animal testing to register a new chemical compound, if no alternative methods are used. This requirement pushes scientific institutions to move towards the better use of existing experimental data. Utilisation of toxicity information in conjunction with modelling techniques contribute to reduce the number of animal tests and decreases the cost of the new chemical compound discovery process.

Developing an interoperable and extensible data governance framework for predictive toxicology is crucial and highly required. First, toxicity data includes information about chemical compounds, *in vivo* and *in vitro* experiments. This large amount of information is distributed through publicly available, and often overlapping data sources. Different data formats, incomplete information about experiments and computational errors increase inconsistency of the collected information and make data governance challenging.

Second, in predictive toxicology, Quantitative Structure-Activity Relationships (QSAR) models relate the chemical compound structures to their measured effect or activity<sup>78</sup>. The quality of chemical and toxicity data has an impact on the process of model development for extracting new information, prediction or classification. Thus, it is crucial to ensure the accuracy and consistency of data. Currently, this process involves manual data curation and expert

---

<sup>1</sup> On the governance of information: Introducing a new concept of governance to support the management of information

<sup>2</sup> NOTITLE!

<sup>3</sup> IBM Data Governance webpage

<sup>4</sup> Data Governance Institute

<sup>5</sup> Designing data governance

<sup>6</sup> REACH

<sup>7</sup> Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints

<sup>8</sup> QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review

judgements of the data content. Automated data quality assessment is still a challenge for predictive toxicology<sup>9</sup><sup>10</sup>.

Third, predictive toxicology is a multidisciplinary subject and the development of its data governance framework requires cross-functional groups/organisations to bring together the expertise needed to make predictive toxicology data-related decisions. This makes data governance in predictive toxicology more important and more difficult.

This paper attempts to offer a brief overview of data governance development in predictive toxicology. In particular, some of the most significant and recent public predictive toxicology data sources are reviewed, as well as a discussion on their data governance aspects. This helps to bridge the significant gap in toxicology data quality assessment and hopefully will attract more attention to develop and refine data governance frameworks for predictive toxicology. Such frameworks will provide standards in data representation and easy access to good quality publicly available information. It will also support data quality assessment, common reporting formats and collaborations in knowledge exchange amongst various organisations. The remainder of this paper is organised as follows. In the next section, the main concepts and components of a proposed data governance framework are introduced and discussed. It sets the boundary for the toxicity data sources discussion, with regard to data governance aspects. Based on these components, some of the most significant current toxicity data repositories are reviewed according to their data governance aspects in Section: *Review of Public Data Sources Supporting Predictive Toxicology*. The final section concludes the paper and identifies important further work.

### 23.3 Data Governance: Main Decision Domains

Data governance receives increasing interest due to its importance and advantages in governing the use of data within and outside an organisation. This is evident in that data governance has recently been given prominence in many leading conferences, such as TDWI (The Data Warehousing Institute) World Conference, DAMA (Data Management Association) International Symposium, DG (Data Governance) Annual Conference and MDM (Master Data Management) Summit.

In addition to this, some different general data governance frameworks<sup>15</sup><sup>11</sup><sup>12</sup> have been proposed by different organisations and researchers with an attempt to provide guidance in designing and developing effective data governance approaches. Different organisations may have their own focus on specific aspects of data governance. The above frameworks all aim to provide support for the most common areas of their particular interests.

It is obvious that due to the wide variety of backgrounds, motivations and expectations, the proposed general frameworks can be different. For example, the framework<sup>5</sup> which inherits an existing IT governance framework<sup>13</sup> places special interests in data principles and data life cycle, whereas the work of<sup>11</sup> focuses more on data warehouses and Business Intelligence (BI). Despite this, they still share some common decision domains, such as data quality, data availability and data privacy. In Figure 1, the main decision domains in predictive toxicology data governance that this paper focuses on are depicted. The scopes of the selected decision domains will be firstly introduced and these will serve as the boundary for the toxicity data sources and applications review provided in the following sections.

The data governance decision domains with particular interest for this paper include (see Figure 1):

- **data principle:** is the top level of data governance framework. An effective data principle provides a clear link between the data and the organisational business. Data principle establishes the goal and the main intended uses of data, thus it identifies the directions for all other decision domains. For example, the data principle identifies the data content and the standards for data quality, and these in turn are the basis for how data is interpreted by users in the format of metadata. Also, if the data principle will be publicly available, the data authorisation will be relatively simple.
- **data quality:** is one of the most important domains in data governance. Poor data quality inevitably has huge negative impacts on enterprise profits and operations. When it comes to predictive modelling, in order to develop

<sup>9</sup> Collaborative development of predictive toxicology applications

<sup>10</sup> OpenTox

<sup>11</sup> The DGI Data Governance Framework

<sup>12</sup> A Model for Data Governance - Organising Accountabilities for Data Quality Management

<sup>13</sup> NOTITLE!

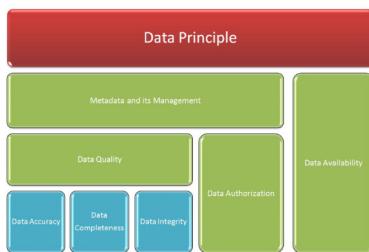


Figure 23.1: Figure 1. Decision domains for data governance  
**Decision domains for data governance.**

accurate predictive models for toxicity values, high quality data is required. In a toxicological sense, data quality may refer to not only the recorded chemical information, but also the quality of experimental results. For example, whether the experiment was performed to Good Laboratory Practice standards (GLP) and whether the identify and purity of the tested chemical compounds were confirmed would affect the quality of experimental results. A more detailed discussion of data quality assessment in *in silico* toxicology can be found in<sup>14</sup>.

Undoubtedly, poor quality chemical and toxicological data with errors and missing information contribute to poor predictive performance and low statistical fit<sup>15 16</sup>. This makes data quality checking a significant task and simultaneously a challenge for predictive toxicology. The quality of a study is judged based on its documentation, e.g. a study report or published paper. Therefore, in this paper, the quality of study data is considered in terms of data provenance under the *metadata and its management* domain. Rather than directly assessing the quality of experimental data, this paper more focuses on the discussion of whether the associated study metadata has been documented and provided together with data sources, which would help the users to make their own judgement about the quality of the underlying toxicology study.

Generally speaking, data quality refers to its fitness for serving its purpose in a given context<sup>5</sup>. In data storage sense, data quality often involves multiple dimensions, such as data accuracy, data integrity and completeness. In most cases, these dimensions need to be defined in the context of data usage. For predictive toxicology, the following three dimensions are of particular interest and they are defined as:

- **data accuracy:** is the fundamental dimension in data quality. Data may come from various internal and external sources, and therefore data accuracy refers to the correctness and consistency of data. For example, given the same chemical name, a poor quality chemical repository may return different chemical structures. This often confuses the user. A curation process is often involved to improve the data accuracy.
- **data completeness:** indicates that the required data are well recorded without missing values. A complete data source covers adequate data in both depth and breadth to meet the defined business information demand.
- **data integrity:** means the wholeness, entirety and soundness of organisational data<sup>17</sup>. It often involves a data integration process which combines data from different sources in an attempt to provide users with a unified view of these data. Data governance concerns some questions in this domain. For example, how to assure the integrity of data and how to determine if the integrity is compromised.
- **metadata and its management:** is defined as data which describes data. Understanding the data context as well as the content and encoding it into meta representation is a core aim of data governance<sup>3</sup>. A concise and consistent metadata representation makes the semantics of data become interpretable to users and ensures the data can be effectively used and tracked.

Organisational metadata can be classified into different categories, including the physical metadata, provenance metadata, domain-specific metadata and user metadata<sup>5</sup>. The physical storage metadata describes the physical storage of data. Provenance metadata refers to the data author and time stamp information, e.g. the source of the data, who is

<sup>14</sup> Chapter 4 Data Quality Assessment for In Silico Methods: A Survey of Approaches and Needs

<sup>15</sup> NOTITLE!

<sup>16</sup> Best Practices for QSAR Model Development, Validation, and Exploitation

<sup>17</sup> Data Integration: A Theoretical Perspective

the owner of this data, when it was created and when has been last modified. If the recorded data is up-to-date for the task at hand and/or comes from a reliable source, it will contribute to better data quality. Domain-specific metadata summarises the semantics of data content. For example, in predictive toxicology, the study metadata indicates its key elements (e.g. study type, species and endpoints). User metadata captures the user information and historical usage record.

Different categories of metadata contain lots of valuable information and play key roles in data management, retrieval, discovery and analysis within organisations. The development of metadata repositories and their efficient stewardship is one of the essential activities of data governance and it will maximise the value of collected data, support decision making processes and business needs. In addition, due to the rapid changing business environment, the way an organisation conducts business and the consequent data also changes. As such, there is also a need to manage changes in metadata. Metadata management includes: data object standards and definitions, identifying relationships between data objects, providing accuracy, completeness and timeliness measurements, standards of documenting and reporting. In short, data governance uses metadata management to impose management discipline on the collection and control of data.

- **data availability:** concerns how users can access the data. It includes the data access and export formats (e.g. xml, pdf and spreadsheet), and how it can be accessed in what way and on what devices (e.g. PC and mobile phone). For example, some toxicity data sources can be downloaded in a dump sql file for local processing, whereas some data sources can only be queried about fixed fields via web/standalone system graphic user interface (GUI).

- **data authorisation:** controls user access to private and sensitive data based on user privileges. The data might have multi-level user access by the support of the pre-defined authorisation policies. The authorisation policy indicates what parts of the data can be accessed/manipulated by whom. For example, given a database, some data may be publicly available, while for some sensitive data, only authorised users can access it.

## 23.4 Review of Public Data Sources Supporting Predictive Toxicology

Collection of sufficient *in vivo* - *in vitro* assays is a starting point to build predictive models in predictive toxicity<sup>1518</sup>. In the literature, there is a large amount of chemical, biological and toxicological data; however, only a small portion of them can be directly used for quantitative toxicity prediction. The current data sources mainly cover the following three major domains (see Figure 2):

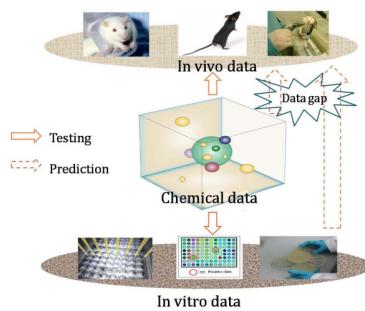


Figure 23.2: Figure 2. Data domains for predictive toxicology  
Data domains for predictive toxicology.

- **chemical data:** this domain contains the chemicals and their associated information such as chemical descriptors, chemical and physical properties and molecular structures, which are used to build predictive models.

<sup>18</sup> Public Databases Supporting Computational Toxicology

• *in vivo* toxicology data: refers to information collected from experiments or studies done in live organisms. Currently, *in vivo* assay data is distributed across various resources such as scientific articles, company internal reports, governmental organisation documents and many institutional services. Integrating this information in publicly available data sources by appropriate extraction, curation and pre-processing is challenging and extremely valuable.

• *in vitro* data: *in vitro* cell and molecular biology technology has attracted more attention of toxicology researchers, due to its relatively low cost. In comparison to *in vivo* technology, *in vitro* research is more suitable for the deduction of biological mechanisms of action<sup>15</sup>. Together with chemical information *in vitro* data are used to predict *in vivo* toxicity and to prioritise animal testing.

The huge amount of varied data impacts on the modelling process. However, there exists a common problem in data quality assessment. Due to the lack of standard measures (such as accuracy, consistency, comprehensiveness, coverage, accessibility, reliability) for quality evaluation, the quality of data sources is either not assessed or just assessed manually. The current assessment is normally quite subjective and mainly depends on assessors' own experiences and preferences. This makes the data governance of current toxicological data sources very important.

Recently, many toxicogenomics and chemical data sources have been reviewed according to needs and challenges in predictive toxicology<sup>18 19 20 21 22 23 24</sup>. In this section, the discussion of several major data sources that are widely applied to predictive toxicology is presented, with particular focus on their data governance aspects.

More specifically, in the context of data governance, the framework which has been proposed in the previous section sets the boundary for discussion. The main framework components, including data accuracy, data completeness, data integrity, metadata and its management, data availability and data authorisation are the main interest of this paper. Their status in different data sources will be discussed respectively. It is also worth noting that the list of data sources discussed here is not exhaustive, but the selected data sources illustrate well the range of the publicly available data sources. Table 1 details the analysis of various data sources according to the previously mentioned components of data governance framework.

### 23.4.1 ChemSpider

The development of the Internet fosters on-line publications of chemical information by various organisations. The broad distribution of chemical and physical properties and molecular structures results in duplicated storage cost, inefficient information retrieval and inconsistency. Thus, there was a need to provide a framework for data integration and access to high quality chemical information.

The free online repository ChemSpider<sup>25</sup> was developed to address this problem. It was launched in 2007 to provide searchable chemical structures and property information in the public domain. The key feature of ChemSpider is the compound structure centric database. All possible available information about a given chemical compound is linked by its molecular structure, though ChemSpider is not an exhaustive database. The aim of ChemSpider is to provide services in chemical information searching, and also to build a crowd-sourcing community. In predictive toxicology, this community is a group of specialists from different organisations that contribute their knowledge to improve the quality of collected information in a database. At present, ChemSpider does not directly support toxicity predictions, however, the content of ChemSpider provides high quality chemical information and links to original resources that are very useful for building QSAR models.

Reflecting back to the data governance framework (see Figure 1), the main domains are discussed as follows:

- **data accuracy:** ChemSpider has realised that ensuring the accuracy of included data is an essential requirement for public data sources, and it is trying to distinguish itself from other public chemical repositories (e.g. PubChem

<sup>19</sup> The Comparative Toxicogenomics Database: update 2011

<sup>20</sup> Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks

<sup>21</sup> Database development in toxicogenomics: issues and efforts

<sup>22</sup> ChemSpider: An Online Chemical Information Resource

<sup>23</sup> Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics dataCEBS

<sup>24</sup> Toxicogenomics and systems toxicology: aims and prospects

<sup>25</sup> ChemSpider

<sup>26</sup> from the National Institutes of Health (NIH)). The NIH currently lacks a data curation mechanism to ensure the quality of included data. It relies totally on the data depositors to curate their own data. As a result, errors from various depositors and multiple representations and formats are included in PubChem. Rather than depending on any individual depositor, ChemSpider has spent much effort in improving data quality by employing the crowd-sourcing activities of the community. The following systematically organised data curation is performed by ChemSpider and this can be achieved both automatically and manually:

- the general curation activities include: removing incorrect names, correcting spellings, adding multilingual names and alternative names.
- to avoid anonymous act of vandalism, only registered users are allowed to edit the records.
- when uploading a chemical structure to ChemSpider, automated chemistry checking is performed.
- domain experts are invited to continuously validate and update included data. The data curation labels include “validated by experts”, “validated by users”, “non-validated”, “redirected by users” and “redirect approved by experts”<sup>27</sup>.

Currently, more than 130 people are involved in data validation and annotation. Over a million chemical structure and identifier relationships have been validated either automatically or manually<sup>22</sup>. This data curation effort will continue with an intention to offer the highest quality online chemical database.

- **data completeness:** ChemSpider currently contains over 25 million compounds (ChemSpider Count 03/2011) and the number is growing daily. It is claimed to be the richest source of structure-based chemistry. The variety of information about a compound provided at ChemSpider is hard to match on any other free website.
- **data integrity:** ChemSpider aims to act as an aggregator of chemical information. Data from nearly 400 different data sources (including Wikipedia, PubChem<sup>26</sup>, ChEBI<sup>28</sup> and etc) are integrated and linked by means of chemical structure. Where possible, each chemical record retains the links to the original source of the material and also links out to other information of particular interests, including where to purchase a chemical, chemical toxicity, metabolism data, etc. Instead of employing classical search engines (e.g. Google) to search individual pieces of information, aggregating such chemical information into a central database saves lots of search effort and time for users.
- **metadata and its management:** ChemSpider uses metadata representation to extract information about a chemical compound, its associated data sources together with its external IDs (if available) and relevant scientific articles. Additionally, the creation date and owners of the included data sources are provided in ChemSpider. The service keeps track of curation updates and makes this information available to the master curators. Also, once registered users login to the system to perform the search, their search history will be recorded and stored. This information can be further used to capture user preferences and then provide more customised services.
- **data availability:** ChemSpider also realised the importance of data availability and accessibility. The website provides various simple and advanced search options via a user friendly web GUI. In addition to this, the system provides web services (including APIs) to allow users to query the content and request physiochemical properties prediction, and the retrieved results can be downloaded as a set. The included structure images and spectra can be easily embedded into external web pages by the use of provided tools. In addition, the newly developed Mobile ChemSpider allows users to access ChemSpider data through mobile devices, such as mobile phone browsers and iPads. Note that users are limited to assemble 5000 ChemSpider records or less to build an in-house data source.
- **data authorisation:** ChemSpider is a publicly available database, the RSC logs the IP address of user's PC to be able to receive and send information on the Internet. However, no multi-level user access is provided, all registered users share the same authorisation schema to access the data.

<sup>26</sup> PubChem

<sup>27</sup> ChemSpider and Its Expanding Web: Building a Structure-Centric Community for Chemists

<sup>28</sup> Chemical Entities of Biological

## 23.4.2 CEBS

Chemical Effects in Biological Systems (CEBS) is the first public repository which captures toxicogenomics data developed by the National Center for Toxicogenomics (NCT) within the National Institute of Environmental Health Science (NIEHS)<sup>23<sup>29</sup></sup>. A distinguishing feature of CEBS is that it contains very detailed animal-level study information including treatment protocols, study design, study time-line, metadata for microarray and proteomics data, histopathology and even raw genomic microarray results<sup>23</sup>.

The objective of CEBS is to provide users easy access to the integrated wide diversity of data types and detailed study information. The embedded rich information makes it possible to develop answers to comprehensive queries posted in the database, and then conduct gene signature and pathway analysis based on the retrieved answers. Users are allowed to query the data using study conditions, subject responses and microarray module.

The main domains in the previously presented data governance framework (as shown in Figure 1) are described as:

- **data accuracy:** the accuracy of CEBS data is handled by collaboration between the data depositors and internal curation staff. Prior to being exported to CEBS database, all study information needs to go through Biomedical Investigation Database (BID) which is a component of CEBS responsible for loading and curating data.
- **data completeness:** as of 2010, CEBS contains 132 chemicals (structure searchable via U.S EPA(Environmental Protection Agency) DSSTox) and their responses which were derived from 34 studies in mice, rats, and *Caenorhabditis elegans*<sup>18</sup>. Most of the included studies have associated microarray data. CEBS welcomes high-quality study data relating to environmental health, pharmacology and toxicology. When submitting such data, a study design and phenotypic anchoring data is required.
- **data integrity:** one of the main objectives of CEBS is to permit users to integrate various data types and studies. By the support of a well designed relational database schema, the CEBS users can effectively retrieve the associated biological and toxicological data based on the selected subject response or study conditions, etc. CEBS has employed controlled vocabularies (i.e. CEBS data dictionary) rather than free text to capture study related data, allowing data to be integrated within a given microarray platform for effective filtering, query and analysis. On the other hand, CEBS has to manage data from a variety of resources and each resource may use different study designs, treatment regimes and measures. Currently, there is lack of a widely accepted public standard for exchange and capture of such study design and assay data. A number of effort is under way to address this need, and CEBS will fully support the development of a standard which will provide better data integration.
- **metadata and its management:** every study included in CEBS has its associated details document containing the following information: the institution, principal investigator, start date of the study, design details and supporting publications (together with their PubMed IDs). Such metadata can be easily retrieved in various graphical representations by clicking the provided links on the web page. This detailed domain-specific metadata helps to provide customised search options and then maximise the included data values.
- **data availability:** CEBS provides easy access for users to retrieve, combine and download customised information. By using the SysTox browser, the CEBS users can combine components of different workflows provided by CEBS to customise their queries and export them to various formats for downloading. Alternatively, users can use the provided FTP service to individually download data sources. In addition, users are allowed to create and manage their own workspaces. The infrastructure has been simulated to be able to support up to 100 concurrent users in various use cases and workflows.
- **data authorisation:** although CEBS is claimed as a public repository integrating study design and assay data, it only allows public access to the fully published data sources. For those not fully published data sources, they are not available for downloading without prior agreements of data owners. It is common to obtain access to unpublished data sources on a collaborative basis.

One unique feature of CEBS is its allowance for users to upload their own data into CEBS in a private mode and only make it visible to their nominated collaborators behind the CEBS firewall<sup>30</sup>. This private data authorisation schema

---

<sup>29</sup> Toward a Checklist for Exchange and Interpretation of Data from a Toxicology Study

<sup>30</sup> CEBS

securely protects sensitive user data and also allows users to integrate their own data with other public data sources in CEBS for analysis.

### 23.4.3 CDT

Environmental chemicals may play a crucial role in the etiology of human diseases. Despite this observation, the mechanism of action and the potential influences of most chemicals on many diseases are not known<sup>19203132</sup>. To gain a better understanding about the impact environmental chemicals have on human health, the Comparative Toxicogenomics Database (CTD)<sup>3334</sup> has been developed by Mount Desert Island Biological Laboratory. It serves as a unique centralised and freely available resource to explore the interactions amongst chemicals, genes or proteins and diseases in diverse species.

Chemicals might interact with various genes and proteins in multiple ways and then affect the mechanisms underlying the etiology of diseases. One of the major goals of CTD is to support the generation of novel hypotheses about chemical actions and environmental diseases. It is worth noting that CTD acts not only as a data repository but also as a discovery tool to generate novel inferences. The inferred interactions can also be evaluated based on local statistics, and the derived ranking scores will help users to prioritise further testing. In addition, a set of tools to visualise, manipulate and analyse different types of data such as comparisons of gene sequences from different species or comparisons of associated data sources for up to three chemicals, diseases or genes are provided by CTD.

In terms of data governance framework components, the following aspects are considered:

- **data accuracy:** the included chemical-gene interactions, chemical-disease and gene-disease relationships in CTD are completely manually curated, thus the data accuracy relies solely on professional experts. Curators should be trained in the data curation according to CTD requirements. Prior to monthly public releases, the curated data submitted by an individual curator still needs to go through a further review conducted by the scientific community, and this helps to ensure the high accuracy of curated data. It is obvious that such manual curation is time consuming and this task becomes more challenging with the increasing scope and volume of available data. With an attempt to speed up and improve the efficiency of manual data curation, a prototype text-mining application has recently been developed to prioritise the available literature<sup>32</sup>.
- **data completeness:** to date, CTD database includes 1.4 million chemical-gene-disease data connections and new data is available monthly. It currently consists of over 240,300 molecular interactions between 5900 unique chemicals and 17,300 gene products, 11,500 direct gene-disease relationships and 8500 direct chemical-disease relationships extracted from over 21,600 publications<sup>19</sup>. The applicability and utility of CTD has been widely recognised and it is evident that CTD is being indexed by many public repositories.
- **data integrity:** all curated data, especially the inferred interactions, are captured in a structured manner to minimise the inconsistency amongst different curators. This is achieved by employing community-accepted controlled vocabularies and ontologies. This not only helps users to retrieve data efficiently, but also provides a useful way of integrating and communicating with external data sources by using consistent terms. CTD data is integrated with a number of external chemical, gene, disease and pathway resources, including ChemIDplus, DrugBank, Gene Ontology Consortium, NCBI Gene, NCBI PubMed, etc<sup>20</sup>.
- **metadata and its management:** the original references of the curated interactions are included in CTD. For quality monitoring and follow-up purposes, the data of the curation such as the curator ID, date of curation and related articles is also recorded. Note that, the metadata of CTD records not just providing the domain-specific information, but also includes some evaluation metrics, such as similarity score and inference score.
- **data availability:** CTD provides a wide variety of ways to access the included data, users can easily download individual data sources. Alternatively, the whole database as a dump file can be downloaded for local analysis. In addition, CTD allows users to perform both detailed query and batch query to find various types of data instead of just

<sup>31</sup> The Comparative Toxicogenomics Database: A Cross-Species Resource for Building Chemical-Gene Interaction Networks

<sup>32</sup> Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD)

<sup>33</sup> Comparative Toxicogenomics Database

<sup>34</sup> The Comparative Toxicogenomics Database (CTD)

those relating to a specific chemical, gene or disease term. The retrieved results can be customised and exported to different formats (such as csv, xml and tsv).

- **data authorisation:** CTD is a community-supported public resource tool that advances understanding of the effects of chemicals on human health. However, no multi-level user access is available. All registered users share the same authorisation schema to access the data, so that there is no authorisation protection of users' private data.

### 23.4.4 DSSTox

The Distributed Structure-Searchable Toxicity (DSSTox) public database network provides a public forum for high-quality, standardised chemical structure files associated with toxicity data<sup>35</sup>. It aims to help in building publicly and easily accessible data collection to improve predictive toxicology capabilities. A major goal of the DSSTox project is to encourage the use of the DSSTox data format (including DSSTox Standard chemical structure fields and standardised SDF (structure data format)) for publishing chemicals and their associated toxicity data files.

Recalling the data governance framework in Figure 1, the related domains are discussed as:

- **data accuracy:** one of the unique features of DSSTox is the quality control of the included data. All documentation and data files considered for DSSTox publication are subject to DSSTox project review, EPA internal review, and in some cases outside peer review<sup>35</sup>. In addition, most data sources are curated by experts and uniformly applied into DSSTox<sup>18</sup>. DSSTox deals with the quality control process only according to the information stored in its own data sources. An extensive and clear information quality review procedure is applied to the annotation of included data. Every data source includes a quality assurance log file. This file summarises all undertaken procedures to ensure accuracy and consistency of chemical structure within a data source<sup>35</sup>. Chemical structures are retrieved from outside sources and cross-validated for internal consistency. All errors and missing values are reported in a log file. Moreover, DSSTox provides procedures for users to report errors in DSSTox data sources. Such community efforts support integration and migration of chemical toxicity information into DSSTox. Finally, data is validated by experts (DSSTox users). It is worth noting that DSSTox does not provide a quality review for information collected from outside resources.
- **data completeness:** the current DSSTox database contains over 8000 chemicals and has been incorporated into several external resources, including: ChemSpider, PubChem inventory<sup>26</sup>, GEO<sup>36</sup>, ACToR database, ArrayExpress Repository, and EMBL-EBI. DSSTox standardised format has gradually gained popularity in recent years. It is anticipated that more public chemical information and related toxicity data will be migrated into the DSSTox for publishing in the near future.
- **data integrity:** publicly available toxicity data sources exist in different locations and widely disparate file formats. They are quite different in terms of toxicity endpoints, test methods, treatment conditions and degrees of result details. Additionally, they are often not downloadable in their entirety and most do not include related chemical structures embedded with rich content. Being aware of this, DSSTox extends the existing SDF and annotates the list with DSSTox Standard Chemical Fields to integrate the molecular structures and toxicity data into standardised DSSTox SDF. This is fully incorporated into U.S EPA ACToR and PubChem.
- **metadata and its management:** metadata of published data sources are well recorded. Every data source is associated with documentation including following information: how and when chemical compounds were collected into a data source, links to the outside repositories, time-stamp of when a data source was created and updated, information about last update; a list of authors and reviewers of DSSTox data source and supporting literature.
- **data availability:** DSSTox data sources and their documentations can be downloaded from the DSSTox website. User contributions are used to build larger DSSTox user communities. Each of the published data files can be freely downloaded in common formats such as pdf, sdf and spreadsheet, and are of potential use for (Q)SAR modelling. Large files are offered in compressed file format. In addition, a DSSTox Structure-Browser has been developed by EPA to provide a simple and handy structure-searching capability. The full collection of DSSTox published data files

---

<sup>35</sup> DSSTox

<sup>36</sup> GEO

is searchable using many options, including chemical text, data file, SMILES, structure, generic test substance level and outputs.

- **data authorisation:** DSSTox allows full and publicly open access to included toxicity data files and no user registration is required. Although the DSSTox website collects and stores user access log file, including access date and time, IP address, the objects/web page requested, access status and etc, no multi-level user access is enabled, all users share the same authorisation schema to access the data. To ensure that the service remains available to all users, EPA also makes effort to identify and block unauthorised attempts to upload or change information on their website.

### 23.4.5 ToxCast

There is a huge data gap between environmental chemicals and their associated toxicity information. This is because of the expense and length of time required to conduct animal testing to obtain such toxicity data. Traditional animal testing provides very limited information on mechanism of action (MOA). MOA is a sequence of events from the absorption of a compound into a living organism to toxic outcome and it is crucial to predicting toxicity in humans. Therefore, there is a pressing need to screen the large backlog of chemicals for their potential toxicity and, ultimately, their contribution to human diseases<sup>37</sup>. Inspired by this, EPA NCCT launched the ToxCast Project<sup>3839</sup> in 2007, in an attempt to develop an effective approach for prioritising the toxicity testing of a large number of environmental chemicals at low cost. The major goals of ToxCast are<sup>37</sup>:

- to detect *in vitro* assays which can reliably indicate alterations in biological processes that may lead to adverse health effects,
- to improve *in vivo* toxicity prediction by developing signatures or computational models from multiple *in vitro* bioassays, together with calculated and available chemical properties, rather than just using a single assay or chemical structure alone,
- to speed up the screening of the large number of untested environmental chemicals by the use of detected signatures from *in silico* and *in vitro* data.

As stated in the data governance framework (as shown in Figure 1), related domains are discussed as follows:

- **data accuracy:** it is claimed by EPA that the data included in the ToxCast project has been subjected to internal technical review and approved for research use. In most cases, the data has also undergone the external peer-review and publication in scientific journals<sup>39</sup>. In addition, the collection of ToxCast Phase I chemicals is carefully chosen subject to the availability of high quality, guideline-based animal toxicity data, and chemical property space coverage. The selected compounds represent the chemical space well and there are only a few compounds with extreme property values. In terms of updates, subsequent chemical analysis and verification of activity is still going on and the findings are added to the current data sources periodically.
- **data completeness:** in its initial stage, Phase I, ToxCast has profiled 320 chemicals, most of which are pesticides due to the availability of their extensive animal testing results, in over 400 high-throughput screening (HTS) endpoints using 9 different assay technologies. The current data source covers various chemical classes and diverse mechanisms of action and it has been deposited in PubChem. A total of 624 *in vitro* assay endpoints ranging from gene to entire organism have been measured for each chemical in ToxCast Phase I. The output of each chemical-assay combination was reported in terms of either half-maximal activity concentration (AC50) or lowest effective concentration (LEC) at which there was a statistically significant change from the concurrent negative control.

ToxCast Phase I is a proof of concept project to demonstrate its impact in various dimensions (e.g. chemical space, HTS assay data (“fast biology”) and *in vivo* bioassay data (“slow biology”)). Later phases will expand its impact to broader coverage of chemical space (over thousands of chemicals) and employ classes to validate the predictive toxicity signatures which built in Phase I. Phase II is currently screening 1,000 chemicals from a wider range of sources, including industrial and consumer products, food additives and drugs<sup>39</sup>.

<sup>37</sup> In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project

<sup>38</sup> The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals

<sup>39</sup> ToxCast

• **data integrity:** it is important to note that ToxCast includes *in vivo* data for ToxCast chemicals and such information is stored in a relational database called ToxRefDB<sup>40</sup> which contains nearly \$2 billion worth of animal toxicity studies (55,950 in total). With the growth of the ToxCast database, the confidence in building statistical and computational models to forecast potential chemical toxicity will increase. This will result in refinement and reduction of animal use for hazard identification and risk assessment.

In addition to this, ToxCast data is well integrated to many other EPA databases. The ToxCast toxicity testing results are also available in the EPA ACToR data warehouse. They can be easily searched and combined with other related testing results derived from other projects. Also, the Phase I chemical structures and associated toxicity data are integrated in SDF files and they are available for download at the DSSTox website.

• **metadata and its management:** due to the availability of ToxCast Phase I data in DSSTox, the associated metadata (see data governance discussion in DSSTox) for these included information is provided. The software packages used to calculate physical and chemical properties from a chemical structure are indicated in the ToxCast data sources. For those *in vitro* assays included in ToxCast Phase I, the associated information, including contractor/collaborator, assay type, date stamp, assays/endpoints and references (publications and websites), are also provided in the EPA ToxCast website. In addition, ToxRefDB is the relational database storing *in vivo* toxicity testing data of ToxCast Phase I chemicals. The study details including time-stamps and owners are recorded in ToxRefDB.

• **data availability:** the ToxCast data sets, including descriptions, supplemental data and associated publications are available from the U.S. EPA ToxCast website. The users are allowed to download them individually in zip files. More complete and detailed information of the ToxCast chemicals, including chemical structures and identifiers, can be found in the EPA DSSTox website. The DSSTox ToXCST files [http://www.epa.gov/ncct/dsstox/sdf\\_toxcst.html](http://www.epa.gov/ncct/dsstox/sdf_toxcst.html) can be downloaded in diverse formats, such as sdf, spreadsheet and pdf. In addition, the ToxRefDB makes it possible to link toxicity information with the HTS and genomic data of ToxCast within the ACToR system. Recently, a new interface to access all ToxCast data has been provided via ToxCastDB<sup>41</sup>. It delivers a clear and easy browse format, such that users can search and download ToxCast chemicals, assays, genes, pathways and endpoints more effectively.

• **data authorisation:** similar to other EPA databases, ToxCast provides full and open access to included data and no user registration is required. For details, readers can refer to DSSTox data authorisation discussion.

### 23.4.6 ACToR

Tens of thousands of chemicals are currently in commercial use, but only a small portion of them have been adequately assessed for their potential toxicological risk due to the expensive and time consuming conventional chemical testing methods. As aforementioned, the ToxCast project has been developed by EPA to speed up the process in filling such data gaps. ToxCast is also a major driver of the development of the Aggregated Computational Toxicology Resource (ACToR) project<sup>42</sup>. It is difficult to find related information about a given chemical. ACToR is therefore developed to support the ToxCast screening and prioritisation effort.

An important goal of the ACToR project is to develop a publicly accessible and widely usable database that gathers toxicity data associated with a large number of environmental chemicals from multiple sources<sup>1842</sup>. The key uses of ACToR are to support predictive toxicology in terms of computational analysis<sup>4243</sup>:

- collect and combine data from different sources to construct training and validation data sources to support high-throughput chemical screening and prioritisation efforts.
- serve as a unique resource which links chemical structure with *in vitro* and *in vivo* assays to support the development of computational models.
- provide EPA and other regulatory agency reviewers with the decision making support for novel chemicals approval.

---

<sup>40</sup> ToxRefDB

<sup>41</sup> ToxCastDB

<sup>42</sup> ACToR - Aggregated Computational Toxicology Resource. 2007 Toxicology and Risk Assessment Conference: Emerging Issues and Challenges in Risk Assessment - 2007 TRAC

<sup>43</sup> The Toxicity Data Landscape for Environmental Chemicals

- support the workflow construction feature enabling users to build customised prioritisation of data capture, quality control and chemical prioritisation scoring tasks.

The related domains in the data governance framework (see Figure 1) are examined as follows:

- data accuracy:** ACToR database itself does not provide any data curation, the accuracy of the included data totally depends on the original sources of the data. Therefore, to ensure high data accuracy, the sources of included data collections are carefully selected by ACToR and the sources' institutional information are well recorded.
- data completeness:** currently, ACToR is made up of over 500,000 environmental chemicals and their associated toxicity data which were derived from more than 500 public data sources including various databases such us: DSSTox, ToxCast and ToxRefDB; NIH, PubChem and TOXNET. ACToR is focused mainly on capturing chemical structures, physical and chemical properties, *in vitro* bioassays and *in vivo* toxicity data. In terms of chemical structures, EPA DSSTox and PubChem are the two main sources for ACToR due to their high data quality and wide data coverage, respectively. Although ACToR currently does not include all toxicity data, it is designed to be flexible enough to incorporate new data from sources with different formats into the system in a straightforward manner.
- data integrity:** ACToR itself is an integration of available chemical toxicity data from a large number of sources. It is common that chemical toxicity data are stored in incompatible formats and in many different locations. In the past, in order to retrieve all relevant information for a given chemical, users needed to come across diverse data sources and aggregate the results manually for use. With the rapid increase of available chemicals and data sources, such a task becomes impractical and even impossible for comprehensive data sources. ACToR aggregates large number of chemical toxicity data and makes the included data searchable by chemical information (e.g. chemical name, chemical structure and identifiers). The data collections are formatted and organised in a consistent manner within ACToR. The clear and flexible ACToR database schema facilitates users to search and query data from other chemical toxicity data sources including ToxRefDB, ToxCastDB, DSSTox, ExpoCastDB and others to be added in the future.
- metadata and its management:** the institutional sources of data collections and assay information are well included in ACToR. The recorded metadata of data collections includes: details of the data collection, source ID, name, description, source type, number of substance, number of generic chemicals, number of assay results, and link-out which provides a direct connection to the external website of a given data collection. In addition, the study and assay time-stamps, source ID, name, units, value type, component type as well as owners are included in the original data source, such as ToxRefDB.
- data availability:** one of the significant advantages of using ACToR is its availability. The included data is provided in an easy accessible and computable manner via its web interface. The ACToR system is implemented using 100% open-source software, MySQL for the database development, Perl for loading data and Java for web interface development. This enables the user to download the entire ACToR database for local analysis. In addition, the ACToR database covers a wide variety of data sources. In particular, due to the use of similar data schemas, all data in PubChem can be easily loaded into ACToR. This feature makes the ACToR database easily expandable and scalable in the future.
- data authorisation:** similar to other EPA databases, ACToR provides full and open access to the included data and no user registration is required. For details, readers can refer to DSSTox data authorisation discussion.

### 23.4.7 OpenTox

OpenTox is a modern predictive toxicology framework under development<sup>10</sup>. Different from other toxicity data sources, OpenTox provides easy access to not only the good quality data, but also a collection of various predictive toxicity applications. OpenTox is designed to support effective data exchange and accurate cross-organisational communication and collaboration. Its flexible architecture and modular design contribute to the development of customised predictive toxicology applications with respect to user requirements. Currently, OpenTox provides two applications for model development and toxicity estimation. *ToxPredict* allows users to predict a toxicity endpoint for a given chemical compound.

*ToxCreat*e supports predictive model generation. Developed models within the OpenTox can be validated, reported

according to the OECD principles<sup>44</sup> and published within the OpenTox framework. Detailed model reporting supports reliable judgements about model validation and gives the possibility to reproduce predictions.

Data access and management, model development, feature construction and selection are core components of the OpenTox framework. Thus, OpenTox supports the creation of dictionaries and ontologies, that describe the relations between chemical and toxicological data and experiments and to develop novel techniques for the retrieval and quality assurance of toxicological information<sup>9</sup>.

Recalling the data governance framework in Figure 1, the related domains are discussed as:

- **data accuracy:** OpenTox provides data quality assessment by assessing validation labels to included data (e.g. 2D chemical structures). It allows for three types of data quality validation: automated, manual and global. For a given chemical compound, its chemical structures can be imported from different sources. Then, these structures are automatically compared and classified into the following groups and each group is associated with a predefined validation label:

- consensus (“OK”) - all structures are identical,
- majority (“ProbablyOK” and “ProbablyError”) - majority of identical structures,
- ambiguous (“Unknown”) - there is no majority of equal structures,
- unconfirmed (“Unknown”) - single source and no comparison available.

The assigned quality labels will be further reviewed by experts according to their knowledge and manual comparison with external sources. The global validation aggregates the validation labels which derived from automated and manual validations for a given data source. Opentox employs a numerical measure, “ProbablyError” rate, to indicate the overall quality of a given data source. Obviously, the lower “ProbablyError%” indicates the better quality of the data source.

- **data completeness:** OpenTox mainly focuses on publicly available toxicology data. As reported in<sup>9</sup>, OpenTox framework currently includes data from ISS ISSCAN, IDEA AMBIT, JRC PRS, EPA DSSTox, ECETOC skin irritation, LLNA skin, and the Bioconcentration Factor (BCF). The additional informations for chemical structures has been collected from public sources such as Chemical Identifier Resolver, ChemIDplus, PubChem.

- **data integrity:** is a current challenge for OpenTox. An ontology and controlled vocabularies have been developed by OpenTox to integrate and organise multidisciplinary data (e.g. chemicals, experiments, and toxicity data). With the support of ontology, OpenTox is currently moving towards the development of the Resource Description Framework (RDF) representation to exchange data from various sources.

- **metadata and its management:** OpenTox database provides means to identify the original sources of the included data by indicating inventor name and reference. By doing this, the user is allowed to select the compounds of interest from a specified inventory. OpenTox also makes the latest updates of the data (e.g. updates of chemical structures and descriptor calculations) which become available. OpenTox includes enhanced metadata for algorithms, models and datasets that are managed by the ontology web service.

- **data availability:** toxicity data is currently publicly available and accessible via the OpenTox Representational State Transfer (REST) web services. The RESTful web service has been chosen because it allows for the combination of different services into multiple applications to satisfy diverse user requirements. Additionally, OpenTox offers workflow architecture that is understandable and easy interpretable to users.

- **data authorisation:** the OpenTox website has public sections that are read accessible by anyone with a web browser. Only the OpenTox Development area requires a user name/password registration and registration approval. Having recognised the importance of multi-level user access, OpenTox have considered different authentication and authorisation solutions for an initial implementation to grant access to protected resources. A set of REST operations (including authentication, authorisation, create policy, delete policy and etc) have been published in OpenTox API.

---

<sup>44</sup> OECD principles for the validation, for regulatory purposes, of QSAR models

## 23.5 Summary

While the above discussed public data sources are well developed, there nevertheless remain some gaps in the development of data governance framework to support predictive toxicology. Firstly, more and more data repositories have realised the importance of included data quality and their inherent impacts on QSAR modelling. In most of the reviewed data sources (except Actor), data curation procedures are applied to ensure accuracy and completeness of included data. These features are important components in data quality assessment. However, the data quality checking is mostly based on human expertise. While there are techniques leading to the automated data quality assessment, the derived results still have to be validated by experts. Moreover, there is a lack of systematic and standard measures of data quality, including data completeness, accuracy and consistency. The current assessments are all based on internal measures and this causes difficulty in comparing data from different sources. A more interpretable and transparent assessment mechanism is highly desirable and such standard measures would also be used to visualise data quality and then to compare and rank different data resources.

Second, data provenance is another substantial issue in predictive toxicology. Provenance is important and valuable to understanding, aggregating and making use of data sources and scientific results. In the presented data governance framework, data provenance is a part of metadata and its management. As shown in this review, the information about data submission, curation and authors are included as metadata in different manners and formats. This makes tracking the provenance of data very challenging. The determination of data provenance procedure and representation requires coordination and cooperation between various data sources, and this is currently very difficult. Automated data quality assessment will provide a more systematic and analytical metadata representation and this could lead to simplified and unified data quality control processes.

Third, the publicly available data sources contain rich and valuable scientific data and they are also of great interest to commercial companies. Combining company in-house data with existing public data will help to discover more hidden ideas and knowledge. However, as shown in the review, although CEBS and OpenTox have recognised the importance and benefits of multi-level user access, this development is still at the early stage. Most public data sources currently do not allow multi-level user access. All (registered) users share the same authorisation schema to access the data. The lack of protection of private and sensitive data becomes a bottleneck which limits commercial companies to contribute their in-house data to public repositories. It is challenging but highly desirable to develop more flexible and customised data authorisation schemas which will allow multi-level user access to both in-house and public data resources.

## 23.6 Conclusions

In this paper, the review of some of the widely used public data sources which support predictive toxicology has been presented with regard to a proposed data governance framework. Current predictive toxicology challenges such as data integration and data quality assessment were authors' motivations to look at the existing solutions. In this review, widely used and well-known data resources were chosen, but the choice was not exhaustive.

It has been well recognised that data quality inherently affects model development. With the increasing amount of varied toxicity data that comes from *in vivo* -*in vitro* studies, there is expected to be a boom in the number of predictive toxicity models. Datasets and models should not be considered in isolation. A more systematic and analytical metadata representation would help users to explore the relationships between datasets and models. This will lead to better management of information and knowledge captured from metadata and help users to choose the most appropriate model for a given task. For example, it would be interesting to investigate who has used which datasets previously and for what purpose, by analysing the captured information. As a result, the most predictive and popular models can be stored, managed and reused for future work. The most important aspect in this context is an extraction of relationships between a large number of objects (chemical compounds, datasets, models and users). This will lead to better management of information and knowledge captured from such predictive toxicology objects. Additionally, it will help the visualisation of the relationship among various objects and will support the utilisation of the existing information. Development of such a framework will support monitoring the model life cycle, automated model selection, chemical

compound identification across various projects, and can lead to speeding up the process of chemical compound virtual screening.

To achieve this, the following actions are foreseen:

- Represent data and models as programmable objects and provide standards for their representations. Existing dataset and model representation schemas (e.g. ToxML, QSAR-ML and PMML) can be employed and extended.
- Enable users to leave comments. In addition, each object (including dataset, model and user) can be associated with a wiki web link to keep notes and historical changes.
- Introduce a score rating schema which will allow users to rate the overall quality and suitability of the datasets and models that they have used. The storage of such metadata (e.g. rating scores and comments) would help to reduce duplication of effort and provide suggestions for subsequent users.
- Introduce a model version control system, which will allow models to be continuously updated whilst providing robust provenance of model predictions.

The authors believe that applying data governance in building information warehouses will provide a good start for data and model quality control. The analytical measurement of object quality, object similarity and object relationships monitoring will make such a framework more trustworthy and transparent for users and regulatory bodies. Standards in data and model representation will allow for effective object categorisation and consistent supporting documentation. This will lead to easy access to high quality information. It will also reduce the cost of information management, and secure the use of available data. Designing and developing a novel data and model governance framework is an important piece of future work. The main idea is to provide a common formatting system for data generation, extraction, curation, model estimation and validation. This will involve extension and unification of existing solutions that are accepted by regulatory bodies, and introduction of new standards in predictive toxicology. It also opens a wide area for various interesting research questions such as data provenance tracking, data and model quality measurements and the capture of object relationships.

## 23.7 Competing interests

The authors declare that they have no competing interests.

## 23.8 Authors' contributions

XF proposed the described data governance framework and contributed to the data sources review and the drafting of the manuscript. AW participated in analysis for database review and helped to prepare the manuscript for this publication. DN and MR firstly proposed the concept of data and model governance. DN, MR and KT have been involved in the review discussions and participated in revising critically this manuscript and proof read the draft. All authors read and approved the final version of the manuscript.

## 23.9 Acknowledgements and funding

The authors would like to thank Syngenta Ltd for partly sponsoring the Knowledge Transfer Partnerships (KTP) Grant (No. 7596) for XF and BBSRC Industrial CASE Studentship Grant (No. BB/H530854/1) for AW. The authors are also grateful to the referees for their invaluable and insightful comments that have helped to improve this work.

# PUBCHEM3D: SHAPE COMPATIBILITY FILTERING USING MOLECULAR SHAPE QUADRUPOLES

## 24.1 Abstract

### 24.1.1 Background

PubChem provides a 3-D neighboring relationship, which involves finding the maximal shape overlap between two static compound 3-D conformations, a computationally intensive step. It is highly desirable to avoid this overlap computation, especially if it can be determined with certainty that a conformer pair cannot meet the criteria to be a 3-D neighbor. As such, PubChem employs a series of pre-filters, based on the concept of volume, to remove approximately 65% of all conformer neighbor pairs prior to shape overlap optimization. Given that molecular volume, a somewhat vague concept, is rather effective, it leads one to wonder: can the existing PubChem 3-D neighboring relationship, which consists of billions of shape similar conformer pairs from tens of millions of unique small molecules, be used to identify additional shape descriptor relationships? Or, put more specifically, can one place an upper bound on shape similarity using other “fuzzy” shape-like concepts like length, width, and height?

### 24.1.2 Results

Using a basis set of 4.18 billion 3-D neighbor pairs identified from single conformer per compound neighboring of 17.1 million molecules, shape descriptors were computed for all conformers. These steric shape descriptors included several forms of molecular volume and shape quadrupoles, which essentially embody the length, width, and height of a conformer. For a given 3-D neighbor conformer pair, the volume and each quadrupole component ( $Q_x$ ,  $Q_y$ , and  $Q_z$ ) were binned and their frequency of occurrence was examined. Per molecular volume type, this effectively produced three different maps, one per quadrupole component ( $Q_x$ ,  $Q_y$ , and  $Q_z$ ), of allowed values for the similarity metric, shape Tanimoto (ST) 0.8.

The efficiency of these relationships (in terms of true positive, true negative, false positive and false negative) as a function of ST threshold was determined in a test run of 13.2 billion conformer pairs not previously considered by the 3-D neighbor set. At an ST 0.8, a filtering efficiency of 40.4% of true negatives was achieved with only 32 false negatives out of 24 million true positives, when applying the separate  $Q_x$ ,  $Q_y$ , and  $Q_z$  maps in a series ( $Q_{xyz}$ ). This efficiency increased linearly as a function of ST threshold in the range 0.8-0.99. The  $Q_x$  filter was consistently the most efficient followed by  $Q_y$  and then by  $Q_z$ . Use of a monopole volume showed the best overall performance, followed by the self-overlap volume and then by the analytic volume.

Application of the monopole-based  $Q_{xyz}$  filter in a “real world” test of 3-D neighboring of 4,218 chemicals of biomedical interest against 26.1 million molecules in PubChem reduced the total CPU cost of neighboring by between 24-38%

and, if used as the initial filter, removed from consideration 48.3% of all conformer pairs at almost negligible computational overhead.

### 24.1.3 Conclusion

Basic shape descriptors, such as those embodied by size, length, width, and height, can be highly effective in identifying shape incompatible compound conformer pairs. When performing a 3-D search using a shape similarity cut-off, computation can be avoided by identifying conformer pairs that cannot meet the result criteria. Applying this methodology as a filter for PubChem 3-D neighboring computation, an improvement of 31% was realized, increasing the average conformer pair throughput from 154,000 to 202,000 per second per CPU core.

## 24.2 Background

PubChem is an open and free resource of the biological activities of small molecules<sup>1234</sup>. PubChem has an integrated theoretical 3-D layer, PubChem3D<sup>567</sup>, which provides a precomputed 3-D neighboring relationship called “Similar Conformers”<sup>7</sup> to help users locate and relate data in the archive. “Similar Conformers” identifies chemicals with similar 3-D shape and similar 3-D orientation of functional groups typically used to define pharmacophores (defined here simply as “features”), complementing a PubChem 2-D neighboring relationship called “Similar Compounds”, which identifies closely related chemical analogs using the PubChem 2-D subgraph fingerprint<sup>8</sup>. Effectively, for each PubChem chemical structure, this 3-D neighboring relationship provides (at the time of writing) the results of a 3-D similarity search against 28.9 million compound records using three diverse conformers per molecule.

The PubChem3D neighboring uses as a measure of molecular shape similarity the shape Tanimoto (ST)<sup>910</sup>, given as the following equation:

where  $V_{AA}$  and  $V_{BB}$  are the self-overlap volumes of conformers A and B, respectively, and  $V_{AB}$  is the common overlap volume between A and B. The 3-D neighboring requires finding the maximum shape similarity between static compound 3-D conformations, as dictated by  $V_{AB}$  in **Equation 1**, to calculate ST, a computationally intensive step. It is highly desirable to avoid this overlap computation, especially if it can be determined with certainty that a conformer pair cannot meet the criteria to be a 3-D neighbor. As such, PubChem employs a series of filters, based on the concept of volume, to effectively ignore approximately 65% of all conformer neighbor pairs during 3-D neighboring, thus dramatically accelerating processing<sup>7</sup>.

Volume, although a rather fuzzy concept, is rather effective as a filter between conformers dissimilar in shape and features<sup>7</sup>. Conceivably there are other aspects of molecular shape beyond volume to “recognize” when two shapes are (dis)similar. A characteristic one can readily imagine are descriptors associated with aspects of length, width, and height. Steric shape quadrupoles embody such a concept and attempts have been made to use their differences as a shape similarity metric<sup>1112</sup>. This leads to the question: can additional simple shape descriptor relationships be identified that improve upon the volume-based filtering efficacy? Or, put another way, can one place an upper bound on shape similarity by identification of some (additional) crude shape compatibility between conformers?

In this paper, we examine the use of shape descriptors as a means to rapidly identify “dissimilar” molecule shapes. As a part of this, we attempt to answer the critical questions: are vague shape descriptors representing the concepts

<sup>1</sup> PubChem: integrated platform of small molecules and biological activities

<sup>2</sup> PubChem: a public information system for analyzing bioactivities of small molecules

<sup>3</sup> An overview of the PubChem BioAssay resource

<sup>4</sup> Database resources of the National Center for Biotechnology Information

<sup>5</sup> PubChem3D: conformer generation

<sup>6</sup> PubChem3D: diversity of shape

<sup>7</sup> PubChem3D: similar conformers

<sup>8</sup> PubChem substructure fingerprint description

<sup>9</sup> A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape

<sup>10</sup> A Gaussian description of molecular shape

<sup>11</sup> Gaussian shape methods

<sup>12</sup> Small molecule shape-fingerprints

of length, width, and height good discriminators of molecular shape? Can 3-D similarity searching speed be further accelerated using shape descriptors more sophisticated than volume? Is it possible to create a “shape compatibility” mapping indexed to shape similarity?

## 24.3 Results and Discussion

### 24.3.1 1. Distribution of shape descriptor components and their volume dependency

The molecular shape quadrupoles in the principal-axes frame<sup>9<sup>13</sup></sup> are given as the following:

where,  $x_Q$ ,  $y_Q$ , and  $z_Q$  are the  $x$ ,  $y$ , and  $z$  components of the quadrupole moment, respectively. The  $x$ ,  $y$ , and  $z$  components are conceptually equivalent to the length, width, and height of a molecule, respectively, with the largest quadrupole component defined as  $x_Q$  and the smallest as  $z_Q$ , by convention. An assumption underlying this study is that there is a point whereby, if the shape quadrupole difference between two conformers is too large, they cannot meet the ST 0.8 threshold required by PubChem3D neighboring, as illustrated in Figure 1. This relationship, if it actually exists, would allow conformer pairs to be filtered out, avoiding the time-consuming shape superposition optimization step for those pairs and enhancing the throughput of the PubChem 3-D neighboring. To attempt to determine if a relationship can be found, the shape quadrupole differences for known 3-D “Similar Conformers” were analyzed.

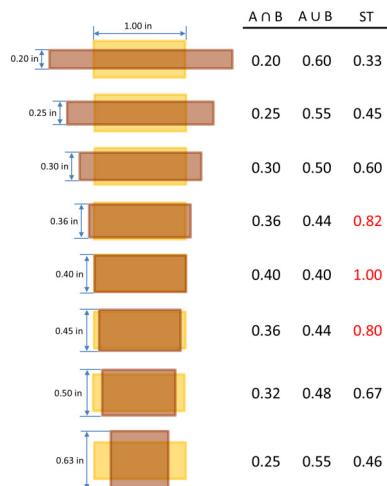


Figure 24.1: Figure 1. Small changes in dimensions can result in large changes in overlap

**Small changes in dimensions can result in large changes in overlap.** Using a 2-D rectangle shape with constant area ( $0.4 \text{ in}^2$ ), one can see that small changes in shape dimensions (length and width) can result in large changes in shape overlap (ST). Note that, for two shapes to be considered similar to each other (with a ST score of 0.8, indicated in red), the difference in length and width between them should be smaller than a certain threshold.

At the time of quadrupole filter project initiation (in October, 2008), 3-D neighboring of 17,143,181 unique molecules, effectively covering the CID range 1-25,000,000, had been completed using a single conformer per compound, yielding 4,182,412,802 3-D neighbors. Table 1 shows the statistics of the three quadrupole components for those 17.1 million molecules. The mean and standard deviation for  $x_Q$ ,  $y_Q$ , and  $z_Q$  were  $15.01 \pm 8.07 \text{ \AA}^5$ ,  $3.81 \pm 1.80 \text{ \AA}^5$ , and  $1.52 \pm 0.65 \text{ \AA}^5$ , respectively. Figure 2 and 3 display the distributions of  $x_Q$ ,  $y_Q$ , and  $z_Q$ , after they were binned into units of  $2.5 \text{ \AA}^5$ ,  $0.5 \text{ \AA}^5$ , and  $0.1 \text{ \AA}^5$ , respectively. All three components showed strongly skewed distributions; however, most of the molecules were populated near the mean and relatively few molecules had quadrupole components much larger than the mean values.

<sup>13</sup> A new class of molecular shape descriptors. 1. Theory and properties

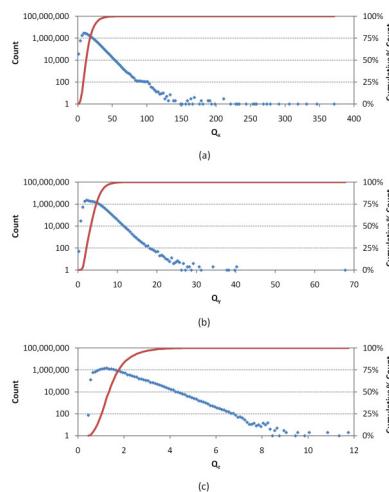


Figure 24.2: Figure 2. Quadrupole distribution

**Quadrupole distribution.** The frequency of occurrence of the three quadrupole moment components for 17.1 million molecules from the PubChem Compound database, where (a)  $xQ$ , (b)  $yQ$ , and (c)  $zQ$  were binned into units of  $2.5 \text{ \AA}^5$ ,  $0.5 \text{ \AA}^5$ , and  $0.1 \text{ \AA}^5$ , respectively.

The molecular volume and quadrupole moments are correlated with each other according to the following equation:

where  $R_g$  is the radius of gyration and  $V_{mp}$  is the monopole volume, which corresponds to the monopole in the shape multipole expansion<sup>13</sup>. **Equation 3** implies that the size of a molecule (represented by the molecular volume) is not completely independent of its quadrupole moment. Therefore, at the beginning of this study, the correlation between molecular volume and quadrupole moment was investigated. Note that, because the molecular volume is not a measurable quantity with a clear, unanimous definition, there are many ways to estimate it<sup>13 14 15 16 17 18</sup>. Therefore, in addition to the monopole volume, the PubChem 3-D information includes two other volumes computed in different ways. One is the analytic volume and the other is the self-overlap volume. The analytic volume is considered to be most consistent to other definitions of molecular volume among the three, but its computation is also the slowest. For this reason, evaluation of the ST score given in **Equation 1** uses the self-overlap volume, whose evaluation is considerably faster than the analytic volume; however, it typically overestimates the molecular volume by a factor of three greater than the analytic volume, as shown in Table 2. Each compound conformer record in the PubChem provides all three volumes and they can be downloaded: individually from the Compound Summary pages, using a list from the PubChem Download Facility ([http://pubchem.ncbi.nlm.nih.gov/pc\\_fetch](http://pubchem.ncbi.nlm.nih.gov/pc_fetch)), or in bulk from the PubChem FTP site (<ftp://ftp.ncbi.nlm.nih.gov/pubchem>). To avoid confusion about these three different volumes used in the present paper, we denote the monopole volume, self-overlap volume, and analytic volume as  $V_{mp}$ ,  $V_{so}$ , and  $V_{an}$ , respectively, whereas the volume in a general sense is denoted as  $V$  (without any subscript).

Figure 4 displays the distribution of the three different volumes of the 17.1 million molecules from the PubChem Compound database. In general,  $V_{so}$  is the largest, and  $V_{an}$  is the smallest. As shown in Figure 5, the quadrupole moment increases with molecular size, implying that the effect of quadrupole difference between two molecules upon their shape similarity may depend on their relative molecular sizes. Therefore, the quadrupole differences of 3-D “Similar Conformer” neighbors as a function of volume need to be considered.

<sup>14</sup> The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state organic-reactivity

<sup>15</sup> Molecular volume calculation using AM1 semi-empirical method toward diffusion coefficients and electrophoretic mobility estimates in aqueous solution

<sup>16</sup> Molecular volumes and Stokes-Einstein equation

<sup>17</sup> Partial molar volumes of ionic and nonionic organic solutes in water: a simple additivity scheme based on the intrinsic volume approach

<sup>18</sup> Correlation of computed van der waals and molecular volumes with apparent molar volumes (AMV) for amino-acid, carbohydrate and sulfate tantant molecules. Relationship between Corey-Pauling-Koltun volumes ( $V_{cpk}$ ) and computed volumes

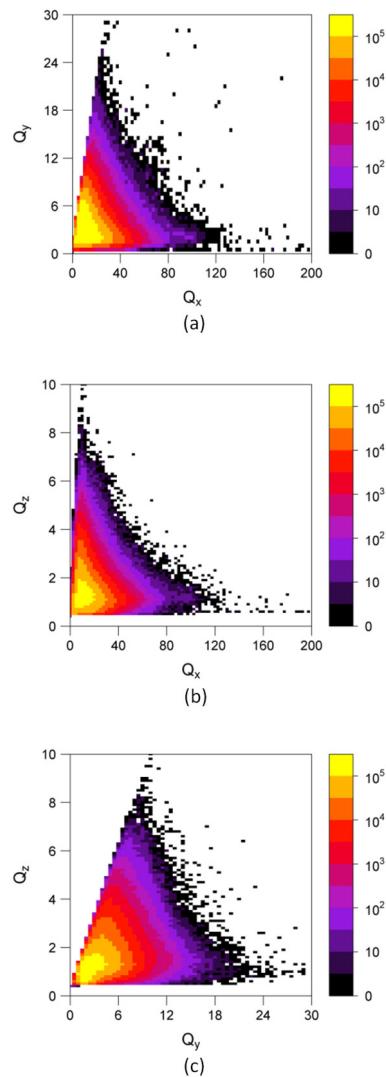


Figure 24.3: Figure 3. Quadrupole interdependence

**Quadrupole interdependence.** The distribution of 17.1 million molecules from the PubChem Compound database as a function of (a)  $x$  Qand  $y$ Q, (b)  $x$  Qand  $z$ Q, and (c)  $y$  Qand  $z$ Q, respectively.  $x$ ,  $y$ Q, and  $z$ Q were binned into units of  $2.5 \text{ \AA}^5$ ,  $0.5 \text{ \AA}^5$ , and  $0.1 \text{ \AA}^5$ , respectively. The legend indicates the frequency of observation.

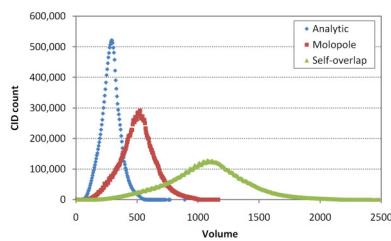


Figure 24.4: Figure 4. Volume distribution

**Volume distribution.** The frequency of occurrence of the three different volume types, analytic volume ( $_{an}V$ , blue), monopole volume ( $_{mp}V$ , red), and self-overlap volume ( $_{so}V$ , green), for 17.1 million molecules from the PubChem Compound database, where all three volumes were binned into units of  $5.0 \text{ \AA}^3$ .

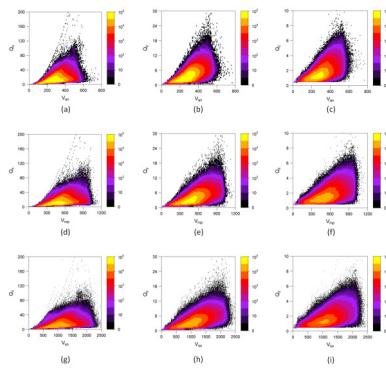


Figure 24.5: Figure 5. Volume-quadrupole interdependence

**Volume-quadrupole interdependence.** The distribution of 17.1 million molecules from the PubChem Compound database as a function of the molecular volume type and quadrupole component.  $V_1$ [in panel (a)-(c)],  $V_{\text{mp}}$ [in panel (d)-(f)], and  $V_{\text{so}}$ [in panel (g)-(i)] indicate the analytic volume, monopole volume, and self-overlap volume, respectively.  $x_Q$ [in panel (a), (d) and (g)],  $y_Q$ [in panel (b), (e) and (h)], and  $z_Q$ [in panel (c), (f) and (i)] indicate the three components of the quadrupole moment. All three volumes were binned into units of  $5.0 \text{ \AA}^3$  and the  $x_Q$ ,  $y_Q$ , and  $z_Q$  were binned into units of  $2.5 \text{ \AA}^5$ ,  $0.5 \text{ \AA}^5$ , and  $0.1 \text{ \AA}^5$ , respectively. The legend indicates the frequency of observation.

### 24.3.2 2. Design of 3-D neighbor filters using quadrupole moment differences

As a general premise, if two molecules with the same volume also have identical values for the quadrupole components, they are likely to be shape similar to each other. In addition, as the quadrupole moment difference deviates from zero, the maximum shape similarity is expected to decrease (see Figure 1). When the quadrupole (and volume) difference becomes greater than some value or threshold, the shape dissimilarity is such that the molecule conformer pair cannot possibly meet the criteria to be a PubChem 3-D neighbor (ST 0.8). Therefore, if we know these quadrupole difference thresholds for a given volume pair, one may be able to preclude conformer pairs that are not sufficiently shape similar, using only knowledge of the volume and quadrupole moments.

In the present study, the quadrupole moment differences of the 4.18 billion 3-D neighbors, identified from the 3-D neighboring of 17.1 million molecules, were analyzed to find the maximum possible quadrupole differences for two molecules to be neighbors (see also the “Materials and Methods” section). The volume and quadrupole moments of the two molecules in each neighbor pair were first converted into an integer value by using the following two equations:

where superscript “bin” is used to distinguish these integers from the original, non-binned values. The denominator *Binsize* was  $5.0 \text{ \AA}^3$  for all the three volumes, and  $2.5 \text{ \AA}^5$ ,  $0.5 \text{ \AA}^5$ , and  $0.1 \text{ \AA}^5$ , for  $x_Q$ ,  $y_Q$ , and  $z_Q$ , respectively. After all 4.18 billion 3-D neighbors were binned according to their  $V_1^{\text{bin}}$  and  $V_2^{\text{bin}}$  values, the 3-D neighbor distribution for a given  $(V_1^{\text{bin}}, V_2^{\text{bin}})$  was calculated.

To illustrate the general premise above that quadrupole deviations from zero result in a reduction in shape similarity, Figure 6 shows the neighbor count as a function of  $\Delta Q_z^{\text{bin}}$  with respect to the ordinate axis (

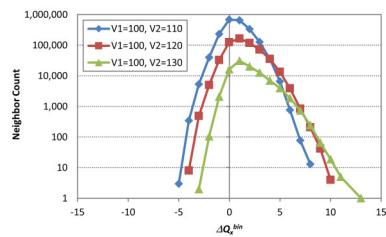


Figure 24.6: Figure 6. Quadrupole difference tolerance

**Quadrupole difference tolerance.** The distributions of the 3-D neighbors as a function of the binned quadrupole differences,  $\Delta Q_z^{\text{bin}}$  rapidly decreases to zero as a function of magnitude.

Figures 7, 8 and 9 show the  $\Delta^{\text{bin}}$  Qthreshold for each quadrupole component as a function of volume for the 4.18 billion 3-D neighbors. Note that, since PubChem regularly gets additional new unique content from its contributors, there is always a possibility that the 3-D neighboring of these new records may identify previously unseen cases of  $\Delta^{\text{bin}}$  Qthreshold. If we use these  $\Delta^{\text{bin}}$  Qthreshold maps [see panels (a) and (b) of Figures 7, 8 and 9] as a filter during neighboring, we would preclude those 3-D neighbors. Therefore, we modified the maps [see panels (c) and (d) of Figures 7, 8 and 9], as described in the “Materials and Methods” section, to extend  $\Delta^{\text{bin}}$  Qdifference values or to add neighboring bins where no population is found in an attempt to mitigate any such issues in the fringe regions on the maps.

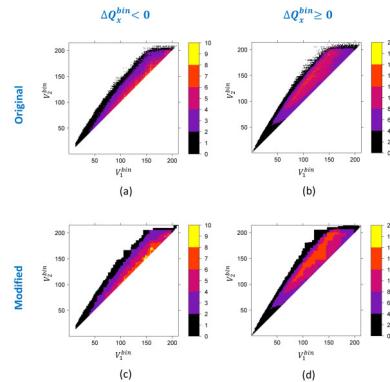


Figure 24.7: Figure 7. Monopole volume  
**Monopole volume**. The absolute value of the maximum possible value of

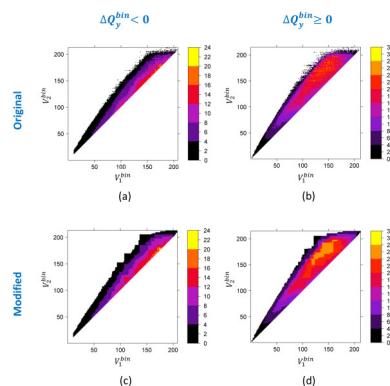


Figure 24.8: Figure 8. Monopole volume  
**Monopole volume**. The absolute value of the maximum possible value of

These modified  $\Delta^{\text{bin}}$  Qthreshold maps are designated as quadrupole filters. For simplicity, we name these filters with a capital letter “F” followed by a subscript, which represents one of the quadrupole components, and a superscript, which represents the type of volume involved. For example, filter “ $x$ ” Qfilter generated with the analytic volume,  $v_n$ .

Given that these quadrupole filters were built using an existing set of 3-D neighbor cases, one needs to validate the extent of their efficacy. To do so, a 13.2 billion molecule conformer pair test set not considered as a part of the original 3-D neighboring training set, is utilized (see the “Materials and Methods” section). After computing the ST scores for the 13.2 billion pairs, the fraction of 3-D neighbors and non-neighbors, which would have been pre-screened if the quadrupole filters were applied, is summarized in Table 3.

Of the three volume types utilized, the monopole-based quadrupole filters,  $F^{\text{mp}}$ , is arguably the best. Filter *feature* similarity as well as *shape* similarity, while the quadrupole filters deal only with shape similarity. As such, the 30 pairs

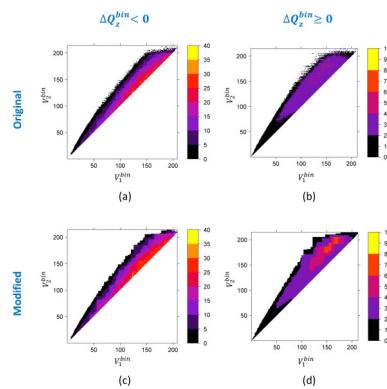


Figure 24.9: Figure 9. Monopole volume  
**Monopole volume**. The absolute value of the maximum possible value of

filtered out had a ST score sufficient to be a 3-D neighbor, making it a “potential” 3-D neighbor.] The false negative count of 30 removed by

Filters  $\mathbf{F}^{\text{mp}}$  filters are used in a series (denoted as  $\mathbf{F}^{\text{so}}$  showed similar performance to  $\mathbf{F}^{\text{mp}}$ , but it filtered out more potential neighbors (288 for  $\mathbf{F}^{\text{an}}$  filters showed the least loss of potential neighbors (4 for

Effects of the ST threshold for PubChem 3-D neighboring upon the efficiency of the quadrupole filters were also investigated by generating a set of quadrupole filters, each using a different ST threshold, ranging from 0.80 to 0.99 with an increment of 0.01. As shown in Figure 10, the fraction of molecule pairs filtered increases almost linearly as a function of the ST threshold. For the entire ST range tested, the

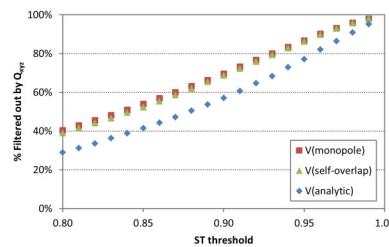


Figure 24.10: Figure 10. Shape compatibility filtering efficiency  
**Shape compatibility filtering efficiency**. Performance of the  $\mathbf{F}_{\text{xyz}}$  quadrupole filter to filter conformer pairs at different ST threshold values.

### 24.3.3 3. Application of 3-D neighbor filters using quadrupole moment differences

Given that filtering conformer pairs using steric shape quadrupoles is effective with minimal loss of potential 3-D neighbors, a “real world” test is made with<sup>7</sup> whereby a set of known drugs and other molecules of keen biomedical interest are neighbored against the 3-D contents of PubChem. Table 4 and Table 5 summarize the results of these tests.

Considering PubChem 3-D neighboring is a precomputed similarity search, one can see that the neighboring throughput improvements using

It is important to note that 5, meaning that there are three other filters utilized before 5. The CT Feature, CT Volume, and ST Volume filters, applied in that order, remove 27.9%, 0.1%, and 0.002% conformer pairs, respectively, when

## 24.4 Conclusion

Simple molecular shape descriptors, volume and steric quadrupole moments (embodying the length, width, and height of a shape), of 4.18 billion 3-D neighbor pairs resulting from PubChem 3-D neighboring of 17.1 million single conformer molecules were analyzed. The maximum quadrupole differences between neighbor conformers were determined. This examination demonstrated a distinct dependency of shape similarity upon quadrupole variation. With some slight modification of fringe regions, the results of this analysis were turned into computationally inexpensive, yet highly effective set of filters capable of removing 3-D conformer pairs that cannot meet a required shape similarity, using only knowledge of the volume and steric quadrupole moments of the conformer pair. When applied in the context of shape similarity searching, these filters can significantly improve throughput performance by avoiding expensive superposition optimization computation of conformer pairs that cannot possibly meet a pre-defined shape similarity search threshold.

The filters devised were tested using a dataset of 13.2 billion compound pairs. The quadrupole filters based on a monopole volume showed the best efficacy, while the filters using an analytic volume had the lowest efficacy. For all the three volume types, the  $x$  Qfilters eliminated a larger portion of the compound pairs than the  $y$  Qand  $z$  Qfilters. When the filters were used in a series simultaneously, they could eliminate 30~40% of non-neighbor pairs, with the removal of a negligible amount of potential neighbors. For example, the  $xyz$  Qfilter based on the monopole volumes (

In summary, the quadrupole filters developed in this study can speed up the PubChem 3-D neighbor processing with a negligible loss of the 3-D neighbors. However, its applicability is not just limited to PubChem 3-D neighboring. The results of the present study also suggest that the shape multipole moments can be applied generally to enhance the speed of 3-D similarity search methods by the rapid preclusion of dissimilar molecules that cannot be a result. This approach may be able to significantly speed up 3-D similarity search, especially if the 3-D shape superposition optimization is a bottleneck of the similarity search.

## 24.5 Materials and methods

### 24.5.1 1. Datasets

At the time of project initiation, PubChem 3-D neighboring of 17,143,181 unique molecules (ranging from CID 1 to CID 25,000,000) had been completed using a single conformer per compound, yielding 4,182,412,802 3-D neighbors. Using the Shape Toolkit from the OpenEye Software<sup>19</sup>, the analytic volume ( $_{an}V$ ), monopole volume ( $_{mp}V$ ), self-overlap volume ( $_{so}V$ ), and steric shape quadrupole moments ( $x$ ,  $Q_yQ$ , and  $zQ$ ) were computed for the theoretical conformer of all 17.1 million molecules. See Figures 2 and 4 for the distributions of the computed values.

### 24.5.2 2. Filter generation

The quadrupole filters developed for pre-screening conformer-pairs based on quadrupole differences as a function of shape similarity ST threshold were generated using the following steps:

1. The 4.18 billion 3-D neighbor pairs and their associated data were obtained from PubChem.
2. The volumes ( $_{mp}V$ ,  $_{so}V$ , and  $_{an}V$ ) and quadrupole components ( $x$ ,  $Q_yQ$ , and  $zQ$ ) of the compound pair for each 3-D neighbor were converted into integers using Equations 4 and 5 to yield *BinSize* was 5.0 Å<sup>3</sup> for all three volume types and 2.5 Å<sup>5</sup>, 0.5 Å<sup>5</sup>, and 0.1 Å<sup>5</sup>, for  $x$ ,  $Q_yQ$ , and  $zQ$ , respectively.
3. For each of the three binned volume types, the following was performed using the 3-D neighbor pairs (in this case using
  1. Of the two conformers in a 3-D neighbor, the one with the smaller

<sup>19</sup> ShapeTK-C++

2. For each of the three binned quadrupole components, and using
  1. 3-D neighbors were binned according to three indices,  $x$  Qdifference between the two molecules.
  2. The neighbor count for all [7, 8 and 9].
  3. The  $x$  Qfilter based on monopole volumes (7, 8 and 9).
  4. To obtain filters effective at an ST threshold other than 0.80, first restrict the original 4.18 billion 3-D neighbor pairs to those at or above the desired ST threshold and repeat step 3.

### 24.5.3 3. Modification of filters

Figure 11 shows a schematic diagram describing how an original difference map is modified at a given  $\Delta^{\text{bin}}$  Qvalue. In an original map [panel (a) of Figure 11], the (11) at the given  $\Delta^{\text{bin}}$  Qvalue. Similarly, any empty bins within the range of 11 for the given  $\Delta^{\text{bin}}$  Qvalue.

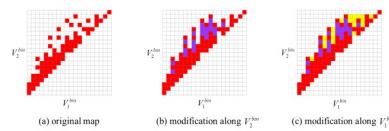


Figure 24.11: Figure 11. Transformation of shape compatibility map into a filter

**Transformation of shape compatibility map into a filter.** Schematic diagram describing modification of an original difference map at a given  $\Delta^{\text{bin}}$  Qvalue: (a) in an original map, neighbor regions are indicated in red, (b) all empty bins between the minimum and maximum values of

This procedure is performed for all unique  $\Delta^{\text{bin}}$  Qvalues starting with the maximum. As lesser  $\Delta^{\text{bin}}$  Qvalues are considered in this correction, greater  $\Delta^{\text{bin}}$  Qvalues are considered at the  $\Delta^{\text{bin}}$  Qvalue being considered. A pseudo-code implementation of this procedure is shown in Figure 12. All quadrupole filters resulting from this modification are available in Additional file

```

Load original map.
min_qdiff = minimum of quadrupole differences.
max_qdiff = maximum of quadrupole differences.

/* Loop over quadrupole difference */
for (qdiff=max_qdiff; qdiff=min_qdiff; qdiff--) {
    min_v1_at_qdiff = minimum of v1 in the neighbor region at qdiff.
    max_v1_at_qdiff = maximum of v1 in the neighbor region at qdiff.
    min_v2_at_qdiff = minimum of v2 in the neighbor region at qdiff.
    max_v2_at_qdiff = maximum of v2 in the neighbor region at qdiff.

    /* Overlay the neighbor region at qdiff to that at qdiff */
    for (i = 0; i < stop; i++) {
        for (j = 0; j < stop; j++) {
            if (neighbor is found for (i,j) at any quadrupole difference bigger than qdiff,
                set the (i,j) pair to neighbor region.
        }
    }

    /* Modify the neighbor region until there is no change in the map. */
    while (1) stop == 0 || {
        /* determine the range of neighbor region */
        for (i from min_v1_at_qdiff to max_v1_at_qdiff) {
            min_v1_at_v1 = minimum of v1 in the neighbor region at v1
            max_v1_at_v1 = maximum of v1 in the neighbor region at v1
        }

        for (j from min_v2_at_qdiff to max_v2_at_qdiff) {
            min_v2_at_v2 = minimum of v2 in the neighbor region at v2
            max_v2_at_v2 = maximum of v2 in the neighbor region at v2
        }

        /* Add additional neighbor bins (in the fringe regions) */
        for (i from min_v1_at_qdiff to max_v1_at_qdiff) {
            for (j from min_v2_at_qdiff to max_v2_at_qdiff) {
                set all bins b/w (min_v1_at_v1, i), (max_v1_at_v1, i) to neighbor region at qdiff.
                set all bins b/w (min_v2_at_v2, j), (max_v2_at_v2, j) to neighbor region at qdiff.
            }
        }

        if (no changes in the map) { stop = 1 }
    }
}

```

Figure 24.12: Figure 12. Pseudo code to transform shape compatibility map into a filter

**Pseudo code to transform shape compatibility map into a filter.**

Additional file 1

**Quadrupole filters.** A zip archive of text files containing information on the maximum quadrupole differences as a function of molecular volumes.

[Click here for file](#)

#### 24.5.4 4. Efficiency test of filters

To test the efficiency of the quadrupole filters devised, two sets of molecules were chosen. One set contains molecules in the PubChem CID range of 1 ~ 25,000,000, and the other contains those in the CID range of 25,000,001~25,001,000. Because a theoretical conformer was not generated for all CIDs or because compound records were not “live”, the two datasets had 17,488,897 and 753 molecules, respectively. All-by-all comparison between the two sets gives 13,169,139,441 CID pairs. Using the first diverse conformer for each compound, the ST values for these 13.2 billion pairs were computed using ROCS<sup>20</sup> from OpenEye software, Inc., consuming ~419 CPU days in total, and stored. These ST scores were used to estimate how many CID pairs would be filtered out when applying the quadrupole filters as a function of volume type and as a function of ST threshold, for example, as demonstrated in Table 3 and Figure 10.

#### 24.5.5 5. Effect of Quadrupole filters on PubChem3D Neighboring

One aspect of this effort is to examine the change in real-world efficiency of PubChem3D neighboring processing when using quadrupole filters while computing the 3-D “Similar Conformers” relationship. To achieve this, the set of 4,218 biologically relevant chemical structures with known pharmacological actions from our earlier efforts<sup>7</sup> was used. These small molecules with known biological action (*Query set*) were neighbored against 26,157,365 compound records (*Search set*), representing the entire “live” PubChem3D contents as of Oct. 2010, using up to 1, 3, 5, 7, and 10 diverse conformers per compound for both compound sets. Timing and efficiency differences with our earlier work are given in Tables 4 and 5.

### 24.6 Competing interests

The authors declare that they have no competing interests.

### 24.7 Authors' contributions

SK analyzed the quadrupole differences of the 3-D neighbors, generated the quadrupole filters, and wrote the first draft. EEB supervised the project and revised manuscript. SHB reviewed the final manuscript. All authors read and approved the final manuscript.

### 24.8 Acknowledgements

We are grateful to the NCBI Systems staff, especially Ron Patterson, Charlie Cook, and Don Preuss, whose efforts helped make the PubChem3D project possible. This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, U.S. Department of Health and Human Services.

<sup>20</sup> ROCS - Rapid Overlay of Chemical Structures



# PUBCHEM3D: BIOLOGICALLY RELEVANT 3-D SIMILARITY

## 25.1 Abstract

### 25.1.1 Background

The use of 3-D similarity techniques in the analysis of biological data and virtual screening is pervasive, but what is a biologically meaningful 3-D similarity value? Can one find statistically significant separation between “active/active” and “active/inactive” spaces? These questions are explored using 734,486 biologically tested chemical structures, 1,389 biological assay data sets, and six different 3-D similarity types utilized by PubChem analysis tools.

### 25.1.2 Results

The similarity value distributions of 269.7 billion unique conformer pairs from 734,486 biologically tested compounds (all-against-all) from PubChem were utilized to help work towards an answer to the question: what is a biologically meaningful 3-D similarity score? The average and standard deviation for the six similarity measures  $ST^{**}ST\text{-opt:sup:}\backslash$ ,  $*CT^{**}ST\text{-opt}\star\backslash\text{:sup:}$ ,  $ComboT^{**}ST\text{-opt:sup:}\backslash$ ,  $*ST^{**}CT\text{-opt}\star\backslash\text{:sup:}$ ,  $CT^{**}CT\text{-opt:sup:}\backslash$ , and  $*ComboT^{**}CT\text{-opt}\star\backslash\text{:sup:}$  were  $0.54 \pm 0.10$ ,  $0.07 \pm 0.05$ ,  $0.62 \pm 0.13$ ,  $0.41 \pm 0.11$ ,  $0.18 \pm 0.06$ , and  $0.59 \pm 0.14$ , respectively. Considering that this random distribution of biologically tested compounds was constructed using a single theoretical conformer per compound (the “default” conformer provided by PubChem), further study may be necessary using multiple diverse conformers per compound; however, given the breadth of the compound set, the single conformer per compound results may still apply to the case of multi-conformer per compound 3-D similarity value distributions. As such, this work is a critical step, covering a very wide corpus of chemical structures and biological assays, creating a statistical framework to build upon.

The second part of this study explored the question of whether it was possible to realize a statistically meaningful 3-D similarity value separation between reputed biological assay “inactives” and “actives”. Using the terminology of noninactive-noninactive (NN) pairs and the noninactive-inactive (NI) pairs to represent comparison of the “active/active” and “active/inactive” spaces, respectively, each of the 1,389 biological assays was examined by their 3-D similarity score differences between the NN and NI pairs and analyzed across all assays and by assay category types. While a consistent trend of separation was observed, this result was not statistically unambiguous after considering the respective standard deviations. While not all “actives” in a biological assay are amenable to this type of analysis, *e.g.*, due to different mechanisms of action or binding configurations, the ambiguous separation may also be due to employing a single conformer per compound in this study. With that said, there were a subset of biological assays where a clear separation between the NN and NI pairs found. In addition, use of combo Tanimoto (ComboT) alone, independent of superposition optimization type, appears to be the most efficient 3-D score type in identifying these cases.

### 25.1.3 Conclusion

This study provides a statistical guideline for analyzing biological assay data in terms of 3-D similarity and PubChem structure-activity analysis tools. When using a single conformer per compound, a relatively small number of assays appear to be able to separate “active/active” space from “active/inactive” space.

## 25.2 Background

Recent advances in combinatorial chemistry<sup>123456</sup> and high-throughput screening technology<sup>7891011121314151617</sup> have made the synthesis and screening of diverse chemical compounds easier, helping to create a demand in the biomedical research community for archives of publicly available screening data. To help satisfy this demand, the U.S. National Institutes of Health launched the PubChem project (<http://pubchem.ncbi.nlm.nih.gov>)<sup>18192021</sup> as a part of its Molecular Libraries Roadmap Initiative. PubChem archives contributed biological screening data and chemical information from various data sources in academia and industry, and offers its contents free of charge to biomedical researchers, helping to facilitate scientific discovery.

PubChem consists of three primary databases: Substance, Compound, and BioAssay. While the PubChem Substance database (unique identifier SID) contains information provided by individual depositors, the PubChem Compound database (unique identifier CID) contains the unique standardized chemical structure contents extracted from the PubChem Substance database. PubChem provides various analysis tools to relate chemical structures to the biological activity data stored in the PubChem BioAssay database (unique identifier AID).

The PubChem3D project<sup>22232425</sup>, launched, in part, to help users identify useful structure-activity relationships, generates a theoretical 3-D conformer model<sup>2223</sup> for each molecule in the PubChem Compound database, whenever it is possible. An all-against-all 3-D neighboring relationship (known as “Similar Conformers”)<sup>24</sup> is pre-computed to help users to locate related data in the archive, augmenting the complementary “Similar Compounds” relationship, based on 2-D similarity of the PubChem subgraph binary fingerprint<sup>26</sup>.

PubChem3D uses two 3-D similarity measures: shape-Tanimoto (ST)<sup>2427282930</sup> and color-Tanimoto (CT)<sup>242728</sup>. The

---

<sup>1</sup> From combinatorial chemistry to cancer-targeting peptides

<sup>2</sup> Recent advances in combinatorial chemistry applied to development of anti-HIV drugs

<sup>3</sup> Dynamic combinatorial chemistry

<sup>4</sup> The interplay between structure-based design and combinatorial chemistry

<sup>5</sup> The synergy between combinatorial chemistry and high-throughput screening

<sup>6</sup> Combinatorial chemistry: oh what a decade or two can do

<sup>7</sup> High-throughput electrophysiology: an emerging paradigm for ion-channel screening and physiology

<sup>8</sup> High-throughput screening assays for the identification of chemical probes

<sup>9</sup> High-throughput RNAi screening in cultured cells: a user’s guide

<sup>10</sup> Statistical practice in high-throughput screening data analysis

<sup>11</sup> Integration of virtual and high-throughput screening

<sup>12</sup> Enzyme assays for high-throughput screening

<sup>13</sup> Flow cytometry for high-throughput, high-content screening

<sup>14</sup> Electrospray ionization tandem mass spectrometry in high-throughput screening of homogeneous catalysts

<sup>15</sup> High-throughput screening: new technology for the 21st century

<sup>16</sup> High-throughput screening in drug metabolism and pharmacokinetic support of drug discovery

<sup>17</sup> High-throughput and ultra-high-throughput screening: solution- and cell-based approaches

<sup>18</sup> PubChem: integrated platform of small molecules and biological activities

<sup>19</sup> PubChem: a public information system for analyzing bioactivities of small molecules

<sup>20</sup> An overview of the PubChem BioAssay resource

<sup>21</sup> Database resources of the National Center for Biotechnology Information

<sup>22</sup> PubChem3D: conformer generation

<sup>23</sup> PubChem3D: diversity of shape

<sup>24</sup> PubChem3D: similar conformers

<sup>25</sup> PubChem3D: shape compatibility filtering using molecular shape quadrupoles

<sup>26</sup> PubChem substructure fingerprint description

<sup>27</sup> ROCS - Rapid Overlay of Chemical Structures

<sup>28</sup> ShapeTK-C++

<sup>29</sup> A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape

<sup>30</sup> A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction

ST score is a measure of shape similarity, which is defined as the following:

where  $V^{**AA} :sub:\backslash$  and  $*V^{**BB*} \backslash :sub:$  are the self-overlap volume of conformers A and B and  $V^{**AB} :sub:$ <sup>31</sup> is the common overlap volume between them. The CT score, given by **Equation (2)**, quantifies the similarity of 3-D orientation of functional groups used to define pharmacophores (henceforth referred to simply as “features”) between conformers by checking the overlap of fictitious “color” atoms<sup>28</sup> used to represent the six functional group types: hydrogen-bond donors, hydrogen-bond acceptors, cation, anion, hydrophobes, and rings.

where, the index “ $f$ ” indicates any of the six independent fictitious feature atom types,  $f$  and  $f$ . The ST and CT scores range between 0 (for no similarity) and 1 (for identical molecules). These similarity metrics can be combined to create a Combo-Tanimoto (ComboT), as specified by **Equation (3)**:

The ST and CT similarity metrics attempt to cover key aspects important for locating chemical structures that may have similar biological activity. ST helps to identify molecules that can adopt a particular 3-D shape, *e.g.*, of an inhibitor bound in a particular conformational orientation in a protein binding pocket. Considering that a hydrocarbon and a drug molecule could adopt the same shape, CT helps to identify molecules with similar 3-D orientation of features, *e.g.*, necessary for making binding interactions between a small molecule and protein binding pocket. This suggests that two molecules with highly similar 3-D shape and 3-D feature orientations may also have similar biological activity. It should be no small wonder that such similarity metrics have garnered widespread use in virtual screening<sup>31,32</sup>. It leads one to wonder: what is a statistically meaningful 3-D similarity score? Or, in other words, if one was to examine the 3-D similarities between biologically tested compounds, what does the distribution look like? In the case of 2-D similarity, one only needs the molecule graph to make a comparison but, in the case of 3-D similarity, molecules can potentially adopt a number of different conformations. Is it sufficient to use only a single conformer per compound and still realize a statistically meaningful difference or separation between the 3-D similarities of reputed actives and inactives from a biological test?

In the present paper, two important questions concerning ST, CT, and ComboT as 3-D similarity measures are investigated. The first question is “if we randomly select any two conformers from the PubChem Compound database, what values of ST, CT, and ComboT scores will be expected on the average?” With knowledge of these values, one can evaluate a statistical significance of the similarity score between any two conformers in PubChem (*e.g.*, if their similarity score becomes greater than what one expects for a random conformer pair, it may be statistically more meaningful).

The second question we seek to answer in this study is “for a given bioassay in PubChem, what is the average difference in similarity scores between the noninactive-noninactive (NN) pairs and the noninactive-inactive (NI) pairs, when a single conformer per compound is used for 3-D similarity computation?” The choice of terminology of NN and NI are necessary considering that the definition of an “active” is not always specified in PubChem. Therefore, for the purposes of this study, we consider “active space” to be anything not specified to be “inactive”, thus the term “noninactive” is used in place of “active”. This may help provide users with an idea on the separation in the 3-D shape and feature spaces between the active and inactive compounds tested in a given bioassay. An additional question we will answer is: does an optimization type affect the similarity scores? Currently, the PubChem 3-D neighboring involves a shape superposition optimization that maximizes the ST scores<sup>24</sup>, but it may be possible to optimize a feature superposition that maximizes the CT score. Will the ST-optimization and CT-optimization make any changes in a 3-D similarity-based bioassay data analysis?

## 25.3 Results and Discussion

### 25.3.1 A. Notations

In the present study, we consider six different similarity measures: ST, CT, and ComboT for two different optimization types (either ST-optimized or CT-optimized). They are denoted with a superscript, which represents the optimization type (either “ST-opt” or “CT-opt”), and a subscript, which specifies the type of CID pairs (“NN” for the NN pairs and

<sup>31</sup> Molecular shape and medicinal chemistry: a perspective

<sup>32</sup> Comparison of topological, shape, and docking methods in virtual screening

“NI” for the NI pairs). The subscript “NN-NI” is used for the similarity score difference between the NN and NI pairs. For example, *i.e.*, *ST*, *CT*, and *ComboT*), or a similarity score in a general sense.

In the second part of this study, we analyze the average and standard deviation of the similarity scores of *CID pairs for a given AID*, and these per-AID average and standard deviation are denoted with Greek letters  $\mu$  and  $\sigma$ , respectively, followed by the corresponding similarity measure in parentheses [*e.g.*,

where *XT* is one of the six similarity measures (*i.e.*, *ST\*\*ST-opt:sup:\*, *\*CT\*\*ST-opt\*\* :sup:, *ComboT\*\*ST-opt:sup:\*, *\*ST\*\*CT-opt\*\* :sup:, *CT\*\*CT-opt:sup:\*, and *\*ComboT\*\*CT-opt\*\* :sup:), and *n\*\*NN :sub:\* and *\*n\*\*NI\* \ :sub:* are the number of the NN pairs and NI pairs for the AID, respectively. When we refer to the average and standard deviation of the per-AID statistical parameters *over a set of AIDs*, we use additional Greek letters  $\mu$  and  $\sigma$ , respectively, followed by the corresponding statistical parameter in brackets. For example,

## 25.3.2 B. 3-D similarity score distribution of random conformer pairs

### B-1. Structural and chemical characteristics of the biologically tested molecules

As of January 2010, the PubChem BioAssay database had 2,008 bioassay records, (ranging from AID 1 to AID 2310) and 734,486 molecules with a 3-D conformer model were tested in at least one of these bioassays. The structural and chemical characteristics of these biologically tested molecules are shown in Figures 1, 2 and 3, and they are compared with those of the entire PubChem3D contents (26,157,365 CIDs as of September 2010) in Table 1. The average and standard deviation of the heavy atom count per-CID are  $24.6 \pm 6.4$ , slightly less than those across the entire PubChem3D contents ( $26.3 \pm 7.0$ ). The conformer monopole volume (V) and three components of the shape quadrupole moments ( $Q_x$ ,  $Q_y$ , and  $Q_z$ , which give a sense of the conformer length, width, and height dimensions, respectively)<sup>25</sup> of the biologically tested molecules default conformer are also slightly less than those across the entire PubChem3D contents ( $474.1 \pm 124.0 \text{ \AA}^3$  vs.  $509.0 \pm 137.1 \text{ \AA}^3$  for V,  $12.6 \pm 7.0 \text{ \AA}^5$  vs.  $13.6 \pm 7.8 \text{ \AA}^5$  for  $Q_x$ ,  $3.3 \pm 1.6 \text{ \AA}^5$  vs.  $3.6 \pm 1.8 \text{ \AA}^5$  for  $Q_y$ ,  $1.3 \pm 0.6 \text{ \AA}$  vs.  $1.5 \pm 0.6 \text{ \AA}^5$  for  $Q_z$ ). As shown in Figure 1(b) and Table 1, the 734,486 biologically tested molecules have  $8.1 \pm 2.6$  features on average, slightly less than the entire PubChem3D contents does ( $8.5 \pm 2.7$ ). The count for each of the six feature types of the biologically tested molecules is equal to or slightly less than those of the entire PubChem3D contents.

### B-2. Distribution of 3-D similarity scores for biologically tested molecules

One key question this study attempts to answer is: what are statistically meaningful 3-D similarity values for biologically tested molecules? By using the entire set of 734,486 biologically tested molecules in PubChem (as of late January 2010) and their 269,734,474,855 unique CID pairs, we believe this to be a sufficient corpus to make such a determination in a general sense. What may be questionable (*to some*) is the intention to use only a single conformer per compound for each of the CID pairs.

The reasons for this choice are rather practical. The use of two diverse conformers per compound yields four times more unique conformer pairs and using three diverse conformers per compound makes the unique conformer pair set nine times larger and so on. In other words, the problem size scales as a square of the conformers per compound considered. We could sample the 734,486 compounds into a smaller set, to say ten percent of the original dataset and then consider three diverse conformers per compound to yield approximately the same count of conformer pairs, but are three diverse conformers per compound sufficient? If we down sampled to 1% of the biologically tested compounds and used ten diverse conformers per compound, would ten diverse conformers per compound be sufficient and would the random 1% of the compound set be sufficient to represent biologically tested compounds? For the purposes of this study, we will ignore the multiple conformer representation issue and consider a single conformer per compound to be sufficiently random to provide a useful set of statistically meaningful 3-D similarity thresholds; however, a more detailed study may be necessary to determine the full effect of using multiple conformers per compound, *e.g.*, when picking the best conformer pair per compound pair.

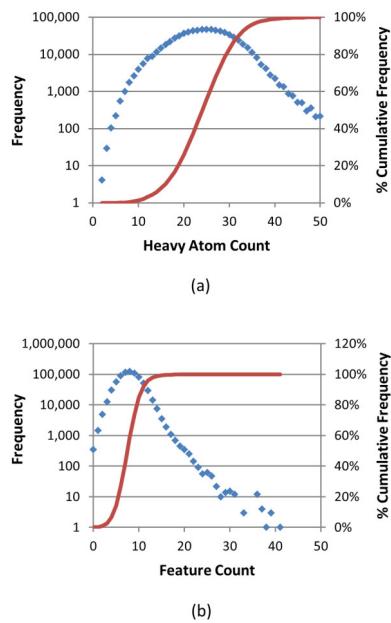


Figure 25.1: Figure 1. Atom and feature count histograms of biologically tested compounds

**Atom and feature count histograms of biologically tested compounds.** Frequency (blue) and percent cumulative frequency (red) of (a) heavy atom count and (b) total feature count for the 734,486 molecules tested in at least one bioassay archived in the PubChem BioAssay database.

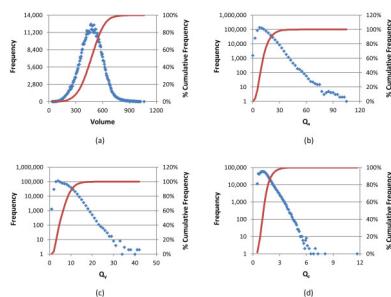


Figure 25.2: Figure 2. Conformer volume and quadrupole histograms of biologically tested compounds

**Conformer volume and quadrupole histograms of biologically tested compounds.** Frequency (blue) and percent cumulative frequency (red) of (a) volume, (b)  $Q_x$ , (c)  $Q_y$ , and (d)  $Q_z$  for the 734,486 molecules tested in at least one bioassay archived in the PubChem BioAssay database.

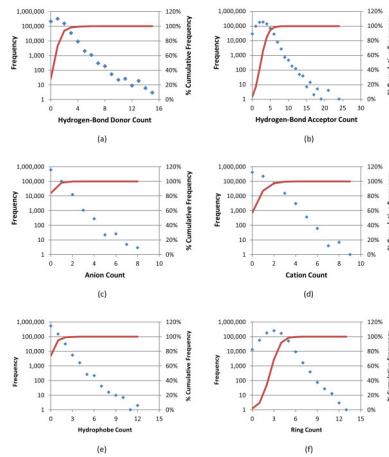


Figure 25.3: Figure 3. Individual feature histograms of biologically tested compounds

**Individual feature histograms of biologically tested compounds.** Frequency (blue) and percent cumulative frequency (red) of respective feature atom count for the 734,486 molecules tested in at least one bioassay archived in the PubChem BioAssay database: (a) hydrogen-bond donor count, (b) hydrogen-bond acceptor count, (c) anion count, (d) cation count, (e) hydrophobe count, and (f) ring count.

To investigate the average values of ST, CT, and ComboT for random conformer pairs, we downloaded all 734,486 biologically tested molecules from PubChem that had a theoretical 3-D description, and the six similarity scores [*i.e.*, `ST**ST-opt:sup:\`, `*CT**ST-opt*\ :sup:\`, `ComboT**ST-opt:sup:\`, `*ST**CT-opt*\ :sup:\`, `CT**CT-opt:sup:\`, and `*ComboT**CT-opt*\ :sup:\`] were computed for all 269,734,474,855 unique CID pairs arising from all possible combination of the 734,486 CIDs, using a single conformer per-CID. The distribution of these scores represents the 3-D similarity scores one would get from any two conformers randomly selected from the PubChem database. The distributions of the similarity scores, binned in 0.01 increments, are shown in Figure 4 and their statistics are summarized in Table 2. The average and standard deviation for `ST**ST-opt:sup:\`, `*CT**ST-opt*\ :sup:\`, `ComboT**ST-opt:sup:\`, `*ST**CT-opt*\ :sup:\`, `CT**CT-opt:sup:\`, and `*ComboT**CT-opt*\ :sup:\` were  $0.54 \pm 0.10$ ,  $0.07 \pm 0.05$ ,  $0.62 \pm 0.13$ ,  $0.41 \pm 0.11$ ,  $0.18 \pm 0.06$ , and  $0.59 \pm 0.14$ , respectively. The conformer pairs whose similarity scores are equal to or smaller than  $\mu + \sigma$  account for 85% to 87% of the 269.7 billion CID pairs, and the corresponding fractions for the  $\mu + 2\sigma$  threshold range from 96% to 98%. This information may be used to evaluate the statistical significance of the similarity score between any two conformers. For example, if the `ST**ST-opt:sup:\` value between two conformers is 0.74, the probability of randomly getting a `*ST**ST-opt*\ :sup:\` score equal to or higher than 0.74 is only 2%, and hence, one may consider that the two conformers have statistically meaningful similarity in terms of `ST**ST-opt:sup:\`.

Note that the PubChem “Similar Conformers” 3-D neighboring requires the `ST**ST-opt:sup:\ \|nonascii_41|\ 0.8` and `*CT**ST-opt*\ :sup:\ 0.5` for two molecules to become neighbors of each other. The conformer pairs whose ST value is smaller than 0.80 correspond to 99.32% of the random ST score distribution. Similarly, the conformer pairs with `CT**ST-opt:sup:\ < 0.50` correspond to 99.98% of the random CT score distribution. Therefore, if the `*ST**ST-opt*\ :sup:\` and `CT**ST-opt:sup:\` scores are assumed to be independent of each other, the probability of two conformers being identified as 3-D “Similar Conformers” neighbors of each other by chance is  $(100 - 99.32) \ \|nonascii_43|\ (100 - 99.98) = 0.0136\%$  (or 1 in 7,353). Note that the `*CT**ST-opt*\ :sup:\` score is not completely independent of the `ST**ST-opt:sup:\` score because it is evaluated at the ST-optimized alignment. Therefore, the probability of random conformers being identified as PubChem 3-D neighbors will be higher than the estimated value of 0.0136%, but it will still be smaller than 1%.

Figures 5, 6, and 7 show the distribution of the average and standard deviation of the 3-D similarity scores per-CID (computed from the similarity scores between one CID of the 734-K conformer set and all the other conformers in

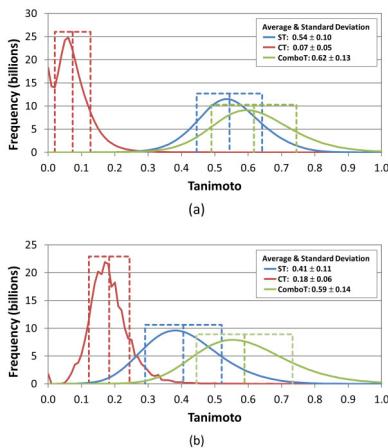


Figure 25.4: Figure 4. Overall 3-D similarity statistics between biologically tested compounds

**Overall 3-D similarity statistics between biologically tested compounds.** Distribution of 3-D similarity scores of 269,734,474,855 conformer pairs, arising from the 734,486 molecules tested in at least one bioassay archived in the PubChem BioAssay database: (a) ST-optimized similarity scores and (b) CT-optimized similarity scores. A single conformer was used for each compound. All values binned in 0.01 increments.

the set) for ST, CT, and ComboT for both ST-optimized and CT-optimized superpositions, representing the similarity scores that one may expect when a conformer in PubChem is compared with a randomly selected conformer. Most conformers have the average and standard deviation similar to those for the random conformers listed in Table 2. However, in the case of  $ST^{**}ST-opt :sup:$  [Figure 5 (a)] there is a bit of skew in the distribution of average ST value per CID towards the maximum value, peaking at 0.58, as opposed to the overall average of 0.54. Also of interest in Figure 5 (a), the ST average per-CID rapidly drops off as the ST average approaches 0.65. Note that a small fraction of biologically tested CIDs in PubChem have low average similarity scores per-CID, which indicates their relative uniqueness in the 3-D shape space (*i.e.*, their 3-D shape and/or feature orientations may be very different from most biologically tested molecules in PubChem, resulting in low similarity scores on average).

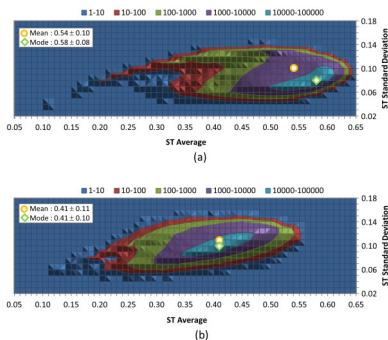


Figure 25.5: Figure 5. Per-CID shape similarity statistics of biologically tested compounds

**Per-CID shape similarity statistics of biologically tested compounds.** Distribution of the average and standard deviation of the ST scores for each of the 734,486 molecules tested in at least one bioassay archived in the PubChem BioAssay database: (a) ST-optimized ST ( $ST^{**}ST-opt:sup:$ ) and (b) CT-optimized ST ( $*ST**CT-opt*\sup:$ ). All values binned in 0.01 increments.

Potentially surprising when looking at feature similarity statistics in Table 2 is that standard deviation values for CT are about half that found for ST. When looking at the per-CID statistics in Figure 6, one sees that the range of standard deviation of CT is comparable to that of ST, although with a significant population of CIDs on the lower end of the standard deviation. Why is this so? Presumably, the 3-D orientation of features is substantially more diverse than the 3-D molecular shape, keeping both the average and standard deviation values low when compared to all other

biologically tested compounds.

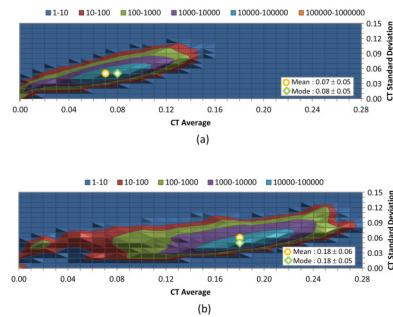


Figure 25.6: Figure 6. Per-CID feature similarity statistics of biologically tested compounds

**Per-CID feature similarity statistics of biologically tested compounds.** Distribution of the average and standard deviation of the CT scores for each of 734,486 molecules tested in at least one bioassay archived in the PubChem BioAssay database: (a) ST-optimized CT ( $CT^{**}ST\text{-opt:sup:}\backslash$ ) and (b) CT-optimized CT ( $*CT^{**}CT\text{-opt}\backslash :sup:$ ). All values binned in 0.01 increments.

An important observation is that the overall  $ComboT^{**}ST\text{-opt:sup:}\backslash$  and  $*ComboT^{**}CT\text{-opt}\backslash :sup:$  scores have very similar average values, as shown in Table 2. Whereas the  $ST^{**}ST\text{-opt:sup:}\backslash$  average was greater by 0.13 than the  $*ST^{**}CT\text{-opt}\backslash :sup:$  average, the CT-optimization results in an average  $CT^{**}CT\text{-opt:sup:}\backslash$  score greater by 0.11 than that of  $*CT^{**}ST\text{-opt}\backslash :sup:$ . As a result, the difference in averages between  $ComboT^{**}ST\text{-opt:sup:}\backslash$  and  $*ComboT^{**}CT\text{-opt}\backslash :sup:$  were only 0.03, implying that the ComboT score is not very sensitive to the type of optimization. A similar optimization-type dependency of the ST, CT, and ComboT scores was observed in Figures 5, 6 and 7. That is, whereas the ST-optimization results in an increased ST and decreased CT scores, the CT-optimization gives a decreased ST and increased CT scores, resulting in the average ComboT score that is relatively constant regardless of the optimization type employed. However, as shown in Figure 7, the  $ComboT^{**}CT\text{-opt:sup:}\backslash$  data had a narrower range of standard deviation variation per-CID than  $*ComboT^{**}ST\text{-opt}\backslash :sup:$  and the standard deviation for  $ComboT^{**}CT\text{-opt:sup:}\backslash$  per-CID appeared to linearly increase as a function of the per-CID average value.

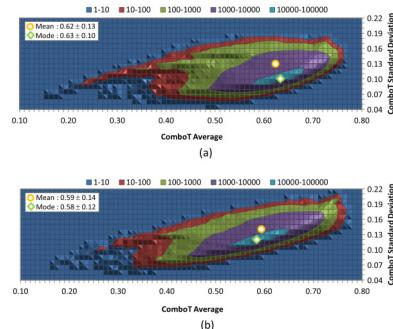


Figure 25.7: Figure 7. Per-CID shape plus feature similarity statistics of biologically tested compounds

**Per-CID shape plus feature similarity statistics of biologically tested compounds.** Distribution of the average and standard deviation of the ComboT scores for each of the 734,486 molecules tested in at least one bioassay archived in the PubChem BioAssay database: (a) ST-optimized ComboT ( $ComboT^{**}ST\text{-opt:sup:}\backslash$ ) and (b) CT-optimized ComboT ( $*ComboT^{**}CT\text{-opt}\backslash :sup:$ ). All values binned in 0.01 increments.

### 25.3.3 C. 3-D similarity score differences for the NN and NI pairs

The second part of this study examines the question: is it sufficient to only use a single conformer per compound and still realize a statistically meaningful difference or separation between the 3-D similarities of reputed actives and inactives? Or, to say this in another way, are noninactive and inactive compounds in a given bioassay well separated in 3-D shape/feature space? If so, one would expect to see some statistically significant separation in 3-D similarity scores between the partitioned noninactive-noninactive (NN) pairs and noninactive-inactive (NI) pairs. This requires 3-D similarity scores for both the NN pairs and NI pairs for each assay considered. This information is already available in the all-by-all similarity score matrices for the 734-K biologically tested molecules computed in the first part of this study. A detailed procedure for extracting the 3-D similarity scores from these matrices on the per-AID basis was described in the **Materials and Methods** section.

It is important to note that 3-D similarity methodologies (or other analysis methodologies, for that matter) are not expected to work for all biological assay data sets. A tacit assumption of 3-D methodologies is that chemical structures with similar shape and binding features will have similar (if not the same) mode of action of “activity”, *e.g.*, of binding to a protein binding pocket in the same fashion. In reality, some assays in PubChem do not have a well-defined target, *e.g.*, being a whole cell, meaning that there could be a number of targets and a number of different mechanisms of action per target for the observed activity in a single assay. In other cases, many chemical structures are active for reasons that have little to do with binding to a protein target, being aggregators, covalent binders, cytotoxic, or some other unintended mode giving rise to the measured “activity” during the biological test (so called “false positives”). As such, 3-D methodology cannot be expected to work for false positives, as reputed “active” molecules may not have any apparent 3-D correlation to each other. This is also true of cases of molecules that would be “active” if not for solubility or some other issue during the biological experiment performed (so called “false negatives”). These issues with biological tests will be nearly completely ignored for the purpose of this analysis. Instead, by looking across a wide set of assays and assay types, there is an expectation that, if there is some effect whereby 3-D similarity averages between “actives” will be greater than the averages between “actives” and “inactives” using a single conformer per compound, a certain subpopulation of assays will show this behavior.

#### C-1. Selection of AIDs from the PubChem BioAssay database

Among the 2,008 AIDs archived in the PubChem BioAssay database at the time of project initiation (January 2010), 1,744 AIDs had at least one molecule with a 3-D theoretical description. The bioassays in the PubChem BioAssay database can be classified into four categories, according to user-provided assay types (*i.e.*, screening, confirmatory, summary, and other) and the assay count for each category in the 1,744 AIDs is shown in Figure 8 (a). Note that there is another category, “Unspecified”, because the assay-type attribute for these AID records are not provided. There were 523 screening assays (30%), 867 confirmatory assays (50%), 57 summary assays (3%), 192 other assays (11%), and 105 unspecified (6%).

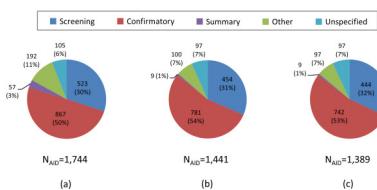


Figure 25.8: Figure 8. Assay counts by category

**Assay counts by category.** Assay count for each assay-type category in the PubChem BioAssay database: (a) for assays that have at least one tested molecule with 3-D information (as of January 28, 2010), (b) for assays that have at least one noninactive-noninactive (NN) pair and one noninactive-inactive (NI) pair, and (c) for assays that have at least six NN pairs and six NI pairs.

For a given AID, comparison of the 3-D similarity scores for the NN pairs with those for the NI pairs requires that the AID has at least one NN pair and one NI pair. Among the 1,744 AIDs, there were 1,441 AIDs that satisfy this condition [Figure 8 (b)]. Further filtering was necessary to remove AIDs in which the number of NN or NI pairs is too small,

because these AIDs may yield biased results. On the contrary, we did not want to filter out more summary assays, if it could be avoided, as there were only nine summary assays at this point. [Summary assays are final stages of lead/probe screening processes and, as such, they have a significantly smaller number of molecules provided (and hence, a smaller number of the NN and NI pairs), compared to other assay types.] Among the nine summary assays in Figure 8 (b), AID 1844 had the smallest number of the NN pairs, which was six, and this number was used as a threshold for further filtering (*i.e.*, AIDs with less than six NN pairs or less than six NI pairs were excluded in any subsequent analysis). After requiring an assay to have a minimum of six compound pairs for each of the NN and NI pairs (that is, 12 pairs per-AID in total), 1,389 AIDs resulted. As shown in Figure 8 (c), there were 444 primary screenings (32%), 742 confirmatory screenings (53%), 9 summary assays (1%), 97 other assays (7%), and 97 unspecified (7%).

## C-2. Differences between the 3-D similarity scores of NN and NI pairs

With the set of 1,389 AIDs decided, the average and standard deviation [*i.e.*, `\nonascii_44|>(*XT)` and `\nonascii_45|>(*XT)`, respectively] of the six different similarity values were determined for the NN and NI pairs per-AID. The complete set of per-AID results is available in Additional File *i.e.*, `\nonascii_46|>(*XT**NN:sub:\ )` and `\nonascii_47|(\ (*XT**NI\ :sub:)`, respectively] across the 1,389 AIDs are shown in Figure 9. The corresponding distributions of differences between the average similarity scores for NN and NI pairs per-AID [*i.e.*, `\nonascii_48|(*XT**NN-NI:sub:“”)`] are provided in Figure 10, while Table 3 and Table 4 summarize by similarity optimization type the per-AID statistics across all 1,389 AIDs [*i.e.*,  $\mu[\mu(*XT)]$ , `\nonascii_51|(*\mu(*XT))`,  $\mu[\sigma(*XT)]$  and `\nonascii_55|(*\sigma(*XT))`], with further break out by assay type category.

Additional file 1

**Similarity Scores** Statistical parameters of similarity scores for each AID.

Click here for file

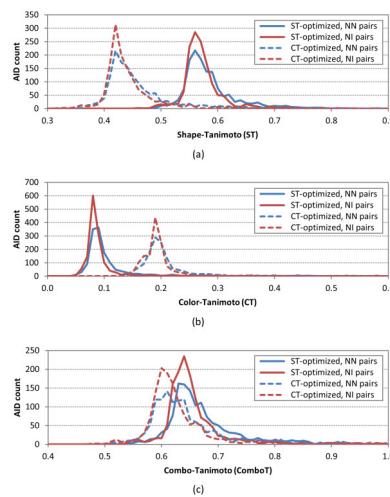


Figure 25.9: Figure 9. `\nonascii_57|(*XT)` per-AID similarity histogram **\nonascii\_58|XT**. The distribution of the average similarity scores for noninactive-noninactive (NN) pairs and noninactive-inactive (NI) pairs of 1,389 AIDs in the PubChem BioAssay database: (a) shape-Tanimoto (ST), (b) color-Tanimoto (CT), and (c) Combo-Tanimoto (ComboT). All values binned in 0.01 increments.

When looking at the distributions in Figure 9 of the per-AID results, it is interesting to see, for a single conformer per compound anyway, that the per-AID average similarity distribution of NN pairs (primarily corresponding to the reputed “active/active” compound space) overlaps extensively with those of the NI pairs (essentially the reputed “active/inactive” compound space). The original hope was that there might be two clearly separated distributions, as this would be a clear signal that 3-D similarity using a single conformer per compound is able to distinguish between “actives” and “inactives” across all PubChem assays, but this is clearly not the case. The average and

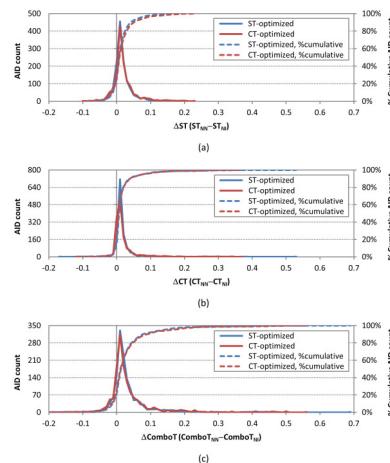


Figure 25.10: Figure 10. `|nonascii_59|(*XT**NN-NI:sub:“”)` per-AID similarity statistics `|nonascii_60|XT****NN-NI:sub:“”` per-AID similarity statistics. The distribution of the difference of the average similarity scores for noninactive-noninactive (NN) pairs and noninactive-inactive (NI) pairs of 1,389 AIDs in the PubChem BioAssay database: (a) shape-Tanimoto (ST), (b) color-Tanimoto (CT), and (c) Combo-Tanimoto (ComboT). All values binned in 0.01 increments.

standard deviation of the 3 were  $0.58 \pm 0.05$  and  $0.57 \pm 0.04$ , respectively. The corresponding values for *not* be considered statistically significant, considering their standard deviations. In fact, the average of averages per-AID for the NN and NI pairs are not significantly different from the  $ST^{ST\text{-opt}}$  and  $CT^{ST\text{-opt}}$  values for random conformers ( $0.54 \pm 0.10$  and  $0.07 \pm 0.05$ , respectively), listed in Table 2. For the same reason, the  $ComboT^{**ST\text{-opt}}$  differences between the NN and NI pairs are also not statistically significant. Note that, although the `*not*` be interpreted to be statistically meaningful, considering that the `\ :ref: '3<table_3>' -\ :ref: '4<table_4>'`. The optimization type (*i.e.*, either ST- or CT-optimization) was also found to not make significant difference in `*\ |nonascii_65|\ * (*XT**NN-NI*\ :sub:)` values.

Despite the significant overlap between the distributions for the NN and NI pairs in Figure 9, there are very subtle differences between them; for all six similarity scores, the NN-pair distributions, compared to the NI-pair distributions, have smaller AID counts at the peak and greater AID counts at the upper-tail region, indicating a small shift of the NN-pair distribution toward high similarity scores. This shift is also reflected in sharp, (mostly) normal distributions of `|nonascii_66|(*XT**NN-NI:sub:“”)`, centered on the positive side just above zero in all cases (Figure 10). This suggests that single conformer per compound 3-D similarity is showing some of the anticipated effect of the “similarity principle”, which states that structurally similar molecules are likely to have similar biological activities<sup>33343536</sup>, such that the “active/active” space is separated from the “active/inactive” space; however, for most assays in PubChem, this effect is simply not large enough to be unambiguous for all biological assays, as reflected in the  $\mu[\mu(XT_{NN-NI})]$  values smaller than  $\sigma[\mu(XT_{NN-NI})]$  for all six similarity measures. Tables 3 and 4 also clearly show that, in general, there is no clear statistically meaningful separation across assays or assay category type using a single conformer per compound. For example, while there is clearly a positive average of NN-NI difference across all similarity score types for all assays and all assay categories, ranging from 0.00-0.13 for  $\mu[\mu(XT_{NN-NI})]$ , the corresponding standard deviation of the average [*i.e.*, “ $\sigma[\mu(XT_{NN-NI})]$ ”] is consistently larger than the average value.

These results lead to a number of questions. Why isn’t there a greater, unambiguous separation in the 3-D similarity scores between the NN and NI pairs? Is it that we are employing a single conformer per compound in the analysis? After all, the current PubChem3D theoretical conformer generation approach does not guarantee that the single (de-

<sup>33</sup> Concepts and Applications of Molecular Similarity

<sup>34</sup> On outliers and activity cliffs - why QSAR often disappoints

<sup>35</sup> Do structurally similar molecules have similar biological activity?

<sup>36</sup> Similarity methods in chemoinformatics

fault) conformer used for each molecule in the NN pairs is a (or “the”) bioactive conformation. A general premise of the interpretation of 3-D similarity between a NN pair requires a “bioactive” conformation surrogate for *both* non-inactive molecules. Estimating 3-D similarity between “non-bioactive” conformers of both molecules, or between a “bioactive” conformer of one molecule and a “non-bioactive” conformer of the other, is essentially identical to 3-D similarity comparison for the NI pairs. Therefore, the use of a single conformer per compound is not likely to result in enough similarity score difference between the NN and NI pairs across a wide set of assays. Using multiple conformers per compound may result in a greater separation in similarity scores between the NN and NI pairs, but performing the same analysis using multiple conformers per compound is prohibitively expensive, considering that we are dealing with 269.7 billion conformer pairs arising from 734 thousand compounds and optimizing each conformer pair by ST and then by CT (9 TB of data gzip compressed). Any increase in the count of conformers also increases the computational complexity (and data storage requirements) by the square of the number of conformers per compound considered.

From a gross statistical approach, there is not sufficient separation across the averages of assays for a single conformer per compound to say definitively there is a clear separation between NN and NI pairs. It could be that, by considering multiple conformers per compound (and picking the best similarity conformer pair per compound pair), a clearer separation may occur, but this is a study for another day (and a bigger computer cluster and a bigger data storage system). There are, however, clear examples where some AIDs do show a clear separation, as shown in the tail regions of Figure 10, using only a single conformer per compound.

### C-3. Outliers

Although the overall average differences in similarity scores between the NN and NI pairs were not statistically significant, some AIDs do have substantial (and statistically meaningful) NN-NI differences. These “outlier” cases correspond to the tail regions of the distribution curves in Figure 10. For each of the six similarity measures, the AIDs that lie outside the 11 shows Venn diagrams detailing the outlier overlap as a function of 3-D similarity score type. To aid in discussion, the AIDs that have a statistically significant positive value of average NN-NI difference are deemed “upper-bound” cases [Figure 11(b) and 11(d)] and the AIDs that have a statistically significant negative value of average NN-NI difference are deemed “lower-bound” cases [Figure 11(a) and 11(c)].

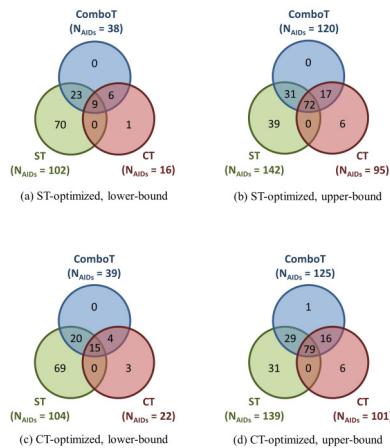


Figure 25.11: Figure 11. Assay `|nonascii_75|*/μ(*XT**NN-NI:sub:“”)`] outlier commonality by 3-D similarity type `|nonascii_77||nonascii_78|XTAssay NN-NI:sub:“”]` **outlier commonality by 3-D similarity type.** The Venn diagrams show the number of AIDs whose difference of the average similarity scores for noninactive-noninactive (NN) pairs and noninactive-inactive (NI) pairs of 1,389 AIDs in the PubChem BioAssay database are out of the range of

The lower-bound cases are when the average 3-D similarity scores for “active/inactive” compound pairs are greater than for “active/active” compound pairs, a counter result to the whole notion of chemical similarity. While the opposite of what one might expect, it can readily occur from a set of chemical structures that are predominately 3-D similar, being on both sides of that subjective and (at times) arbitrary line of being “active” or “inactive”, and where most

compounds in the compound series are considered “inactive”, as can be the case with well defined “activity cliffs”<sup>34</sup><sup>37</sup><sup>38</sup><sup>39</sup><sup>40</sup>.

Among the 109 unique, lower-bound *11 (a)*]. A similar trend is found in the case of lower-bound *11 (c)*]. Perhaps this should not be a surprise as shape alone (ignoring features) might not be expected to be a good discriminator of “actives” and “inactives”. On the other hand, as shown in Figure *11 (b)*, there are relatively few unique upper-bound outlier cases solely attributable to *11 (d)*]. This suggests, for the upper-bound AID outlier cases, use of ComboT similarity score is most efficient at finding most of the outlier cases when using a single conformer per compound.

Figure *12* compares the

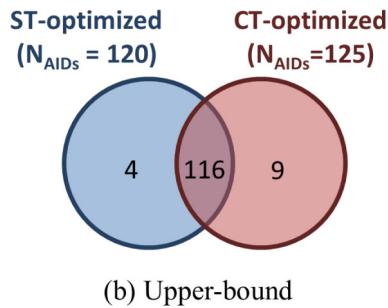
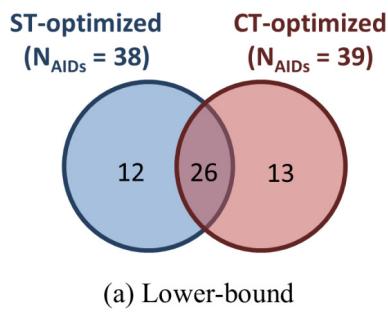


Figure 25.12: Figure 12. Assay `|nonascii_79|*[μ(*ComboT**NN-NI:sub:*)]` outlier commonality by superposition optimization type

**|nonascii\_81||nonascii\_82|ComboTAssay NN-NI:sub:\*** outlier commonality by superposition optimization type. The Venn diagrams show the number of AIDs whose difference of the average ComboT similarity scores for noninactive-noninactive (NN) pairs and noninactive-inactive (NI) pairs of 1,389 AIDs in the PubChem BioAssay database that are out of the range of

Table 5 gives the top 25% of the common *ComboT\*\*NN-NI :sub:\ upper-bound AID outliers*, yielding the largest magnitude difference in average NN-NI separation, and Table \ :ref: '6<table\_6>' gives all common \*ComboT\*\*NN-NI\* \ :sub:lower-bound AID outliers. Table 7 lists the count of assay outliers broken down by optimization type and similarity metric type. Exploring the top five assays in Table 5, the first three represent trivial examples of a compound series easily identifiable using 2-D similarity or 3-D similarity or by eye. AID 672, with the fourth largest NN-NI positive difference found, is somewhat more interesting.

AID 672 is a secondary confirmatory assay with four active compounds, shown in Figure *13 (a)*, that comprise the NN pairs. Of these four structures, three have a similar substructure but only two of the structures (CIDs 647501 and 653297) might be considered “similar” with a 0.76 2-D similarity using the PubChem subgraph fingerprint [Figure

<sup>37</sup> Design of multitarget activity landscapes that capture hierarchical activity cliff distributions

<sup>38</sup> Chemical substitutions that introduce activity cliffs across different compound classes and biological targets

<sup>39</sup> Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs

<sup>40</sup> Use of structure-activity landscape index curves and curve integrals to evaluate the performance of multiple machine learning prediction models

*13 (b)]*; however, using  $\text{ComboT}^{\text{ST-opt}}$  3-D similarity, all four compounds have pair-wise similarity beyond random (*i.e.*,  $\text{ComboT}^{\text{ST-opt}} > \{\mu + \sigma\} = 0.74$  from Table 2) except for one compound pair (CIDs 66541 and 787437). An example of one of these pair-wise superpositions [Figure 13 (c)] shows one way these different chemical structures can be superimposed relative to their shape and feature complements. While a relatively small example, and easy to examine in detail, there readily exists much larger examples.

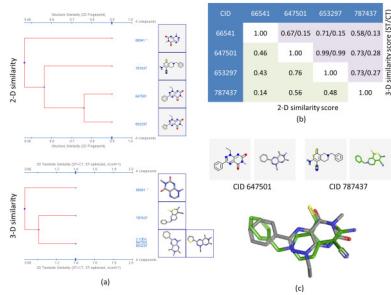


Figure 25.13: Figure 13. Separation between actives and inactives  
**Separation between actives and inactives.** An example of clear separation between

AID 2230, also a secondary confirmatory assay and fifth in the list found in Table 5, possesses a much larger NN set with 92 compounds. When examining these by 2-D cluster analysis using the PubChem Structure Clustering tool, as shown in Figure 14, there are clearly two compound series, one with 51 compounds and the other with 31 compounds, representing the majority of the “active” chemical structures. Switching to 3-D  $\text{ComboT}$  similarity, all but four of the 92 compounds, as shown in Figure 15, are inter-related at a  $\text{ComboT}^{**\text{CT-opt}}$  value above 1.04. As shown in Table 2, a value of 1.04 is more than three standard deviations away from the random average of 0.59 for  $\text{ComboT}^{**\text{CT-opt}}$ . As one goes to a  $\text{ComboT}^{**\text{CT-opt}}$  value of 1.2, several different clusters appear with the largest containing 46 compounds and second largest containing 20 compounds. This demonstrates how 3-D similarity is able to relate chemical series distinct in 2-D similarity, as representing similar shape and feature space even with a single conformer per compound.

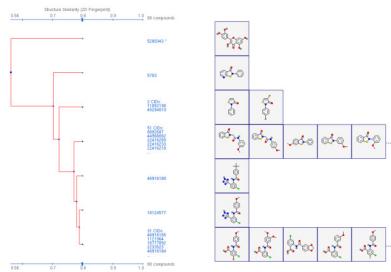


Figure 25.14: Figure 14. 2-D similarity isolates related chemical series  
**2-D similarity isolates related chemical series.** Dendrogram from the PubChem Structure Clustering tool for 88 of the 92 noninactive pairs from AID 2230 showing two primary clusters (containing 51 and 31 compounds, respectively) at 0.8 Tanimoto using 2-D similarity. Note that all but one compound is related above 0.7 Tanimoto.

## 25.4 Conclusion

Six 3-D similarity measures ( $\text{ST}^{\text{ST-opt}}$ ,  $\text{CT}^{\text{ST-opt}}$ ,  $\text{ComboT}^{\text{ST-opt}}$ ,  $\text{ST}^{\text{CT-opt}}$ ,  $\text{CT}^{\text{CT-opt}}$ , and  $\text{ComboT}^{\text{CT-opt}}$ ) in conjunction with 734,486 biologically tested compounds from PubChem were utilized to help answer the question: what is a biologically meaningful 3-D similarity score? The distribution of the six similarity measures for biologically tested

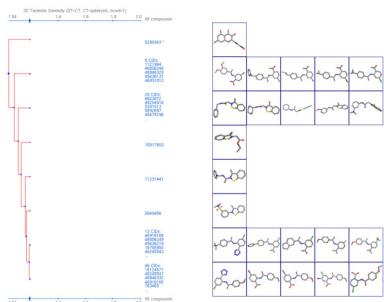


Figure 25.15: Figure 15. 3-D similarity interrelates chemical series

**3-D similarity interrelates chemical series.** Dendrogram from the PubChem Structure Clustering tool for 88 of the 92 noninactive pairs from AID 2230 showing three primary clusters (containing 46, 20, and 13 compounds, respectively) at 1.2 combo Tanimoto (ComboT) using 3-D similarity, CT-optimized. All structures are interrelated at a ComboT of 1.04, more than 3.2 standard deviations beyond the random pair average of 0.59.

compound pairs, resulting from computation of all-against-all similarity scores (269.7 billion unique conformer pairs), yielded an average and standard deviation for  $ST^{ST\text{-opt}}$ ,  $CT^{ST\text{-opt}}$ ,  $ComboT^{ST\text{-opt}}$ ,  $ST^{CT\text{-opt}}$ ,  $CT^{CT\text{-opt}}$ , and  $ComboT^{CT\text{-opt}}$  of  $0.54 \pm 0.10$ ,  $0.07 \pm 0.05$ ,  $0.62 \pm 0.13$ ,  $0.41 \pm 0.11$ ,  $0.18 \pm 0.06$ , and  $0.59 \pm 0.14$ , respectively. These values represent valuable benchmarks for the 3-D similarity values provided by PubChem and those computed by some commercial software packages. One can now know when a statistically meaningful superposition between a conformer pair occurs, potentially helping to improve their ability to analyze bioactivity information.

This random distribution of biologically tested compounds was constructed using a single theoretical conformer per compound (the “default” conformer provided by PubChem). If one were to use multiple diverse conformers per compound and pick the best 3-D similarity score, the average random distribution values may well be higher (perhaps significantly so); however, if one considers the continuum of all similarity values produced in the use of multiple diverse conformers per compound to yield a similar random distribution values, the averages (and standard deviations) above may still be applicable or, perhaps, treated as a conservative lower bound result. Further study is clearly warranted using multiple diverse conformers per compound. This work is a critical first step covering a very wide corpus of chemical structures and biological assays and creating a statistical framework to build upon.

The second part of this study explored the question of whether it was possible to realize a statistically meaningful 3-D similarity value separation between reputed biological assay “inactives” and “actives”. Using the terminology of noninactive-noninactive (NN) pairs and the noninactive-inactive (NI) pairs to represent comparison of the “active/active” and “active/inactive” spaces, respectively, each of the 1,389 biological assays were examined by their 3-D similarity score differences between the NN and NI pairs and analyzed across all assays and assay category types. Regardless of the optimization type employed (*i.e.*, either of ST- or CT-optimization), the overall average difference between the `\nonascii_91|(*XT**NN:sub:\ )` and `*\ \nonascii_92|\ \ *(*XT**NI*\ :sub:)` values, while consistently positive (as hoped), were not statistically unambiguous after considering their large standard deviations. Similarly, an increase in the

The negligible difference in 3-D similarity between the NN and NI pairs may be due to employing a single conformer per compound in this study. Conceivably the 3-D similarity between two noninactive molecules should be evaluated using the “bioactive” conformer for each molecule, being the conformer giving rise to the observed biological activity; however, the single conformers per compound used in the present study are not guaranteed to be sufficiently similar to the bioactive conformers, and the average similarity scores per-AID for the NN pairs were not much different than those from the NI pairs. Considering the negligible difference in the 3-D similarity scores between the NN and NI pairs, it may not be appropriate to analyze bioassay data with a single conformer per compound in a general sense. With that said, there were a subset of biological assays where a clear separation between the NN and NI pairs were found. In addition, use of combo Tanimoto (ComboT) alone, independent of superposition optimization type, appears to be the most efficient 3-D score type in identifying these cases.

## 25.5 Materials and methods

### 25.5.1 1. Datasets

At the time of project initiation (late January of 2010), there were 2,008 bioassays (unique identifier AID) deposited in the PubChem BioAssay database, ranging from AID 1 to AID 2310. Among the chemical structures tested in these assays, those with associated PubChem Compound records (unique identifier CID) with theoretical 3-D conformer models available<sup>22</sup> were considered in the present study. Note that the 3-D information is only available for CIDs that satisfy the following restrictions<sup>22,23</sup>:

1. is a single covalent component.
2. contains only organic [H, C, N, O, F, P, S, Cl, Br, and I] elements
3. possess only typical bonding situation (*e.g.*, no hyper valent situations)
4. not too big (*e.g.*, 50 non-hydrogen atoms or less) and not too flexible (*e.g.*, 15 effective rotors or less)
5. have five undefined stereocenters or less

There are 734,486 CIDs satisfying the above conditions for the 2,008 AIDs. All data is accessible from the PubChem website (<http://pubchem.ncbi.nlm.nih.gov>). Bulk download of data is also available from the PubChem FTP site (<ftp://ftp.ncbi.nlm.nih.gov/pubchem>). The AIDs considered are provided in Additional File

### 25.5.2 2. Similarity Score Computation

In the first part of this study, the first diverse conformer<sup>24</sup> for each of the 734,486 CIDs were downloaded. A total of six different 3-D similarity scores were computed, resulting from three different similarity metrics computed for conformer pairs superpositions optimized in two different ways. The three similarity metrics are: shape Tanimoto [ST, **Equation (1)**], measuring the shape similarity; color Tanimoto [CT, **Equation (2)**], measuring the similarity of 3-D orientation of functional groups used to defined pharmacophores (specified simply as features); and combo Tanimoto (ComboT), the simple sum of ST and CT [**Equation (3)**]. The two conformer superposition methods used optimize: by shape similarity (ST-optimized), where conformer shape overlap is maximized; and feature similarity (CT-optimized), where conformer feature overlap is maximized. Feature definitions and all similarities were computed using the C++ Shape toolkit<sup>28</sup> from OpenEye Scientific Software, Inc.

There were a total of 269,734,474,855 conformer pair similarity sets from all possible unique combinations of the 734,486 conformers. Histograms of the computed similarity scores were generated after binning all similarity scores in 0.01 increments [using the C function “rint(float)”). Note that we used only the first diverse conformer for each compound, being the PubChem default conformer. Considering the total size of data files (9.0 TB compressed, when storing only the two conformer IDs, the two similarity scores, the  $3 \times 3$  rotational matrix, and translation vector per conformer pair computed), employing additional conformers per compound in this study would quickly overwhelm the available computational resources and disk space to consider.

Many of the compounds in the present study were biologically tested in multiple assays, and hence, a substantial fraction of conformer pairs appear in multiple assays. Therefore, since consideration is given to one assay at a time, extracting the similarity scores for the conformer pairs tested in each AID from the all-by-all similarity score matrices computed and stored in the first part of study is described in Figure 16.

## 25.6 Competing interests

The authors declare that they have no competing interests.

```

Pre-compute ST-optimized similarity score matrices and store them into distributed files.
Repeat for all distributed score matrix files.

/** Read a list of AIDs that tested a particular CID. */
cid_aid_list1; // a list of AIDs in which a particular cid is tested to be noninactive.
cid_aid_list2; // a list of AIDs in which a particular cid is tested to be inactive.

/** Read a score file one line at a time */
while(FILE >> cid1 >> cid2 >> ST >> CT) {
    Combo1 = ST + CT;

    if (cid1 == cid2) { // remove a self-pair
        /* cid1-cid2 pair was tested to be noninactive and inactive, respectively, in any AIDs? */
        Get common AIDs that tested both cid1 and cid2, by comparing cid_aid_list1 & cid_aid_list2.

        Loop over common AIDs.
        Increase the counter for the ST scores. // binned with 0.01 increment;
        Increase the counter for the CT scores. // binned with 0.01 increment;
        Increase the counter for the Combo1 scores. // binned with 0.01 increment.

        /* cid1-cid2 pair was tested to be inactive and noninactive, respectively, in any AIDs? */
        Get common AIDs that tested both cid1 and cid2, by comparing cid_aid_list1 & cid_aid_list1.

        Loop over common AIDs.
        Increase the counter for the ST scores. // binned with 0.01 increment;
        Increase the counter for the CT scores. // binned with 0.01 increment;
        Increase the counter for the Combo1 scores. // binned with 0.01 increment.
    }
}

Collect similarity score histograms generated from all distributed similarity score files.
Get average and standard deviation of the similarity scores.

```

Figure 25.16: Figure 16. Analysis method overview

**Analysis method overview.** Pseudo code that describes the process by which the average and standard deviation of the ST-optimized similarity scores for noninactive-inactive (NI) pairs for individual bioassay were computed. This process was repeated for the CT-optimized similarity scores. For the average and standard deviation of the similarity scores for the noninactive-noninactive (NN) pairs were also computed in a similar manner, except that only the cid\_aid\_list1 (for noninactives) was searched both for cid1 and cid2.

## 25.7 Authors' contributions

EEB computed the similarity score matrices. SK analyzed the data and wrote the first draft. SHB reviewed the final manuscript. All authors read and approved the final manuscript.

## 25.8 Acknowledgements

We are grateful to the NCBI Systems staff, especially Ron Patterson, Charlie Cook, and Don Preuss, whose efforts helped make the PubChem3D project possible. This research was supported (in part) by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, U.S. Department of Health and Human Services. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD. (<http://biowulf.nih.gov>).



# THEORETICAL NMR CORRELATIONS BASED STRUCTURE DISCUSSION

## 26.1 Abstract

The constitutional assignment of natural products by NMR spectroscopy is usually based on 2D NMR experiments like COSY, HSQC, and HMBC. The actual difficulty of the structure elucidation problem depends more on the type of the investigated molecule than on its size. The moment HMBC data is involved in the process or a large number of heteroatoms is present, a possibility of multiple solutions fitting the same data set exists. A structure elucidation software can be used to find such alternative constitutional assignments and help in the discussion in order to find the correct solution. But this is rarely done. This article describes the use of theoretical NMR correlation data in the structure elucidation process with  $W^{13}C$  chemical shift prediction.

## 26.2 Findings

Nuclear Magnetic Resonance allied with Elemental analysis or high resolution Mass Spectroscopy are the most common tools used for the structure elucidation of new compounds. The used 2D NMR experiments like COSY, HSQC, and  $^{13}C$ -HMBC deliver correlation information between atoms that can be translated into connectivity information. Out of these, correlation information from COSY and HSQC experiments can be transcribed directly into connectivity between atoms. But the  $^{13}C$ -HMBC correlations need more attention because of their ambiguity and complexity. Hence the difficulty of the structure elucidation problem depends more on the type of the investigated molecule than on its size<sup>1</sup>. Saturated compounds can usually be assigned unambiguously using mainly COSY and some  $^{13}C$ -HMBC data, whereas condensed heterocycles are problematic due to their lack of protons that could show interatomic connectivities. This ambiguity has driven the development of different software packages to aid in the interpretation of

---

<sup>1</sup> Computer-assisted constitutional assignment of large molecules: COCON analysis of ascomycin

the  $^{13}\text{C}$ -HMBC correlation data [234567891011121314151617181920](#) as much as the development of additional correlation experiments [2122](#).

Most of these approaches have in common that they work only based on experimental NMR correlation data. COCON [142324](#) has recently been extended with the capability to create a theoretical NMR correlation data set, based on a molecule's suggested constitution. The theoretical data set is used as input data for the structure elucidation software COCON. The resulting set of constitutional assignments indicates how unambiguous NMR would have been able to describe the originally suggested molecule. The freely accessible online version of COCON (W '<http://cocon.nmr.de>' \_) offers this analysis as "Alternative Constitutions".

The data derived from the NMR correlation spectra is the result of magnetization transfer via scalar coupling between the atoms in the molecule of interest. Since the scalar coupling is based on the interatomic bonds, the correlation data will reflect those bonds. Hence, a set of all feasible NMR correlation data (theoretical correlation data) can be derived from the molecular constitution. This is done by iteratively looking for all protons in the molecule, then building a list of their atoms in 2-bond and 3-bond distance. From each proton all connectivities are inspected recursively up to three bonds distance. If a carbon is found in a two bond distance, a  $^2J$  and a 1,1-ADEQUATE correlation are added to the list. If a carbon is found in a three bond distance, a HMBC correlation is added to the list, if a proton is found, a COSY correlation is added. In principle  $^4J$  correlations for COSY and HMBC could be generated, as sometimes they are observable in experiments as well. But, COCON can not handle  $^4J$  COSY correlations, therefore those are left out. The generation of  $^4J$  HMBC correlations is not used, because when the HMBC correlations are allowed to be  $^4J$  in the structure generation process, the process takes much more time and many more results are produced. Finally carbon chemical shifts are generated by table lookup, a table reverse generated based on the chemical shift rules that COCON uses. These values are not comparable to a chemical shift prediction, but enough to ensure that COCON will generate the starting structure.

For online use, the MarvinSketch applet from ChemAxon is available for drawing or loading of the molecule. The resulting MDL file contains all atoms, their connectivity and multiplicity information. Based on this file, the recently developed Module "Alternative Constitutions" in W

The actual magnitude of the scalar coupling, and therefore the observability of a correlation, depends on the atoms involved, their chemical environment and relative geometry. For  $^1J$  and  $^2J$  couplings mainly the atoms involved and their chemical environment are of importance, since the geometry varies little. That is different with  $^3J$  coupling, which depends on the dihedral angle, hence the actual molecular conformation decides on the magnitude of the coupling. The creation of theoretical correlation data disregards the molecule's real conformation, assuming that all correlations are observable. Hence the data set represents the upper limit of correlations that may be experimentally available for

<sup>2</sup> Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation

<sup>3</sup> Computer-assisted structure elucidation: Application of CISOC-SES to the resonance assignment and structure generation of betulinic acid

<sup>4</sup> COCON: From NMR correlation data to molecular constitutions

<sup>5</sup> Computer-aided structure elucidation of organic compounds: Recent advances

<sup>6</sup> Fuzzy structure generation: A new efficient tool for computer-aided structure elucidation (CASE)

<sup>7</sup> Computer-aided determination of relative stereochemistry and 3D models of complex organic molecules from 2D NMR spectra

<sup>8</sup> Automated structure elucidation of two unexpected products in a reaction of an alpha, beta-unsaturated pyruvate

<sup>9</sup> Recent developments in automated structure elucidation of natural products

<sup>10</sup> Applications of a HOUDINI-based structure elucidation system

<sup>11</sup> SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry

<sup>12</sup> Recent advancements in the development of SENECA, a computer program for Computer Assisted Structure Elucidation based on a stochastic algorithm

<sup>13</sup> Computer aided method for chemical structure elucidation using spectral databases and C-13 NMR correlation tables

<sup>14</sup> SESAMI: An integrated desktop structure elucidation tool

<sup>15</sup> LUCY - A program for structure elucidation from NMR correlation experiments

<sup>16</sup> Combinatorial Problems in the Treatment of fuzzy C-13 NMR Spectral Information in the Process of Computer-Aided Structure Elucidation - Estimation of the Carbon-Atom Hybridization and Alpha-Environment States

<sup>17</sup> Computer-Assisted Structure Elucidation for Organic-Compound

<sup>18</sup> Computer Method of Fragmentary Formula Prediction of an unknown by its Mass and NMR-Spectra

<sup>19</sup> Structure Elucidation of organic-compounds aided by the Computer-Program System Scannet

<sup>20</sup> Computer-Aided Spectral Assignment in NMR Spectroscopy

<sup>21</sup> ADEQUATE, a new set of experiments to determine the constitution of small molecules at natural abundance

<sup>22</sup> Impact of the H-1, N-15-HMBC experiment on the constitutional analysis of alkaloids

<sup>23</sup> 2D-NMR-guided constitutional analysis of organic compounds employing the computer program COCON

<sup>24</sup> A COCON analysis of proton-poor heterocycles - Application of carbon chemical shift predictions for the evaluation of structural proposals

the constitution.

Calculations were run with three molecules (Figure 1) on the publicly available W\*\*1\*\* and Oroidin **2** in runs with theoretical and experimental data are shown in table 1. Also, a webpage allowing direct access to the results shown here has been set up on the W‘<http://cocon.nmr.de/StructureDiscussion/>‘\_ (The results are mirrored at <http://science.jotjot.net/StructureDiscussion/>).

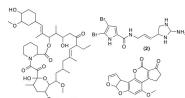


Figure 26.1: Figure 1. Ascomycin 1, Oroidin 2 and Aflatoxin B1 3 are used to evaluate the use of theoretical data  
**Ascomycin 1, Oroidin 2 and Aflatoxin B1 3 are used to evaluate the use of theoretical data.**

Ascomycin **1** is a well known ethyl derivative of Tacrolimus, it serves as example of a large natural product, featuring 43 Carbon atoms. Using theoretical NMR correlation data (COSY and  $^{13}\text{C}$ -HMBC correlations) COCON generates only one solution, independent of whether atom types are defined or not. Using experimental COSY and  $^{13}\text{C}$ -HMBC correlation data the structure generator comes up with 100 structural assignments, which are reduced to one when the atom types are fixed as well. In this case NMR correlation data was able to define the constitution unambiguously.

Oroidin **2** has been frequently used for the demonstration of COCON. The use of theoretical COSY and  $^{13}\text{C}$ -HMBC correlations leads to a total of 16 possible constitutional assignments, also predefining the atom types reduces this set to one constitutional assignment. The experimental data set leads to 252,566 structural assignments generated, which reduce to 1,486 when atom types are predefined as well. Hence the structure can not be safely determined by NMR alone. The original structure determination was carried out by chemical derivatization and total synthesis <sup>2526</sup>.

The pictures change with Aflatoxin B1 **3** with 17 Carbon atoms. Using theoretical COSY and  $^{13}\text{C}$ -HMBC data alone, COCON generates 1,048 structures, compared to 1,932 solutions using experimental data. When the atom types are predefined, COCON generates 55 constitutional assignments, compared to 108 with experimental data. The molecule set generated contains constitutions with the element cyclobutadiene, a structural element that is very uncommon in natural products. COCON has several built-in rules that eliminate certain constitutional elements, like cyclobutadiene, cyclopropene and peroxides. By default these rules are not used, but in this special case we observed a substantial difference in the number of results.

When these rules are activated the number of solutions drops to 58 for the experimental correlation data set and 33 for the theoretical data set. All planar molecules suggested are shown in Figure 2, the correct constitution and starting point of the analysis is **6**. For the small number of interesting constitutions a back-calculation on the carbon chemical shifts was made (ChemDraw v11), that were compared to the experimental values (see table 2). The last line in the table contains the sum of the absolute chemical shift differences for all carbons, exposing molecule **6** as the one that best fits the experimental data <sup>242728</sup>.

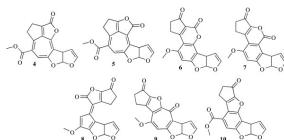


Figure 26.2: Figure 2. Planar constitutions suggested for Aflatoxin B1  
**Planar constitutions suggested for Aflatoxin B1.** Suggestions 4 - 6 are obtained using theoretical data, 5 - 10 using experimental data. Constitution 6 is the correct one.

The theoretical NMR correlation dataset is the upper limit of number of correlations that are possible with a given

<sup>25</sup> Reinvestigation into structure of Oroidin, a bromopyrrole derivative from marine sponge

<sup>26</sup> New bromo-pyrrole derivatives from sponge Agelas-Oroides

<sup>27</sup> Validation of structural proposals by substructure analysis and C-13 NMR chemical shift prediction

<sup>28</sup> Novel methods of automated structure elucidation based on C-13 NMR spectroscopy

constitution. Therefore all alternative constitutions generated with this data are “NMR-identical” with regard to correlation data. A careful analysis of this alternatives might be used to direct further investigations needed to confirm the proposed constitution. Whilst Ascomycin’s structure can be confirmed by NMR correlations, Oroidin’s structure can not. The results obtained would direct further work towards chemical derivatization and synthesis<sup>2526</sup> or x-ray crystallography. The results obtained for Aflatoxin B1 show nicely how carbon chemical shift prediction can be used as tool for the structure discussion, exposing one suggested constitutional assignment as best fitting.

## 26.3 Availability

The W ‘<<http://cocon.nmr.de>>’ \_.

## 26.4 Competing interests

The author declares that they have no competing interests.

## 26.5 Authors’ contributions

JJ maintains the W

## 26.6 Acknowledgements

The author wishes to acknowledge Rainer Haessner and the Technische Universität München for providing the Hardware for the W

# AZORANGE - HIGH PERFORMANCE OPEN SOURCE MACHINE LEARNING FOR QSAR MODELING IN A GRAPHICAL PROGRAMMING ENVIRONMENT

## 27.1 Abstract

### 27.1.1 Background

Machine learning has a vast range of applications. In particular, advanced machine learning methods are routinely and increasingly used in quantitative structure activity relationship (QSAR) modeling. QSAR data sets often encompass tens of thousands of compounds and the size of proprietary, as well as public data sets, is rapidly growing. Hence, there is a demand for computationally efficient machine learning algorithms, easily available to researchers without extensive machine learning knowledge. In granting the scientific principles of transparency and reproducibility, Open Source solutions are increasingly acknowledged by regulatory authorities. Thus, an Open Source state-of-the-art high performance machine learning platform, interfacing multiple, customized machine learning algorithms for both graphical programming and scripting, to be used for large scale development of QSAR models of regulatory quality, is of great value to the QSAR community.

### 27.1.2 Results

This paper describes the implementation of the Open Source machine learning package AZOrange. AZOrange is specially developed to support batch generation of QSAR models in providing the full work flow of QSAR modeling, from descriptor calculation to automated model building, validation and selection. The automated work flow relies upon the customization of the machine learning algorithms and a generalized, automated model hyper-parameter selection process. Several high performance machine learning algorithms are interfaced for efficient data set specific selection of the statistical method, promoting model accuracy. Using the high performance machine learning algorithms of AZOrange does not require programming knowledge as flexible applications can be created, not only at a scripting level, but also in a graphical programming environment.

### 27.1.3 Conclusions

AZOrange is a step towards meeting the needs for an Open Source high performance machine learning platform, supporting the efficient development of highly accurate QSAR models fulfilling regulatory requirements.

## 27.2 Background

Machine learning is applied within a vast range of disciplines such as economical forecasting, robotics, image analysis and risk assessment. Scientists using machine learning are not in general machine learning experts themselves and the algorithmic understanding for the various methods could be rather limited. Additionally, within many of these disciplines, low level programming knowledge is not abundant and scientist are often restricted to predefined machine learning protocols, wrapped in some graphical environment.

The new European chemical legislation, REACH, requires the chemical industry to provide information on, for example, ecotoxicity and human safety for chemicals used on the European market. However, the legislation does not support an increased usage of laboratory animals, but rather advocates the sharing of data and the development of alternative in vitro and in silico methods. Machine learning is routinely used in the prediction of chemical properties based on molecular structure information, so called quantitative structure activity relationships (QSARs). Hence, the ratification of the REACH legislation emphasizes the need to develop new methods for building and validation of QSAR models. The increased importance of QSAR modeling is manifested by the establishment of the OECD (Q)SAR project in 2004. The project aims to promote regulatory acceptance of QSAR approaches and it is in the process of establishing the “OECD Principles for the Validation for Regulatory Purpose of QSAR Models”<sup>1</sup>.

Economical necessities and the concern for laboratory animals have driven the pharmaceutical industry in the same direction, replacing in vivo studies with in vitro experiments and in silico methods. Hence, QSAR modeling is becoming increasingly important also within drug discovery<sup>2</sup>. Information related to Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) is relevant to all pharmaceutical projects and the aim is often to build models intended to be applicable within vast ranges of chemical space. In developing such global models, as much of chemical diversity as possible is included in the training sets, often encompassing tens of thousands of compounds obtained from proprietary internal databases<sup>3</sup>.

Several studies have shown that it is not possible to identify a single machine learning algorithm which will be the most accurate for all data sets, even restricted to QSAR applications<sup>4</sup>. Hence, a data set specific choice of modeling algorithm and perhaps also the usage of combined model predictions, has the potential of increasing the model applicability beyond what is achievable with a single algorithm<sup>3</sup>. Multivariate linear modeling algorithms, such as Partial Least Squares (PLS), are well established within the QSAR community. However, the often non-linear relationship between descriptors and biological responses is recognized and the application of non-linear machine learning algorithms for QSAR modeling is increasing<sup>56</sup>. In general, non-linear machine learning algorithms represent a more complex optimization problem than linear methods and therefore require more training examples. Hence, the non-linear machine learning methods are of particular interest for global QSAR modeling, for which they need to be used in conjunction with thorough statistical validation and assessment of the applicability domain. In addition, non-linear methods are considered more difficult to use because of the tweaking of model hyper-parameters, such as the number of hidden neurons in an artificial neural network, often required to build accurate models.

Commercial mathematical packages, for example MATLAB, interfaces several machine learning algorithms, while Simca and TreeNet are commercial packages, well established within the QSAR community, thought developed around a single modeling algorithm. The Open Source statistical package R<sup>7</sup> has several third party modules with

---

<sup>1</sup> The report from the expert group on (quantitative) structure-activity relationships [(Q)SARs] on the principles for the validation of (Q)SARs

<sup>2</sup> ADMET in silico modelling: towards predictin paradise?

<sup>3</sup> Greater Than the Sum of its Parts: Combining Models for ADMET Prediction

<sup>4</sup> Contemporary QSAR Classifiers Compared

<sup>5</sup> Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review

<sup>6</sup> ADMET Property Prediction: The State of the Art and Current Challenges

<sup>7</sup> NOTITLE!

state-of-the-art machine learning methods. Orange<sup>8</sup> and Weka<sup>9</sup> are developed to be machine learning platforms providing multiple algorithms together with preprocessing and validation methods, while KNIME<sup>10</sup> is a general pipe-lining system with several machine learning plug-ins.

Model hyper-parameter selection aims to find the parameters with the greatest generalization accuracy for a given data set by comparing the accuracy for different combinations of hyper-parameters. A few machine learning packages implement semi-automated model hyper-parameter selection. The Random Forest (RF) module of R monotonously increases the number of active variables until the out-of-bag (OOB) error no longer decreases. The libSVM<sup>11</sup> authors recommend a grid search to find the *C* and *nonascii\_1*\* parameters with the best cross validation (CV) accuracy, while the grid search in the OpenCV [#B12]\_ SVM implementation can optimize the \**C*, *nonascii\_2*\*, *p*, *nu*, *coeff* and *degree* parameters. Weka offers the possibility to calculate the CV accuracy varying a single parameter within a user defined interval. In addition, Weka implements a grid search restricted to two model parameters.

Despite the diversity of available machine learning packages, there is no package fulfilling all of the requirements on an Open Source state-of-the-art QSAR modeling platform. Such a system needs to include all tools necessary within a work flow encompassing database communication, data preprocessing, descriptor calculation and selection, model building and validation. It should be possible to build flexible machine learning applications in a graphical programming environment, as well as in a scripting mode. Because data sets often contain tens of thousands of compounds and the size of available data sets is expected to grow rapidly, the machine learning algorithms need to be highly numerically efficient. To exhaust the statistical aspects of model development, multiple and complementary machine learning algorithms should be made available. Complex modeling algorithms need to be customized for non-expert users and model hyper-parameters selected in an automated work flow to increase accuracy and efficiency in the model development process. Finally, the system should make it easy to develop models compliant with the OECD principals for validation of QSAR models.

AZOrange is a general Open Source platform for machine learning, however developed to meet the increasing demand for ADMET models in drug discovery in particular. AZOrange customizes several high performance state-of-the-art machine learning algorithms. The automated and generalized model hyper-parameter selection is a unique feature of AZOrange. The customization and the automated model hyper-parameter selection provide the tools necessary for automated model development, batch generation of models and the assessment of multiple model hypothesis. In addition, a graphical programming environment makes development of flexible high performance machine learning applications possible without scripting requirements.

## 27.3 Implementation

The Open Source foundation of AZOrange gives complete algorithmic transparency, allows further development of the algorithms and reduces license costs. Furthermore, the Open Source solution grants the fundamental scientific principal of reproducibility, which is recognized in the OECD principals for QSAR modeling as an advantage over commercial packages. Making AZOrange itself an Open Source code reaches out to a larger group of users, thereby assuring a more extensive validation of the code.

The “Architecture” subsection describes the AZOrange architecture and the major Open Source dependencies, while the “Extension of Orange functionality” subsection gives a detailed overview of the functionality by which AZOrange complements the Orange package to facilitate ADMET modeling in particular.

### 27.3.1 Architecture

Because of its diversity, quality and architecture, AZOrange uses the Orange machine learning platform as a foundation. Orange implements the demanding numerical computations in C, while wrapping the top level objects in a

<sup>8</sup> Orange: From Experimental Machine Learning to Interactive Data Mining

<sup>9</sup> The WEKA Data Mining Software: An Update

<sup>10</sup> KNIMEtech

<sup>11</sup> NOTITLE!

Python scripting environment, as illustrated in Figure 1. The Python application programming interface (API) is used in a graphical user interface (GUI), providing a highly flexible framework for tailored machine learning application development. AZOrange interfaces Orange with a set of other Open Source codes to extend its functionality, in particular for QSAR modeling. The OpenCV package <sup>12</sup> adds a set of computationally efficient, non-linear machine learning algorithms. Although non-linear machine learning algorithms usually results in more accurate models for large descriptive QSAR data sets, a linear method constitutes a baseline. The PLearn <sup>13</sup> interface makes a partial least squares (PLS) algorithm executable from within the AZOrange framework. APPSPACK <sup>14</sup> was integrated for automated derivative free optimization of the model hyper-parameters, while Cinfony <sup>15</sup> provides AZOrange with a set of publically available molecular descriptors.

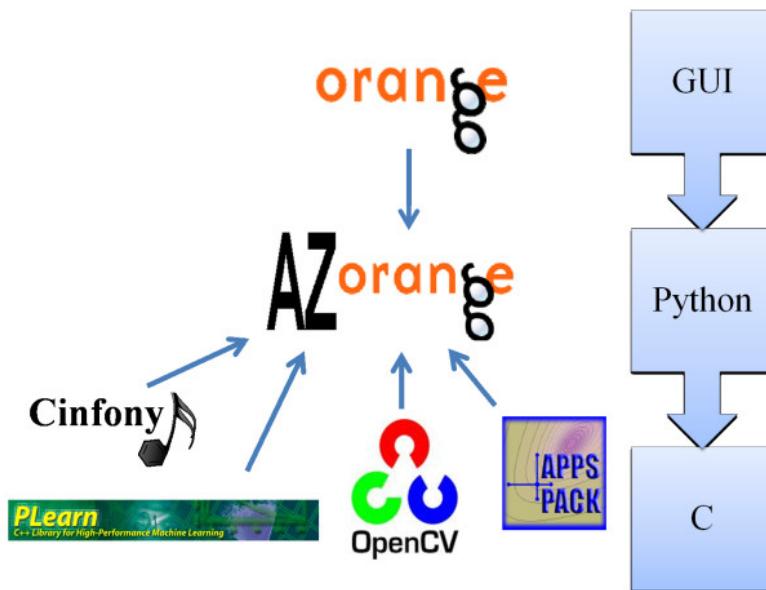


Figure 27.1: Figure 1. The architecture of AZOrange

**The architecture of AZOrange.** The architecture and the major Open Source codes constituting AZOrange.

### 27.3.2 Extension of Orange functionality

The major interfaces of AZOrange extend the functionality of Orange by incorporating descriptor calculation, additional persistent learners and generalized, automated model hyper-parameter selection. Further modifications are made to enhance feature ranking, prediction of external test sets and model persistency.

#### Molecular Descriptors

As AZOrange is intended to be a complete platform for QSAR modeling, a set of Open Source molecular descriptors is interfaced. Provided with SMILES, AZOrange calculates any descriptor within the Cinfony package and makes them available in Orange data objects. Cinfony is a mutual Python API for CDK <sup>16</sup>, RDkit <sup>17</sup> and Open Babel <sup>18</sup>, thereby efficiently interfacing the descriptors of these packages with AZOrange.

<sup>12</sup> OpenCV

<sup>13</sup> PLearn

<sup>14</sup> Algorithm 856: APPSPACK 4.0: Asynchronous parallel pattern search for derivative-free optimization

<sup>15</sup> Cinfony - combining Open Source cheminformatics toolkits behind a common interface

<sup>16</sup> The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics

<sup>17</sup> NOTITLE!

<sup>18</sup> OpenBabel

## Feature ranking and selection

The Orange methods available for global ranking of features have been extended by the Random Trees (RT) variable importance assessment method <sup>19</sup> in OpenCV. The OpenCV implementation randomly permutes the values of one variable within the out-of-bag (OOB) set of examples of each tree. The OOB error of all trees, with and without permuted values, is used to quantify the importance of each variable. This RT variable importance assessment can be used to rank the importance of variables in a data set and consecutively in a wrapper variable selection algorithm.

## Learners

The Orange learners are complemented by five new learners. These learners are implemented to comply with the Orange learner object standard and encompasses all functionality of these objects. The integrated learners are customized versions of the RT, Support Vector Machine (SVM), CvBoost and Artificial Neural Networks (ANN) implementations in OpenCV and the PLS algorithm in PLearn. The default model parameter values are those of OpenCV, but these values can be changed within AZOrange. All models except CvBoost, which is solely for binary classification, can be used with any dimensionality of the response variable. Furthermore, they are persistent, making AZOrange model predictions accessible from within other environments.

By default AZOrange imputes missing values with the average or the most frequent value of the training set, as implemented by the corresponding Orange method. Imputation is used on both the training set and on examples being predicted by AZOrange models. However, for Random Forest (RF) models, imputation can be replaced by defining surrogate nodes upon training, as originally proposed by Breiman <sup>20</sup>. The SVM and ANN algorithms require scaling of the variable values for the optimization algorithms to operate smoothly. Unless scaling is explicitly deselected, the ANN algorithm will use OpenCV functions to scale both the attribute values and the response variable. The OpenCV implementation of SVM does not have this inherent scaling. Hence, it is performed in AZOrange, transforming the variable values into the range between -1 and 1, using the same expression as in libSVM <sup>11</sup>.

AZOrange implements a simple generalized consensus model, combining the predictions from AZOrange learners by averaging or by using the majority vote. A consensus prediction can be made even with an even number of classifiers if the individual classifiers calculate prediction probabilities. The class with the greatest sum of probabilities is predicted.

## ANN customization

The OpenCV ANN algorithm is customized to reduce the risk for overfitting and to increase the chances of finding an optimal network. This is achieved by supporting early stopping based on the accuracy on a validation set <sup>21</sup> and by providing generalized methods for building multiple networks using different initial weights <sup>22</sup>.

The ANN implementation in OpenCV supports two stopping criteria, reaching a predefined maximum number of epochs or a decrease in training set accuracy between two consecutive epochs ( $*\epsilon*$ ) below a user defined threshold. Using the  $*\epsilon*$  criteria will stop the training when the first of these two criteria is met, while the maximum number of epochs disregards the change in training set accuracy.

The OpenCV stopping criteria have been complemented by an early stop criteria. When early stopping is used, 20% of the data will be selected by stratified random sampling to constitute a validation set, which is left outside of the updating of the weights. Lutz <sup>21</sup> examined three classes of early stopping criteria. For robustness with respect to noisiness on the accuracy surface, the third class of stopping criteria was selected. Hence, the accuracy is evaluated on the validation set every fifth epoch and the early stopping criteria is triggered when the performance does not improve over a user defined number of consecutive evaluations (defaulting to 5). The network with the best performance on the validation set is selected as the final model. When early stopping is enabled, the training of the network stops when the early stop criteria is triggered or when the maximum number of epochs is reached. The default maximum number of epochs has been increased to 3000.

---

<sup>19</sup> NOTITLE!

<sup>20</sup> Random Forests

<sup>21</sup> Automatic early stopping using cross validation: quantifying the criteria

<sup>22</sup> NOTITLE!

The difficulty of finding the global minimum on any multi dimensional surface is well recognized, also in the context of optimization of the network weights of an ANN<sup>22</sup>. The chances of finding a more accurate network increases when training multiple networks while varying the initial weights, thus starting in different points on the surface. The initial weights in AZOrange are varied by controlling the seed of the pseudo random sampling in the Nguyen-Widrow initialization function used by OpenCV. The user can control the number of networks built and a final network is selected based on the accuracy on the validation set. The network resulting from the smallest number of iterations is selected when several networks have the same accuracy.

## Model parameter selection

A general automated model parameter optimizer has been developed within AZOrange. Any number of parameters can be optimized simultaneously for the RF, SVM, ANN, CvBoost and PLS algorithms. For computational efficiency, the pattern search algorithm in APPSPACK is used to provide a derivative free search algorithm. Before starting the pattern search, the generalization accuracy is always assessed with the default model parameter configuration. Additionally, the mid point of each model parameter range is evaluated to provide an initial point for the pattern search. To reduce the risk of ending up in a local minimum, the pattern search can be complemented by an optional sparse grid search that could select an initial point other than the mid range point.

For model parameter selection purposes, the objective function needs to quantify the difference in generalization accuracy when varying the model parameter settings. Hence, an accurate generalization error is not critical, while correct relative generalization errors is paramount. The objective function used with the automated parameter optimizer is a double CV loop with any number of folds, however defaulting to a single 5-fold CV.

In an automated model parameter optimization scheme, special care should be taken to avoid overfitting as a result of the selection of too complex models. The generalization accuracy increases with increased model complexity up until a point where model flexibility can no longer be accounted for by the data set. Thus, this optimal model complexity is dependent on the size of the data set. Using a CV scheme to assess the generalization accuracy reduces the risk of overfitting, as compared to considering solely a training set accuracy. The tendency to select model parameters resulting in complex models could be moderated by introducing a regularization term, penalizing solutions with greater model complexity<sup>23</sup> or by considering the Akaike Information Criterion (AIC)<sup>24</sup>. The pragmatic approach controlling the parameter optimization in AZOrange thus far simply restricts the search intervals. Furthermore, the model parameter point with the greatest generalization accuracy could be disregarded if the improvement in accuracy is smaller than the variance originating from data sampling effects.

Multiple parameters control the architecture and complexity of machine learning algorithms. Even though the parameter optimizer handles any number of parameters simultaneously, a comprehensive optimization would in general be far too computationally expensive. Hence, for each machine learning algorithm, the parameters with the greatest impact on model accuracy need to be identified. Table 1 displays the parameters of the AZOrange machine learning algorithms selected for optimization by default. The ranges within which the parameters are optimized are also specified. The selection is supported by experience and results from literature. However, a more comprehensive study on the improvements in generalization accuracy upon optimizing various model parameters would be desirable.

## Miscellaneous

In addition to the major interfaces described above, AZOrange extends the functionality of Orange by various modifications to the Orange code.

AZOrange makes extensive use of automated model parameter selection to tune the machine learning algorithms for individual data sets. There is a clear risk of overestimating the generalization accuracy when the model hyperparameters have been selected using the same data set. Hence, AZOrange has generalized methods to perform a double CV loop around the model parameter optimization. The generalization accuracy is assessed on the left out

---

<sup>23</sup> Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters

<sup>24</sup> Information theory and an extension of the maximum likelihood principle

folds of the external loop, while model parameter optimization is performed on the corresponding training data, also using CV.

When a machine learning model is used for an extensive time period and new data is being made available during this time, it is important to be able to assess model performance on the new data. Alternatively, when there is a known time dependence in the data available at the time of developing the model, a temporal test set is a complement to other data sampling strategies used to assess the generalization accuracy of the model. Thus, AZOrange makes methods to quantify the performance on such separate test sets available in the GUI.

Methods for assessment of the applicability domain are crucial to a QSAR platform and an important area for further development. AZOrange includes a module for calculating the Mahalanobis distance in descriptor space between an example being predicted and the training set. The training set can either be represented by the nearest neighbors of the training set or the center of the set. An applicability domain can be estimated by considering the distribution of such Mahalanobis distances of compounds in an external test set. An example falling into the first quartile would be considered inside the applicability domain, while predictions of compounds in the last quartile would be considered unreliable. Using an external test set allows for assessment of the correlation between the Mahalanobis distance and the prediction error. The Mahalanobis distance based method already available within AZOrange is currently being complemented with multiple reliability methods in a collaborative effort with the Orange group. To enhance compatibility of Orange data objects while concatenating various data sets, the domain data objects are automatically harmonized. For example, type conversion is tried for variables with the same name but of different type and the order of enumeration variables is always forced to be aligned. The user is provided with information about the conversions required for compatibility. This domain compatibility enhancement is also applied to examples being predicted, for compliance with the model domain.

When developing a classification model it could be more important to have a high sensitivity for one class at the expense of a greater number of corresponding false predictions. Furthermore, classifiers could be biased and show a greater tendency to predict one of the classes, in particular for unbalanced data sets. In both these cases class weights can be used to shift the prediction distribution towards the desired class. The RF, SVM and ANN algorithms of AZorange implements support for weighting the importance of classes by setting the priors in the underlying OpenCV algorithm.

## 27.4 Results and Discussion

The AZOrange package is intended to aid scientist with limited knowledge in machine learning to accurately use state-of-the-art machine learning algorithms. The customization of the interfaced learners includes, for example, descriptor scaling, optimization of model hyper-parameters and appropriate stopping criteria. Currently, AZOrange provides a framework by which fully automated QSAR pipelines can be constructed, while a generalized process is being developed. This process will accept a data set as the input and automatically return the most accurate model, selected amongst models based on all the statistical methods interfaced with Orange. To reduce the risk of overfitting and to avoid overestimating the generalization accuracy, the process under development uses multiple re-sampling of external test sets, for which no selection of methods or parameters has been done. In addition, the variation in accuracy between the folds in the external validation loop constitutes a stability criteria, also used in the model selection process. This process will further aid users with little experience in machine learning to build high quality QSAR models.

AZOrange is supported on the Ubuntu Lucid Lynx platform and freely accessible under the general public license (GPL) on a git hub. In addition, a version of the source code is provided together with this manuscript as “Additional file

Additional file 1

**The AZOrange source code.** A zipped version of the AZOrange source code is provided. This version corresponds to the “chemistryCentral” tag in the git repository.

[Click here for file](#)

This section gives a brief tutorial of the GUI, as well as references to example code for Python scripting with AZOrange. The examples are designed to cover functionality unique for AZOrange.

### 27.4.1 Using the GUI

The GUI of Orange has been complemented by several widgets to enhance functionality important to make Orange a complete platform for predictive QSAR model development. The following Canvases will illustrate how to use some of the added methods.

#### Generalization accuracy with model parameter optimization

Figure 2 shows an AZOrange Canvas used to calculate RDK descriptors through the Cinfony widget for bioactivity data on the Estrogen receptor from PubChem (assay ID 639). The “Test Optimized Learners” widget is connected to the data set and to the AZOrange RF learner. Any model parameter can be selected for optimization and the number of folds in the inner and outer CV loops can be defined by the user. The presented accuracy metrics is the summation over the test sets of the outer loop, while optimizing the selected model parameters using CV on the corresponding training sets.

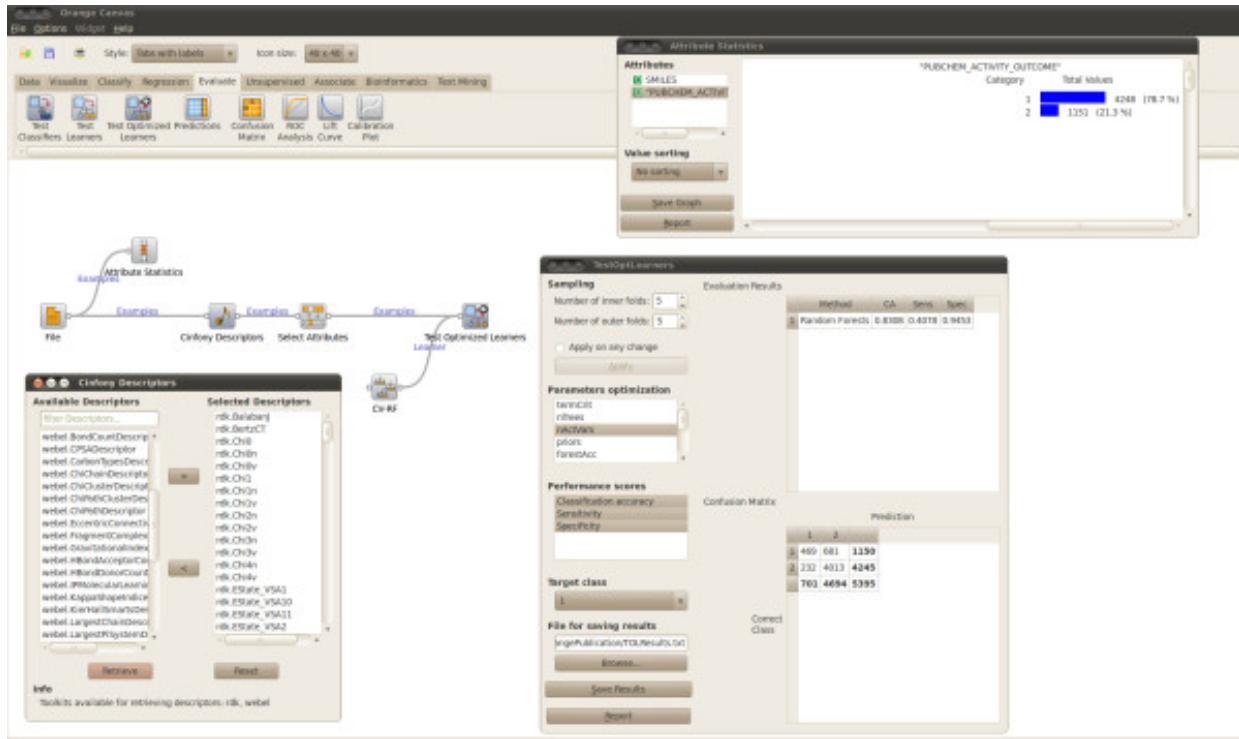


Figure 27.2: Figure 2. Generalization accuracy

**Generalization accuracy.** Assessing the generalization accuracy of a learner with optimized model hyper-parameters with the double loop data sampling algorithm.

#### Building and saving a model with optimized parameters

Figure 3 and 4 show model parameter optimization and consecutive training of a model with the optimized parameters. In the “Parameter Optimizer” widget interface the sampling technique used to assess the generalization accuracy at each model parameter configuration can be selected. In addition, the mid range model parameter point used as

the default initial point for the pattern search, can be replaced by the best point obtained with a grid search. The parameters displayed on the right hand side depend on the learner collected to the “Parameter Optimizer” widget. Figure 3 shows the optimizer interfaces while connected to the RF and SVM learners. The right hand side table defines the parameter default values, the allowed values and the ranges within which the parameters are optimized. The “Parameter Optimizer” widget returns an Orange learner object with optimized rather than default model parameters. The actual model is built by connecting the “Train Learner” widget with the “Parameter Optimizer” and the train data set. The “Train Learner” widget interface can be used to save the trained model to disk, as displayed in Figure 4.

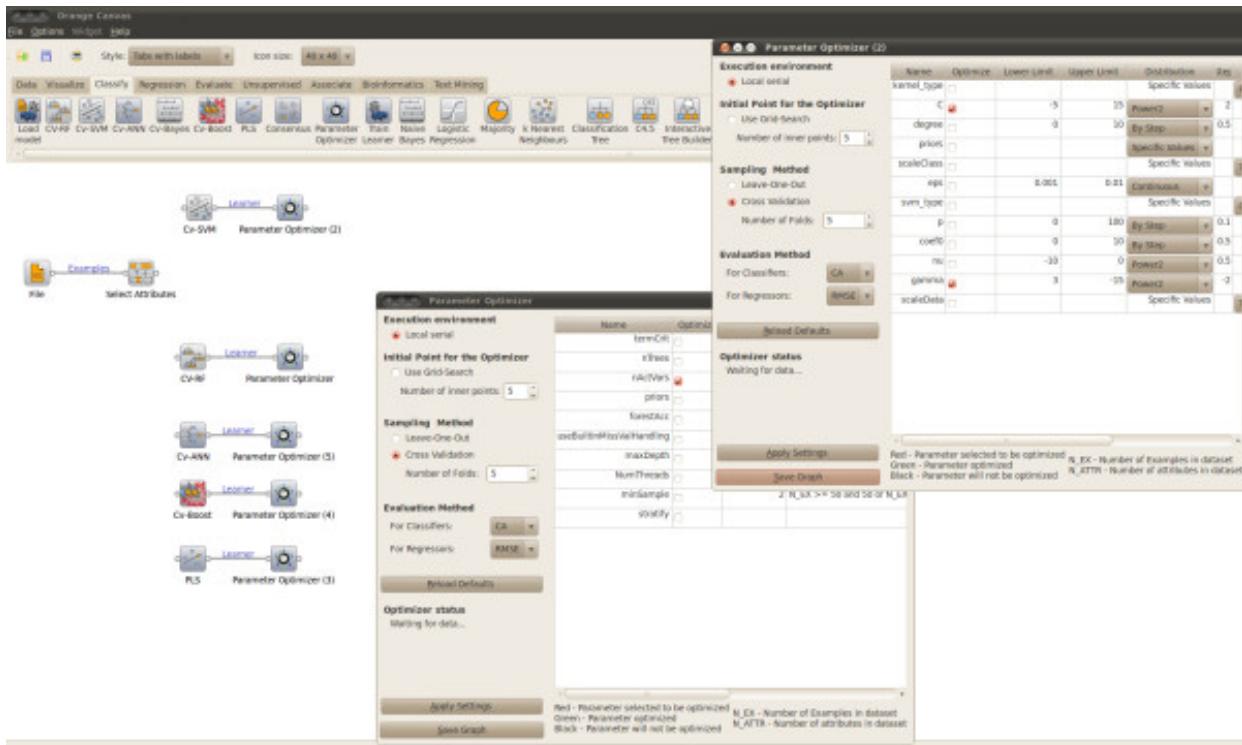


Figure 27.3: Figure 3. Parameter optimization

**Parameter optimization.** Using the “Parameter Optimizer” widget to optimize the parameters of any AZOrange learner.

## Performance on external test sets

Saved models can be loaded into AZOrange and used to predict temporal test sets as displayed in Figure 5. The temporal test set must contain all variables used while training the model. However, additional variables will be ignored upon predicting an example. Use the “Test Classifiers” widget to calculate accuracies of a trained model on the temporal data. “Test Classifiers” returns the Orange results object which can be used with for example the “Confusion Matrix” widget.

### 27.4.2 Scripting with AZOrange

Even more flexible machine learning applications can be tailored using the Python scripting API of AZOrange. The basic principles of AZOrange scripting follow those of the Orange Python API. Three well commented scripts reproducing the work flows of the Canvases above, are available in the following sub directory of AZOrange: \$AZORANGEHOME/doc/openExampleScripts

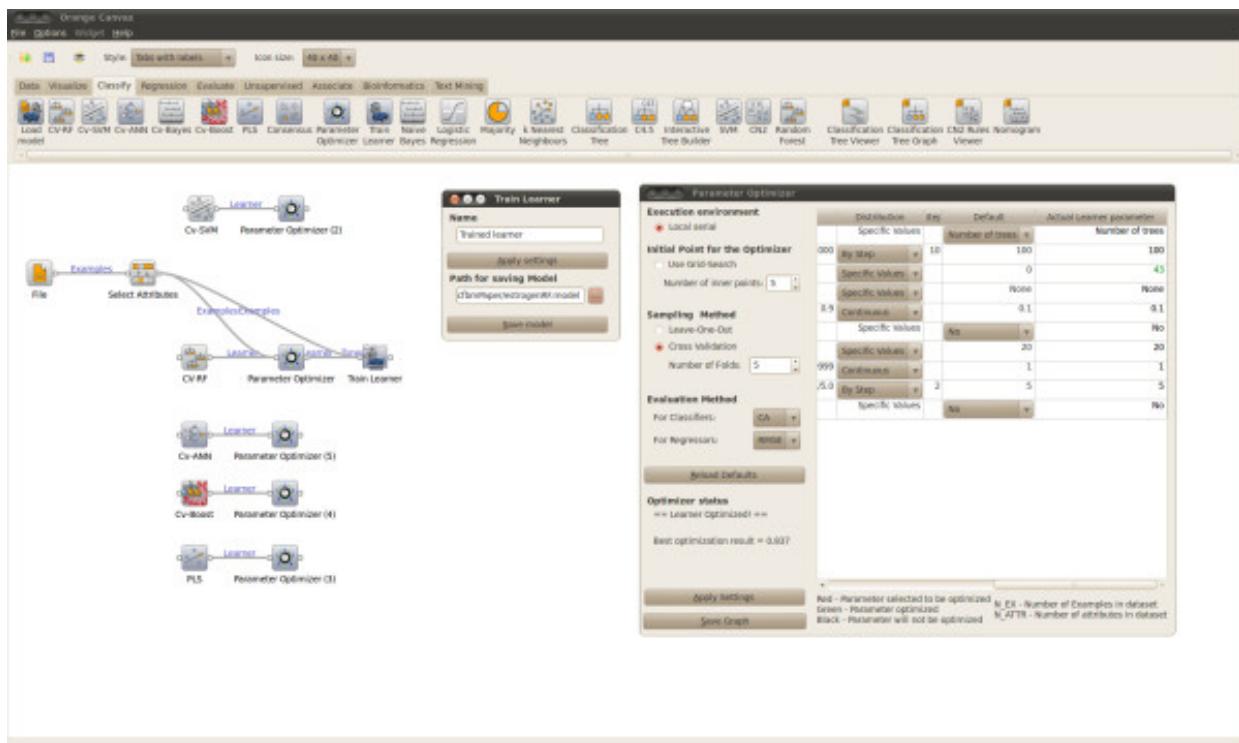


Figure 27.4: Figure 4. Saving a model  
Saving a model. Saving a trained model with optimized model hyper-parameters.

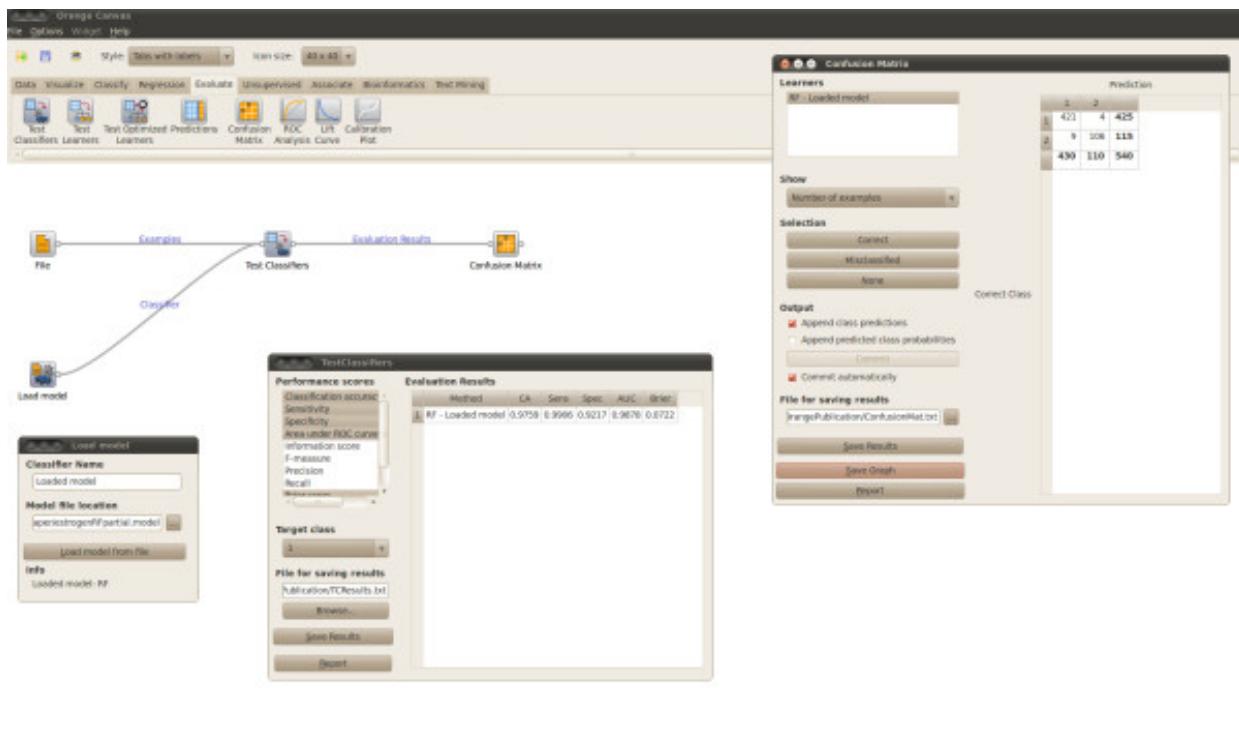


Figure 27.5: Figure 5. External test set  
External test set. Test the performance of a saved model on an external test set.

## 27.5 Conclusions

AZOrange complements already available machine learning packages by interfacing and customizing several state-of-the-art machine learning algorithms. Multiple methods within the same package makes a data set specific selection of algorithm simple, potentially increasing accuracies beyond what is achievable in packages based on a single machine learning algorithm. The customization reduces the algorithmic knowledge requirements on users and allows users to concentrate on model development and data analysis, rather than programming and compatibility issues. For example, AZOrange transforms data formats, scales descriptor values where appropriate, accommodates missing values and selects stopping criteria. Model hyper-parameter selection is particularly important for non-linear machine learning algorithms and AZOrange optimizes any number of model hyper-parameters automatically and simultaneously for all its methods. This assures that a wider range of model parameters can be searched and makes the model development process more efficient as compared to manual tweaking of parameters. The AZOrange methods are accessible, not only at a scripting level, but also for development of flexible machine learning applications within a graphical user interface.

AZOrange is a complete platform for QSAR modeling, integrating data retrieval, descriptor calculation and selection, with training and validation. AZOrange is intended to aid in developing models compliant with the OECD principals for validation of QSAR models, by providing established methods for performance assessment and by granting the principal of transparency in being based solely upon Open Source codes. Furthermore, the tools for automated QSAR model development makes AZOrange suitable for large scale batch generation of QSAR models.

## 27.6 Availability and Requirements

- Project name: AZOrange
- Project home page: <http://github.com/AZcompTox/AZOrange>
- Operating system: Ubuntu 10.04 LTS
- Programming language: C and Python
- Other requirements: All dependencies are automatically installed by the AZOrange installation procedure.
- License: GPL

## 27.7 Competing interests

The authors declare that they have no competing interests.

## 27.8 Authors' contributions

SB and LAC initially identified the need for the AZOrange platform and they explored the opportunities for using an Open Source solution. In addition, they identified the Orange package as a comprehensive machine learning tool already providing much of the desired functionality. JCS evaluated the computational efficiency of the OpenCV package and its potential to be used for QSAR applications. JCS, PA and LAC together explored the possibilities for using a pattern search algorithm for generalized and automated model hyper-parameter selection. JCS and PA have been responsible for the customization of the learners and the interface to the Cinfony package. JCS has been the main author of the manuscript, while all authors have contributed by revising and further developing the content.

## 27.9 Acknowledgements

We would like to acknowledge the Open Source community and in particular the Orange developers at the Artificial Intelligence Laboratory at the University of Ljubljana, without whom this project had not been possible.

# MULTIPLE SEARCH METHODS FOR SIMILARITY-BASED VIRTUAL SCREENING: ANALYSIS OF SEARCH OVERLAP AND PRECISION

## 28.1 Abstract

### 28.1.1 Background

Data fusion methods are widely used in virtual screening, and make the implicit assumption that the more often a molecule is retrieved in multiple similarity searches, the more likely it is to be active. This paper tests the correctness of this assumption.

### 28.1.2 Results

Sets of 25 searches using either the same reference structure and 25 different similarity measures (similarity fusion) or 25 different reference structures and the same similarity measure (group fusion) show that large numbers of unique molecules are retrieved by just a single search, but that the numbers of unique molecules decrease very rapidly as more searches are considered. This rapid decrease is accompanied by a rapid increase in the fraction of those retrieved molecules that are active. There is an approximately log-log relationship between the numbers of different molecules retrieved and the number of searches carried out, and a rationale for this power-law behaviour is provided.

### 28.1.3 Conclusions

Using multiple searches provides a simple way of increasing the precision of a similarity search, and thus provides a justification for the use of data fusion methods in virtual screening.

## 28.2 Background

The constantly increasing costs of drug discovery have resulted in the development of many techniques for virtual screening<sup>1234</sup>. One of the simplest, and most widely used, techniques is similarity searching, in which a known bioactive reference structure is searched against a database to identify the nearest-neighbour molecules, since these are the most likely to exhibit the bioactivity of interest<sup>56789</sup>.

A quarter of a century has passed since the first descriptions of similarity searching<sup>1011</sup>, but it has still not proved possible to identify some single similarity method that is consistently superior (in terms of quantitative measures of screening effectiveness such as enrichment factor or cumulative recall) to the many others that have been developed over the years<sup>71213</sup>. Indeed, we would agree with Sheridan<sup>6</sup> that it is unlikely that it will ever be possible to identify such an optimal solution. There has hence been much interest in the use of *data fusion* methods, in which multiple searches are carried out and the resulting database rankings combined to yield an overall ranking (in order of decreasing probability of activity) that is the final search output presented to the user. The many studies that have been carried out have suggested that the fusion of multiple search outputs can provide an effective, and robust, alternative to conventional, single-search approaches<sup>14</sup>. Most of these studies have been empirical in character and have not sought to provide a theoretical rationale for the fusion procedures that have been used. There is, however, an underlying assumption that is common to all approaches to the use of data fusion for virtual screening. This assumption is that the availability of information resulting from multiple searches will increase the likelihood of detecting active molecules when compared to the use of just a single search. The assumption seems entirely reasonable but it has not, to our knowledge, been tested systematically: this article reports such a test.

The starting point for our work was a paper by Spoerri that investigated the extent to which the assumption applies when a query is matched against a database of textual documents using multiple search engines<sup>15</sup>. In brief, Spoerri showed that a given document was more likely to be relevant to a user's query the more search engines retrieved that document, with this likelihood increasing very rapidly as the number of search engines retrieving it increased. Spoerri called this phenomenon the Authority Effect: here, we seek to determine whether the Effect also applies in the context of similarity-based virtual screening systems, since this would provide a firm basis for the use of fusion methods.

## 28.3 Results and Discussion

We have considered both of the two principal types of data fusion that have been used for virtual screening: *similarity fusion* and *group fusion* (which we refer to subsequently as SF and GF, respectively)<sup>14</sup>. SF involves searching a single reference structure against a database using multiple different similarity measures, and the output is obtained by combining the rankings resulting from these different measures. GF involves searching multiple reference structures against a database using a single similarity measure, and the output is obtained by combining the rankings resulting from these different reference structures. The reader should note that while we refer in this paper to the SF experiments and the GF experiments, real data fusion using either of the two approaches requires a procedure to combine the multiple ranked search outputs to give the final ranking that is presented to the searcher. Here, we have merely

<sup>1</sup> Integrating virtual screening in lead discovery

<sup>2</sup> Virtual Screening in Drug Discovery

<sup>3</sup> Quo vadis, virtual screening? A comprehensive survey of prospective applications

<sup>4</sup> Virtual screening: an endless staircase?

<sup>5</sup> Molecular similarity analysis in virtual screening: foundations, limitation and novel approaches

<sup>6</sup> Chemical similarity searches: when is complexity justified?

<sup>7</sup> Similarity methods in chemoinformatics

<sup>8</sup> Molecular similarity measures

<sup>9</sup> How similar are those molecules after all? Use two descriptors and you will have three different answers

<sup>10</sup> Atom pairs as molecular-features in structure activity studies - definition and applications

<sup>11</sup> Implementation of nearest-neighbour searching in an online chemical structure search system

<sup>12</sup> Why do we need so many chemical similarity search methods?

<sup>13</sup> Similarity metrics and descriptor spaces - which combinations to choose?

<sup>14</sup> Data fusion in ligand-based virtual screening

<sup>15</sup> Authority and ranking effects in data fusion

considered the molecules retrieved in the top-1% or top-5% of the rankings (see Experimental Methods), with no attempt being made to produce a final output ranking from the top-ranked subset of the database.

We consider first the results of the SF searches. Figure 1(a) shows the overlap plot for the WOMBAT database with a top-1% cut-off. It will be seen that the same basic pattern of behaviour is obtained for all of the activity classes, *viz* a very large number of molecules that are retrieved by just a single search, and then rapidly decreasing numbers of molecules as more searches are considered. For example, if we consider the COX-2 searches, then there were (averaged over the ten different reference structures for this activity class) 2195 different molecules retrieved once in the 25 searches, 1749 different molecules retrieved twice, 1345 different molecules retrieved thrice etc. Entirely comparable plots are obtained with the top-1% cut-off for the MDDR activity classes (Figure 1(b)) and for the top-5% searches for both datasets (data not shown). For comparison with these data, selecting WOMBAT molecules completely at random with a probability of 0.01 (for top-1% searches) in the Binomial Distribution would yield 27,128 molecules that were retrieved once; however, the numbers then drop off very rapidly so that only a single molecule would be expected to be retrieved five times and no molecules at all for greater numbers of similarity searches.

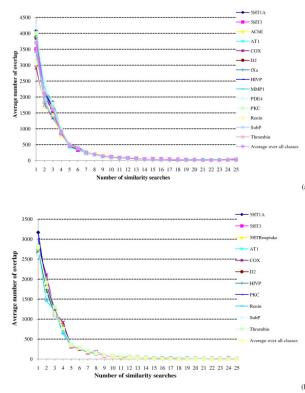


Figure 28.1: Figure 1. Search overlap using similarity fusion

**Search overlap using similarity fusion.** Plots of the mean numbers of molecules retrieved in a given number of similarity searches for: (a) WOMBAT top-1% searches; (b) MDDR top-1% searches.

The skewed nature of the data in Figure 1 suggests that there may a power law relationship between the overlap and the number of searches, with a few observations (*i.e.*, molecules being retrieved in the present context) occurring very frequently and the great majority occurring only once. Such relationships have been widely discussed in library and information science, where the Bradford, Lotka and Zipf distributions have been used for many years to discuss the dispersion of the scholarly literature, author productivity and word-usage frequencies respectively<sup>16,17</sup>. However, such relationships have been observed across the physical and social sciences: published applications include phenomena as diverse as the populations of cities, casualty figures in wars, and the sizes of lunar craters *inter alia*<sup>18</sup>, with Benz *et al.* reviewing applications in chemoinformatics<sup>19</sup>.

A power law relationship in the current context has the general form

where  $O$  is the overlap (see Experimental Methods),  $n$  is the number of similarity searches and  $a$  and  $b$  are constants. Plotting  $\log(O)$  against  $\log(n)$  should then give a straight line with a slope of  $-b$ , and this has been tested in Figure 2 for the top-1% searches, where the overlap figures have been averaged over all of the activity classes for simplicity and ease of viewing. There are clear deviations from straight line behavior in both plots, especially at the largest and smallest numbers of searches. This is not unexpected since inspection of the log-log plots that comprise Figure four of the review by Newman<sup>18</sup> shows that the twelve highly disparate datasets considered there all exhibit at least some degree of curvature analogous to that observed in Figure 2. The slopes ( $b$ ) and the  $r^2$  values for the WOMBAT and MDDR datasets (both top-12% and top-5%) are listed in the upper part of Table 1 in the column headed ‘Molecules’.

<sup>16</sup> Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction

<sup>17</sup> Informetrics

<sup>18</sup> Power laws, Pareto distributions and Zipf’s law

<sup>19</sup> Discovery of power-laws in chemical space

It will be seen that the slopes range from -1.75 (WOMBAT top-5%) to -2.17 (MDDR top-1%) and thus cluster around the value of -2 that characterizes a classical Lotka plot<sup>20</sup>. Mitzenmacher has noted that log-linear plots often give results that are comparable to log-log plots in power-law studies<sup>21</sup>. For the SF searches in Table 1, the log-linear plots gave better  $r^2$  values for the two top-5% results and worse values for the two top-1% values. Similarly inconsistent sets of values were obtained when the scaffold overlap and GF results were considered (*vide infra*).

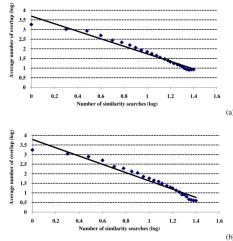


Figure 28.2: Figure 2. Search overlap using similarity fusion

**Search overlap using similarity fusion.** Log-log plots of the mean numbers of molecules retrieved in a given number of similarity searches for: (a) WOMBAT top-1% searches; (b) MDDR top-1% searches.

Figures 1 and 2 consider the overlap of individual molecules. Comparable analyses were conducted in which we counted the overlap of individual ring systems, specifically the Murcko scaffolds identified by the Pipeline Pilot software. Very similar results to those above were obtained, with the numbers of distinct scaffolds again dropping off very quickly with an increase in the number of searches. The  $b$  and  $r^2$  values for the scaffold log-log plots are included in the upper part of Table 1.

When applied to virtual screening, the Authority Effect would suggest that a given molecule is more likely to be active the more searches that retrieve it. From the results presented thus far, it is clear that multiple searches retrieve decreasingly small numbers of molecules; if the Effect holds then these decreasingly small numbers will contain increasingly large percentages of actives. That this enrichment occurs in practice is clearly demonstrated in Figures 3 and 4. There are often marked differences between the various activity classes comprising a dataset but the plots are at one in showing that the precision (see Experimental section) is very low for molecules retrieved by just a few searches but that it then increases very rapidly as the number of searches moves towards the maximum. As in the overlap experiments, the skewed nature of the data suggests that a power law relationship may be appropriate to describe the relationship. Averaging over all of the activity classes, the precision figures are shown as log-log plots in Figure 5. The plots all curve upwards to the right: fitting the log-log data to power and exponential trends, the former always gave the better fit, with the continuous curves in the figures representing a cubic relationship. Comparable results to those shown in Figures 3, 4, 5 were again obtained when we considered the active molecules' scaffolds that were retrieved, rather than the active molecules that were retrieved.

We hence conclude that a molecule is more likely to be active the more frequently it is retrieved when multiple similarity measures are available for carrying out a similarity search for a bioactive reference structure. The Authority Effect would thus appear to hold, at least for the datasets and similarity measures used here.

Turning now to the GF searches, the overlap plots that were obtained are very similar in form to those shown in Figures 1 and 2, and we have hence included just the top-1% log-log plots in Figure 6. The  $b$  and  $r^2$  values for these plots are included in the lower part of Table 1, and it will be seen that the magnitudes of the slopes are larger than in the upper part of this table, i.e., the numbers of molecules retrieved drops off more rapidly than in the similarity fusion searches. However, this drop-off is from a much larger starting point, as can be seen by comparing the intercepts on the y-axis in, e.g., Figures 2(a) and 6(a), i.e., the single similarity measure and 25 reference structures in the GF search identify a notably larger number of molecules than the 25 similarity measures and single reference structure in the SF search. This behaviour is detailed in Table 2, which shows the mean numbers of common molecules and common scaffolds for SF and GF searches using 1, 5, 10, 15, 20 and 25 similarity searches.

<sup>20</sup> An empirical examination of Lotka's Law

<sup>21</sup> A brief history of generative models for power law and lognormal distributions

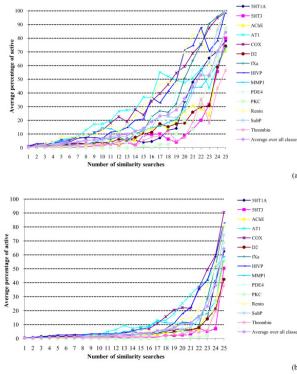


Figure 28.3: Figure 3. Search precision using similarity fusion

**Search precision using similarity fusion.** Plots of the percentage of the molecules retrieved in a given number of similarity searches that were active for: (a) WOMBAT top-1% searches; (b) WOMBAT top-5% searches.

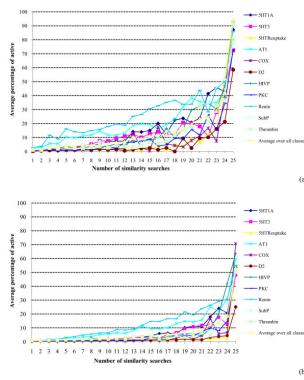


Figure 28.4: Figure 4. Search precision using similarity fusion

**Search precision using similarity fusion.** Plots of the percentage of the molecules retrieved in a given number of similarity searches that were active for: (a) MDDR top-1% searches; (b) MDDR top-5% searches.

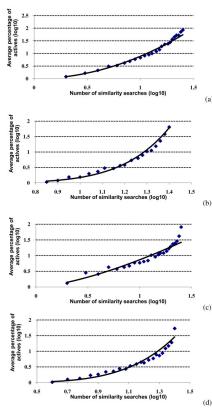


Figure 28.5: Figure 5. Search precision using similarity fusion

**Search precision using similarity fusion.** Log-log plots of the percentage of the molecules retrieved in a given number of similarity searches that were active for: (a) WOMBAT top-1% searches; (b) WOMBAT top-5% searches; (c) MDDR top-1% searches; (d) MDDR top-5% searches.

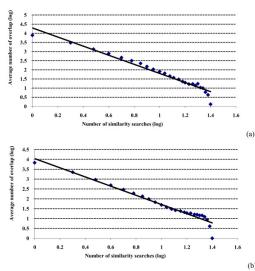


Figure 28.6: Figure 6. Search overlap using group fusion

**Search overlap using group fusion.** Log-log plots of the mean numbers of molecules retrieved in a given number of similarity searches for: (a) WOMBAT top-1% searches; (b) MDDR top-1% searches.

We believe that there are two factors that may explain the observed difference between SF and GF. First, the very different natures of the two types of search. In an SF search, the same reference structure is used in all 25 searches. Now, the substructures present within that structure are encoded in different ways by the five different fingerprints, and those encodings are processed in different ways by the five different similarity coefficients; however, it is the same basic structural information that is being used in each and every search. In the GF searches, conversely, a totally different reference structure (and hence different structural information) is used in each search. Second, some of the similarity measure components are quite closely related to each other; thus the Tanimoto and cosine coefficients are known to give very similar (though not monotonic) rankings<sup>22</sup>, and the Unity and Daylight fingerprints use a similar fragment encoding scheme. Thus, not only is the same basic structural information being used for all the SF searches, but in some cases this information is being processed in a similar manner. Taking these two effects together, the top-ranked molecules resulting from the SF searches hence have a greater degree of commonality than the top-ranked molecules from the GF searches, making it relatively easier for a molecule to be retrieved multiple times using SF (and relatively more difficult using GF). In like vein, a still more steeply angled plot (albeit one that is not based on a log-log relationship) is obtained when searching is simulated by drawing molecules at random using the Binomial Distribution, resulting in sets of molecules having minimal structural commonality.

The differences between the two types of fusion are still more marked when we consider the precision of the GF searches, as can be seen by comparing the results in Figures 7 and 8 with those in Figures 3 and 4. The general GF trend is for the precision to rise steeply (as in the SF searches) but then to fall rapidly away, giving an inverted bell-shape rather than the constantly increasing plots observed previously (see also the log-log plot for the MDDR top-1% data in Figure 9). The low precision values observed towards the right-hand parts of the Figures 7 and 8 plots follow naturally from the discussion above since if the 25 reference structures in a GF search are quite disparate then it is unlikely that many, or even any, molecules will be retrieved by large numbers of these reference structures. The precision (when averaged over the ten sets of GF searches for each activity class) is hence expected to be low, and there is some evidence to support this view from consideration of the individual activity classes. Specifically, there is a tendency for the more homogeneous activity classes (such as the renin inhibitors) to exhibit their maximum precision at larger numbers of searches than for the less homogeneous (i.e., more heterogeneous) activity classes, where we approximate the homogeneity of an activity class by the mean pair-wise similarity when averaged across all the pairs of molecules in that class. For example, consider the MDDR top-1% GF searches. We have ranked the eleven activity classes in order of decreasing mean pair-wise similarity and noted for each such class the number of similarity searches (in brackets) that gives the maximum precision. The resulting order is: Renin (23) > HIVP (12) > Thrombin (16) > AT1 (11) > SubP (11) > 5HT3 (12) > 5HTReuptake (9) > D2 (8) > 5HT1A (11) > PKC (11) > COX (6). Thus the differences in behavior between the GF and SF searches tend to increase the more diverse the activity class that is being sought, i.e., the more disparate the reference structures that are used for the searches. Support for this view comes from previous studies by Hert *et al.*<sup>23</sup> and by Whittle *et al.*<sup>24</sup>, who showed that GF, using the MAX and SUM fusion rules respectively, gave comparable levels of screening effectiveness to conventional similarity searching (and hence, by implication, to similarity fusion) when structurally homogeneous activity classes were searched; however,

<sup>22</sup> A fast algorithm for selecting sets of dissimilar molecules from large chemical databases

<sup>23</sup> New methods for ligand-based virtual screening: use of data-fusion and machine-learning techniques to enhance the effectiveness of similarity searching

<sup>24</sup> Enhancing the effectiveness of virtual screening by fusing nearest-neighbour lists: A comparison of similarity coefficients

there were noticeable differences in screening effectiveness (with GF the superior approach) when more heterogeneous classes were searched.

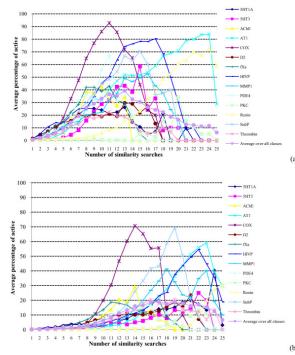


Figure 28.7: Figure 7. Search precision using group fusion

**Search precision using group fusion.** Plots of the percentage of the molecules retrieved in a given number of similarity searches that were active for: (a) WOMBAT top-1% searches; (b) WOMBAT top-5% searches.

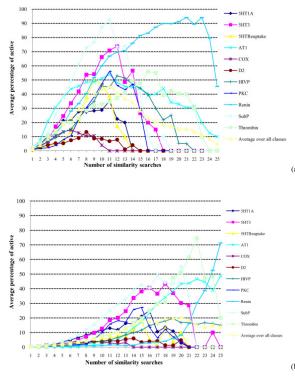


Figure 28.8: Figure 8. Search precision using group fusion

**Search precision using group fusion.** Plots of the percentage of the molecules retrieved in a given number of similarity searches that were active for: (a) MDDR top-1% searches; (b) MDDR top-5% searches.

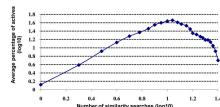


Figure 28.9: Figure 9. Search precision using group fusion

**Search precision using group fusion.** Log-log plot of the percentage of the molecules retrieved in a given number of similarity searches that were active for MDDR top-1% searches.

We hence conclude that the Authority Effect applies to GF searches when the reference structures are structurally quite similar; when this is not the case, it is applicable when relatively small numbers of reference structures are used, i.e., when meaningful numbers of molecules are being retrieved in all of the searches. It should be emphasized that this does not mean that GF is in some way inferior to SF as a technique for ligand-based virtual screening. First, the discussion here has focussed on the numbers of active molecules that are retrieved, without consideration of their diversity, and the previous studies mentioned above demonstrated the applicability of GF when structurally diverse molecules are sought<sup>2324</sup>. Second, it must be remembered that whilst we refer to SF and GF, practical implementations of these techniques entail a subsequent step in which a fusion rule combines the sets of nearest neighbours from the individual

searches. Finally, if 25 different, active reference structures were available, one should probably be using a more sophisticated, machine learning method<sup>25</sup> for database screening, e.g., a naive Bayesian classifier or a support vector machine, rather than simple, similarity-based approaches.

The results above show that Spoerri's Authority Effect holds - to some extent - for the chemical datasets and biological activity classes considered here. Specifically, a molecule is more likely to be active the more frequently it is retrieved in multiple similarity searches using a single reference structure or in multiple similarity searches using structurally similar multiple reference structures. This observation hence provides a justification for the use of data fusion methods in ligand-based virtual screening. In saying that, we must emphasise that our experiments have been conducted specifically to investigate the Authority Effect, and that rather different procedures are normally applied when data fusion procedures are used in operational virtual screening systems. For example, a common approach to GF is to use the MAX (or 1-NN) fusion rule, where the similarity for a database structure is taken to be the maximum of the similarities between that structure and each of the reference structures. Whittle *et al.* have shown that the numbers of retrieved actives increase approximately monotonically with the number of GF reference structures even when many of them are employed (see Figure six in Ref.<sup>24</sup>). Again, if one were to use SF in practice, one would choose similarity measures that differed in character, as exemplified by the work of Muchmore *et al.* on belief theory<sup>26</sup>, rather than the similar 2D fingerprint measures used here. Thus, while the results that we have presented provide a basis for the use of data fusion methods in principle, they do not provide a guide as to the effectiveness of any specific fusion method in practice.

It would clearly be desirable if we could not only demonstrate, but also rationalize, the frequency plots that we have presented. There has recently been interest in the underlying mathematical models that could generate power law distributions (see, e.g.,<sup>1821</sup>). Mitzenmacher has identified five broad types of generative model, and applied them to the analysis of both log-log and log-normal distributions<sup>21</sup>. In what follows, we apply a modification of one of his types - which he refers to as 'preferential attachment' - to the analysis of our virtual screening data.

Assume that there are  $n$  similarity search methods available, each of which models the possible activity of a molecule in a similar manner. Without loss of generality, assume also that the search methods for a given query (i.e., a single reference structure in similarity fusion or a set of reference structures in group fusion) are run sequentially. At each time step, a search is conducted of the  $M$  molecules in a database and a set of  $m$  possibly active molecules is returned (e.g., those in the top-5% of the ranking resulting from that search method). Thus, at time step 1, the first search is run and a set of  $m$  potentially active molecules is returned; at time step 2, a second search is run and another set of  $m$  possibly active molecules is returned, and so on. We now make the following assumption: that the second search returns the molecules that have been already returned by the first search with some probability proportional to  $\gamma$  ( $\gamma < 1$ ) while the rest of the molecules are returned with a probability proportional to  $(1 - \gamma)$ . Then, when the third search is conducted, a molecule is retrieved with probability proportional to the number of searches that have already returned that molecule. We are using here retrieval methods that are basically very similar (e.g., all using the same basic 2D substructural components and closely related association coefficients in a similarity fusion search), and it is hence not unreasonable to assume that a molecule satisfying the search criterion for one method is also likely to satisfy the criteria for other, related methods. If the different search methods are all equally similar to each other then a single  $\gamma$  is able to capture this similarity independently of the order of the methods used for searching the database. This is the strongest assumption we make here.

At the end of all the  $n$  searches, a total of  $n*m$  molecules will have been retrieved (though some of these will have been retrieved more than once). Let  $s$  denote the fraction of molecules returned by exactly  $s$  searches (i.e. the overlap between  $s$  similarity searches): we now demonstrate that  $s$  follows a power-law distribution. However, before providing a mathematical derivation of the distribution, we shall illustrate the approach using the example of four searches each retrieving three molecules as shown in Table 3. The set of molecules returned by three searches is {C}, while the set of molecules returned by two searches is {A, B, D}. If the current search (Search 5) returns any of A, B or D then the size of the set of molecules returned by three searches will increase by one. The chances of one of the three molecules being selected by Search 5 is  $\gamma(2*3/12)$ , since out of the 12 molecules already returned there are  $2*3 = 6$  instances of molecules already returned twice. If the current search returns C then the size of the set of molecules returned by exactly three searches will decrease by one since C now belongs to the set of molecules returned by exactly four searches. The chance that the molecule C is returned is  $\ln(\text{non-ascii\_6}) * (3*1/12)$  since out of the

<sup>25</sup> Machine learning in computational chemistry

<sup>26</sup> Application of belief theory to similarity data fusion for use in analog searching and lead hopping

12 molecules already returned there are  $3*1 = 3$  instances of molecules already returned thrice. If the growth of the set of molecules returned  $s$  times can be expressed mathematically then we shall be able to model the distribution of the fraction  $\text{O}_s$ , as we now demonstrate. In saying that, the reader should note that the following derivation excludes the special case of  $s = 1$ : this is not only to simplify the explanation but also because  $s = 1$  is the extreme end of the distribution, corresponding to molecules retrieved just once in any of the  $n$  searches and thus unlikely to be of practical interest in a screening context. The full derivation is presented by Mitzenmacher<sup>21</sup>.

Let  $\text{sX}(t)$  be a random variable describing the number of molecules returned by  $s$  searches at time step  $t$ . Then for  $s \geq 2$  the increase in  $\text{sX}(t)$  is described by the following formula

This is the probability that the current search returns one of the molecules retrieved in  $s-1$  of the previous searches. The denominator  $m*t$  is the total number of all retrieved molecules up to time step  $t$ ,  $(s-1)**\text{X}**s*-1$  is the total number of instances already retrieved  $s-1$  times, and thus  $(s-1)**\text{X}**s*-1/m*t$  is the fraction of the complete set of retrieved molecules that has been previously retrieved by  $s$  searches. Since the current search returns  $m$  molecules, the probability that a molecule is retrieved given that it has already been retrieved  $s-1$  times is hence

and

The decrease of  $\text{sX}(t)$  is described by the following formula

hence it is equal to the probability that the current search returns one of the molecules previously retrieved by  $s$  searches. Here,  $s*\text{X}$  is the total number of molecules already retrieved by  $s$  searches, and thus  $s*\text{X}/m*t$  is the fraction of the complete set of retrieved molecules that has been previously retrieved by  $s$  searches. The probability that a molecule is retrieved given that it has already been retrieved  $s$  times is hence

The growth of  $\text{X}$  is hence given approximately by

After all  $n$  searches have been executed  $\text{sX}(t) = s*m*t$ , i.e., the molecules retrieved by  $s$  searches constitute a fraction  $\text{O}_s$  of all the molecules retrieved. In the general case (for  $s \geq 2$ )

Solving for the fraction of molecules returned by  $s$  searches gives

For large  $s$ ,  $|nonascii\_9|*s+1 \sim |nonascii\_10|*s$  and thus,

Asymptotically, for the above to hold, we have  $a$ , giving a power law distribution for the fraction  $\text{O}_s$  and hence a rationale for the behavior observed in the MDDR and WOMBAT searches (see Figures 1, 2 and 6).

The reader should note that  $b = -(1+1/\gamma)$  can only give rise to exponents (slopes) that are less than -2, i.e.  $b = -(1+1/|nonascii\_12|)*|nonascii\_13|-2$ , for  $0 < \gamma < 1$ , so that some of the exponents (slopes) shown in the upper part of Table 1 cannot be explained by the proposed model. These are the slopes empirically derived from the data for the molecules and scaffolds using similarity fusion, regarding which we make two comments. First, the goodness of fit, as measured by  $r^2$ , is not as high as the goodness of fit for the rest of the empirical data, suggesting that the slope  $b$  may not be accurate enough. Second, the number of searches may not be large enough for accurate use of the approximation  $|nonascii\_16|*s+1 \sim |nonascii\_17|*s$  in the derivation. In particular, using the formula before this approximation and simulating the overlaps for different values of  $|nonascii\_18|$  based on the formulae above we obtain: for  $|nonascii\_19| = 0.9$ ,  $b = -1.923 > -2$ , while for  $|nonascii\_20| = 0.99$ ,  $b = -1.845 > -2$ . This can explain most of the slopes in Table 1 with the exception of those for WOMBAT top-5%.

It must be emphasized that this derivation considers only the overlap of the search outputs and says nothing about the precision of the searches. There is, however, an analogy that suggests one way in which the precision distributions might be modeled in future work. The overlap plots show that there is a distinct lack of consistency, i.e., that the different search methods generally retrieve very different sets of molecules. This situation has also been shown to pertain in many analogous retrieval contexts, such as the assignment of indexing terms<sup>27</sup>, the creation of links in hypertext systems<sup>28</sup> and the selection of search strategies [#B29]\_\*inter alia\*. In particular, it has been suggested that while indexers often differ considerably as to which indexing terms should be assigned to documents, where there is a high degree of consistency then this should result in enhanced search effectiveness. Whilst generally dubious of the correctness of this suggestion in practice, Cooper has shown, using a highly simplified model of the retrieval process,

<sup>27</sup> Inter-indexer consistency tests

<sup>28</sup> On the creation of hypertext links in full-text documents: measurement of inter-linker consistency

that effectiveness gains are obtainable in principle<sup>29</sup>, and it may be that analogous procedures could be applied to the modeling of the search results in Figures 3, 4, 7 and 8.

## 28.4 Conclusions

Data fusion, or consensus, methods are being increasingly used to combine the rankings that result from multiple virtual screening searches, with the hope that the combined ranking will contain a greater number of active molecules than will the original rankings. Our experiments with the MDDR and WOMBAT datasets demonstrate that different ranking methods result in markedly different sets of retrieved molecules, with the numbers of retrieved molecules common across a set of search outputs dropping off rapidly as the number of searches is increased. Specifically, we find an inverse log-log relationship between the numbers of searches carried out and the numbers of molecules common to those searches, with this power-law relationship being obtained when both similarity fusion and group fusion consensus approaches are used. However, whilst the numbers of retrieved molecules in common drop away very rapidly as more searches are carried out, the fraction of those that are active increases in the case of similarity fusion, or increases to a maximum before falling away in the case of group fusion. We also describe a generative model for the overlap between different screening searches, which provides a quantitative basis for the observed power-law behaviour. Thus, while the work presented here does not immediately suggest any new way of carrying out virtual screening, it does provide a rationale, both empirical and theoretical, for the use of a practice that is widely used in virtual screening, i.e. data fusion.

## 28.5 Experimental Methods

Testing the applicability of Spoerri's Authority Effect to virtual screening requires test datasets containing molecules of known (in)activity in one or more bioassays, and a range of different measures that can be used to carry out similarity searches on those datasets. Two separate datasets were used, these being the MDL Drug Data Report (MDDR) and World of Molecular Bioactivity (WOMBAT) databases. The versions used here were those that we have employed in many previous studies of virtual screening in this laboratory and that are described in detail by Arif *et al.*<sup>30</sup>. In brief, the MDDR file contained 102,535 molecules, with searches being carried out for 11 activity classes; and the WOMBAT file contained 138,127 molecules, with searches being carried out for 14 activity classes. The databases were searched using the two types of data fusion: similarity fusion (SF) and group fusion (GF).

In the SF experiments, ten compounds were randomly selected from each of the chosen activity classes to act as the reference structure for a similarity search, with each reference structure being searched using a total of 25 different similarity measures. These were obtained by combining five different 2D binary fingerprints with five different similarity coefficients. The 2D fingerprints used for describing the reference structure and the database structures were 166-bit MDL Keys, 1052-bit BCI bit-strings, 2048-bit Daylight fingerprints, 998-bit Unity fingerprints, and 1024-bit Pipeline Pilot ECFP\_4 fingerprints. The similarity coefficients used to measure the similarity between the reference structure's fingerprint and the database structures' fingerprints were the Cosine, Forbes, Russell-Rao, Simple Match and Tanimoto coefficients<sup>31</sup>. In the GF experiments, 25 compounds were randomly selected from each of the chosen activity classes to act as the reference structures, and these were then searched using ECFP\_4 fingerprints and the Tanimoto coefficient; ten such sets of 25 compounds were used for each activity class.

Given a specific reference structure, a similarity search was carried out using each of the different similarity measures in turn, yielding a total of 25 rankings (SF) or a similarity search was carried out using the ECFP\_4/Tanimoto measure for each of the 25 reference structures (GF). A threshold was then applied to each of the resulting database rankings to obtain the nearest neighbours of the reference structure, i.e., the top-ranked database structures. The thresholds here were the top-1% and the top-5% of the rankings. For each of the molecules in a database, a note was made as to the number of times that it was identified as a nearest neighbour, so that each database structure had an associated integer

<sup>29</sup> Is interindexer consistency a hobgoblin?

<sup>30</sup> Analysis and use of fragment occurrence data in similarity-based virtual screening

<sup>31</sup> Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings

value between 0 (meaning that it was retrieved in none of the searches) and 25 (meaning that it was retrieved in all of the searches). The resulting sets of integers, which are independent of the order in which the searches were carried out, were then processed to identify the *search overlap* and the *search precision*: the overlap measures the extent of the overlap between the search outputs, in terms of the numbers of molecules retrieved by some specific number of different searches; and the precision measures the percentage of the molecules retrieved by some specific number of different searches that are active. The nearest-neighbour data was collected for each reference structure in turn, and the results for each activity class were obtained by averaging over the set of ten searches for that class (and some of the results that are discussed are averaged over the set of 11 (for MDDR) or 14 (for WOMBAT) activity classes).

## 28.6 Competing interests

The authors declare that they have no competing interests.

## 28.7 Authors' contributions

NM carried out all of the virtual screening experiments with assistance from JH. EK carried out the mathematical power-law analysis. PW conceived the study, participated in its design and coordination, and drafted the manuscript with assistance from JH, NK and NM. All authors read and approved the final manuscript.

## 28.8 Acknowledgements

We thank the following: the Government of Malaysia for funding Nurul Malim; the European Union Seventh Framework Programme for funding Evangelos Kanoulas under the grant agreement FP7-PEOPLE-2009-IFF-254562. Dr S. Joshua Swamidass for suggesting the use of generative models; and Accelrys Inc., Daylight Chemical Information Systems Inc., Digital Chemistry Limited, the Royal Society, Tripos Inc. and the Wolfson Foundation for data, software and laboratory support.



# STRUCTURAL DIVERSITY OF BIOLOGICALLY INTERESTING DATASETS: A SCAFFOLD ANALYSIS APPROACH

## 29.1 Abstract

### 29.1.1 Background

The recent public availability of the human metabolome and natural product datasets has revitalized “metabolite-likeness” and “natural product-likeness” as a drug design concept to design lead libraries targeting specific pathways. Many reports have analyzed the physicochemical property space of biologically important datasets, with only a few comprehensively characterizing the scaffold diversity in public datasets of biological interest. With large collections of high quality public data currently available, we carried out a comparative analysis of current day leads with other biologically relevant datasets.

### 29.1.2 Results

In this study, we note a two-fold enrichment of metabolite scaffolds in drug dataset (42%) as compared to currently used lead libraries (23%). We also note that only a small percentage (5%) of natural product scaffolds space is shared by the lead dataset. We have identified specific scaffolds that are present in metabolites and natural products, with close counterparts in the drugs, but are missing in the lead dataset. To determine the distribution of compounds in physicochemical property space we analyzed the *molecular polar surface area*, the *molecular solubility*, the *number of rings* and the *number of rotatable bonds* in addition to four well-known Lipinski properties. Here, we note that, with only few exceptions, most of the drugs follow Lipinski’s rule. The average values of the *molecular polar surface area* and the *molecular solubility* in metabolites is the highest while the *number of rings* is the lowest. In addition, we note that natural products contain the maximum *number of rings* and the *rotatable bonds* than any other dataset under consideration.

### 29.1.3 Conclusions

Currently used lead libraries make little use of the metabolites and natural products scaffold space. We believe that metabolites and natural products are recognized by at least one protein in the biosphere therefore, sampling the fragment and scaffold space of these compounds, along with the knowledge of distribution in physicochemical property space, can result in better lead libraries. Hence, we recommend the greater use of metabolites and natural products

while designing lead libraries. Nevertheless, metabolites have a limited distribution in chemical space that limits the usage of metabolites in library design.

## 29.2 Background

An established idea of similarity-based virtual screening is that similar structures tend to have similar properties<sup>1</sup>. Diversifying the compound library collection for *in silico* and *in vitro* high-throughput screening without compromising biological activity remains an active research area. Chemical space is enormous but mostly biologically insignificant<sup>2</sup> and therefore, uninteresting from a drug design perspective. Given the large number of currently available chemical compounds in one of the largest public databases, PubChem<sup>3</sup>, it is impossible and irrational to screen all known compounds for potential ligands. One key methodology, fragment-based virtual screening (FBVS) or fragment-based drug discovery (FBDD), is an emerging area to identify novel, small molecules for preclinical studies. In FBDD, the starting points are small low molecular weight, drug-like fragments. Examples of such fragments are ring systems, functional groups, side chains, linkers and fingerprints.

Over the past decade, substructures contributing to drug-like or lead-like properties have governed library design<sup>4</sup>. In one of the pioneering works to understand the distribution of common fragments in drugs, Bemis and Murcko<sup>5</sup> fragmented a drug dataset (taken from the Comprehensive Medicinal Chemistry database) into rings, linkers, frameworks and side chains. Using two-dimensional topological graph-based molecular descriptors, they found 2506 different frameworks for a set of 5120 drug compounds, with the top 32 accounting for the topologies of 50% of the database compounds. They concluded a skewed distribution of molecular frameworks in drugs. Metabolite-likeness is increasingly being used as filter to design lead libraries similar to metabolites with better absorption, distribution, metabolism, elimination and toxicology (ADMET) properties<sup>6</sup>. Many recent studies have compared chemical space occupied by compounds of pharmaceutical interest<sup>789101112</sup>. Grabowski and Schneider<sup>7</sup> studied the molecular properties and chemotype diversity of drugs, pure natural products (NPs), and natural product derived compounds. Following the approach described by Bemis and Murcko<sup>5</sup>, they virtually dissected the molecules into frameworks, corresponding to scaffolds and side-chains. The drug dataset was ranked most structurally diverse, followed by marine and plant derived NPs, respectively. However, in contrast to the observation of Bemis and Murcko, that only 32 frameworks form the basis of nearly 50% of the compounds in CMC drug database, they found that 160 graph-based frameworks are needed to explain the chemotype of 50% of the compounds in the Collection of Bioactive Reference Analogues (COBRA) dataset<sup>13</sup> which contains drug-like reference molecules for ligand-based library design. In the same year, Siegel and Vieth<sup>8</sup> examined a set of 1386 marketed drugs and found that 15% of the drugs are embedded within other larger drugs, differing by one or more chemical fragments while 30% of drugs contain other drugs as building blocks. Recently, Franco *et al.*<sup>9</sup> analyzed scaffold diversity of 16 datasets of active compounds, targeting five protein classes, using an entropy-based information metric. They found that compounds targeted to the vascular endothelial growth factor receptor kinase, followed by compounds targeted to HIV reverse transcriptase and phosphodiesterase V, are maximally diverse. On the other hand, molecules in the glucocorticoid receptor, neuraminidase and glycogen phosphorylase  $\beta$  datasets are least diverse. Singh *et al.*<sup>10</sup> employed multiple criteria to compare libraries of drugs, small molecules and NPs, in terms of physicochemical properties, molecular scaffolds and fingerprints. The degree of overlap between libraries was assessed using the R-NN curve technique and the biologically relevant chemical space occupied by various compound datasets delineated. Hert *et al.*<sup>11</sup> compared a comprehensive dataset of 26 million

---

<sup>1</sup> Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors

<sup>2</sup> Chemical space and biology

<sup>3</sup> PubChem: a public information system for analyzing bioactivities of small molecules

<sup>4</sup> Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings

<sup>5</sup> The properties of known drugs. 1. Molecular frameworks

<sup>6</sup> ‘Metabolite-likeness’ as a criterion in the design and selection of pharmaceutical drug libraries

<sup>7</sup> Properties and Architecture of Drugs and Natural Products Revisited

<sup>8</sup> Drugs in other drugs: a new look at drugs as fragments

<sup>9</sup> Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure

<sup>10</sup> Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository

<sup>11</sup> Quantifying biogenic bias in screening libraries

<sup>12</sup> Physicochemical property space distribution among human metabolites, drugs and toxins

<sup>13</sup> Collection of Bioactive Reference Compounds for Focused Library Design

compounds (i.e. a representative sample of the full chemical space) with 25810 purchasable screening compounds, metabolites, and natural product dataset. They found that almost 1300 ring systems present in NPs are missing in current day screening or lead libraries and suggest introducing bias in screening libraries towards molecules that are likely to bind protein targets. Khanna and Ranganathan<sup>12</sup> compared current day drugs with toxics and metabolites and found that drugs are more similar to toxics than to metabolites in physicochemical property space distribution.

As discussed above, there are many studies analyzing the scaffolds and physicochemical properties of the various chemical datasets. However, none of the studies contains a comprehensive comparison of the compounds obtained from publically available datasets of human metabolites, toxics, drugs, natural products and currently used lead libraries. In addition, we believe that inclusion of the experimental compounds from National Cancer Institute open database and the recently released ChEMBL database would enhance our analysis and prove useful in recognizing fragments in biologically interesting compounds.

In this study, we aim to answer questions such as 1) What is the physicochemical property space distribution of compounds for the datasets under comparison? 2) Are there any pharmaceutically relevant scaffolds or fragments present in metabolites and natural products that are missing in current lead libraries? 3) Are there any preferred or frequently occurring fragments and scaffolds in these datasets? 4) What is the percentage similarity of the scaffolds and fragments found in drugs to those found in other datasets?

We found patterns of commonly occurring fragments using extended connectivity functional class fingerprint (FCFP\_4; details in Methods section). FCFP is a variant of extended connectivity atom type (ECFP) fingerprint, differing from the latter in the assignment of initial code<sup>14</sup>. The highly specific initial atoms types in ECFP fingerprints are replaced with more general atom types, with functional meaning in the FCFP fingerprints. For example, a single initial code is assigned for all halogen atoms in the FCFP fingerprints as they can often substitute each other functionally. In accord with their definition, ECFP fingerprints are a better choice to measure diversity. Therefore, we used ECFP fingerprints for diversity analysis while the more generic FCFP fingerprints were selected for Tanimoto analyses.

## 29.3 Results and discussion

Five different types of pharmaceutically relevant public molecular datasets were selected for this study: drugs, human metabolites, toxics, natural products and a sample of currently used lead compounds. Furthermore, we have also considered two popular small molecule databases *viz.* National Cancer Institute (NCI) database and ChEMBL database (details in the Methods section). Our results are presented in three sections, *viz.* preliminary analysis (measuring diversity and Tanimoto similarity), calculating physicochemical properties and scaffold analysis.

After carefully pruning and filtering the datasets, all the datasets were clustered (see Methods section) to avoid biased results due to overrepresentation of similar molecules.

### 29.3.1 1. Preliminary analysis

#### 1.1 Diversity analysis

In order to compare the diversity of features (fragments) present in each dataset, we have plotted the total number of non-redundant fingerprint features calculated, using ECFP fingerprints, up to order 8 (Figure 1). Our results indicate that overall, the ChEMBL dataset generates the maximum number of fragments and is highly diverse, while the metabolite dataset is the least diverse. From Figure 1a, we note that initially toxics outnumber other molecular datasets in generating features. This could be due to the high heteroatom content in toxics, resulting in large numbers of ECFP features generated during the first iteration step of fingerprinting. Similarly, the NCI dataset contains a large number of features during the initial iteration step of fingerprint feature generation. Metabolites, on the other hand, produce the least number of features, which suggests a limited occupancy of chemical space. Drugs were moderately diverse throughout and we find an increase in fragment diversity with increasing order of fingerprints.

<sup>14</sup> Extended-connectivity fingerprints

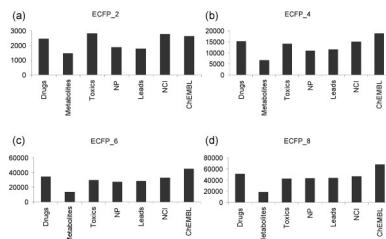


Figure 29.1: Figure 1. The number of non-redundant fingerprint features as a function of ECFP fingerprint order  
**The number of non-redundant fingerprint features as a function of ECFP fingerprint order.** Fingerprints of orders 2, 4, 6 and 8 for datasets comprising drugs, metabolites, toxics, natural products, leads, NCI and ChEMBL are presented.

## 1.2 Tanimoto analysis

The Tanimoto similarity coefficient compares two molecules, A and B, having  $A_N$  as the number of features in A,  $B_N$  as the number of features in B, and  $AB_N$  as the number of features common to both A and B as given in equation 1. This value is usually reported in the binary form, represented as  $b_T$ , and reported for simple comparisons between molecules. However, the Tanimoto coefficient can also encompass nonbinary data<sup>15</sup>; for example, if a fingerprint encodes not just the fragment incidences but also the frequencies of occurrence, as in the case of comparison between two compound datasets. In this case, the Tanimoto coefficient ( $nb_T$ ) is given by equation 2 where  $x_iA$ ,  $x_iB$  are the number of times the  $i^{th}$  fragment occurs in A and B, respectively, summed over  $n$  elements of each fingerprint.

We extend this concept to compare different datasets used in this study. To calculate how similar two datasets are, we first calculated the Scitegic Pipeline Pilot connectivity fingerprints, FCFP\_4 (details in the Methods section) for all the datasets. Subsequently, the sum of squares of the frequency of fingerprint features was calculated over the  $n$  elements for each dataset. Finally, the common features present in both datasets were counted and their frequencies multiplied, to determine  $nb_T$ .

For the five different datasets described in the Methods section, as well as the two reference datasets, NCI and ChEMBL, the Tanimoto coefficient values are shown in Table 1. We note that the FCFP fingerprint patterns (of order 4; FCFP\_4) found in drugs are most similar to toxics (FCFP\_4: 0.91) than to any other dataset, except for the fingerprint patterns found in reference datasets. On the other hand, drugs are least similar to metabolites (FCFP\_4: 0.72). These observations are consistent with our earlier study on smaller datasets<sup>12</sup>. We also note that ChEMBL contains more drug-like fragments than any other biologically relevant fragment type present in this study (FCFP\_4: 0.94). Further, we note that the fragments found in metabolites are least similar to the fragments present in NPs and lead dataset. Additionally, with the increasing order of fingerprints (FCFP\_6 and so on), although the number of fragments generated increases, the Tanimoto similarity coefficient values fall slightly for all the datasets compared (data not shown). This suggests an inverse relationship between the size of the fragment and the probability of its occurrence in two separate datasets, i.e. the larger the fragment, the less likely that it will be found in the two datasets being compared.

## 29.3.2 2. Physicochemical property analysis

### 2.1 Lipinski's properties for "rule of five" (Ro5) compliance

Ro5 has dominated drug design since 1997 and therefore, we believe it would be useful to analyze these datasets for compliance with the Ro5 test. Ro5 predicts passive and oral absorption based on log P, molecular weight, hydrogen bond donors and hydrogen bond acceptors. We report in Table 2, the percentage of molecules “failing” the Ro5 test, i.e. at least not meeting one condition of the Ro5 test. The results are comparable for both kinds of datasets, showing that randomly selected subsets are representative of the clustered datasets. Also, for the clustered datasets, initially, over 25% of drugs do not adhere to Ro5 while 68% of the metabolites are outside Lipinski's universe.

<sup>15</sup> Development of a compound class-directed similarity coefficient that accounts for molecular complexity effects in fingerprint searching

However, by removing lipids from metabolites we note that the percentage of molecules failing Ro5 test drops to 20% indicating that majority of the lipids do not follow Lipinski's rule. Further, we found that similar to drugs, only 26.5% of the toxics fail the Ro5 test. Lipinski's rule was originally designed to estimate bioavailability of compounds rather than toxicity. Therefore, the above result suggests that empirical rules such as Ro5 can be supplemented with toxicity information in order to reduce high attrition rates during drug discovery programs as has been reported in the literature<sup>16,17</sup>. Further, we found that only 16% of NPs failed Lipinski's test. Many other related studies on NPs have reported similar results<sup>7,18</sup>. Grabowski and Schneider<sup>7</sup> analyzed pure natural products (isolated exclusively from plants and terrestrial microorganisms) from MEGAbolite and Interbioscreen, natural products derivatives (isolated and synthesized natural products and derivatives from natural sources like plants, fungi, bacteria and sea organisms) from BioSpecs and marine natural products (isolated from sponges (41%), Coelenterates (21%), marine microorganisms and phytoplankton (10%)) from the literature. They found that 18% of the pure natural products, 30% of the marine natural products and only 8% of the natural product derived compounds violate Lipinski's rule, averaging 18.7%. While Grabowski and Schneider have reported results very similar to ours, Ganesan<sup>18</sup> analyzed a focused set of 24 natural products that were the starting point for marketed drugs in the 25-year period from 1981-2006 and found that 50% of these failed Lipinski's rule. In general, NPs do not necessarily abide by Lipinski's rule because they are thought to enter the human body not by passive diffusion but by more complex mechanisms such as active transportation, and so are not expected to comply with the rules for bioavailability. The probable explanation of our results could be the manner in which the NP dataset is pooled at the ZINC database. ZINC is a public database for commercially available compounds and NPs present in ZINC are pre-filtered to cover more drug-like space, contributing towards Ro5-like characteristics. Lead molecules on the other hand also did well in the Ro5 test as only 19.5% of the molecules violated one or more than one condition of the Lipinski's rule. This is in accordance with the lead-likeness concept proposed earlier<sup>19</sup> which states that leads should be simple, low molecular weight molecules and thus, should fall well within Lipinski's universe. Further, our results indicate that, NCI compounds follow Lipinski's rule more strictly than compounds present in ChEMBL dataset.

## 2.2 Lipinski's properties as boxplots

Box plots for Lipinski properties for random subsets are available from Figure 2. We find that the mean value for the molecular weight in the metabolite dataset is relatively low when compared to the other datasets such as drugs, leads and natural products. We also observe that the lead dataset is well within Lipinski's universe and covers a fair amount of drug space. Further, we find a noticeable difference in lipophilicity values of metabolites as compared to drugs and leads. The mean value of lipophilicity (measured as AlogP) suggests that metabolites prefer a hydrophilic environment. Our results are comparable to the recent study using similar datasets<sup>6</sup>. In this study, lipophilicity (measured by a similar parameter, clogD) for drugs, metabolites and library compounds showed that the distribution of library compounds is similar to that of drugs, but differ markedly from metabolites and that metabolites are more hydrophilic than both drugs and library compounds.

## 2.3 Other physicochemical properties

For a comprehensive study on the physicochemical property space distribution, we computed four more common whole molecule descriptors: the *molecular polar surface area*, the *number of rotatable bonds*, the *molecular solubility* and the *number of rings* (details in the Methods section). Distributions of these physicochemical properties as box plots are available from Figure 3. We note that metabolites show relatively higher solubility, higher molecular polar surface area but lower complexity (less rings, less rotatable bonds and lower molecular weight) compared to drugs. Further, our results indicate that, in general, NCI molecules are also low molecular weight compounds with less complexity and slightly higher solubility than drug molecules. In addition, we note that a large part of the ChEMBL database contain drug-like compounds with a biasness towards higher molecular weight and more complex molecules than drugs.

<sup>16</sup> Why drugs fail—a study on side effects in new chemical entities

<sup>17</sup> Theragenomic knowledge management for individualised safety of drugs, chemicals, pollutants and dietary ingredients

<sup>18</sup> The impact of natural products upon modern drug discovery

<sup>19</sup> Is there a difference between leads and drugs? A historical perspective

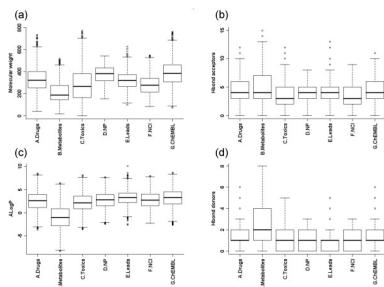


Figure 29.2: Figure 2. Box plots for the Lipinski physicochemical properties

**Box plots for the Lipinski physicochemical properties:** (a) molecular weight, (b) the number of hydrogen bond acceptors, (c) AlogP and (d) the number of hydrogen bond donors.

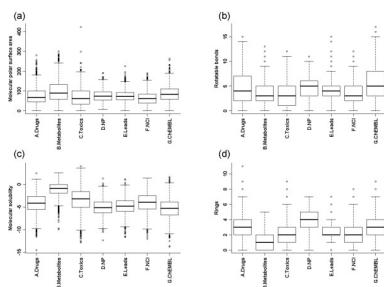


Figure 29.3: Figure 3. Box plots for other significant physicochemical properties

**Box plots for other significant physicochemical properties:** (a) molecular polar surface area, (b) the number of rotatable bonds, (c) molecular solubility and (d) the number of rings.

### 29.3.3 3. Scaffold or cyclic system analysis

It is quite informative to study the molecular frameworks while comparing different datasets of chemical compounds. Since the publication of Bemis and Murcko<sup>5</sup>, many attempts have been made to explore the chemical space occupied by bioactive scaffolds<sup>20</sup> as scaffold hopping remains an active area under research<sup>21</sup>. In this study, we define scaffolds as the core structure of the molecule after removing side chains but not the linkers, similar to the definition of *atomic frameworks* used by Bemis and Murcko. A detailed analysis of the total number of non-redundant scaffolds present in the different datasets is available in Table 3. The percentage of singletons (scaffolds occurring only once) relative to the total number of scaffolds in a dataset has also been reported. In addition, we have tabulated the proportion of non-redundant scaffolds containing aromatic and non-aromatic rings.

The drug dataset generates the largest proportion of non-redundant scaffolds (50.0%) relative to the dataset size, followed by the toxics (42%), ChEMBL (33.4%), leads (32%) and NCI dataset (28%). Exceptionally low number of scaffolds in metabolites (14.3%) and natural products (21.2%) suggest lower scaffold diversity in these datasets. The higher scaffold diversity in drugs could be attributed to the fact that drugs are derived from various biologically relevant compounds. The drug scaffold diversity is probably also due to the patenting requirements, to position functionality in the same way as an existing drug but outside of its patent space, that is often achieved by a minor change in the scaffold. Similarly, a large number of scaffolds in the toxic compound set is indicative of the high diversity of compounds with toxicity potential. Further, we note that distribution of scaffolds in all the datasets is highly skewed with large number of them occurring only once (singletons). In fact, almost 70% of the scaffolds in drugs, toxics, NCI and ChEMBL dataset occur only once. We also found that natural products comprise maximum number of recurring scaffolds (100 - % of singletons = 64%) followed by metabolites (38.9%) and leads (35.7%) suggesting that the compounds in these datasets revolve around certain preferred types of scaffolds. Our results agree with the recent study using similar natural product and drug dataset<sup>10</sup>. In their study, authors found high scaffold diversity in drugs (39.7%) while low

<sup>20</sup> Interactive exploration of chemical space with Scaffold Hunter

<sup>21</sup> Scaffold-hopping potential of fragment-based de novo design: the chances and limits of variation

diversity in natural products (17.9%) which is in accordance with our results. By counting the number of aromatic rings in non-redundant scaffolds, we note that metabolites contain least number of aromatic rings (only 47.3% contain one or more aromatic rings in a scaffold) as compared to other datasets. 85% of the drugs on the other hand have scaffolds with aromatic rings. Furthermore, we note that 97.4% of the scaffolds found in lead dataset contain aromatic rings. There seems to be a bias towards aromatic ring containing scaffolds in presently used lead libraries.

The top five scaffolds and their relative percentages based on the total number of scaffolds found in each dataset are shown in Figure 4. Benzene is the most abundant scaffold system in all the datasets, particularly in metabolites (over 36%). Apart from metabolites, toxics (15%) and NCI compounds (13%) also contain benzene in high percentages. Drugs and leads, on the other hand contain benzene in moderate amounts (10% and 7% respectively). While benzene is the most common scaffold type in NP (2.2%) and ChEMBL datasets (3.4%), the relative abundance of benzene in these datasets is far lower than that in the other datasets. Following benzene, pyridine is the second most commonly occurring scaffold type in the top five scaffolds. It is found in four out of seven datasets: metabolites (5.2%), drugs (1%), leads (1%), and NCI (1.2%). We also note that steroid derivatives are largely present in drugs and NPs. Similarly, most of the fused large scaffolds are found in NPs (four of the top five scaffolds) followed by drugs and the ChEMBL dataset. Metabolites, on the other hand, seem to prefer smaller, less complex systems. Likewise, toxics and lead compounds also have few complex fused systems. Other commonly occurring scaffold systems are purine and purine derivatives (found mainly in metabolites and ChEMBL dataset), imidazole and biphenyls.

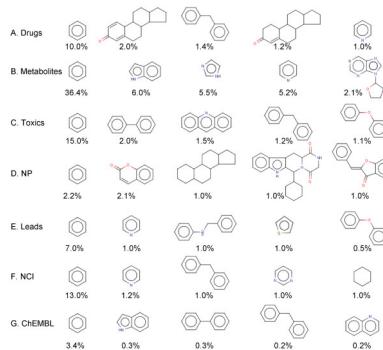


Figure 29.4: Figure 4. Top 5 scaffolds derived from A. drugs, B. metabolites, C. toxins, D. natural products, E. leads, F. NCI and G. ChEMBL

#### Top 5 scaffolds derived from A. drugs, B. metabolites, C. toxins, D. natural products, E. leads, F. NCI and G. ChEMBL.

The extent of occurrence of the scaffold relative to the total number of scaffolds in the dataset (as %) are listed.

In Table 4, we tabulate the percentages of non-redundant shared scaffolds between pairs of different datasets. From Table 4 we note that drugs and metabolites share 6% of the total non-redundant scaffolds whereas NPs, leads and toxics share overall 2.4%, 1.4% and 7.5% of scaffolds with drugs, respectively. It is interesting to note that metabolites and leads do not share as many scaffolds (0.3%) as drugs and metabolites (6%). Due to the uneven size of the datasets, we have also reported the contribution of each dataset to the set of shared scaffolds. We find that of the total 296 non-redundant scaffolds found in metabolites (Table 3), 123 (42%) are shared by drugs whereas only 68 (23%) are shared by the lead dataset. This suggests that lead compounds need further optimization to become more metabolite-like. Similarly, there seems to be little overlap between the scaffolds of presently used lead libraries and NPs (2.1%). Since metabolites and NPs are recognized by at least one protein in the biosphere, they seem to be appropriate candidates in lead library design. Our results however, indicate that neither metabolites nor NP scaffolds are being sampled enough while designing lead libraries. In addition, we note that over 7% of scaffolds are shared between drugs and toxics while metabolites and toxics share over 6% of the scaffolds, suggesting the recurrence of common scaffolds between these datasets. Compounds in the NCI and ChEMBL datasets are quite diversified; however, the NCI dataset clearly contains more toxic scaffolds than the ChEMBL dataset. Furthermore, we note that large part of the drug scaffold space is present in NCI (45%) and ChEMBL (72%) implying that these datasets cover good amount of drug-like compounds. We also note that a large part of metabolite scaffold space is present in natural product (47%), NCI (78%) and ChEMBL (73%) datasets.

We expect that lead libraries biased towards molecules that biological targets have evolved to recognize, would yield

better hits rates, than unbiased or universal libraries. Metabolites and NPs could potentially provide suitable lead molecules. Consequently, we further analyzed these datasets for the type of scaffolds that are currently missing in lead libraries. In fact, we note a very slight overlap in the scaffold space of lead libraries and these datasets as discussed above. We therefore, suggest that with the optimum coverage of biologically relevant scaffold space, hit rates in high throughput screening experiments can be improved. We report a set of scaffolds that occur in NPs (Figure 5) and metabolites (Figure 6), with a minimum Tanimoto similarity of 0.9 to the scaffolds found in drugs, which are actually missing in currently used lead datasets.

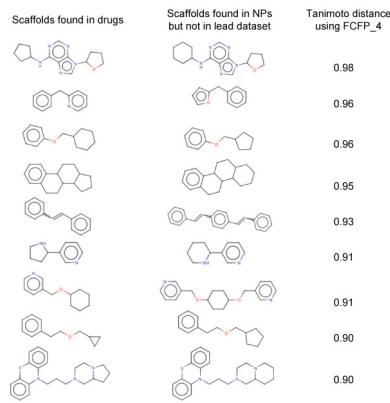


Figure 29.5: Figure 5. A set of scaffolds present in NPs but are missing in lead dataset

**A set of scaffolds present in NPs but are missing in lead dataset.** The Tanimoto distance of these scaffolds with the closest counterparts in drugs is also reported.

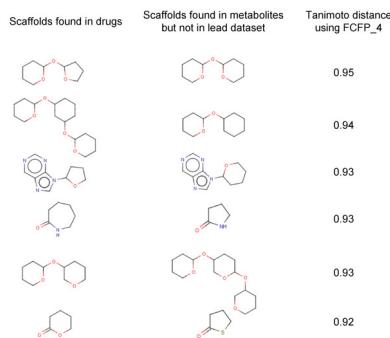


Figure 29.6: Figure 6. A set of scaffolds present in metabolites but are missing in lead dataset

**A set of scaffolds present in metabolites but are missing in lead dataset.** The Tanimoto distance of these scaffolds with the closest counterparts in drugs is also reported.

## 29.4 Conclusions

In this study, we have carried out a detailed analysis of commonly occurring fragments in various datasets of biological interest. Dataset comparison using the Tanimoto coefficient shows that drugs and toxics share a large number of topological fragments whereas drugs are least similar to metabolites than to any other dataset studied. However, in scaffold analysis we found that drugs and metabolites share 6% of the total non-redundant scaffolds, i.e. over 42% of the metabolite scaffolds are present in drugs, whereas only 23% of the metabolite scaffolds are represented in current leads. This shows that although drugs and metabolites share many scaffolds, they largely differ in topological fragment space. Further, we conclude that current lead libraries do not cover much of metabolite scaffold space.

Library design is a multi-class optimization problem. It often presents a trade-off between several factors, including diversity and ADMET properties. Since metabolites and NPs are already optimized by millions of years of evolution to bind to at least one biological macromolecule therefore, it is highly likely that libraries designed based on the scaffolds and fragments occurring in metabolite and NP space will result in molecules with better ADMET properties. Hence, the use of metabolites and NPs while designing lead libraries would be beneficial. However, metabolites occupy a limited space in chemical universe that limits their usage in library design.

From physicochemical properties analysis, we note that there is a need to diversify present day lead libraries in order to optimize the coverage of chemical space. We also note that with the exception of few compounds, most of the drug molecules follow Lipinski's rule whereas over 68% of metabolites are outside Lipinski's universe. On a closer examination of metabolites, we found that the compounds that do not follow Lipinski rule are mainly lipids and large molecules. Further, we note that lipid-free metabolite dataset contains low molecular weight and less complex molecules as compared to other datasets. Our studies on scaffolds systems suggest that drugs are most diverse (50% scaffolds relative to the dataset size) and prefer aromatic to non-aromatic ring-containing scaffolds. Metabolites, on the other hand, have a very narrow distribution of scaffolds (only 14.3% scaffolds relative to the dataset size) of which 38.9% recur. The exceptionally low number of cyclic systems in metabolites implies lower scaffold diversity in metabolites. Further, we confirm earlier reports of skewed distribution of scaffolds, with many more singletons than recurring scaffolds.

## 29.5 Methods

### 29.5.1 Preparation of datasets

Five different types of biologically relevant molecular datasets have been considered in this study. Beside these, the contents of public databases like NCI and ChEMBL were also analyzed. Table 5 presents a summary of all the databases used in this study. The drug dataset was assembled by merging molecules obtained from the DrugBank<sup>22</sup> and a subset of Kyoto Encyclopedia of Genes and Genomes database (KEGG DRUG)<sup>23</sup>. DrugBank is a comprehensive resource on drugs and includes over 1350 FDA-approved small drugs. KEGG is a bioinformatics resource and currently provides 19 databases; we used the KEGG DRUG subset as it contains all the drugs approved in the USA and Japan. It not only contains prescription drugs but also “over-the-counter” (OTC) drugs. Similarly, for metabolite dataset we used the Human Metabolome Database (HMDB)<sup>24</sup>, HumanCYC<sup>25</sup> database and BiGG<sup>26</sup> database. HMDB contains information on nearly 8,000 metabolites found in the human body. HumanCYC is a bioinformatics database that combines human metabolic pathway and genome information, providing KEGG, PubChem and ChEBI identifiers for the metabolites present in this database. BiGG stores manually annotated human metabolic network information, with links to KEGG metabolites.

Likewise, for the toxics dataset, compounds from various public sources were integrated to make a single dataset focusing largely on carcinogenic molecules. The Distributed Structure-Searchable Toxicity (DSSTox) Carcinogenic Potency Database<sup>27</sup> contains experimental results and carcinogenicity information for 1547 substances tested against different species. Contrera *et al.*<sup>28</sup> published a dataset of 282 human pharmaceuticals obtained from FDA database for carcinogenicity studies on mouse and rat. They reported 125 (44% of the above 282) of the positive chemicals that were used in this study. Toxicology Excellence for Risk Assessment (TERA) is an independent non-profit organization dedicated to the public health. Since 1996, TERA has maintained an International Toxicity Estimate for Risk database<sup>29</sup> which provides chronic human risk assessment data from organization around the world for over 650 chemicals

<sup>22</sup> DrugBank: a knowledgebase for drugs, drug actions and drug targets

<sup>23</sup> KEGG for linking genomes to life and the environment

<sup>24</sup> HMDB: a knowledgebase for the human metabolome

<sup>25</sup> Computational prediction of human metabolic pathways from the complete human genome

<sup>26</sup> BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions

<sup>27</sup> Distributed structure-searchable toxicity (DSSTox) public database network: a proposal

<sup>28</sup> Carcinogenicity testing and the evaluation of regulatory requirements for pharmaceuticals

<sup>29</sup> International Toxicity Estimate for Risk database (TERA)

<sup>30</sup>. Finally, ~1000 molecules with medium and high toxicity were downloaded from the SuperToxic database <sup>31</sup>. The dataset for NPs was obtained from the ZINC database <sup>32</sup>. These molecules can be searched under the subset tab, as “Meta subsets”. For lead dataset, we merged two independent screening sets obtained from BioNET <sup>33</sup> and Maybridge database <sup>34</sup>. The molecules in these two databases are well diversified and we integrated them to form a dataset of lead compounds as found in pharmaceutical collections. Further, we included molecules from NCI open database <sup>35</sup>. The latest September 2003 release of the database stores 260071 organic compounds tested by NCI for anticancer activity. Since many of the compounds are experimental, have not been tested for human consumption and covers high diversity therefore, we believe it would be good choice to include this dataset in our study. One other public dataset, ChEMBL <sup>36</sup> was used as the reference dataset for biologically interesting molecules. ChEMBL is a chemogenomics data resource with over 8000 targets and about 622,884 bioactive compounds.

All datasets are current as of 10-November-2010.

## 29.5.2 Cleaning and processing of the datasets

We followed a standard cleaning procedure (see additional file <sup>6</sup>). Clusters were generated, using the Cluster “Clara” algorithm embedded in the Pipeline Pilot (PP) software <sup>37</sup> by employing an atom type fingerprint as a chemical descriptor and Euclidean distance was the distance metric selected. Cluster centers served as the representatives for clusters containing more than one molecule while singletons were directly used as cluster centers. This resulted in 30% decreases of each dataset. Upon further analysis, we found that clustered metabolite set contains lipids in large numbers. In order to remove the bias towards lipids and large molecules, we filtered out lipids resulting in 2072 molecules in the “lipid-free” metabolite dataset, used for analysis in this study.

Additional file 1

**Supplementary figure S1.** Flowchart adapted for the overall methodology.

Click here for file

To simplify the analysis, we randomly selected 2000 compounds from each of the clustered datasets and lipid-free metabolite dataset in case of metabolites. The majority of the analysis was carried out using the clustered datasets and lipid-free metabolite dataset, except for preliminary analysis, where these randomly selected molecules were used and in the case of Ro5 test, where both datasets were compared.

## 29.5.3 Molecular descriptors

All the descriptors were calculated using PP. Beside the four Lipinski properties: molecular weight, the number of hydrogen bond acceptors, AlogP (a hydrophobicity measure) and the number of hydrogen bond donors <sup>4</sup>, other descriptors such as molecular polar surface area (MPSA), molecular solubility (MS), the number of rings (NR) and the number of rotatable bonds (NRB) were also computed. AlogP was calculated using the Ghose-Crippen method <sup>38</sup> which takes into account the group’s contribution to Log P. MPSA is defined as the sum over all the polar atoms. This descriptor is correlated with drug transport capabilities and is important in penetrating the blood-brain barrier. The NRB is a direct measure of the flexibility of molecules thus related to MPSA. Binary descriptors (ECFP\_4 and FCFP\_4) were calculated using a structural property calculator embedded in PP. Initially, each atom is assigned a code based on its properties and connectivity. With increasing iteration, each atom code is combined with the code of its

<sup>30</sup> Resources for global risk assessment: the International Toxicity Estimates for Risk (ITER) and Risk Information Exchange (RiskIE) databases

<sup>31</sup> SuperToxic: a comprehensive database of toxic compounds

<sup>32</sup> ZINC—a free database of commercially available compounds for virtual screening

<sup>33</sup> BioNET

<sup>34</sup> Maybridge

<sup>35</sup> National Cancer Institute (NCI)

<sup>36</sup> ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr

<sup>37</sup> SciTegic Pipeline Pilot

<sup>38</sup> Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions

immediate neighbours to produce the next order code. This process is repeated until the desired number of iterations has been achieved, typically to four iterations, generating ECFP\_4, or FCFP\_4 fingerprints.

#### 29.5.4 Cyclic systems

In addition to examining the physicochemical properties, each dataset was also explored for the frequent scaffold systems. We used an inbuilt PP protocol to identify the most common fragments, by setting “FragmentType” to MurckoAssemblies and adjusting “MaxFragSize” parameter at the required level.

### 29.6 Competing interests

The authors declare that they have no competing interests.

### 29.7 Authors' contributions

VK curated the datasets and conducted the analysis work; SR directed the study and both authors prepared and approved the manuscript.

### 29.8 Acknowledgements

VK is grateful to Macquarie University for the award of MQRES research scholarship.



# STATISTICAL FILTERING FOR NMR BASED STRUCTURE GENERATION

## 30.1 Abstract

The constitutional assignment of natural products by NMR spectroscopy is usually based on 2D NMR experiments like COSY, HSQC, and HMBC. The difficulty of a structure elucidation problem depends more on the type of the investigated molecule than on its size. Saturated compounds can usually be assigned unambiguously by hand using only COSY and  $^{13}\text{C}$ -HMBC data, whereas condensed heterocycles are problematic due to their lack of protons that could show interatomic connectivities. Different computer programs were developed to aid in the structural assignment process, one of them C

## 30.2 Findings

Nuclear Magnetic Resonance is the most common tool used for the structure elucidation of new compounds. The used 2D NMR experiments like COSY, HSQC, and  $^{13}\text{C}$ -HMBC deliver correlation information between atoms that can be translated into connectivity information. Out of these, correlation information from COSY and HSQC experiments can be transcribed directly into connectivity between atoms. But the  $^{13}\text{C}$ -HMBC correlations need more attention because of their ambiguity and complexity. Hence the difficulty of the structure elucidation problem depends more on the type of the investigated molecule than on its size. Saturated compounds can usually be assigned unambiguously using mainly COSY and some  $^{13}\text{C}$ -HMBC data, whereas condensed heterocycles are problematic due to their lack of protons that could show interatomic connectivities. This ambiguity has driven the development of different software

packages to aid in the interpretation of the  $^{13}\text{C}$ -HMBC correlation data [12345678910111213141516171819](#) as much as the development of additional correlation experiments <sup>2021</sup>.

When the observed connectivity information is used as input for the structure generation program C<sup>3222324</sup> it will create all compatible constitutional assignments. In the case of unsaturated molecules C<sup>2526</sup> integrated to C‘<<http://cocon.nmr.de>>‘\_), since it uses data protected by Intellectual Property. A different way of handling the result set had to be chosen, and the statistical filter was implemented.

The idea behind the filter is, to compare the suggested constitutions against existing molecules, like the ones contained in the PubChem (PubChem can be found at <http://pubchem.ncbi.nlm.nih.gov/>) database. For each C‘<<http://www.chembiogrid.org/cheminfo/smi23d/>>‘\_) has been used to generate 3D coordinates for almost 13M compounds contained in PubChem (The corresponding 3D coordinates generated by smi23d can be found at <http://www.chembiogrid.org/cheminfo/p3d/>; the error observed is ~ 0.4% (= 53.000) false negatives for 13M compounds) and succeeded on generating coordinates for 99.6% of the molecules contained in the Database. The filtering application actually uses smi23d to generate 3D coordinates for all constitutional assignments generated by C\*\*1\*\* and 2. For 3 the longest running time was 3 days for the generation of the 523.668 constitutional assignments using COSY,  $^{13}\text{C}$ -HMBC correlations and open atom types. A webpage allowing direct access to the results of the structure generator runs presented here has been set up on the <http://cocon.nmr.de/StatisticalFilter/> (The results are also mirrored at <http://science.jotjot.net/StatisticalFilter/>).

Ascomycin is a well known ethyl derivative of Tacrolimus, it serves as example of a large natural product, featuring 43 Carbon atoms. Using experimental NMR correlation data (COSY and  $^{13}\text{C}$ -HMBC correlations) together with fixed atom types, C

The results change with the second example molecule, Aflatoxin B1 with 17 Carbon atoms. Using COSY and  $^{13}\text{C}$ -HMBC data alone, C<sup>1</sup>) all contain oxet-2-one as structural element, that can be found in 6 basic variations in 85 compounds in PubChem (see Figure 2). Until now, no natural product has been described with this substructure. The numbers of results for the different C\*\*1\*\* and 2 are summarized in table 1. Oroidin **3** has been frequently used for the demonstration of  $^{13}\text{C}$ -HMBC correlations lead to a total of 523.668 possible constitutional assignments, out of which only 1904 belong to the correct atom type combination. After the statistical filtering there are still 252.566 respectively 1486 suggestions left. In this case the reliable structure elucidation by NMR needs  $^{15}\text{N}$ -HMBC or 1,1-ADEQUATE correlations. For calculations with open atom types, only when using both kinds of correlation information and filtering, a reasonable amount of 275 suggested constitutions is generated.

<sup>1</sup> Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation

<sup>2</sup> Computer-assisted structure elucidation: Application of CISOC-SES to the resonance assignment and structure generation of betulinic acid

<sup>3</sup> COCON: From NMR correlation data to molecular constitutions

<sup>4</sup> Computer-aided structure elucidation of organic compounds: Recent advances

<sup>5</sup> Fuzzy structure generation: A new efficient tool for computer-aided structure elucidation (CASE)

<sup>6</sup> Computer-aided determination of relative stereochemistry and 3D models of complex organic molecules from 2D NMR spectra

<sup>7</sup> Automated structure elucidation of two unexpected products in a reaction of an alpha,beta-unsaturated pyruvate

<sup>8</sup> Recent developments in automated structure elucidation of natural products

<sup>9</sup> Applications of a HOUDINI-based structure elucidation system

<sup>10</sup> SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry

<sup>11</sup> Recent advancements in the development of SENECA, a computer program for Computer Assisted Structure Elucidation based on a stochastic algorithm

<sup>12</sup> Computer aided method for chemical structure elucidation using spectral databases and C-13 NMR correlation tables

<sup>13</sup> SESAMI: An integrated desktop structure elucidation tool

<sup>14</sup> LUCY - A program for structure elucidation from NMR correlation experiments

<sup>15</sup> Combinatorial Problems in the Treatment of fuzzy C-13 NMR Spectral Information in the Process of Computer-Aided Structure Elucidation - Estimation of the Carbon-Atom Hybridization and Alpha-Environment States

<sup>16</sup> Computer-Assisted Structure Elucidation for Organic-Compound

<sup>17</sup> Computer Method of Fragmentary Formula Prediction of an unknown by its Mass and NMR-Spectra

<sup>18</sup> Structure Elucidation of organic-compounds aided by the Computer-Program System Scannet

<sup>19</sup> Computer-Aided Spectral Assignment in NMR Spectroscopy

<sup>20</sup> ADEQUATE, a new set of experiments to determine the constitution of small molecules at natural abundance

<sup>21</sup> Impact of the H-1,N-15-HMBC experiment on the constitutional analysis of alkaloids

<sup>22</sup> 2D-NMR-guided constitutional analysis of organic compounds employing the computer program COCON

<sup>23</sup> A COCON analysis of proton-poor heterocycles-Application of carbon chemical shift predictions for the evaluation of structural proposals

<sup>24</sup> Computer-assisted constitutional assignment of large molecules: COCON analysis of ascomycin

<sup>25</sup> Novel methods of automated structure elucidation based on C-13 NMR spectroscopy

<sup>26</sup> Validation of structural proposals by substructure analysis and C-13 NMR chemical shift prediction

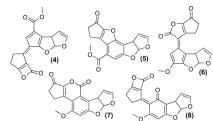


Figure 30.1: Figure 1. The constitutions 4-8 shown here are excluded by the statistical filter  
**The constitutions 4-8 shown here are excluded by the statistical filter.** Each constitution appears with two Different  $^{13}\text{C}$  chemical shift assignments in the solution set.

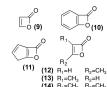


Figure 30.2: Figure 2. Basic variations of the structural element Oxet-2-one that is excluded by the statistical filter, as found in 85 hits from PubChem

**Basic variations of the structural element Oxet-2-one that is excluded by the statistical filter, as found in 85 hits from PubChem.**

When the  $^{15}\text{N}$ -HMBC correlations and fixed atom types are added to the COSY and  $^{13}\text{C}$ -HMBC based calculation the statistical filter excludes only the constitutional assignments containing the 1-nitro-prop-2-en-Z-ylidene substructural element (see Figure 3). According to Beilstein, this structural element appears only 14 times, always in conjunction with an aromatic ring, as depicted in Figure 4. When 1,1-ADEQUATE correlations are added instead, and atom types are fixed, the filter excludes 16 constitutions, shown in Figure 5. All resulting numbers of constitutional assignments for the Different combinations of correlation data are summarized in table 2.

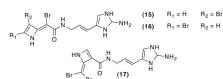


Figure 30.3: Figure 3. Constitutional assignments excluded by the statistical filter when the structure generator runs with COSY, HMBC,  $^{15}\text{N}$ -HMBC correlation data and atom types for Oroidin  
**:sup: '15'N-HMBC correlation data and atom types for Oroidin Constitutional assignments excluded by the statistical filter when the structure generator runs with COSY, HMBC, .**

The results from tables 1 and 2 show that the filter excludes more constitutional assignments when the atom types are undefined (45% - 65%) than when the atom types are defined (~ 20%). In neither case the correct constitutions were excluded, and in the case of Ascomycin the filter did not exclude any constitutional assignment. The calculation time increases depending on the number of possible constitutional assignments, as smi23d runs about 0.5 s per structure. This explains the observed 3 days for the generation of the 523,668 constitutional assignments for Oroidin using COSY,  $^{13}\text{C}$ -HMBC correlations and open atom types, C2), looking at a mere 275 constitutional assignments instead of 716 is a considerable improvement. When looking at the excluded constitutions, and checking for the common structural elements, it turns out that they are not stable or do not exist in PubChem. Run times for the filter could be cut down in the future by restricting the MD run to just the generation of the parameters, but this would need changing the existing smi23d software package, and throwing away the possibility of improved visualization of the results with the 3D structures. The new statistical filtering presented here has already been made available in **1**, Aflatoxin B1 **2** and Oroidin **3** (Figure 6), example molecules that have already been used on other occasions. The molecules are available as examples on the

### 30.3 Availability

The <http://cocon.nmr.de>.



Figure 30.4: Figure 4. The 14 molecules found in Beilstein containing the 1-nitro-prop-2-en-Z-ylidene substructural element all have the substitution pattern of 18 and 19

**The 14 molecules found in Beilstein containing the 1-nitro-prop-2-en-Z-ylidene substructural element all have the substitution pattern of 18 and 19.**  $\text{R}_2$  is either a polyaromatic or polyhalogenic substituent.

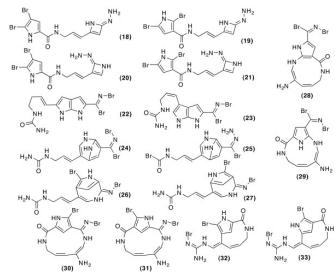


Figure 30.5: Figure 5. Constitutional assignments excluded by the statistical filter when the structure generator runs with COSY, HMBC, 1,1-ADEQUATE correlation data and atom types for Oroidin

**Constitutional assignments excluded by the statistical filter when the structure generator runs with COSY, HMBC, 1,1-ADEQUATE correlation data and atom types for Oroidin.**

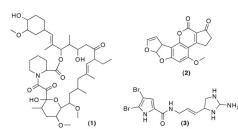


Figure 30.6: Figure 6. Ascomycin 1, Aflatoxin B1 2 and Oroidin 3 are used to evaluate the statistical filter  
**Ascomycin 1, Aflatoxin B1 2 and Oroidin 3 are used to evaluate the statistical filter.**

## 30.4 Competing interests

The author declares that they have no competing interests.

## 30.5 Authors' contributions

JJ maintains the

## 30.6 Acknowledgements

The authors wishes to thank Rainer Haessner and the Technische Universität München for providing the Hardware for the



# PUBCHEM3D: A NEW RESOURCE FOR SCIENTISTS

## 31.1 Abstract

### 31.1.1 Background

PubChem is an open repository for small molecules and their experimental biological activity. PubChem integrates and provides search, retrieval, visualization, analysis, and programmatic access tools in an effort to maximize the utility of contributed information. There are many diverse chemical structures with similar biological efficacies against targets available in PubChem that are difficult to interrelate using traditional 2-D similarity methods. A new layer called PubChem3D is added to PubChem to assist in this analysis.

### 31.1.2 Description

PubChem generates a 3-D conformer model description for 92.3% of all records in the PubChem Compound database (when considering the parent compound or salts). Each of these conformer models is sampled to remove redundancy, guaranteeing a minimum (non-hydrogen atom pair-wise) RMSD between conformers. A diverse conformer ordering gives a maximal description of the conformational diversity of a molecule when only a subset of available conformers is used. A pre-computed search per compound record gives immediate access to a set of 3-D similar compounds (called “Similar Conformers”) in PubChem and their respective superpositions. Systematic augmentation of PubChem resources to include a 3-D layer provides users with new capabilities to search, subset, visualize, analyze, and download data.

A series of retrospective studies help to demonstrate important connections between chemical structures and their biological function that are not obvious using 2-D similarity but are readily apparent by 3-D similarity.

### 31.1.3 Conclusions

The addition of PubChem3D to the existing contents of PubChem is a considerable achievement, given the scope, scale, and the fact that the resource is publicly accessible and free. With the ability to uncover latent structure-activity relationships of chemical structures, while complementing 2-D similarity analysis approaches, PubChem3D represents a new resource for scientists to exploit when exploring the biological annotations in PubChem.

## 31.2 Background

PubChem<sup>1234</sup> (<http://pubchem.ncbi.nlm.nih.gov>) is an open repository for small molecules and their experimental biological activities. The primary goal of PubChem is to be a public resource containing comprehensive information on the biological activities of small molecules. PubChem provides search, retrieval, visualization, analysis, and programmatic access tools in an effort to maximize the utility of contributed information. The PubChem3D project adds a new layer to this infrastructure. In the most basic sense, PubChem3D<sup>5678910</sup> generates a 3-D conformer model description of the small molecules contained within the PubChem Compound database. This 3-D description can be employed to enhance existing PubChem search and analysis methodologies by means of 3-D similarity. Prior to PubChem3D, this similarity approach was limited to a 2-D dictionary-based fingerprint ([ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem\\_fingerprints.txt](ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt)) to help relate chemical structures. With the advent of PubChem3D, this is now expanded to use a Gaussian-based similarity description of molecular shape<sup>111213</sup> used in software packages such as ROCS<sup>14</sup> and OEShape<sup>15</sup> from OpenEye Scientific Software, Inc.

It is reasonable to ask, why do we consider 3-D similarity methodologies at all? To put it simply, 2-D methods, while very useful and far cheaper computationally, may not be enough. A pitfall of most 2-D similarity methods is a general lack of ability to relate chemically diverse molecules with similar biological efficacy and function. For example, if a small molecule adopts an appropriate 3-D shape and possesses compatible functional groups properly oriented in 3-D space, it will likely bind to the biological moiety of interest. This “lock and key” binding motif is a major premise of structure-based drug design, docking, and molecular modelling applied with varying degrees of success over the past twenty years or more<sup>1617181920212223</sup>. These “compatible functional groups” involved in binding small molecules to proteins, which are typically used to define pharmacophores, are referred to here simply as “features”. Therefore, in this context, 3-D similarity considering both shape and feature complementariness may be useful to find or relate chemical structures that may bind similarly to a protein target.

In its essence, 3-D similarity adds another dimension to data mining and it can provide some degree of orthogonality from 2-D similarity results. With 2-D similarity, one can typically see by eye increased changes in the chemical structure molecular graph with increasing dissimilarity<sup>810</sup>. With 3-D similarity, it is not always obvious by looking only at the molecular graph, often requiring one to visualize 3-D conformer alignments to relate diverse chemistries. In all, 3-D similarity is complementary to 2-D similarity and provides an easy-to-grasp understanding (*i.e.*, one can readily see by examining a conformer pair superposition that both shape and features are similar) that may help to provide a contrast or new insight to the same (biological) data.

This work gives an overview of the PubChem3D project and its current capabilities. The technology and background that allowed 3-D methodologies to be economically applied to the tens of millions of chemical structures in the

<sup>1</sup> PubChem: integrated platform of small molecules and biological activities

<sup>2</sup> PubChem: a public information system for analyzing bioactivities of small molecules

<sup>3</sup> An overview of the PubChem BioAssay resource

<sup>4</sup> Database resources of the National Center for Biotechnology Information

<sup>5</sup> PubChem3D Thematic Series

<sup>6</sup> PubChem3D: conformer generation

<sup>7</sup> PubChem3D: diversity of shape

<sup>8</sup> PubChem3D: similar conformers

<sup>9</sup> PubChem3D: shape compatibility filtering using molecular shape quadrupoles

<sup>10</sup> PubChem3D: biologically relevant 3-D similarity

<sup>11</sup> A Gaussian description of molecular shape

<sup>12</sup> A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape

<sup>13</sup> Gaussian shape methods

<sup>14</sup> ROCS - Rapid Overlay of Chemical Structures, Version 2.2

<sup>15</sup> ShapeTK - C++, Version 1.8.0

<sup>16</sup> Synopsis of some recent tactical application of bioisosteres in drug design

<sup>17</sup> Molecular shape and medicinal chemistry: a perspective

<sup>18</sup> Docking: successes and challenges

<sup>19</sup> Structure-based discovery of antibacterial drugs

<sup>20</sup> Structure-based drug design strategies in medicinal chemistry

<sup>21</sup> Structure-based design of molecular cancer therapeutics

<sup>22</sup> Structure-based drug metabolism predictions for drug design

<sup>23</sup> Structure-based strategies for drug design and discovery

PubChem Compound database are described elsewhere<sup>5678910</sup> covering various aspects of the project, including conformer model generation validation<sup>6</sup>, the relative uniqueness of molecular shape<sup>7</sup>, and 3-D neighboring methodology<sup>8</sup>.

## 31.3 Construction and Content

### 31.3.1 1. PubChem3D Coverage

As one can imagine, it does not make sense nor is it possible to compute a 3-D description for all chemical structures in PubChem (*e.g.*, complexes and mixtures). PubChem provides a 3-D conformer model description for each record in the PubChem Compound database that satisfies the following conditions:

1. Not too large (with 50 non-hydrogen atoms).
2. Not too flexible (with 15 rotatable bonds).
3. Consists of only supported elements (H, C, N, O, F, Si, P, S, Cl, Br, and I).
4. Has only a single covalent unit (*i.e.*, not a salt or a mixture).
5. Contains only atom types recognized by the MMFF94s force field<sup>242526</sup>.
6. Has fewer than six undefined atom or bond stereo centers.

Figure 1 shows the PubChem3D coverage as of June 2011. Out of more than 30.3 million chemical structure records in the PubChem Compound database, there are nearly 27.2 million records with a 3-D description. This represents 89.6% of the PubChem Compound contents (92.3% when considering that 2.7% are salts whose parent structure has a 3-D description). Of the remaining 7.7% of chemical structures in PubChem devoid of a 3-D description, the largest category (representing 1.48 million or 4.9% of the total archive) consists of structures with more than 15 rotatable bonds. The next largest unique count (*i.e.*, those not already represented by structures with more than 15 rotatable bonds) is the cases of MMFF94s non-supported elements and non-supported atom environments (representing 280 thousand or 0.9% of the total archive, with an overlapping absolute count of 389 thousand). The remaining unique counts are the cases of large structures with +50 non-hydrogen atoms (representing 253 thousand or 0.8% of the total archive, with an overlapping absolute count of 882 thousand), excessive undefined stereo (representing 129 thousand or 0.4% of the total archive, with an overlapping absolute count of 234 thousand), chemical structures involving complexes or mixtures (representing 105 thousand or 0.3% of the total archive, with an overlapping absolute count of 324 thousand), and conformer generation failure (representing 79 thousand or 0.3% of the total archive). While the reasons for missing a 3-D description categories sometimes overlap, the ordering above is such that the one with the largest overall population is chosen first, with each subsequent category picking the largest remaining unique subpopulation not already covered, until all categories were exhausted.

### 31.3.2 2. Conformer Models

The computed coordinates for the 3-D representations are the essence of the PubChem3D project. Creation of the stored conformational models consists of multistep processes involving separate conformer generation, sampling, and post processing steps.

All conformers were generated by the OpenEye Scientific Software, Inc., OMEGA software<sup>2728293031</sup> using the C++

<sup>24</sup> Merck molecular force field. 1. Basis, form, scope, parameterization, and performance of MMFF94

<sup>25</sup> Merck molecular force field. 2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions

<sup>26</sup> MMFF VI. MMFF94s option for energy minimization studies

<sup>27</sup> OMEGA, Version 2.0

<sup>28</sup> OMEGA, Version 2.1

<sup>29</sup> OMEGA, Version 2.2

<sup>30</sup> OMEGA, Version 2.3

<sup>31</sup> OMEGA, Version 2.4

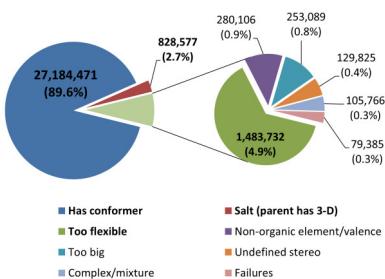


Figure 31.1: Figure 1. PubChem Compound database 3-D coverage

**PubChem Compound database 3-D coverage.** As one can see, 89.6% of all records have a 3-D conformer model. If one includes the parent compound of salts, this coverage can be considered to be 92.3%. Of the cases not having a 3-D conformer model, the majority are due to the flexibility of the chemical structure being too great to be suitable for conformer generation.

interface, the MMFF94s force field<sup>242526</sup> minus coulombic terms, and an energy filter of 25 kcal/mol. (Removal of coulombic terms<sup>632333435</sup> eliminated a bias towards conformations with energy-lowering intra-molecular interactions that tend not to be important for inter-molecular interactions, an important consideration given that the 3-D coordinates are generated in vacuo. Removal of attractive van der Waals terms did not have any noticeable effect<sup>6</sup>.) A maximum of 100,000 conformers per chemical structure stereo isomer were allowed. When undefined stereo centers were present, each stereo isomer was enumerated and conformers independently generated. These stereo isomer conformers were then combined ( $2^{**}5 = 32$  maximum stereo permutations,  $32 * 100,000 =$  maximum 3.2 million conformers).

Limiting to 100,000 conformations per stereo isomer can be a significant factor in limiting exploration of the conformational space. Ideally, one would want to explore the conformational space of a molecule exhaustively. In reality, it is not tractable to do so. For example, if one considers only three angles per rotatable bond and there are eleven rotatable bonds, this would yield  $3^{**}11 (= 177,147)$  possible conformers. If one considers four torsion angles per rotatable bond, and there are nine rotatable bonds, this would yield  $4^{**}9 (= 262,144)$  possible conformers. One can see how quickly systematic approaches can run into trouble with such exponential growth in the count of conformations and why there is a limit on how flexible a molecule is allowed to be.

With conformers generated, another important consideration is immediately obvious. It is not practical to store many thousands of conformers per compound. Therefore, after conformer generation is complete, the conformation count is reduced by sampling using root-mean-square-distance (RMSD) of pair-wise comparison of non-hydrogen atomic coordinates using the OEChem<sup>36</sup> OERMSD function with the automorph detection (which considers local symmetry equivalence of atoms such that, for example, rotation of a phenyl ring does not yield an artificially high RMSD) and overlay (which minimizes RMSD between conformers by rotation and translation of one conformer to the other) options selected. In some rare cases, the automorph detection was prohibitively expensive computationally and not used.

The sampling procedure employed is described elsewhere<sup>7</sup> but involves a two-stage clustering approach with an initial pass to partition-cluster conformers using an exclusion region hierarchy of decreasing dissimilarity (NlogN computational complexity, each cluster representative forms an exclusion region at a particular RMSD), followed by a step to remove edge-effects from the partition clustering ( $N^2$  computational complexity using only the cluster representatives at the desired RMSD). The RMSD value used when sampling was dependent on the size and flexibility of the chemical structure.

**Equations 1** and **2** were developed<sup>6</sup> to help prevent using a conformer sampling RMSD that was less than the capability of the OMEGA software to reproduce bioactive ligand conformations. The equations were intended to ensure that 90% of the sampled conformer models of 25,972 small-molecule ligands, whose 3-D structures were

<sup>32</sup> Biasing conformational ensembles towards bioactive-like conformers for ligand-based drug design

<sup>33</sup> Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database

<sup>34</sup> MIMUMBA revisited: Torsion angle rules for conformer generation derived from X-ray structures

<sup>35</sup> Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools

<sup>36</sup> OEChem, Version 1.7.4

experimentally determined, should contain at least one conformer within the RMSD sampling value to a bioactive conformation. The resulting *RMSD\_pred* value was rounded to the nearest 0.2 increment. The smallest RMSD value used was 0.4. If more than 500 conformers resulted after sampling, the RMSD was incremented by a further 0.2 and the conformer model was re-clustered. This process was repeated as many times as necessary to restrict the overall count of conformers to be 500 or less.

where “*nha*” is the count of non-hydrogen atoms in the molecule, “*er*” is the effective rotor count, and “*RMSD\_pred*” is the predicted average accuracy for a given “*nha*” and “*er*” value.

where “*er*” is the effective rotor count, “*rb*” is the rotatable bond count (computed using the OEChem “IsRotor” function) and “*nara*” is the count of non-aromatic ring atom count (OEChem OpenEye aromaticity model) excluding bridgehead atoms and SP2 hybridized atoms.

A post processing step was performed, after conformer model RMSD sampling, to completely relax the hydrogen atom locations by performing a full energy minimization where all non-hydrogen atoms were kept frozen. A subsequent “bump” check removed any conformers that had MMFF94 atom-atom interactions greater than 25 kcal/mol. Finally, each conformer was rotated and translated to their principal steric axes (*i.e.*, non-mass weighted principal moments of inertia axes) considering only non-hydrogen atoms.

It is important to note that the conformers produced are not stationary points on a potential energy hypersurface. In fact, one can readily achieve lower-energy conformations of a given chemical structure by performing an all-atom energy minimization to remove any bond, angle, or torsion strain present in vacuo. The PubChem3D conformer model for a chemical structure is meant to represent all possible biologically-relevant conformations that the molecule may have. In theory, one should have a reasonable chance to find any biologically accessible conformation within the RMSD sampling distance of the conformer model.

### 31.3.3 3. Conformer Model Properties

After a conformer model is produced, a series of properties are computed for each compound and each associated conformer. Table 1 lists the compound- and conformer-level properties provided by PubChem3D. The compound properties include: the sampling RMSD used to construct the conformer model; the MMFF94 partial charges per atom<sup>36</sup>; the functional group atoms that define each pharmacophore feature<sup>15</sup>; and the diverse conformer ordering, always starting with the default conformer per compound.

The feature definition lists the set of non-hydrogen atoms that comprise a given fictitious feature atom. The feature definitions are computed using the OEShape “ImplicitMillsDeans” forcefield<sup>15,37</sup>. Care is taken to (iteratively) merge feature definitions of common type that are within 1.0 Å distance of each other. Each feature definition is used to generate a fictitious “color” atom, whose 3-D coordinates are at the steric center of the atoms that comprise it (*i.e.*, at the average {X, Y, Z} value). There are six feature types used: anion, cation, (hydrogen-bond) acceptor, (hydrogen-bond) donor, hydrophobe, and ring.

The conformer properties include: the global conformer identifier (GID); conformer volume<sup>15</sup>; steric shape moments (monopole, quadrupole {Q<sub>x</sub>, Q<sub>y</sub>, Q<sub>z</sub>}, and octopole {O<sub>xxx</sub>, O<sub>yyy</sub>, O<sub>zzz</sub>, O<sub>xxz</sub>, O<sub>yyx</sub>, O<sub>yyz</sub>, O<sub>zzx</sub>, O<sub>zzy</sub>, and O<sub>xyz</sub>})<sup>15</sup>; shape self-overlap volume used in shape similarity computations<sup>11</sup>; feature self-overlap volume used in feature similarity computations<sup>11</sup>; MMFF94s energy with coulombic terms removed<sup>38</sup>; and the PubChem shape fingerprint<sup>8</sup>.

where *ST* is the measure of shape similarity (shape Tanimoto), <sub>AA</sub>V and <sub>BB</sub>V are the respective self-overlap volume of conformers A and B, and <sub>AB</sub>V is the common overlap volume between them.

where *CT* is the measure of feature similarity (color Tanimoto), the index “*f*” indicates any of the six independent fictitious feature atom types, *f*, and *f*.

where *ComboT* is the combo Tanimoto, *ST* is the shape Tanimoto, and *CT* is the color Tanimoto.

<sup>37</sup> Three-dimensional hydrogen-bond geometry and probability information from a crystal survey

<sup>38</sup> Szybki TK, Version 1.5.0

A diverse ordering of conformers is provided for each compound conformer ensemble<sup>83940</sup>. Using the lowest energy conformer in the ensemble as the initial default conformer, the conformer most dissimilar to the first is selected as the second diverse conformer. The conformer most dissimilar to the first two dissimilar conformers is chosen as the third diverse conformer. This process is repeated until there are no more conformers to be assigned a dissimilarity ordering. Similarity is measured by ST (**Equation 3**) and CT (**Equation 4**), involving a conformer superposition optimization<sup>1136</sup> to maximize the shape volume overlap between two conformers by means of rotation and translation of one conformer to the other. This is followed by a single point CT computation at the ST-optimized conformer pair overlay. The ST and CT are then added to yield a combo Tanimoto (**Equation 5**). The conformer with the smallest sum of combo Tanimoto to all *assigned* dissimilar conformers is selected as the next most dissimilar. In the case of a tie, the one with the largest sum of combo Tanimoto to *unassigned* conformers is used.

Note that PubChem has another source of 3-D information of small molecules, besides PubChem3D. The PubChem Substance database (unique identifier: SID) contains 3-D structures of small molecules deposited from individual depositors, which can be either experimentally determined or computationally predicted. For clarification, these depositor-provided structures are called “*substance* conformers”, and the theoretical conformers generated by PubChem3D for each PubChem Compound record (unique identifier: CID) are called “*compound* conformers”. For an efficient use of the PubChem3D resources, it is necessary to assign a unique identifier to each of compound conformers in the PubChem Compound database and substance conformers in the PubChem Substance database. The global conformer identifier (GID) uniquely identifies each conformer and is stored as a hex-encoded 64-bit unsigned integer, where the first 16-bits (0x000000000000FFFF) correspond to the local conformer identifier (LID), which is specific to a given conformer ensemble, the next 16-bits (0x00000000FFFF0000) are the version identifier (always zero for PubChem3D compound conformers, but nonzero for deposited substance conformers) and the last 32 bits (0xFFFFFFFF00000000) correspond to the structure identifier. This identifier is a compound identifier (CID), if the version identifier is zero, and a substance identifier (SID), when the version identifier is non-zero (the version identifier indicates the substance version to which the conformer corresponds). Substance conformer identifiers allow deposited 3-D coordinates to be utilized effectively by the PubChem3D system. As one can see, the GID provides global conformer identification system across all PubChem conformers.

A shape fingerprint is computed for the first ten diverse conformers. To generate this property, each conformer is ST-optimized to a set of reference conformers that describe the entire shape space diversity of the contents of PubChem3D. If the conformer is shape similar beyond a particular threshold to a reference conformer, the reference conformer identifier (CID and LID) and a packed rotational/translational matrix (64-bit integer) are retained. This makes each set reference conformer like a bit in a binary fingerprint, however; in this case, additional information (the superposition) is also retained. One can imagine that these shape fingerprints are a little like coordinates in shape space, mapping where a given conformer is located.

This shape fingerprint can be used in several ways during 3-D similarity computation and was born out of our earlier research<sup>841</sup> on “alignment recycling.” This work demonstrated that similar conformers align to a reference shape in a similar way. This means that, if one is interested only in finding similar shapes, conformer pairs that do not have common shape fingerprint “bits” can be ignored (*i.e.*, there is no need to perform a computationally intensive conformer alignment overlap optimization between two conformers when no common shape fingerprint reference exists, because the two conformer shapes are dissimilar to the extent that they may not need to be considered further). Additionally, when a common shape fingerprint reference exists between two conformers, one can “replay” the alignments of the two conformers to the common reference shape to yield a conformer alignment overlap between conformers that is (typically) very close to the optimal overlay; thus speeding up any conformer alignment overlap optimization but also providing an opportunity to further skip overlap optimization, when the best pre-optimized alignment overlap is not sufficient.

### 31.3.4 4. Similar Conformer Neighboring Relationship

Analogous to the precomputed “Similar Compounds” relationship for 2-D similarity, PubChem3D now provides a “Similar Conformers” neighboring relationship<sup>8</sup> using 3-D similarity. This neighboring takes into account both con-

<sup>39</sup> Clustering of chemical structures on the basis of two-dimensional similarity measures

<sup>40</sup> Implementing drug screening programs using molecular similarity methods

<sup>41</sup> Fast 3D shape screening of large chemical databases through alignment-recycling

former shape similarity and conformer pharmacophore feature similarity. Essentially, this is equivalent to performing a shape-optimized similarity search using ROCS<sup>1415</sup> at a threshold of ST > 0.795 and CT > 0.495, when both conformers have defined pharmacophore features. To allow for compounds devoid of features to be neighbored, a threshold of ST > 0.925 is used, but with the caveat that both conformers must not have any defined pharmacophore features. Currently, three diverse conformers per compound are neighbored; however, this may change, with up to ten conformers per compound used as computational resources allow. The conformers used for neighboring correspond to the first “N” conformers in the diverse conformer list property. (See the **Conformer Model Properties** section.) This ensures maximal coverage of the unique shape/feature space of a chemical structure as additional conformers are considered in neighboring.

### 31.3.5 5. FTP Site

PubChem3D data is available on the PubChem FTP site ([ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound\\_3D](ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound_3D)). One may download in bulk 3-D descriptions of PubChem Compound records. On average there are approximately 110 conformers per compound in the PubChem3D system; however, not all data is provided for public download, in part due to the overall size being many terabytes, more data than one can readily share publicly. Therefore, two different subsets are provided in various file formats (SDF, XML, and ASN.1) that correspond to either the default conformer or the first ten conformers in the diverse conformer list property. (See the **Conformer Model Properties** section.) Beyond these two conformer subsets of PubChem3D, one may also find a description of the conformers that comprise the PubChem3D shape fingerprint. These conformers represent all shape diversity present in the PubChem3D system for a given analytic volume range and a given level of shape similarity ST threshold.

The “Similar Conformers” neighboring relationship is also provided for download. This conformer pair relationship (one per line) includes the respective conformer identifiers, ST, CT, and the 3 × 3 rotation matrix and translation vector (applied in that order) to superimpose the second conformer to the first. The rotation/translation refers to the coordinates provided in the download set of ten diverse conformers or otherwise available for download from our PubChem download facility. (See the **Utility: Download** section.)

## 31.4 Utility

### 31.4.1 1. NCBI Entrez Interface

The primary search interface for PubChem is Entrez<sup>4</sup>, *e.g.*, for the PubChem Compound database, accessible by means of the PubChem homepage (<http://pubchem.ncbi.nlm.nih.gov>) or the URL: <http://www.ncbi.nlm.nih.gov/pccompound?Db=pccompound>. There are fourteen Entrez indexes available to query PubChem Compound records based on 3-D information detailed in Table 2. For example, to find which compound conformer models were sampled in the RMSD range between 0.4 and 0.6, one would perform the query

The indexes for “Volume3D”, “XStericQuadrupole3D”, “YStericQuadrupole3D”, and “ZStericQuadrupole3D” correspond, respectively, to the analytic volume and the three steric quadrupole moments<sup>91242</sup> for only the first conformer in the diverse conformer list (*i.e.*, the default conformer). The steric quadrupoles essentially correspond to the extents of the compound, where X, Y, and Z correspond to the length, width, and height. For example, to find very long, near-linear compounds, one may give the PubChem Compound Entrez query [http://pubchem.ncbi.nlm.nih.gov/help.html#PubChem\\_index](http://pubchem.ncbi.nlm.nih.gov/help.html#PubChem_index).

PubChem also provides filtering capabilities. Unlike indexes, which hold discrete values, filters are Boolean-based (*i.e.*, either a record is in the list or it is not). PubChem3D provides some additional filtering capabilities. In the case of the PubChem Compound database, there is a filter “has 3d conformer” that will indicate whether a given compound record has a 3-D conformer model by means of the PubChem Compound query: “

Filtering capabilities were also expanded in the PubChem Substance database. Two filters were added: “has deposited 3d” and “has deposited 3d experimental” to indicate when a substance record has 3-D coordinates and when the

<sup>42</sup> A new class of molecular shape descriptors. 1. Theory and properties

contributed 3-D coordinates were determined experimentally, respectively. For example, to find all experimentally determined 3-D structures for substance records, one would use the PubChem Substance databases query: “

### 31.4.2 2. Visualization

Each PubChem Compound (and Substance) record has a summary page as depicted in Figure 2 (<http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=681> for dopamine). When a 3-D conformer model can be produced for a compound record (or a depositor-provided 3-D coordinates for the substance record), a 3-D image of the structure will be available by clicking the “3D” tab. In the case of a PubChem Compound record, this corresponds to the first diverse conformer, which is the default conformer. As shown in Figure 3, if one clicks the image, a popup menu appears allowing one to invoke the “Web-based 3D Viewer” or to send the 3-D information to the “Pc3D Viewer Application”.

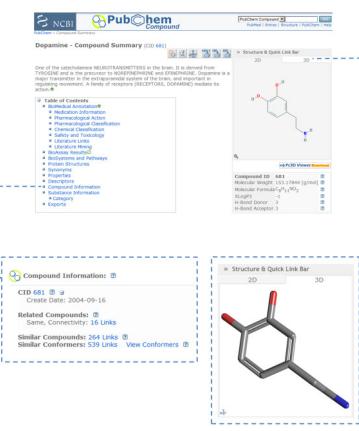


Figure 31.2: Figure 2. Summary page enhancements

**Summary page enhancements.** A snapshot of the PubChem Compound summary page of dopamine (CID 681). Clicking on the “3D” tab on the right side of the page shows the 3-D structure of the molecule. Clicking the “Compound information” in the “Table of Contents” box directs users to 2-D neighbors (“Similar Compounds”) and 3-D neighbors (“Similar Conformers”).

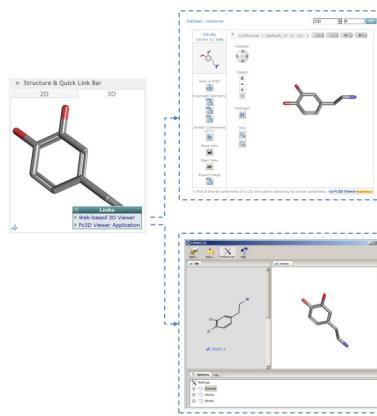


Figure 31.3: Figure 3. Visualization of a 3-D structure conformer

**Visualization of a 3-D structure conformer.** Clicking on the 3-D image on the PubChem Compound summary page (left) shows links to the web-based 3-D viewer (top right) and the Pc3D desktop helper application (bottom right).

The Pc3D viewer application can be downloaded and installed on PC, Mac, or Linux computers. A link to download this application can be found below the image on a given summary page or other PubChem3D aware pages (e.g., see

the “Pc3D Viewer Download” icon in Figure 2). The viewer provides an interface for rendering 3-D structures of PubChem Compound records and visualizing their superpositions. With a customizable 3-D rendering engine that provides dynamic molecular visualization experience, it has the ability to create high-resolution, publication-quality images. It allows use of XYZ model files and SDF files and supports PubChem native formatted files (with the .pc3d or .asn extension).

The web-based 3-D viewer, like the Pc3D viewer application, allows one to browse 3-D conformers available for substances or compounds and their superpositions. This interactive tool (accessible via <http://pubchem.ncbi.nlm.nih.gov/vw3d/>) operates without the need for a web browser plug-in (and does not use Java, for support related reasons) by means of displaying a series of images to simulate molecule rotation. As shown in Figure 4, besides providing immediate access to the “Similar Conformer” neighboring relationship per compound (and per compound conformer), users can access various controls to perform such tasks as: superposition or conformer navigation, data export, conformer rotation type, conformer rotation speed, conformer image resize, conformer filtering, and sorting. The viewer allows any arbitrary set of 3-D compound conformers or conformer pairs (substance and compound) that exists within PubChem to be viewed or superimposed. This tool is also the primary resource to visualize and manage 3-D information from various PubChem3D-aware tools, including 3-D conformer search and 3-D structure clustering.

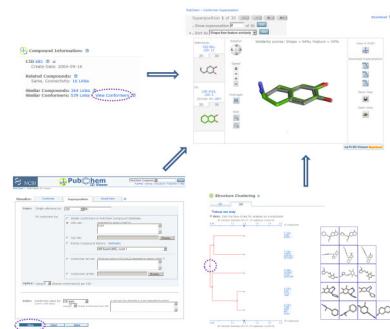


Figure 31.4: Figure 4. Visualization of 3-D structure conformer superpositions

**Visualization of 3-D structure conformer superpositions.** Superpositions between compound conformers are accessible from various PubChem3D-aware applications. The PubChem Compound summary page (top left) allows the “Similar Conformers” neighboring relationship to be visualized. The PubChem3D web-based viewer (bottom left) allows arbitrary superpositions to be generated. The PubChem Structure Clustering tool (bottom right) enables all pair-wise superpositions to be examined.

### 31.4.3 3. Search

The PubChem Structure Search system<sup>1</sup> (accessible via <http://pubchem.ncbi.nlm.nih.gov/search/>) allows one to search the PubChem Compound database using a chemical structure in various formats. PubChem3D adds a new capability to this system by allowing one to perform a 3-D similarity search and to visualize the results. At the time of writing, this similarity search is essentially equivalent to that described in the **Similar Conformer Neighboring Relationship** section. If 3-D coordinates are not provided for a chemical structure query, they are generated automatically, as-is possible, while keeping in mind that not all chemical structures can be covered by the PubChem3D system. (See the **PubChem3D Coverage** section for more details.) To aid in performing automated queries, a programmatic interface is available. (See the **Programmatic Interface** section for more details.)

A 3-D conformer search currently considers the first three diverse conformers per compound as candidates for “Similar Conformers”. (See diverse conformer ordering in the **Conformer Model Properties** section.) Given that there are more than 27 million CIDs and three conformers per compound are being considered, this means that there are around 81 million conformers considered by each 3-D query. This count will change as a function of time as data is added to PubChem and as the count of conformers per compound are increased. To achieve adequate query throughput, an “embarrassingly parallel divide-and-conquer” strategy is employed. The PubChem Compound conformer data set is subdivided into multiple evenly-sized subsets. Each subset is then searched in parallel. If more query throughput

is desired, and the computational capacity exists, the solution is simple; one simply needs to increase the count of evenly-sized subsets to simultaneously process.

#### 31.4.4 4. Download

The PubChem Download facility<sup>1</sup> ([http://pubchem.ncbi.nlm.nih.gov/pc\\_fetch](http://pubchem.ncbi.nlm.nih.gov/pc_fetch)) allows one to download PubChem records resulting from a search or a user-provided identifier list. With the advent of the PubChem3D layer, there is now the ability to download up to ten diverse conformers per compound. Alternatively, 3-D images may be downloaded (for the default conformer, only). A programmatic interface is available. (See the **Programmatic Interface** section for more details.)

#### 31.4.5 5. Similarity Computation

The PubChem Score Matrix facility ([http://pubchem.ncbi.nlm.nih.gov/score\\_matrix](http://pubchem.ncbi.nlm.nih.gov/score_matrix)) allows one to compute pairwise similarities of a set of PubChem compound records (up to 1,000,000 similarity pairs per request). The PubChem3D layer adds the ability to compute 3-D similarities using up to ten conformers (either the first  $N$ -diverse conformers or a user-provided conformer set) per compound per request. Additionally, this service allows one to select the type of superposition optimization (shape or feature) to perform. A programmatic interface is available. (See the **Programmatic Interface** section.)

#### 31.4.6 6. Clustering and Analysis

The PubChem Structure Clustering tool<sup>10</sup> (<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=clustering>) allows one to perform single-linkage clustering for up to 4,000 compounds at a time. This interactive tool provides visualization, subset, selection, and analysis capabilities. For example, the dendrogram allows compounds to be grouped into clusters by clicking the Tanimoto bar provided above and below the dendrogram (see the bottom right panel in Figure 4). One can then click on the cluster to view the individual compounds or perform other operations. The PubChem3D layer adds the ability to cluster compounds according to their 3-D similarities, with up to ten diverse conformers per compound. This service allows one to select: the superposition optimization type (shape or feature); whether to cluster all conformers or just the most similar conformer pair; and the conformer similarity metric.

#### 31.4.7 7. Programmatic Interface

PubChem provides a programmatic interface called the Power User Gateway (PUG)<sup>1</sup>. This extends the capabilities provided by the NCBI eUtils programmatic interface<sup>43</sup>, which interfaces the NCBI Entrez search engine contents. PUG can be used to send programmatic requests (*e.g.*, to perform queries or other tasks). If a request does not complete, a request ID is returned. One uses this to “poll” whether the request is completed, at which point an URL is provided to obtain the results. This is necessary, considering that most user requests are queued and may not be executed or completed immediately. A PUG/SOAP interface exists to allow the SOAP-based protocol to be used to route requests. SOAP-interfaces are readily available for most programming (*e.g.*, Java, C#, VisualBasic) and scripting languages (*e.g.*, Perl, Python), as well as workflow applications (*e.g.*, Taverna<sup>44</sup>, Pipeline Pilot<sup>45</sup>). The PubChem3D layer extensions are now available in individual PUG-aware interfaces and by means of the PUG/SOAP interface.

---

<sup>43</sup> Entrez Programming Utilities Help

<sup>44</sup> Taverna: a tool for building and running workflows of services

<sup>45</sup> Pipeline Pilot, Version 8.5

## 31.5 Examples of use

To assist in understanding how PubChem3D can be useful to locate additional biological annotation and enhance one's ability to identify potential structure-activity relationships, a series of illustrative examples were prepared. These examples benefit from a recent study<sup>10</sup> of the statistical distribution of random 3-D similarities of more than 740,000 biologically tested small molecules in PubChem using a single conformer per compound, where the average ( $\mu$ ) and standard deviation ( $\sigma$ ) of the shape-optimized ST, CT, and ComboT scores between two randomly selected conformers were found to be  $0.54 \pm 0.10$ ,  $0.07 \pm 0.05$ , and  $0.62 \pm 0.13$ , respectively. The probability of two random conformers having a ST-optimized similarity score greater than or equal to the  $\mu + 2\sigma$  threshold (*i.e.*, 0.74, 0.17, and 0.88 for ST, CT, and ComboT, respectively) was 2%, 4%, and 3% for ST, CT, and ComboT, respectively. This statistical information is meaningful to provide reasonable 3-D similarity thresholds, whereby one can be confident that most of the 3-D similarities between chemical structures is not simply by chance. When a group of chemical structures with similar biological activity and function are shown to have 3-D similarity to each other above these thresholds, it suggests that a common macromolecule binding interaction orientation exists and, furthermore, that the features required for such binding are present.

### 31.5.1 1. Finding additional biological annotation

In a data system such as PubChem, with a very uneven amount of biological annotation, it is helpful to find related chemical structures where more information is known. PubChem provides two precomputed neighboring relationships to locate similar chemical structures. The “Similar Conformers” neighboring relationship precomputes the 3-D similarity between all chemical structures in PubChem, while the “Similar Compounds” neighboring relationship precomputes the 2-D similarity. Using dopamine (CID 681) as an example, Figure 5 shows there can be relatively little commonality between 2-D and 3-D similarities; however, both relationships find chemicals that are related, with the 2-D similarity being good at finding chemical analogs of a given chemical while the 3-D similarity is skilled at locating molecules with similar shape and similar 3-D orientation of binding features. Therefore, use of both neighboring relationships allows a larger number of related chemicals to be found with associated biomedical literature (MeSH Links), biologically tested (BioAssay Tested), or bound to a protein 3-D structure (Protein3D Links).

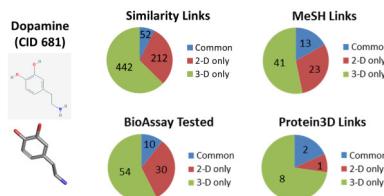


Figure 31.5: Figure 5. 3-D similarity relationship finds additional biological annotation

**3-D similarity relationship finds additional biological annotation.** Comparison of the 2-D “Similar Compound” and 3-D “Similar Conformer” neighboring relationships using dopamine to demonstrate how both neighboring relationship complement each other when locating related chemical structures with unique biological annotation.

### 31.5.2 2. Relating chemical probes for same biological target

ML088 (CID 704205) and ML087 (CID 25199559), shown in Figure 6, are chemical probes reported<sup>46</sup> in a PubChem BioAssay (AID 1548) with EC50s of  $6.19 \mu\text{M}$  and  $0.20 \mu\text{M}$ , respectively. Both probes target a common protein, the tissue non-specific alkaline phosphatase (TNAP, GI 116734717), the deficiency of which is associated with defective bone mineralization in the form of rickets and osteomalacia. At first glance, these two chemical structures are rather dissimilar, with a 2-D subgraph similarity of 0.43 using the PubChem fingerprint. This suggests the two chemical structures are unrelated to each other, giving no hint as to why they have similar biological function and efficacy.

<sup>46</sup> HTS identification of compounds activating TNAP at an intermediate concentration of phosphate acceptor detected in luminescent assay

Using 3-D similarity, by means of the PubChem3D web-based viewer as shown in Figure 6, the shape, feature, and combo similarities (0.80, 0.23, and 1.03 for ST, CT, and ComboT, respectively) tell a very different story. The two chemical structures are 3-D similar, suggesting that the two chemical structures can adopt a similar shape and have some binding features in a common 3-D orientation, thus helping to relate the observed biological activity by providing a hypothesis that the two inhibitors may bind in a similar manner. While this could be interpreted as simply highlighting a deficiency in the PubChem 2-D similarity metric, in this case, PubChem 3-D similarity complements the PubChem 2-D similarity by allowing such a similarity relationship to be found between these two chemical probes.

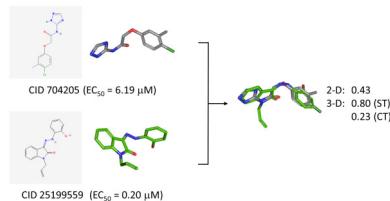


Figure 31.6: Figure 6. Relating biologically active compounds by means of PubChem3D

**Relating biologically active compounds by means of PubChem3D.** Chemical probes ML088 (CID 704205) and ML087 (CID 25199559) from PubChem BioAssay 1548 against tissue non-specific alkaline phosphatase (TNAP, GI:116734717) are not similar by 2-D similarity but are by 3-D similarity.

### 31.5.3 3. Relating chemically diverse structures with same pharmacological action

Figure 7 shows the 2-D and 3-D similarity score matrices for a carefully selected set of eight anti-inflammatory drug molecules having the same MeSH<sup>47</sup> pharmacological action annotation of “Histamine H1 Antagonists” (MeSH ID 68006634). Figure 8 depicts a subset of 3-D ST-optimized superpositions resulting from the 28 unique compound pairs. The 2-D Tanimoto similarity values between these compounds are quite low, with only three compound pairs above 0.75, indicating that the 2-D similarity method based on the PubChem fingerprint fails to interrelate their common biological activity as histamine H1 receptor antagonists. On the contrary, the 3-D similarity between these eight molecules is rather high, with a ST 0.74 and ComboT 1.0 for all but eight of the 28 compound pairs. As illustrated in Figure 8, even if the 2-D Tanimoto value between a pair of molecules is as low as 0.31, they can still have significant structural overlap in 3-D shape/feature space, resulting in relatively larger ST and CT similarity scores. The structure clustering tool is specifically geared towards helping to identify such structure-activity trends in 3-D similarity (as well as 2-D similarity) space and, in combination with the PubChem3D viewer, allow them to be visualized. If one thinks about this, it shows how easy it might be to “scaffold hop” or relate diverse chemical structures with similar biological function by examining 3-D similar chemicals in PubChem. It may also suggest that one may be able to better understand additional biological functions of known drugs (*i.e.*, so called “side effects”) by examining their PubChem 3-D similarity to other chemicals with known biological roles.

## 31.6 Conclusions

A new resource for scientists, PubChem3D, layered on top of PubChem, provides a new dimension to its ability to search, subset, export, visualize, and analyze chemical structures and their associated biological data. With a broad suite of tools and capabilities, 3-D similarity is given equal footing to assist in finding non-obvious trends in experimentally observed biological activity. As a complement to 2-D similarity, 3-D similarity demonstrates a capability to relate chemical series that are not sufficiently 2-D similar.

<sup>47</sup> Medical Subject Headings

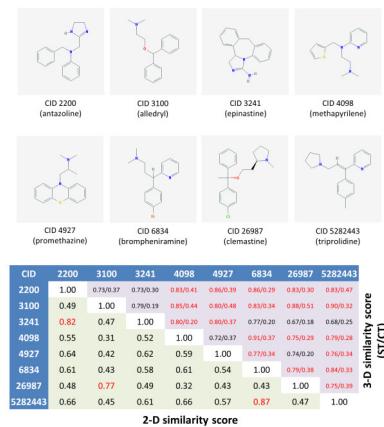


Figure 31.7: Figure 7. Similarity score matrix for selected histamine H1 receptor antagonist anti-inflammatory drugs

**Similarity score matrix for selected histamine H1 receptor antagonist anti-inflammatory drugs.** The lower triangle of the score matrix corresponds to the 2-D similarity computed using the PubChem fingerprint. The upper triangle corresponds to the 3-D similarity ST/CT scores. The matrix elements in red text indicate a 2-D similarity 0.75 or 3-D similarity with ST 0.74 and ComboT 1.0. The first ten diverse conformers per molecule were superimposed using shape-based optimization and the single conformer-pair per compound-pair with the largest ComboT retained.

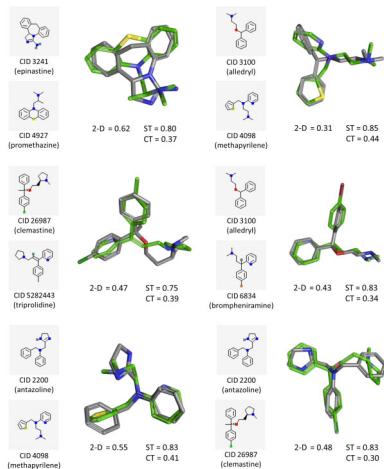


Figure 31.8: Figure 8. 3-D superposition of selected histamine H1 receptor antagonist anti-inflammatory drugs

**3-D superposition of selected histamine H1 receptor antagonist anti-inflammatory drugs.** Although there is little 2-D similarity, using the PubChem fingerprint, substantial 3-D similarity is found between various structurally diverse anti-inflammatory drugs.

## 31.7 Abbreviations

2-D: (2-dimensional); 3-D: (3-dimensional); MMFF: (Merck Molecular Force Field); RMSD: (root-mean-square distance).

## 31.8 Competing interests

The authors declare that they have no competing interests.

## 31.9 Authors' contributions

All authors contributed in a material way to the PubChem3D project. Specific attributable contributions are as follows: EEB drafted the manuscript and performed all major project aspects not directly attributed to others; JC implemented the web-based viewer and search interfaces; SK drafted the manuscript and helped develop neighboring accelerators; LH and YS developed analysis database components; WS and SH developed storage and neighboring database components; VS developed the image generation methodology, PC3D viewer application, search backends, and viewer backends; PAT developed the download and score matrix facilities; JW developed the structure clustering and heat-map facilities; BY and PAT developed the identifier exchange facilities; JZ developed the compound/substance summary facilities; and SHB heads the PubChem project. All authors read and approved the final manuscript.

## 31.10 Acknowledgements

This research was supported (in part) by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, U. S. Department of Health and Human Services. This effort utilized the high-performance computational capabilities of the Helix Systems at the National Institutes of Health, Bethesda, MD (<http://helix.nih.gov>).

EEB is very appreciative of the folks at OpenEye Scientific Software, Inc., for allowing their tools and methodologies to be utilized in the PubChem3D project and for the many fruitful suggestions/discussions/insights. EEB is also very thankful for the significant and continuing support of the NCBI systems staff, specifically Ron Patterson, Charlie Cook, and Don Preuss, without which the PubChem3D project would not exist.

We are very grateful to the reviewers for their careful consideration of this manuscript and useful suggestions.

# OPEN BABEL: AN OPEN CHEMICAL TOOLBOX

## 32.1 Abstract

### 32.1.1 Background

A frequent problem in computational modeling is the interconversion of chemical structures between different formats. While standard interchange formats exist (for example, Chemical Markup Language) and *de facto* standards have arisen (for example, SMILES format), the need to interconvert formats is a continuing problem due to the multitude of different application areas for chemistry data, differences in the data stored by different formats (0D versus 3D, for example), and competition between software along with a lack of vendor-neutral formats.

### 32.1.2 Results

We discuss, for the first time, Open Babel, an open-source chemical toolbox that speaks the many languages of chemical data. Open Babel version 2.3 interconverts over 110 formats. The need to represent such a wide variety of chemical and molecular data requires a library that implements a wide range of cheminformatics algorithms, from partial charge assignment and aromaticity detection, to bond order perception and canonicalization. We detail the implementation of Open Babel, describe key advances in the 2.3 release, and outline a variety of uses both in terms of software products and scientific research, including applications far beyond simple format interconversion.

### 32.1.3 Conclusions

Open Babel presents a solution to the proliferation of multiple chemical file formats. In addition, it provides a variety of useful utilities from conformer searching and 2D depiction, to filtering, batch conversion, and substructure and similarity searching. For developers, it can be used as a programming library to handle chemical data in areas such as organic chemistry, drug design, materials science, and computational chemistry. It is freely available under an open-source license from <http://openbabel.org>.

## 32.2 Introduction

The history of chemical informatics has included a huge variety of textual and computer representations of molecular data. Such representations focus on specific atomic or molecular information and may not attempt to store all possible chemical data. For example, line notations like Daylight SMILES<sup>1</sup> do not offer coordinate information, while

---

<sup>1</sup> SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules

crystallographic or quantum mechanical formats frequently do not store chemical bonding data. Hydrogen atoms are frequently omitted from x-ray crystallography due to the difficulty in establishing coordinates, and are often ignored by some file formats as the “implicit valence” of heavy atoms that indicates their presence. Other types of representations require specification of atom types on the basis of a specific valence bond model, inclusion of computed partial charges, indication of biomolecular residues, or multiple conformations.

While attempts have been made to provide a standard format for storing chemical data, including most notably the development of Chemical Markup Language (CML)<sup>23456</sup>, an XML dialect, such formats have not yet achieved widespread use. Consequently, a frequent problem in computational modeling is the interconversion of molecular structures between different formats, a process that involves extraction and interpretation of their chemical data and semantics.

We outline for the first time, the development and use of the Open Babel project, a full-featured open chemical toolbox, designed to “speak” the many different representations of chemical data. It allows anyone to search, convert, analyze, or store data from molecular modeling, chemistry, solid-state materials, biochemistry, or related areas. It provides both ready-to-use programs as well as a complete, extensible programmer’s toolkit for developing cheminformatics software. It can handle reading, writing, and interconverting over 110 chemical file formats, supports filtering and searching molecule files using Daylight SMARTS pattern matching<sup>7</sup> and other methods, and provides extensible fingerprinting and molecular mechanics frameworks. We will discuss the frameworks for file format interconversion, fingerprinting, fast molecular searching, bond perception and atom typing, canonical numbering of molecular structures and fragments, molecular mechanics force fields, and the extensible interfaces provided by the software library to enable further chemistry software development.

Open Babel has its origin in a version of OELib released as open-source software by OpenEye Scientific under the GPL (GNU Public License). In 2001, OpenEye decided to rewrite OELib in-house as the proprietary OEChem library, so the existing code from OELib was spun out into the new Open Babel project. Since 2001, Open Babel has been developed and substantially extended as an international collaborative project using an open-source development model<sup>8</sup>. It has over 160,000 downloads, over 400 citations<sup>9</sup>, is used by over 40 software projects<sup>10</sup>, and is freely available from the Open Babel website<sup>11</sup>.

## 32.3 Features

### 32.3.1 File Format Support

With the release of Open Babel 2.3, Open Babel supports 111 chemical file formats in total. It can read 82 formats and write 85 formats. These encompass common formats used in cheminformatics (SMILES, InChI, MOL, MOL2), input and output files from a variety of computational chemistry packages (GAMESS, Gaussian, MOPAC), crystallographic file formats (CIF, ShelX), reaction formats (MDL RXN), file formats used by molecular dynamics and docking packages (AutoDock, Amber), formats used by 2D drawing packages (ChemDraw), 3D viewers (Chem3D, Molden) and chemical kinetics and thermodynamics (ChemKin, Thermo). Formats are implemented as “plugins” in Open Babel, which makes it easy for users to contribute new file formats (see Extensible Interface below). Depending on the format, other data is extracted by Open Babel in addition to the molecular structure; for example, vibrational frequencies are extracted from computational chemistry log files, unit cell information is extracted from CIF files, and property fields are read from SDF files.

---

<sup>2</sup> Chemical markup, XML, and the Worldwide Web. 1. Basic principles

<sup>3</sup> Chemical Markup, XML and the World-Wide Web. 2. Information Objects and the CMLDOM

<sup>4</sup> Development of chemical markup language (CML) as a system for handling complex chemical content

<sup>5</sup> Chemical Markup, XML, and the World Wide Web. 4. CML Schema

<sup>6</sup> Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions

<sup>7</sup> NOTITLE!

<sup>8</sup> NOTITLE!

<sup>9</sup> NOTITLE!

<sup>10</sup> NOTITLE!

<sup>11</sup> NOTITLE!

A number of “utility” file formats are also defined; these are not strictly speaking a way of storing the molecular structure, but rather present certain functionality through the same interface as the regular file formats. For example, the *report format*<sup>12</sup> that presents a summary of the molecular structure of a molecule; the *fingerprint format*<sup>13</sup> and *fastsearch format*<sup>14</sup> are used for similarity and substructure searching (see below); the *MolPrint2D* and *Multilevel Neighborhoods of Atoms* formats calculate circular fingerprints defined by Bender *et al.*<sup>15</sup><sup>16</sup> and Filimonov *et al.*<sup>17</sup><sup>18</sup> respectively.

Each format can have multiple options to control either reading or writing a particular format. For example, the InChI format has 12 options including an option “K” to generate an InChIKey, “T <param>” to truncate the InChI depending on a supplied parameter and “w” to ignore certain InChI warnings. The available options are listed in the documentation, are shown in the Graphical User Interface (GUI) as checkboxes or textboxes, and can be listed at the command-line. In fact, all three are generated from the same source; a documentation string in the C++ code.

### 32.3.2 Fingerprints and Fast Searching

Databases are widely used to store chemical information especially in the pharmaceutical industry. A key requirement of such a database is the ability to index chemical structures so that they can be quickly retrieved given a query substructure. Open Babel provides this functionality using a path-based fingerprint. This fingerprint, referred to as *FP2* in Open Babel, identifies all linear and ring substructures in the molecule of lengths 1 to 7 (excluding the 1-atom substructures C and N) and maps them onto a bit-string of length 1024 using a hash function. If a query molecule is a substructure of a target molecule, then all of the bits set in the query molecule will also be set in the target molecule. The fingerprints for two molecules can also be used to calculate structural similarity using the Tanimoto coefficient, the number of bits in common divided by the union of the bits set.

Clearly, repeated searching of the same set of molecules will involve repeated use of the same set of fingerprints. To avoid the need to recalculate the fingerprints for a particular multi-molecule file (such as an SDF file), Open Babel provides a *fastindex* format that solely stores a fingerprint along with an index into the original file. This index leads to a rapid increase in the speed of searching for matches to a query - datasets with several million molecules are easily searched interactively. In this way, a multi-molecule file may be used as a lightweight alternative to a chemical database system.

### 32.3.3 Bond Perception and Atom Typing

As mentioned above, many chemical file formats offer representations of molecular data solely as lists of atoms. For example, most quantum chemical software packages and most crystallographic file formats do not offer definitions of bonding. A similar situation occurs in the case of the Protein Data Bank (PDB) format; while standardized<sup>19</sup> files contain connectivity information, non-standard files exist that often do not provide full connectivity information. Consequently, Open Babel features methods to determine bond connectivity, bond order perception, aromaticity determination, and atom typing.

Bond connectivity is determined by the frequently used algorithm of detecting atoms closer than the sum of their covalent radii, with a slight tolerance (0.45 Å) to allow for longer than typical bonds. To handle disorder in crystallographic data (e.g., PDB or CIF files), atoms closer than 0.63 Å are not bonded. A further filtering pass is made to ensure standard bond valency is maintained; each element has a maximum number of bonds, if this is exceeded then the longest bonds to an atom are successively removed until the valence rule is fulfilled.

<sup>12</sup> NOTITLE!

<sup>13</sup> NOTITLE!

<sup>14</sup> NOTITLE!

<sup>15</sup> NOTITLE!

<sup>16</sup> Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier

<sup>17</sup> NOTITLE!

<sup>18</sup> Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors

<sup>19</sup> NOTITLE!

After bond connectivity is determined, if needed or requested by the user, bond order perception is performed on the basis of bond angles and geometries. The method is similar to that proposed by Roger Sayle<sup>20</sup> and uses the average bond angle around an un-typed atom to determine sp and sp<sup>2</sup> hybridized centers. 5-membered and 6-membered rings are checked for planarity to estimate aromaticity. Finally, atoms marked as unsaturated are checked for an unsaturated neighbor to give a double or triple bond. After this initial atom typing, known functional groups are matched, followed by aromatic rings, followed by remaining unsatisfied bonds based on a set of heuristics for short bonds, atomic electronegativity, and ring membership.

Atom typing is performed by “lazy evaluation,” matching atoms against SMARTS patterns to determine hybridization, implicit valence, and external atom types. Atom type perception may be triggered by adding hydrogens (which requires determination of implicit and explicit valence), exporting to a file format that requires atom types, or as requested by the user. To minimize the amount of typing required, when importing from a format with atom types specified, a lookup table is used to translate between equivalent types.

An important part of atom typing is aromaticity detection and assignment of Kekulé bond orders (kekulization). In Open Babel, a central aromaticity model is used, largely matching the commonly used Daylight SMILES representation<sup>1</sup>, but with added support for aromatic phosphorous and selenium. Potential aromatic atoms and bonds are flagged on the basis of membership in a ring system possibly containing  $4n+2\pi$  electrons. Aromaticity is established only if a well-defined valence bond Kekulé pattern can be determined. To do this, atoms are added to a ring system and checked against the  $4n+2\pi$  electron configuration, gradually increasing the size to establish the largest possible connected aromatic ring system. Once this ring system is determined, an exhaustive search is performed to assign single and double bonds to satisfy all valences in a Kekulé form. Since this process is exponential in complexity, the algorithm will terminate if more than 30 levels of recursion or 15 seconds are exceeded (which may occur in the case of large fused ring systems such as carbon nanotubes).

### 32.3.4 Canonical Representation of Molecules

In general, for any particular molecular structure and file format, there are a large number of possible ways the structure could be stored; for example, there are  $N!$  ways of ordering the atoms in an MOL file. While each of the orderings encodes exactly the same information, it can be useful to define a canonical numbering of the atoms of a molecule and use this to derive a canonical representation of a molecule for a particular file format. For a zero-dimensional file format without coordinates, such as SMILES, the canonical representation could be used to index a database, remove duplicates or search for matches.

Open Babel implements a sophisticated canonicalization algorithm that can handle molecules or molecular fragments. The atom symmetry classes are the initial graph invariants and encode topological and chemical properties. A cooperative labeling procedure is used to investigate the automorphic permutations to find the canonical code. Although the algorithm is similar to the original Morgan canonical code<sup>21</sup>, various improvements are implemented to improve performance. Most notably, the algorithm implements heuristics from the popular nauty package<sup>2223</sup>. Another aspect handled by the canonical code is stereochemistry as different labelings can lead to different parities. This is further complicated by the possibility of symmetry-equivalent stereocenters and stereocenters whose configuration is interdependent. The full details will be the subject of a separate publication.

### 32.3.5 Coordinate Generation in 2D and 3D

Open Babel, version 2.3, has support for 2D coordinate generation (Figure 1) through the donation of code by Sergei Trepalin, based on the code used in the MCDL chemical structure editor<sup>242526</sup>. The MCDL algorithm aims to layout

---

<sup>20</sup> NOTITLE!

<sup>21</sup> The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service

<sup>22</sup> NOTITLE!

<sup>23</sup> Practical graph isomorphism

<sup>24</sup> Modular Chemical Descriptor Language (MCDL): Composition, connectivity, and supplementary modules

<sup>25</sup> A Java Chemical Structure Editor Supporting the Modular Chemical Descriptor Language (MCDL)

<sup>26</sup> Modular Chemical Descriptor Language (MCDL): Stereochemical modules

the molecular structure in 2D such that all bond lengths are equal and all bond angles are close to 120°. The layout algorithm includes a small database of around 150 templates to help layout cages and large fragment cycles. To deal with the problem of overlapping fragments, the algorithm includes an exhaustive search procedure that rotates around acyclic bonds by 180°.

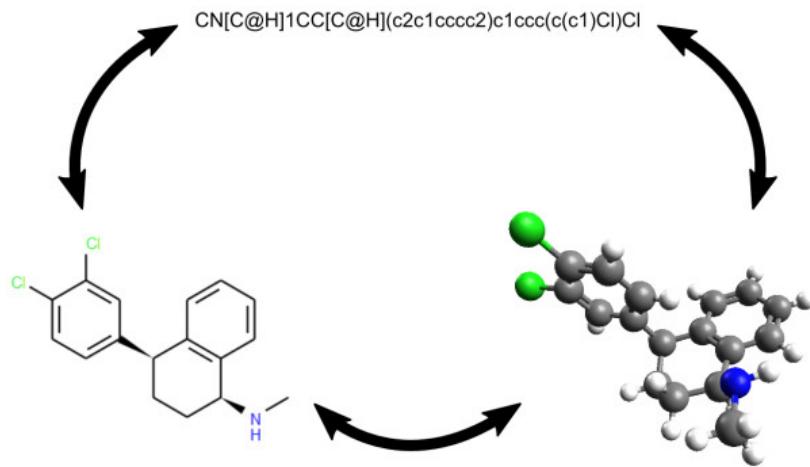


Figure 32.1: Figure 1. Interconversion of 0D, 2D and 3D structures

**Interconversion of 0D, 2D and 3D structures.** The structures shown are of sertraline, a selective serotonin reuptake inhibitor (SSRI) used in the treatment of depression. A SMILES string for sertraline is shown at the top; this can be considered a 0D structure (only connectivity and stereochemical information). From this, Open Babel can generate a 2D structure (bottom left, depicted by Open Babel) or a 3D structure (bottom right, depicted by Avogadro), and all of these can be interconverted.

Coordinate generation in 3D was introduced in Open Babel version 2.2, and improved in version 2.3, to enable conversion from 0D formats such as SMILES to 3D formats such as SDF (Figure 1). The 3D structure generator builds linear components from scratch following geometrical rules based on the hybridization of the atoms. Single-conformer ring templates are used for ring systems. The template matching algorithm iterates through the templates from largest to smallest searching for matches. If a match is found, the algorithm continues but will not match any ring atoms previously templated except in the case of a single overlap (the two ring systems of a spiro group) or an overlap involving exactly two adjacent atoms (two fused ring systems). After an initial structure is generated, the stereochemistry (cis/trans and tetrahedral) is corrected to match the input structure. Finally, the energy of the structure is minimized using the MMFF94 forcefield<sup>27</sup><sup>28</sup><sup>29</sup><sup>30</sup><sup>31</sup> and a low energy conformer found using a weighted rotor search.

While the 3D structure builder produces reasonable conformations for molecules without rings or with ring systems for which a template exists, the results may be poor for molecules with more complex ring systems or organometallic species. Future work will be performed to compare the results of Open Babel with other programs with respect to both speed and the quality of the generated structures<sup>32</sup>.

### 32.3.6 Stereochemistry

A recent focus of Open Babel development has been to ensure robust translation of stereochemical information between file formats. This is particularly important when dealing with 0D formats as these explicitly encode the perceived stereochemistry. Open Babel 2.3 includes classes to handle cis/trans double bond stereochemistry, tetrahedral

<sup>27</sup> Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94

<sup>28</sup> Merck molecular force field .2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions

<sup>29</sup> Merck molecular force field .3. Molecular geometries and vibrational frequencies for MMFF94

<sup>30</sup> Merck molecular force field .4. Conformational energies and geometries for MMFF94

<sup>31</sup> Merck molecular force field .5. Extension of MMFF94 using experimental data, additional computational data, and empirical rules

<sup>32</sup> Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress

stereochemistry and square-planar stereochemistry (this last is still under development), as well as perception routines for 2D and 3D geometries, and routines to query and alter the stereochemistry.

The detection of stereogenic units starts with an analysis of the graph symmetry of the molecule to identify the symmetry class of each atom. However, given that a complete symmetry analysis also needs to take stereochemistry into account, this means that the overall stereochemistry can only be found iteratively. At each iteration, the current atom symmetry classes are used to identify stereogenic units. For example, a tetrahedral center is identified as chiral if it has four neighbors with different symmetry classes (or three, in the case where a lone pair gives rise to the tetrahedral shape).

### 32.3.7 Forcefields

Molecular mechanics functions are provided for use with small molecules. Typical applications include energy evaluation or minimization, alone or as part of a larger workflow. The selection of implemented force fields allows most molecular structures to be used and parameters to be assigned automatically. The MMFF94(s) force field can be used for organic or drug-like molecules<sup>2728293031</sup>. For molecules containing any element of the periodic table or complex geometry (i.e. not supported by MMFF94), the UFF force field can be used instead<sup>33</sup>. Recently, code implementing the GAFF force field<sup>3435</sup> was also contributed and released as part of version 2.3. All of the forcefields allow the application of constraints on particular atom positions, or particular distances.

Several conformer searching methods have been implemented using the forcefields, all based on the “torsion-driving” approach. This approach involves setting torsion angles from a set of predefined allowed values for a particular rotatable bond. The most thorough search method implemented is a systematic search method, which iterates over all of the allowed torsion angles for each rotatable bond in the molecule and retains the conformer with the lowest energy. Since a systematic search may not be feasible for a molecule with multiple rotatable bonds, a number of stochastic search methods are also available: the random search method, which tries random settings for the torsion angles (from the predefined allowed values), and a weighted rotor search, a stochastic search method that converges on a low energy conformer by weighting particular torsion angles based on the relative energy of the generated conformer. With Open Babel 2.3, conformer search based on a genetic algorithm is also available which allows the application of filters (e.g. a diversity filter) and different scoring functions. This latter method can be used to generate a library of diverse conformers, or like the other methods to seek a low energy conformer<sup>36</sup>.

## 32.4 Implementation

### 32.4.1 Technical Details

Open Babel is implemented in standards-compliant C++. This ensures support for a wide variety of C++ compilers (MSVC, GCC, Intel Compiler, MinGW, Clang), operating systems (Windows, Mac OS X, Linux, BSD, Windows/Cygwin) and platforms (32-bit, 64-bit). Since version 2.3, it is compiled using the CMake build system<sup>3738</sup>. This is an open-source cross-platform build system with advanced features for dependency analysis. The build system has an associated unit test framework CTest, which allows nightly builds to be compiled and tested automatically with the results collated and displayed on a centralized dashboard<sup>39</sup>.

To simplify installation Open Babel has as few external dependencies as possible. Where such dependencies exist, they are optional. For example, if the XML development libraries are not available, Open Babel will still compile successfully but none of the XML formats (such as Chemical Markup Language, CML) will be available. Similarly,

<sup>33</sup> UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations

<sup>34</sup> Development and testing of a general amber force field

<sup>35</sup> Automatic atom type and bond type perception in molecular mechanical calculations

<sup>36</sup> Confab - Systematic generation of diverse low-energy conformers

<sup>37</sup> NOTITLE!

<sup>38</sup> Mastering CMake: A Cross-Platform Build System

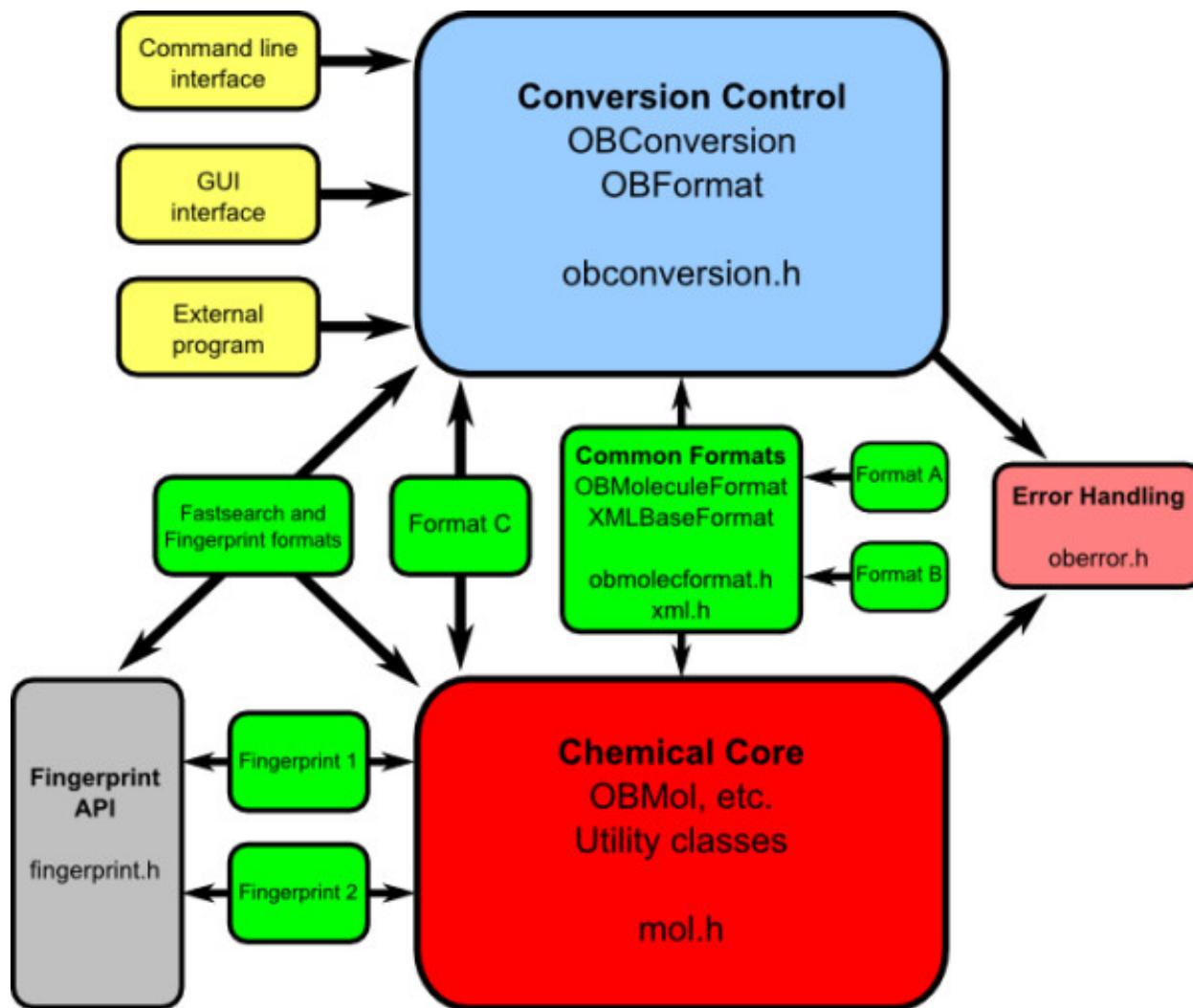
<sup>39</sup> NOTITLE!

if the Eigen matrix and linear algebra library is not found, any classes that require fast matrix manipulation (such as OBAAlign, which performs least squares alignment) will not be compiled.

While the majority of the Open Babel library is written in C++, bindings have been developed for a range of other programming languages, including Java and the .NET platform, as well as the so-called “dynamic” scripting languages Perl, Python, and Ruby. These are automatically generated from the C++ header files using the SWIG tool. As described previously<sup>40</sup>, in the case of Python an additional module is provided named Pybel that simplifies access to the C++ bindings. These interfaces facilitate development of web-enabled chemistry applications, as well as rapid development and prototyping.

### 32.4.2 Code Architecture

The Open Babel codebase has a modular design as shown in Figure 2. The goal of this design is threefold:



**Architecture of the Open Babel codebase.**

---

<sup>40</sup> Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit

1. To separate the chemistry, the conversion process and the user interfaces reducing, as far as possible, the dependency of one upon another.
2. To put all of the code for each chemical format in one place (usually a single file) and make the addition of new formats simple.
3. To allow the format conversion of not just molecules, but also any other chemical objects, such as reactions.

The code base can be considered as consisting of the following modules (Figure 2):

- The Chemical Core, which contains OBMol etc. and has all of the chemical structure description and manipulation. This is the heart of the application and its API can be used as a chemical toolbox. It has no input/output capabilities.
- The Formats, which read and write to files of different types. These classes are derived from a common base class, OBFormat, which is in the Conversion Control module. They also make use of the chemical routines in the Chemical Core module. Each format file contains a global object of the format class. When the format is loaded the class constructor registers the presence of the class with OBConversion. This means that the formats are plugins - new formats can be added without changing any framework code.
- Common Formats include OBMoleculeFormat and XMLBaseFormat from which most other formats (like Format A and Format B in the diagram) are derived. Independent formats like Format C are also possible.
- The Conversion Control, which also keeps track of the available formats, the conversion options and the input and output streams. It can be compiled without reference to any other parts of the program. In particular, it knows nothing of the Chemical Core: mol.h is not included.
- The User Interface, which may be a command line application, a Graphical User Interface (GUI), or may be part of another program that uses Open Babel's input and output facilities. This depends only on the Conversion Control module (obconversion.h is included), but not on the Chemical Core or on any of the Formats.
- The Fingerprint API, as well as being usable in external programs, is employed by the fastsearch and fingerprint formats.
- The Fingerprints, which are bit arrays that describe an object and which facilitate fast searching. They are also built as plugins, registering themselves with their base class OBFingerprint which is in the Fingerprint API.
- Other features such as Forcefields, Partial Charge Models and Chemical Descriptors, although not shown in the diagram, are handled similarly to Fingerprints.
- The Error Handling can be used throughout the program to log and display errors and warnings.

### 32.4.3 Extensible Interface

The utility of software libraries such as Open Babel depends on the ability of the design to be extended over time to support new functionality. To facilitate this, Open Babel implements a *plugin interface* for file formats, fingerprints, charge models, descriptors, “operators” and molecular mechanics force fields. This ensures a clean separation of the implementation of a particular plugin from the core Open Babel library code, and makes it easy for a new plugin (e.g. a new file format) to be contributed; all that is needed is a single C++ file and a trivial change to one of the build files. The operator plugins provide a very general mechanism for operating on a molecule (e.g. energy minimization or 3D coordinate generation) or on a list of molecules (e.g. filtering or sorting) after reading but before writing.

Plugins are dynamically loaded at runtime. This decreases the overall disk and memory footprint of Open Babel, allowing external developers to choose particular functionality needed for their application and ignore other, less relevant features. It also allows the possibility of a third-party distributing plugins separately to the Open Babel distribution to provide additional functionality.

### 32.4.4 Open-Source License and Open Development

Open Babel is open-source software, which offers end users and third-party developers a range of additional rights not granted by proprietary chemistry software. Open-source software, at its most basic level, grants users the rights to study how their software works, to adapt it for any purpose or otherwise modify it, and to share the software and their modifications with others. In this sense, Open Source functions in similar ways to the processes of open peer review, publication, and citation in science. The rights granted by open source licenses largely coincide with the norms of scientific ethics to enable verifiability, repeatability, and building on previous results and theories.

Beyond these rights, Open Babel (like most other open-source projects) offers open development – that is, all development occurs in public forums and with public code repositories. This results in greater input from the community as any user can easily submit bug reports or feature suggestions, get involved in discussions on the future direction of Open Babel or even become a developer him/herself. In practice, the number of active contributors has increased over time through this level of open, public development (Figure 3). Moreover, it means that the development of the code is completely transparent and the quality of the software is available for public scrutiny. Indeed, since its inception, over 658 bugs have been submitted to the public tracker and fixed<sup>41</sup>.

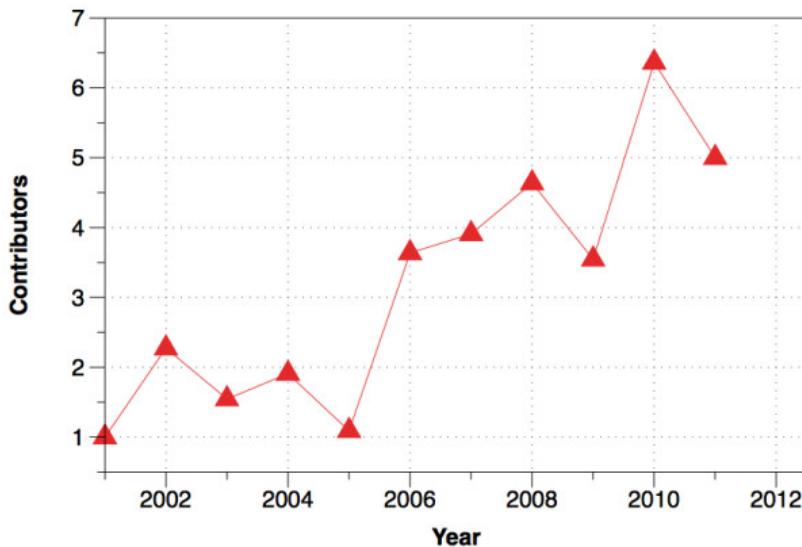


Figure 32.3: Figure 3. Number of contributors over time

**Number of contributors over time.** Note that this graph only includes developers who directly committed code to the Open Babel source code repository, and does not include patches provided by users.

### 32.4.5 Validation and Testing

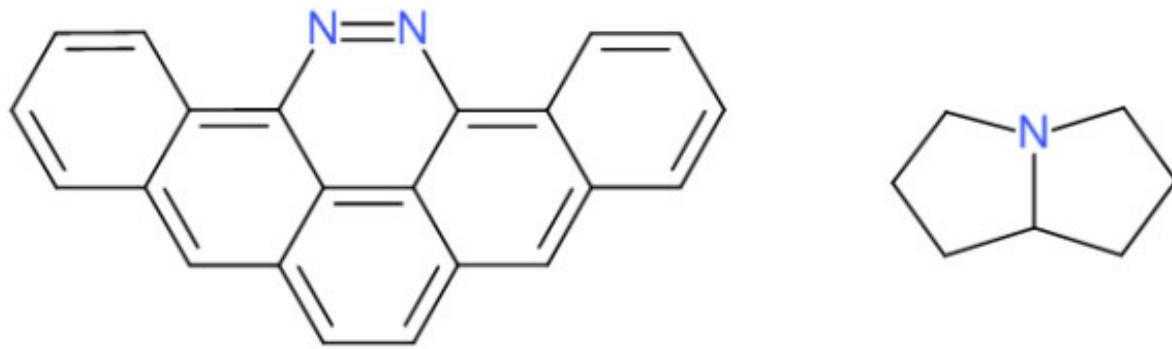
Open Babel includes an extensive test suite comprising 60 different test programs each with tens to hundreds of tests. In early 2010, a nightly build infrastructure and dashboard was put in place with support from Kitware, Inc. This has greatly improved code quality by catching regressions, and also ensures that the code compiles cleanly on all platforms and compilers supported by Open Babel. Some examples of tests that are run each night are:

1. The MMFF94 forcefield code is tested against the MMFF94 validation suite.
2. The OBAccurateAlign class, which was developed using Test-Driven Development (TDD) methodology, is run against its test suite.
3. Handling of symmetry is validated by converting several test cases between SMILES, 2D and 3D SDF, and InChI (there are also several test programs with unit tests for the individual stereo classes in the API).

<sup>41</sup> NOTITLE!

4. The SMARTS parser is tested using over 250 valid and invalid SMARTS patterns, and the SMARTS matcher is tested using 125 basic SMARTS patterns.
5. The LSSR (Least Set of Smallest Rings) code is tested for invariance against changing the atom order for a series of polycyclic molecules.

Recently the development team has placed a major focus on increasing the robustness of file format translation particularly in relation to the commonly used SMILES and MDL Molfile formats. Translating between these formats requires accurate stereochemistry perception, inference of implicit hydrogens, and kekulization of delocalized systems. While it is difficult to ensure that any complex piece of code is free of bugs, and Open Babel is no exception, validation procedures can be carried out to assess the current level of performance and to find additional test cases that expose bugs. The following procedure was used to guide the rewriting of stereochemistry code in Open Babel, a project that began in early 2009. Starting with a dataset of 18,084 3D structures from PubChem3D as an SDF file, we compared the result of (a) conversion to SMILES, followed by conversion of that to Canonical SMILES to (b) conversion directly to Canonical SMILES. This procedure can be used to flush out errors in reading the original SDF file, reading/writing SMILES (either due to stereochemistry errors or kekulization problems), and is also a test (to some extent) of the canonicalization code. At the time of starting this work (March 2009), the error rate found was 1424 (8%); by Oct 2009, combined work on stereochemistry, kekulization and canonicalization had reduced this to 190 (~1%), and continued improvements have reduced the number of errors down to two (shown in Figure 4) for Open Babel 2.3.1 (~0.01%). The first failure is due to a kekulization error in a polycyclic aromatic molecule incorporating heteroatoms: (a) gave c1ccc2c(c1)c1[nH][nH]c3c4c1c(c2)ccc4cc1c3cccc1 while (b) gave c1ccc2c(c1)c1nnc3c4c1c(c2)ccc4cc1c3cccc1. This error led to confusion over whether or not the aromatic nitrogens have hydrogens attached (they do not). The second failure involves confusion over the canonical stereochemistry at a bridgehead carbon: (a) gave C1CN2[C@@H](C1)CCC2 while (b) gave C1CN2[C@H](C1)CCC2. This is actually a meso compound and so both SMILES strings are correct and represent the same molecule. However the canonicalization algorithm should have chosen one stereochemistry or the other for the canonical representation.



PubChem CID 9107

PubChem CID 12558

Figure 32.4: Figure 4. The two failures found in the validation test for reading/writing SMILES  
**The two failures found in the validation test for reading/writing SMILES.**

Another area of focus was the canonicalization algorithm, which can be used to generate canonical SMILES as well as other formats. The algorithm can be tested by ensuring that the same canonical SMILES string is obtained even when the order of atoms in a molecule is changed (while retaining the same connection table). The test stresses all areas of the library, including aromaticity perception, kekulization, stereochemistry, and canonicalization. The development of the canonicalization code in Open Babel was guided by applying this test to the 5,151,179 molecules in the eMolecules catalogue (dated 2011-01-02) with 10 random shuffles of the atom order. At the time of the Open Babel 2.2.3 release, there were 24,404 failures of the canonicalization algorithm; this has now been reduced to only four (shown in Figure 5, < 0.001%). The Open Babel nightly test suite ensures that this test passes for a number of problematic molecules. Although the canonicalization algorithm is still not perfect, we believe that the current level of

performance (99.99992% success on the eMolecules catalogue) is acceptable for general use and with time we intend to improve performance further.

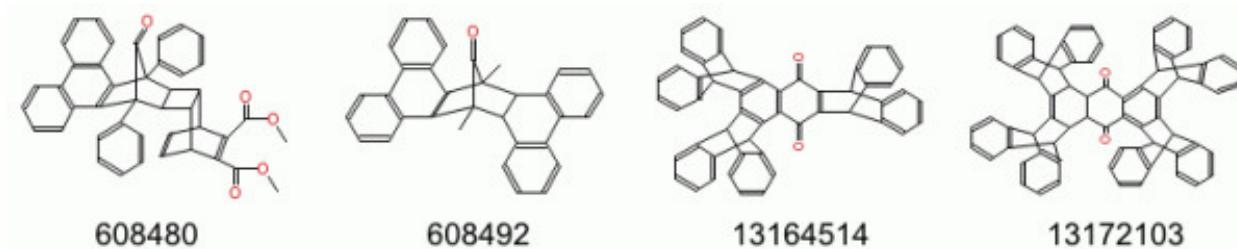


Figure 32.5: Figure 5. The four failures found in the validation test for canonicalization  
**The four failures found in the validation test for canonicalization.**

Given that the error rate for canonicalization and handling of stereochemistry is now quite low, the next area of focus for the Open Babel development team is to improve the handling of implicit valence for “unusual atoms.” This is particularly important for organometallic species and inorganic complexes.

## 32.5 Using Open Babel

### 32.5.1 Applications

The Open Babel package is composed of a set of user applications as well as a programming library. The main command line application provided is *obabel* (a small upgrade on the earlier *babel*), which facilitates file format conversion, filtering (by SMARTS, title, descriptor value, or property field), 3D or 2D structure generation, conversion of hydrogens from implicit to explicit (and vice versa), and removal of small fragments or of duplicate structures. A number of features are provided to handle multi-molecule file formats (such as SDF or MOL2) and to use or manipulate the information in property fields and molecule titles. Here is an example of using *obabel* to convert from SDF format to SMILES:

A more complicated use would be to extract all molecules in an SDF file whose titles start with “active”:

The *copy* format specified by “-o copy” is a utility format that copies the exact contents of the input file (for the filtered molecules) directly to the output, without perception or interpretation. The “-aT” indicates that only the title of the input SDF file should be read; full chemical perception is not required.

The Open Babel graphical user interface (GUI) provides the same functionality. Figure 6 is a screenshot of the GUI carrying out the same filtering operation described in the *obabel* example above. The left panel deals with setting up the input file, the right panel handles the output and the central panel is for setting conversion options. Depending on whether a particular option requires a parameter, the available options are displayed either as check boxes or as text entry boxes. These interface elements are generated dynamically directly from the text description and help text provided by each format plugin.

### 32.5.2 Programming Library

The Open Babel library allows users to write chemistry applications without worrying about the low-level details of handling chemical information, such as how to read or write a particular file format, or how to use SMARTS for substructure searching. Instead, the user can focus on the scientific problem at hand, or on creating a more easy-to-use interface (e.g. a GUI) to some of Open Babel’s functionality. The Open Babel API (Application Programming

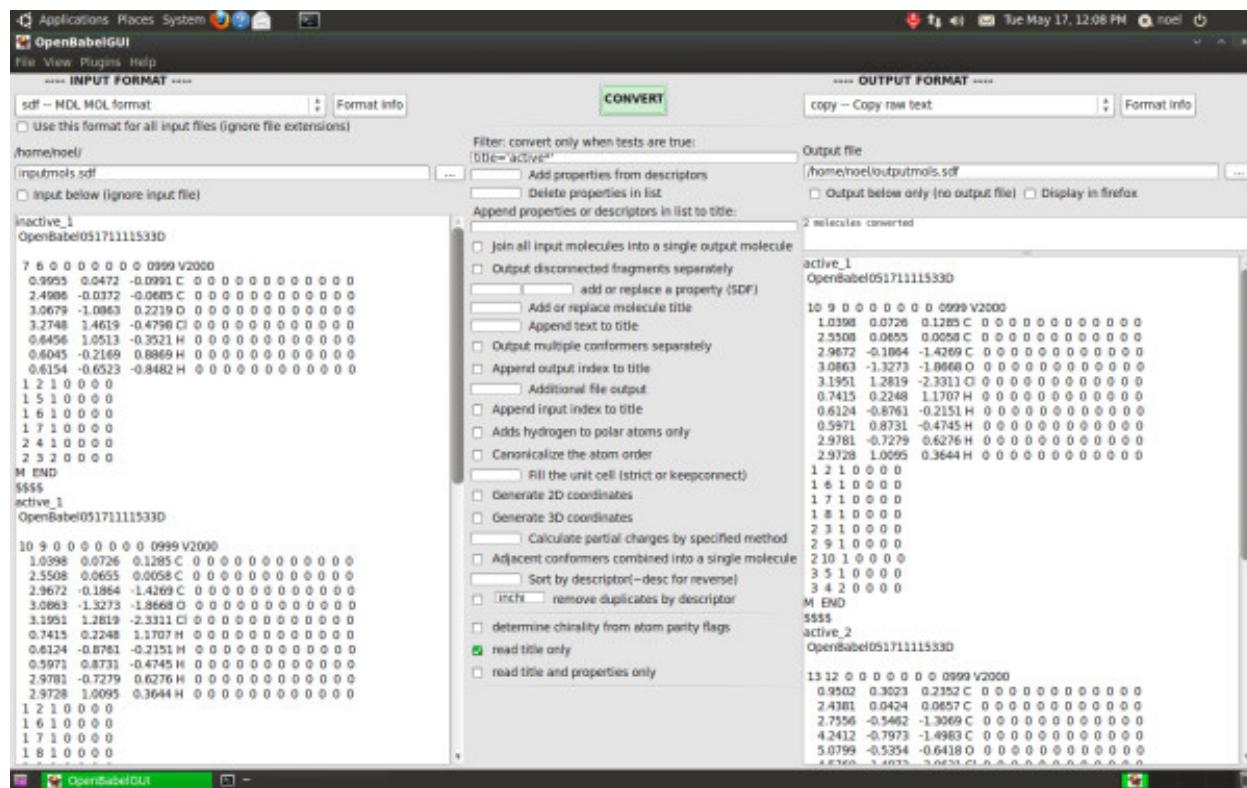


Figure 32.6: Figure 6. Screenshot of the Open Babel GUI

**Screenshot of the Open Babel GUI.** In the screenshot, the Open Babel GUI is running on Bio-Linux 6.0, an Ubuntu derivative.

Interface) is the set of classes, methods and variables provided by Open Babel to the user for use in programs. Documentation on the complete API (generated using Doxygen<sup>42</sup>) is available from the Open Babel website<sup>43</sup>, or can be generated from the source code.

The functionality provided by the Open Babel library is relied upon by many users and by several other software projects, with the result that introducing changes to the API would cause existing software to break. For this reason, Open Babel strives to maintain API stability over long periods of time, so that existing software will continue to work despite the release of new Open Babel versions with additional features, file formats and bug fixes. Open Babel uses a version numbering system that indicates how the API has changed with every release:

- Bug fix releases (e.g. 2.0.0 versus 2.0.1) do not change API at all
- Minor version releases (e.g. 2.0 versus 2.1) will add to the API, but will otherwise be backwards-compatible
- Major version releases (e.g. 2 versus 3) are not backwards-compatible, and have changes to the API (including removal of deprecated classes and functions)

Figure 7 shows an example C++ program that uses the two main classes OBConversion and OBMol to print out the molecular weight of all of the molecules in an SDF file. This could be used, for example, to investigate differences in the molecular weight distribution between two databases. The same program is shown in Figure 8 but implemented using the Python bindings.

### 32.5.3 Examples of Use

Open Babel has already been referenced over 400 times for various uses. The most common use of Open Babel is through the *obabel* command line application (or the corresponding graphical user interface) for the interconversion of chemical file formats. Such conversions may also involve the calculation or inference of additional molecular information or application of a filter. Some published examples of these include the following:

- interconversion of chemical file formats or representations<sup>44454647</sup>
- addition of hydrogens<sup>484950</sup>
- generation of 3D molecular structures<sup>515253</sup>
- calculation of partial charges<sup>5455</sup>
- generation of molecular fingerprints<sup>56575859</sup>
- removal of duplicate molecules from a dataset<sup>60</sup>

<sup>42</sup> NOTITLE!

<sup>43</sup> NOTITLE!

<sup>44</sup> A collaborative informatics infrastructure for multi-scale science

<sup>45</sup> A Database-Centric Virtual Chemistry System

<sup>46</sup> A general approach for developing system-specific functions to score protein-ligand docked complexes using support vector inductive logic programming

<sup>47</sup> A virtual library of constrained cyclic tetrapeptides that mimics all four side-chain orientations for over half the reverse turns in the protein data bank

<sup>48</sup> A Mining Minima Approach to Exploring the Docking Pathways of p-Nitrocatechol Sulfate to YopH

<sup>49</sup> A Gibbs free energy correlation for automated docking of carbohydrates

<sup>50</sup> An Evaluation of Explicit Receptor Flexibility in Molecular Docking Using Molecular Dynamics and Torsion Angle Molecular Dynamics

<sup>51</sup> A Series of Natural Flavonoids as Thrombin Inhibitors: Structure-activity relationships

<sup>52</sup> A Structure-Based Approach for Mapping Adverse Drug Reactions to the Perturbation of Underlying Biological Pathways

<sup>53</sup> Molecular modeling of the human serotonin1A receptor: role of membrane cholesterol in ligand binding of the receptor

<sup>54</sup> TMACC: Interpretable Correlation Descriptors for Quantitative StructureActivity Relationships

<sup>55</sup> AMMOS: Automated Molecular Mechanics Optimization tool for *in silico* Screening

<sup>56</sup> An Efficiently Computable Graph-Based Metric for the Classification of Small Molecules

<sup>57</sup> Bioisosteric Similarity of Molecules Based on Structural Alignment and Observed Chemical Replacements in Drugs

<sup>58</sup> Application of kernel functions for accurate similarity search in large chemical databases

<sup>59</sup> Binary Classification of Aqueous Solubility Using Support Vector Machines with Reduction and Recombination Feature Selection

<sup>60</sup> Automated procedure for candidate compound selection in GC-MS metabolomics based on prediction of Kovats retention index

```
#include <iostream>

#include <openbabel/obconversion.h>
#include <openbabel/mol.h>

int main(int argc, char **argv)
{
    OBConversion conv;
    conv.SetInFormat("sdf");
    OBMol mol;

    bool notatend = conv.Readfile(&mol, "dataset.sdf");
    while (notatend)
    {
        std::cout << "Molecular Weight: "
              << mol.GetMolWt() << std::endl;

        mol.clear();
        notatend = conv.Read(&mol);
    }

    return 0;
}
```

Figure 32.7: Figure 7. Example C++ program that uses the Open Babel library  
**Example C++ program that uses the Open Babel library.** The program prints out the molecular weight of each molecule in the SDF file “dataset.sdf”.

```

import openbabel as ob

conv = ob.OBConversion()
conv.SetInFormat("sdf")
mol = ob.OBMol()

notatend = conv.ReadFile(mol, "dataset.sdf")
while notatend:
    print "Molecular weight: %f" % mol.GetMolWt()

    mol.Clear()
    notatend = conv.Read(mol)

```

Figure 32.8: Figure 8. Example Python program that uses the Open Babel library

**Example Python program that uses the Open Babel library.** The program prints out the molecular weight of each molecule in the SDF file “dataset.sdf”.

- calculation of MOL2 atom types <sup>61</sup>

An interesting example that shows how a particular chemical representation may be used to facilitate a scientific study is the crystallographic study of Fábián and Brock who used Open Babel to generate InChI strings for molecules in the Cambridge Structural Database <sup>62</sup>. Exploiting the fact that InChIs of enantiomers are identical except at the enantiomer sublayer (“/m0” or “/m1”), they used the InChIs as part of a workflow to identify kryptoracemates (a class of racemic crystals where the enantiomers are not related by space-group symmetry) in the database.

To implement new methods, or access additional molecular information, it is necessary to use the Open Babel library directly either from C++ or using one of the supported language bindings. Some examples of published studies that have done this include the following:

- Dehmer *et al.* implemented molecular complexity measures based on information theory <sup>63</sup>.
- Langham and Jain developed a model for chemical mutagenicity based on atom pair features <sup>64</sup>.
- Fontaine *et al.* implemented a method, anchor-GRIND, that uses an anchor point of a molecular scaffold to compare molecular interaction fields when different substituents are present <sup>65</sup>.
- Konyk *et al.* have developed a plugin for Open Babel that adds support for the Web Ontology Language (OWL) to allow automated reasoning about chemical structures <sup>66</sup>.
- Kogej *et al.* (AstraZeneca) implemented a 3-point pharmacophore fingerprint called TRUST <sup>67</sup>.
- Many other examples exist <sup>68697071</sup>.

<sup>61</sup> Very fast prediction and rationalization of pKa values for protein-ligand complexes

<sup>62</sup> A list of organic kryptoracemates

<sup>63</sup> A Large Scale Analysis of Information-Theoretic Network Complexity Measures Using Chemical Structures

<sup>64</sup> Accurate and Interpretable Computational Modeling of Chemical Mutagenicity

<sup>65</sup> Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent Descriptors

<sup>66</sup> Chemical knowledge for the semantic web

<sup>67</sup> Multifingerprint Based Similarity Searches for Targeted Class Compound Selection

<sup>68</sup> Designing Focused Chemical Libraries Enriched in Protein-Protein Interaction Inhibitors using Machine-Learning Methods

<sup>69</sup> DG-AMMOS: A New tool to generate 3D conformation of small molecules using Distance Geometry and Automated Molecular Mechanics Optimization for in silico Screening

<sup>70</sup> The environmental fate of organic pollutants through the global microbial metabolism

<sup>71</sup> Substructure Mining Using Elaborate Chemical Representation

The vital role that a cheminformatics toolkit plays in the development of scientific resources is shown by Tables 1 and 2. Table 1 lists examples of stand-alone applications or programming libraries that rely on Open Babel, either calling the library directly or via one of the command-line executables. Table 2 contains examples of web applications and databases that either use Open Babel on the server or where Open Babel was used in the preparation of the data.

## 32.6 Conclusions

In November 2011, Open Babel will mark 10 years of existence as an independent project, and for the first time, we have discussed its development and features. As shown by more than 400 citations, it has become an essential tool for handling the myriad of molecular file formats encountered in diverse branches of chemistry. While more work remains to be done, through validation processes such as those described above and the recent introduction of a nightly build and testing framework, we aim to improve the quality and robustness of the toolkit with each new release.

Looking forward to the future, one of the goals of the project is to extend support to molecules that currently are not handled very well by existing cheminformatics toolkits. Typically toolkits focus on the types of molecules of principal importance to the pharmaceutical industry, namely stable organic molecules comprising wholly of 2-center 2-electron covalent bonds. Molecules outside this set - such as radicals, organometallic and inorganic molecules, molecules with coordinate bonds or 3-center 2-electron bonds - are poorly supported in general. Future releases of Open Babel will provide substantially improved handling of such species. We also seek to improve speed and coverage of important methods such as structure generation, kekulization and canonicalization.

Open Babel is freely available from <http://openbabel.org>, and new community members are very welcome (users, developers, bug reporters, feature requesters). For information on how to use Open Babel, please see the documentation at <http://openbabel.org/docs> and the API documentation at <http://openbabel.org/api>.

## 32.7 Availability and Requirements

**Project Name:** Open Babel

**Project home page:** <http://openbabel.org>

**Operating system(s):** Cross-platform

**Programming language:** C++, bindings to Python, Perl, Ruby, Java, C#

**Other requirements (if compiling):** CMake 2.4+

**License:** GNU GPL v2

**Any restrictions to use by non-academics:** None

## 32.8 Competing interests

The authors declare that they have no competing interests.

## 32.9 Authors' contributions

GRH is the lead developer of the Open Babel project. CAJ, CM, MB, NMOB, and TV are developers of Open Babel. All authors read and approved the final manuscript.

## 32.10 Acknowledgements and Funding

We would like to thank all users and contributors to the Open Babel project over its history, including OpenEye Scientific Software Inc. for their initial OELib code. We also thank the Blue Obelisk Movement for ideas, comments on this manuscript, and support. We thank SourceForge for providing resources for issue tracking and managing releases, and Kitware for additional dashboard resources. NMOB is supported by a Health Research Board Career Development Fellowship (PD/2009/13).



# ADVENTURES IN PUBLIC DATA

## 33.1 Abstract

This article contains the slides and transcript of a talk given by Dan Zaharevitz at the “Visions of a Semantic Molecular Future” symposium held at the University of Cambridge Department of Chemistry on 2011-01-19. A recording of the talk is available on the University Computing Service’s Streaming Media Service archive at <http://sms.cam.ac.uk/media/1095515> (unfortunately the first part of the recording was corrupted, so the talk appears to begin at slide 6, ‘At a critical time’). We believe that Dan’s message comes over extremely well in the textual transcript and that it would be poorer for serious editing. In addition we have added some explanations and references of some of the concepts in the slides and text. (Charlotte Bolton; Peter Murray-Rust, University of Cambridge)

### 33.1.1 Editorial preface

The following paper is part of a series of publications which arose from a Symposium held at the Unilever Centre for Molecular Informatics at the University of Cambridge to celebrate the lifetime achievements of Peter Murray-Rust. One of the motives of Peter’s work was and is a better transport and preservation of data and information in scientific publications. In both respects the following publication is relevant: it is about public data and their representation, and the publication represents a non-standard experiment of transporting the content of the scientific presentation. As you will see, it consists of the original slides used by Dan Zaharevitz in his talk “Adventures in Public Data” at the Unilever Centre together with a diligent transcript of his speech. The transcribers have gone through great effort to preserve the original spirit of the talk by preserving colloquial language as it is used at such occasions. For reasons known to us, the original speaker was unable to submit the manuscript in a more conventional form. We, the Editors, have discussed in depth whether such a format is suitable for a scientific journal. We have eventually decided to publish this “as is”. We did this mostly because it was Peter’s wish that this talk was published in this form and because we agreed with his notion that this format transmits the message just as well as a formal article as defined by our instructions for authors. We, the Editors, wish to make clear however that this is an exception that we made because we would like to preserve the temporal unity and message of this set of publications. Insisting on a formal publication would have meant losing this historical account as part of the thematic series of papers or disrupting the series. We hope that this will find the consent of our readership.

## 33.2 Introduction

(Figure 1) This article contains the slides and transcript of a talk given by Dan Zaharevitz at the “Visions of a Semantic Molecular Future” symposium held at the University of Cambridge Department of Chemistry on 2011-01-19. A recording of the talk is available on the University Computing Service’s Streaming Media Service archive at <http://sms.cam.ac.uk/media/1095515> (Endnote 1). We believe that Dan’s message comes over extremely well in the

textual transcript and that it would be poorer for serious editing. In addition we have added some explanations and references of some of the concepts in the slides and text. (Charlotte Bolton; Peter Murray-Rust, University of Cambridge)

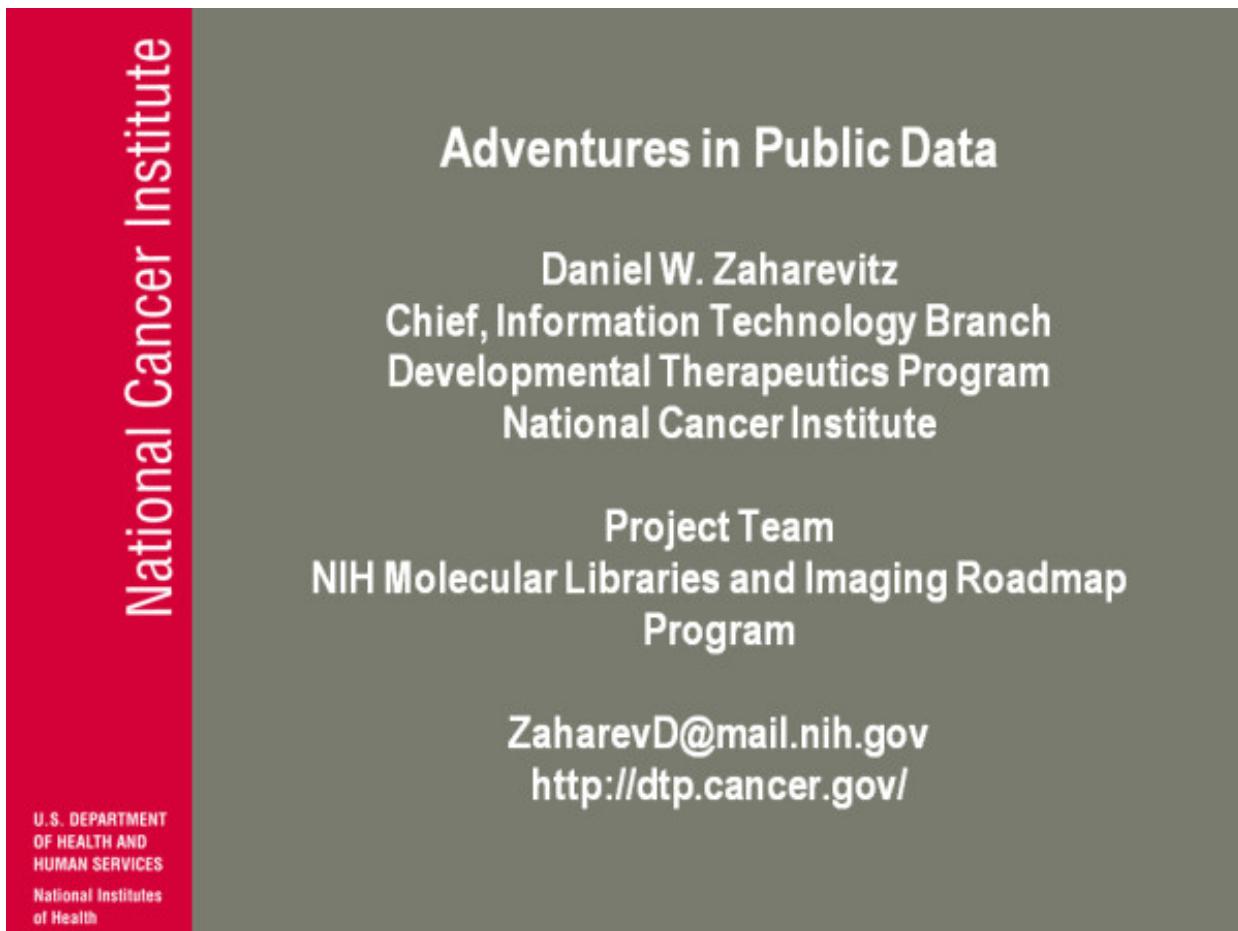


Figure 33.1: Figure 1. Introduction  
Introduction.

### 33.3 Discussion

(Figure 2) The history of the DTP (Developmental Therapeutics Program) starts in 1955 with a US Congress specific appropriation to create a national chemotherapy service centre (Endnote 2). The rationale for this was that at the time there was no interest in attempting to develop anti-cancer drugs within the pharma industry. It was thought not possible to alter the course of the disease. Compounds were screened for anti-cancer activity. The primary screen was transplantable mouse models. Over the course of >50 years that the NCI has been acquiring compounds (Endnote 3), we have registered more than 550,000 compounds. Roughly half of these (280,000) were acquired without confidentiality agreement so that data can be publicly made available.

(Figures 3 and 4) What are the sources of data? This is all on the DTP webpage. For the majority of the history, the primary screens were L1210 and P388 mouse leukemias. Some 300,000-400,000 compounds were run through these models. The screens required significant amounts of the compound. We looked at preliminary toxicology, which involved lots of animals and lots of doses, so we needed them to get lots of compound-a gram or two. This has present

## National Cancer Institute

U.S. DEPARTMENT  
OF HEALTH AND  
HUMAN SERVICES  
National Institutes  
of Health

### DTP History

- Begins in 1955 with Congressional appropriation for Cancer Chemotherapy National Service Center (CCNSC)
  - screen compounds for potential anti-cancer activity
  - no interest in anti-cancer drugs at this time in the pharmaceutical industry
  - Very interesting historical information see <http://dtp.nci.nih.gov/timeline/flash/index.htm>
  - 554917 compounds registered as of Jan 12, 2011
    - 280854 with no confidentiality agreement

Figure 33.2: Figure 2. DTP History  
**DTP History.**

day implications-if a compound was dropped early (e.g. not enough activity, too much toxicity), a large amount of the compound was left in our inventory. And this material is now publicly available.

**National Cancer Institute**

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES  
National Institutes of Health

**Assays run**  
(<http://dtp.nci.nih.gov/webdata.html>)

- **1955-1990 primary screen (300K-400K compounds)**
  - survival in transplantable mouse tumor models
  - L1210, P388
  - required significant amount (1 gram or more) of test compound
- **1990-present primary screen (100K compounds)**
  - 60 human tumor cell lines in culture (NCI-60)
- **1995-present secondary screen (4K compounds)**
  - Hollow fiber model
- **1985-present secondary screen (1K compounds)**
  - human tumor cell line xenografts in nude mice

Figure 33.3: Figure 3. Assays run  
**Assays run.**

In the 1980s it was publicly recognised that the mouse models were not general enough to pick up solid tumor agents that people were interested in. So we developed the human tumor cell-line inhibition screen, known as NCI-60. We've run roughly 100,000 compounds in last 20 years, and this screen is still active. There were a number of secondary screens dating from the mid90s to the present: hollow fibre model, where tumor cells are implanted in a semi-permeable fibre which is implanted in the mouse. Multiple fibres can be implanted in one mouse so it's possible to test multiple cell-lines per mouse. This gives us a hint of in vivo activity in an efficient and cost-effective assay. We also use human tumor xenografts in nude mouse: 1500-2000 screens in the last few years.

Because of the NCI infrastructure for acquiring and testing interesting compounds, with the sources of compounds and the data, and having the infrastructure already set up to test large scale compounds and assays when the AIDS epidemic hit, the screening for anti-HIV compounds ended up in DTP. In roughly 10 years 1990-2000, DTP assayed roughly 100,000 compounds in AIDS antiviral screens, looking for survival of cells in the presence of the virus.

There was also an attempt to create a yeast anti-cancer screen. This took yeast with known mutations, generally in the DNA repair pathway, and treated them with drugs, looking for toxicity for defined mutations. Specificity for a particular mutation gives mechanistic information.

Lastly, with the NCI-60 cell-lines, there has been an effort to characterise all the cells in these panels in wide variety of ways. This effort is ongoing. We now have 8-10 separate measures using microarrays of gene expression, so there

## National Cancer Institute

U.S. DEPARTMENT  
OF HEALTH AND  
HUMAN SERVICES  
National Institutes  
of Health

### Assays run (continued)

- 1993-2000 AIDS Antiviral Screen(100K compounds)
  - looks for cell survival in the presence of virus
- 1995-2000 Yeast Anticancer Drug Screen (100K compounds)
  - yeast cells with defined mutations (mainly in DNA repair pathways)
- Molecular Targets
  - characterization of NCI-60 cell panel
  - many microarray measurements of gene expression

Figure 33.4: Figure 4. Assays run (continued)  
**Assays run (continued).**

is lots of this available for NCI-60. It's very useful to correlate with growth inhibition patterns.

(Figure 5) So all of this created a large amount of data available to the public. Before 1995 the policy was to avoid data release if possible. The thinking behind this was several-fold. It created a lot of extra work. What data format do you use? The representation was not well settled, so it's problematic, even if you had a reasonable electronic representation, how was best to transport it to others? There are physical (tapes) and format considerations. It's very difficult to conceive of a way for widespread distribution of data without an awful lot of hands-on work. And the extra work doesn't help you to accomplish your 'real job'. The extra work doesn't get compounds into the clinic, or give you further information about how to do your job. It just sucks up time and resources.

National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES  
National Institutes of Health

## Data Release Policies (pre 1995)

- avoid any data release if at all possible
  - it is considerable extra work
    - data format (computer representation of chemical structures not widespread)
    - data transport (pieces of paper?, computer tape; which one?)
  - the extra work doesn't help accomplish the "real" job
    - won't be able to justify extra cost
  - there is limited value anyway
    - data was generated to make specific decisions; decisions that have already been made and in a production environment there is very limited ability to examine alternate decisions.

Figure 33.5: Figure 5. Data Release Policies (pre 1995)  
Data Release Policies (pre 1995).

And it's of limited value anyway. The data was generated to make specific decisions, which are already made. It's a production environment, you can't easily examine alternate decisions. People from outside might ask-why didn't you do this? That's not wrong but in the production environment you have to make decisions and move on. The next couple of thousand compounds are on the way, you must move forward. Why look at data just to rehash a decision that couldn't be re-examined anyway?

(Figure 6) In the mid 1990s there were dramatic changes. These were driven by the development of the internet-now distribution is not an issue. You can quite easily distribute data to hundreds of thousands of people all over world with virtually no effort. The formats for chemical structure were more developed, more useful, embedded in software. It was easier to give people documentation. HTML made it much easier to give people a way to collect data with clear and accessible documentation for that data: what it is, how to use it etc. You could spend less time on the phone having to explain all this!

## National Cancer Institute

U.S. DEPARTMENT  
OF HEALTH AND  
HUMAN SERVICES  
National Institutes  
of Health

### Things Change

- Internet
  - distribution vastly easier
  - formats more developed, especially chemical structure
  - easier to write documentation for a wide audience
- COMPARE
  - Ken Paull  
(<http://dtp.nci.nih.gov/docs/compare/compare.html>)
  - looking at not just one assay result, but the pattern of results across a number of assays results in a powerful tool for evaluating a compound's mechanism of action

Figure 33.6: Figure 6. Things Change  
**Things Change.**

Ken Paull (Endnote 4), former chief of the IT branch, developed COMPARE, which was looking at the NCI-60 cell-line data, not as individual assay results, is one cell-line sensitive another not, but at the overall pattern of activity. If the correlation was high between two compounds, it's likely to mean that the compounds shared the mechanism of action. It was a powerful tool to take a gross empirical assay to give a biochemical idea of what's going on. Using assay results as a pattern, to give an overall finger print of activity is a very powerful tool compared to looking at these things one at a time.

(Figure 7) And so the very detailed review came about. It was a year-long review-forty people-it was massive. And for the purposes here to talk about it I demonstrated searching and displaying structures and data on the DTP web pages and showed that you can also run COMPARE via your web interface. One of the big guns chased me out of the room where the presentation was given, was totally excited, he could see that you could sit in your living room or you could sit in your office and you could explore all kinds of ideas by just logging on to a web page, and in 1997 that wasn't exactly the most widespread notion in the world.

National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES  
National Institutes of Health

## At a critical time

- 1995 review
  - DTP was only a small part of its focus, but internally it was generally perceived as set up to severely cut or eliminate DTP
  - Realized that outside perception of DTP didn't exactly match reality and recommended a more detailed review. COMPARE had a lot to do with this
- 1997 review
  - demonstrated searching and displaying structures and data on the DTP web pages. Also could run COMPARE via a web interface
  - presented web stats for how much the pages were used. Weren't that impressive by today's web standards, but were enormous compared to people's expectations for individual service (phone, reprint request)

Figure 33.7: Figure 7. At a critical time  
At a critical time.

I also presented web stats for how many pages were accessed. The number of hits weren't that impressive by today's standards but again you're talking with people that were used to thinking of contacts as phone, reprint requests, fax requests, something like that and it was clear that your ability to respond to requests for outside information via a website was just enormous compared to the things you could think about in the 1980s. I also point out my boss at the time made the specific challenge to some of the people, the reviewers in the room, talking about the worth of the developmental therapeutics programme and asked them, you know, big drug company guys 'Can you point me to your web page where I can download your data?', and so it drives home the point that there was a difference.

(Figure 8) Lessons learned. I'm gonna call these lessons learned. You might say exaggerated extrapolations from limited knowledge but this is the thing that I take home from it. So if you think narrowly about your job, don't be surprised when the broader community thinks narrowly about what your job is worth. On the other hand if you think broadly about what your job is, it can lead to not only better tools for your specific job, you can be better at what you think you need to do but you also end up having better integration into the larger community. And a point that will come back to you time and time again is, infrastructure development is critical to enable the ability to take advantage of this broad thinking. All the good intentions in the world in the mid nineties would not have enabled us to do some of these things without the infrastructure of the internet.

**National Cancer Institute**

U.S. DEPARTMENT  
OF HEALTH AND  
HUMAN SERVICES  
National Institutes  
of Health

## Lessons Learned: I

- thinking narrowly about your job can lead to the broader community thinking narrowly about what your job is worth.
- thinking broadly about your job can lead not only to better tools for your specific job, but better integration into a larger community (i.e. science)
- Infrastructure development is critical to the ability to take advantage of thinking more broadly

Figure 33.8: Figure 8. Lessons Learned: I  
**Lessons Learned: I.**

(Figure 9) So I go now into some details of the chemical structures we collected (Endnote 5). You can download an SDF file from us and say great I'll take my structures and go on. Here is all the stuff I have to deal with to get it to an SDF file. So in 1955 collecting chemical structures meant sort of ink drawings on  $3 \times 5$  cards; that's the beginnings of our compound collection. In the 1970s there was this SANSS (Structure and nomenclature search system) which essentially was a connection table format. It gave you the atoms and which atoms were connected and what the bond order was but it had no coordinate information, no display information. This I think was partly due to CAS. There was also the EPA NIH chemical information system that was coming about here. For about twenty years starting about 1980 we had what we called the drug information system. The connection tables were stored in CAS but you also had a picture. The picture was stored in the database as HP plotter pen movement commands, so you can get a picture, you can get a connection table but you couldn't put them together at least in any useful way.

From about 2000 to present we went to a fully integrated relational database, did all those conversions and right now

## National Cancer Institute

U.S. DEPARTMENT  
OF HEALTH AND  
HUMAN SERVICES  
National Institutes  
of Health

### Data Details Chemical structures

- 1955 - ink drawings on 3X5 cards
- ~1970s - SANSS (Structure and Nomenclature Search System)
- ~1980-2000 - Drug Information System. Connection tables stored in SANSS, graphics stored as HP plotter pen movement commands.
- ~2000-present - Web interface for compound submission. Only accept structures in MDL MolFile format

Figure 33.9: Figure 9. Data Details Chemical structures  
**Data Details Chemical structures.**

we have an online submission where we only accept the structures at the moment in an MDL molfile format. At least from the beginning we have a computer representation that comes in. I should point out the entire time up to the institution of this online request system, the procedure for asking us to test a compound was the supplier would send in a picture, would send in a piece of paper, a graphic so we did not have an electronic interaction between the requester and our systems; it was all us doing transfers from some kind of picture.

(Figure 10) Considerations on what to do; how to get this into something we can make public. There were many, many format inter conversions throughout the fifty years this was going on. One thing to note, and I think it's again important, when you see it's easy to say here's a chemical structure, all chemical structures are alike, they all came from somewhere. If you don't understand where they came from you're not necessarily gonna understand what the strengths and weaknesses of various sets are. The first computer representation of the SANSS was explicitly for sub-structure searching. In some cases, for example polymers, there was no attempt to have the connection table represent a full molecule. The idea was you don't need to all that information if you're not gonna model; it was not for modelling, it was not for computing properties, it was for doing sub-structure searching. If you take a polymer, if you had say a dimer: most of the kind of substructure elements that you might search for are probably gonna be represented in the dimer. You can argue trimer or what not, but you don't need the whole thing because you're not gonna have a sub-structure that says search for a linear chain of 200 atoms or something like that. Most sub-structure searches are more limited so you don't need to bother to put the whole molecule in. So what you end up doing now is having perfectly wonderful SANSS files, that look perfectly complete, that in fact never ever had any intention of representing what that molecule was or what that substance was in a vial.

The slide has a dark grey background. On the left, a vertical red sidebar contains the text 'National Cancer Institute' and 'U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES National Institutes of Health'. The main title 'Structure Considerations' is centered at the top in white font. Below the title is a bulleted list of points in white font:

- many format interconversions
- first computer representations were explicitly for substructure searching only
  - in some cases (polymers) no attempt at all to have the connection table represent a full molecule
- display representations have fair amount of non-structural “atoms”
  - one dummy atom with the label “No Structure available”
  - labels with various composition, stereochemical and other information (“1:2”, “racemic mixture”)

Figure 33.10: Figure 10. Structure Considerations  
Structure Considerations.

The display representations also had a fair amount of what you'd call non-structural features and the one that drives me up the wall today is a structure that comes out as a perfectly legal molfile which has one dummy atom with a label "no structure available". You know, I mean, enough said about that, it still drives me up the wall... But you also have labels so there is a dummy atom that has a label that actually has something that you might want to capture. So, composition of the two parts of the substance: label it as a racemic mixture, label it as something else so maybe you don't wanna completely just delete it,' but at the same time it's a pollution of the structure with other information in a format that's hard to disentangle.

(Figure 11) Structure release. The first one we put on the NIH page, not a page, just for anonymous FTP. I think it's generally called the NCI 127 k. They were open structures for which there was a CAS number. We figured the other thing to realise is that a lot of people say 'can you give us the chemical names of all these structures?' The vast majority of these structures were not published on, or at least we don't know that they were published on: no one ever bothered to name it, there's certainly not a trivial name. And so for a lot of the structures the only identifier we had was the NSC number and of course back in 1994 nobody knew what an NSC number was except for a handful of people interacting with NCI. We didn't think that was very useful so we sub selected a set where we also had the CAS numbers. Historical aside: CAS was our input contractor for about 6 or 8 years about 1975-1983 or something, so they automatically assigned a CAS number for everything that came in. And so there were CAS numbers: you figured you might be able to search on that and so that's where this group came from. Where we got the coordinates-it was actually the SANSS connection tables, they were converted to a SD format and the programme CORINA<sup>1</sup> from Johann Gasteiger was used to generate 3D coordinates, so that was the first stage of the release.

The screenshot shows a dark grey web page with white text. At the top right, the title 'Structure release' is centered. On the left side, there is a vertical red sidebar containing the text 'National Cancer Institute' and 'U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES National Institutes of Health'. The main content area lists several bullet points about the structure release:

- First was in 1994 (via anonymous FTP); the NCI 127K
  - open structures for which there was a CAS number
  - SANSS connection tables were converted to MOL format and Corina was used to generate 3D coordinates
- Major conversion (Kekule) in internal system finished about 2000. Now extracting 2D MDL format files from internal system
- Releases about once a year on web page
  - Latest Dec 2010 - 265824 structures
- Many versions of "NCI structures" around, including multiple depositions in PubChem

Figure 33.11: Figure 11. Structure release  
Structure release.

<sup>1</sup> CORINA-Fast Generation of High-Quality 3D Molecular Models

We had our major conversion, a program called Kekule<sup>2</sup> to look at graphics and try and to chemical structure. The internal system was finished in 2000 and so now when we pull from our company database we are actually pulling MDL format files; so at least it's a format that tries to recognise chemicals structure. Right now we have releases about once a year; we're hoping that in the next year we'll go to a little bit more often than that. The latest release was a few weeks ago and 265 almost 266 thousand structures. The other thing I'll make a point about is, we've been releasing these sets for a long, long time. A lot of people have pulled them up and we have PubChem<sup>3</sup> now. A lot of people, their deposition in the PubChem was basically from a file that they pulled from us that's not documented, and that's fine, it's legal but there are certainly inconsistencies and differences in all kind of things in this data. If you go to PubChem and say 'what's the structure of "something-or-other-amycin"?' , and you could look it up and maybe you find ten versions in PubChem. Ten depositions for a compound with that name and maybe you say seven of them have the same actual chemical structure but there's these others that are different. Well I can believe that if seven people think it's this and only two people think it's that, it's probably the seven people that are correct. But it might be that those seven versions have a mistake in them that are propagated because all of them go back to downloading our structures. So it's just a heads up that without the background, without the metadata about where these structures came from, you can potentially get into problems or you can potentially be misled.

(Figure 12) So lessons learned number two. It's one of my pet peeves-a fixed set of fields are a disaster. People are gonna find a place for the information that they need to store no matter what, and if you fixed your set of fields to some gigantic number ('I'm gonna think of everything possibly people are gonna store'), there's always gonna be something you forgot and there's always gonna be a huge number of fields in that case that are never used. So what's the problem? Just stick them in as a field that's never used...but now all your careful documentation of what that field is for is polluted because people don't use it like that! Information will be appended to the expected information in existing fields, so again you have your case when you are plopping some kind of composition data, some kind of stereochemical data into a label on a dummy atom in a structure picture. If that's the place where you can put it, people put it there and you're not gonna stop them. Use XML!

So we go back to, where did I first meet Peter Murray-Rust and why am I so high on XML? It's all because of Peter. I think the first time I interacted with Peter was in 1995, maybe one of the earlier attempts at an internet chemistry poster session, having a chemistry meeting over the internet and I put in a presentation about 3D database searching. And Peter started asking questions basically along the lines of 'can we get to the point where we don't have to do the experiments?' Well gee, if you don't have all the stereochemical information, this and that, and I said 'well I don't think that's a real goal' and I'm thinking to myself 'good God man be reasonable'. Of course in the last fifteen years, that's simply not a thing you say to Peter! I mean he's never gonna be reasonable although he does it in a way that always pushes us. Thinking about this-I'm not sure 100% comes through in this talk-that a lot of this stuff really is Peter's influence, making sure we are driven in directions that are gonna be useful.

Internal database keys should be internal. When you start to make your internal keys meaningful in the external world you lose your flexibility and maintaining really good internal consistency-you should make that primary. The compound structure is an empirical result; it is not an identifier and again I don't know whether PubChem got the terminology right but their distinction between a compound and the substance I think is extraordinarily important when you talk about chemical structure data and bioassay data-I'll give an example of that

The other thing I've learned is identifier equivalencies. So you have a CAS number, you have a NSC number, you have this, you have a name, you have all kinds of stuff. Identifier equivalencies are pivotal too: there are claims people made-NSC27 is the same as CAS number blah blah blah. We can use those labels interchangeably—that's a claim, and again various people make various claims and sometimes the claim is wrong and sometimes the claim is misleading. So if you don't understand and can't manage where those claims come from and have access to them you're gonna eventually run into problems. I have an aside:

(Figure 13) Think about the difference of writing in a laboratory notebook '50 milligrams of methotrexate' or '50 milligrams of a powder from vial number 123'. To give a concrete example, there's a paper published in Science about MDMA (3,4-Methylenedioxymethamphetamine, ecstasy). They retracted the paper after they found out the bottle they had used that was labelled with MDMA did not actually contain MDMA, but methamphetamine.

The biggest problem when you make mistakes: it's embarrassing, you should double check, but the biggest problem

<sup>2</sup> Kekule: OCR-optical chemical (structure) recognition

<sup>3</sup> PubChem

## National Cancer Institute

U.S. DEPARTMENT  
OF HEALTH AND  
HUMAN SERVICES  
National Institutes  
of Health

### Lessons learned: II

- **fixed set of fields are a disaster!**
  - people will find a place for "extra" information
    - little used fields will be overloaded
    - information will be appended to the expected information in existing fields
  - use XML!
- **internal databases keys should be internal**
- **compound structure is an empirical result not an identifier**
  - PubChem compound vs. substance
- **identifier equivalences are empirical and subject to change**

Figure 33.12: Figure 12. Lessons Learned: II

**Lessons Learned: II.**

## National Cancer Institute

U. S. DEPARTMENT  
OF HEALTH AND  
HUMAN SERVICES  
National Institutes  
of Health

### Aside

- 50 mg of methotrexate was dissolved in ...
- 50 mg of powder from vial #123 was dissolved in ...
- Severe dopaminergic neurotoxicity in primates after a common recreational dose regimen of MDMA
  - Science 297:2260-3 (2002)
- retracted after they found out the bottle labeled MDMA didn't contain MDMA  
[http://en.wikipedia.org/wiki/Retracted\\_article\\_on\\_dopaminergic\\_neurotoxicity\\_of MDMA](http://en.wikipedia.org/wiki/Retracted_article_on_dopaminergic_neurotoxicity_of MDMA)
- biggest problem is that that they couldn't reliably say which experiments used compound from the problem bottle.
- **chemical structure or chemical name is a lousy primary identification field!**

Figure 33.13: Figure 13. Aside  
Aside.

here was is they couldn't even reliably say which experiments were affected by this bottle because they had done this in their notebooks. They had not identified the bottle, they just said 'yeah we used...' so chemical structure, chemical name is just a lousy primary identification field, and you're really gonna run the risk of corrupting data and not having full control of data if you don't understand this difference.

(Figure 14) A couple of these things in the last few slides are labelled community priorities and/or involvement. What I mean by that is these are things I've been thinking of that I think are useful or can potentially be useful, and obviously in this day and age if you can get people to help you actually do it that's fantastic, jump in, let's go. But for our purposes even expressing a notion 'is this a high priority or low priority', 'what kind of thing is useful to us', that helps us with our limited resources say 'well gee a whole lot of people wanna do this so maybe that is where we put our cut off' so when I say this I really would like feedback on any level.

**National Cancer Institute**

**U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES**  
National Institutes of Health

## Structure Cleanup (Community priorities/involvement)

- How to make the structure set most useful
  - 80 different elements in the set, so really tests chemical software
  - currently use CDK to compare MW generated from molecular formula to MW generated from structure
  - lots of inconsistencies in formal charge assignments
  - use or just delete labels with extra information
- How to document structures
  - when and how was the data extracted from the DTP system
  - how was the structure standardized
  - most usable way to code this information in CML
- Help in correcting/redrawing bad structures?

Figure 33.14: Figure 14. Structure Cleanup (Community priorities/involvement)  
**Structure Cleanup (Community priorities/involvement).**

One of the things I have been worried about is how to make the structure set, the data structure, the chemical structure set, more useful. There are at least eighty different elements in the set so it really does exercise any kind of chemical software. It's not just carbon, nitrogen, oxygen, blah blah blah-there's I think 300 tin compounds. We currently use the Chemical Development Kit<sup>4</sup> to compare the molecular weight generated from the molecular formula to the molecular weight generated from the structure. You go back to this problem with SANSS: did it try and represent the complete molecule or not, only some little bit? The molecular formula in our data base was always entered independent of the structure, and so if these two things match you have a little bit of added confidence that the structure that came out really does intend to represent the full structure. If they don't match then well maybe you have a problem. A lot of

<sup>4</sup> The Chemistry Development Kit

times when they don't match, it comes down to this: inconsistencies in formal charge assignments. A lot of times it is easy to see how you would clean that up: the molecular formula says 'dot-CL', the structure says 'CL-minus': OK, I understand that. Some of them are not so clear. Do you try and use what I mentioned before-do you try and use the information from these dummy atom labels or do you just forget about them?

How to document the structures-when and how was the data extracted, did it come from us, what kind of algorithms were used to do any kind of clean up or any kind of comparison. Are there beginning to get ways that people would like to see structures standardised more? The most useful way to code this is in the chemical mark-up language-I'm beginning to think that the best way to just to do it and let things evolve. But if people have strong opinions on how to represent some of this and potentially help and correcting or withdrawing bad structures from the community-we had a student in the summer crank out about 400 compounds in a couple of weeks-it might be something people may be interested in.

(Figure 15) Other data. We have our NCI screening data-this is growth inhibition in human tumour cell lines. We have this both as calculated parameters from a full dose response curve, and as a full dose response curve. I mentioned molecular target data: there's microarray of gene expression in vivo in xenografts, which I think could be very important; that data is not quite public yet but it should be soon. We have in vivo survival screens curves from those old mouse model screens, so if anybody's interested in developing software or display tools for looking at survival curves we probably have something like 200 or 250 thousand survival curves. That's publicly available.

**National Cancer Institute**

**U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES**  
National Institutes of Health

## Other Data

- NCI-60 screening data
  - parameters from dose response (GI50, TGI, LC50)
  - dose response data
- Molecular target data
- Microarray measurements of gene expression in xenografts (coming soon)
- In vivo screen survival curves
  - community priority/involvement?
- ~6500 compounds from DTP in the Molecular Libraries screening deck.
  - community priority/involvement?

Figure 33.15: Figure 15. Other Data  
Other Data.

I probably don't have time to talk about this but we have about 65 hundred compounds from our inventories now in the molecular library screening deck, so we have the ability not only to associate NCI-60 data (let's say a pattern of activity

in the NCI-60 cells) but then in some cases we have 2 or 3 hundred assays in the molecular library in PubChem that can be related to them. We haven't really started to develop ways to bring all those things together and try to find ways to best utilise them. Again, they came from us so again we have a guarantee that the NCI-60 data and the molecular library screening data actually all came from the same sample.

(Figure 16) There are other DTP resources. We have a compound repository, as I mentioned before. For a good chunk of our screening, we wanted a couple of grams. If the molecule was abandoned fairly early on we had a gram left, so there are about 80,000 compounds for which I think at least formally we have a gram in our inventory. We have an online sample request for that. We have not for the most part identified to run any analytics on this, although that's changing. You can submit compounds to the screen-we have an online submission form. In a few months we'll have web services and one of the things I'm excited about here is talking to, say, Alex [Wade]and a few people about ways to use these web services where you can manage your submissions and accounts in a Word program or something similar. We also have a COMPARE service server set up so you can do these calculations. We have put them up as web services but we haven't really taken advantage of that, building alternative interfaces to this and alternative ways of putting it together with other things. We haven't really gone beyond that.

The slide has a red vertical header bar on the left containing the text 'National Cancer Institute' and 'U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES National Institutes of Health'. The main content area has a grey background. At the top, the title 'Other DTP resources' is displayed in bold. Below the title is a bulleted list of three items: 'Compound repository', 'Submitting compounds to NCI-60 screen', and 'COMPARE server'. Each item has a corresponding list of sub-points.

- Compound repository
  - online request for samples from our repository
  - about 80K compounds available
  - purity/identity not guaranteed
- Submitting compounds to NCI-60 screen
  - online submission form
  - soon (few months) will have web services available for submission
- COMPARE server
  - also accessible via web services

Figure 33.16: Figure 16. Other DTP resources  
**Other DTP resources.**

(Figure 17) In terms of what my fantasy-well, hopefully it's not a fantasy (but you know...)-how about an anti cancer discovery workbench? Sort of a Bioclipse<sup>5</sup> where you can do structure activity for growth inhibition, connect to COMPARE to look for compounds with similar patterns and mechanism. You could order compounds from DTP, you could connect to COMPARE, you can do correlations to COMPARE, not only look at growth inhibitions versus growth

<sup>5</sup> Bioclipse

inhibition but growth inhibition versus molecular target data so you can prepare growth inhibition to gene expression, see if a compound has been tested in NCI-60, submit compounds: basically a platform for testing new ways to analyse structure and data, and also to connect people that do computational work with synthetic chemists. If the synthetic chemist is sitting at his bench using this to submit the compounds to NCI the other tools there would be available to them to do ways of hopefully prioritising and maximising their chances of getting good structures submitted.

Figure 33.17: Figure 17. Anti-cancer Discovery Workbench (community priority/involvement)  
**Anti-cancer Discovery Workbench (community priority/involvement).**

(Figure 18) And in the last few minutes just talk a little bit about sort of philosophy. So the it's a critical time for the research community-the funding outlook is dismal, everybody knows that grant applications success rates are very low, dissatisfaction with therapeutic pipelines, nobody is happy with the rate and what looks like the therapeutics. You can say 'well, more money could be a big help', and it could save some things in some ways but the question is how you justify it.

(Figure 19) Let me give you a specific problem and this is very dear to my heart because this should be exactly what the developmental therapeutics programme is enabling. There is a paper, published in Nature Medicine in 2006, and it looked for genomic signatures to guide the use of chemotherapeutics. They started with our NCI-60 data and our microarray data, fantastic! They downloaded it and they could do something: they went beyond it, they got a Nature Medicine paper and it led to clinical trials. Other questions arose-a letter to Nature Medicine back and forth, but then a group at MD Anderson [Cancer Center] couldn't quite figure out how they got the results they got. The group at MD Anderson can download our data as well but couldn't quite balance it; the first group would cooperate but they couldn't figure things out. It led finally to this Annals of Applied Statistics paper where the group at MD Anderson laid out (they called it 'forensic bioinformatics'), what they tried to do, how they tried to go about it and the fact that they

## A critical time for the research community

- funding outlook is dismal
- grant application success rates are very low
- dissatisfaction with therapeutic pipelines
- more money could be an enormous help, but how can it be justified?

Figure 33.18: Figure 18. A critical time for the research community  
**A critical time for the research community .**

couldn't really be sure in what was going on. In taking public data, publishing a paper but not being able to connect the dots so you know you can... Here's one particular URL that has a fairly reasonable overview: resume problems arose, clinical trials halted, a very big mess and it's a huge problem.

National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES  
National Institutes of Health

## A Specific problem

- Genomic signatures to guide the use of chemotherapeutics. *Nat Med.* 2006 Nov;12(11):1294-300. Leads to clinical trials
- Questions Arise - *Nat Med.* 2007 Nov;13(11):1276-7; author reply 1277-8.
- Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.* Volume 3, Number 4 (2009), 1309-1334.
- Resume problems, clinical trials halted, big mess  
[http://journals.lww.com/oncology-times/Fulltext/2010/09100/Duke\\_Scandal\\_Shines\\_Light\\_on\\_Systemic\\_Problems\\_in.1.aspx](http://journals.lww.com/oncology-times/Fulltext/2010/09100/Duke_Scandal_Shines_Light_on_Systemic_Problems_in.1.aspx)

Figure 33.19: Figure 19. A Specific problem  
A Specific problem.

(Figure 20) What you have is an example of what could be a general problem: it's a well funded group in a well respected institution, published results in a peer review journal, favourably reviewed clinical trial plan. What do you say to the patients when a clinical trial is halted? ‘Oops, sorry, my bad, no problem. By the way can you write your congressman and tell them they got to double our NIH budget?’ There is a real disconnect there. So what kind of answers can you give? Oh you know “it was a bad apple”; “this guy is bad”; “he lied so he probably faked it”; blah blah blah. I don’t think any of these are acceptable, they are primarily ways to avoid responsibility by making it a specific problem, but there’s too many interactions with the entire research community here to avoid a more general responsibility. The basic fundamental problem is not any of these (no it’s not my problem excuses); the problem is the research community did not demand full accountability, the parameters for all this review was not something that was at all able at all to catch this.

(Figure 21) So I call it, Peter says something sometimes about ‘take back our scholarship’, I say ‘take responsibility for our scholarship’: insist the data supporting publications be accessible, useable, documented and complete, and recognise adhering to this standard. Is what science is, it’s not an add-on, it’s not an extra burden from on high. I claim if you don’t understand your data well enough to export it, you don’t understand your data well enough to use it, and it’s in all scientists’ interests to prevent sloppy scholarship. I don’t think anybody in the research community benefits from the mess of the genomic signature paper, whether you were directly a part of it or not. I work in the molecular

## The General Problem

- Well funded group at a well respected institution published results in a well respected peer reviewed journal that led to a favorably reviewed clinical trial plan. What do you say to the patients when the clinical trial is halted?
- Unacceptable answers (if you expect to argue successfully for more money)
  - "bad apple"
  - peer review is the journal's problem
  - not my field
  - not my job
  - too much work

Figure 33.20: Figure 20. The General Problem  
**The General Problem.**

library. There is, and again this is indirectly a tribute to Peter, the ethos and the setup of how data was handled in the molecular library. I strongly argued for [this] and my arguments were influenced by Peter's, but the idea is to make sure that from the beginning the data was released as soon as it was verified. But you need to budget for it and you need to work at it and administratively you need to keep at these people to do it and it's been constant: it works but it doesn't 'just happen'.

**National Cancer Institute**

**U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES**  
National Institutes of Health

## Take responsibility for our scholarship

- insist that data in support of a publication be accessible, usable, documented, and complete
- recognize that adhering to this standard is an integral part of scientific scholarship, not an irrelevant add-on to be resisted
  - if you don't understand your data well enough to export it, you don't understand it well enough to use it.
  - it is in all scientist's interest to prevent sloppy scholarship
  - Molecular Library experience - you need to budget for it and work at it
- think broadly and carefully about what is promised to the public in return for their support. Make sure that community standards, policies, procedures, etc. work toward this goal first and narrower goals secondarily.

Figure 33.21: Figure 21. Take responsibility for our scholarship  
**Take responsibility for our scholarship.**

So in general you know I think you need to think broadly and carefully about what is promised to the public in return for their support and you have to make sure that all these community standards policies and procedure work toward that goal first the goal of delivering what you are claiming the public benefits from, and all the other goals are secondary: all the prestige, the money and all that stuff.

(Figure 22) Concrete steps: Open Scholarship, Open Access. There's a lot of people here that are gonna talk more than I about that, but I think that from my perspective, in addition to a kind of philosophical approach, it's a very practical approach where you get better error detection, potentially better results, better able to connect.

The other thing I'm thinking about is whether can we actually take that [genomic signature] fiasco and turn it around and say 'here's how we would do it with Open data', 'here's how we would do it in a more documented way' and have maybe an Open Genomic Signature Workbench, so we have in vitro gene expression, we have growth inhibition data, we're going to publicly soon have in vivo gene expression data so we have all the pieces. NCI has the xenograft testing possibilities so we have all the pieces for not only generating drug expression, drug sensitivity relations but *testing* them before you start to go to the clinic and you can do it in a transparent, documented and reproducible way. We can show people how it should be done.

## Concrete Steps

- Open Access, Open Scholarship
  - Better error detection (even prior to submission)
  - Better results
- Open Genomic Signature Workbench?
  - in vitro gene expression data
  - growth inhibition data
  - in vivo gene expression data coming soon
  - possibility for not only generating gene expression - drug sensitivity relations, but testing them in vivo.
  - can do it in a transparent, documented, reproducible way.

Figure 33.22: Figure 22. Concrete Steps  
**Concrete Steps.**

(Figure 23) So here's my email address. All those things-remember whenever your priorities, interests, needs: I'm from the government, I'm here to help you.



Figure 33.23: Figure 23. Conclusion  
Conclusion.

[Applause]

Question from Egon Willighagen: Are all the characterizations other than the gene expression data of the NCI-60 publicly available?

Dan Zaharevitz: Yes-we have lots of other different characterisations so we have metabolomics data we have enzyme activity measurements but in terms of number of data points the largest set of data in the molecular targets set of data is gene expression data just by number but there's a lot of other things in there as well

## 33.4 Endnotes

### 33.4.1 Endnote 1

Unfortunately the first part of the recording was corrupted, so the talk appears to begin at slide 6, 'At a critical time'.

### 33.4.2 Endnote 2

PMR: The NCI research was for many years the outstanding example of Open, publicly financed research and data collection. It stemmed from President Nixon's "war on cancer" which captured the spirit of the moonshots but also shows that biology is tougher than physics.

### 33.4.3 Endnote 3

PMR: The systematic testing of public and private compounds was a key strategy for NCI DTP.

### 33.4.4 Endnote 4

PMR: Ken Paull's contribution to DTP was dramatic. The COMPARE program is one of those archetypal tools which is both very simple and very powerful. It's a table "browser" for the DTP data with compounds == rows and screens == columns. By tabulating hits compounds can be compared by activity in screens and screen can be compared \*\*by activity\*\* of compounds. And it emphasizes the importance of having lots of data, carefully aligned, and the tools to manipulate it.

### 33.4.5 Endnote 5

PMR: In the early years data was "paper". Chemical structures were hand drawn.

# THREE STORIES ABOUT THE CONDUCT OF SCIENCE: PAST, FUTURE, AND PRESENT

## 34.1 Abstract

In this piece I would like to tell a few stories; three stories to be precise. Firstly I want to explain where I am, where I've come from and what has led me to the views that I hold today. I find myself at an interesting point in my life and career at the same point as the research community is undergoing massive change. The second story is one of what the world might look like at some point in the future. What might we achieve? What might it look like? And what will be possible? Finally I want to ask the question of how we get there from here. What is the unifying idea or movement that actually has the potential to carry us forward in a positive way? At the end of this I'm going to ask you, the reader, to commit to something as part of the process of making that happen.

## 34.2 A story of the past

But let us start in the past. My scientific career starts with a book by Isaac Azimov, “*Life and Energy*”<sup>1</sup> that sat as a child on my parent’s bookshelves. I’ve never seen another copy of it. I couldn’t even remember the title until I went digging on Amazon. It was about biochemistry, about how energy is obtained, transformed, stored and used in living systems. Even when I read it in the early 1980s it was woefully out of date, first published in the early 60s. But I was hooked, I wanted to be a biochemist, and I wanted to do research. I did well at school and I did well at University. I started my first real research project in 1994 at UWA, looking at which molecules human platelets would select from plasma to generate ATP<sup>2</sup>. You can draw a straight line from that research project back to the ideas I absorbed from the Azimov book.

I can also vividly remember my first research supervisor explaining to the new intake of students how research worked, how lab books needed to be kept but above all his view on keeping abreast of literature:

“You need to spend half a day each week, reading all the new journals that have come in”.

That statement dates me. For half the people reading I would guess that it is totally incomprehensible on at least two levels. How could you read that much? And who ever leafed through the pages of a journal? For the other half I would guess it raises a nostalgia for the days when it was possible to do just that, when perhaps there was time to dedicate half a day or more to keeping up with the literature and when it involved the pleasant task of sitting in a quiet space, paging through the 10 or 20 new issues that might have come in.

---

<sup>1</sup> NOTITLE!

<sup>2</sup> Fuel choices by human platelets in human plasma

I went from this, somewhat cozy world, into a PhD at the Australian National University and the world started to change. Over the course of my PhD the web became central to doing research. Going to the library went from the weekly excursion to an occasional trip. Paging through journals changed to clicking through emailed tables of contents and as that became untenable shifted to search. Medline had appeared online and this changed the game for someone in the biological sciences. The year I started my PhD was the year the *E. coli* genome was released (not the year it was published incidentally, that waited until 1996<sup>3</sup> but the sequence was available). I remember having to manually change the memory allocation for Netscape Navigator so we could download it.

The world had changed, papers were available online, email was now essentially ubiquitous within the university and data was becoming more and more readily available online, as the PDB and human genome projects lead the way in pushing data into publicly accessible repositories. But at the same time not much had changed, in fact not much has still changed another fifteen years later. The PDF, the version of the paper everyone seems to want was still (mostly!) a dead document. It was just a digital dead document rather than a paper one. The business models hadn't shifted that much. The attitudes and culture of academics hadn't really changed with the media and the wider public still often held in suspicion or even contempt. Yet at the same time big changes were afoot, changes that we are still struggling to work through today.

In retrospect I probably did the wrong PhD. I went to a lab trying to do something wildly, excitingly, and perhaps naively, ambitious. I didn't get the experience of working in a lab that churns out papers, that has that well-oiled machine running and that remains a hole in my experience. A good PhD delivers both experience and a good set of papers that can then provide a bit of a cushion as a researcher explores more ambitious and speculative postdoctoral projects. But in 1999 it was still possible to apply for fellowships and postdocs with only a single paper without being laughed at. I was lucky enough to get a Wellcome Trust fellowship and then 14 months later I got a lectureship position at the University of Southampton.

The next five or six years was full of extremes. I got my first grant but it, again, was really too ambitious. Again that lack of experience was playing out. I was involved in some big projects, some parts of which worked well, some parts of which did not. In amongst all of this I'd jumped at the option of applying for some BBSRC funding in the e-science area, working on developing a laboratory notebook system with Jeremy Frey's group. My motivation in doing this was purely selfish. I wanted to raise some money to support a PhD student. However looking back at that proposal now I had at some level seen that there was a problem of data sharing. The area I was working in, directed evolution, had a lot of papers, a lot of positive results but no theory, no real understanding of how things worked at an abstract level. There simply wasn't any data to help build the models that would predict how to do the experiments so people just did lots of experiments, reporting the ones that worked. The idea of our project was to provide a framework to enable people to share data, particularly data around unsuccessful experiments to support the development of a theoretical framework for the field.

I wrote that grant in 2005, but I was as yet unaware of open access, or the open data movement. In fact I wrote a scathing reply back to a survey from Nucleic Acids Research, that was at the time proposing to move to author processing charge supported open access. They of course ignored this<sup>4</sup>. But as I went down the track of exploring the ideas of data availability, of what the web can do you pretty quickly become a convert. It is difficult not to be struck by the potential of the web once you get your head out of the tunnel vision that a research career creates. Many people have been struck by these possibilities, I wasn't the first, and I certainly wasn't the last. The potential to improve the process of research is immense but it remains largely unrealised. And the reasons it is unrealised are pretty well established. There is no short term motivation, beyond a desire to do the right thing, to build the tools, and to change practice. All of these things require work, work that is not rewarded, or rather is not rewarded in a way that maps well onto getting a research position or getting promotion, or indeed in today's world, just keeping your job.

2005 marked another departure for me. I moved to the Rutherford Appleton Lab where I now head up biological sciences at the ISIS neutron source. I wanted to work somewhere where working *with* people was valued more, but the main reason was because I saw a big potential for neutron scattering to contribute to structural biology in a unique and valuable way. This would require some significant investment but the time was right in terms of the capabilities of new instruments, computational infrastructure, and data analysis tools, to make a real difference. Strategically it was a great opportunity to really do something significant and to make a big difference.

---

<sup>3</sup> The complete genome sequence of Escherichia coli k-12

<sup>4</sup> Editorial

Fast forward five years and that opportunity again remains largely unrealised. The resources haven't really been there due fundamentally to restrictions in research budgets, to work at the level required to make the breakthrough, not the scientific breakthrough, but the breakthrough in terms of awareness of the possibilities and willingness to try these techniques out amongst the wider community.

So is the failure my fault? Well certainly in part. I focused too much on strategy and not enough on tactics. We spread ourselves too thin and raised expectations too far as to what we would provide. But at the end of the day the strategic opportunity that I see doesn't map onto the strategic priorities of the funders enough to make it happen. And I don't have the stature, as a structural biologist, to make the case and make it stick because I don't have the Nature papers that are needed to even get into the room.

In the area of web technology and scholarly communications I do have the stature to get into the room. And I think it's interesting to ask what the difference is. Is it simply the standards are lower in this new area or did I just get in early enough to get in at the ground floor. Is there something particular about my skill set that is a better fit to web science, or is it down to different styles and means of communication? Papers vs. blogs? Referees reports vs. twitter?

This matters because I've reached a point where I realise that what matters to me is working to make the biggest difference I can given the resources I have to deploy. If writing papers is the way to do it, then I'll write papers. If writing blog posts is more efficient I'll do that. Obviously the real answer lies in a balance of the two, reaching different audiences for different purposes but finding that balance is important if I'm to make the most of the limited resources that are my time and energy. And in particular if I am to deliver the most benefit for the public investment in those resources.

For me this is brutally pragmatic. I advocate open approaches and help to develop open tools because I believe that they will ultimately deliver the best return on the public investment in research. If someone can convince me that subscription based business models and hiding data behind pay walls is the most effective way of delivering that return then I will man the barricades with directors of the subscription based publishers. I don't think this is likely. I don't think those approaches offer good value for money either in economic terms or for social and community returns but in my opinion we should remain focused on the need to responsibly discharge the public trust granted in us in spending research funding. And we live in very interesting times when it comes to both the level of that trust, and the view on how well we are discharging it.

### 34.3 A story of the future

I've taken you from the mid 1970s to today, now let me jump 30 or 40 years into the future. About the time when I might hope to be retiring. This is a somewhat utopian vision albeit one hopefully grounded in reality but I think it is important to note the possibility of a dystopian future. There is a possible future in which the US congress is controlled by the Tea Party, leading to the destruction of US federal research funding. A future in which stagnant economic recovery in Europe is accompanied by continuing crises of confidence in the honesty of the research community leading to another flat cash or worse settlement in future spending reviews. There is a future in which the whole scientific research process does not retain (or perhaps regain) public trust. We should acknowledge that, and act accordingly.

But there is a brighter future as well. One which is more efficient, if perhaps smaller. One where there is more central coordination of resources but greater federation and distribution of research work and of research roles. This is a future that takes advantage of the fact that enabling specialization in particular tasks and skills can improve efficiency and it is therefore not a future in which all researchers take on the same set of roles, but one in which groups, perhaps institutions, perhaps even countries, specialise in specific tasks in data collection, analysis, building and maintaining infrastructure, and effective communication. This is a future in which most research projects will be international in scope but with centralised resources and frameworks that support these collaborations and make them work efficiently.

Let us think of a young researcher, one with relatively little experience because I think this is where the real interest is. How do you both train and enable less experienced researchers to contribute effectively? It is likely that we will have a smaller funded research community so making the most of everyone's abilities and time is crucial. A young researcher might start their day logging on and checking what new data has come in overnight. This is a researcher who is starting

out so they'll be probably be doing relatively basic data management. They might be doing categorisation, or perhaps some simple analysis, spotting interesting cases that can be pushed up the chain to more experienced analysts. Some of these might in turn be tagged as learning opportunities that come back down again for our young researcher to take on themselves.

Our researcher is probably some time away from collecting the data themselves as this is a specialised and highly skilled role, one in which particular people excel and are therefore encouraged to focus. Similarly they probably didn't design this particular experiment but signed on to a project created and managed somewhere else. Projects looking for this kind of support will be easy to find because the problems of metadata collection and standardization that we face today will largely be solved by having them embedded in the systems that collect the data. Nonetheless these systems will still have limitations and human categorisation and spotting of edge cases will still be necessarily, an area where our young researcher can contribute effectively, probably in parallel with a number of others. Each process that they carry out will be logged, the provenance recorded, and the metadata automatically captured via the context of their actions.

Our researcher is motivated and interested. Maybe they want to get into data collection, or into building the software systems that support their work. Maybe they're just interested in getting more to grips with the underlying science. They will be tracking a wide range of relevant communications, all openly accessible. There might be a new paper published in Australia, a conference keynote in Brazil, or a discussion panel in Utrecht they want to catch. The timezones make this difficult but all of these communications are available and discoverable. They have all been linked to each other, and conversations about all of them are available online.

Our researcher doesn't understand a point made by the speaker in Brazil and asks a question. It turns out to be a common misunderstanding so the question is handled by a professional educator based in South Africa rather than being sent to the speaker themselves. They have a more interesting question for the discussion panel and a moderator sends this to the panel itself. Our researcher gets a credit for asking a good question and the answer helps them to build their case for getting more responsibility in their data analysis project. They download the Australian paper in audio form for their commute home later in the day and then set off a quick re-run of the paper's data analysis but with the parameters changed so as to compare it to the work they did on their own project this morning.

There are lots of things that could happen next, we could talk about how the data is marked up and integrated, what systems are required to manage the data markup, or who is paying for the moderators and educators to do their work, but at this moment all of this needs to stop. Because our young researcher's dad has come into their bedroom and told them to stop mucking around on the computer and get ready for school.

None of this should be surprising because almost all of it is already here today. It is certainly all possible today. Tracking remote events via video streaming and twitter is commonplace. Data can be obtained from online repositories and analyses re-run via workflow engines. Analysis can be distributed to systems that are part human and part computational. What is different in my story is the ability to integrate these systems. The sharing of common vocabularies and APIs can allow a multitude of such systems to interact. A key difference is a system of reputation that transcends one single service but can be used to gain access to people, to use their time, because in the past you've offered good value. This works on a small scale, at the level of a StackExchange<sup>5</sup> or a GalaxyZoo<sup>6</sup>, but not in a way in which we can barter with people's time. People's time, expert attention, remains the most valuable resource we have in research. We are still some way from good systems for helping us to decide how best to use it at the level of research systems.

What is different is a shared framework with a *stable* and *trusted* infrastructure, rather than the rubber bands and string systems that we often use today to jury rig demonstrations of what might be possible online. Today you can give a talk remotely at a conference, but you don't want to be relying on it. Backups are required and even sending in a pre-recorded video can be a risk. But at a higher level, the question of strategic allocation of research resources we also have no shared infrastructure. It isn't possible to test my opinion on strategic priorities versus that of a traditional structural biologist in a systematic way beyond asking the opinion of trusted people. You can't model the choice to support this rather than that or tension my record on strategic thinking with the domain knowledge of a top person in structural biology.

It is the framework, the trust, and the systems that could help us to apportion valuable resources that make the dif-

---

<sup>5</sup> Stack Overflow

<sup>6</sup> Galaxy Zoo

ference between where we are today and this future vision. In a world where the physical experiments are probably largely done by robots (humans don't generate reproducible enough results) and computational systems have enough capacity that you can choose to simply try every possibility on the basis that someone might want it someday. The central issue therefore becomes pushing the right problem to the right available person depending in their skills, availability, and interest.

This world requires a different approach to the design of research projects, with much more modular parts, standardised inputs and outputs. We have to be careful that this standardisation doesn't limit the science that can be done, and remember that there will always be bespoke efforts pushing the boundaries, but the benefits of such an approach are enormous. Anyone can ask a question and see whether it has already been answered. If it hasn't it can be tested to see if it is a good question and how it relates to current knowledge. If it is worth doing then automated systems can be brought into action to determine whether the results are interesting.

The difference between the utopian and dystopian futures described here is public engagement in science. My suspicion is that if we can't bring interested members of the public into the process of research then we won't be looking at a happy future in terms of funding. Galaxy Zoo<sup>6</sup> and Foldit<sup>7</sup> show that these approaches can work, and although these may be relatively low hanging fruit many of the lessons learnt can be applied more widely. Smaller scale projects also work without the exciting interface, high profile subject area, or a need for huge numbers of volunteers. The Open Dinosaur Project<sup>8</sup> is making real progress simply by asking people to copy the length of leg bones from research papers into a Google spreadsheet.

The key is to always be identifying the opportunities for more people to become involved and how to reconfigure research to make it more modular and easier to divide up. If standards across data, samples, analysis and frameworks are used then much more of this can be done by people at home than you might think. Treat the public with contempt and they will do the same for us. Treat them with respect and invite the interested ones in and they will become our strongest advocates. They can be much better for public relations than anything our own communication systems could ever achieve. Authenticity and personal interest are what matter in the networked world, not who has the phone number of the science correspondent at the BBC.

The future of course will be totally different. Prediction is a mug's game, but the key themes of standardisation, modularity, sustainability, and open frameworks are what make a positive future possible, regardless of what form it takes. And all of these things can enable genuine engagement in a way which is only just possible today and would have been unimaginable ten or twenty years ago. A positive future depends on pulling these strands together and actually making the web work for science both in the way Tim Berners-Lee intended and in the way that Tim O'Reilly, Jon Udell and Clay Shirky saw was possible as the social web emerged over the past decade. But the key aspects, engagement, standards, open approaches, solid infrastructure are what will take us forward in a positive direction.

## 34.4 A story of the present

So if we return to the present, the space we sit in now, how can we take this vision forward? What can we rally around, what can we agree on, that will provide the focus, and the necessary incentives for us to be more efficient and more effective? I think there is something, but it's not what most of you expect. I think the thing that can take us forward as a community is Research Impact.

Now hear me out here. "Impact" has become something of a dirty word amongst the research community. I don't think the introduction of impact statements by government funders has been handled as well it might have been, and the message has become a bit muddled, but impact is just a word, and an agenda, and if we re-focus on the real agenda and reclaim the word then I think we can actually make it something we can all agree on. The UK science minister, David Willets has a quite sophisticated understanding of what he means by impact. It's not just economic impact, and it's not just short term practical outcomes. It is about the capacity to innovate, capacity to use innovation from overseas, as well as long term and unexpected outcomes from research that might not look to have practical outcomes at the outset. What we are really talking about is maximizing the opportunity for research outputs to be re-used. We

---

<sup>7</sup> FoldIt

<sup>8</sup> Open Dinosaur Project

need to re-structure the research enterprise so as to maximise re-use and the potential for re-use. Re-use might be by other researchers, it might be by industry, or it might be in educational settings or in public health. But re-use *is* impact in a very real sense.

Researchers, like any human being are motivated to a certain extent by fame and fortune, but equally most researchers are also motivated by the wish for their research to make a difference. A real difference; not the difference of publishing a paper, but the difference of seeing that paper cited, seeing its findings used by other researchers, and seeing it applied to real world problems. This is impact; seeing our research re-used. And we should be configuring our research efforts, ruthlessly if need be, to maximise the ability of our research to be re-used. Not just by other researchers, although this is an important audience, but by small and medium enterprise, large companies, patients, schoolchildren, teachers, doctors, engineers, and government.

How do we maximise re-use? Largely through open approaches. The unexpected uses far outweigh the expected one so protecting and hiding results is for the most part counterproductive. It serves only the short term interest of the researcher. “I haven’t finished analysing this data”? Tough. If someone else can do it faster, they should. In the worst-case scenario someone is dying because that data wasn’t made available or an opportunity to avert environmental catastrophe is missed. Or perhaps just some poor postdoc somewhere is replicating your work again, wasting money that could be spent on more productive work. Yes we need replication, yes we will need to configure systems so that some of it can be done blinded, but these are easy things to arrange.

We will also need a portfolio of research without clear applications. If we believe that this kind of exploratory, non-applied research is where the big unexpected advances come from then we need to support it to maximise impact. We need to accept that much of this work will have only small benefits, much of it will be incremental. And that it is essentially impossible to pick winners in advance. But someone has to fill in the tables before we move on to the next theory, the next model. Maximising impact is not just about research published in Nature. It’s not about publishing papers at all. Publishing is the start of the story, not the end. And it’s not just about “the best science”, not if you take a long term view; it’s about the right blend. It’s not just that not everyone can be in the top 50% but that not everyone *should* be in the top 50%. But we need a ruthless focus on configuring our research work so as to maximise its re-usability. Open approaches, standardised approaches, high standards of replicability.

If we focus on the potential impact of research and maximising it we can see a clear route towards more efficient and effective, more open and more standardised research approaches. We would be engineering systems and configuring a community that was both more federated and perhaps in some ways more centrally supported through infrastructure provision. But how do we get from the individualistic, secretive, personality driven present to this future? How do we reconfigure the incentives in our research culture to drive this change?

Again I think, if we think of impact as re-use the answer becomes obvious. Currently we measure and reward outputs. How many papers? How many patents? How many successful grants? If impact and re-use are our goals then this is what we should measure. At some level we already do this, citation counts, and H-factors are measures of re-use, if extremely crude and somewhat misleading ones. If we could measure the re-use of data, the application of new theories, the development of products and services out of research results and value people’s contribution on this basis then we can both satisfy the government agenda, address the public engagement agenda, drive cultural change in the research community and provide real incentives for people to work on the infrastructure, both technical and cultural, that will make the vision of the future possible. If the incentives align with optimizing research for downstream re-use then the community will optimize their outputs to ensure re-usability.

It will directly drive a move to open access, open data, and open process because these directly support re-use. It will directly drive improvements in reproducibility because reproducibility supports re-use. It will directly drive standardisation and modularisation because these support the ability of others, as well as ourselves, to re-use and apply the results of our research. Measure people on the basis of the re-use of their research, reward them for that and the rest will follow.

So I promised audience participation. What I want you to do is look at the following statements. Absorb them. If you feel so moved stand up where you are and say them aloud. Share them with others and above all think about how they apply to your work:

I want my work to make a difference.

I will act to optimise the potential for my work to make a difference.

I will persuade others to optimise the potential for their work to make a difference.

Ok, you can sit down now. I'm not asking you to adopt these today, or to change what you are doing here and now. What I'm asking is that you think about how the choices we make in how we discharge the public trust invested in us to spend public money in a sensible and informed way should shape the way we do research. Think, and discuss with others how best to take that investment and turn it into a public good over the long, and also the short, term.

We don't really have a choice about the Impact Agenda, but we have a choice about how we approach it. We don't really have a choice about improving public engagement, but we have a choice about how we think about and interact with the wider public. We do have a choice about how we act as a community to discharge the public trust vested in us, to optimise the efficiency and effectiveness of the public investment in research. And we have a moment in time where we need to seize the opportunity to make that choice.



# OPENNESS AS INFRASTRUCTURE

## 35.1 Abstract

The advent of open access to peer reviewed scholarly literature in the biomedical sciences creates the opening to examine scholarship in general, and chemistry in particular, to see where and how novel forms of network technology can accelerate the scientific method. This paper examines broad trends in information access and openness with an eye towards their applications in chemistry.

## 35.2 Commentary

Science, like culture, is grounded in stories. Science has long sought to make sense of the information we receive from the world around us, resolved to tell stories that are supported by data, that explain why the sun comes up in the east, and goes down in the west, and does so every day, without recourse to mystical beings. And the way we communicate science belongs to this long storytelling tradition: we write papers, and publish them, so that others might know the stories of what happened in a given laboratory at a given time, that someone found the crystal structure of DNA, or that light behaves like a wave and a particle at the same time.

These stories are validated by their presence in journals, collections of stories, bound up and published monthly, many physically printed and mailed out even in the digital age. If the story is in a famous journal, it's trustable. If the story is in John's Journal Of Chicken-Fried Science-or, less facetiously, a journal that is bought and paid for by a pharmaceutical company<sup>1</sup>-it is not. This trust comes from the brand of the journal, built over the years through the recruitment of trusted scientists to serve as peer reviewers. And this entire method encases the idea that individual scientists, the principal investigators, are romantic entities at the core of the laboratory, shouting Eureka and running naked through the halls after proving a new theorem.

The truth is of course a lot more complex. Principal investigators depend on postdoctoral and graduate students. The paper is merely an advertisement for years of research(Endnote 1), a snapshot of a far more complex knowledge generation process, but for hundreds of years, it's been the best knowledge compression technology available to us. The papers have become finely tuned objects where some of the text is used to show the author understands the existing paradigm of the field, some of the text is used to describe the methods and results, and some is used to describe the implications. Each of these sections needed to be terse, as paper was expensive to print and ship.

This hid the fact that science was, in fact, actually much more like a wiki. Every topic in science is open for back and forth, and new discoveries spark rounds of editing and re-editing, and the print equivalent of flame wars in biting letters to the editor. But it was a wiki that was camouflaged as physical media. And in an era of increasingly computerized science, with automated and massively parallel lab equipment pumping data into massively parallel processing power, we're starting to see an absolutely overwhelming increase in the number of digital papers. Leaving behind the irony of digital paper, there is a strong parallel in science today to when cities crested ahead of their sewer systems and

---

<sup>1</sup> Merck Published Fake Journal- from The Scientist

highways-industrial knowledge production capacity, pre-colombian recycling capacity. Science is drowning in its own outputs, and a lot of those outputs are turning out to be either non-reproducible <sup>2</sup> or downright false <sup>3</sup>.

What we need is a full-scale revolution in the way that we publish knowledge, and there are many claimants to carry the standard of that revolution. Some are from the “radical incrementalism” school(Endnote 2)-into which I would put Open Access, a movement that puts literature online, free of charge, and free of copyright restrictions other than providing credit to the author (there are several core definitions of Open Access, but I am quoting here from the Budapest Open Access Initiative <sup>4</sup>), as well as the movement to separate the subjective judgement of impact from a more objective judgement of scientific validity in the peer review process <sup>5</sup>. Others go farther, arguing for the abandonment of the article as the core unit of knowledge transfer, for nano-publication of individual assertions <sup>6</sup>, for the publication of figures or data rather than articles <sup>7</sup>, for the rise of wiki science and the end of peer review entirely <sup>8</sup>.

It’s an explosion in our capacity to capture data that is a large player in the explosion of papers, and in the various claimants to the revolution in publishing knowledge. We now have massively parallel ways to measure reactions, run experiments, capture information about the state of the world. But the publication revolution (that is, beyond radical incrementalism) will not occur without some new help. The promised Fourth Paradigm of Science <sup>9</sup> will require that we build new systems into the existing *data* infrastructure that we have for science.

Infrastructure used to be something physical-highways, in the common world, or big buildings and expensive machines in the science world (such as the Large Hadron Collider, or the Hubble Telescope). The rise of the network has brought a new layer of physical infrastructure, from the fiber across which bits flow to the server farms and compute clusters and clouds where processing now takes place, all connected by yet another crucial element-the standard protocols by which data and documents and music files and more are broken up into packets, routed, transported, and reassembled. And one of the most important sets of protocols is the set we know generally speaking as the Web. It’s the stuff that lets us share documents, and it’s changed the world.

But in the case of complex adaptive systems-like the body, the climate, or our national energy usage-the data are usually not part of a document. They exist in massive databases which are loosely coupled, and are accessed by humans not through search engines but through large-scale computational models. There are so many layers of abstraction between user and data that it’s often hard to know where the actual data at the base of a a set of scientific claims reside.

This is at odds with the fundamental nature of the Web. The Web is a web of documents. Those documents are all formatted the same way, using a standard markup language, and the same protocol to send copies of those documents around. Because the language allows for “links” between documents, we can navigate the Web of documents by linking and clicking. Because the right to link is granted to creators of web pages, we get lots of links. And because we get lots of links (and there aren’t fundamental restrictions on copying the web pages) we get innovative companies like Google that index the links and rank web pages, higher or lower, based on the number of links referring to those pages <sup>10</sup>. Google doesn’t know, in any semantic sense, what the pages are about, or what they mean. It simply has the power to do clustering and ranking at a scale never before achieved, and that turns out to be good enough.

But in the data world, very little of this applies. The data exist in a world almost without links. There is no accepted standard language, though some are emerging <sup>11</sup>, to mark up data. And if you had that, then all you get is another problem-the problem of semantics and meaning. So far at least, the statistics aren’t good enough to help us really structure data the way they structure documents.

There is one emerging world of data, often location-based data, where we can make a lot of progress. It’s the world of apps that help you know when the bus will be at a given stop in Boston, and thus avoid the cold <sup>12</sup>. It’s one that doesn’t

---

<sup>2</sup> The Truth Wears Off: Is there something wrong with the scientific method?

<sup>3</sup> An Epidemic of False Claims: Competition and conflicts of interest distort too many medical findings

<sup>4</sup> Budapest Open Access Initiative

<sup>5</sup> Open-access journal will publish first, judge later

<sup>6</sup> Nano-Publication in the e-science era

<sup>7</sup> Introducing FigShare: a new way to share open scientific data- blog post at the Open Knowledge Foundation, retrieved on 5/31/11 at

<sup>8</sup> The End of Peer Review and Traditional Publishing as We Know It

<sup>9</sup> The Fourth Paradigm: Data-Intensive Scientific Discovery

<sup>10</sup> PageRank: Bringing Order to the Web

<sup>11</sup> Resource Description Framework

<sup>12</sup> Catch the Bus, iPhone application

worry much about data integration, or data interoperability, or data infrastructure, because it's simple data-where is the bus and how fast is it going?-and because it's mapped against a knowledge system we have had for hundreds of years, that we understand, and which is...well, a map.

But the world of modern science isn't so simple. Doing deeply complex modeling of climate events, of energy usage, of cancer progression-these are not so easy to turn into iPhone apps. The way we treat them shouldn't be with the output of a document. It's the wrong metaphor. We don't need a "map" of cancer-at least not in the classical sense of a 2-dimensional representation. We need a model that tells us, given certain inputs, what our decision matrix looks like. And the infrastructure for documents doesn't get us there.

So, I have made the argument for more infrastructure. That imposes the requirement that I say what I mean by infrastructure. I believe there to be at least three essential elements missing.

First is the infrastructure to collaborate scientifically. Laboratories are natural breeding grounds for collaboration and conversation-reagents are shared, coffee and tea are drunk, journal club is hosted. Virtual collaboration lacks these elements that form the circadian rhythms of a group, and this absence of shared rhythm dogs collaborative projects far beyond the sciences (Endnote 3). We have seen some infrastructure for distributed collaboration in software, like github, but as yet this has not emerged in the sciences (and indeed may need to evolve discipline by discipline as needs and local context dictate).

Another missing link is that of classification. Before the web, classification was a library or taxonomical function, imposed from above by hierarchical authority, famously subject to bias, prejudice, and sheer incompetence (Endnote 4). But with the advent of the web, we see the rise of "categories, links, and tags" as emergent systems of classification, ones that are plenty good enough to help us fine web pages about ourselves, ratings of local restaurants, or lengthy rants against ontology. We no longer need a file system, we just need the right search string (and of course, services that provide us the search capacity).

But science actually fits many of the elements where expert classification and formal ontology actually make some sense-formal categories, expert users, authoritative sources of judgement, etc. And in particular, the problem that automated machine-generated data imposes of an explosion of unstructured content means that the emergent classification on which the Web runs doesn't emerge, *because there aren't any people tagging it and linking it*. We have to have at least some formal classifications to impose to help us deal with big data, but science doesn't like to fund that sort of work nearly as much as it does the creation of new (you guessed it) papers.

The last one is thankfully the easiest of the three. It is the infrastructure for *data openness*. It's composed of open data (Endnote 5) licenses (Endnote 6)(covering not only copyright and database rights, but issues of privacy, identity, and more (Endnote 7)), legal user interfaces to make sure users understand the terms, and technological implementations for licenses, so that machines can negotiate and discover the terms under which a given piece of data is (or isn't) available. This infrastructure for openness draws on successes in free software and free culture, where open licenses have been part of the creation of entire ecosystems of co-creation that would otherwise have been impossible <sup>13</sup>.

Open data also helps us address the first two elements of missing infrastructure. It's highly unlikely that any one scientific funder, or any one company, will develop the right system for collaboration across sciences, or even across a single discipline in the sciences like chemistry. Open data means that the disciplines can each evolve towards their own systems of collaboration, that the marketplace of ideas can take place without high transaction costs to try, and often fail, at new methods to work together. Open data also helps address the classification problem, again by lowering the cost at which one group attempts to organize their information, and by creating a culture in which classification schemes are themselves shared, remixed, hacked, and subjected to incremental improvement-but also ready to be torn down and rebuilt when the data indicate.

There are two striking examples of open data that we can look to as inspiration for chemistry. One is in astronomy, where there is a longstanding tradition (caused in part by scarce, and thus shared, physical resources like radio telescopes) of sharing open data, as well as an evolved, open source infrastructure for virtual collaboration (Endnote 8). Openness has become the norm, and has allowed for classification and collaboration to emerge over time, so that now the serious work of astronomical science takes place in the open.

<sup>13</sup> The Wealth of Networks: How Social Production Transforms Markets and Freedom

A second is more emergent, and more scattered, in biology. Biology has for years been like chemistry-laboratory focused, principal investigator driven-and subject to enormous competitive pressures with the boom of the biotechnology industry. But the larger the data become, and the more complex the human body is discovered to be, the more open data becomes the only tractable methodological approach that accelerates science. Chemistry itself is seeing a flowering of expertise in novel methods of publication and knowledge construction, though only time will tell which approach will become part of the infrastructure of the discipline (Endnote 9).

Thus, the pharmaceutical industry itself has systematically invested in the public domain of data, from the Single Nucleotide Polymorphism Consortium (Endnote 10) to the Structural Genomics Consortium (Endnote 11). As the pharmaceutical industry is well known to embrace patent rights in many areas, its decade-long investment in, and support of, open data is a telling example of the market finding its own way towards openness as infrastructure that simply accelerates science. The recent advent of Sage Bionetworks, another non profit data sharing project, promises to bring the same kind of benefits to disease biology, moving from “fundamental” data like sequences and structures to experimental and clinical information.

Taken together, these three skeins of collaboration, classification, and openness draw us inevitably towards the long-claimed, but rarely-achieved, goal of the scientific method: to make claims that are reproducible under similar circumstances by someone other than the claimant, to be reproducible.

The road to implementing the three new levels of data infrastructure face barriers. Science is complex, and even if we implement on all three levels, that won’t magically create new insights. The Alzheimer’s Disease Neuroimaging Initiative ran for nearly a decade as an open data, open collaboration project, with standardized ways to classify the images, before its research breakthroughs made it into the peer-reviewed (wait for it) papers <sup>14</sup>. There is a lag time between when we invest in infrastructure and when we see the results, and we will have to be patient.

But open data will in the end win out, just as open systems have won out for networking, for document sharing, for software, and are beginning to win for culture and education. It is, in the end, the better way to do science, one in which there is less duplication of effort, less fraud, more reproducibility, more return on investment, and faster times to market of knowledge. It is, moreover, one that returns scientific data to its most natural state, one that is a pure public good, that gains more value as more people possess it.

## 35.3 Endnotes

### 35.3.1 Endnote 1

Apocryphal, but told to the author by Victoria Stodden.

### 35.3.2 Endnote 2

I owe this phrase to a conversation with Christine Borgman of the University of California of Los Angeles.

### 35.3.3 Endnote 3

See The World Opera project for a fascinating example at <http://theworldopera.org/>- debates that never occur in a normal opera, such as “should we have a real conductor at one location, an avatar, or just a metronome?” must be resolved before a collaborative performance in real time can be achieved.

---

<sup>14</sup> Sharing of Data Led To Results On Alzheimer’s

### 35.3.4 Endnote 4

Clay Shirky has written a lovely deconstruction of classification called “Ontology is Overrated”—available at [http://www.shirky.com/writings/ontology\\_overrated.html](http://www.shirky.com/writings/ontology_overrated.html). This paragraph draws on his arguments at multiple points, but I encourage readers to read the whole article, including his high praise of the periodic table of the elements as a high-water mark in classification.

### 35.3.5 Endnote 5

See the Open Knowledge Definition at <http://www.opendefinition.org/okd/>—although I dispute the idea that data necessarily equals knowledge, I still like the definition’s spirit.

### 35.3.6 Endnote 6

See Creative Commons’ CC0 legal tool at <http://creativecommons.org/publicdomain/zero/1.0/> for an example of an implementation of the OKD for data.

### 35.3.7 Endnote 7

This is a space where the naive “porting” of open infrastructure for software and culture fails. Privacy constraints, especially around human subjects data, are totally orthogonal to the right to make and distribute copies. This is a key area for future work and research.

### 35.3.8 Endnote 8

See the International Virtual Observatory Alliance, at <http://www.ivoa.net/>, for a remarkable example of international virtual science based on public domain data.

### 35.3.9 Endnote 9

For example, <http://www.openphacts.org/>,  
[http://en.wikipedia.org/wiki/Blue\\_Obelisk](http://en.wikipedia.org/wiki/Blue_Obelisk),  
<http://chem2bio2rdf.org>

[http://semanticweb.com/semantic-chemistry\\_b10684](http://semanticweb.com/semantic-chemistry_b10684),  
[http://en.wikipedia.org/wiki/Open\\_Notebook\\_Science](http://en.wikipedia.org/wiki/Open_Notebook_Science),

### 35.3.10 Endnote 10

The SNP Consortium (TSC) was established in 1999 as a collaboration of several companies and institutions to produce a public resource of single nucleotide polymorphisms (SNPs) in the human genome. The initial goal was to discover 300 000 SNPs in two years, but the final results exceeded this, as 1.4 million SNPs had been released into the public domain at the end of 2001. In the end, 1.8 million SNPs were released. More than \$50,000,000 was contributed to fund this project, the majority by for-profit companies from <sup>15</sup> and from “The SNP Fact Sheet” at [http://www.ornl.gov/sci/techresources/Human\\_Genome/faq/snps.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml).

---

<sup>15</sup> The SNP Consortium website: past, present and future

### 35.3.11 Endnote 11

The SGC is a public-private partnership whose mandate is to promote the development of new medicines by carrying out basic science of relevance to drug discovery and placing all information, reagents and know-how into the public domain without restriction. The core mandate of the SGC is to determine 3D structures on a large scale and cost-effectively-targeting human proteins of biomedical importance and proteins from human parasites that represent potential drug targets. In these two areas, the SGC is now responsible for >25% and >50% of all structures deposited into the Protein Data Bank each year. It is funded by public and private institutions, including three of the world's largest pharmaceutical companies. From the SGC FAQ at [http://www.thesgc.org/about/faqs.php#faq\\_3](http://www.thesgc.org/about/faqs.php#faq_3)

# OPEN DATA, OPEN SOURCE AND OPEN STANDARDS IN CHEMISTRY: THE BLUE OBELISK FIVE YEARS ON

## 36.1 Abstract

### 36.1.1 Background

The Blue Obelisk movement was established in 2005 as a response to the lack of Open Data, Open Standards and Open Source (ODOSOS) in chemistry. It aims to make it easier to carry out chemistry research by promoting interoperability between chemistry software, encouraging cooperation between Open Source developers, and developing community resources and Open Standards.

### 36.1.2 Results

This contribution looks back on the work carried out by the Blue Obelisk in the past 5 years and surveys progress and remaining challenges in the areas of Open Data, Open Standards, and Open Source in chemistry.

### 36.1.3 Conclusions

We show that the Blue Obelisk has been very successful in bringing together researchers and developers with common interests in ODOSOS, leading to development of many useful resources freely available to the chemistry community.

## 36.2 Background

The Blue Obelisk movement was established in 2005 at the 229<sup>th</sup> National Meeting of the American Chemistry Society as a response to the lack of Open Data, Open Standards and Open Source (ODOSOS) in chemistry. While other scientific disciplines such as physics, biology and astronomy (to name a few) were embracing new ways of doing science and reaping the benefits of community efforts, there was little if any innovation in the field of chemistry and scientific progress was actively hampered by the lack of access to data and tools. Since 2005 it has become evident that a good amount of development in open chemical information is driven by the demands of neighbouring scientific fields. In many areas in biology, for example, the importance of small molecules and their interactions and reactions in

biological systems has been realised. In fact, one of the first free and open databases and ontologies of small molecules was created as a resource about chemical structure and nomenclature by biologists<sup>1</sup>.

The formation of the Blue Obelisk group is somewhat unusual in that it is not a funded network, nor does it follow the industry consortium model. Rather it is a grassroots organisation, catalysed by an initial core of interested scientists, but with membership open to all who share one or more of the goals of the group:

**|nonascii\_1| Open Data in Chemistry.** One can obtain all scientific data in the public domain when wanted and reuse it for whatever purpose.

**|nonascii\_2| Open Standards in Chemistry.** One can find visible community mechanisms for protocols and communicating information. The mechanisms for creating and maintaining these standards cover a wide spectrum of human organisations, including various degrees of consent.

**|nonascii\_3| Open Source in Chemistry.** One can use other people's code without further permission, including changing it for one's own use and distributing it again.

Note that while some may advocate also for Open Access to publications, the Blue Obelisk goals (ODOSOS) focus more on the availability of the underlying scientific data, standards (to exchange data), and code (to reproduce results). All three of these goals stem from the fundamental tenants of the scientific method for data sharing and reproducibility.

The Blue Obelisk was first described in the CDK News<sup>2</sup> and later as a formal paper by Guha et al.<sup>3</sup> in 2006. Its home on the web is at <http://blueobelisk.org>. This contribution looks back on the work carried out by the Blue Obelisk over the past 5 years in the areas of Open Data, Open Source, and Open Standards in chemistry.

### 36.3 Scope

The Blue Obelisk covers many areas of chemistry and chemical resources used by neighbouring disciplines (*e.g.* biochemistry, materials science). Many of the efforts relate to cheminformatics (the scope of this journal) and we believe that many of the publications in Journal of Cheminformatics could be completely carried out using Blue Obelisk resources and other Open Source chemical tools. The importance of this is that for the first time it would allow reviewers, editors and readers to validate assertions in the journal and also to re-run and re-analyse parts of the calculation.

However, Blue Obelisk software and data is also used outside cheminformatics and certainly in the five main areas that, for example, Chemical Markup Language (CML)<sup>4</sup> supports:

1. **Molecules:** This is probably the largest area for Blue Obelisk software and data, and is reflected by many programs that visualise, transform, convert formats and calculate properties. It is almost certain that any file format currently in use can be processed by Blue Obelisk software and that properties can be calculated for most (organic compounds).
2. **Reactions:** Blue Obelisk software can describe the semantics of reactions and provide atom-atom matching and analyse stoichiometric balance in reactions.
3. **Computational chemistry:** Blue Obelisk software can interpret many of the current output files from calculations and create input for jobs. The Quixote project (see below and elsewhere in this issue) shows that Open Source approaches based on Blue Obelisk resources and principles are increasing the availability and re-usability of computational chemistry.
4. **Spectra:** 1-D spectra (NMR, IR, UV etc.) are fully supported in Blue Obelisk offerings for conversion and display. There is a limited amount of spectral analysis but the software gives a platform on which it should be straightforward to develop spectral annotation and manipulation. However, currently the Blue Obelisk lacks support for multi-dimensional NMR and multi-equipment spectra (*e.g.* GC-MS).

---

<sup>1</sup> Chemical Entities of Biological Interest: an update

<sup>2</sup> The Blue Obelisk

<sup>3</sup> The Blue Obelisk - Interoperability in Chemical Informatics

<sup>4</sup> Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles

**5. Crystallography:** The Blue Obelisk software supports the bi-directional processing of crystal structure files (CIF) and also solid-state calculations such as plane-waves with periodic boundary conditions. There is considerable support for the visualisation of both periodic and aperiodic condensed objects.

Many of the current operations in installing and running chemical computations and using the data are integration and customisation rather than fundamental algorithms. It is very difficult to create universal platforms that can be distributed and run by a wide range of different users, and in general, the Blue Obelisk deliberately does not address these. Our approach is to produce components that can be embedded in many environments, from stand-alone applications to web applications, databases and workflows. We believe that a chemical laboratory with reasonable access to common software engineering techniques should be able to build customised applications using Blue Obelisk components and standard infrastructure such as workflows and databases. Where the Blue Obelisk itself produces data resources they are normally done with Open components so that the community can, if necessary, replicate them.

Much of the impetus behind Blue Obelisk software is to create an environment for chemical computation (including cheminformatics) where all of the components, data, specifications, semantics, ontology and software are Openly visible and discussable. The largest current uses by the general chemical community are in authoring, visualisation and cheminformatics calculations but we anticipate that this will shortly extend into mainstream computational chemistry and solid-state. Although many of the authors are employed as research scientists, there are also several people who contribute in their spare time and we anticipate an increasing value and use of the Blue Obelisk in education at all levels.

## 36.4 Open Source

The development of Open Source software has been one of the most successful of the Blue Obelisk's activities. The following sections describe recent work in this area, and Table 1 provides an overview of the projects discussed and where to find them online.

### 36.4.1 Cheminformatics toolkits

Open Source toolkits for cheminformatics have now existed for nearly ten years. During this period, some toolkits were developed from scratch in academia, whereas others were made Open Source by releasing in-house codebases under liberal licenses. When the Blue Obelisk was established five years ago, the primary toolkits under active development were the Chemistry Development Kit (CDK)<sup>5</sup>, Open Babel<sup>6</sup>, and JOELib<sup>8</sup>. Of these, both the CDK and Open Babel continue to be actively developed.

The CDK project has been under regular development over the last five years. Several features have been implemented ranging from core components such as an extensible SMARTS matching system and a new graph (and subgraph) isomorphism method<sup>9</sup>, to more application oriented components such as 3D pharmacophore searching and matching, and a variety of structural-key and hashed fingerprints. In addition, there have been a number of second generation tools developed on top of the CDK (see below). As well as the use of the CDK in various tools, it has been deployed in the form of web services<sup>10</sup> and has formed the basis of a variety of web applications.

Since 2006, major new features of Open Babel include 3D structure generation and 2D structure-diagram generation, UFF and MMFF94 forcefields, and significantly expanded support for computational chemistry calculations. In addition, a major focus of Open Babel development has been to provide for accurate conversion and representation in areas of stereochemistry, kekulisation, and canonicalisation. The project has also grown, in terms of new contributors, new support from commercial companies, and second-generation tools applying Open Babel to a variety of end-user applications, from molecular editors to chemical database systems.

<sup>5</sup> The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics

<sup>6</sup> Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics

<sup>7</sup> Open Babel

<sup>8</sup> JOELib

<sup>9</sup> Small Molecule Subgraph Detector (SMSD) Toolkit

<sup>10</sup> A Web Service Infrastructure for Chemoinformatics

Two new Open Source cheminformatics toolkits have appeared since the original paper. In 2006 Rational Discovery, a cheminformatics service company (since closed down), released RDKit<sup>11</sup> under the BSD License. This is a C++ library with Python and (more recently) Java bindings. RDKit is actively developed and includes code donated by Novartis. Recent developments include the Java bindings, as well as performance improvements for its database cartridge.

More recently, GGA Software Services (a contract programming company) released the Indigo toolkit<sup>12</sup> and associated software in 2009 under the GPL. Indigo is a C++ library with high-level wrappers in C, Java, Python, and the .NET environment. Like RDKit and other toolkits, Indigo provides support for tetrahedral and cis-trans stereochemistry, 2D coordinate generation, exact/substructure/SMARTS matching, fingerprint generation, and canonical SMILES computation. It also provides some less common functionality, like matching tautomers and resonance substructures, enumeration of subgraphs, finding maximum common substructure of  $N$  input structures, and enumerating reaction products.

### 36.4.2 Second-generation tools

Although feature-rich and robust cheminformatics toolkits are useful in and of themselves, they can also be seen as providing a base layer on which additional tools and applications can be built. This is one of the reasons that cheminformatics toolkits are so important to the open source ‘ecosystem’; their availability lowers the barrier for the development of a ‘second generation’ of chemistry software that no longer needs to concern itself with the low-level details of manipulating chemical structures, and can focus on providing additional functionality and ease-of-use. Although a wide range of chemistry software has been built using Blue Obelisk components (see for example, the “Related Software” link on the Open Babel website,<sup>13</sup> listing over 40 projects as of this writing, or “Software using CDK” at the CDK website), in this section we focus on second-generation tools which themselves have been developed by members of the Blue Obelisk.

Bioclipse<sup>14</sup> (v2.4 released in Aug 2010) and Avogadro<sup>15</sup> (v1.0 in Oct 2009) are two examples of such software, based on the CDK and Open Babel, respectively. Bioclipse (Figure 1) is an award-winning molecular workbench for life sciences that wraps cheminformatics functionality behind user-friendly interfaces and graphical editors while Avogadro (Figure 2) is a 3D molecular editor and viewer aimed at preparing and analysing computational chemistry calculations. Both projects are designed to be extended or scripted by users through the provision of a plugin architecture and scripting support (using Bioclipse Scripting Language<sup>16</sup>, or Python in the case of Avogadro). An interesting aspect of both Avogadro and Bioclipse is that they share some developers with the underlying toolkits and this has driven the development of new features in the CDK and Open Babel.

Both products in turn act as extensible platforms for other software. Bioclipse, for example is used by software such as Brunn<sup>17</sup>, a laboratory information system for microplate based high-throughput screening. Brunn provides a graphical interface for handling different plate layouts and dilution series and can automatically generate dose response curves and calculate IC<sub>50</sub>-values. Avogadro is used by Kalzium<sup>18</sup>, a periodic table and chemical editor in KDE, and XtalOpt<sup>19</sup><sup>20</sup>, an evolutionary algorithm for crystal structure prediction. XtalOpt provides a graphical interface using Avogadro and submits calculations using a range of solid-state simulation software to predict stable polymorphs.

A final example of second-generation Blue Obelisk software is the AMBIT2<sup>21</sup><sup>22</sup> software, which was developed to facilitate registration of chemicals for the REACH EU directive, and is based on the CDK. It was distributed initially as a standalone Java Swing GUI, and more recently as downloadable web application archive, offering a web services

---

<sup>11</sup> RDKit

<sup>12</sup> Indigo

<sup>13</sup> Open Babel - Related Software

<sup>14</sup> Bioclipse: an open source workbench for chemo- and bioinformatics

<sup>15</sup> Avogadro: an open-source molecular builder and visualization tool

<sup>16</sup> Bioclipse 2: A scriptable integration platform for the life sciences

<sup>17</sup> Brunn: An open source laboratory information system for microplates with a graphical plate layout design process

<sup>18</sup> Kalzium - Periodic Table and Chemistry in KDE

<sup>19</sup> XtalOpt - Evolutionary Crystal Structure Prediction

<sup>20</sup> XtalOpt: An open-source evolutionary algorithm for crystal structure prediction

<sup>21</sup> AMBIT RESTful web services: an implementation of the OpenTox application programming interface

<sup>22</sup> Open Source Tools for Read-Across and Category Formation

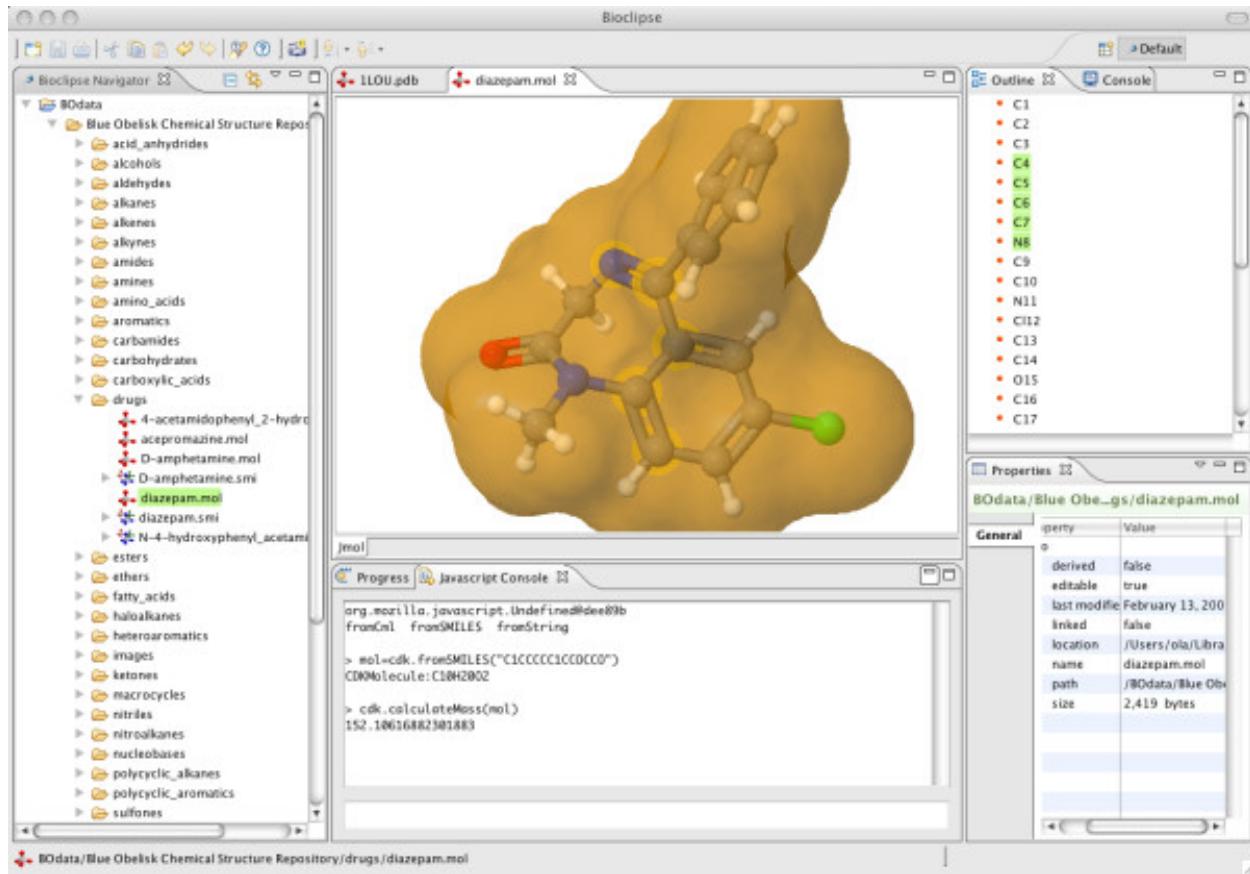


Figure 36.1: Figure 1. Screenshot of Bioclipse using Jmol to visualise a molecular surface  
Screenshot of Bioclipse using Jmol to visualise a molecular surface.

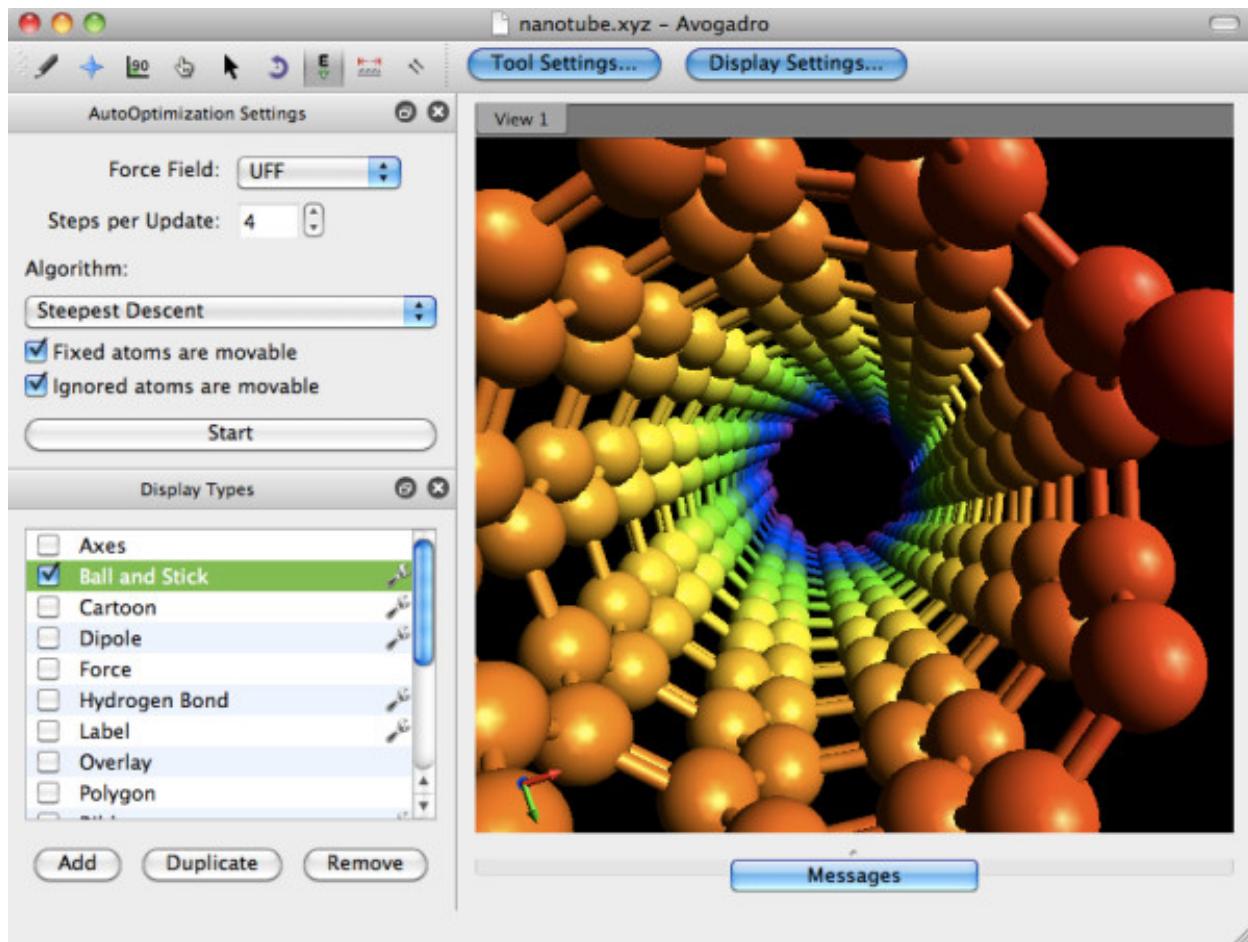


Figure 36.2: Figure 2. Screenshot of Avogadro showing a depiction of a carbon nanotube  
Screenshot of Avogadro showing a depiction of a carbon nanotube.

interface to a searchable chemical structures database. Also integrated are descriptor calculations, as well as the ability to run and build predictive models, including modules of the open source Toxtree<sup>22</sup><sup>23</sup><sup>24</sup> software for toxicity prediction.

### 36.4.3 Computational chemistry analysis

Another area where the Blue Obelisk has had a significant impact in the past five years is in supporting quantum chemistry calculations and in interpreting their results. Electronic structure calculations have a long tradition in the chemistry community and a variety of programs exist, mostly proprietary software but with an increasing number of open source codes. However, since each program uses different input formats, and the the output formats vary widely (sometimes even varying between different versions of the same software), preparing calculations and automatically extracting the results is problematic.

Avogadro has already been mentioned as a GUI for preparing calculations. It uses Open Babel to read the output of several electronic structure packages. Avogadro generates input files on the fly in response to user input on forms, as well as allowing inline editing of the files before they are saved to disk. It also features intuitive syntax highlighting for GAMESS input files, allowing expert users to easily spot mistakes before saving an input file to disk.

In addition to this, significant development of new parsing routines took place in an Avogadro plugin to read in basis sets and electronic structure output in order to calculate molecular orbital and electron density grids. This code was written to be parallel, using desktop shared memory parallelism and high level APIs in order to significantly speed up analysis. Most of this code was recently separated from the plugin, and released as a BSD licensed library, OpenQube, which is now used by the latest version of Avogadro. Jmol (see below) can also depict computational chemistry results including molecular orbitals.

In 2006, the Blue Obelisk project cclib<sup>25</sup> was established with the goal of parsing the output from computational chemistry programs and presenting it in a standard way so that further analyses could be carried out independently of the quantum package used. cclib is a Python library, and the current version (version 1.0.1) supports 8 different computational chemistry codes and extracts over 30 different calculated attributes. Two related Blue Obelisk projects build upon cclib. GaussSum<sup>26</sup>, is a GUI that can monitors the progress of SCF and geometry convergences, and can plot predicted UV/Vis absorption and infrared spectra from appropriate logfiles containing energies and oscillator strengths for easy comparison to experimental data. QMForge<sup>27</sup> provides a GUI for various electronic structure analyses such as Frenking's charge decomposition analysis<sup>28</sup> and Mulliken or C-squared analyses on user-defined molecular fragments. QMForge also provides a rudimentary Cartesian coordinate editor allowing molecular structures to be saved via Open Babel.

The Quixote project epitomises the full use of the Blue Obelisk software and is described in detail in another article in this issue. Here we observe that it is possible to convert legacy chemistry file formats of all sorts into semantic chemistry and extract those parts which are suitable for input to computational chemistry programs. This chemistry is then combined with generic concepts of computational chemistry (*e.g.* strategy, machine resources, timing, accuracy etc.) into the legacy inputs for a wide range of programs. Quixote itself follows Blue Obelisk principles in that it does not manage the submission and monitoring of jobs but resumes action when the jobs have been completed, and then applies a range of parsing and transformation tools to create standardised semantic chemical content. A major feature of Quixote is that it requires all concepts to validate against dictionaries and the process of parsing files necessarily generates communally-agreed dictionaries, which represent an important step forward in the Open specifications for Blue Obelisk. When widely-deployed, Quixote will advertise the value of Open community standards for semantics to the world.

The Quixote project is not dependent on any particular technology, other than the representation of computational chemistry in CML and the management of semantics through CML dictionaries. At present, we use JUMBO-

<sup>23</sup> ToxTree

<sup>24</sup> An evaluation of the implementation of the Cramer classification scheme in the Toxtree software

<sup>25</sup> cclib: A library for package-independent computational chemistry algorithms

<sup>26</sup> GaussSum

<sup>27</sup> QMForge

<sup>28</sup> Investigation of Donor-Acceptor Interactions: A Charge Decomposition Analysis Using Fragment Molecular Orbitals

Converters<sup>29</sup> for most of the semantic conversion, Lensfield2<sup>30</sup> for the workflow and Chempound (chem#)<sup>31</sup> to store and disseminate the results.

#### 36.4.4 Web applications

While desktop software has composed the majority of scientific tools since the computer was introduced, the internet continues to change how applications and content are distributed and presented. The web presents new opportunities for scientists as it is an open and free medium to distribute scientific knowledge, ideas and education. Web applications are software that runs within the browser, typically implemented in Java or JavaScript. Recently, a new version of the HTML specification, HTML5, defined a well-developed framework for creating native web applications in JavaScript and this opens up new possibilities for visualising chemical data.

Jmol, the interactive 3D molecular viewer, is one of the most widely used chemistry applets, and indeed has seen widespread use in other fields such as biology and even mathematics (it is used for 3D depiction of mathematical functions in the Sage Mathematics Projects<sup>32</sup>). It is implemented in Java, and has gone from being a “Rasmol/Chime” replacement to a fully fledged molecular visualisation package, including full support for crystallography<sup>33</sup>, display of molecular orbitals from standard basis set/coefficient data, the inclusion of dynamic minimisation using the UFF force field, and a full implementation of Daylight SMILES and SMARTS, with extensions to conformational and biomolecular substructure searching (Jmol BioSMARTS).

In 2009, iChemLabs released the ChemDoodle Web Components library<sup>34</sup> under the GPL v3 license (with a liberal HTML exception). This library is completely implemented in JavaScript and uses HTML5 to allow the scientist to present publication quality 2D and 3D graphics (see Figure 3) and animations for chemical structures, reactions and spectra. Beyond graphics, this tool provides a framework for user interaction to create dynamic applications through web browsers, desktop platforms and mobile devices such as the iPhone, iPad and Android devices.

#### 36.4.5 The business end

Open Source provides a unique opportunity for commercial organisations to work with the cheminformatics community. Traditional business models rely on monetisation of source code, causing companies to repeat work done by other companies. This model is sometimes combined with a free (gratis) model for people working at academic institutes, to increase adoption and encourage contributions from academics. This solution defines the return on investment as the IP on the software, but has the downside of investment losses due to duplication of software and method development, which become visible when proprietary companies merge. Some authors have argued that in the chemistry field few contributors are available to volunteer time to improve codes and IP considerations may prevent contributions from industry<sup>35</sup>. If true, this would hamper adoption of Open Source and Open Data in chemistry, and greatly slow the growth of projects such as those in the Blue Obelisk.

The Blue Obelisk community, however, takes advantage of the fact that much of the investment needed for development is either paid for by academic institutes and funding schemes, or by volunteers investing time and effort. In return, contributors get full access to the source code, and the Open Source licensing ensures that they will have access any time in the future. In this way, the license functions as a social contract between everyone to arrange an immediate return on investment. Effectively, this approach shares the burden of the high investment in having to develop cheminformatics software from scratch, allowing researchers and commercial partners alike to focus on their core business, rather than the development of prerequisites. In the case of the Blue Obelisk, the rich collection of Open

---

<sup>29</sup> JUMBO-Converters

<sup>30</sup> Lensfield 2

<sup>31</sup> Chempound

<sup>32</sup> NOTITLE!

<sup>33</sup> Jmol - a paradigm shift in crystallographic visualization

<sup>34</sup> ChemDoodle Web Components: HTML5 Chemistry

<sup>35</sup> Open-source software: not quite endsiville

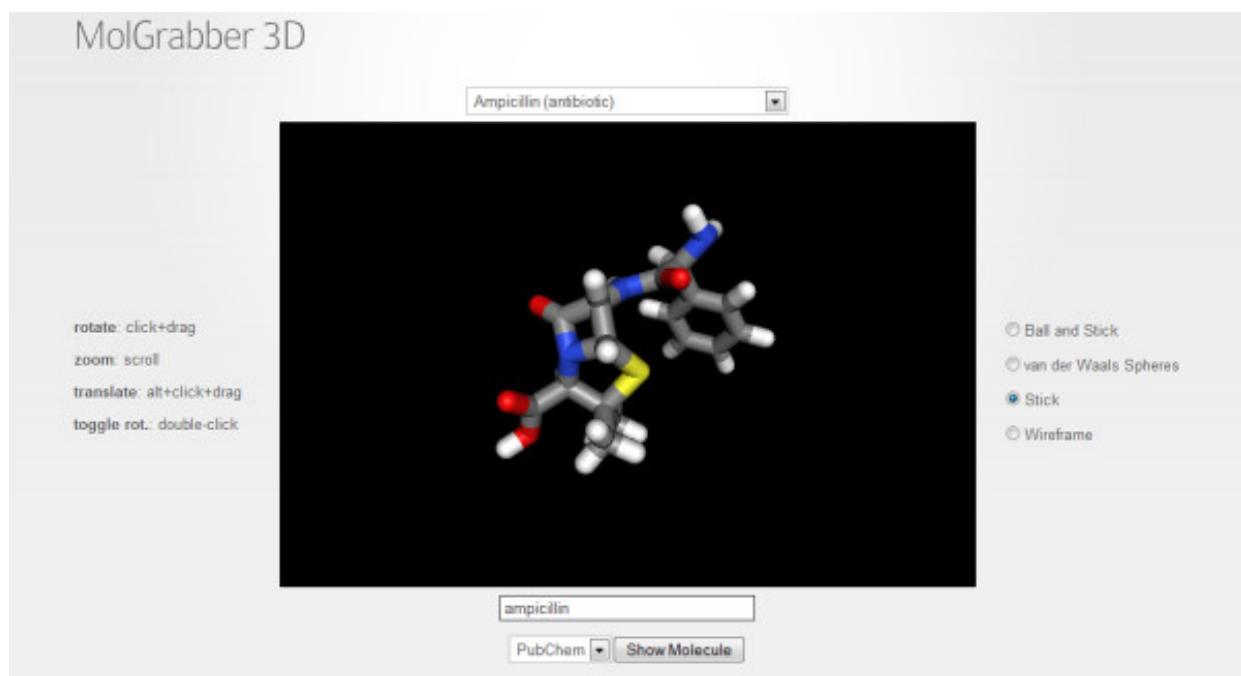


Figure 36.3: Figure 3. Screenshot of the MolGrabber 3D demo from ChemDoodle Web Components  
**Screenshot of the MolGrabber 3D demo from ChemDoodle Web Components.**

Source cheminformatics tools provided greatly reduces investment up front for new companies in the cheminformatics market. Such advantages have also been noted in the drug discovery field <sup>363738</sup>.

The use of Open Standards allows everyone to select those Blue Obelisk components they find most useful, as they can easily replace one component with another providing the same functionality, taking advantage that they use the same standards for, for example, data exchange. This way, licensing issues are becoming a marginal problem, allowing companies to select a license appropriate for their business model. This too, allows a company to create a successful product with significantly reduced cost and effort.

At the time of writing there are many commercial companies developing chemistry solutions around Open Source cheminformatics components provided by the Blue Obelisk community. Examples of such companies include iChemLabs, IdeaConsult, Wingu, Silicos, GenettaSoft, eMolecules, hBar, Metamolecular, and Inkspot Science. Some of these merely use components, but several actively contribute back to the Blue Obelisk project they use, or donate new Open Source cheminformatics projects to the community.

For example, iChemLabs released the ChemDoodle Web Components library under the GPL v3 license, based on the upcoming HTML5 Open Standard. It allows making web and mobile interfaces for chemical content. The project is already being adopted by others, including iBabel <sup>39</sup>, ChemSpotlight <sup>40</sup> and the RSC ChemSpider <sup>4142</sup>.

Silicos has released several Open Source utilities <sup>43</sup> based on Open Babel, such as Pharao, a tool for pharmacophore searching, Sieve for filtering molecular structure by molecular property, Stripper for removing core scaffold structures from a molecule set, and Piramid for molecular alignment using shape determined by the Gaussian volumes as a descriptor. Additionally, contributions have been made to the Open Babel project itself.

<sup>36</sup> The case for open-source software in drug discovery

<sup>37</sup> Can open-source R&D reinvigorate drug research?

<sup>38</sup> Optimizing the use of open-source software applications in drug discovery

<sup>39</sup> iBabel

<sup>40</sup> ChemSpotlight

<sup>41</sup> ChemSpider - the free chemical database

<sup>42</sup> iChemLabs and RSC ChemSpider announce partnership

<sup>43</sup> Silicos Open Source Software

Other companies use Blue Obelisk components and contribute patches, smaller and larger. For example, IXELIS donated the isomorphism code in the CDK, eMolecules donated canonicalisation code to Open Babel, Metamolecular improved the extensibility and unit testing suite of OPSIN, and AstraZeneca contributed code to the CDK for signatures. This is just a very minor selection, and the reader is encouraged to contact the individual Blue Obelisk projects for a detailed list.

In May 2011, a Wellcome Trust Workshop on Molecular Informatics Open Source Software (MIOSS) explored the role of Open Source in industrial laboratories and companies as well as academia (several of the presenters are among the authors of this paper). The meeting identified that Open Source software was extremely valuable to industry not just because it is available for free, but because it allows the validation of source code, data and computational procedures. Some of the discussion was on business models or other ways to maintain development of Open Source software on which a business relied. Companies are concerned about training and support and, in some cases, product liability. There are difficulties for software for which there is no formal transaction other than downloading and agreeing to license terms. One anecdote concerned a company that wished to donate money to an Open Source project but could not find a mechanism to do so.

Industry participants also pointed out that there is a considerable amount of contribution-in-kind from industry, both from enhancements to software and also the development of completely new software and toolkits. Companies are now finding it easier to create mechanisms for releasing Open Source software without violating confidentiality or incurring liability. A phrase from the meeting summed it up: “The ice is beginning to melt”, signifying that we can expect a rapid increase in industry’s interest in Open Source.

### 36.4.6 Converting chemical names and images to structures

The majority of chemical information is not stored in machine-readable formats, but rather as chemical names or depictions. The OSRA and OPSIN projects focus on extracting chemical information from these sources. Such software plays a particularly important role for data mining the chemical literature, including patents and theses.

Optical Structure Recognition Application (OSRA)<sup>44</sup> was started in early 2007 with the goal to create the first free and open source tool for extraction and conversion of molecular images into SMILES and SD files. From the very beginning the underlying philosophy was to integrate existing open source libraries and to avoid “reinventing the wheel” wherever possible. OSRA relies on a variety of open source components: Open Babel for chemical format conversion and molecular property calculations, GraphicsMagick for image manipulation, Potrace for vectorisation, GOCR and OCRAD for optical character recognition. The growing importance of image recognition technology can be seen in the fact that only a few years ago there was only one widely available software package for chemical structure recognition - CLiDE (commercially developed at Keymodule, Ltd), but today there are as many as seven available programs.

OPSIN (Open Parser for Systematic IUPAC Nomenclature)<sup>45</sup> focuses instead on interpreting chemical names. The chemical name is the oldest form of communication used to describe chemicals, predating even the knowledge of the atomic structure of compounds. Chemical names are abundant in the scientific literature and encode valuable structural information. Through successive books of recommendations<sup>4647</sup>, IUPAC has tried to codify and to an extent standardise naming practices. OPSIN aims to make this abundance of chemical names machine readable by translating them to SMILES, CML or InChI. The program is based around the use of a regular grammar to guide tokenisation and parsing of chemical names, followed by step-wise application of nomenclature rules. It is able to offer fast and precise conversions for the majority of names using IUPAC organic nomenclature, and is available as a web service, Java library and standalone application for maximum interoperability.

---

<sup>44</sup> OSRA

<sup>45</sup> Chemical Name to Structure: OPSIN, an Open Source Solution

<sup>46</sup> NOTITLE!

<sup>47</sup> NOTITLE!

### 36.4.7 Chemical database software

Registration, indexing and searching of chemical structures in relational databases is one of the core areas of cheminformatics. A number of structure registration systems have been published in the last five years, exploiting the fact that Open Source cheminformatics toolkits such as Open Babel and the CDK are available. OrChem<sup>48</sup>, for example, is an open source extension for the Oracle 11G database that adds registration and indexing of chemical structures to support fast substructure and similarity searching. The cheminformatics functionality is provided by the CDK. OrChem provides similarity searching with response times in the order of seconds for databases with millions of compounds, depending on a given similarity cut-off. For substructure searching, it can make use of multiple processor cores on today's powerful database servers to provide fast response times in equally large data sets.

Besides the traditional and proven relational database approach with added chemical features ('cartridges'), there is growing interest in tools and approaches based on the web philosophy and practice. Several groups<sup>49 50</sup> are experimenting with the Resource Description Framework (RDF) language on the assumption that generic high-performance solutions will appear. RDF allows everything to be described by URIs (data, molecules, dictionaries, relations). The Chempound system<sup>31</sup>, as deployed in Quixote and elsewhere, is an RDF-based approach to chemical structures and compounds and their properties. For small to medium-sized collections (such as an individual's calculations or literature retrieval), there are many RDF tools (e.g. SIMILE, Apache Jena) which can operate in machine memory and provide the flexibility that RDF offers. For larger systems, it is unclear whether complete RDF solutions (e.g. Virtuoso) will be satisfactory or whether a hybrid system based on name-value pairs (e.g. CouchDB, MongoDB) will be sufficient.

### 36.4.8 Collaboration and interoperability

One of the successes of the Blue Obelisk has been to bring developers together from different Open Source chemistry projects so that they look for opportunities to collaborate rather than compete, and to leverage work done by other projects to avoid duplication of effort. As an example of this, when in March 2008 the Jmol development team were looking to add support for energy minimisation, rather than implement a forcefield from scratch they ported the UFF forcefield<sup>51</sup> implementation from Open Babel to Jmol. This code enables Jmol to support 2D to 3D conversion of structures (through energy minimisation). In a similar manner, efficient Jmol code for atom-atom rebonding has been ported to the CDK. Figure 4 shows the collaborative nature of software developed in the Blue Obelisk, as one project builds on functionality provided by another project.

Another collaborative initiative between Blue Obelisk projects was the establishment in May 2008 of the ChemiSQL project. This brought together the developers of several open source chemistry database cartridges (PgChem<sup>52</sup>, Mychem<sup>53</sup>, OrChem<sup>48</sup> and more recently Bingo<sup>54</sup>) with a view to making their database APIs more similar and collaborating on benchmark datasets for assessing performance. For two of these projects, PgChem and Mychem, which are both based on Open Babel, there is the additional possibility of working together on a shared codebase.

In the area of cheminformatics toolkits, two of the existing toolkits Open Babel and RDKit are planning to work together on a common underlying framework called MolCore<sup>55</sup>. This project is still in the planning stage, but if it is a success it will mean that the two libraries will be interoperable (while retaining their existing focus) but also that the cost of maintaining the code will be shared among more developers, freeing time for the development of new features.

One of the goals of the Blue Obelisk is to promote interoperability in chemical informatics. When barriers exist to moving chemical data between different software, the community becomes fragmented and there is the danger of vendor lock-in (where users are constrained to using a particular software, a situation which puts them at a disadvantage).

<sup>48</sup> OrChem - An open source chemistry search engine for Oracle(R)

<sup>49</sup> Resource description framework technologies in chemistry

<sup>50</sup> Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data

<sup>51</sup> UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations

<sup>52</sup> PgChem

<sup>53</sup> Mychem

<sup>54</sup> Bingo

<sup>55</sup> MolCore

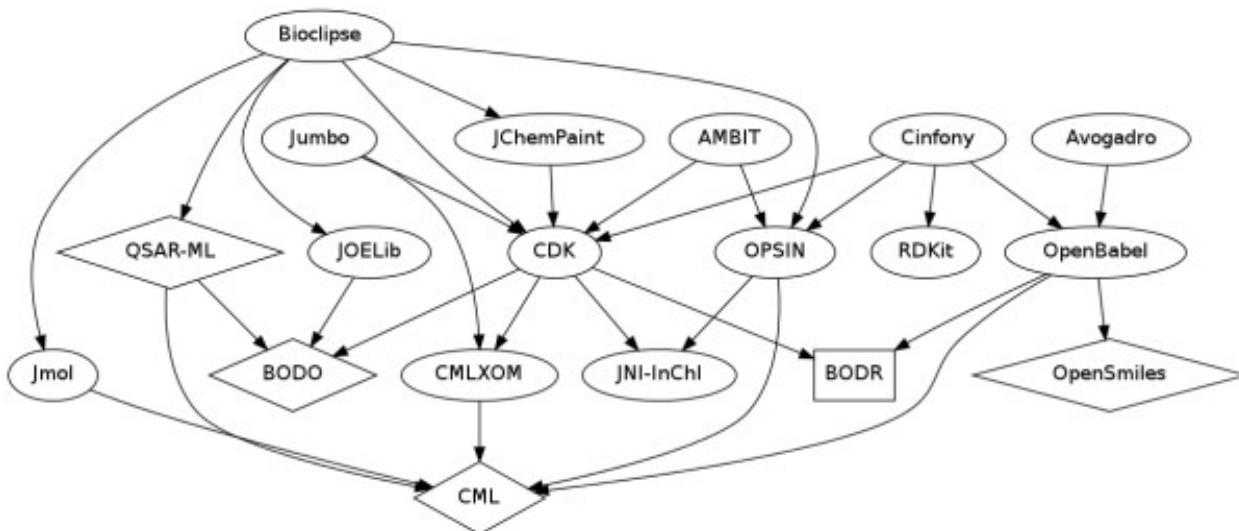


Figure 36.4: Figure 4. Dependency diagram of some Blue Obelisk projects

**Dependency diagram of some Blue Obelisk projects.** Each block represents a project. Square blocks show Open Data, ovals are Open Source, and diamonds are Open Standards.

This applies as much to Open Source software as to proprietary software. Cinfony is a project (first release in May 2008) whose goal is to tackle this problem in the area of cheminformatics toolkits<sup>56</sup>. It is a Python library that enables Open Babel, the CDK, and RDKit (and shortly, Indigo and OPSIN) to be used using the same API; this makes it easy, for example, to read a molecule using Open Babel, calculate descriptors using the CDK and create a depiction using RDKit.

Another way through which interoperability of Blue Obelisk projects has been promoted and developed is through integration into workflow software such as Taverna<sup>57</sup> and KNIME<sup>58</sup> (both open source). Such software makes it easy to automate recurring tasks, and to combine analyses or data from a variety of different software and web services. A combination of the Chemistry Development Kit and Taverna, for instance, was reported in 2010<sup>59</sup>. In the case of KNIME, it comes with built-in basic collection of CDK-based and Open Babel-based nodes, while other nodes for the RDKit and Indigo are available from KNIME’s “Community Updates” site.

## 36.5 Open Standards

### 36.5.1 Chemical Markup Language, CML

Chemical Markup Language (CML) is discussed in several articles in this issue, and a brief summary here re-iterates that it is designed primarily to create a validatable semantic representation for chemical objects. The five main areas (molecules, reactions, computational chemistry, spectra and solid-state (see above)) have now all been extensively deployed and tested. CML can therefore be used as a reference for input and output for Blue Obelisk software and a means of representing data in Blue Obelisk resources.

CML, being an XML application, can inter-operate with other markup languages and in particular XHTML, SVG, MathML, docx and more specialised applications such as UnitsML and GML (geosciences). We believe that it would be possible using these languages to encode large parts of, say, first year chemistry text books in XML. Similarly, it is possible to create compound documents with word processing or spreadsheet software that have inter-operating text,

<sup>56</sup> Cinfony-combining Open Source cheminformatics toolkits behind a common interface

<sup>57</sup> Taverna: a tool for building and running workflows of services

<sup>58</sup> KNIME

<sup>59</sup> CDK-Taverna: an open workflow environment for cheminformatics

graphics and chemistry (as in Chem4Word). Being a markup language, CML is designed for re-purposing, including styling, and therefore a mixture of these languages can be used for chemical catalogues, general publications, logbooks and many other types of document in the scientific process.

CML describes much of its semantics through conventions and dictionaries, and the emerging ecosystem (especially in computational chemistry) is available as a semantic resource for many of the applications and specifications in this article.

### 36.5.2 InChI

The IUPAC InChI identifier is a non-proprietary and unique identifier for chemical substances designed to enable linking of diverse data compilations. Prior to the development of the InChI identifier chemical information systems and databases used a wide variety of (generally proprietary) identifiers, greatly limiting their interoperability. Although its development predates the Blue Obelisk, software such as Open Babel has included InChI support since 2005, and support for InChI in Indigo is due in 2011.

Since the official InChI implementation is in C, it is difficult to access from the other widely used language for cheminformatics toolkits, Java. Early attempts to generate InChI identifiers from within Java involved programmatically launching the InChI executable and capturing the output, an approach that was found to be fairly unreliable and broke the ‘write once, run anywhere’ philosophy of Java. The Blue Obelisk project JNI-InChI<sup>60</sup> was established in 2006 to solve this problem by using the Java Native Interface framework to provide transparent access to the InChI library from within Java and other Java Virtual Machine (JVM) based languages, supporting the wider adoption of this standard identifier by the chemistry community.

The Java Native Interface framework provides a mechanism for code running inside the JVM, to place calls to libraries written in languages such as C, C++ and Fortran, and compiled into native, machine specific, code. JNI-InChI provides a thin C wrapper, with corresponding Java code, around the IUPAC InChI library, exposing the InChI library’s functionality to the JVM. To overcome the need to have the correct InChI library pre-installed on a system, JNI-InChI comes with a variety of precompiled native binaries and automatically extracts and deploys the correct one for the detected operating system and architecture. The JNI-InChI library comes with native binaries supporting a range of operating systems and architectures; the current version has binaries for 32- and 64-bit Windows, Linux and Solaris, 64-bit FreeBSD and 64-bit Intel-based Mac OS X - a number of which are not supported by the original IUPAC distribution of InChI. The JNI-InChI project has matured to support the full range of functionality of the InChI C library: structure-to-InChI, InChI-to-structure, AuxInfo-to-structure, InChIKey generation, and InChI and InChIKey validation. JNI-InChI provides the InChI functionality for a number of Open Source projects, including the Chemistry Development Kit, Bioclipse and CMLXOM/JUMBO, and is also used by commercial applications and internally in a number of companies. Through its widespread use and Open Source development model, a number of issues in earlier versions of the software have been identified and resolved, and JNI-InChI now offers a robust tool for working with InChIs in the JVM.

### 36.5.3 OpenSMILES

One of the most widely used ways to store chemical structures is the SMILES format (or SMILES string). This is a linear notation developed by Daylight Information Systems that describes the connection table of a molecule and may optionally encode chirality. Its popularity stems from the fact that it is a compact representation of the chemical structure that is human readable and writable, and is convenient to manipulate (e.g. to include in spreadsheets, or copy from a web page).

Despite its widespread use, a formal definition of the language did not exist beyond Daylight’s SMILES Theory Manual and tutorials. This caused some confusion in the implementation and interpretation of corner cases, for example the handling of cis/trans bond symbols at ring closures. In 2007, Craig James (eMolecules) initiated work on the OpenSMILES specification<sup>61</sup>, a complete specification of the SMILES language as an Open Standard developed

<sup>60</sup> JNI-InChI

<sup>61</sup> The OpenSMILES specification

through a community process. The specification is largely complete and contains guidelines on reading SMILES, a formal grammar, recommendations on standard forms when writing SMILES, as well as proposed extensions.

### 36.5.4 QSAR-ML

The field of QSAR has long been hampered by the lack of open standards, which makes it difficult to share and reproduce descriptor calculations and analyses. QSAR-ML was recently proposed as an open standard for exchanging QSAR datasets<sup>62</sup>. A dataset in QSAR-ML includes the chemical structures (preferably described in CML) with InChI to protect integrity, chemical descriptors linked to the Blue Obelisk Descriptor Ontology<sup>63</sup>, response values, units, and versioned descriptor implementations to allow descriptors from different software to be integrated into the same calculation. Hence, a dataset described in QSAR-ML is completely reproducible. To allow for easy setup of QSAR-ML compliant datasets, a plugin for Bioclipse was created with a graphical interface for setting up QSAR datasets and performing calculations. Descriptor implementations are available from the CDK and JOELib, as well as via remote web services such as XMPP<sup>64</sup>.

### 36.5.5 Remaining challenges

A core requirement for chemical structure databases and chemical registration systems in general is the notion of structure standardisation. That is, for a given input structure, multiple representations should be converted to one canonical form. Structure canonicalisation routines partially address this aspect, converting multiple alternative topologies to a single canonical form. However, the problem of standardisation is broader than just topological canonicalisation. Features that must be considered include

- topological canonicalisation
- handling of charges
- tautomer enumeration and canonicalisation
- normalisation of functional groups

Currently, most of the individual components of a ‘standardisation pipeline’ can be implemented using Blue Obelisk tools. The larger problem is that there is no agreed upon list of steps for a standardisation process. While some specifications have been published (e.g., PubChem) and some standardisation services and tools are available (for example, PubChem provides an online service to standardise molecules<sup>65</sup>) each group has their own set of rules. A common reference specification for standardisation would be of immense value in interoperability between structure repositories as well as between toolkits (though the latter is still confounded by differences in lower level cheminformatic features such as aromaticity models).

We have already discussed the development of an Open SMILES standard. While much progress has been made towards a complete specification, more remains to be done before this can be considered finished. After that point, the next logical step would be to start work on a standard for the SMARTS language, the extension to SMILES that specifies patterns that match chemical substructures.

## 36.6 Open Data

A considerable stumbling block in advocating the release of scientific data as Open Data has been how exactly to define “Open.” A major step forward was the launch in 2010 of the Panton Principles for Open Data in Science<sup>66</sup>. This formalises the idea that Open Data maximises the possibility of reuse and repurposing, the fundamental basis of

---

<sup>62</sup> Towards interoperable and reproducible QSAR analyses: Exchange of datasets

<sup>63</sup> The Blue Obelisk Descriptor Ontology

<sup>64</sup> XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous web services

<sup>65</sup> PubChem Standardization Service

<sup>66</sup> Panton Principles - Principles for Open Data in Science

how science works. These principles recommend that published data be licensed explicitly, and preferably under CC0 (Creative Commons ‘No Rights Reserved’, also known as CCZero) <sup>67</sup>. This license allows others to use the data for any purpose whatsoever without any barriers. Other licenses compatible with the Panton Principles include the Open Data Commons Public Domain Dedication and Licence (PDDL), the Open Data Commons Attribution License, and the Open Data Commons Open Database License (ODbL) <sup>68</sup>.

Despite this positive news, little chemical data compatible with these principles has become available from the traditional chemistry fields of organic, inorganic, and solid state chemistry. Table 2 lists a few notable exceptions, some of which are discussed further below. There is also data available using licenses not compatible with the Panton Principles, but where the user is allowed to modify and redistribute the data. A new data set in this category is the data from the ChEMBL database <sup>69</sup>, which is available under the Creative Commons Share-Alike Attribution license. The RSC ChemSpider database <sup>41</sup>, although not fully Open, also hosts Open Data; for example, spectral data when deposited can be marked as Open.

Importantly, publishing data as CC0 is becoming easier now that websites are becoming available to simplify publishing data. Two projects that can be mentioned in this context are FigShare <sup>70</sup>, where the data behind unpublished figures can be hosted, and Dryad <sup>71</sup> where data behind publications can be hosted. Initiatives like this make it possible to host small amounts of data, and those combined are expected to become soon a substantial knowledge base.

### 36.6.1 Reaction Attempts

Although there are existing databases that allow for searching reactions, those using Open Data are harder to find. The Reaction Attempts database <sup>72</sup>, to which anyone can submit reaction attempts data, consists mainly of reaction information abstracted from Open Notebooks in organic chemistry, such as the UsefulChem project from the Bradley group <sup>73</sup> and the notebooks from the Todd group <sup>74</sup>. Key information from each experiment is abstracted manually, with the only required information consisting of the ChemSpider IDs of the reactants and the product targeted in the experiment; and a link to the laboratory notebook page. Information in the database can be searched and accessed using the web-based Reaction Attempts Explorer <sup>75</sup>.

Since the database reflects all data from the notebooks, it includes experiments in progress, ambiguous results and failed runs. Unlike most reaction databases that only identify experiments successfully reported in the literature, the Reaction Attempts Explorer allows researchers to easily find patterns in reactions that have already been performed, and since the data are open and results are reported across all research groups, intersections are easily discovered and possible Open Collaboration opportunities are easily found <sup>7677</sup>.

### 36.6.2 Non-Aqueous Solubility

Although the aqueous solubility of many common organic compounds is generally available, quantitative reports of non-aqueous solubility are more difficult to find. Such information can be valuable for selecting solvents for reactions, re-crystallization and related processes. In 2008, the Open Notebook Science Solubility Challenge was launched for the purpose of measuring non-aqueous solubility of organic compounds, reporting all the details of the experiments in an Open Notebook and recording the results as Open Data in a centralized database <sup>7879</sup>. This crowdsourcing project

<sup>67</sup> About CC0 - “No Rights Reserved”

<sup>68</sup> Open Licenses - Data

<sup>69</sup> ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr

<sup>70</sup> FigShare

<sup>71</sup> Dryad

<sup>72</sup> Reaction Attempts Database

<sup>73</sup> Useful Chemistry: Reaction Attempts Book Edition 1 and UsefulChem Archive

<sup>74</sup> Useful Chemistry: The Synaptic Leap Experiments on Reaction Attempts

<sup>75</sup> Useful Chemistry: Reaction Attempts Explorer

<sup>76</sup> Useful Chemistry: Visualizing Social Networks in Open Notebooks

<sup>77</sup> Collaboration using Open Notebook Science in Academia

<sup>78</sup> Open Notebook Science Challenge: Solubilities of Organic Compounds in Organic Solvents

<sup>79</sup> Beautifying Data in the Real World

was also supported by Submeta, Sigma-Aldrich, Nature Publishing Group and the Royal Society of Chemistry. The database currently holds 1932 total measurements and 1428 averaged solute/solvent measurements all of which are available under a CC0 license. Several web services and feeds are available to filter and re-use the dataset<sup>80</sup>. In particular, models have been developed for the prediction of non-aqueous solubility in 72 different solvents<sup>81</sup> using the method of Abraham et al<sup>82</sup> with descriptors calculated by the Chemistry Development Kit. These models are available online and will be refined as more solubility data is collected.

### 36.6.3 The Blue Obelisk Data Repository (BODR)

The Blue Obelisk has created a repository of key chemical data in a machine-readable format<sup>83</sup>. The BODR focuses on data that is commonly required for chemistry software, and where there is a need to ensure that values are standard between codes. Examples are atomic masses and conversions between physical constants. These data can be used by others for any purpose (for example, for entry into Wikipedia or use in in-house software), and should lead to an enhancement in the quality of community reference data. The Blue Obelisk provides also a complementary project, the Chemical Structure Repository<sup>83</sup>. It aims to provide 3D coordinates, InChIs and several physico-chemical descriptors for a set of 570 organic compounds.

### 36.6.4 NMRShiftDB

NMRShiftDB<sup>8485</sup> represents one of the earliest resources for Open community-contributed data (first released in 2003). Research groups that measure NMR spectra or extract it from the literature can contribute that information to NMRShiftDB which provides an Open resource where entries can be searched by chemical structure or properties (especially peaks). Although it is difficult to encourage large amounts of altruistic contribution (as happens with Wikipedia), an alternative possible source of data could come from linking data capture with data publication. For example, the Blue Obelisk has enough software that it is possible to create a seamless chain for converting NMR structures in-house into NMRShiftDB entries. If and when the chemistry community encourages or requires semantic publication of spectra rather than PDFs, it would be possible to populate NMRShiftDB rapidly along the lines of CrystalEye (see below). A similar approach has been demonstrated earlier using the Blue Obelisk components Oscar and Bioclipse using text mining approaches<sup>86</sup>.

### 36.6.5 CrystalEye

CrystalEye<sup>87</sup> is an example of cost-effective extraction of data from the literature where this is published both Openly and semantically. Software extracts Openly-published crystal structures from a variety of scholarly journals, processes them and then makes them available through a web interface. It currently contains about 250,000 structures. CrystalEye serves as a model for a high-value, high-quality Open data resource, including the licensing of each component as Panton-compatible Open data.

## 36.7 Other areas of activity

While each Blue Obelisk project has its own website and point of contact (typically a mailing list), because of the breadth of Blue Obelisk projects it can be difficult for a newcomer to understand which of them, if any, can best

<sup>80</sup> Open Notebook Solubility Web Services

<sup>81</sup> Useful Chemistry: General Transparent Solubility Prediction using Abraham Descriptors

<sup>82</sup> Prediction of solubility of drugs and other compounds in organic solvents

<sup>83</sup> Blue Obelisk Data Repository

<sup>84</sup> NMRShiftDB

<sup>85</sup> NMRShiftDB - compound identification and structure elucidation support through a free community-build web database

<sup>86</sup> Chemical Archeology: OSCAR3 to NMRShiftDB.org

<sup>87</sup> CrystalEye

address a particular problem. To address this issue, members of the Blue Obelisk established a Question & Answer website<sup>88</sup> (see Figure 5). This is a website in the style of Stack Overflow<sup>89</sup> that encourages high quality answers (and questions) through the use of a voting system. In the year since it was established, over 200 users have registered, many of whom had no previous involvement with the Blue Obelisk, showing that the Q&A website complements earlier existing channels of communication.

**Recent questions (235)**

what's your question? be descriptive.

Votes	Answers	Views	Tags	Asked By	Time Ago
3	3	19	structure software sdf tool	rich apedaca	434 • 7 about 5 hours ago
1	1	20	Annotated Table of Atomic Radii and Bond Lengths for Use in 3D Molecular Viewer?	steve wathen	384 • 1 • 7 1 day ago
1	0	36	Search for large Blood Brain Barrier dataset	chem-bla-ics	1410 • 2 • 11 24 minutes ago
1	1	86	Making grid file and problem for opening rigid and flexible file in autodock	kirtandave7	17 • 2 5 days ago
-	-	-	What Open Source tools for Aqueous Solubility Prediction are available?	-	-

**Blue Obelisk eXchange**

The Blue Obelisk Exchange is the place to ask about the use and development of Open Data, Open Source, and Open Standards: how to perform tasks and solve chemical problems with these, or if an ODOOS tool is available for some task. Or even to ask if someone can provide such a tool. The questions do not require to be about Blue Obelisk solutions itself, they can be about any ODOOS chemistry tool, service, or database.

**Top 5 users**

User	Reputation
chem-bla-ics	1410
baoilleach	1027
mattie floris	687
tony27587	603

Figure 36.5: Figure 5. Screenshot of the Blue Obelisk eXchange Question and Answer website  
Screenshot of the Blue Obelisk eXchange Question and Answer website.

The rise of self-publishing and print-on-demand services has meant that publishing a book is now as straightforward as uploading to an appropriate website. Unlike the traditional publishing route where books with projected low sales volume would be expensive, websites such as Lulu<sup>90</sup> allow the sale of low-priced books on chemistry software, and books are now available for purchase on Jmol<sup>91</sup>, the Chemistry Development Kit<sup>92</sup> and Open Babel<sup>93</sup>.

## 36.8 Conclusions

We have shown that the Blue Obelisk has been very successful in bringing together researchers and developers with common interests in ODOOS, leading to development of many useful resources freely available to the chemistry community. However, how best to engage with the wider chemistry community outside of the Blue Obelisk remains

<sup>88</sup> Blue Obelisk Q&A

<sup>89</sup> Stack Overflow

<sup>90</sup> Lulu

<sup>91</sup> NOTITLE!

<sup>92</sup> NOTITLE!

<sup>93</sup> NOTITLE!

an open question. If the Blue Obelisk is truly to make an impact, then an attempt must be made to reach beyond the subscribers to the Blue Obelisk mailing list and blogs of members.

We hope to see this involvement between the Blue Obelisk and the wider community grow in the future. To this end, we encourage the reader to visit the Blue Obelisk website <sup>94</sup>, send a message to our mailing list, investigate related projects or read our blogs.

## **36.9 Competing interests**

The authors declare that they have no competing interests.

## **36.10 Authors' contributions**

The overall layout of the manuscript grew from discussions between NMOB, RG and ELW. The authorship of the paper is drawn from those people connected with fully Open Data/Standards/Source (OSI-compliant or OKF-compliant) projects associated with the Blue Obelisk. There are a large number of people contributing to these projects and because those projects are published in their own right it is not appropriate to include all their developers by default. We invited a number of ‘project gurus’ who have been active in promoting the Blue Obelisk, to be authors on this paper and most have accepted and contributed.

## **36.11 Acknowledgements**

NMOB is supported by a Health Research Board Career Development Fellowship (PD/2009/13). The OSRA project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views of the policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organisations imply endorsement by the US Government.

---

<sup>94</sup> Blue Obelisk web site

# THE QUIXOTE PROJECT: COLLABORATIVE AND OPEN QUANTUM CHEMISTRY DATA MANAGEMENT IN THE INTERNET AGE

## 37.1 Abstract

Computational Quantum Chemistry has developed into a powerful, efficient, reliable and increasingly routine tool for exploring the structure and properties of small to medium sized molecules. Many thousands of calculations are performed every day, some offering results which approach experimental accuracy. However, in contrast to other disciplines, such as crystallography, or bioinformatics, where standard formats and well-known, unified databases exist, this QC data is generally destined to remain locally held in files which are not designed to be machine-readable. Only a very small subset of these results will become accessible to the wider community through publication.

In this paper we describe how the Quixote Project is developing the infrastructure required to convert output from a number of different molecular quantum chemistry packages to a common semantically rich, machine-readable format and to build repositories of QC results. Such an infrastructure offers benefits at many levels. The standardised representation of the results will facilitate software interoperability, for example making it easier for analysis tools to take data from different QC packages, and will also help with archival and deposition of results. The repository infrastructure, which is lightweight and built using Open software components, can be implemented at individual researcher, project, organisation or community level, offering the exciting possibility that in future many of these QC results can be made publically available, to be searched and interpreted just as crystallography and bioinformatics results are today.

Although we believe that quantum chemists will appreciate the contribution the Quixote infrastructure can make to the organisation and exchange of their results, we anticipate that greater rewards will come from enabling their results to be consumed by a wider community. As the repositories grow they will become a valuable source of chemical data for use by other disciplines in both research and education.

The Quixote project is unconventional in that the infrastructure is being implemented in advance of a full definition of the data model which will eventually underpin it. We believe that a working system which offers real value to researchers based on tools and shared, searchable repositories will encourage early participation from a broader community, including both producers and consumers of data. In the early stages, searching and indexing can be performed on the chemical subject of the calculations, and well defined calculation meta-data. The process of defining more specific quantum chemical definitions, adding them to dictionaries and extracting them consistently from the results of the various software packages can then proceed in an incremental manner, adding additional value at each stage.

Not only will these results help to change the data management model in the field of Quantum Chemistry, but the methodology can be applied to other pressing problems related to data in computational and experimental science.

## 37.2 Background

### 37.2.1 Quantum Chemical calculations and data

High-level quantum chemical (QC) methods have become increasingly available to the broader scientific community through a number of software packages such as Gaussian<sup>1</sup>, GAMESS(US)<sup>2</sup>, GAMESS-UK<sup>3</sup>, NWChem<sup>4</sup>, MOLCAS<sup>5</sup> and many more. Additionally, the cost of computer power has experienced an exponential reduction in recent decades and, more importantly, sophisticated approximations have been developed that pursue (and promisingly approach) the holy grail of linear scaling methods<sup>67</sup>. This has enabled any researcher, with no specific QC training, to perform calculations on large, interesting systems using very accurate methods, thus generating a large amount of valuable and expensive data. Despite the scientific interest of this data and its potential utility to other groups, its lack of homogeneity, organization and accessibility has been recognized as a significant problem by important agents within the scientific community<sup>89</sup>.

These problems, and specially the ones related to the accessibility of data have many consequences that reduce the efficiency of the field. As mentioned, QC methods are computationally expensive: the scaling of the computer effort and storage of high-level computations with the size of the system ( $N$ ) is harsh, reaching, for example,  $N^7$ , for the most expensive and most accurate wavefunction-based methods, such as Coupled Cluster<sup>101112</sup>. This makes it very difficult for groups that cannot use supercomputing facilities to have access to high-quality results, even if they possess the expertise to analyze and use the data. Even groups that do have access to powerful computational resources, given the lack of access to previously computed data by other researchers, often face the choice between *two inefficient* options: either they spend a lot of human time digging in the literature and contacting colleagues to find out what has already been calculated, or they spend a lot of computer effort (and also human time) calculating the needed data themselves, with the risk of needlessly duplicating work.

Another problem originating in the lack of access to computed QC data and the very large number of methods available, is that users typically do not have the integrated information about which method presents the best accuracy *vs.* cost relation for a given application. The reason is that comparing one quantum chemical method with another, with classical force fields or with experimental data is non-trivial, the answer frequently depending on the studied molecular system and on the physical observable sought. Moreover, all the details and parameters that define what John Pople termed a *model chemistry*<sup>13</sup>, *i.e.*, the exact set of rules needed to perform a given calculation do not obey a continuous monotonic function. Thus increasing the expense and “accuracy” of a calculation may not always converge to the “correct” solution. As a consequence, the quality of the results does not steadily grow with the computational effort invested, but rather there exist certain tradeoffs that render the relation between them more involved<sup>141516</sup>. Hence, not only the choice of the more efficient QC method for a given problem among the already existing ones, but also the design of novel model chemistries becomes ‘more an art than a science’<sup>17</sup>, based more on know-how and empiricism than in a set of systematic procedures.

---

<sup>1</sup> Gaussian 03, Revision C.02

<sup>2</sup> Advances in electronic structure theory: GAMESS a decade later

<sup>3</sup> The GAMESS-UK electronic structure package: algorithms, developments and applications

<sup>4</sup> NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations

<sup>5</sup> MOLCAS: A program package for computational chemistry

<sup>6</sup> A mathematical and computational review of Hartree-Fock SCF methods in Quantum Chemistry

<sup>7</sup> Advances in methods and algorithms in a modern quantum chemistry program package

<sup>8</sup> Towards a European e-Infrastructure for e-Science digital repositories - Final Report

<sup>9</sup> European Computational Science Forum: The “Lincei Initiative”: from computers to scientific excellence

<sup>10</sup> Parallel calculation of CCSD and CCSD(T) analytic first and second derivatives

<sup>11</sup> NOTITLE!

<sup>12</sup> NOTITLE!

<sup>13</sup> Nobel lecture: Quantum chemical models

<sup>14</sup> Efficient model chemistries for peptides. I. General framework and a study of the heterolevel approximation in RHF and MP2 with Pople split-valence basis sets

<sup>15</sup> Intrinsically stable secondary structure elements of proteins: A comprehensive study of folding units of proteins by computation and by analysis of data determined by X-ray crystallography

<sup>16</sup> Stability issues of covalently and noncovalently bonded peptide subunits

<sup>17</sup> Computational quantum chemistry: A primer

### 37.2.2 Design of Scientific data repositories

In this paper we describe a novel, flexible, multipurpose repository technology. It arises out of a series of meetings and projects in the computational chemistry (compchem) community which have addressed the desire and need to have repositories available for capturing and disseminating the results of QC calculations. It is also strongly influenced by the eScience (“cyberinfrastructure”, “eResearch”) programs which have stressed the value of instant semantic access to research information from many disciplines, and by the Open Innovation vision supported by the Scientific Software Working Group of CECAM (Centre Européen de Calcul Atomique et Moléculaire) <http://www.cecam.org/>, which seeks an innovation model based on sharing, trust and collaboration, and which recognizes the important role played by the availability of reference data and archives of outputs of calculations and simulations. It also coincides with the increasing mandates for data publication from a wide range of funders; our repository can address a large part of these requirements.

This paper describes a distributed repository technology and the social aspects associated with developing its use. The technology is robust and deployed but the way it may be used is at a very early stage. We address known social issues (sustainability, quality, etc.) but expect that deployment, even in the short term, may look very different from what is reported.

The development and acceptance of Wikipedia may act as a valuable guide and it represents a community-driven activity with community-controlled quality. Although variable, we believe that articles for most mainstream physical sciences are reliable. Thus to help understand and represent moments of inertia in computational chemistry we can link to Wikipedia [http://en.wikipedia.org/wiki/Moment\\_of\\_inertia](http://en.wikipedia.org/wiki/Moment_of_inertia). This contains many hundreds of edits over eight years from many authors - it is almost certainly “correct”. Quixote has many of the same features - anyone can contribute content and repurpose it. We expect a culture to emerge where the community sets guidelines for contributions and corrections/annotations. We are building filters (“lenses”) so that the community can identify subcollections of specific quality or value.

The background to Quixote includes a number of meetings and projects which specifically addressed the development of infrastructure in computational chemistry and materials. The goal of these was to explore the commonality between approaches and see how data and processes could interoperate. One (Materials Grid) also addressed the design and implementation of a repository for results.

- 2004: A meeting under the UK eScience program “Toward a common data and command representation for quantum chemistry” <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=394>.
- 2006: A meeting under the auspices of CECAM “Data representation and code interoperability for computational materials physics and chemistry” <http://www.cecam.org/workshop-50.html>
- 2005-2010: A 5-year project under the COST D37 program to develop various aspects of interoperability both within the calculation (Q5COST) and between programs (WG5).
- A funded project in computational materials (“Materials Grid”) <http://www.materialsgrid.org/> which resulted in considerable development of CML specifications and trial implementations in a number of codes (CASTEP, DL POLY).

These meetings and projects were exploratory and localized. Within them there was a general agreement that interoperability and access to results would be a great benefit. But they also highlighted the problem that infrastructure development is expensive and, if public, requires political justification for funding. Such funding is perhaps most likely to come from supranational efforts such as computational Grids, where there is a clear imperative for making services as accessible as possible. In COST-D37 the funding was for meetings and interchange visits; the WG5 community made useful but limited progress without dedicated developer or scientist funding.

There is often a vicious circle here - a frequent reason for not adopting a new technology in chemistry is “there is no demand for it”. This becomes a self-fulfilling prophecy and naturally limits innovation. It is also true that people are often only convinced by seeing a “working system” - hypothetical linkages and implementations have often been wildly optimistic. Therefore without seeing a working repository it is difficult to know what its value is, or the costs of sustaining it.

However the Internet age shows that it is much easier, cheaper and quicker to get new applications off the ground. It should be possible, in a short time and with modest effort, to create a system which demonstrates semantic interop-

erability and to convince a community of its value. We have successful examples of this reported elsewhere in this issue (OSCAR, Open Bibliography) where an early system has caught the imagination and approval of a section of the community.

### 37.2.3 Existing related projects

These issues, and undoubtedly more that will appear in the future, together with a wealth of scientific problems in neighbouring fields, could be tackled by public, comprehensive, up-to-date, organized, on-line repositories of computational QC data. Additionally, several fields reporting experimental data require it to be presented in a standard validatable form. The crystallography community has long required deposition of data as a prerequisite for publication, and this is now enhanced by machine validation (the CheckCIF philosophy and program <http://checkcif.iucr.org/>). When data are submitted, the system can comment on whether all appropriate data are present, inspect their values and compare either with known ranges or re-compute relationships between them based on accepted theoretical principles. In this way reviewers and readers can expect that a very large number of potential errors in experiment and publication have been eliminated.

This requirement for deposition of data as part of the publication process is increasingly common in bioscience, like genetics or proteomics, where the NCBI GenBank <http://www.ncbi.nlm.nih.gov/genbank/> or the Protein Data Bank (PDB) <http://www.rcsb.org/pdb/home/home.do> constitute very successful examples of data sharing and organization. In an age in which both the monetary cost and the accuracy of QC calculations rival those of experimental studies, the need to extrapolate the model to this field seems obvious. We also note that funders are requiring that data be deposited as part of the condition of funding. On the one hand, there exist some in-house solutions that individual research groups or firms have built in order to implement a local-scale data management solution. This is the case of David Feller's Computational Results Database <http://tyr3.chem.wsu.edu/~feller/Site/Database.html><sup>18</sup>, an intra-lab database to store and organize more than 100,000 calculations on small to medium-sized molecules, with an emphasis on very high levels of the theory. Also, the commercial standalone application SEURAT <http://www.synapticscience.com/seurat/> can open and parse QC data files and allows for metadata customization by the user, thus providing some limited, local databasing capabilities. In the same family of solutions, ChemDataBase<sup>19</sup> is a data management infrastructure mainly focused on virtual screening which presents the distinctive feature of being able to create and retrieve databases over grid infrastructures. Packages for interacting with QC codes (launching, retrieving and analyzing calculations), such as ECCE <http://ecce.emsl.pnl.gov/index.shtml> or Ampac <http://www.semichem.com/ampac/afeatures.php>, have modest data management capabilities too, although only insofar as it helps to perform their main tasks, and they can be regarded as intra-lab solutions as well. Probably the most complete in-house infrastructure of which we are aware of is the RC<sup>3</sup> (Regional Computational Chemistry Collaboratory) developed by the group of David Dixon at the Department of Chemistry of the University of Alabama. The main objective of RC<sup>3</sup> is to perform the everyday data backup, collection and metadata assignment for calculations, and to organize them for research purposes. At the time of writing, RC<sup>3</sup> has been tested by 36 users for more than a year, and backed-up and organized 1.6 million files, amounting to 1.5TB of data storage. The database contains 144,000 records and it can currently parse multiple QC data formats.

### 37.2.4 Heterogeneous data repositories

A different category of data management solutions from the one discussed above is that constituted by a number of online web-based repositories of QC calculations, normally developed by one research group with a very specific scientific objective in mind. Among them, we can mention the NIST Computational Chemistry Comparison and Benchmark DataBase (CCCBDB) <http://cccbdb.nist.gov/>, which contains a collection of experimental and calculated *ab initio* thermochemical, vibrational, geometric and electrostatic data for a set of gas-phase atoms and small molecules; the Benchmark Energy and Geometry DataBase (BEGDB) [www.begdb.com](http://www.begdb.com)<sup>20</sup>, which includes geometry and energy CCSD(T)/CBS calculations as well as other high-level calculations,

---

<sup>18</sup> The role of databases in support of Computational Chemistry

<sup>19</sup> ChemDataBase 2: An enhanced chemical database management system for virtual screening

<sup>20</sup> Quantum chemical benchmark energy and geometry database for molecular clusters and complex molecular systems [www.begdb.com](http://www.begdb.com): A users manual and examples

with a special emphasis on intermolecular interactions; the DFT Database for RNA Catalysis (QCRNA) [21](http://theory.rutgers.edu/QCRNA/), which contains high-level density-functional electronic structure calculations of molecules, complexes and reactions relevant to RNA catalysis; the Atomic Reference Data for Electronic Structure Calculations [22](http://www.nist.gov/pml/data/dftdata/index.cfm) compiled at NIST, containing total energies and orbital eigenvalues for the atoms hydrogen through uranium, as computed in several standard variants of density-functional theory, or the thermochemistry database at the Computational Modeling Group of Cambridge's Department of Chemical Engineering [23](http://como.cheng.cam.ac.uk/index.php?Page=cmcc), collecting thermochemical data of small molecules, powered by RDF and SPARQL and offering the output files of the calculations, together with the parsed CML [24](http://cml.sourceforge.net).

Apart from these solutions (either local or web-based), in which one or a few groups build a complete data management infrastructure, one can also consider the possibility of adopting a modular approach, in which different researchers tackle different parts of the problem, whilst always enforcing the maximum possible interoperability between the modules. The Blue Obelisk group [24](http://www.blueobelisk.org) has been championing this approach for a number of years now, and many of the developers of the tools discussed below are members of it. In this category of solutions, we can also mention the Basis Set Exchange (BSE) [18<sup>25</sup>](https://bse.pnl.gov/bse/portal), which provides an exhaustive list and definition of the most common basis sets used in QC calculations, thus facilitating the definition and implementation of semantic content regarding the method used, as well as improving the interoperability among codes at the level of the input data; modern tagging and markup technologies like XML and RDF together with the building of semantic dictionaries, not only to promote interoperability, but to do it in a web-friendly manner that allows one to easily plug modules and build complex online data management projects; the CML language (a chemical extension of XML)<sup>23</sup> is also one of the few cases in which a common semantics has been widely adopted by the chemistry community, and its extension to the QC field is one of the cornerstones of the Quixote project described here. Also on the interoperability front, we can mention the cclib [26](http://cclib.sf.net) and CDK [27](http://cdk.sf.net) libraries, as well as the OpenBabel toolbox [28](http://openbabel.org), which provide many capabilities for reading, converting and displaying QC data in many formats. Regarding the ease of use of possible data management solutions, the Open Source molecular editor and visualizer Avogadro [29](http://avogadro.openmolecules.net) can certainly be used as a useful module in complex projects, and in fact the design of Quixote is being carried out in collaboration with the developers of Avogadro, with the intention of efficiently interfacing it in future versions. The Java-based viewer Jmol [29](http://jmol.sourceforge.net/) performs similar tasks.

All in all, and despite the numerous efforts described above, it is clear that a global, unified, powerful solution to the management of data in QC does not exist at present; at the same time that the new internet-based technologies, the existence of vibrant communities, and the wide availability of powerful software to perform the calculations, and to convert and analyze the results, all seem to indicate that the field is ripe to produce a revolutionary (and much needed) change in the model. In this article, we present the beginnings of an attempt to do so.

### 37.2.5 The Quixote solution

The catalyst for Quixote was a meeting on interoperability and repositories in QC held at ZCAM (Zaragoza Scientific Center for Advanced Modeling), Zaragoza (Spain) in September 2010. There was general agreement on the need for collection and re-dissemination of data. In the final discussion a number of participants felt that there was now enough impetus and technology that something could and should be done. This wasn't a universal view, and we are aware that Quixote is unconventional in its genesis and aspirations - hence the name, reflecting a difficult but hopefully not impossible dream.

We decide to pursue this as an informal “unsponsored” project. It is not actually “unfunded”, in that we recognize the critical and valuable cash and in-kind support of several bodies, including CECAM, STFC Daresbury Laboratory,

<sup>21</sup> QCRNA 1.0: A database of quantum calculations for RNA catalysis

<sup>22</sup> Local-density-functional calculations of the energy of atoms

<sup>23</sup> Chemical markup, XML, and the Worldwide Web. 1. Basic principles

<sup>24</sup> The Blue Obelisk - Interoperability in Chemical Informatics

<sup>25</sup> Basis Set Exchange: A community database for computational sciences

<sup>26</sup> cclib: a library for package-independent computational chemistry algorithms

<sup>27</sup> The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics

EPSRC, JISC, ZCAM, and the employers of many of the participants. In particular we have been able to hold, and continue to hold, meetings. But there are no sponsor-led targets or requirements. In this it has many of the features of successful virtual projects in ICT (such as Apache, Linux, *etc.*) and communal activities such as Wikipedia and Open Street Map.

Speed and ambition were critical and project management has been by deadlines - external events fixed in time for which the project had to have something to show. These have included:

- An *ad hoc* meeting in 2010-10 in Cambridge where a number of the participants happened to be. This was to convince ourselves that the project was feasible in our eyes
- The PMR symposium 2011-01 that has catalysed this set of articles
- A workshop 2011-03 at STFC Daresbury Laboratory to demonstrate the prototype to a representative set of QC scientists and code developers
- Open repositories (OR11) 2011-06 where the technology was presented to the academic repository community as an argument for the need for domain repositories
- (planned) A meeting in Zaragoza 2011-08 where the argument for domain repositories will be demonstrated by Quixote.

As of 2011-06 we have a working repository with over 6000 entries, which are searchable chemically, by numeric properties and through metadata.

Our primary goal has been to build working, flexible technology without being driven by specific use-cases. This can be seen as heresy, and indeed we might regard it as such ourselves, if it were not that we have spent about 10 years working in semantic chemistry, computational chemistry and repositories and so have anticipated many of the possible use cases and caveats. To help show Quixote's flexibility we now list a number of use cases, any one of which may serve to convince the reader that Quixote has something to offer:

The Quixote system (Figure 1 shows the workflow, Figure 2 shows the distributed heterogeneity) is very flexible in that it can be installed in several different ways. Here we give a number of possible uses of the system, some of which we have deployed and several more we expect to be useful.

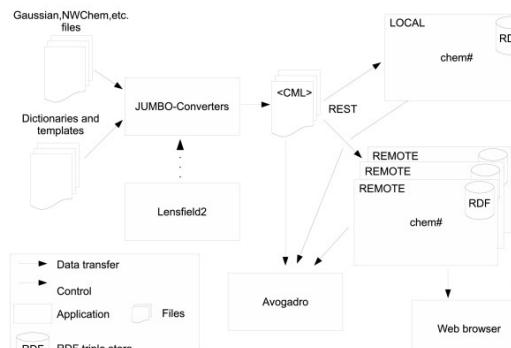


Figure 37.1: Figure 1. Quixote architecture and conversion workflow

**Quixote architecture and conversion workflow.** The user instructs Lensfield2 to convert output files of different computational chemistry codes into semantically rich CML files. The conversion is performed by JUMBO-Converters following the hints provided in the dictionaries and templates. The generated CML files are then transferred to one or more local and remote chem# repositories using a RESTful web API. The user can search and browse those repositories with a web browser, and can also manipulate and visualize the CML files with Avogadro.

- Collection of results within a group or laboratory. There is a growing desire to capture scientific results at the time of creation, and we have been involved in several projects (CLaRION, JISC XYZ) the impetus of which is to see whether scientists can capture their data as they create it. Computational chemistry is one of the simplest types of results and

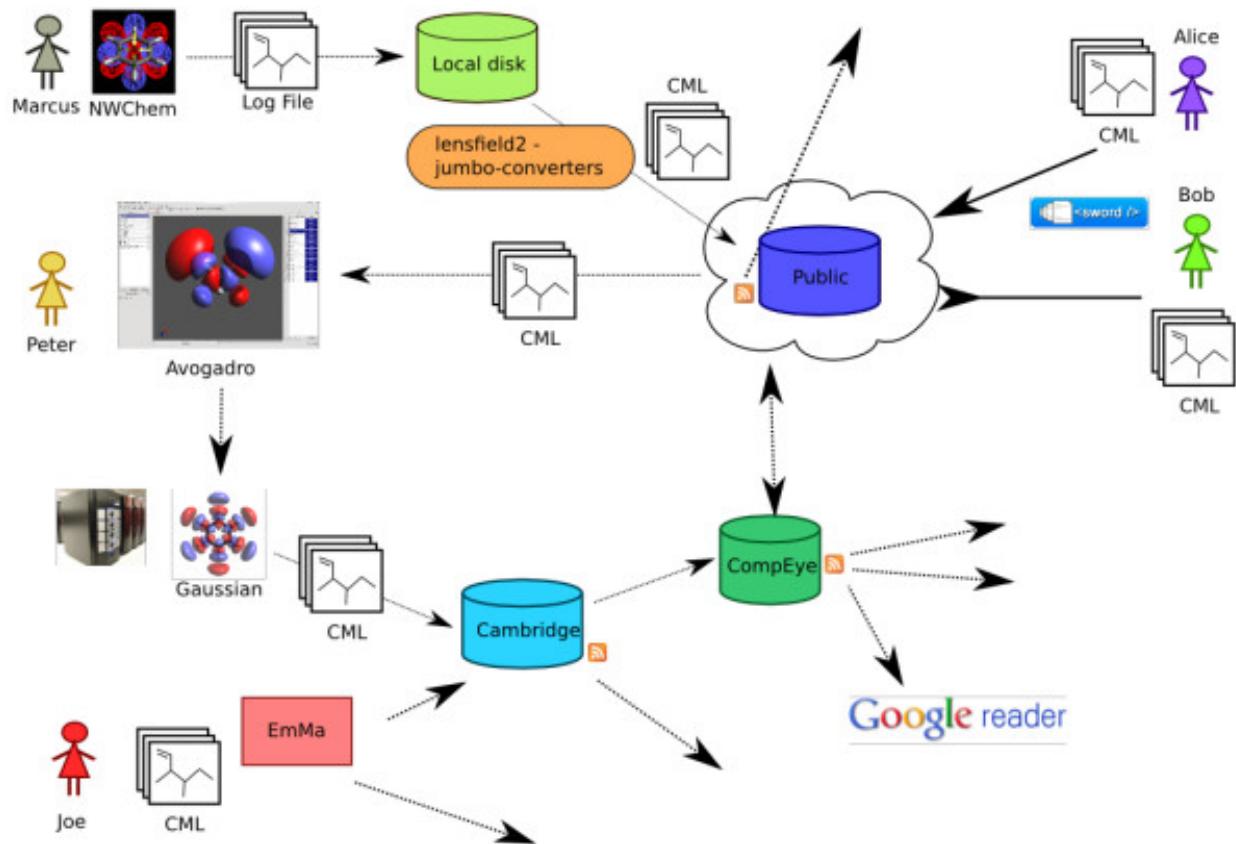


Figure 37.2: Figure 2. Quixote distributed repositories

**Quixote distributed repositories.** A schematic view of distributed Quixote repositories. Some repositories push documents to the public web, others aggregate from it. There is (deliberately) no check on whether repositories have identical documents. Users can build search strategies that look for individual entries with specific data or make collections of documents that share or contrast properties.

Quixote has been designed so that a single log file provides most of the input to the repository. This system allows groups and individual researchers to “pick up their results” and transport them to different environments.

- Formal publication in journals and theses. Results in a Quixote repository can be made available to other people and parties in the publication process. For example an author could make their results available to a journal before review so that the editors and reviewers could use the data to assess the value of the science. Similarly a graduate student could make their results available as part of their thesis submission and these could be assessed by the examiners. If the thesis and accompanying data are also published in the institutional repository then this provides a simple but very effective way of capturing and preserving the record of scientific experiments.
- Teaching and learning resources. Quixote can collect resources used for teaching and can also be used to provide subsets of research objects which are valuable for teaching and learning. For example in the current set there are 75 calculations on benzene, mainly from Henry Rzepa’s laboratory and these have been deposited by students carrying these out as part of their undergraduate work. This resource allows us to compare methods and to get information and experience which may help us do similar calculations.
- A collaborative central repository for a project. An increasing number of projects are distributed over geography and discipline. (The current Quixote project is an example.) A repository allows different people and groups in the project to share a central resource in an analogous manner to the use of Bitbucket and similar repositories for sharing code.
- A set of reference data and molecules. Quixote allows us to search for different parameters used in a given problem (*e.g.* level of theory, number of orbitals, convergence of results, algorithms, *etc.*).
- Validation sets for software and methods. In a similar manner datasets within Quixote can be used by different groups as reference input to compare results from different programs or different approaches.
- Enrichment of data through curation. Quixote is annotatable, so that it is possible for the world community to add their comments to particular entries. If a result is suspect, an annotation can be added. Similarly it is possible to point out related entries highlighting different scientific aspects.
- Building blocks for calculations. It is often valuable to start from an unknown program resource (*e.g.* a molecule whose structure is known and where the calculations are verified) and to modify it slightly for a related calculation, *e.g.* by adding additional atoms or by refining the calculation parameters.
- Combining data from different sources. As Quixote can also store experimental structures such as crystallographic ones, or experimental data such as spectra it is possible to enhance and combine components of the calculation.
- Data-driven science. Now that computational chemistry is relatively cheap and relatively accessible for a very large number of scientists, we foresee that literally millions of processors will be used routinely to calculate theoretical chemistry results. This allows us to carry out data mining from the Quixote repositories with the possibility of discovering new scientific patterns.
- Indexing the web. In a similar way to our indexing of crystallography through CrystalEye <http://wwmm.ch.cam.ac.uk/crystaleye/> we anticipate that web crawlers can increasingly discover and retrieve published computational chemistry.
- Developing software tools. Since Quixote represents an abstraction of many codes, developers writing software for computational chemistry will be able to see the type of semantics which are captured and the structure of the document.

## 37.2.6 Quality

The collection of the scientific computational record through Quixote could be regarded as an objective process in that each logfile is sufficiently described from the view of repeatability. Any user of Quixote could, if they had access to the code(s), re-run the calculation and “get the same output”. The examples of student calculations on benzene in the current content illustrate this view.

On the other hand it can be objected that unless a calculation is carried out with professional care then it can not only be meaningless but seriously misleading. Non-experts in QC can obtain these results and can misinterpret them. This is true, but it is a fact of modern Open science - results should be and are available to anyone. Science must evolve

social and technical methods to guide people to find the data they want. We can buy a kit and in our garages determine the sequence of a gene or protein without realising the potential experimental errors, or the difficulty of describing the species or strain that it came from. We can buy table-top crystallography sets that will automatically solve the structure of almost all crystalline materials. The results of these experiments are valuable if interpreted correctly and much of the time there is little room for serious error. However we might not realise that one lanthanide might be mistaken for another, that crystals can be twinned, and that certain spacegroups are problematic. Similarly the neophyte may not appreciate the difficulty of getting accurate energies, spin densities, non-bonded interactions, and many more subtleties of computational chemistry. But Pandora's box has been opened and computational chemistry is a commodity open to all. Quixote will help us in making our communal judgments.

There are a few objective concerns about quality. The Quixote system converts legacy computational chemistry (log-files) into semantic form. Automatic conversion will usually have a small number of errors, but mainly in that fields will not be recognized, rather than corrupted. In the early stages the semantics of some quantities may be misinterpreted (many are often laconic "E = 1.2345" - what exactly is E? and what are the units?) Given the exposure of the system to "many eyes" such problems will be few and should be relatively rapid to remove.

The fuzzier concern is whether Quixote can grow to gain the confidence of the QC and the non-QC community. Computational chemistry has the unique feature that anyone in the world, given the same input, will create the same output. The question is not whether the log file is an accurate record of the calculation but whether the calculation is valuable. It is quite possible to create junk, often unknowingly, and the commonest way is by inputting junk. A typical example is that many chemoinformatics programs can garble hydrogen counts and formal charges. However there are several criteria that the Quixote user and community can apply:

- If the methodology is very standard, then the results are likely to be usable in a similar way to other results using the same method. For example a very common combination of method and basis for organic molecules is B3LYP + 6-31G\*\*. If another group has successfully employed this for a set of molecules similar to the user's it is likely to be a useful starting point. This does not of course absolve the user from critical judgement but it is better than having nowhere to start.
- Automated methods can be used to compare the results of calculations for similar molecules or with varied parameters.
- We particularly encourage collections provided by specified individuals or groups. We have made two available in the current release (Dr. Anna Croft, Prof. Henry Rzepa). The user can browse through collections and get an idea of the type of calculation and the quality of metadata.
- Are the data coupled to publication? In CrystalEye almost all records are coupled to primary publications which can be read by the user (assuming that they have access to the journal). There is no technical barrier why this should not be done for articles and theses in computational chemistry. This is harder in compchem until the community develops a culture of publishing data concurrently with articles.
- Have the entries been annotated? This feature will shortly be available in Quixote, probably through blogging tools.
- Are there criteria for depositing an entry in the particular Quixote repository? Since we expect there to be many repositories, some of them can develop quality criteria for deposition. Some, perhaps the majority, may have human curators. In the first instance it will be important that users can assess the quality of a particular Quixote repository and we are appealing to any scientist who have collections of computational chemistry data that they would be prepared to make available. We expect that there will be a range of levels of quality in Quixote repositories. For example a crawler visiting random web sites for data might store these in an "unvalidated" repository. Users could examine this for new interesting entries and make their own decisions as to their value. The web has many evolved systems for the creation of quality metrics (popularity, usage, recommendations, *etc.*) and many of these would make sense for compchem. A journal might set up their own repository (as is done for crystallography). A department could expose its outputs (and thereby gain metrics and esteem) and the contents would be judged on the assessment of the creators.

## 37.3 Methods

All materials and methods mentioned here are available as Open Source/Data from the Quixote site or the WWMM Bitbucket repository. A small amount is added as appendixes to guide the reader.

### 37.3.1 Concepts and vocabulary

In any communal system requiring interoperability and heterogeneous contributions it is critical to agree concepts and construct the appropriate infrastructure. Chemistry has few formal shared ontologies and Quixote explores the scope and implementation of this for QC.

We draw inspiration from formal systems such as the Crystallographic Information File (CIF) created over many years by the International Union of Crystallography (IUCr). This is a community activity with medium-strong central management - the community has an input but there are formal procedures. It works extremely well and is universally adopted by crystallographers, instrument manufacturers, and publishers. The vocabulary and semantics have been developed over 20 years, are robust and capable of incremental extension. We take this as a very strong exemplar for Quixote and more widely QC.

We believe that almost all QC codes carry out calculations and create outputs which are isomorphic with other codes in the community. Thus an “electric dipole”, “heat of formation” or a “wavefunction” is basically the same abstract concept across the field. The values and the representation will be code-dependent but with the appropriate conversions of (say) units, coordinate systems and labelling, it is possible to compare the output of one code with another. This is a primary goal of Quixote, and we work by analysing the inputs and outputs of programs as well as top-down abstractions. It also means that Quixote is primarily concerned with what goes into and comes out of a calculation rather than what is held inside the machine (the data model and the algorithms).

### 37.3.2 Community development

From the human resource point of view, the Quixote project operates on a decentralised approach with no central site and with all participants contributing when available, and in whatever quantity they can donate at a particular time. For that reason, different parts of the project progress at variable speeds and technically independently. This means that there is very little effort required in collating and synthesising other than the general ontological problem of agreeing within a community the meaning deployment and use of terms and concepts.

The work is currently driven (*cf.* use cases) by datasets which are available. This drives the need to write parsers, collate labels into dictionaries, and collate results. In the week of 2011-05-09, for example, we ran daily Skype conferences, with Openly editable Etherpads <http://quixote.wikispot.org/> generously provided by the Open Knowledge Foundation (OKF) <http://okfn.org/>. The participants created tutorial material, wiki pages, examples and discussions which over the week focused us to a core set of between 20-50 dictionary entries that should relate to any computational chemistry output. The input to this effort was informed by logfiles from the Gaussian, NWChem, Jaguar and GAMESS-UK programs.

The initial approach has been to parse logfiles with JUMBO-Parser, as this can be applied to any legacy logfiles and does not require alterations of code. (At a later date we shall promote the use of CML-output libraries in major codes.) At this stage it is probably the best approach to analyse the concepts and their structure. A JUMBO-Parser is written for each code and run over a series of example logfiles. Ideally every part of every line is analysed and the semantic content extracted. In practice each new logfile instance can bring novel structure and syntax but it is straightforward to determine which sections have been parsed and which have not. Parsing failure may be because a parser has not been written for those sections, or because the syntax varies between different problems and runs. The parser writer can then determine whether the un-parsed sections are important enough to devote effort to, or whether they are of minor importance and can be effectively deleted.

The process is highly iterative. The parser templates do not cover all possible document sections and initially some parts remain unparsed. The parsers are then amended and re-run; it is relatively simple in XML to determine which

parts still need work.

Currently (2011-06) there are about 200 templates for NWChem, 150 for Gaussian and a small number for Jaguar, GAMESS-UK, GAMESS(US), AMBER and MOPAC <http://openmopac.net>. Each time a parse fails, the section is added as a failing unit test to the template and these also act as tutorial material and a primary source of semantics for the dictionary entries.

Quixote is designed as a bottom-up community project and co-ordinated through the modern metaphors of wikis, mailing lists, Etherpads and distributed autonomous implementations. The primary entry point is currently <http://quixote.wikispot.org/> which gives details of current resources and how to get involved.

### 37.3.3 Quixote components

#### JUMBO-Converters

The JUMBO-Converters are based on a templating approach, matching the observed output to an abstraction of the QC concepts. They have been hand-crafted for a number of well-structured output files (Gaussian archive files, MOPAC and various punchfiles) but the emphasis is now on writing JUMBO-Parsers for the logfiles for each code. We have explored a wide range of technologies for parsing logfiles including machine learning, formal grammars (lex/yacc), ANTLR <http://www.antlr.org/>, but all of these have problems when confronted with unexpected output, variations between implementations, error messages and many other irregularities. The JUMBO-Parser will not be described in detail here but in essence consists of the following approach:

- Recognition of common document *fragments* in the logfile (*e.g.*, tables of coords, eigenvalues, atomic charges, *etc.*) which appear to be produced by record-oriented (FORTRAN format) routines in the source code. We create a *template* for each such *chunk*, which contains *records*, with regexes for each record that we wish to match and from which we will extract information. These templates can be nested, often representing the internal structure of the program (*e.g.*, nested subroutine calls).
- Each template is then used to match any chunks in the document, which are then regarded as completed and unavailable to other templates. The strategy allows for nesting and a small amount of back-tracking.
- Chunks of document that are not parsed may then be extracted by writing additional parsers, very often to clean up records such as error messages or timing information.

At the end of this process a good parse will results in a highly-structured document with CML module providing the structure and CML scalar, array and matrix providing the individual fields [http://quixote.wikispot.org/Tutorials\\_and\\_problems](http://quixote.wikispot.org/Tutorials_and_problems).

This document is rarely fit for purpose in Quixote or other CML conventions and a second phase of transformation is applied. This carries out the following:

- Removal of unwanted fields.
- Removal of unnecessary hierarchy (often an artifact of the parsing strategy)
- Addition of
- Addition of units (often not explicitly mentioned in the logfile but known to the parser writer)
- Grouping of sibling elements into a more tractable structure (unflattening)
- Annotation of modules to reflect semantic purpose, *e.g.*, initial coordinates, optimizations, *etc.*
- Re-structuring of the modules in the parsed output to fit the *compchem* convention <http://www.xml-cml.org/convention/compchem>

This is carried out by a domain-specific declarative language which makes heavy use of XPath and a core set of Java routines for generic operations (delete/create/move elements, transform (matrix/molecule/strings etc.)). This approach means that failures are relatively silent (a strange document does not crash the process) and that changes can be made

external to the software (by modifying the transformation files). As with the templates this should make it easier for the community to maintain the process (*e.g.* when new syntax or vocabulary occurs).

A typical template is shown in Appendix A.

JUMBO-Parser has been designed for portability, in that most of the instructions are declarative (XML). It still requires the JUMBO-Parser interpreter to be ported, but this is written in mainstream Java and should not be particularly problematic for most object-oriented targets such as Python, C++ and C#. To help in the parsing, there are a large number of unit and regression tests.

### CML Conventions and Dictionaries

The final output is CML compliant to the *compchem* convention and validated against the current validator <http://validator.xml-cml.org/>. The dictionaries are in a constant state of update and consist of a reference implementation on the CML site and a working dictionary associated with the JUMBO-Converters distribution. As concepts are made firm in the latter, they are transferred to the reference dictionary.

The current compchem dictionary is shown in Appendix B. It contains about 90 terms which are independent of the codes. We expect that about the same amount again will be added to deal with other properties and solid state concepts.

### Lensfield2

Lensfield2 <https://bitbucket.org/sea36/lensfield2/> is a tool for managing file transformation workflows and can be thought of as a

Lensfield2 requires a build file, defining the various sets of input files and the conversions to be applied to them. Like Lensfield2 is designed to run workflow steps written in Java and build using Apache Maven <http://maven.apache.org/>, utilising Maven's dependency management system to pull in the required libraries for each build step.

Lensfield2 has been successfully used in running the parser and subsequent software over the 40,000 files in the test datasets 1-4 (*v.i.*).

### RESTful uploading

It is important that the methods for “uploading” and “downloading” files are as flexible as possible. Some collaborators may not have privileges to run their own server, so they need to be able to upload material to a resource run by other collaborators. However, if the protocols are complex then they may be put off taking part. Similarly, others may wish to delegate this to software agents which poll resources and aggregate material for uploading. Similar variability exists in the download process. Web-based collaborators are becoming used to very lightweight solutions such as Dropbox <http://www.dropbox.com/> where files can be uploaded, and where permitted, downloaded by anyone. We do not expect a single solution to cover everything, and the more emphasis on security, the more effort required. In this phase of Quixote, we are publishing our work to the whole world and do not expect problems of corruption or misappropriation. We have therefore relied on simple proven solutions such as RESTful systems. Some of this is covered in the semantic architecture paper in this issue, and here we simply illustrate that initial systems at Cambridge have been implemented with AtomPub <http://tools.ietf.org/html/rfc5023>. Because the academic repository system has invested effort in the SWORD system <http://www.ukoln.ac.uk/repositories/digirep/index/SWORD> (which runs over AtomPub), this allows us to deposit/upload aggregations of files.

### Chempound repository

Quixote is built on CML compchem and, in our system, is further transformed to provide RDF used for accessing sub-components and expressing searches. The Chempound (chem#) repository system <https://bitbucket.org/chempound/><sup>28</sup>

<sup>28</sup> Chempound - a Web 2.0-inspired repository for physical science data

(see Figure 3) has been built to support this. We expect that the first wave of distributed repositories will be using Chempound, and a publically accessible prototype repository is already in use within the Quixote project <http://quixote.ch.cam.ac.uk/>

### Institutional repositories, DSpace

Institutional repositories (running software such as DSpace <http://www.dspace.org/> or Fedora <http://fedoraproject.org/>) may be responsible for storing the raw output files that are transformed into CML by the JUMBO-Converters. Alongside, they will also store basic metadata (authorship, usage rights, related works, *etc.*).

This usage of institutional repositories distributes data management responsibilities among the institutions where the creators of the raw output files work. This provides an efficient basic data management support to the creators, and lets topic-specific repositories (such as Quixote's chem#) to focus on leveraging the specialized CML semantics extracted from the raw files, while still linking back to the original raw files at the institutional repositories. This schema also favors re-use of the same primary data by different specialized research topic repositories.

Yet another temporary advantage of this approach is that, as the data collection increases, resource discoverability becomes a real challenge - even for the researcher herself. Even if much data can be extracted from the datafiles, some title and description metadata could be very useful to issue searches and can be provided by the person submitting the files to the repository. In the development phase, other researchers - as well as the dataset creator - would be able to discover and access a given unprocessed dataset without needing to wait for it to get processed and transferred into the final Chempound data repository.

Designing a DSpace-based raw data repository will also allow for defining a *de facto* standardized metadata collection for compchem data description that may be very useful for harmonisation of data description in this specific research area - and might eventually evolve into some kind of standard for the discipline. At the present stage, we have done some preliminary work along metadata collection definition. A set of metadata has been defined and is being discussed in order to provide thorough descriptions of raw compchem datasets (potentially extendable to data from other research areas). Once the metadata set for bibliographical description of raw datasets is agreed, fields contained therein will be mapped to existing or new qualified DublinCore (QDC) metadata and a draft format will thus be defined. This format will be implemented at a DSpace-based repository, where trial-and-error storing loops with real datasets will be performed for metadata collection completion and fine-tuning - besides accounting for particular cases.

### Avogadro

Avogadro is an open source, cross-platform desktop application to manipulate and visualize chemical data in 3D. It is available on all major operating systems, and uses Open Babel for much of its file input and output as well as basic forcefields and cheminformatics techniques. Avogadro was already capable of downloading chemical structures from the NIH structure resolver service, editing structures and optimizing those structures.

Input generation from these structures is present for many of the major computational chemistry codes Quixote targets such as GAMESS(US), GAMESS-UK, Gaussian, NWChem, MOPAC and others. These dialogs allow the user to change input parameters before producing input files to be run by the code. The output files from several of these codes can also be read directly, this functionality was recently split out into OpenQube - a library to read quantum computational code log files, and calculate molecular orbitals, electron density and other output.

Ultimately, much of this functionality will move into the Quixote parsers, with the OpenQube library concentrating on multithreaded calculation of electronic structure parameters. A native CML reader plugin has also been developed for Avogadro, to read in CML files directly and display the tree structure allowing visual exploration of CML files. As JUMBO and other tools can extract electronic structure, spectra and vibrational data, this plugin is being developed to extract them from the CML document.

Avogadro is already network aware, with a network fetch extension interacting with the NIH structure resolver and the Protein Data Bank (PDB). Experimental support for interacting with a local queue manager is also being actively developed, sending input files to the queue manager, and retrieving log files once the calculation is complete. Some data management features are being added, and as Chempound has a web API a plugin for upload, searching and

**A**

chempound data repository

Home | Browse | Search | SPARQL | Feeds

- Gaussian Calculation  
**CH4**
- Gaussian Calculation  
**Cl**
- Gaussian Calculation  
**CH3**
- Gaussian Calculation  
**chlorobz/cl complex basis**
- Gaussian Calculation  
**Gaussian Calculation**

**B**

chempound data repository

Home | Browse | Search | SPARQL | Feeds

**nitrobz/cl basis**

[http://daedal.ch.cam.ac.uk:8080/content/output.log\\_15/](http://daedal.ch.cam.ac.uk:8080/content/output.log_15/)

**Calculation**

Package	Gaussian 03 (Sun64-SVR4-Unix-G03RevB.04)
Method	UB3LYP
Basis Set	6-311+G(d,p)

**Molecule**

Formula	C <sub>6</sub> H <sub>5</sub> ClNO <sub>2</sub> (2)
Point Group	C <sub>1</sub>
Electronic State	2-A

**Calculated Properties**

Hartree-Fock Energy	-897.052
---------------------	----------

Click to activate Jmol

Figure 37.3: Figure 3. Chempound repository graphical interface

**Chempound repository graphical interface.** Chempound accepts either converted compchem CML or logfiles (which are then parsed by the JMD) and stores them in a database. It can be queried through a SPARQL endpoint (I) and a Jmol viewer (II), host, dates, etc.) (III) initialization (molecular structure, basis sets, methods, algorithms, parameters, etc. in the Internet stage progression of optimization) (IV) finalization (molecular structure, properties, times, etc.) (a) Each entry is displayed with a thumbnail and key metadata (b) Properties and parameters for each entry, all searchable through SPARQL endpoint.

downloading of structures will be added. A MongoDB-based application has been prototyped, using a document store approach to storing chemical data. This approach coupled with Chempound repositories and seamless integration in the GUI will significantly lower barriers for both deposition and retrieval of relevant computational chemistry output.

Avogadro forms a central part of the computational chemistry workflow, but is in desperate need of high quality chemical data. The data available from existing online chemical repositories is a good start, but having high quality, discoverable computational chemistry output would significantly improve efficiency in the field. Widespread access to optimized chemical structures using high level theories and large basis sets would benefit everyone from teaching right through to academic research and industry.

### 37.3.4 Installing a Quixote repository

The Quixote system is based on the Chempound package, which provides a complete set of components for ingestion of CML, conversion to RDF and customisable display of webpages (using Freemarker templates). This installation has already been satisfactorily carried out at Zaragoza in less than 24 hours using the current Chempound distribution <http://bitbucket.org/chempound> and a number of calculations have been ingested. The Chempound system contains customisable modules for many types of chemical object and, in this case, is supported by the compchem module. This provides everything necessary for the default installation but, if customisation is required, the configuration and resource files in compchem-common, compchem-handler and compchem-importer can be edited. Chempound uses the SWORD2 protocol for ingest and so can accept input from any SWORD2- compliant client system.

## 37.4 Results and Discussion

The Quixote project can manage input and output from any of the main compchem packages including plane-wave and solid-state approaches. The amount of semantic information in the output files can vary from a relatively small amount of metadata for indexing to a complete representation of every information output in the logfile. The community can decide at which point on the spectrum it wishes to extract information and can also retrospectively enhance this by running improved parsers and converters over the archived logfiles and output files.

The current test datasets in the Murray-Rust group are generated by parsing existing logfiles into CML using the JUMBO-Converters software. The amount of detail depends at the moment on the amount of effort that has been put into the parser. The current project is working hard to ensure inter-operability of dictionary terms and concepts by collating a top-level dictionary resource. When this is complete, the files will be re-parsed to reflect the standard semantics.

In the first pass, with the per-code parsers, we have been able to get a high conversion rate and a large number of semantic concepts from the most developed parsers. The use cases below represent work to date showing that the approach is highly tractable and can be expected to scale across all types of compchem output and types of calculation.

A typical final CML document (heavily truncated for brevity) is shown in Appendix C. This shows the structure of jobs and the typical fields to be found in most calculations.

### 37.4.1 Test dataset 1

The first use case consisted of 1095 files in Gaussian logfile format contributed by Dr. Anna Croft of the University of Bangor. These were deliberately sent without any human description with the challenge that we could use machine methods to determine their scope and motivation. We have applied the JUMBO-Parser to these, of which all except 5 converted without problems. The average time for conversion was between 3-10 seconds depending on the size of file. These files have now been indexed, mainly from the information in the archive section of the logfile but also with the initial starting geometry and control information. A large number of the files appear to be a systematic study of the attack by halogen radicals on aromatic nuclei.

### 37.4.2 Test dataset 2

This use case comprised of over 5000 files which Henry Rzepa and collaborators have produced over the years and which have been stored Openly in the Imperial College repository (helix). They are much more varied than the Croft sample and include studies on Möbius computational chemistry, transitional metal complexes and transition state geometries. A considerable proportion of the files emanate from student projects, many of which tackle hitherto novel chemical problems. It is our intention to create a machine-readable catalogues of these files and to determine from first principles their content and, where possible, their intent.

### 37.4.3 Test dataset 3

The NWChem distribution (NWChem-6.0) contains a directory (

### 37.4.4 Test dataset 4

In the group of Pablo Echenique, at the Institute of Physical Chemistry “Rocasolano” (CSIC) and the University of Zaragoza, a large number of calculations were performed in peptide systems using the Gaussian quantum chemistry package. These calculations represent an exhaustive study (whose results and aims have been discussed elsewhere<sup>14</sup>), of more than 250 *ab initio* potential energy surfaces (PESs) of the model dipeptide HCO-L-Ala-NH<sub>2</sub>. The model chemistries investigated are constructed as homo- and heterolevels involving possibly different RHF and MP2 calculations for the geometry and the energy. The basis sets used belong to a sample of 39 representants from Pople’s split-valence families, ranging from the small 3-21G to the large 6-311++G(2 df, 2 pd). The conformational space of this molecule is scanned by defining a regular 12×12 grid from -165° to 165° in 30° steps in the 2D space spanned by its Ramachandran angles  $\phi$  and  $\psi$ . This totals more than 35000 Gaussian logfiles, all generated at the standard level of verbosity, some of them corresponding to single-point energy calculations, some of them to energy optimizations. The use of JUMBO-converters through Lensfield 2 has allowed to parse the totality of these files, through a complicated folder tree, generating the corresponding raw XML and structured compchem CML with a very high rate of captured concepts. The total time required to do the parsing was about five hours in an iMac desktop machine with a 2.66 GHz Intel Core 2 Duo processor, and 4 GB of RAM memory, running the Mac OS X 10.6.7 operating system.

### 37.4.5 Quixote repository at Cambridge

The first repository (Figure 3) has been built at Cambridge <http://quixote.ch.cam.ac.uk> and has been viewable and searchable. In the spirit of Quixote this is not intended to be a central permanent resource but one of many repositories. It is available for an indefinite time as a demonstration of the power and flexibility of the system but not set up as a permanent “archive”. It may be possible to couple such repositories to more conventional archive-oriented repositories which act as back-end storage and preservation.

## 37.5 Conclusions

Each day, countless calculations are run by thousands of computational chemistry researchers around the world, on everything from ageing, dusty desktops to the most powerful supercomputers on the planet. It might be supposed that this would lead to a deluge of valuable data, but the surprising fact remains that most of this data, if it is archived at all, usually lies hidden away on hard disks or buried on tape backups; often lost to the original researcher and never seen by the wider chemistry community at all.

However, it is widely accepted that if the results of all these calculations were publicly accessible it would be extremely valuable as it would:

- avoid the costly duplication of results,

- allow different codes to be easily validated and benchmarked,
- provide the data required for the development of new methods,
- provide a valuable resource for data mining,
- provide an easy, automated way of generating and archiving supporting information for publications.

In the rare cases when data is made openly available, the output of calculations are inevitably produced in a code-specific format; there being no currently accepted output standard. This means that interpreting or reusing the data requires knowledge of the code, or the use of specific software that understands the output. A standard semantic format will:

- allow tools, (*e.g.* GUIs) to operate on the input and output of any code supporting the format, vastly increasing their utility and range,
- enable different codes to interoperate to create complex workflows,
- additionally, if a semantic model underlies the format, data can easily be validated.

The benefits of a common data standard and results databases are obvious, but several previous efforts have failed to address them, largely because of an inability to settle on a data standard or provide any useful tools that would make it worthwhile for code developers to expend the time to make their codes compatible.

The Quixote project aims to tackle both of these problems in a pragmatic way, building an infrastructure that can be used to both archive and search calculations on a local hard-drive, or expose the data on publicly accessible servers to make it available to the wider community.

The vision with which we started the Quixote project some months ago is one in which all data generated in computational QC research projects is used with maximal efficiency, is immediately made available online and aggregated into global search indexes, a vision in which no work is duplicated by researchers and everyone can get an overall picture of what has been calculated for a given system, for a given scientific question, in a matter of minutes, a vision in which all players collaborate to achieve maximum interoperability between the different stages of the scientific process of discovery, in which commonly agreed, semantically rich formats are used, and all publications expose the data as readable and reusable supplementary material, thus enforcing reproducibility of the results; a vision in which good practices are wide spread in the community, and the greatest benefit is earned from the effort invested by everyone working in the field.

With the prototype presented in this article, which has been validated by real use cases, we believe this vision is beginning to be accomplished.

The methodological approach in Quixote is novel: The data standard will be consolidated around the tools and encourage its adoption by providing code and tool developers with an obvious reason for adopting the data standard; the “If you build it, they will come” approach. The project is rooted in the belief that scientific codes and data should be “Open”, and we are therefore focussing our efforts on using existing Open Source solutions and standards where possible, and then developing any additional tools within the project. The Quixote project is itself completely Open, de-centralised and community-driven. It is composed of passionate researchers from around the globe that are happy to collaborate with anyone who shares our aims.

## 37.6 Competing interests

The authors declare that they have no competing interests.

## 37.7 Authors' contributions

SA has participated in the design of the Quixote system, is the main developer of Chempound and collaborated in the development of the compchem dictionaries and conventions. PdeC has written the manuscript, and collaborated in the

design of the D-Space-based solution for metadata. PE has written the manuscript, participated in the design of the Quixote system and help develop some of the tools contained in it. JE has written the manuscript, participated in the design of the Quixote system and help develop some of the tools contained in it. MH has written the manuscript, participated in the design of the Quixote system and is a core developer of Avogadro. PM-R has written the manuscript, participated in the design of the Quixote system and he has been the main developer of the software tools. PS has written the manuscript, and collaborated in the design of the Quixote system. JTh has written the manuscript, participated in the design of the Quixote system and help develop some of the tools contained in it. JTo has participated in the design of the Quixote system, developed the CML validator and collaborated in the development of the compchem dictionaries and conventions. All authors have read and approved the final manuscript.

## 37.8 Appendixes

### 37.8.1 Appendix A. Template for parsing a link from Gaussian log files

A template to parse the output from the 601 link output in Gaussian logfiles. (The code for beta eigenvalues has been omitted for clarity.)

`http://www.xml-cml.org/schema" xmlns:cmlx="http://www.xml-cml.org/schema/cmlx">`

### 37.8.2 Appendix B. Dictionary for Computational chemistry

The current dictionary for (code-independent) computational chemistry. A few entries are shown in full; most show the id's and the terms. The full dictionary is maintained within the current Bitbucket content.

`http://www.xml-cml.org/schema" <!-- CML -->`  
`http://www.xml-cml.org/schema/cmlx" <!-- CML extensions -->`  
`http://www.w3.org/1999/xhtml" <!-- XHTML -->`  
`http://www.xml-cml.org/convention/" <!-- convention namespace -->`  
`http://www.xml-cml.org/unit/unitType/" <!-- CML unitType namespace -->`  
`http://www.xml-cml.org/unit/si/" <!-- SI units -->`  
`http://www.xml-cml.org/unit/nonSi/" <!-- other units -->`  
`http://purl.org/dc/elements/1.1/" > <!-- Dublin Core -->`  
`http://>" <!-- namespace of the dictionary -->`  
`http://www.xml-cml.org/convention/compchem/">compchem convention</h:a>`

### 37.8.3 Appendix C. CML produced from a Gaussian log file

A complete semantic parse for a Gaussian log file (Dr Anna Croft, for methane CH4). The log files describes two chained jobs, the first an optimization and the second the calculation of frequencies and thermochemistry. All significant information is captured, but much is repetitious and much is omitted here for brevity. Some fields have been truncated for clarity - no precision is lost in parsing.

The complete parse can be found at

`http://www.xml-cml.org/schema"`

`http://www.xml-cml.org/schema/cmlx"`

[http://www.xml-cml.org/convention/“](http://www.xml-cml.org/convention/)  
[http://“](http://)  
[http://“](http://)  
[http://www.w3.org/2001/XMLSchema“](http://www.w3.org/2001/XMLSchema)  
[http://www.xml-cml.org/unit/nonSi/“](http://www.xml-cml.org/unit/nonSi/)  
[http://www.xml-cml.org/dictionary/cml/“>](http://www.xml-cml.org/dictionary/cml/)

## 37.9 Acknowledgements

We thank all the many researchers that have contributed to the work discussed here with their ideas, testing and support; particularly Egon Willighagen, Anna Croft, Henry Rzepa, Lance Westerhoff, Luis Martínez-Uribe, Tamás Beke, Valera Veryazov, Weerapong Phadungsukanan, José Luis Alonso, Fermín Serrano, Isabel Bernal, and, of the library of Universidad de Zaragoza, Roberto Soriano, Miguel Martín, Teresa Muñoz and Ramón Abad (director). We also thank the ZCAM, and especially its Director, Michel Mareschal, for hosting and co-organizing the vibrant workshop in which the Quixote project was born. We thank as well the Daresbury Laboratory of the UK Science and Technology Facilities Council (STFC), which sponsored the First Quixote Conference, and EPSRC for supporting for the contribution of Jens Thomas through the Service Level Agreement with STFC. Both ZCAM and Daresbury are nodes of CECAM. Finally, thanks to Charlotte Bolton for the careful editing of the manuscript.

16. Echenique aknowledges support from the research grants E24/3 (DGA, Spain), FIS2009-13364-C02-01 (MICINN, Spain). P. Echenique and J. Estrada aknowledge support from the research grant 200980I064 (CSIC, Spain), and and ARAID and Ibercaja grant for young researchers (Spain). The mentioned meeting has been funded by ZCAM, the University of Zaragoza, Piregrid, the Aragón Government, and the Spanish Ministry of Science and Innovation. P. de Castro aknowledges support by Joint Information Systems Committee (JISC). Peter Murray-Rust acknowledges funding from JISC (CLaRION, XYZ) and EPSRC (Pathways to Impact).



# CMLLITE: A DESIGN PHILOSOPHY FOR CML

## 38.1 Abstract

CMLLite is a collection of definitions and processes which provide strong and flexible validation for a document in Chemical Markup Language (CML). It consists of an updated CML schema (schema3), conventions specifying rules in both human and machine-understandable forms and a validator available both online and offline to check conformance. This article explores the rationale behind the changes which have been made to the schema, explains how conventions interact and how they are designed, formulated, implemented and tested, and gives an overview of the validation service.

## 38.2 Introduction

There is an on-going need for formal, computable representations of scientific data and documents which are also accessible to humans<sup>1,2,3</sup>. The challenge is to devise systems that people will not only use but for which they will, critically, develop additional tools and content. Our approach for chemistry is Chemical Markup Language (CML) (whose evolution and philosophy is described elsewhere in this issue<sup>4</sup>) which has been developed to support five main areas of chemistry (molecules, reactions, solid-state, spectroscopy and computational chemistry).

The strengths and weaknesses of CML have been recently analysed by Dumontier<sup>3</sup> and we quote directly:

*Chemical Markup Language, backed by a controlled vocabulary, has been rather successful in specifying most aspects of chemistry, from small molecules and their connectivity to polymers and crystal structures.*

*Unfortunately, while most elements of this specification can be parsed out using one of the many XML libraries, certain elements do not render themselves to facile interpretation. Consider the sample CML specification of a water molecule [...]. In order to identify the member atoms in a given bond, it is necessary to carry out string processing as an intermediate step. Further, while many of the elements of CML are defined in a controlled vocabulary, the lack of explicit, consistent, and formal axiomatization of the involved concepts gives rise to difficulties in inferring connections between chemical concepts where no such connections are stated explicitly, something that is possible in formal ontology-backed RDF-based information specifications. Although CML specifications have been increasingly evolving to incorporate elements of the Semantic Web, the lack of widespread adoption of the format, and the limited availability of large-scale CML-based chemical knowledge repositories, have somewhat limited CML-assisted federation of the world of chemical data. Furthermore, the implementation of coverage of additional chemical concepts in*

---

<sup>1</sup> Overcoming the Limitations of a Connection Table Description: A Universal Representation of Chemical Species

<sup>2</sup> The semantic smart laboratory: a system for supporting the chemical \*e\*Scientist

<sup>3</sup> Chemical Entity Semantic Specification: Knowledge representation for efficient semantic cheminformatics and facile data integration

<sup>4</sup> CML: Evolution and Design

*most chemical representations requires a formal, rigorous representation specification, complicating the incorporation of data represented using domain-specific representation extensions. We believe that an ideal chemical representation would require no specialized wrapper or interpreter, would be generic such as to allow for facile and conflict-free extensions, would be based on a formal ontology, and would be encoded in a machine-\*understandable \*(as opposed to simply machine-readable, as in CML) manner and therefore facilitates automated reasoning and data integration.*

This article addresses these points and describes a system we have built for managing explicit and implicit semantics. It was initially developed during the Chem4Word (C4W) project <sup>5</sup> (which creates or edits CML documents in a.NET/Word context) and which we have now generalised to any CML deployment. In C4W we agreed that a fundamental part of the design was that the semantics could be verified. Any document input to the system must be semantically valid so that the C4W system would not break for invalid input. Essentially we designed a contract between the importing system, and the editing/display system.

Rather than rewrite JUMBO <sup>6</sup> and other CML libraries, we designed a set of rules for conformant input documents and tools to process validation. These tools (CML schema3 and CMLValidator) are platform-independent and are reported in this article.

### 38.2.1 Semantics in CML

We agree with Dumontier's analysis and in this article show how our current approach to semantics in CML is both achievable and largely compatible with his and others' <sup>7<sub>8</sub></sup> ideal chemical representations. As noted, CML has a small, but important, set of elements (molecule, atom, bond, crystal, spectrum and a few others) where some semantics are implicit and the rules hardcoded. This approach is pragmatic; translating the implicit rules to formal semantics is a considerable effort and makes it more difficult to write libraries to support them.

However most CML concepts can be automatically expressed in equivalent semantic form, *e.g.* using RDF format <sup>9</sup> for the document and RDFSchema <sup>10</sup> or (if appropriate) the OWL language to specify an ontology <sup>11</sup> (see Figure 1) and managed with generic (non-chemical) semantic tools. The use of RDF in this manner is advocated by *e.g.* the Bio2RDF project <sup>12</sup>. In particular property and parameter can be completely represented in RDF and we already use this extensively in Quixote <sup>13<sub>14</sub></sup> and similar projects (where CML is imported as RDF).

We have explored a full RDF implementation of CML through ChemAxiom <sup>16</sup> (an OWL-compatible representation of physical chemical properties) and we have also explored full RDF in Open Bibliography <sup>17</sup>). Both of these have shown that the entry overhead is high as the tools are at an early stage. For example there is no support for RDFSchema-based approaches in chemistry. At this stage in chemical informatics, therefore, we feel that CML as a mixture of explicit and implicit semantics provides a useful infrastructure accessible to a large number of implementers and users.

### 38.2.2 Implicit semantics

As a typical example of implicit semantics CML schema2.4 requires the formalCharge on a molecule to be consistent with the formalCharges on descendant molecules and atoms. We can express this in pseudocode:

if (not molecule[@formalCharge]) then

molecule@formalCharge: ==

---

<sup>5</sup> Chemistry Add-in for Word

<sup>6</sup> Chemical Markup Language

<sup>7</sup> Model Tool to Describe Chemical Structures in XML Format Utilizing Structural Fragments and Chemical Ontology

<sup>8</sup> Design and Development of Chemical Ontologies for Reaction Representation

<sup>9</sup> Resource Description Framework, RDF

<sup>10</sup> RDF Vocabulary Description Language 1.0: RDF Schema

<sup>11</sup> Web Ontology Language (OWL)

<sup>12</sup> Bio2RDF: Towards a mashup to build bioinformatics knowledge systems

<sup>13</sup> The Quixote project: Collaborative and Open Quantum Chemistry data management in the Internet age

<sup>14</sup> Quixote project on QC databases

<sup>16</sup> ChemAxiom-An Ontological Framework for Chemistry in Science

<sup>17</sup> Open Bibliography for Science, Technology, and Medicine

```

<property dictRef="compchem:hfenergy">
  <scalar dataType="xsd:double" units="nonsi:hartree">-39.1202429</scalar>
</property>

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix cml: <http://www.xmlcml.org/rdf-schema#> .
@prefix compchem: <http://www.xml-cml.org/dictionary/compchem/> .
@prefix nonSi: <http://www.xml-cml.org/unit/nonSi/> .

<http://example.com/a/calculation>
  compchem:hfenergy [
    a cml:Property ;
    rdf:value "-39.1202429"^^xsd:double ;
    cml:units nonSi:hartree
  ] .

```

Figure 38.1: Figure 1. A property represented in CML (top) and the equivalent RDF (bottom)-as used in the Quixote repository

**A property represented in CML (top) and the equivalent RDF (bottom)-as used in the Quixote repository**<sup>15</sup>. The dictionary is referenced by **compchem:hfenergy** for which there must be an entry in an online dictionary. These are completely equivalent and can be translated in both directions without semantic loss. (There are minor syntactic variants such as the capitalization varying systematically.) It is always possible to generate RDF from CML; the reverse may not be possible for arbitrary RDF. The challenge is to create communally acceptable dictionaries/ontologies-the syntax (CML or RDF) is immaterial.

```
sum (./molecule@formalCharge * ./molecule@count) +
sum (./atomArray/atom(@formalCharge * @occupancy * @count))
```

(If the formal charge (an integer) is missing on a molecule, calculate it by recursively summing its descendants. This is more complex in practice as we have to apply semantics for atoms without formalCharge.)

CML has thousands of relationships like this, and they are relatively straightforward to implement through procedural code (in libraries such as JUMBO, Chem4Word, the Chemistry Development Kit (CDK)<sup>18</sup>, Open Babel [#B18]\_\*etc\*). This article describes how their combination with unit tests and other validation procedures creates strong accessible semantics.

### 38.2.3 Choice of semantic system

It is commonly believed that the (perceived) ease of use of a new technology will affect its adoption by communities<sup>19</sup>. A successful deployed system needs to have the following interconnected components:

- Accessibility for humans
- Proven infrastructure
- Authoring tools
- Reading tools
- Editors
- Domain libraries
- Critical mass of content
- Agreed concepts and vocabularies
- Critical mass of users.

This requires a large investment to which we also have to add *Postel's Law*<sup>20</sup>: “be conservative in what you send, liberal in what you accept” *i.e.* do extra implementation work to make it forgiving to use. However we believe the investment in CML<sup>21</sup> has been sufficient to make the semantic approach valuable and tractable.

There seems to be a conservation law which trades ease of implementation and deployment for semantic power. At one of the spectrum is natural language (NL) which is almost infinitely expressive. It relies on an implicit fluid vocabulary and the burden on interpretation is almost completely on the accepter. Its flexibility also generates ambiguity. At the other end are completely hardcoded unambiguous systems with very limited scope (such as InChI<sup>22</sup>); this works because there is a single global implementation of a canonical InChI generator. NL can transmit the concept of “boiling point” because “everybody knows what a boiling point is”; InChI cannot represent the concept at all. To represent “boiling point” formally, however, is by no means trivial—we have to think about units, pressure, error estimations, *etc.* Does boiling point apply to vapour- > liquid transitions? There are thousands of similar chemical concepts all of which must be formalized. There is no escape from this labour.

CML trades full semantic representability for (relative) ease of implementation together with clarity for humans. CML takes a pragmatic view that a large number of chemical concepts are implicitly very well understood (most were formulated 100+ years ago) and the semantics can be hardcoded. This allows us to write software libraries for analysing orbital energies, balancing reactions, finding moments of inertia, *etc.* using the common representation that CML provides.

<sup>18</sup> The Chemistry Development Kit, CDK

<sup>19</sup> Perceived usefulness, perceived ease of use, and user acceptance of information technology

<sup>20</sup> NOTITLE!

<sup>21</sup> CMLValidator service

<sup>22</sup> IUPAC International Chemical Identifier, InChI

Here we explore how CML, which represents a set of basic chemical “nouns” (objects), can be combined in flexible, yet rigorous ways. In particular it has to be possible to write software systems that support these developments. We do not set *a priori* constraints on how these nouns can be used, but we require that these usages are documented and validatable, allowing us to write conformant software for each usage.

CML deliberately does not attempt to represent relationships between objects leaving that to RDF; nor does it represent processes (we are still searching for a good, common, formulation). CML is designed to interoperate with other markup languages (XHTML<sup>23</sup>, MathML<sup>24</sup>, SVG<sup>25</sup>, etc.) and is incorporated in some approaches, e.g. BioPAX<sup>26</sup>.

At the present time, therefore, CMLLite represents a cost-effective system which can validate a wide range of chemical documents.

### 38.2.4 Community requirements and CMLLite conventions

CML is now largely developed by communities who build prototypes and provide feedback on how well they work; CMLLite has been created and deployed in this way (Table 1).

The greater flexibility introduced with CML schema3 allows users to create valid documents almost as they want but requires a greater effort understanding for both humans and machines to understand the document. Here are typical community requirements:

- CMSpect. “All spectra MUST contain x-data and y-data”.
- CMLOpen. “only the following CML elements are allowed: module, molecule, atom, property,...” “bond MUST NOT appear as it is not a QM concept”
- molecular (from the Chem4Word project). “a bond MUST contain references to two distinct atoms, the atoms MUST exist, and be in the same ancestor molecule”.
- compchem (from Quixote). “a document MUST have a list of jobs, and each job MUST describe environment, initialization, calculation, and finalization”. All molecules MUST obey the molecular convention.
- dictionary. “all entries MUST have a definition and MAY have one description.”
- Unit-dictionary. “there SHOULD be a specific dictionary for SI units and unitTypes.”

The terms are used as in the IETF’s RFC 2199<sup>27</sup>: “MUST”, “MUST NOT”, “REQUIRED”, “SHALL”, “SHALL NOT”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY”, and “OPTIONAL”. This approach is central to CMLLite.

These domains of chemistry think about chemistry differently from each other; often this means a very tight specification of rules in one particular area of expertise and very little if any applied to the rest. The loosening of the content model in schema3 allows users to combine the elements and attributes as they need. However, users still need to be able to specify a set of rules (constraints) which model their particular domain. The entire set of constraints which the CML should conform to is called a *convention*. Every convention requiring another recursively inherits (aggregates) the requirements from that convention.

A convention should be the result of community engagement and discussion reflecting historical practice and experience. The social aspects of the process of agreeing conventions are discussed in the companion articles<sup>13,28</sup>.

A convention is:

- A description to a human reader of the purpose of the convention, its scope and its implementation. A human MUST be able to hand-craft a compliant document by reading the specification.

<sup>23</sup> XHTML specification

<sup>24</sup> Mathematical Markup Language (MathML) specification

<sup>25</sup> W3C Scalable Vector Graphics (SVG) Working Group

<sup>26</sup> The BioPAX community standard for pathway data sharing

<sup>27</sup> IETF RFC 2119: Key words for use in RFCs to Indicate Requirement Levels

<sup>28</sup> The semantics of Chemical Markup Language (CML): dictionaries and conventions

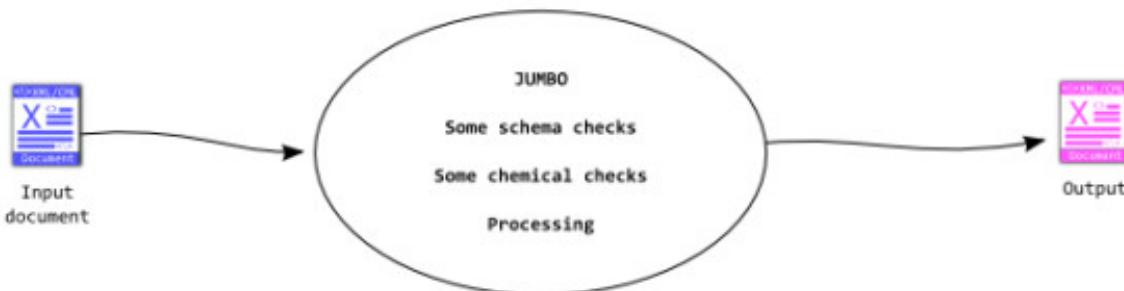
- A description to an implementer of exactly how software SHOULD, MAY, MUST and MUST NOT behave when given any possible input. For example software validating a document purporting to be compliant to a particular convention MUST raise an error if it encounters a node defined in the convention but used incorrectly. If it encounters a node not in the convention, its behaviour is undefined but the default should be to inform the user.

- A statement of interest in a particular subset of CML by a community.

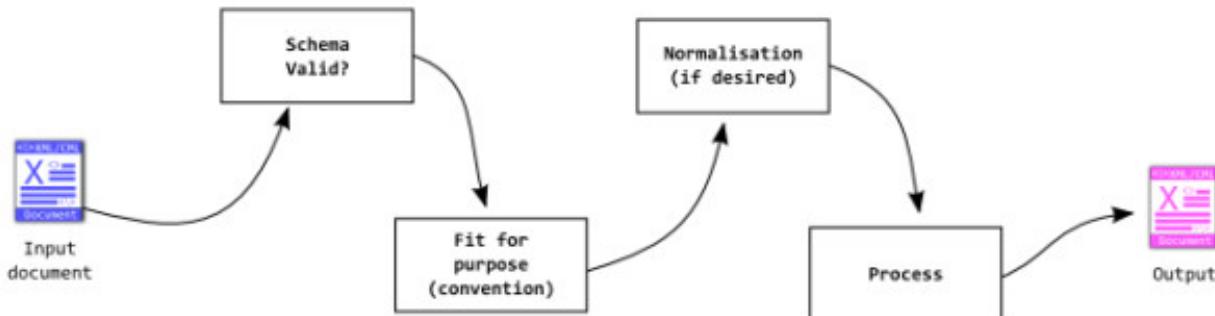
The prime purpose of the convention is validation of documents before transferring them to software. As a result the software is more straightforward to implement and test.

### 38.2.5 The Need for Validation

Validation of input documents is at the heart of the CMLLite approach. There are two complementary approaches to validation (see also Figure 2). Both components (Schema and convention) are *validators* and are normally run sequentially



a) Historical approach to CML processing with the software expected to perform a wide variety of disparate tasks.



b) New CMLLite modular approach to CML processing, separating different functions into discrete modules.

Figure 38.2: Figure 2. (a) historical approach to CML processing

- (a) **historical approach to CML processing.** Software was expected to perform a wide variety of tasks including validation and transformations (processing). (b) the CMLLite approach: each module performs only one task *i.e.* validation, normalisation or transformation (processing). This makes each of the modules more straightforward to understand and produces cleaner code.

- **XSD Schema**<sup>2930</sup>. CML has used this for many years. It works well for isolated elements and attributes with

<sup>29</sup> XML Schema Definition Language (XSD) 1.1 Part 1: Structures

<sup>30</sup> XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes

uncomplicated child content. It breaks or is inappropriate for several chemical concepts, complicated content and relationships. In this article we report schema3 where many of the broken and inflexible constraints have been removed. Note that all CML documents now should validate against schema3.

- **Conventions.** These add power to schema3 and allow many complicated concepts to be represented (in XSLT<sup>31</sup>/XPath<sup>32</sup>).

These components are now described in more detail.

### 38.2.6 CML Schemas-Evolution to schema3

In themselves, schema constraints can provide little chemical validation but provide good support for other simple concepts (*e.g.* numeric, date, and containers) and are the platform on which further constraints are built.

#### Content Models

Schema2.4 introduced flexibility through more relaxed content models than previous incarnations (an unordered child set with no enforced cardinality), and re-usable attributes. Schema3 has even more flexible content (effectively “any” for most elements) and much of the burden of validation has been devolved to conventions. Specific issues are described below (and are also addressed in the CML retrospective paper). The move away from the ‘one-size-fits-all’ model imposed by schema2.4 to a more modular, flexible approach, with supporting tools for implementation, has been driven by challenges in the general areas listed below in approximate order of importance:

- **Content Model** (what elements are allowed as children of which elements). Schema3 explicitly removes as much of the content model as possible.
- **Attribute names and attributeGroups.** Schema2.4 allowed attributes to be defined independently of elements. For maintenance purposes each attribute was defined in its own *attributeGroup*. Unfortunately some attributes used polymorphic names (*e.g.* “type”) and were not re-locatable. The desire to maintain backward compatibility with the majority of existing software means that we were unable to redesign the attribute names..
- **Union of enumerated values.** Some attributes (and string content) used the XSD union approach to express both controlled (enumerated) and uncontrolled vocabulary. Here enumerated values are of one data type whilst the UNIONed value is of a different type, and has to be processed differently. The CMLLite approach restricts attribute values to a single data type, and uses the dictionary (dictRef) mechanism to provide additional information as required.
- **Mixed content** (text and element children). This was used to support free text but is technically challenging and we are deprecating its use in favour of (*say*) XHTML constructs.
- **Aliases** (*e.g.* ‘1’ and ‘S’ for the order of single bonds). These cause a huge overhead in software, are deprecated, and will trigger warnings when the CML is validated against the validation service based on schema3. Normalisation is advised at this stage.

CML has grown to have a collection of approximately 100 elements and approximately 100 attributes. Most of these are in common use, but there are very few documents which use more than about 20 elements and 20 attributes at a time. For example a solid-state calculation has relatively little in common with the textual report of a chemical synthesis. Most elements in CML can be used independently of most other elements and schema3 explicitly supports this. For example a spectrum might occur with a molecule, with a crystal structure or with a computational chemistry output. Some elements have a more restricted use, for example bonds and atoms normally only occur within the context of a molecule. Attributes are more varied, in that some are specific to particular elements (*e.g.* atomRefs2 normally only occurs on the bond elements) while others are very generic (*e.g.* title, id, dictRef). The CML schema determines some of the pattern of attribute occurrence, but leaves others up to the individual conventions.

Almost all changes are backward-compatible as schema3 is more forgiving than schema2.4; a few elements contained mixed content and have been obsoleted (annotation, appinfo, documentation, relatedEntry).

<sup>31</sup> XSL Transformations (XSLT)

<sup>32</sup> XML Path Language

## Attributes

Attributes define string values and can constrain syntax, dataTypes, lists and other constructs. In schema2.4 many attributes had data types defined by a union of enumerations and a “namespaceRef” pattern (effectively a QName). This has now been relaxed to the enumeration with the addition of “other”. Constraints are then added with XPath/XSLT rules. The polymorphism of attributes with names such as type (Appendix A) has not yet been addressed. In schema3 attributes are used in the same way as in schema2.4 and rely on additional constraints added through conventions. Elements with text-only content (scalar, array, matrix) are polymorphic (*e.g.* can be numeric, string, date) and are not supported well by schema constraints.

### 38.2.7 Conventions

A convention specifies and can enforce the relationships between schema components and consists of (often a large number of) statements (rules) that can be understood by humans and enforced by machines. The choice of language for implementation is in principle, arbitrary. We initially used the Schematron<sup>33</sup><sup>34</sup> approach but have since moved to XSLT making heavy use of XPath, an extremely expressive language. XSLT has the advantage that it is implemented in all major languages and highly portable.

CMLLite has to support documents in a rigorous manner whilst accepting that these could come from a variety of sources and describe a wide range of possible chemical concepts. Therefore any CML element in the schema should be allowed, but would not by default have specific constraints. Any foreign XML elements would also be allowed and again would not have any specific constraints.

- An element can have text-only or element-only content (which may be empty, but there are no specifically empty models). For elements described in a defined convention constraints may apply. There are no restrictions on the order of elements in most content models.
- A document MAY contain more than one convention. Conventions are allowed to mandate other conventions.
- An element not specifically mentioned in a convention is effectively ignored by any tools that process after validation has succeeded (*i.e.* treated as any other foreign element), but is not removed from the document.
- Attribute data types are validated by their constraints in the CML schema but further constraints including *e.g.* required/forbidden, scoping of uniqueness and co-occurrence MAY be specified by conventions. These may restrict, but not alter the schema3 interpretation.

The interpretation of an element should not normally be affected by a convention. It constrains inputs and outputs but not the meaning of concepts. For example the `atom/@x3` attribute always defines Cartesian coordinates, and in a right-handed system. A convention can insist that they do or do not exist, that other nodes must or must not exist, but it cannot change the primary semantics.

## 38.3 Methodology of Validation

### 38.3.1 Validation-driven Development

Our approach to validation is strongly informed by test-driven development (TDD), a well-used methodology for building modern software systems<sup>35</sup>. The schema and the validator have been built by creating tests and refining the schema and software such that the tests produce conformant behaviour. To illustrate the philosophy of TDD, we show

---

<sup>33</sup> Schematron, a language for making assertions about patterns found in XML documents

<sup>34</sup> XSLT UK 2001 Report

<sup>35</sup> NOTITLE!

a typical unit test before describing the construction of the validator. There are thousands of unit tests using CML in JUMBO, JUMBO-Converters<sup>36</sup>, Bioclipse<sup>37</sup>, CDK, Open Babel, *etc.*).

### 38.3.2 A typical example of TDD

The following XML and Java snippets define the semantics of moleculeTool.getAverageBondLength() using the JUnit<sup>38</sup> framework:

```
<molecule id = 'mol5'>
<atomArray>
<atom id = 'a1' elementType = 'C' x3 = '0.0' y3 = '0.0' z3 = '0.0' />
<atom id = 'a2' elementType = 'N' x3 = '0.0' y3 = '1.3' z3 = '0.0' />
<atom id = 'a3' elementType = 'O' x3 = '1.0' y3 = '2.2' z3 = '0.0' />
<atom id = 'a4' elementType = 'H' x3 = '0.85' y3 = '-0.54' z3 = '0.5' />
<atom id = 'a5' elementType = 'H' x3 = '-0.85' y3 = '-0.54' z3 = '0.5' />
</atomArray>
</molecule>
```

The function is described in words (in this case the method name is sufficient) and we implement a test which runs the code against an expected valid output:

```
@Test
public void testGetAverageBondLength() {
    molTool5.calculateBondedAtoms();
    Assert.assertEquals("average length", 1.2235,
        molTool5.getAverageBondLength(CoordinateType.*CARTESIAN*), .0001);
}
```

This test passes the assertEquals statement if it can calculate the averageBondLength and also if the result is equal to the expected values within a given tolerance (0.0001). The test gives an example of conformant input and besides being a useful pedagogic and reference document it also implicitly defines semantics (“an average bond length calculation requires all atoms to have 3-D coordinates; these can be supplied as x3/y3/z3”).

### 38.3.3 General aspects of TDD

Test-driven development not only provide a method for verifying the behaviour of existing software-it also provides examples of typical use cases for anyone using CML. Note that unit tests provide implicit rather than explicit semantics-we can define any number of valid input and the outputs required for these, but the actual transformation can be performed by any means.

---

<sup>36</sup> JUMBO-Converters

<sup>37</sup> Bioclipse

<sup>38</sup> NOTITLE!

### 38.3.4 Schemas and CMLValidator

Schemas and conventions are systems to validate documents (*validators*). The basic strategy used throughout the validator design process is to create documents to test them (*validatorTests*). The choice of tests is critical-ideally the implementer should think of every possible distinct case, but in practice this is reduced to generic cases. It is important to generate broken documents as well as valid ones, and this is often surprisingly difficult. In practice edge cases crop up unexpectedly in large corpora and these must then be added to the *validatorTests*. Appendix B shows the effort required to create tests for even a simple convention.

Once the *validatorTests* are created the convention or schema is then coded. In line with test-driven development this starts with the tests failing (deliberately) as there is no code. The validator is then coded until it passes the tests. Frequently during this process the author will gain insight and inspiration and refine the *validatorTests*.

### 38.3.5 A schema3 validatorTest

As an example of how to test schema3 we take the definition of molecule in schema3 (Figure 3).

```
<xsd:element name="molecule" id="el.molecule" substitutionGroup="anyCml">
  <xsd:complexType>
    <xsd:choice minOccurs="0" maxOccurs="unbounded">
      <xsd:element ref="anyCml"/>
      <xsd:any namespace="#other" processContents="lax"/>
      <xsd:any namespace="#local" processContents="lax"/>
    </xsd:choice>
    <!-- attributes omitted for clarity -->
  </xsd:complexType>
</xsd:element>
```

Figure 38.3: Figure 3. A snippet from schema3 showing the typical relaxed content model of the molecule element container

**A snippet from schema3 showing the typical relaxed content model of the molecule element container.**

This is read as:

*A molecule can have zero (minOccurs = 0) or many (maxOccurs = unbounded) child elements in any order; it has no mixed content. The children can be any CML elements ("anyCml"), any elements with a foreign namespace (#other or #local (default namespace when not CML)). All the elements in the CML namespace are part of the anyCml substitution group.*

There are currently *ca.* 300 *validatorTests* (which test both schema and conventions) and we show examples that can validate the schema snippet for molecule. Each *validatorTest* is run against the schema, which only emits messages for invalid constructs.

1.

```
<cml:molecule xmlns:cml = "http://www.xml-cml.org/schema">
<element-in-default-namespace>
```

This is fine. The null prefix is not bound

to anything and therefore is associated with

the default namespace

```
</element-in-default-namespace>
</cml:molecule>
```

This is valid because the null prefix is not explicitly bound to anything and therefore associated with the default namespace and the CML namespace is bound to the cml prefix. This construct is permitted because of the xsd:any namespace = ‘##local’ in the schema.

2.

```
<molecule xmlns = “http://www.xml-cml.org/schema“ xmlns:other = “http://www.example.net“>
<other:foreign-element>
```

This is fine. The null prefix is bound to the  
CML namespace and the “other” prefix  
is bound to a non-CML namespace

```
</other:foreign-element>
</molecule>
```

This document is valid because the null prefix is bound to the CML namespace and the other prefix is bound to a non-CML namespace. This construct is permitted because of the xsd:any namespace = ‘##other’ in the schema.

3.

```
<molecule xmlns = “http://www.xml-cml.org/schema“>
<non-cml-element>
```

This is invalid. The null prefix is bound  
to the CML namespace and the element  
“non-cml-element” is not part of this

```
</non-cml-element>
</molecule>
```

This document is *not* valid because the null prefix is bound to the CML namespace and the element non-cml-element does not appear in the schema which defines CML.

4.

```
<cml:molecule xmlns:cml = “http://www.xml-cml.org/schema“>
<cml:non-cml-element>
```

This is invalid. The cml prefix is bound  
to the CML namespace and the element  
“non-cml-element” does not form part of this

```
</cml:non-cml-element>
</cml:molecule>
```

This document is *not* valid because the cml prefix is bound to the CML namespace and the element non-cml-element does not appear in the schema which defines CML.

### 38.3.6 CMLValidator report language

Because we have taken a unit-test-based approach the initial design of our convention verification software used Schematron, an ISO Standard for testing assertions about the structure of XML documents. After initial testing we found that Schematron scaled poorly with the complexity of the rules and was difficult to debug. We also desired a report language that could better support partial validation required to reflect the MUST, SHOULD, MAY approach to defining rules adopted by CMLLite.

The validating rules are now expressed directly as XSLT which gives greater flexibility and control structure. To support the MUST, SHOULD, MAY style of rules we have developed a small report language (Figure 4) to indicate the different levels of severity.

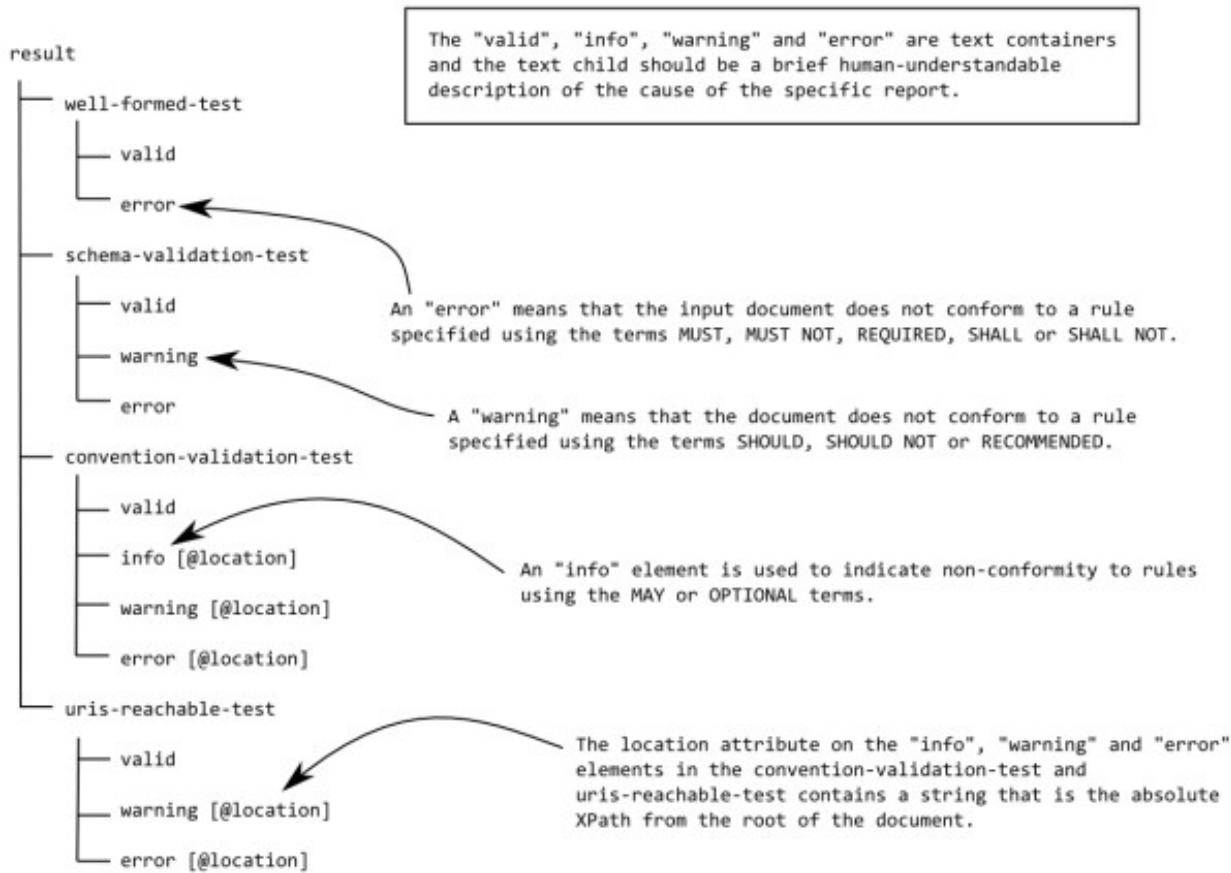


Figure 38.4: Figure 4. An outline of the CML report language

**An outline of the CML report language.** If a test (e.g. well-formed-test or URIs-reachable-test) contains a valid element child then it MUST NOT contain any warning or error element children. There is no such restriction on info elements and these may occur for input documents that otherwise conform completely to the convention.

Figure 5 shows how we use XSLT to encode a typical rule in the molecular convention:

- An atom must have an id attribute.
- The value of the id of an atom must be unique within the eldest containing molecule.

The XPath expression

`count(ancestor::cml:molecule//cml:atom[@id = current()]/@id) > 1`

can be decomposed into a set of steps which define a set of elements to query over and the query itself:

```

<xsl:template match="cml:atom" mode="molecular">
  <xsl:choose>
    <xsl:when test="@id">
      <xsl:if test="count(ancestor::cml:molecule//cml:atom[@id = current()/@id]) > 1">
        <report:error>
          the ids of atoms MUST be unique within the eldest containing molecule ...
        </report:error>
      </xsl:if>
    </xsl:when>
    <xsl:otherwise>
      <report:error>
        atoms MUST have an id attribute
      </report:error>
    </xsl:otherwise>
  </xsl:choose>
</xsl:template>

```

Figure 38.5: Figure 5. Example rules expressed in XSLT: an atom must have an id attribute and the value of the id must be unique among the ids of all the atoms in the eldest containing parent molecule

**Example rules expressed in XSLT: an atom must have an id attribute and the value of the id must be unique among the ids of all the atoms in the eldest containing parent molecule.** The cml prefix is bound to <http://www.xml-cml.org/schema> and the report prefix is bound to <http://www.xml-cml.org/report/>. The error reporting has been simplified for clarity (the location attribute is omitted).

- ancestor::cml:molecule selects any molecule element of which the current atom is a descendent (child, grandchild etc.).
- //cml:atom then selects every single atom element that is a descendent of any of the set of molecules. Note that this must by definition include the original atom.
- [@id = current()/@id] restricts the set of atoms to only include those that have an id that is identical to the original atom (matched in the template).
- The count(...) > 1 expression forms the query and evaluates the number of atoms left in the set. If this is greater than 1 then multiple atoms in the same ancestor molecule have the same id.

The conventions in CMLLite have built-in rules which are generally not explicitly stated in the specification of conventions:

- A convention is applied through an element carrying the convention attribute. The convention applies to that element and all its descendants.
- The value of the convention attribute MUST be a QName that expands to the URI identifying the convention to be applied.
- A convention can require other conventions which must be explicitly specified on appropriate elements.
- If no conventions are declared a warning is issued.

We do not intend conventions to replace the CML schema and they are not a general schema language.

CMLValidator uses normal XSLT processing rules but makes special use of the mode attribute to allow validation of different conventions within the same document. < apply-templates mode = “mode-name” > limits subsequent validation to templates with mode = “mode-name”. An apply-templates call without a mode will only call those templates without a mode (*i.e.* not governed by a convention in the document).

### 38.3.7 An example-simpleUnit

The current conventions contain many hundreds of validatorTest and to illustrate them we create a very simple sub-convention: simpleUnit. There is already a mature convention for units using the schema3 elements unitList and unit. (Schema3 also defines a variety of attributes on unit which are still relevant but as they have default schema3 semantics they do not need explicit redefinition.) simpleUnit explores a small portion of this.

### 38.3.8 The ruleset

- 1 The simpleUnit convention is specified with the <http://www.xml-cml.org/convention/simpleUnit> namespace.
- 2 The simpleUnit convention MUST be specified on a cml:unitList element using the convention attribute.
- 3 A cml:unitList element MUST contain at least one cml:unit child element.
- 4 A cml:unitList element MAY contain other child elements from the CML namespace or from foreign namespaces.

There are no constraints on where in a document the unitList element may appear.

### 38.3.9 ValidatorTest

We start by creating an exhaustive set of tests against which the validator will be developed. These tests (Appendix B) are independent of the actual implementation of the validator. We can be confident that any validator that passes all these tests is likely to be useful in determining whether any of a wide range of documents is valid or invalid against the simpleUnit convention.

### 38.3.10 Validator

Figure 6 shows the XSLT required to encode the ruleset of the simpleUnit convention.

We now address the purpose of each of the templates in the validator in detail.

**Template 1:** creates the root report:result element

**Template 2:** match = “`*|@*|text()`” matches any element, attribute or text node when not in simpleUnit mode. The match expressions for the three node types are the most general possible and will therefore be overridden by any more specific matches. This template takes no action but allows recursive traversal to find elements covered by the simpleUnit conventions arbitrarily deep in the input document.

**Template 3:** carries out the same operations as template 2 but only when in simpleUnit mode. Non-CML elements may be interspersed with CML in the text document and will not cause the validator to emit warnings.

**Template 4:** Only elements from the cml namespace will be matched; the element MUST have a convention attribute, with namespace <http://www.xml-cml.org/convention/> and the local name simpleUnit. The schema enforces that the value of the convention attribute must be a namespaceRefType.

If the element matched is cml:unitList this triggers mode = “simpleUnit” which remains in scope for all descendants.

If the element matched is not cml:unitList the validator informs the users that it is an error to specify the simpleUnit convention and apply-templates is called but not in simpleUnit mode.

**Template 5:** matches any cml:unitList element. If this does not have at least one cml:unit child element then an error is reported. Any child nodes are then processed in simpleUnit mode.

**Template 6:** matches any cml:unit element in simpleUnit mode. XSLT rules dictate that it has higher priority than template 7.

```

<xsl:stylesheet version="2.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
    xmlns:cml="http://www.xml-cml.org/schema" xmlns:report="http://www.xml-cml.org/report/">

    <!-- template 1 -->
    <xsl:template match="/">
        <report:result>
            <xsl:apply-templates/>
        </report:result>
    </xsl:template>

    <!-- template 2 -->
    <xsl:template match="*|@*|text()">
        <xsl:apply-templates/>
    </xsl:template>

    <!-- template 3 -->
    <xsl:template match="*|@*|text()" mode="simpleUnit">
        <xsl:apply-templates mode="simpleUnit"/>
    </xsl:template>

    <!-- template 4 -->
    <xsl:template match="cml:[@convention and namespace-uri-for-prefix(
        substring-before(@convention, ':'),.) = 'http://www.xml-cml.org/convention/' 
        and substring-after(@convention, ':') = 'simpleUnit']">
        <xsl:choose>
            <xsl:when test="self::cml:unitList">
                <xsl:call-template name="simpleUnit-template"/>
            </xsl:when>
            <xsl:otherwise>
                <report:error>
                    the only valid cml element that can specify the simpleUnit
                    convention is "unitList"
                </report:error>
                <xsl:apply-templates/>
            </xsl:otherwise>
        </xsl:choose>
    </xsl:template>

    <!-- template 5 -->
    <xsl:template match="cml:unitList" mode="simpleUnit"
        name="simpleUnit-template">
        <xsl:if test="not(cml:unit)">
            <report:error>
                A unit list MUST contain child cml:unit elements
            </report:error>
        </xsl:if>
        <xsl:apply-templates mode="simpleUnit"/>
    </xsl:template>

    <!-- template 6 -->
    <xsl:template match="cml:unit" mode="simpleUnit">
        <xsl:apply-templates mode="simpleUnit"/>
    </xsl:template>

    <!-- template 7 -->
    <xsl:template match="cml:*" mode="simpleUnit">
        <report:info>
            <xsl:value-of select="local-name()"/>
            is not a part of the http://www.xml-cml.org/convention/simpleUnit
            convention and may be ignored by some processors.
        </report:info>
    </xsl:template>

```

**Template 7:** matches any element from the CML namespace in simpleUnit mode. The match is more specific than template 3 but less specific than templates 4, 5 and 6. This will therefore catch any CML namespaced elements other than unitList and unit. The elements matched by this template are covered by rule 4-they are allowed but they are not really part of the convention, hence the output contains information to this effect.

This template is primarily for information, not errors-it is therefore appropriate to warn when CML elements might be ignored. Note that the presence of report:info elements in the report document does not mean that the input document is invalid.

## 38.4 Interaction and extension of conventions

Conventions are generally designed so that they can be mixed in a document, typically as discrete sections of a document (*i.e.* they do not overlap (instance 2 in Figure 7)). Thus the CMSpect convention does not involve molecular, and molecular does not involve CMSpect. The CMLValidator will engage the appropriate modes when processing each convention.

It is sometimes desirable to nest conventions (a subtree with one convention being found completely within a larger tree with a different convention-*e.g.* instances 3 and 4 in Figure 7). We use this approach in the current CMLLite conventions (Table 1) which may (recursively) validate subtrees labelled as having known conventions. The rules for nesting are under community review and Figure 7 shows the currently allowed interaction of conventions. The scope of a convention is thus similar to that of a namespace in that it “extends from the beginning of the start-tag in which it appears to the end of the corresponding end-tag” <sup>39</sup>.

Some of the specifications from the molecular convention <http://www.xml-cml.org/convention/molecular> are given below;

- A molecule MUST contain at least one of the following elements: molecule, atomArray, name, label, formula.
- A molecule MUST NOT contain both a child molecule and a child atomArray.
- An atomArray MUST contain at least one atom.
- A molecule MAY contain zero or one bondArray children and a bondArray MUST contain at least one bond.
- Every atom MUST have an id which is unique within the eldest containing parent molecule.
- If an atom has an x3 coordinate it MUST also have y3 and z3 (and similarly if y3 or z3 are present).

The compchem convention <http://www.xml-cml.org/convention/compchem> has been developed as part of the Quixote project. It requires that the *initialization module* contains exactly one molecule and that all the atoms in this molecule MUST have three dimensional coordinates. Rather than create a new convention for molecules it was decided that these requirements were compatible with the molecular convention but required a tightening of some constraints.

Some of the rules from the compchem convention are shown below:

- There MUST be an initialization module which is a module element with a specific value of its dictRef attribute (cml:module/@dictRef = ‘compchem:initialization’ where the compchem prefix is bound to <http://xml-cml.org/dictionary/compchem/>).
- The initialization module MUST contain exactly one molecule.
- molecules MUST declare that they conform to the molecular convention by declaring this in the convention attribute (cml:molecule/@convention = ‘conventions:molecular’ where the conventions prefix is bound to <http://www.xml-cml.org/convention/>).
- The molecule in the initialization module is REQUIRED to have an atomArray child.
- All the atoms in the molecule in the initialization module MUST have three dimensional coordinates.

---

<sup>39</sup> Namespaces in XML

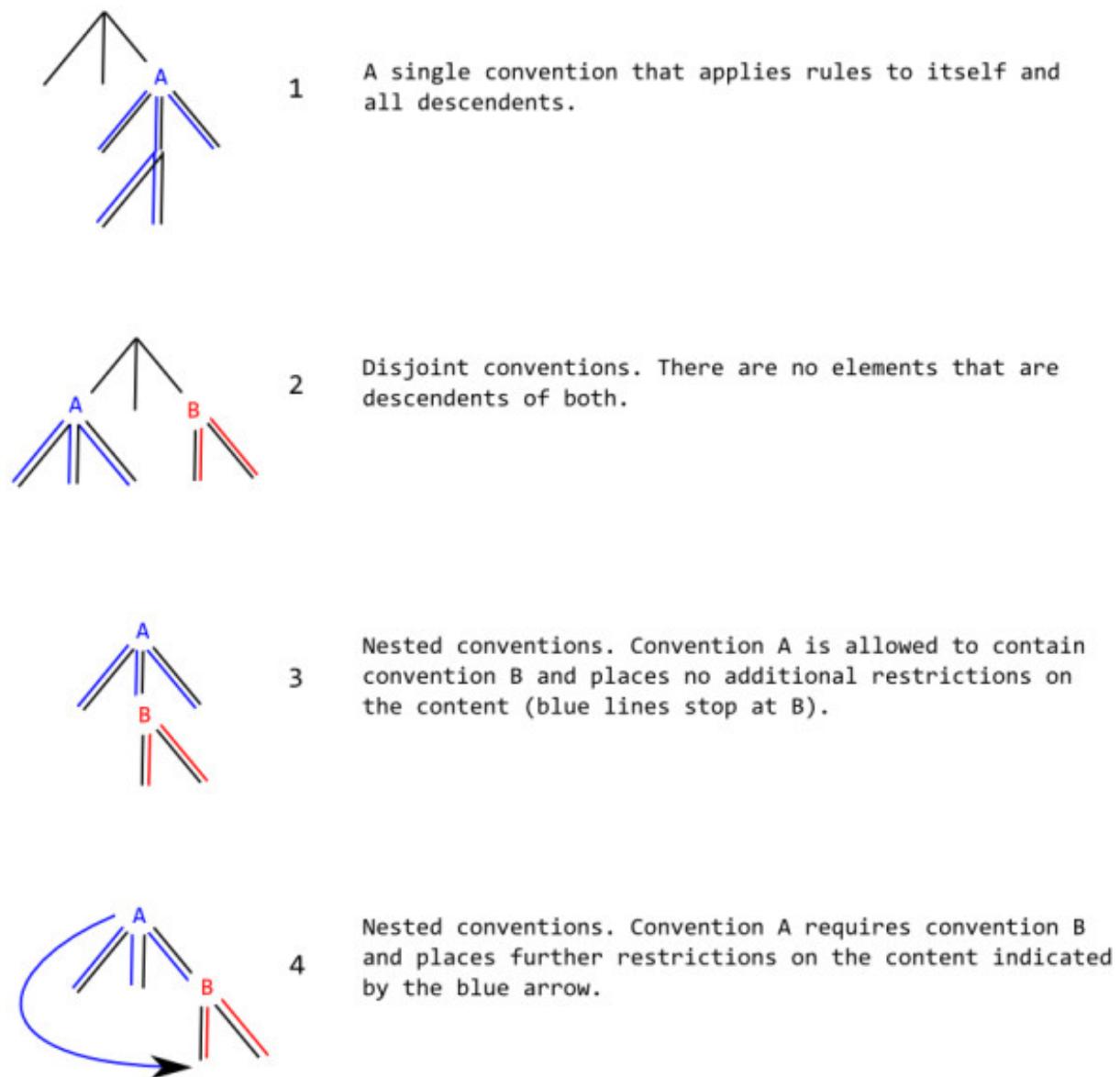


Figure 38.7: Figure 7. Documents with multiple conventions

**Documents with multiple conventions.** The black lines represent the XML tree (DOM) and are shadowed by constraints imposed by conventions A (blue) and B (red).

Figure 8 shows the part of the XSLT that will enforce the requirements on the molecule in the initialization module described above. The first template tests that there is only a single molecule child of the initialization module and that the molecule must specify the molecular convention. The existence of this convention statement will trigger the CMLValidator to apply the relevant rules to the molecule and its descendant nodes.

The second template is in a separate mode (test-atoms-have-3d-coordinates) and tests that an atomArray is present and that all the child atoms of this have 3D coordinates. Note that we do not need to check whether or not the molecule has child molecules or other atomArray children because this will be done by the molecular convention.

### 38.4.1 Validation Service

Following the W3C validation tools<sup>40</sup> (specifically Unicorn<sup>41</sup>), we have created the CML validation service<sup>44</sup>. The validator is available in the following forms: an interactive form-based<sup>42</sup> webpage, a RESTful<sup>43</sup> web service and as a Java library.

The Java library is the same as the backend engine for the web-based services. The program consists of validator classes, an overall workflow control class and a ValidationReport class. The validation report class encapsulates both an XML document containing information about which tests have passed, failed or caused warnings and a ValidationResult property. The ValidationResult can be VALID, VALID\_WITH\_WARNINGS or INVALID.

The checks performed by the validator are shown below in order of application. If a particular check results in an INVALID ValidationResult no further processing is performed and the ValidationReport is created and returned.

1. It is well-formed XML. The control class can takes as input either an *InputStream* or a *nu.xom.Document* (xomDoc) and produces a ValidationReport. If input is an *InputStream* the program checks that it is well-formed XML (this is not necessary for a xomDoc as it is necessarily well-formed). A xomDoc is built from the *InputStream* and further processing is identical regardless of input format.
2. The xomDoc conforms to the CML schema3.
3. Deprecated constructs are not used. The use of deprecated constructs will give a VALID\_WITH\_WARNINGS result.
4. Any conventions specified in the document are obeyed.
5. All the prefixes used in namespaceRefTypes (effectively QNames) have been bound to namespaces and are resolvable URLs.

The final check has been put in place as a reminder to users that sharing information is preferable and they can only “code to the green bar”<sup>37</sup> by making their dictionaries and conventions *etc.* publicly available. The workflow is shown in Figure 9 and 10.

The RESTful webservice implementation is accessed by POSTing the XML/CML document to <http://validator.xml-cml.org/validate> which returns a ValidationReport. This must then be queried by the user to determine whether the overall validation resulted in VALID, VALID\_WITH\_WARNINGS or INVALID. Informal feedback from users indicated that it was more useful to send the complete ValidationReport rather than just a ValidationResult as feedback as this would allow the calling tool to do more.

The website is effectively an instance of a tool that uses the RESTful implementation to do the actual validation but then interprets the results and displays them in the most human-user friendly fashion. Figures 11, 12 and 13 show the interactive form-based service in use.

---

<sup>40</sup> W3C Quality Assurance Tools

<sup>41</sup> Unicorn-W3C’s Unified Validator

<sup>42</sup> HTML4 Recommendation-Forms

<sup>43</sup> Architectural Styles and the Design of Network-based Software Architectures

```

<xsl:template match="cml:module[@dictRef and namespace-uri-for-prefix(
    substring-before(@dictRef, ':'),.) = 'http://www.xml-cml.org/dictionary/compchem'
    and substring-after(@dictRef, ':') = 'initialization']" mode="compchem">
    <xsl:choose>
        <xsl:when test="count(cml:molecule) = 1">
            <xsl:choose>
                <xsl:when test="cml:molecule[@convention and namespace-uri-for-prefix(
                    substring-before(@convention, ':'),.) = 'http://www.xml-cml.org/convention/'
                    and substring-after(@convention, ':') = 'molecular']">
                    <xsl:apply-templates mode="test-atoms-have-3d-coordinates" />
                </xsl:when>
                <xsl:otherwise>
                    <report:error>
                        The molecule in the initialization module MUST conform to the
                        molecular convention
                    </report:error>
                </xsl:otherwise>
            </xsl:choose>
        </xsl:when>
        <xsl:otherwise>
            <report:error>
                The initialization module MUST contain exactly one molecule child
            </report:error>
        </xsl:otherwise>
    </xsl:choose>
</xsl:template>

<xsl:template match="cml:molecule" mode="test-atoms-have-3d-coordinates">
    <xsl:choose>
        <xsl:when test="cml:atomArray">
            <xsl:for-each select="cml:atomArray/cml:atom">
                <xsl:if test="not(@x3) or not(@y3) or not(@z3)">
                    <report:error>
                        An atom MUST have x3, y3 and z3 attributes
                    </report:error>
                </xsl:if>
            </xsl:for-each>
        </xsl:when>
        <xsl:otherwise>
            <report:error>
                It is REQUIRED that the molecule in the initialization
                module has an atomArray child.
            </report:error>
        </xsl:otherwise>
    </xsl:choose>
</xsl:template>

```

Figure 38.8: Figure 8. A snippet showing how the compchem convention can rely on the molecular convention and add further restrictions

A snippet showing how the compchem convention can rely on the molecular convention and add further restrictions.

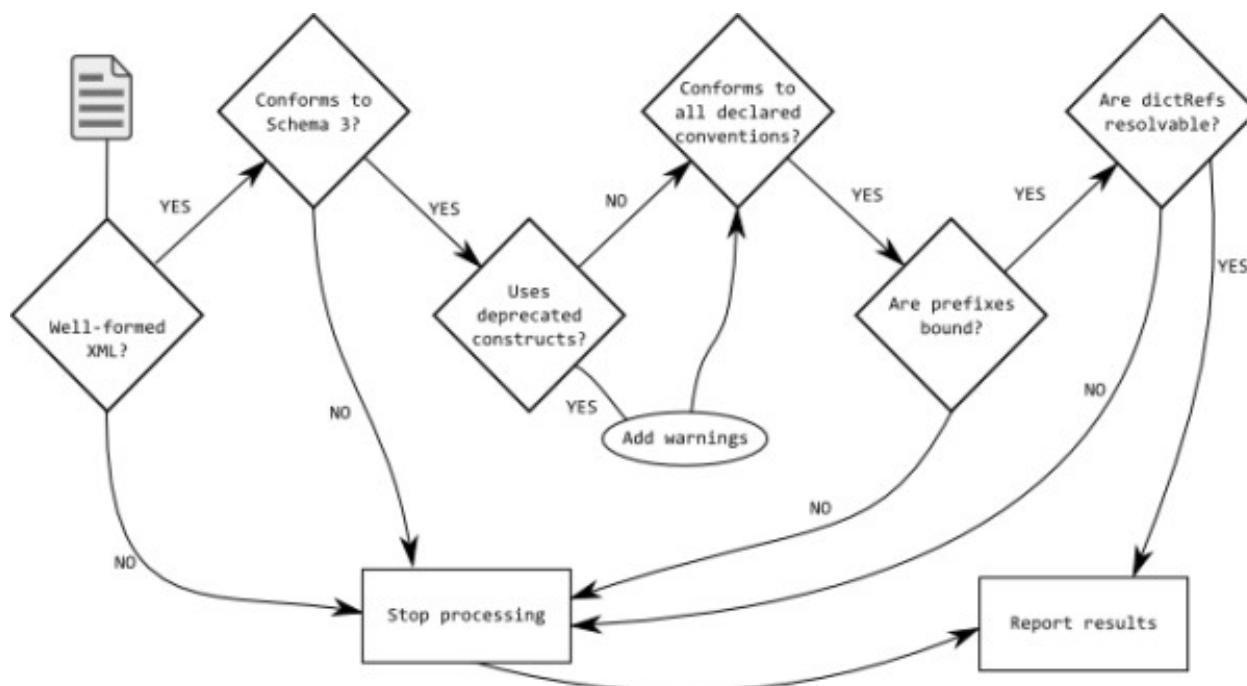


Figure 38.9: Figure 9. Workflow of validation in the CMLValidator  
**Workflow of validation in the CMLValidator.**

## 38.5 Conclusions

We have developed an approach to extensible semantics for Chemical Markup Language, where we assume that the current schema (schema3) is stable and expressive. There is enough software and data that this approach has been widely deployed and tested, even if it is not yet mainstream. Semantics are defined in the XSD schema, with additional natural language and validated using a unit test approach (Java and .NET). It works in the main fields of chemistry for which CML has been developed (molecules, reactions/syntheses, crystallography/solid-state, spectroscopy and computational chemistry). The approach encourages sub-communities in chemistry to create *conventions* which can be as rigid or fluid as they wish. The conventions can be rigorously unit tested using CMLValidator.

The convention-based approach is intermediate between natural language and formal systems. It relies, in part, on the wider community agreeing the semantics in schema3 (in several years deployment we have not yet had any disagreement with the basic elements and attributes and unit-tested examples). Sub-communities are starting to build their extensions of which the compchem convention being developed by the Quixote project is a prime example.

We believe the convention-based approach will help developers to create better software quicker. The tests/conventions define clear, testable APIs and these are essential for any distributed development.

The system interoperates fully with RDF-based systems. Many elements (especially value containers) can be algorithmically translated to RDF. A few core elements (primarily molecule, spectra, crystal) can be held in a more atomic form with bespoke semantics and software (it is, however, always possible to map into the details of these using URIs and to provide fine-grained links). By using this mixture of approaches we believe this is a cost-effective approach to interoperability within chemistry for those who wish to interoperate.

The screenshot shows the CML Validator webform interface. At the top, there is a navigation bar with 'CML' and 'Validator' followed by 'Input' (which is highlighted in orange). Below the navigation bar, a message states: 'This validator checks the document you paste in below to determine if ...'. A numbered list follows: 1. it is well-formed XML, 2. it conforms to the CML Schema 3, 3. it conforms to any conventions specified in the document (click [here](#) for a list of conventions currently checked), 4. all the prefixes used in dictRefs have been bound, 5. all the dictRef values are URLs. Below this, another message says: 'The validator will also give warnings if deprecated constructs are used (e.g. orders should not use numeric values)'. A large text area labeled 'Enter the CML to validate' contains the following XML code:

```

<unitlist xmlns="http://www.xml-cml.org/schema" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dummyDictionary="http://www.xml-cml.org/dictionary/" xmlns:convention="http://www.xml-cml.org/convention/" xmlns:xhtml="http://www.w3.org/1999/xhtml" convention="convention:unit-dictionary" namespace="http://www.xml-cml.org/dictionary/dummy/" title="example units dictionary">
    <description>
        <xhtml:p>A dummy units dictionary.</xhtml:p>
    </description>
    <unit
        title="unit 1"
        symbol="unit1"
        parentSI="dummyDictionary:dummy"
        multiplierToSI="1"
        constantToSI="0"
        unityType="dummyDictionary:dummy">
        <dc:source>a dummy source</dc:source>
        <definition>
            <xhtml:p>

```

At the bottom right of the text area is a 'Submit Query' button. Below the text area, there are links for 'Contact Us' and 'Blog'. On the far right, there is a vertical scroll bar. At the very bottom right of the page is a copyright notice: '© CMLC 2010'.

Figure 38.10: Figure 10. The CMLValidator webform interface  
**The CMLValidator webform interface.** The input claims that it should conform to the unit-dictionary convention but unit 1 does not have an id attribute.

The screenshot shows the CML Validator results page. At the top, there is a navigation bar with 'CML' and 'Validator' followed by 'Result' (which is highlighted in orange). Below the navigation bar, the word 'INVALID' is displayed in a large red header. Underneath, a red banner displays the text '1 Error found'. A bulleted list follows: • **Convention conformance** - a unit must have an id. Below this, an XPath expression is shown: 'Found at: /\*[local-name()='unitList' and namespace-uri()='http://www.xml-cml.org/schema'][1]/\*[local-name()='unit' and namespace-uri()='http://www.xml-cml.org/schema'][1]'.

Figure 38.11: Figure 11. Part of the CMLValidator results page showing that the input (in Figure 10) was invalid  
**Part of the CMLValidator results page showing that the input (in Figure 10) was invalid.** The report contains a human-understandable message and an XPath (machine-understandable) expression giving the location of the error.

The screenshot shows a navigation bar with 'CML', 'Validator', and 'Result'. The 'Result' tab is active and highlighted in orange. Below the navigation bar, a large orange header bar displays the text 'VALID WITH WARNINGS' in white. Underneath this, a section titled '1 Warning found' lists a single warning: 'Convention conformance - The dictionary element SHOULD have a title attribute intended for human-readability'. A detailed XPath expression is provided: 'Found at: /\*[local-name()='cml' and namespace-uri()='http://www.xml-cml.org/schema'][1]/\*[local-name()='dictionary' and namespace-uri()='http://www.xml-cml.org/schema'][1]'.

Figure 38.12: Figure 12. Part of the CMLValidator results page showing that the input is valid but has warnings *i.e.* **Part of the CMLValidator results page showing that the input is valid but has warnings**. The convention states that the dictionary element SHOULD have a title but none was found in the document. The warning gives a human-understandable message and an XPath (machine-understandable) expression giving the location of the warning.

The screenshot shows a navigation bar with 'CML', 'Validator', and 'Result'. The 'Result' tab is active and highlighted in green. Below the navigation bar, a large green header bar displays the text 'VALID' in white. Underneath this, a section titled 'All checks passed' lists two items: 'xml is well formed' and 'document conforms to the schema'. Below this, a note says 'The following information may be useful' followed by a list of checks: 'molecule is not a part of the http://www.xml-cml.org/convention/unit-dictionary convention'. A detailed XPath expression is provided: 'Found at: /\*[local-name()='unitList' and namespace-uri()='http://www.xml-cml.org/schema'][1]/\*[local-name()='molecule' and namespace-uri()='http://www.xml-cml.org/schema'][1]'.

Figure 38.13: Figure 13. Part of the CMLValidator results page showing that the document is valid and which checks have been performed

**Part of the CMLValidator results page showing that the document is valid and which checks have been performed.** Further information is also given because the input document contained a molecule element (which is not part of the unit-dictionary convention).

## 38.6 Availability of Code

The CMLValidator and associated tests are available at <http://bitbucket.org/cml/cmllite-validator-code> and the web-based implementation is available at <http://bitbucket.org/cml/cmllite-validator-ws> both under an Apache 2.0 licence.

## 38.7 Competing interests

The authors declare that they have no competing interests.

## 38.8 Authors' contributions

JAT developed the CMLLite approach, created the CMLValidator and the test corpus, and wrote the manuscript. PMR is the original author of CML, created the test corpus and wrote the manuscript.

## 38.9 Appendix A

The attributes in CML are defined in attributeGroups which must have unique names allowing them to be disambiguated within the schema. The attributeGroup defines an attribute, its datatype/allowed values, and the name of the attribute in the document (these do not have to be unique).

Element declarations in the schema specify which attributeGroups are allowed on them (which in this case caused polymorphism).

Table 2 shows all the attributes in CML schema3 that appear in the document as type. Values in “quotes” are enumerated allowed values, xsd:string means that any string content is permitted.

## 38.10 Appendix B

The documents below are a subset of the documents used to test the behaviour of the simpleUnit convention validator. After every alteration (new test, bug fix *etc.*) of the convention, the validator is run against this test set to verify that its behaviour still conforms to expectations.

In all the examples below the `cml` prefix is bound to <http://www.xml-cml.org/schema> and the `conventions` prefix is bound to <http://www.xml-cml.org/convention/>.

### 38.10.1 Documents Valid Against the simpleUnit Convention

The input documents below should all result in a ValidationResult of VALID and the ValidationReport MUST NOT contain info elements. The tests in this section are primarily concerned with ensuring that the convention is recognised wherever it appears in a document and that non-CML elements do not give rise to info reports.

1.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unit/>
</cml:unitList>
```

This produces the following ValidationReport:

```
<report xmlns = “http://www.xml-cml.org/report/“>
<well-formed-test>
<valid>xml is well formed</valid>
</well-formed-test>
<schema-validation-test>
<valid>document conforms to the schema</valid>
</schema-validation-test>
<convention-validation-test>
<valid>document conforms to all the conventions specified</valid>
</convention-validation-test>
<uris-reachable-test>
<valid>All appropriate URIs were reachable</valid>
<valid>all dictRefs are resolvable </valid>
</uris-reachable-test>
</report>
```

The subsequent inputs produce exactly the same report and we therefore choose to explain what the document is testing for rather than show the output.

2.

```
<x:p xmlns:x = “http://www.w3.org/1999/xhtml“>
<cml:unitList convention = “conventions:simpleUnit”>
<cml:unit/>
</cml:unitList>
</x:p>
```

Tests that simpleUnit convention can be declared on a unitList that is not the root element of the document and is a child of a foreign namespaced element.

3.

```
<cml:module>
<cml:unitList convention = “conventions:simpleUnit”>
<cml:unit/>
</cml:unitList>
</cml:module>
```

The simpleUnit convention can be declared on a unitList that is not the root element of the document and is a child of a CML element (module).

4.

```
<element-in-default-namespace>
<cml:unitList convention = “conventions:simpleUnit”>
<cml:unit/>
```

```
</cml:unitList>
</element-in-default-namespace>
```

Tests that simpleUnit convention can be declared on a unitList that is not the root element of the document and is a child of an element from the default-namespace.

22.

```
<x:p xmlns:x = “http://www.w3.org/1999/xhtml“>
the unitList need not be the root element
<cml:unitList convention = “conventions:simpleUnit”>
<cml:unit/>
</cml:unitList>
</x:p>
```

Test that although CML does not have a mixed content model the unitList can occur within non-CML mixed content.

6.

```
<x:p xmlns:x = “http://www.w3.org/1999/xhtml“>
there are multiple instances of the simpleUnit
convention in this document
<cml:unitList convention = “conventions:simpleUnit”>
<cml:unit/>
</cml:unitList>
<cml:unitList convention = “conventions:simpleUnit”>
<cml:unit/>
</cml:unitList>
</x:p>
```

Test that there can be multiple disjoint simpleUnit convention declarations in the same document.

7.

```
<cml:unitList convention = “conventions:simpleUnit”>
<cml:unit/>
<x:p xmlns:x = “http://www.w3.org/1999/xhtml“>
non cml child-this is fine
</x:p>
</cml:unitList>
```

Tests that the unitList element can contain foreign namespaced elements without giving rise to info reports.

8.

```
<cml:unitList convention = “conventions:simpleUnit”>
<element-in-default-namespace/>
<cml:unit/>
```

```
</cml:unitList>
```

Tests that the unitList element can contain elements from the default namespace without giving rise to info reports.

9.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unit>
<x:p xmlns:x = "http://www.w3.org/1999/xhtml">
non cml child-this is fine
</x:p>
</cml:unit>
</cml:unitList>
```

Tests that the unit element can contain foreign namespaced elements without giving rise to info reports.

24.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unit>
<element-in-default-namespace/>
</cml:unit>
</cml:unitList>
```

Tests that the unit element can contain elements from the default namespace without giving rise to info reports.

11.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unit/>
<cml:unit>
<x:p xmlns:x = "http://www.w3.org/1999/xhtml">
multiple cml:unit elements are allowed
</x:p>
</cml:unit>
</cml:unitList>
```

Tests that a unitList may contain more than one unit child.

### 38.10.2 Documents Valid (with info reports) Against the simpleUnit Convention

The input documents below should all result in a ValidationResult of VALID and the ValidationReport should contain a single info element and MUST NOT contain either error or warning elements. info elements are used to give information about rules in a convention involving the MAY clause. Note that the complete ValidationReport is given for the first example but subsequent examples only contain the error message for brevity.

1.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unit/>
<cml:molecule/>
</cml:unitList>
```

Produces:

```
<report xmlns = "http://www.xml-cml.org/report/">
<well-formed-test>
<valid> xml is well formed </valid>
</well-formed-test>
<schema-validation-test>
<valid> document conforms to the schema </valid>
</schema-validation-test>
<convention-validation-test>
<info location = "/*[local-name() = 'unitList' and namespace-uri() = 'http://www.xml-cml.org/schema'] /*[local-name() = 'molecule' and namespace-uri() = 'http://www.xml-cml.org/schema'] ">
```

molecule is not a part of the <http://www.xml-cml.org/convention/simpleUnit> convention and may be ignored by some processors.

```
</info>
<valid>
document conforms to all the conventions specified
</valid>
</convention-validation-test>
<uris-reachable-test>
<valid> All appropriate URIs were reachable </valid>
<valid> all dictRefs are resolvable </valid>
</uris-reachable-test>
</report>
```

2.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unit>
<cml:atom/>
</cml:unit>
</cml:unitList>
```

Produces: “atom is not a part of the <http://www.xml-cml.org/convention/simpleUnit> convention and may be ignored by some processors”

3.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unit/>
<cml:unit>
<cml:bond/>
</cml:unit>
</cml:unitList>
```

Produces: “bond is not a part of the <http://www.xml-cml.org/convention/simpleUnit> convention and may be ignored by some processors”

Note that if the bond specified a numeric bond order (e.g. order = ‘1’) the test result would be VALID\_WITH\_WARNINGS because numeric bond orders are deprecated.

4.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unit>
<x:p xmlns:x = "http://www.w3.org/1999/xhtml">
this is still going to be processed in
unitList mode.
<cml:bond/>
</x:p>
</cml:unit>
</cml:unitList>
```

Produces: “bond is not a part of the <http://www.xml-cml.org/convention/simpleUnit> convention and may be ignored by some processors”

22.

```
<x:p xmlns:x = "http://www.w3.org/1999/xhtml">
the unitList need not be the root element
<cml:unitList convention = "conventions:simpleUnit">
<cml:molecule/>
<cml:unit/>
</cml:unitList>
</x:p>
```

Produces: “molecule is not a part of the <http://www.xml-cml.org/convention/simpleUnit> convention and may be ignored by some processors”.

### 38.10.3 Documents Invalid Against the simpleUnit Convention

The documents below should all result in a ValidationResult of INVALID. The ValidationReport should contain a single error element and should not contain either info or warning elements.

1.

```
<cml:molecule convention = "conventions:simpleUnit">
<cml:unitList>
<cml:unit/>
</cml:unitList>
</cml:molecule>
```

Produces the following ValidationReport:

```
<report xmlns = "http://www.xml-cml.org/report/">
<well-formed-test>
<valid> xml is well formed </valid>
</well-formed-test>
<schema-validation-test>
<valid> document conforms to the schema </valid>
</schema-validation-test>
<convention-validation-test>
<error location = "/*[local-name() = 'molecule' and namespace-uri() = 'http://www.xml-cml.org/schema'] [#B1]_@*[local-name() = 'convention' and namespace-uri() = ']">
the only valid cml element that can specify the simpleUnit convention is "unitList"
</error>
</convention-validation-test>
<uris-reachable-test>
<valid> All appropriate URIs were reachable </valid>
<valid> all dictRefs are resolvable </valid>
</uris-reachable-test>
</report>
```

2.

```
<x:p xmlns:x = "http://www.w3.org/1999/xhtml">
<cml:molecule convention = "conventions:simpleUnit">
<cml:unitList>
<cml:unit/>
</cml:unitList>
</cml:molecule>
</x:p>
```

Produces: “the only valid cml element that can specify the simpleUnit convention is ‘unitList’”. (Illustrating that the document is still being correctly traversed.)

3.

```
<cml:unitList convention = "conventions:simpleUnit"/>
```

Produces: “A unit list MUST contain child cml:unit elements”.

4.

```
<cml:unitList convention = "conventions:simpleUnit">
<!-- not valid, a unitList must have at least one unit child -->
</cml:unitList>
```

Produces: “A unit list MUST contain child cml:unit elements”.

22.

```
<cml:unitList convention = "conventions:simpleUnit">
<x:p xmlns:x = "http://www.w3.org/1999/xhtml">
no unit child of unitList
</x:p>
</cml:unitList>
```

Produces: “A unit list MUST contain child cml:unit elements”.

6.

```
<cml:unitList convention = "conventions:simpleUnit">
<x:p xmlns:x = "http://www.w3.org/1999/xhtml">
<cml:unit/>
This unit is not a direct child of unitList
and therefore should cause an error.
</x:p>
</cml:unitList>
```

Produces: “A unit list MUST contain child cml:unit elements”.

7.

```
<cml:unitList convention = "conventions:simpleUnit">
<cml:unitList>
<cml:unit>
<x:p xmlns:x = "http://www.w3.org/1999/xhtml">
the outer unitList does not have at least
one unit child
</x:p>
</cml:unit>
</cml:unitList>
</cml:unitList>
```

Produces: “A unit list MUST contain child cml:unit elements”.

## 38.11 Acknowledgements

We thank Microsoft Research for a grant for Chem4Word and EPSRC (Pathways to Impact) for dissemination. We also heartily thank Charlotte Bolton for all her help in preparing this manuscript.



# MINING CHEMICAL INFORMATION FROM OPEN PATENTS

## 39.1 Abstract

Linked Open Data presents an opportunity to vastly improve the quality of science in all fields by increasing the availability and usability of the data upon which it is based. In the chemical field, there is a huge amount of information available in the published literature, the vast majority of which is not available in machine-understandable formats. PatentEye, a prototype system for the extraction and semantification of chemical reactions from the patent literature has been implemented and is discussed. A total of 4444 reactions were extracted from 667 patent documents that comprised 10 weeks' worth of publications from the European Patent Office (EPO), with a precision of 78% and recall of 64% with regards to determining the identity and amount of reactants employed and an accuracy of 92% with regards to product identification. NMR spectra reported as product characterisation data are additionally captured.

## 39.2 Background

The enormous increase in the output of scientific data in recent times now requires radical changes in the way in which it is handled. The CAplus database<sup>1</sup> holds more than 32 million references to patents and journal articles and indexes more than 1500 current journals on a weekly basis, while the CAS REGISTRY<sup>2</sup> holds more than 54 million chemical compounds and the CASREACT<sup>3</sup> database more than 39 million single and multi-step reactions. Such resources are created by a labour-intensive process of manual curation with the consequence that a researcher must pay to access them, and the data themselves become a valuable commercial entity. By necessity, this is *closed data*.

The availability of data is vital for data-driven science such as spectra prediction and Quantitative Structure-Activity Relationship (QSAR) modelling, which has become increasingly important to the pharmaceutical industry as it seeks to control the spiralling costs of drug development. *Open data* - data that is freely available to the community-supports and enables such work. The more the culture of Open data spreads, the more such work becomes viable.

This use of Open data for research, though powerful, is not the end of the story. Tim Berners-Lee first described the concept of the Semantic Web<sup>4</sup>. The idea is simple-the World Wide Web comprises a vast collection of information, but information that is largely meaningless to a computer. If it were to be made machine-understandable, then software agents could be developed that would be able to use this information as a basis for reasoning and to make decisions. This concept, tied to that of Open data, would allow for computerised scientists conducting their own data-driven research and reporting their conclusions back to humans. The concept of a machine performing research is not one for the

---

<sup>1</sup> CAS Databases-CAPlus, Journal and Patent References

<sup>2</sup> CAS REGISTRY-The gold standard for substance information

<sup>3</sup> CAS Databases-CASREACT, Chemical Reactions

<sup>4</sup> The Semantic Web

world of science fiction-indeed, the robot scientist Adam has conducted its own hypothesis-driven research, reaching conclusions that were later validated by human researchers<sup>5</sup>.

In order to make our information machine-understandable, it is necessary to formalise the semantics of the medium in which it is stored. For the semantic web, such formalisation is typically performed by encoding the data using eXtensible Markup Language (XML). The *de facto* standard XML dialect for chemistry is Chemical Markup Language (CML) <sup>6</sup><sup>7</sup><sup>8</sup><sup>9</sup><sup>10</sup>. By rendering chemical information machine-understandable, CML allows for the creation of systems that integrate data of a variety of types and from a variety of sources to perform novel research-a semantic web for chemistry. Datuments<sup>11</sup><sup>12</sup>, hyperdocuments for transmitting ‘complete’ information including content and behavior, can record and reproduce experiments and act as a lossless way of publishing science. Conventional publication paths discourage the full publication of the scientific record-the process itself militates against datuments. Although there is no technical reason for the separation of ‘full-text’ and ‘supporting information’, the author is required to recast their information into models that conform to the publisher’s technology and business model. A common feature of all mainstream science publication is the universal destruction of high-quality information. Spectra, graphs, etc., are semantically rich but are either never published or must be reduced to an emasculated chunk of linear text to fit the paper model. But now we have the technology to address this. Machine-understandability requires both ontological (meaning) and semantic (behaviour) support, and XML is now mature enough that this is possible. Many information components in a datument can be recast as context-free XML and integrated with XML text and XML graphics. Some publishers are actively embracing enhancements to journal articles (see e.g. the RSC’s Project Prospect<sup>13</sup><sup>14</sup><sup>15</sup>), and the Chemistry Add-in for Microsoft Word (sometimes referred to as Chem4Word)<sup>16</sup> supports the authoring of chemical datuments using one of the world’s most popular word processing packages. The evolution of ‘(hyper)activated’ journal articles is discussed in a further article in this issue<sup>17</sup>.

In the absence of author or publisher-led markup, one way in which semantic data collections can be created is through the application of text mining software to the available literature. Chemistry-specific text mining software has been under continuous development at the Unilever Centre over the past decade. A suite of tools have been developed and released, including the named entity recognition tool OSCAR<sup>18</sup><sup>19</sup><sup>20</sup>, the syntactic analysis tool ChemicalTagger<sup>21</sup><sup>22</sup> and the name-to-structure conversion tool OPSIN<sup>23</sup><sup>24</sup>. The availability of these mature, Open packages allow for the large-scale extraction of chemical data from the published literature.

During this work, we have particularly concentrated on chemical texts which share a common style and vocabulary. The most frequently published chemical “chunks” occur in records of chemical synthesis in journal articles, lab books, theses, reports and chemical patents. Of these, legal and contractual restrictions forbid our text-mining of most scientific articles, while lab books and theses are disorganised and difficult to find, even in institutional repositories. We have therefore developed our chemical reaction text mining on the corpus of public patent data. It is a built-in feature of the patent process that the contents of a patent must be published and Openly available after the appropriate period, so in this sense it is an excellent corpus.

A project at the Chemical Abstracts Service (CAS) in the 1980s aimed to produce a system capable of automating or

<sup>5</sup> The Automation of Science

<sup>6</sup> Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles

<sup>7</sup> Chemical Markup, XML, and the World-Wide Web. 2. Information Objects and the CMLDOM

<sup>8</sup> Chemical Markup, XML, and the World-Wide Web. 3. Towards a Signed Semantic Chemical Web of Trust

<sup>9</sup> Chemical Markup, XML, and the Worldwide Web. 4. CML Schema

<sup>10</sup> Chemical Markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators

<sup>11</sup> The Next Big Thing: From Hypermedia to Datuments

<sup>12</sup> (Hyper)activating the chemistry journal

<sup>13</sup> Computers learn chemistry

<sup>14</sup> Project Prospect

<sup>15</sup> Semantic enrichment of journal articles using chemical named entity recognition

<sup>16</sup> Chemistry Add-in for Word

<sup>17</sup> The past, present and future of scientific discourse

<sup>18</sup> High-Throughput Identification of Chemistry in Life Science Texts

<sup>19</sup> Cascaded classifiers for confidence-based chemical named entity recognition

<sup>20</sup> OSCAR4: a flexible architecture for chemical text-mining

<sup>21</sup> ChemicalTagger: A tool for semantic text-mining in chemistry

<sup>22</sup> ChemicalTagger Demonstration

<sup>23</sup> Chemical name to structure: OPSIN, an open source solution

<sup>24</sup> OPSIN: Open Parser for Systematic IUPAC Nomenclature

partially automating the indexing process by application of Natural Language Processing (NLP) technologies<sup>252627</sup>. This system was claimed to “satisfactorily” process 36 out of 40 synthetic paragraphs from the Journal of Organic Chemistry<sup>24</sup> and to produce “usable results” for 80-90% of simple synthesis paragraphs and 60-70% of complex paragraphs<sup>26</sup>, where complex paragraphs are defined as describing general procedures, instances of general procedures, analogous syntheses and parallel syntheses. The size of the corpus used to produce this second set of results was not given, nor in either case was the procedure used for corpus creation. Accordingly, it is not possible to regard this area as a solved problem.

The era in which the aforementioned technology was developed was very different. As a division of the American Chemical Society, CAS was in the privileged position of having access to a large body of published work in an electronic format. The situation today is different—the ubiquity of electronic publication and explosion of the scale of publication has granted such access far more widely, though publishers may very well supply the works subject to restrictive terms of use. The chemical patents used in the current work, however, are subject to no such restrictions and so the time for a re-examination of the subject of automated extraction of chemical reactions has come.

The automated extraction of reaction information from the literature will prove highly useful to, for example, the EPSRC’s “Dial-a-Molecule” grand challenge, which aims to make the synthesis of a novel compound a quick and efficient process that can be completed in days, not years. The automated prediction of synthetic pathways will require an appropriate reaction database which is not currently available. We estimate that around 10 million syntheses per year are currently published in the literature, and so text-mining is an obvious means by which such information can be obtained.

The approach applied here is similar to that of CrystalEye<sup>28</sup>, where we have built tools that retrieve and extract crystallographic data from public sources. This activity has now generated about 250,000 datasets and runs essentially automatically every night. There is no technical reason why an Open patent service should not run in the same way, downloading the incremental updates on the sites at appropriate intervals according to the publishing schedule of the patent organisation in question. The main difference between these activities is that the crystallographic data is already in quasi-semantic form (*i.e.* CIF) and the process is completely algorithmic. With patents there is a variability due to the different styles of natural language and approaches to document layout taken by applicants, and the different technologies used to create the patent itself. However, in practice, most chemical patents have a very closely-defined structure and style of presentation.

### 39.3 Current systems for automatic analysis of patents

Referees have asked us to comment on the current approaches of Chemical Abstracts and other commercial organizations. Since this is a competitive area it is likely that precise methodology is a trade secret. When one of us (PM-R) heard CAS present at ACS in 2009, the presentation showed that patent analysis was through experts annotating patents with handwriting. PM-R asked about machine methods and was not given a public indication that they were significant. The major semi-public organization is the Fraunhofer Institute which develops its own in-house methodology (which includes OPSIN and OSCAR3<sup>29</sup>) and publishes accounts of some of its work.

The motivation of the current work is to explore the semantic structure of patents and parts of the semantic chemical information. We have deliberately not used information which would require OCR of text and to this extent we are likely to show an improvement over the documents published in the last century. We have also not addressed Markush structures as there are other publications describing these and some early commercial applications<sup>3031</sup>. The primary

<sup>25</sup> Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistics Techniques. 1. Lexical and Syntactic Phases

<sup>26</sup> Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistics Techniques. 2. Semantic Phase

<sup>27</sup> Extraction of Chemical Reaction Information from Primary Journal Text

<sup>28</sup> CrystalEye

<sup>29</sup> OSCAR3

<sup>30</sup> InfoChem, ChemProspector

<sup>31</sup> Markush structure reconstruction: A prototype for their reconstruction from image and text into a searchable, context sensitive grammar based extension of SMILES

emphasis has been to show that chemical reactions can be turned into semantic form with an acceptable success rate. This article outlines the steps and likely success of others wishing to enter this area.

## 39.4 PatentEye

The liberation of scientific data and its conversion to machine-understandable forms holds great promise. A key part of the chemical sciences are the reactions that chemists perform and report in great number, and the goal of PatentEye is to demonstrate the potential to create an automated system capable of extracting reactions from the literature, creating machine-understandable representations using CML and sharing them as Open Data. This system is presented as a proof-of-concept, not as a sustainable resource. To increase the reliability of the extracted syntheses, PatentEye attempts to validate the identified product molecules. This validation is achieved by comparison of a candidate product molecule with any accompanying structure diagram using the package OSRA<sup>32333435</sup> for image interpretation and with any accompanying NMR and mass spectra, using the OSCAR3 data recognition functionality. The identified NMR spectra are considered to be valuable data in their own right and are extracted and retained for use in later works.

### 39.4.1 Patent documents

Patents are made available on the World Wide Web by a number of patent offices. For legal reasons, they are frequently published as image-based facsimile reproduction of the original document, and are commonly also available as recovered, free-text documents. While the World Intellectual Property Organisation (WIPO) publishes such documents in HTML with minimal markup indicating the position of document sections and headings, both the European Patent Office (EPO) and United States Patent and Trademark Office (USPTO) employ XML formats in which major sections and heading titles are explicitly delimited. The XML formats used by the USPTO and EPO are similar though not identical, and reflect the structure of a patent as agreed by the Common Application Format (CAF)<sup>36</sup>. While only EPO documents were used in the current work, much of the methodology employed is applicable to alternative document sources. In particular, USPTO documents are available for bulk download via Google patents<sup>37</sup> and present an attractive target for text mining. At the time of writing the documents available for download date from 1976 to the present day, and are claimed to number approximately 7 million, across all subjects.

CAF, agreed in 2007 by the EPO, USPTO and Japan Patent Office (JPO), is intended to “simplify and streamline application filing requirements in each Office to allow applicants to prepare a single application in the common application format for acceptance in each of the three Offices”<sup>32</sup>. It mandates the section titles, and their ordering, that are to be used in patent applications. These are shown in Figure 1, in which those titles shown in bold indicate titles that must be included, and those shown in both bold and parentheses must be included where corresponding information is present.

#### Anatomy of a patent and tractability for linguistic tools

Patents are generally large documents, often running to several hundred pages in length. For that reason automated analysis tools are potentially extremely valuable in rapidly exploring their content. Chemical patents are remarkable in that they not only form a large subject domain within the patent literature, but also in that certain sections exhibit a high degree of similarity across the field, particularly for those that discuss the synthesis and properties of organic molecules. This homogeneity makes them very tractable to linguistic analysis.

The ‘Summary of Invention’ section is often very long and formulaic. In chemical patents, the subjects of the invention are generally presented in the form of Markush structures, generic chemical structures typically defined by a specific

<sup>32</sup> Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution

<sup>33</sup> Extracting Chemical Structure Information: Optical Structure Recognition Application

<sup>34</sup> Improvements in Optical Structure Recognition Application. In Ninth IAPR International Workshop on

<sup>35</sup> OSRA: Optical Structure Recognition Application

<sup>36</sup> Common Application Format, United States Patent and Trademark Office

<sup>37</sup> USPTO Bulk Downloads: Patent Grant Full Text

**Description**

**Title of Invention or Title**

**Technical Field or Field**

**Background Art or Background**

**Summary of Invention or Summary**

    Technical Problem

    Solution to Problem

    Advantageous Effects of Invention

**(Brief Description of Drawings)**

**Description of Embodiments**

    Examples

    Industrial Applicability

    Reference Signs List

    Reference to Deposited Biological Material

**(Sequence Listing Free Text)**

Citation List

    Patent Literature

    Non Patent Literature

**Claims**

**Abstract**

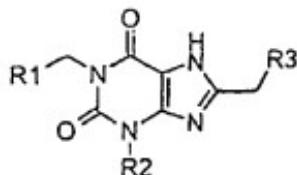
**(Drawings)**

**(Sequence Listing)**

Figure 39.1: Figure 1. The standardised patent heading titles, as mandated by the Common Application Format  
The standardised patent heading titles, as mandated by the Common Application Format.

scaffold bearing a number of variable substituent groups, such as that shown in Figure 2. At present, this is too complex for analysis, except for localised sections where OSCAR and OPSIN can recognise catalogues of substituent groups.

[0006] The present invention is directed to a compound of formula



wherein

R<sup>1</sup> is selected from the group consisting of straight chain or branched chain alkenyl having from 2 to 7 carbon atoms, straight chain or branched chain alkynyl having from 2 to 7 carbon atoms, straight chain or branched chain alkenyl having from 2 to 7 carbon atoms substituted by

Figure 39.2: Figure 2. Typical usage of Markush structures in chemical patents

**Typical usage of Markush structures in chemical patents**<sup>38</sup>. Definitions of pseudoatoms (*e.g.* “R1” are frequently several pages in length and are commonly iterative.

The examples of the invention that are required to be presented in the description section typically consist of reports of the synthesis of specific compounds that correspond to one of these Markush structures. Such reports appear very much as they would in other parts of the chemical literature such as journal articles and theses, and sometimes, though not always, are accompanied by chemical structure diagrams or characterisation data. An example of such a report is shown in Figure 3. For many patents, these reports can be automatically interpreted with a high degree of precision and recall, and this task represents the major body of work reported in this paper.

Many reactions are described as small variants on a common theme, and so full detail is omitted from the patent document. Typical formulations used for this purpose include “following the procedure for...” or “prepared as in example X...”. The challenges to automatically interpreting such examples are to identify the archetypal reaction from the linguistic form and to determine which components of the reaction have been changed in the synthesis. We have made significant progress in interpreter and resolver for this type of language, and in a limited number of cases we have shown that it is possible to not only follow the back-references but replace the chemical structures in context. The process involves a large number of steps and the technology is currently insufficiently mature to be considered a production system.

It is worth noting that identifying sections of the document is not trivial because different applicants use different terminology and often do not announce major sections with the accepted phraseology. Therefore we rely heavily on linguistic processing to determine where sections in the patent begin and end. The patent is also relatively ‘flat’ in that the humans marking up the patent are only required to identify paragraphs and not subject sections, though some of the high-level document structure illustrated in is explicitly defined in the EPO’s XML patent documents. The content of these files is governed by a Document Type Definition (DTD) file that can be downloaded from the EPO website<sup>39</sup>. The root element of the XML documents is *e.g.* patent search reports.

The abstract element can be composed either of an

The Sub-DocuMENT for BIbliographic (SDOBI) data uses proprietary tags to encode a wealth of metadata related to the patent, *e.g.* the tag

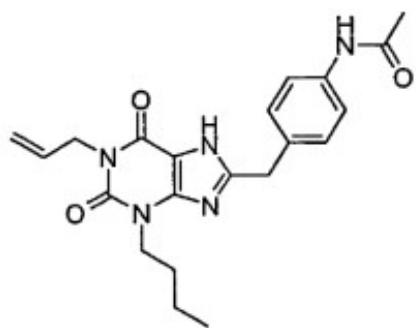
By convention, each document contains three

The

<sup>39</sup> EBD ST.36 (XML) DATA INFORMATION

Step 6: *N*-[4-(1-Allyl-3-butyl-2,6-dioxo-2,3,6,7-tetrahydro1*H*-purin-8-ylmethyl)-phenyl]-acetamide.

[0128]



[0129] This compound was prepared by a method similar to that reported by Müller et al. in Synthesis 1995, 1295. 2-(4-Acetylaminophenyl)-*N*-(3-allyl-6-amino-1-butyl-2,4-dioxo-1,2,3,4-tetrahydro-pyrimidin-5-yl)-acetamide (310 mg) was taken up in 3N aqueous sodium hydroxide (11 mL) and methanol (11 mL) added until the solid had dissolved. The resulting solution was heated in a 50 °C oil bath until TLC indicated the reaction to be complete. The solution was cooled and acidified with 6N aqueous hydrochloric acid until acidic to pH paper, a precipitate formed at pH 6. The resulting mixture was extracted with chloroform. Extracts were dried and concentrated to give a yellow solid (190 mg). <sup>1</sup>H NMR (DMSO-d6) 0.87 (t, 3H), 1.26 (m, 2H), 1.61 (m, 2H), 2.00 (s, 3H), 3.94 (t, 2H), 3.97 (s, 2H), 4.44 (br d, 2H), 5.06 (m, 2H), 5.82 (m, 1H), 7.17 (d, 2H), 7.47 (d, 2H), 9.89 (s, 1H) and 13.4 (s, 1H). MS, m/z(M+)=395.1947.

Figure 39.3: Figure 3. Typical synthesis report <sup>34</sup>

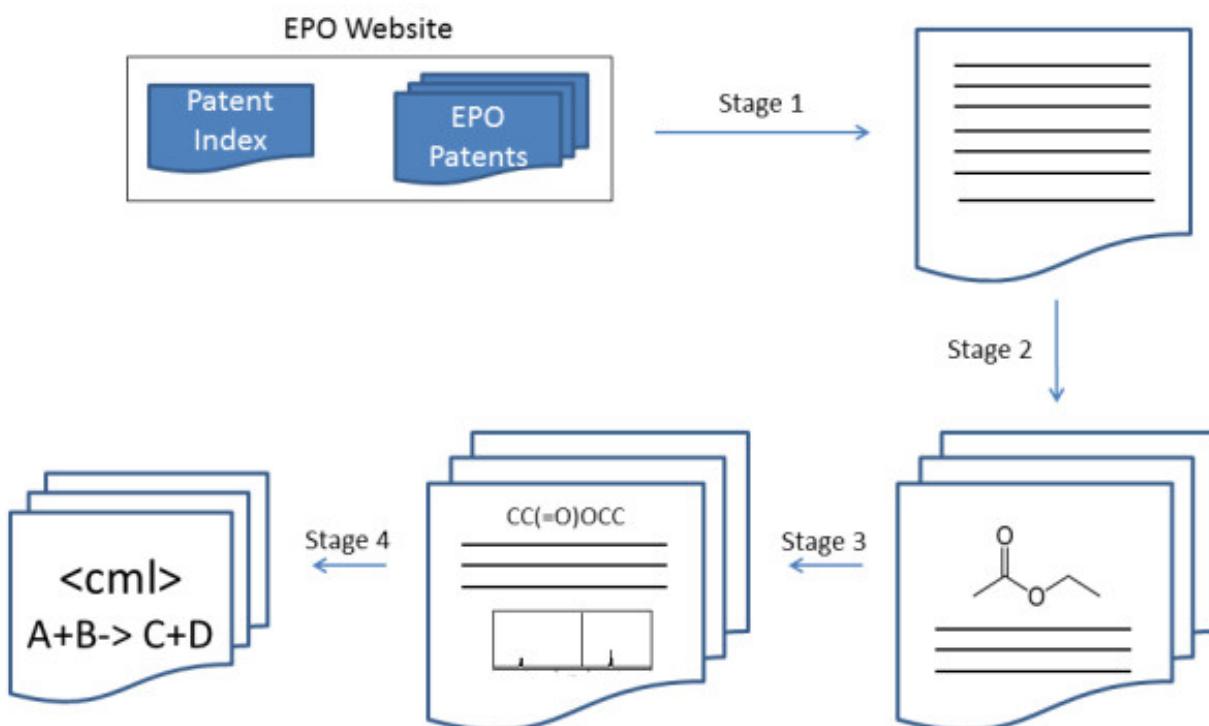
**Typical synthesis report** <sup>34</sup>. The numbers in square brackets indicate sequential numbering of each of the paragraphs in the document.

The e.g.

The XML patent documents are available for download from the EPO website as part of a ZIP package that also contains the PDF facsimile representation and individual TIFF image files that contain the individual figures from the document. The names of these files are numbered sequentially to give identifiers that are referenced in the XML patent document, to indicate which figure occurs at which position in the document.

### 39.4.2 PatentEye workflow

The implemented system is automated to the degree that it is capable of operating with minimal user interaction, and consequently the PatentEye workflow consists of a number of stages of processing. First, chemical patents are identified within the online archive of the European Patent Office (EPO) and are downloaded. The XML documents supplied by the EPO are then semantically enhanced so as to delimit sections and subsections of the text and to introduce additional metadata such as SMILES strings representing the content of structure diagrams and OSCAR3 data markup to describe identified spectra. Finally, reactions are extracted from these semantically enhanced documents using ChemicalTagger and are converted to CML. The overall workflow is depicted in Figure 4.



**Figure 39.4: Figure 4. Schematic workflow for extraction and interpretation of chemical reactions in patents**

**Schematic workflow for extraction and interpretation of chemical reactions in patents.** Stage 1 -the patent is identified and downloaded. Stage 2-the document is flattened and segmented. Stage 3-various tools (OPSIN, OSRA, OSCAR3) are used to identify key elements in the reaction and convert them to semantic form. Stage 4-ChemicalTagger is applied to the language of the chemical reaction to determine the roles and processes. Where successful, the extracted information is converted to reactions expressed in CML.

### 39.4.3 Automated identification and download of patents

The European Patent Office (EPO) publishes patent documents through the European Publication Server, hosted at <https://data.epo.org/publication-server/>. An interactive search using various parameters (a patent ID, a date range

within which to search and a list of document kinds) may be performed; alternatively, a weekly digest of patent index files is also provided <https://data.epo.org/publication-server/data-coverage>. These summaries include the International Patent Classification (IPC) codes assigned to each document. The IPC is a subject-based, hierarchical classification scheme describing the topics covered in a patent. This allows automatic identification of documents relevant to the current work, as listed in Table 1.

Once a list of relevant documents has been determined, PatentEye uses functionality provided by the CrystalEye webcrawler to interact with the EPO search interface. Where full-text XML is available for a relevant patent document, the corresponding ZIP file is retrieved. This is notably absent in the case of patents that have been published under the Patent Cooperation Treaty (PCT) instead of filed directly with the EPO.

In order to create a corpus for the current work, chemical patents from the EPO website for the ten weeks dated from 2009-05-06 to 2009-07-08 were downloaded. Duplicate patent documents were deleted such that only one document remained for each patent ID within the corpus, which then totalled 690 zipfiles. Of these 690 files, it was found that 23 did not contain the XML version of the patent under the expected file name. The subsequent work using the downloaded patent corpus is therefore based on a reduced corpus of 667 unique, full-text patent documents where the XML files are used as input.

#### 39.4.4 Enhancement of document semantics

As discussed previously, the different sections of the XML-formatted patent documents are not always clearly defined. The content of the 5.

```
<description>
  <p>...</p>
  <heading>Heading 1</heading>
  <p>...</p>
  <p>...</p>
  <heading>Heading 1.1</heading>
  <p>...</p>
  <heading>Heading 1.2</heading>
  <p>...</p>
  <heading>Heading 2</heading>
  <p>...</p>
  <p>...</p>
</description>
```

Figure 39.5: Figure 5. Flat document structure as received from the EPO  
**Flat document structure as received from the EPO.**

To a human reader, it is a simple task to realise that the headings 1.1 and 1.2 are subsections of Heading 1, and that the each of the paragraphs belongs to a section of the document that begins with the preceding heading. Since this is not made explicit in the structure of the XML, however, it is not trivially obvious to a machine that the document should be read in such a way. For this reason it is desirable to deflatten the XML-to rewrite the document such that as much of the implicit structure is made explicit as possible. This rewritten document is then saved to disk in order to prevent unnecessary repetition of the task.

A number of other semantic enhancements are performed on the patent documents at this stage. These tasks include the application of OSCAR3 data recognition to identify spectral data within the text, the application of OSRA to add SMILES representations of the chemical structure images contained within the documents, the recognition and annotation of references in the text to other sections of the document, *e.g.* “the reaction was performed as in example 12” and the identification and labelling of the paragraphs in the text that form part of an experimental section.

### Paragraph Deflattening

In this step, the 6.

```
<description>
  <p>...</p>
  <heading>Heading 1
    <p>...</p>
    <p>...</p>
  </heading>
  <heading>Heading 1.1
    <p>...</p>
  </heading>
  <heading>Heading 1.2
    <p>...</p>
  </heading>
  <heading>Heading 2
    <p>...</p>
    <p>...</p>
  </heading>
</description>
```

Figure 39.6: Figure 6. Reordered document showing explicit structure  
**Reordered document showing explicit structure.**

Before this reformatting, the 7.

### Document Segmentation

As previously discussed, the EPO do not attempt to explicitly demarcate in their XML the existence of sections of a patent document. Headings in the document are denoted by use of the 1 to permit slight variation on the author’s part. These headings are renamed (*e.g.* to “disclosureOfInvention” or “summaryOfInvention”) to enable trivial location of them within the document, and the child elements of *e.g.* “example 1” and “example 2”) are identified by finding those headings that have identical text content, disregarding incrementable strings (*e.g.* “1” and “1a”) and chemical names, as identified by OSCAR. The structure of the document is then rewritten to reflect the fact that a heading that intervenes in such a list is logical a subheading of the preceding heading. This process is illustrated in Figure 8.

```

<description>
  <p>...</p>
  <heading title='Heading 1'>
    <p>...</p>
    <p>...</p>
  </heading>
  <heading title='Heading 1.1'>
    <p>...</p>
  </heading>
  <heading title='Heading 1.2'>
    <p>...</p>
  </heading>
  <heading title='Heading 2'>
    <p>...</p>
    <p>...</p>
  </heading>
</description>

```

Figure 39.7: Figure 7. Reordered document showing explicit structure and avoiding mixed content  
**Reordered document showing explicit structure and avoiding mixed content.**

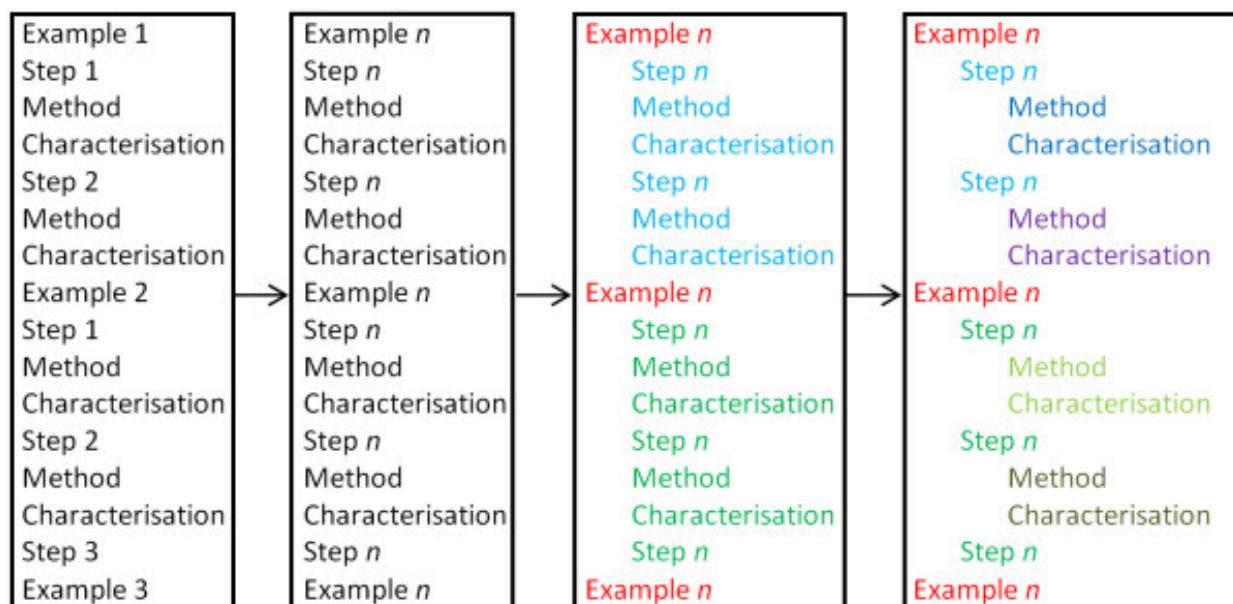


Figure 39.8: Figure 8. Identification of and Document Restructuring Using Consecutive Headings  
**Identification of and Document Restructuring Using Consecutive Headings.**

## Data Annotation

To facilitate its later use in the workflow, characterisation data is at this stage identified and annotated using the OSCAR3 data functionality. The text of each paragraph is passed to OSCAR3 for data recognition, which applies inline annotation to label the various parts of the spectrum. Where data is found, these annotations are inserted into the patent XML document in the appropriate places. In this way, the original text content of the patent document remains intact, and is rendered machine-understandable.

## Classification of Synthesis Sections

While it is common for the experimental sections, *i.e.* those that describe the process and results of a chemical reaction, of a patent to occur as examples of the invention, it is not necessarily the case that the method of identifying document sections described previously will result in their occurrence as part of an *experimental* or *non-experimental* by use of a naïve Bayesian classifier. This classification is achieved using the third-party Java library Classifier4J<sup>40</sup>, version 0.6., and allows for a greater proportion of the experimental sections within the patent corpus to be recognised as such and treated appropriately during the later stages of the workflow.

A corpus was assembled by selecting 800 *i.e.* paragraphs, in the most part) from those patents that had successfully passed through the paragraph deflattening and document segmentation phases of the semantic enrichment procedure, using a random process in which each paragraph had an equal chance of selection. These paragraphs were manually inspected and determined to be *experimental*, *non-experimental* or *empty* according to the following criteria;

- The paragraph is *empty* if it has no text content. Such empty paragraphs generally occur in the patent documents as containers for images.
- The paragraph is *experimental* if:
  1. It is an account of a reaction or a part of a reaction, including by way of reference to another section of text *e.g.* “The reaction was carried out as in example 12”.
  2. It is a report of spectral or other characterisation data.
  3. It is some combination of the above.
- The paragraph is *non-experimental* if it is not *empty* or *experimental*.

The manually-classified paragraphs may be summarised as follows (Table 2).

In order to produce experimental and non-experimental sets of equal size, non-experimental paragraphs after the 238<sup>th</sup> occurrence were ignored for the remainder of this work. The first 119 (50% of the full set) experimental and non-experimental paragraphs were then used to train the Bayesian classifier before it was asked to predict probabilities of the remaining experimental and non-experimental paragraphs belonging to the experimental class. The predicted likelihoods may be summarised as follows (Table 3).

Thus, when classifying paragraphs as experimental if  $p < 0.5$  and non-experimental if  $p > 0.5$ , the experimental paragraphs were correctly classified at a rate of 96.6% and the non-experimental paragraphs at a rate of 89.9%. These rates were deemed high enough to continue into production.

Heading elements in the patent documents are identified by use of the XPath “

## Image Analysis

As previously discussed, the EPO patents frequently feature chemical structure diagrams that illustrate the example compounds for which syntheses are reported. These images therefore contain useful information that can be used to identify the product of a reaction. While USPTO patents supply the connection tables for such structures in the form of ChemDraw and MOL files, in the case of the EPO patents it is necessary to use image-to-structure software to interpret

---

<sup>40</sup> Classifier4J

the supplied TIFF files. As the only such Openly available package at the time, OSRA was used for this task. As the most recent version available at the time that the work commenced, version 1.2.2 was employed. Applying OSRA to a chemical image resulted in a SMILES string, which was attached to the patent XML document as the value of an

In order to validate the performance of OSRA on the patent images, a corpus of two hundred images of single chemical structures was formed by random selection from the patent corpus. The chemical structures contained within these images were manually converted to SMILES strings, chiefly by redrawing the structure using ChemDraw 12.0 and exporting the structure as SMILES or by manual conversion in the case of simple structures, which were recorded in an index of the corpus. OSRA was used to analyse each of the 200 single chemical structure images, and the results of this analysis was appended to the index.

Previous authors in the field have suggested subjective metrics of success such as less than 30 seconds of human editing being required to correct errors in the structure<sup>41</sup>, while Filippov and Nicklaus<sup>28</sup> propose measuring success by calculating a similarity metric between the machine-produced structure and the correct structure. Such measures are of limited utility in the present work; manual correction of structures or determination of correct structures cannot be implemented within a fully automated workflow. What is desired of the image analysis process is the correct identification of the product molecules of chemical syntheses, and while a high similarity between a structure believed to be the product (the “candidate product”) and a structure produced by OSRA may be indicative that the image analysis has made a minor error and the candidate product should be accepted, it may equally indicate that the image analysis is correct and the candidate product should be rejected. As a result, there is no threshold of similarity below the two structures being identical at which the structure derived from the image analysis becomes “good enough”.

The manually-generated and OSRA-generated SMILES strings for each image were thus used to generate the canonical identifier InChI using JUMBO<sup>42</sup>. The performance of OSRA was measured by comparing these InChIs by string equivalence; where the two InChIs were identical, it was counted as OSRA having correctly deduced the chemical structure contained within the image and considered a match. Where the InChIs differed it was considered a non-match. In a number of cases, it was not possible to generate an InChI from the SMILES string produced by OSRA. The causes of these problems were also examined and determined to be primarily that the SMILES string contained the wildcard character, \*, which is valid SMILES but is not supported by the JUMBO SMILES parser. In a further two cases the SMILES string returned by OSRA was found not to be valid, suggesting a bug within the OSRA program itself.

The results from this work were as follows (Table 4).

The agreement between the OSRA-produced structure and the manually-produced structure is, at 34%, significantly lower than that reported for OSRA 1.1.0 by Filippov and Nicklaus<sup>43</sup>, in which the rate was reported as 26 matches out of 42 (61.9%) structures and 107 matches out of 215 (50.0%) structures on two data sets. Such rates will of course be highly dependent upon the images that form the test corpus, and the images supplied by the EPO are of highly variable quality. Many of the images that form the test corpus used in this work are severely pixelated, indistinct or contain background noise; some are only barely legible to a human skilled in the art. Such an example, together with the structure as interpreted by OSRA, is shown in Figure 9.

### 39.4.5 Extraction of reactions

Chemical patents are a rich source of chemical reactions due to the requirement for a patent claimant to detail examples of the invention. The reactions published in this way are routinely manually indexed and added to databases such as CASREACT. In order to devise a system for automated extraction from reported syntheses, it is important to first consider the nature and common structure of such text. Fortunately, the reporting of chemical syntheses is highly stylised. By convention, chemists report syntheses using the past tense and the agentless passive voice, which simplifies the process. Descriptions of syntheses may be conceptually divided into three parts—the *primary reaction*, in which the target compound is completely or substantially produced; the *work-up*, in which the reaction is quenched and neutralised, solvents are removed, the product purified and suchlike; and the *characterisation*, in which spectral

<sup>41</sup> Kekulé: OCR-Optical Chemical (Structure) Recognition

<sup>42</sup> JUMBO6

<sup>43</sup> Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution

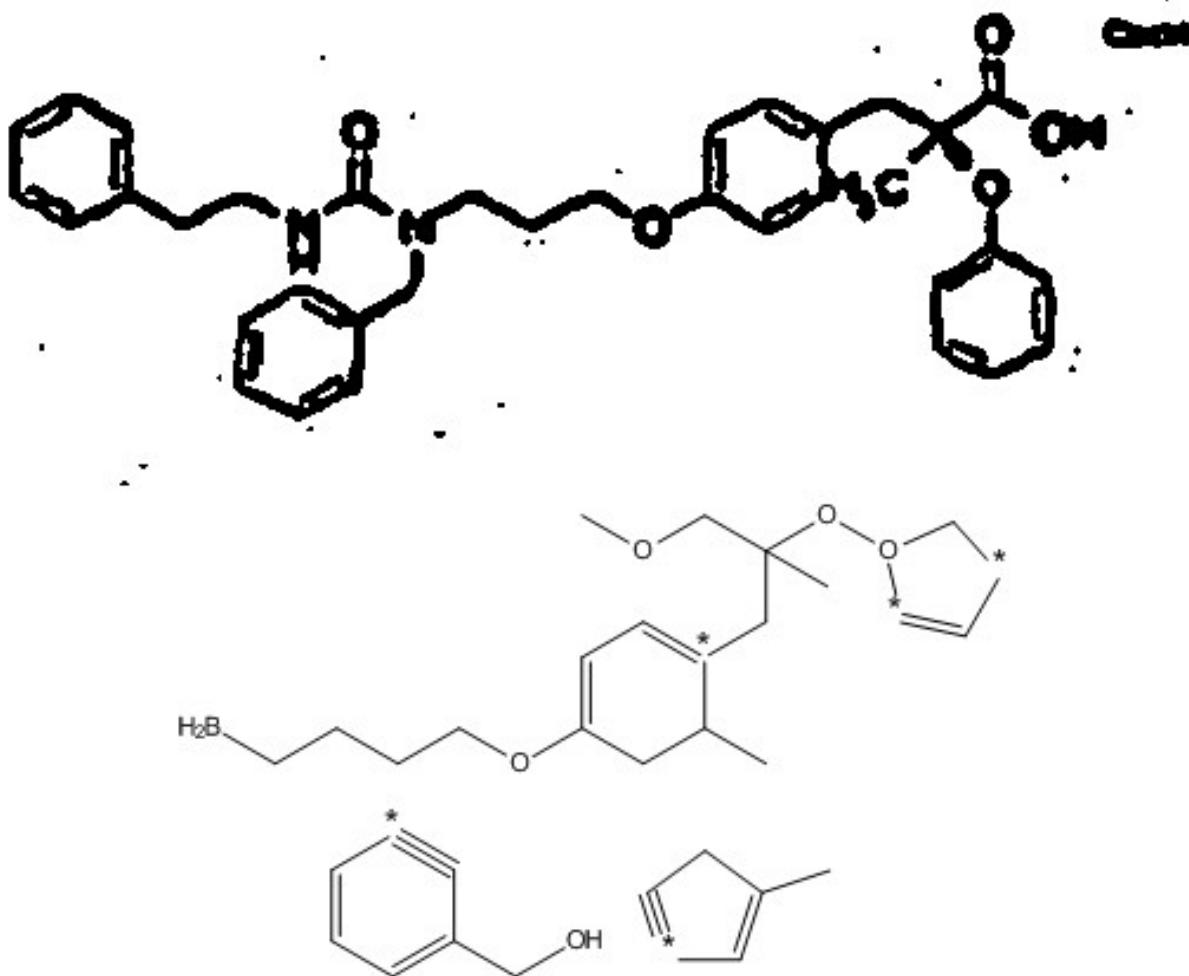


Figure 39.9: Figure 9. Input image (top) and unbuildable result (bottom)  
**Input image (top) and unbuildable result (bottom).**

data is afforded to demonstrate that the product of the reaction is that intended. In the description of the primary reaction, reactants (“a substance that is consumed during the course of a reaction”<sup>44</sup>) are detailed by giving a name or other reference (*e.g.* “ketone 12b” or “the compound from step 2”) together with the quantity used, generally stated by mass and by molar amount. Solvents are typically detailed by giving a name and the volume used. In the description of the work-up these quantities are commonly omitted. The identity of the product of the synthesis may be specified in one of two typical ways; in the heading of the section, or by statement at the end of the description of the work-up, *e.g.* “to yield 1,6-naphthyridine-8-carboxylic acid”.

The enhanced patent XML documents are read into memory, and the headings that have been classed as experimental by the ParagraphClassifier or that are descended from example headings are identified by means of XPath. The sections of the document either contained within the<sup>45</sup>. The product of the reaction is identified by using OSCAR3 to identify chemical names in the heading title. The product identity is then validated by comparison with the results of the OSRA analysis of any image present, and with any  $^1\text{H}$  NMR or mass spectrum that is reported. The results of these processes are combined into a CML Reaction which is saved to disk. This workflow is expanded in greater detail in the following sections and summarised graphically in Figure 10.

### Identification of reagents

Reagents used during the primary reaction section of a chemical synthesis are, by convention, reported along with the quantity used. Such lexical patterns are easily identified using ChemicalTagger.

### Identification of products

In order to identify the product of a reaction, the title text of the document section under examination is passed to OSCAR3 for named entity annotation. If OSCAR3 does not identify a single chemical name (CM) in the title text, then the process of reaction extraction fails and the ExperimentParser throws a RuntimeException. If a single CM is found in the title text, then the name is resolved to a CML Molecule, which is added to the

### Attachment of spectral data

The most common spectra types found in the patent corpus were  $^1\text{H}$  NMR,  $^{13}\text{C}$  NMR and mass spectra. The reports of mass spectra generally report only the mass of the molecular ion, optionally plus or minus a defined offset, and so provide a useful source of information for validating a candidate product molecule but little information worth preserving. The NMR spectra, however, in addition to providing a means by which the product molecule may be verified, are themselves data of potential importance and are worth preserving for future re-use. The format in which they are preserved in the enhanced patent XML documents, using inline annotation to identify features within the original patent text, is ideal in that context as it retains the original document text. It does not, however, enable trivial machine interpretation of the spectrum since it is not valid CML and tools do not exist for its easy manipulation. The OSCAR-annotated spectrum is therefore converted into a CML Spectrum by use of the OSCAR2CMLSpectConverter class in the JUMBO-Converters library. It does not attempt to perform any further text-mining on its input, instead relying entirely on the OSCAR3 annotations to fully identify features of interest such as peaks, integrals, multiplicities and coupling constants.

### Automated verification of product-checking against embedded images/mass spectrum/ $^1\text{H}$ -NMR

It is desirable to be able to automatically verify the product in some way. This can be achieved by comparing the determined product to the extracted spectral data and, if present, any accompanying chemical images. The process of acquiring these sources of information must also be regarded as potentially inaccurate, and so it is not possible to

---

<sup>44</sup> NOTITLE!

<sup>45</sup> JUMBO-Converters

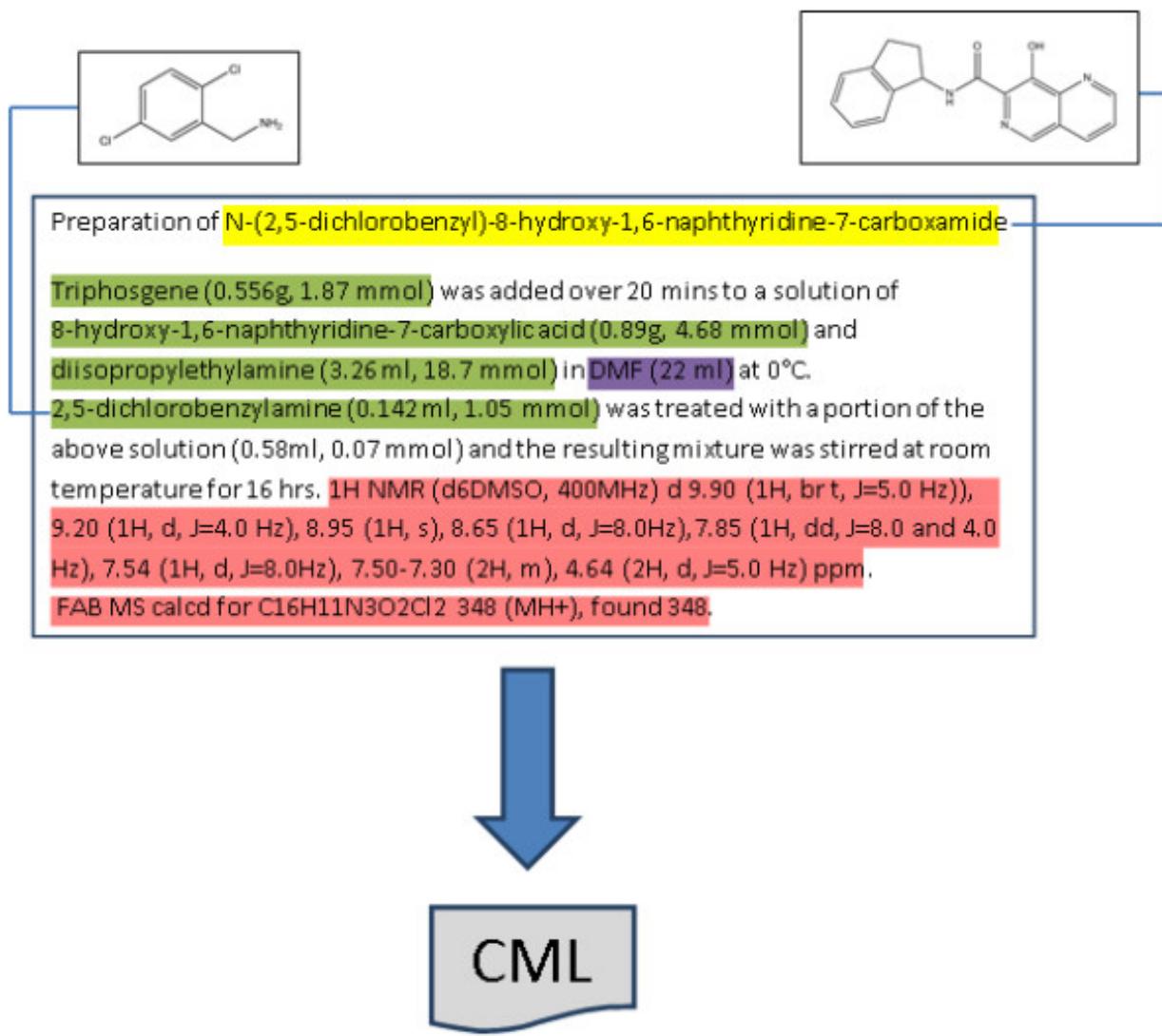


Figure 39.10: Figure 10. Schematic workflow diagram illustrating the extraction of reactions from the patent content. **Schematic workflow diagram illustrating the extraction of reactions from the patent content.** Text is analysed to identify the title compound (yellow), reagents (green), solvents (purple) and data (red), chemical names are used to construct connection tables and the reaction is saved as CML.

definitively confirm or refute any candidate product. Nonetheless, these checks provide potentially useful information regarding the validity of the assigned product and of the assigned spectral data.

Given the  $^1\text{H}$  NMR spectrum of an unknown compound, it is possible for one skilled in the art to discount certain candidate structures. Most trivially, the proton count in the candidate structure should agree with total integral of the NMR spectrum. Each unique chemical environment in the candidate structure should give rise to a distinct peak in the NMR spectrum and it should be possible to assign for each of the chemical environments a peak that is in the correct region of the NMR spectrum. The peak multiplicities should be explained by potential couplings in the candidate molecule, and protons that couple to one another should share coupling constants. The application of these rules is subject to a large amount of subtlety, however. While each chemical environment should give rise to an individual peak, these peaks can overlap and be indistinguishable from one another, most notably in the case of aromatic protons. The determination of unique chemical environments is complicated by the need to consider three dimensional effects, such as in the case of two protons sited on inequivalent faces of a ring system. As a result of these effects, it is not possible to compute chemical environments based solely on the 2D connectivity of a molecule and confidently assert that this will equal the number of peaks reported in the molecule's NMR spectrum. Whether there are more or fewer peaks in the NMR spectrum than predicted, the candidate molecule may be correct. Conversely, if the prediction matches the observation the candidate molecule may still be incorrect. The resolution of this problem falls outside of the scope of this project, and so the checking of structures against  $^1\text{H}$  NMR spectra is limited to the first method mentioned—ensuring that the proton count of the molecule agrees with the total integral of the spectrum. The result of this check is recorded in the automatically generated CML Reaction by adding a

When the experimental section includes a chemical image it is possible to compare the connection table of the candidate product with the results from the OSRA analysis of that image. If a chemical image is found within the source experimental section, the recorded SMILES strings for the image are built into CML Molecules which are then used to generate InChIs, using the SMILESTool and InChIGeneratorTool classes from JUMBO respectively. An InChI for the candidate product molecule is similarly generated, and the InChIs are subsequently compared.

Since the image included in the experimental section may be a reaction diagram it is possible for OSRA to have identified more than one molecule. Since the analysis of the often low-quality images is an error prone procedure, it is possible that the structures identified in the image may not contain the correct product. As previously discussed, when OSRA fails to correctly deduce a connection table from a drawn structure, it frequently reports a result containing the wildcard character, “\*”. This character is not recognised by JUMBO’s SMILESTool, causing it to throw an Exception. As a result, the following rules are applied when checking the candidate product against embedded images:

1. If the InChI generated for the candidate product matches one generated for the structures identified by OSRA, the product is considered to match the image.
2. If all of the structures identified by OSRA can be built into InChIs and the InChI for the candidate product does not match one of these, the product is considered not to match the image.
3. If some or all of the structures identified by OSRA cannot be built into InChIs and the InChI for the candidate product is not matched by one of those generated from the chemical image, no conclusion is drawn.

If a conclusion is drawn from this process, it is recorded in the CML Reaction by the addition of a

### 39.4.6 Performance of the Reaction Extraction Process

Using the methods described above, 26287 input sections for the reaction extraction procedure were derived from the corpus of 667 patent documents. From these inputs, a total of 4444 CML Reactions were derived, representing around 17% of the total input. The principle causes of failure to generate a CML Reaction included the input section containing no text-containing paragraph children; the input section containing more than one text-containing paragraph children (in which case the system backs off since this may describe a single or a multi-step reaction); the failure to identify the product molecule; and the failure to identify any reagent phrases in the source text.

To assess the accuracy of the semantified reactions, the output of the reaction extraction process was manually examined and compared with the source text. Each CML Reaction was assessed on a number of criteria to determine the

performance of the different modules of the reaction extractions system. These criteria included the accuracy of identified products, reagents and spectra, and the performance of the systems for automated product verification was tested by comparing the results of the automated verification with those of the manual verification. The methods employed for this process and the results obtained are subsequently discussed.

Since the manual inspection of each and every reaction extracted from the patent texts was not a feasible task, a subset was selected to serve as a corpus from which to derive performance metrics. From the 4444 reactions successfully extracted from the 667 unique, full-text patent documents, 100 reactions were selected at random. This reaction corpus was then used in the subsequently described validation procedures.

During the manual inspection of the reaction corpus, it was discovered that two of the 100 CML Reactions were derived from multi-step syntheses that were described within a single paragraph. Since these cases did not reflect the kind of input for which the current software was designed, they were excluded from the analysis process. A further two CML Reactions in the reaction corpus were found to have been derived from examples of their respective inventions that did not describe chemical syntheses-instead describing assays. These CML Reactions were similarly excluded from the analysis process; consequently, the process is based upon a reduced corpus of 96 CML Reactions.

The source from which the reaction was extracted was examined to determine whether the chemical name identified in the heading text by OSCAR3 and from which the product CML molecule was generated agreed with that stated in the heading text. Since the name to structure conversion process is not a perfect procedure, this is no guarantee that the attached connection table is also correct. However, the development of OPSIN was not a part of the current work and is reported to operate at an extremely high rate of performance<sup>22</sup> and so it was not considered necessary to measure the accuracy of this process. The manual inspection of the reaction corpus showed that the correct product was identified on 88 of the 96 occasions, a success rate of around 92%. It was further noted that on each of the 8 occasions on which the correct product was not identified, the term identified as the product name could not be successfully resolved to a connection table, suggesting a means by which the errors may be automatically removed. Generally, the cause of the failure to identify the correct product was due to the product of the reaction being named in the accompanying text, and hence not being present in the section heading of the source; instead, a term from the heading was falsely identified as a chemical name, which allowed for the creation of a CML Reaction from the source.

The sources from which the reaction corpus was extracted were examined, and for each the reagents employed and the amounts thereof were identified. These were then compared with those automatically extracted; instances where the same chemical name and amount were both manually identified and automatically extracted were counted as true positives, where the automatically extracted reagent list contained an instance that was not matched by both chemical name and amount in those manually identified a false positive was counted, and where a reagent was manually identified that was not automatically extracted, a false negative was counted.

This work required the formalisation of the concept of a reagent to a sufficient degree that any subjectivity in determining what did and did not constitute a reagent could as far as possible be minimised. The IUPAC definition, “a test substance that is added to a system in order to bring about a reaction or to see whether a reaction occurs”<sup>38</sup>, does not match the common usage of the term which further includes the chemical species involved in a reaction, *i.e.* reactants, solvents, catalysts, *etc.* It is this wider definition that fits the goal of the current work-to automatically determine how a reaction is carried out.

It was observed when considering this task that the chemical literature frequently underspecifies the work-up stage of a reaction. That is to say, the reagents employed may be stated without reference to their amounts, such as in;

*“The reaction mixture was stirred at 25°C for 4 days and then diluted with ethyl acetate. The mixture was then washed with a dilute aqueous hydrochloric acid solution. At this time, methanol was added to the organic layer. A precipitate formed and was removed by filtration. The organics were further washed with a saturated aqueous sodium chloride solution, dried over magnesium sulfate, filtered, and concentrated in vacuo. The resulting solid was triturated with diethyl ether. The solid was collected by filtration and washed again with diethyl ether to afford...”*<sup>34</sup>

While the work-up is an undeniably important phase of a reaction, the techniques used in the current work are reliant on the specification of amounts in order to identify reagents. This technique is well-suited to identification of primary reagents but not those used in work-up, and so in order to produce a metric that indicates the performance of the software in the role for which it was designed it was decided to entirely omit reagents mentioned in the work-up phase, and inert atmospheres under which reactions were performed, from the current analysis.

The manual inspection of the reaction corpus identified 249 true positives, 71 false positives and 139 false negatives—the system having a precision of around 78% and recall of around 64%. When considering these results, it should be remembered that the requirement for an identified reagent to be considered a true positive—that not only the chemical name but also the amounts employed in the reaction be identical to those described in the source text—is a rigorous standard. It was commonly the case during the analysis that the system identified the correct chemical name as a reagent but failed to correctly add one or more amounts, creating both a false positive and a false negative. These situations occurred where one or more of the amounts in the source text were not recognised by ChemicalTagger. Frequently these situations were caused by the patent author employing a structure that may be considered incorrect, *e.g.* “triphenylphosphine (3.08 g., 11.78 mmol)” or “1-Phenylpiperazine (16.2 g, 0.10 mole)”. The non-standard full stop indicating the abbreviation of “grams” in the first example and the failure to contract the unit “mole” to its standard symbol “mol” in the second result in the failure to recognise and convert these amounts to CML. The data gathered in the current exercise permit the improvement of the ChemicalTagger grammar to recognise a greater variety of the reporting formats used by authors and thereby improve the precision and recall for the identification of reagents as measured by the current methods.

These improvements, however, are not sufficient on their own to produce a system that operates at the level of a human operator. The current system requires further development before the data it produces are of sufficient quality to be considered reliable by the community at large.

The extracted reactions contain, where identified and successfully converted to CML, the  $^{13}\text{C}$  and  $^1\text{H}$  NMR spectra of the products. In the patents used for this work,  $^1\text{H}$  NMR spectra are far more common than  $^{13}\text{C}$ -indeed, the manually examined subset of the reaction corpus was found to contain only two  $^{13}\text{C}$  NMR spectra. Consequently, only the validity of the attached  $^1\text{H}$  NMR spectra in the reaction corpus was considered. Where these spectra were present, the content was compared to the reported spectra in the original sources. In order to be considered correct, the attached spectra were required to fully describe the original spectra in terms of the shifts, integrals, multiplicities and coupling constants of each peak-any deviation from what was reported in the original text resulted in the attached spectrum being judged to be incorrect.

The manual inspection identified 25 occasions on which the  $^1\text{H}$  NMR spectrum attached to a product molecule precisely replicated the information presented in the source text and 8 occasions on which it did not, *i.e.* a success rate of around 76%. The primary causes of the inclusion of incorrect  $^1\text{H}$  NMR spectra were the failure to fully convert peak metadata, *e.g.* multiplicities, as identified by OSCAR3 to CML and the conversion to CML spectra of sections of input text that did not indicate  $^1\text{H}$  NMR spectra, *i.e.* false positives in the data recognition procedure. The first of these issues indicates a bug in JUMBO-Converters that could be relatively trivially identified and fixed while it is expected that the second issue should produce  $^1\text{H}$  NMR that could be automatically distinguished from a genuine NMR spectrum in a majority of cases, since false positives will rarely contain expected peak metadata such as integrals and multiplicities. Though the  $^1\text{H}$  NMR spectra validation is based on a small set of data, it is believed that the spectra identified by PatentEye are of nearly sufficient quality that they constitute a resource of value to the community.

## 39.5 The Green Chain Reaction: are chemical reactions in the literature getting greener?

The development of an automatable patent extraction and interpretation system gave us an opportunity to include the scientific community in an *ad hoc* public project. ScienceOnline (2010) was a gathering of bloggers, information specialists, information providers, publishers and funders related to the communication of science. We set up a one month project dedicated to providing “a scientific result” by the time of the meeting. This highly ambitious idea relied on a critical mass of collaborators in a virtual community installing programs, running them to collect chemical information and aggregating it for presentation at the meeting.

The focus of the experiment was to see if a well-defined question could be answered by extracting information from the patent literature, using PatentEye.

As the basis of the project, named “The Green Chain Reaction” (GCR), we chose to focus on the use of solvents in chemical reactions to determining the “greenness” of chemical reactions in manufacturing and research. Tradi-

tional methods of chemical synthesis are becoming increasingly unacceptable because the processes are hazardous (explosion, toxicity), they consume scarce resources (metals, petrochemicals, *etc.*), the by-products (unwanted materials, which are often discharged to the environment) are hazardous (toxic, *etc.*) and they are energy-intensive. Both machines and humans were employed to collect and systematize chemical syntheses and to analyse the results.

The Green Chain Reaction was “Open Notebook Science” in that all discussions, code and results were publically viewable on the web at all stages. Moreover, anyone could volunteer to participate in the project. The planned methodology was:

1. A volunteer downloads the GCR software and installs it on their machine.
2. They run it against a given week of patent data from the 500 weeks available on the EPO website.
3. The software analyses the occurrence of solvents in the patents and records each instance of a particular solvent.
4. The software provides an aggregation and uploads this to a common site (an Open server at Cambridge).
5. The Cambridge software makes a further aggregation and presents the results.

In one sense this is a human analogy of a map-reduced project where a given task is farmed out to a large number of “computers” and the results are aggregated. In practice, we found a number of problems in distributing the software. The OSCAR package did not run “out-of-the-box” on all architectures, and it was some time before we discovered the cause of this (OSCAR’s workspace). For this reason some volunteers were not able to participate in the complete project. As we discovered bugs, new releases were made, sometimes on a daily basis. Nevertheless, we were ultimately able to analyse about 100,000 patents and to tabulate the results.

The GCR PatentEye workflow is as follows:

- Analyses a weekly patent index and downloads all the chemical patents
- Trawls through the patents to see which contain experimental sections
- Analyses the text to extract mentions of solvents, including chemical formula and amount (where given)
- Aggregates all the solvent data from a single patent into a summary file (
- Uploads the summary file to the GreenChainReaction website <http://greenchain.ch.cam.ac.uk/patents/results/>

The results were communicated onto the Cambridge server using a RESTful process. The solvents were identified by their linguistic context (using ChemicalTagger), and validated against Wikipedia pages of the same name. Thus, for example, ethyl acetate would have been determined as a solvent because of its linguistic environment (*e.g.* “dissolved in ethyl acetate” or “in 50 ml of EtOAc”). Sometimes the solvents were given as textual names (*e.g.* dichloromethane), and sometimes as compositional formulae (*e.g.* CH<sub>2</sub>Cl<sub>2</sub>). The first observation is that the extraction of solvents is extremely high precision *i.e.* there are very few entities retrieved which are not solvents. We have no information about the recall but it is clear that a large amount of data has been extracted. The solvents were then listed on the server with their aggregate counts and the chemical structure diagram retrieved from Wikipedia. Note that there needs to be a further disambiguation of names, so that there are entries for both dichloromethane and CH<sub>2</sub>Cl<sub>2</sub>, which should be summed, but in the time available for the project and with the given volunteers it was not possible to include this stage. The precision would appear to be > 99.9%.

We had hoped that there might be a large change in solvent usage over a decade. However, the most commonly used solvents (THF, dichloromethane) have remained at approximately the same frequency. These solvents are not completely green being a) potentially explosive and b) containing toxic C-Cl bonds, so there is no particular evidence in increasing greenness. However we caution this interpretation as there are many dates associated with the patent, and we cannot be sure how these relate to the actual dates on which the syntheses were carried out. Moreover there is a considerable lag between the actual synthesis and the publication of the patent so that recent changes in use have probably not been picked up.

## 39.6 Competing interests

The authors declare that they have no competing interests.

## 39.7 Authors' contributions

DMJ wrote the manuscript, developed the PatentEye software and carried out the analysis of its performance.

SEA provided structure analysis routines, helped develop the Green Chain Reaction software, advised on software architecture and helped write the manuscript.

PMR developed OSCAR and CML, organised and ran the Green Chain Reaction experiment and wrote the manuscript.

## 39.8 Acknowledgements and funding

The invaluable assistance of Dr. Charlotte Bolton in the production of this manuscript is acknowledged.

We also thank Unilever (DMJ's PhD studentship) and the EPSRC (Pathways to Impact Award) for funding.

Green Chain Reaction contributors: Jean-Claude Bradley, Dan Hagon, Bob Hanson, Cameron Neylon, Heather Pi-wowar, Diego Riaño, Alberto Sicilia, Mat Todd, Anita de Waard, Richard West, Mark Woodbridge; Henry Rzepa for suggesting the 'green' focus.



# OSCAR4: A FLEXIBLE ARCHITECTURE FOR CHEMICAL TEXT-MINING

## 40.1 Abstract

The Open-Source Chemistry Analysis Routines (OSCAR) software, a toolkit for the recognition of named entities and data in chemistry publications, has been developed since 2002. Recent work has resulted in the separation of the core OSCAR functionality and its release as the OSCAR4 library. This library features a modular API (based on reduction of surface coupling) that permits client programmers to easily incorporate it into external applications. OSCAR4 offers a domain-independent architecture upon which chemistry specific text-mining tools can be built, and its development and usage are discussed.

## 40.2 Introduction

*In keeping with the historical and methodological aspects of this special issue, we recount the history and motivation of OSCAR.*

A large amount of factual data in chemistry and neighbouring disciplines is published in the form of text and components within text rather than as structured semantic information. If we can discover and extract this information, the textual literature becomes an enormous additional chemical resource. As an example, we estimate that about 10 million chemical syntheses per year are published in the public literature (articles, patents, theses) and the conventional method is a natural language narrative (most commonly in English). It is extremely tedious and error-prone to extract information from this narrative manually, and for this reason many chemical abstracting services limit their scope and also frequently lag behind the current publication list.

The discipline of text-mining has now reached a state where much natural language in textual form can be analysed rapidly and with high precision and recall. Methodologies applied to the problem of chemical named entity recognition include dictionary- and rule-based methods, as well as machine learning and hybrid approaches <sup>1234567891011</sup>. We have been working in this area for approximately 10 years and the OSCAR4 software, together with OPSIN (the Open Parser

---

<sup>1</sup> Extraction of Information from the Text of Chemical Patents. 1. Identification of Specific Chemical Names

<sup>2</sup> Analysis of Biomedical Text for Chemical Names: A Comparison of Three Methods

<sup>3</sup> A scalable machine-learning approach to recognize chemical names within large text databases

<sup>4</sup> Detection of IUPAC and IUPAC-like chemical names

<sup>5</sup> A dictionary to identify small molecules and drugs in free text

<sup>6</sup> Extraction of CYP Chemical Interactions from Biomedical Literature Using Natural Language Processing Methods

<sup>7</sup> Chemical Names: Terminological Resources and Corpora Annotation

<sup>8</sup> Identification of Chemical Entities in Patent Documents

<sup>9</sup> Automatic vs manual curation of a multi-source chemical dictionary: the impact on text mining

<sup>10</sup> Abstracts versus Full Texts and Patents: A Quantitative Analysis of Biomedical Entities

<sup>11</sup> Identifying, Indexing and Ranking Chemical Formulae and Chemical Names in Digital Documents

for Systematic IUPAC Nomenclature)<sup>12</sup><sup>13</sup> and ChemicalTagger<sup>14</sup><sup>15</sup>, represent the public state-of-the-art in chemical text analysis and extraction.

The OSCAR (Open-Source Chemistry Analysis Routines) software has been developed over a period of years and a number of projects. Between 2002 and 2004, sponsors including the Royal Society of Chemistry (RSC), Nature and the International Union of Crystallography (IUCr) supported a number of summer studentships. These projects were focused on the development of software with limited capacity for the automated interpretation of chemical documents, and resulted in two main software components—the Experimental Data Checker<sup>16</sup><sup>17</sup> and OSCAR2.

The Experimental Data Checker was conceived as a tool to be used as part of the RSC’s publication process. The tool is capable of recognising sections of reported experimental data within plain text input using regular expressions to match the highly-stylised and journal-mandated formats in which they are reported in the literature (as shown in Figure 1). Once this information has been identified and interpreted, the tool performs elementary checks on the characterisation data (spectra, analytical) where molecular structures are reported, and attempts to ensure that the data does not conflict with the structure.

The Experimental Data Checker application relied upon a core library of analysis routines, and it was this library that was the first to bear the name OSCAR. Further development of this library in the summer of 2004 resulted in OSCAR2, which used XML formatting to represent the document undergoing processing, and applied XML annotations to the document to indicate recognised sections of text. OSCAR2 implemented a naïve Bayesian system based on *n*-grams and a simple grammar in order to identify chemical names within a text. These improvements were later extended as part of the OSCAR3 project.

In 2005, the EPSRC awarded a grant (“Sciborg”) to develop natural language processing (NLP) tools for chemistry and science. The chemistry component of this project focused on the development of the OSCAR2 methodology and resulted in the creation of OSCAR3<sup>18</sup>. OSCAR3 focuses on the recognition of and, where appropriate, the resolution of connection tables for chemical named entities. OSCAR3 employs a naïve Bayesian model to identify “chemical” tokens in text and offers a choice of two methods for the identification of multi-token named entities. The first of these, the PatternRecogniser, uses predetermined regular-expression style heuristics while the second, the MEMMRecogniser<sup>19</sup>, employs machine learning in the form of a Maximum Entropy Markov Model (MEMM). OSCAR3 uses these methods to identify four classes of named entity (Chemical, Reaction, Chemical Adjective and Enzyme) as well as dictionary lookup to identify a pre-determined set of ontology terms and a discrete finite automaton based method to identify chemical prefixes.

In order to convert chemical names to connection tables (Figure 2), OSCAR3 uses dictionary-based methods and, where this is not successful, OPSIN. Early versions of OSCAR directly included the OPSIN code, but this was later re-factored into a separate library.

By 2008, OSCAR was in common use in many laboratories for the identification and extraction of chemical terms (chemical named entities) in a variety of texts. Our original metrics<sup>18</sup> showed that the precision and recall were domain-dependent and varied considerably with the purpose and style of chemical texts. Feedback from users was informal but it was clear that they were modifying OSCAR for their particular purposes both in vocabulary and recognition methods. As a result we embarked on a major re-factoring program in order to robustify the OSCAR software and simplify the API, and this paper describes the results.

#### 40.2.1 Historical Funding and Collaboration

It is very difficult to get funding for software engineering projects, especially when apparently little changes on the surface. We are grateful to the following bodies for their funding and interest:

---

<sup>12</sup> Chemical name to structure: OPSIN, an open source solution

<sup>13</sup> OPSIN, Open Parser for Systematic IUPAC Nomenclature

<sup>14</sup> ChemicalTagger: A tool for semantic text-mining in chemistry

<sup>15</sup> ChemicalTagger

<sup>16</sup> Experimental data checker: better information for organic chemists

<sup>17</sup> RSC Experimental Data Checker

<sup>18</sup> High-Throughput Identification of Chemistry in Life Science Texts

<sup>19</sup> Cascaded classifiers for confidence-based chemical named entity recognition

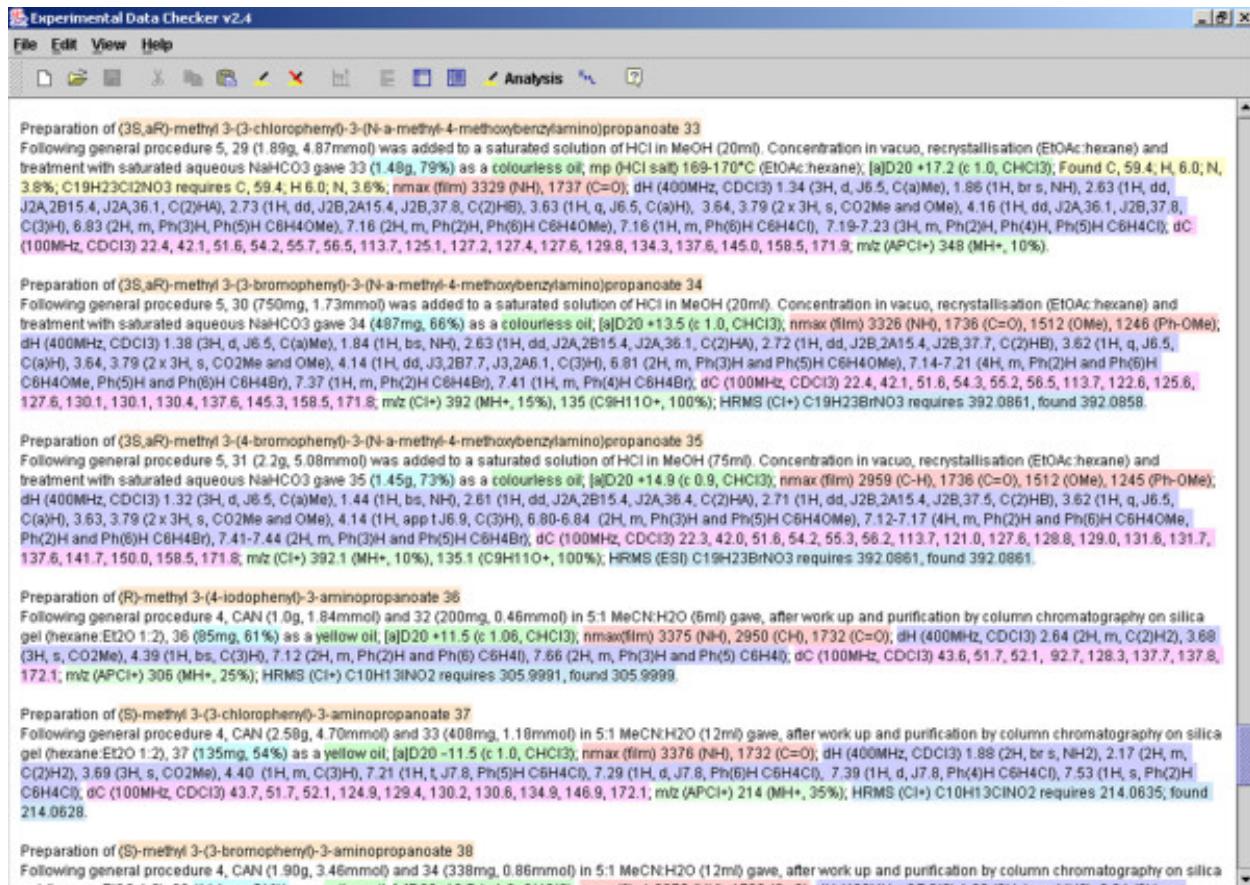


Figure 40.1: Figure 1. A screenshot of the Experimental Data Checker (OSCAR-Data) showing identification and markup of plain text experimental data

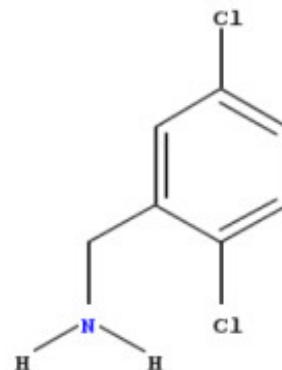
**A screenshot of the Experimental Data Checker (OSCAR-Data) showing identification and markup of plain text experimental data.** The initial application of OSCAR was to parse the highly stylised data used to report spectra and other analytical proofs of synthesis. This functionality is very widely-used (pers. comm. from RSC staff) and has been re-integrated into OSCAR4 rather than being a separate application.

### Preparation of N-(2,5-dichlorobenzyl)-8-hydroxy-1,6-naphthyridine-7-carboxamide

Triphosgene (0.556g, 1.87 mmol) was added over 20 mins to a solution of 8-hydroxy-1,6-naphthyridine-7-carboxylic acid (0.89g, 4.68 mmol) and diisopropylethylamine (3.26 ml, 18.7 mmol) in DMF (22 ml) at 0°C.

2,5-dichlorobenzylamine (0.142 ml, 1.05 mmol) was treated with a portion of the above solution (0.58ml, <sup>1</sup>H NMR id = o11; surface = 2,5-dichlorobenzylamine; type = CM; confidence = 0.9293289445146479; SMILES = [H]C1=C([H])C(=C([H])=C1Cl)C([H])[H]N([H])[H]C1=CC(Cl)=CC(Cl)=C1) for 16 hrs. <sup>1</sup>H NMR id = o11; surface = 2,5-dichlorobenzylamine; type = CM; confidence = 0.9293289445146479; SMILES = [H]C1=C([H])C(=C([H])=C1Cl)C([H])[H]N([H])[H]C1=CC(Cl)=CC(Cl)=C1) and InChI = InChI=1/C7H7Cl2N/c8-6-1-2-7(9)5(3-6)4-10/h1-3H,4,10H2; smiRef = smi6; d, J=4.0 Hz), 8.04 (d, J=4.0 Hz), 7.54 (<sup>1</sup>H, d, J=8.0Hz), 7.50-7.30 (<sup>2</sup>H, m), 4.64 (<sup>2</sup>H, d, J=5.0 Hz) ppm.

FAB MS calcd for C<sub>16</sub>H<sub>11</sub>N<sub>3</sub>O<sub>2</sub>Cl<sub>2</sub> 348 (MH<sup>+</sup>), found 348.



- Experimental data
- Ontology term
- Chemical (etc.) with structure
- Chemical (etc.), without structure
  - Reaction
- Chemical adjective
- enzyme -ase word
- Chemical prefix

Figure 40.2: Figure 2. OSCAR3 markup displaying recognised chemical entities (CM)

**OSCAR3 markup displaying recognised chemical entities (CM).** A mouse-over action on an annotated term displays the associated metadata, in this case for 2,5-dichlorobenzylamine, and displays an image representing the structure generated by the Chemistry Development Kit (CDK) <sup>202122</sup> (right). OSCAR3 concentrated on the identification and interpretation of chemical entities in text (named entity recognition, NER). The primary purpose was to identify and extract the following types of object: chemicals (CM), ontology terms (ONT; looked-up from ChEBI <sup>232425</sup>, FIX <sup>26</sup> and REX [#B42]\_\*etc\*.), reactions (RN; as identified by linguistic constructs, e.g. “methylated”), chemical adjectives (CJ) mainly formed from chemical nouns), enzymes (ASE) and chemical prefixes (CPR), highlighted in different colours. These concepts are maintained in OSCAR4.

1. OMII-UK. This organisation existed to support and robustify the products of the UK eScience program. Many of these were middleware products but OSCAR was seen by the UK eScience community as an example of a widely-deployable component that could be used in a modern manner in many branches of science. The OMII-UK project carried out an initial scoping and re-factoring of the OSCAR3 source.
2. The OSCAR-ChEBI project. This was a competitive funding resource for eScience products and we worked with the European Bioinformatics Institute (EBI) to develop OSCAR as an appropriate tool for the extraction and verification of chemistry in the ChEBI ontology.
3. CheTA. This was a JISC-funded project led by our group in conjunction with the National Centre for Text Mining (NaCTeM) to evaluate the relative merits of human annotation and machine annotation of documents. Part of this project involved OSCAR running under the UIMA <sup>27</sup>/U-Compare <sup>2829</sup> framework and required a re-factoring <sup>30</sup>.

As a result of these projects, which probably amounted to two person-years of effort in the re-factoring, OSCAR4 has now been released in a usable form.

#### 40.2.2 Limitation of OSCAR3 and design goals for OSCAR4

OSCAR3 is a powerful tool for chemical natural language processing, but early attempts to develop software using it as a library rather than as a standalone application—the ChemicalTagger <sup>14</sup> and PatentEye <sup>3132</sup> projects—exposed weaknesses in the code in this regard. The architecture of the software was built around the principle that the software would be running as a server on the user’s local machine. In order to function correctly, it required a properly configured workspace. Many key components were implemented as mutable singletons (static objects), compromising the thread-safety of the application and meaning that safe reconfiguration of a workflow required a complete shutdown and restart of the Java virtual machine (JVM). Furthermore, the implementations of the various OSCAR components required that a document be formatted in SciXML as it underwent processing. Consequently, the use of OSCAR3 by a client programmer to build secondary applications was unintuitive, and the distribution and successful use of such applications was found, as part of the Green Chain Reaction, to require an unacceptably high level of support.

Early attempts to resolve these problems <sup>23</sup> involved the extraction of the OSCAR3 tokeniser, MEMMRecogniser and PatternRecogniser components from the main OSCAR3 codebase and their conversion into modules suitable for use in the popular text-mining framework U-Compare. This work allowed the use of OSCAR as part of a drag-and-drop workflow, but not its direct integration into another application. Consequently, a comprehensive overhaul of the OSCAR3 code began in autumn 2010 with the aim of producing a well-engineered, simple, modularised version of OSCAR that retained the core OSCAR3 functionality and could be easily integrated into external applications. This most recent development has been designated OSCAR4 and is discussed in the remainder of this paper.

The development of OSCAR4 sought to address a number of specific issues. These are summarised below (and in Appendix A) and subsequently discussed in greater detail.

1. To produce an OSCAR library with a simple API, suitable for use by client programmers who may not be familiar with the internal workings of OSCAR. Consequently, while it is desirable for users to be able to customise the behaviour of OSCAR in a number of ways, initialisation of OSCAR components must by default produce configurations that “just work”—the ‘convention over configuration’ paradigm (Appendix B).
2. In order to run, OSCAR3 required the existence of a properly configured workspace—a directory on the executing machine that contains the OSCAR chemical name dictionary, the InChI <sup>3334</sup> binary file and a properties file along with subdirectories intended to contain further resource files. When OSCAR3 is first run this workspace

---

<sup>27</sup> UIMA

<sup>28</sup> U-Compare

<sup>29</sup> U-compare: Share and compare text mining tools with UIMA

<sup>30</sup> Using workflows to explore and optimise named entity recognition for chemistry

<sup>31</sup> Information extraction from chemical patents

<sup>32</sup> Mining chemical information from Open patents

<sup>33</sup> The IUPAC International Chemical Identifier

<sup>34</sup> IUPAC International Chemical Identifier

is automatically created, and when OSCAR3 is used as a library the workspace is automatically created in the working directory. This behaviour was deemed undesirable, unnecessary and found to be a cause of difficulties in producing distributable OSCAR-dependent software. Consequently, the removal of the requirement for a workspace was considered a high priority of the OSCAR4 project.

3. Much of the OSCAR3 code required that a document undergoing processing is formatted in SciXML. Though converters are provided to transform HTML into plain text and plain text into SciXML, the requirement to perform this transformation is frustrating to the client programmer in that it prevents him from working directly with plain text or with a custom XML format which may very well be the native format of a document that he wishes to process. Consequently, the removal of this SciXML dependence was considered important.
4. In addition to its core functionality—the recognition and interpretation of chemical named entities—OSCAR3 included a wide range of secondary functions including the OSCAR3 server. This server runs on the local machine and provides an interactive demonstration of the capacity of OSCAR3 for text processing as well as a number of other utilities including the capacity to manually annotate a text from within a browser window, a servlet for the interconversion and depiction of chemical names and formats and an experimental Hearst pattern<sup>35</sup> based system for the extraction of chemical relations from text. The OSCAR3 codebase had the resemblance of a ‘treasure trove’ which made code maintenance a more complex task than necessary. The separation of a library containing the core OSCAR functionality from these secondary functions was therefore considered desirable.
5. Much of the architecture of OSCAR3 lacked clear definition. Excessive use is made of mutable singletons which, while aiding performance by eliminating the need for re-initialisation of components, allows for complex interactions in the code, making it difficult to understand, debug and re-factor. This problem was compounded by the manner in which program logic is partially controlled by a properties object backed by a serialised file. Some of the property values can be modified at runtime while others, once accessed by the objects that rely upon them, are duplicated in memory and cannot be further changed. Attempts to resolve these complex interactions can have unintended consequences since the unit test coverage in OSCAR3 is sparse. Consequently, the improvement of the architecture of the OSCAR software was considered a vital part of the OSCAR4 project.
6. It has been known for some time that the speed of OSCAR3 operation could be improved by introducing certain optimisations into the code. Using the YourKit Java profiler<sup>36</sup>, a number of performance blackspots were identified and subsequently eliminated. This work was started after the final version of OSCAR3 (OSCAR3 alpha 5<sup>37</sup>) and continued as part of the OSCAR4 project.

### 40.2.3 Library as a design

OSCAR4 has been deliberately written as a Java library, rather than an application or service. Consequently, the decoupling of the core OSCAR functionality from applications that use this functionality has been achieved. The usage of the library has been simplified as much as possible with the introduction of the Oscar API object—a class intended to wrap the functionality of the wider library and provide default implementations of the various components. As a result, OSCAR4 can be called from external software, as shown in the examples in Figure 3.

In the first of the examples in Figure 3, OSCAR4 is used to detect named entities in an input string, returning a List of NamedEntity objects. In the second, it is used to both detect named entities and, where these named entities correspond to chemical names, to resolve these names to chemical structures—returning a List of ResolvedNamedEntity objects. The ResolvedNamedEntity class links a NamedEntity to a list of chemical structures in a number of formats—SMILES, InChI and CML—while the NamedEntity class stores such information as the surface (raw text) and type (e.g. compound or reaction) of the named entity and the indices that define its position within the source text. The outputs of these examples are illustrated in Figure 4.

The examples above show how OSCAR4 can be used without the need for any understanding of the underlying technology or implementations. An overview of the workflow managed by the Oscar API object is shown in Figure 5.

<sup>35</sup> Automatic acquisition of hyponyms from large text corpora

<sup>36</sup> YourKit java profiler

<sup>37</sup> OSCAR3 alpha 5

```

a) String text = "The quick brown ethyl acetate jumps over the lazy bromine";
Oscar oscar = new Oscar();
List <NamedEntity> neList = oscar.findNamedEntities(text);

b) String text = "The quick brown ethyl acetate jumps over the lazy bromine".
Oscar oscar = new Oscar();
List <ResolvedNamedEntity> rneList = oscar.findAndResolveNamedEntities(text);

```

Figure 40.3: Figure 3. Java code using the OSCAR4 API to a) identify chemical named entities (CNEs) in a block of text and b) identify CNEs and resolve their connection tables where possible

**Java code using the OSCAR4 API to a) identify chemical named entities (CNEs) in a block of text and b) identify CNEs and resolve their connection tables where possible.**

The input is first passed to the Tokeniser to produce a list of TokenSequence objects, each of which roughly corresponds to a paragraph of text and contains a list of Token objects. The Token represents a string of characters that mostly correspond to words but also to punctuation or other discrete units of text *e.g.* “C<sub>2</sub>H<sub>6</sub>O” or “42”. In NLP tools, tokenisation commonly occurs at whitespace or punctuation boundaries, however due to the form of some of the domain-specific entities found in chemical texts such as “C-H” a custom Tokeniser is used. The TokenSequences are then passed to a ChemicalEntityRecogniser-an interface for a class capable of identifying a list of NamedEntities, which are subsequently passed to the ChemNameDictRegistry to create a list of ResolvedNamedEntities if required.

This workflow can be customised by the user, who can use the set() methods of the Oscar class to replace the components of the default configuration with suitable customised or custom-built alternatives. Specifically, the user can select which implementation of ChemicalEntityRecogniser to use or can specify which set of ontology terms are to be recognised and which model the default ChemicalEntityRecogniser should use, and which dictionary registry, *i.e.* set of chemical name dictionaries, to use for name to structure resolution. In addition to this, the public APIs of the individual components can be used to assume a greater degree of control over the execution of the workflow.

OSCAR4 provides three implementations of the ChemicalEntityRecogniser. The first, the RegexRecogniser, finds terms that match a given regular expression and is intended to find serial numbers corresponding to compounds *e.g.* “NSC-2648”. The others, the PatternRecogniser and the MEMMRecogniser, use more complex strategies to identify chemical named entities and feature subcomponents that can be customised by the user to produce the desired behaviour.

The architecture of the PatternRecogniser is shown in Figure 6. A list of “chemical” words is drawn from an internal dictionary composed mostly of words derived from the ChEBI database and from a corpus of manually-annotated documents, while a list of “non-chemical” words is determined by removing those words that occur in the chemical word list from a standard English dictionary. These lists are used to build an *n*-gram model which is used by a naïve Bayesian classifier to determine whether novel tokens are “chemical” or “non-chemical”. Multi-token named entities, *e.g.* “ethyl acetate”, that occur within the input text are then identified by regex-style matching of chemical tokens to a set of pre-specified pattern definitions such as “\*yl \*ate”.

The architecture of the MEMMRecogniser is shown in Figure 7, in which chemical named entities are identified using a Maximum Entropy Markov Model (MEMM). The feature set that is generated for each token includes features that describe the token in question, such as the *n*-grams that describe it and the probability that it is chemical as predicted by the *n*-gram model as previously, as well as contextual features that describe its neighbouring tokens. Using these features, the MEMM model assigns a chemical token as being either the first token in a named entity or a subsequent token in a named entity. Given these assignments, multi-token named entities can be constructed. Novel MEMM models can be built from a corpus of hand-annotated documents by the user, and OSCAR4 is supplied with two pre-

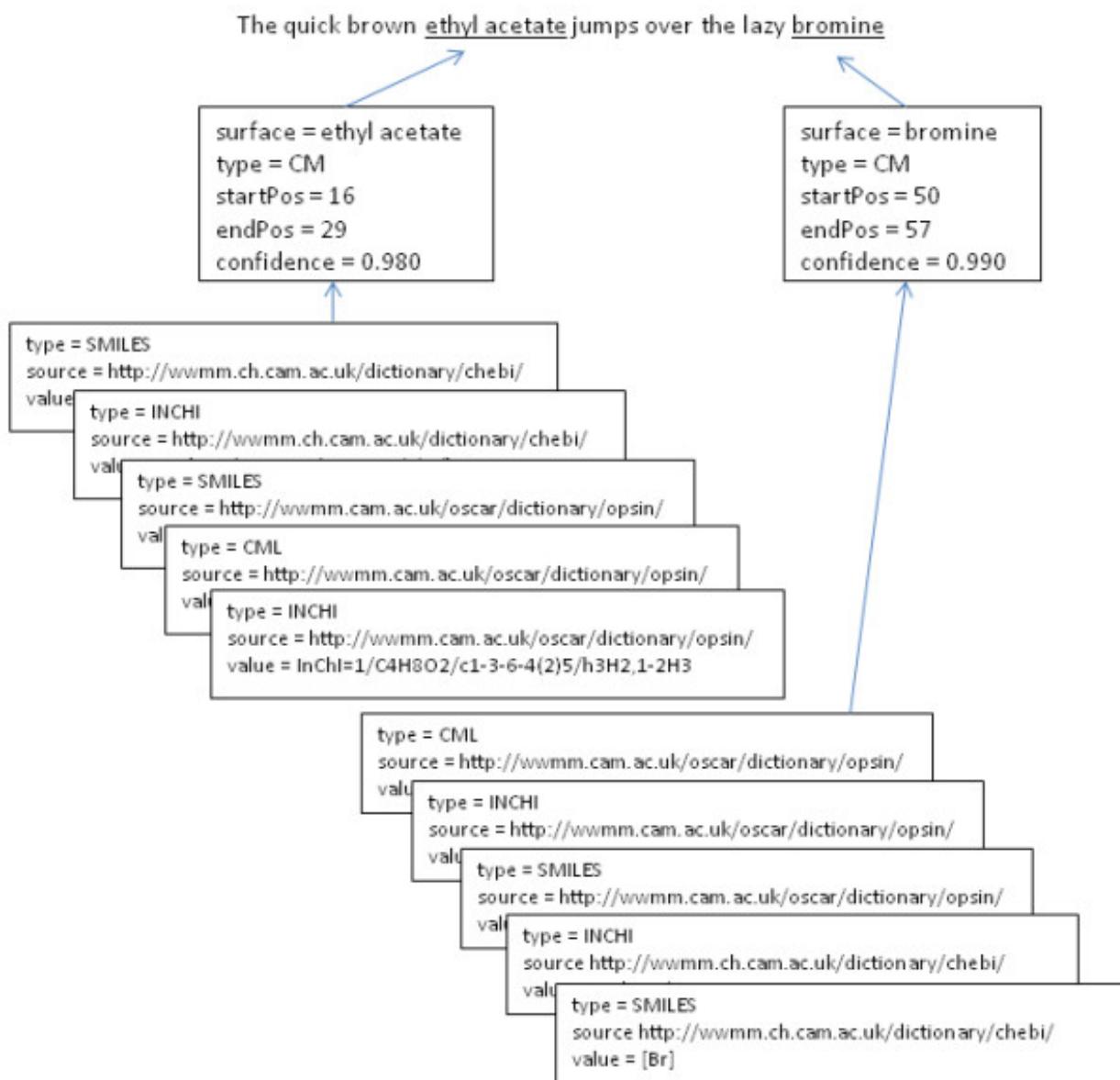


Figure 40.4: Figure 4. Graphic representing the structure of the OSCAR4 API output object

**Graphic representing the structure of the OSCAR4 API output object.** Named entities reference their position in the input text, the confidence in their identification and resolved structures in various formats (SMILES<sup>3839</sup>, InChI, CML [#B45]\_\*etc\*.).

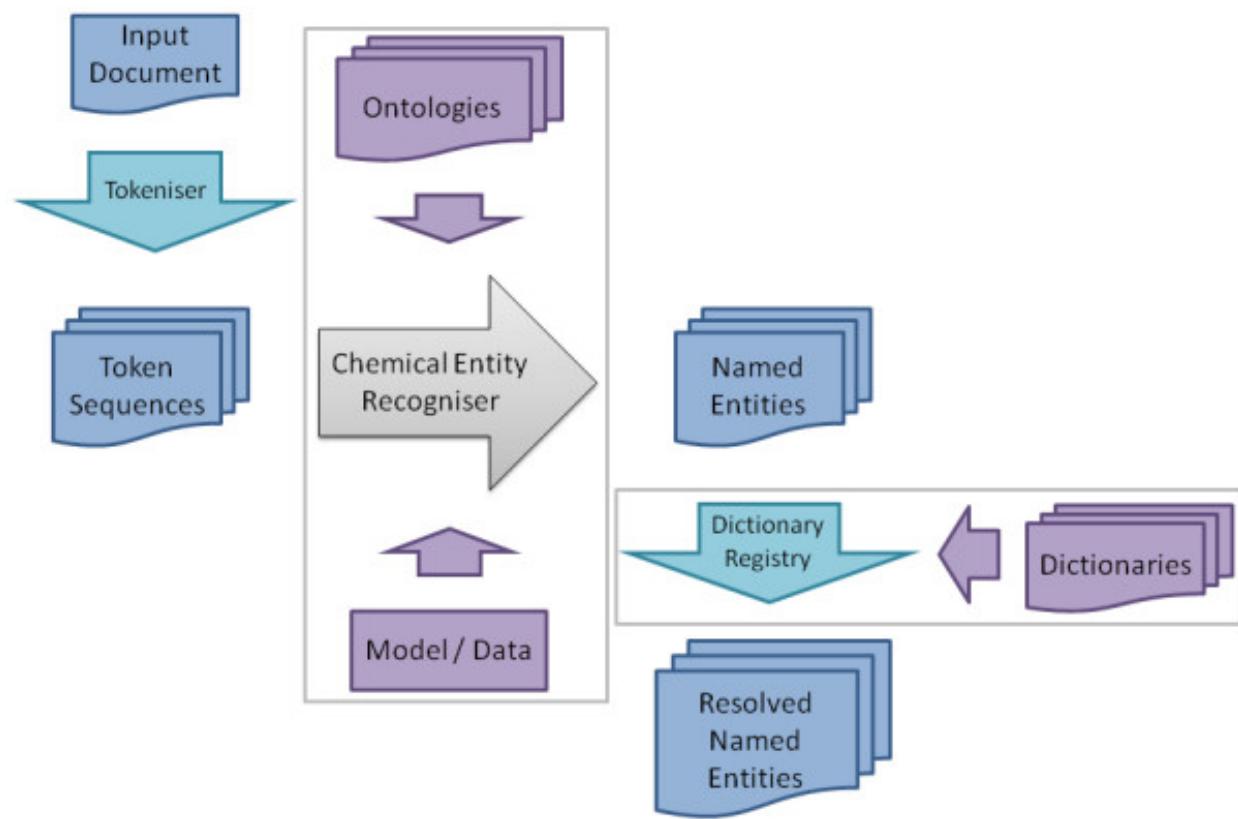


Figure 40.5: Figure 5. Workflow of the OSCAR4 API object  
Workflow of the OSCAR4 API object.

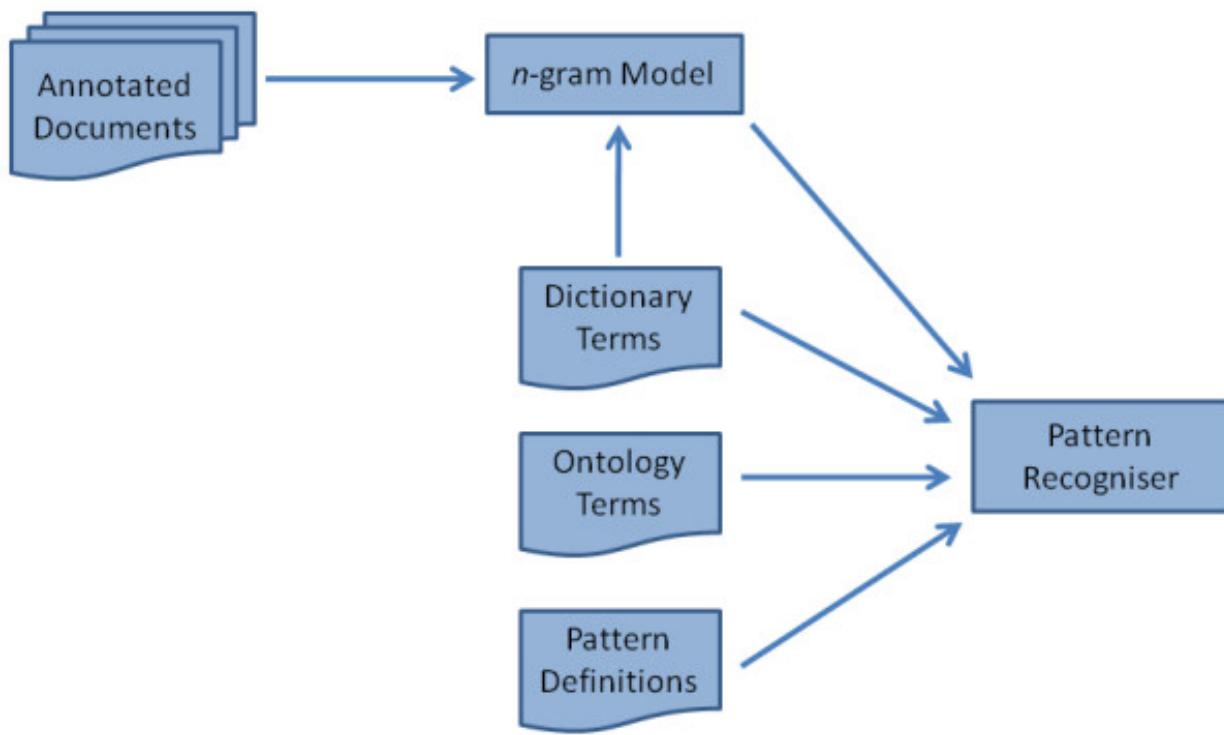


Figure 40.6: Figure 6. PatternRecogniser architecture  
PatternRecogniser architecture.

generated models. One of these models was built from a set of papers from RSC journals<sup>40</sup>, while the other was built from a set of abstracts retrieved from PubMed<sup>19</sup>.

#### 40.2.4 Architecture and tests

The OSCAR4 library has been separated into a number of modules with each performing a defined role in the operation of the OSCAR code, such as the tokenisation of text or the provision of chemical name dictionaries. This allows client programmers to use as much or as little of OSCAR in their applications as required, without the need to unnecessarily pull in a large, comprehensive, single JAR. The process of creating the sub-projects had the additional advantage of highlighting the ways in which the separate components interact. During this process, the readability of the OSCAR code was improved by imposing a number of the idioms of ‘clean code’, and the reliability of the code was improved by the creation of appropriate unit and regression tests. At the time of writing, OSCAR4 has nearly 500 tests. As a result, the OSCAR4 code is far more robust than OSCAR3, so a developer can work both with and on the core OSCAR code with a far greater degree of confidence.

The mutable singletons that were commonplace in OSCAR3 have been largely removed. Instead, when setting up custom workflows, a user has the choice of either calling the getDefaultInstance() method or the default constructor as appropriate-each of which returns a preconfigured instance of the class-or using the custom constructor which uses dependency injection to supply the OSCAR components upon which the class depends. For example, the OntologyTerms class represents a set of ontology terms and their corresponding ontology IDs. The following two methods of obtaining an OntologyTerms object are available:

```

OntologyTerms.getDefaultInstance();
new OntologyTerms(ListMultimap<String, String> terms);
  
```

<sup>40</sup> Annotation of Chemical Named Entities

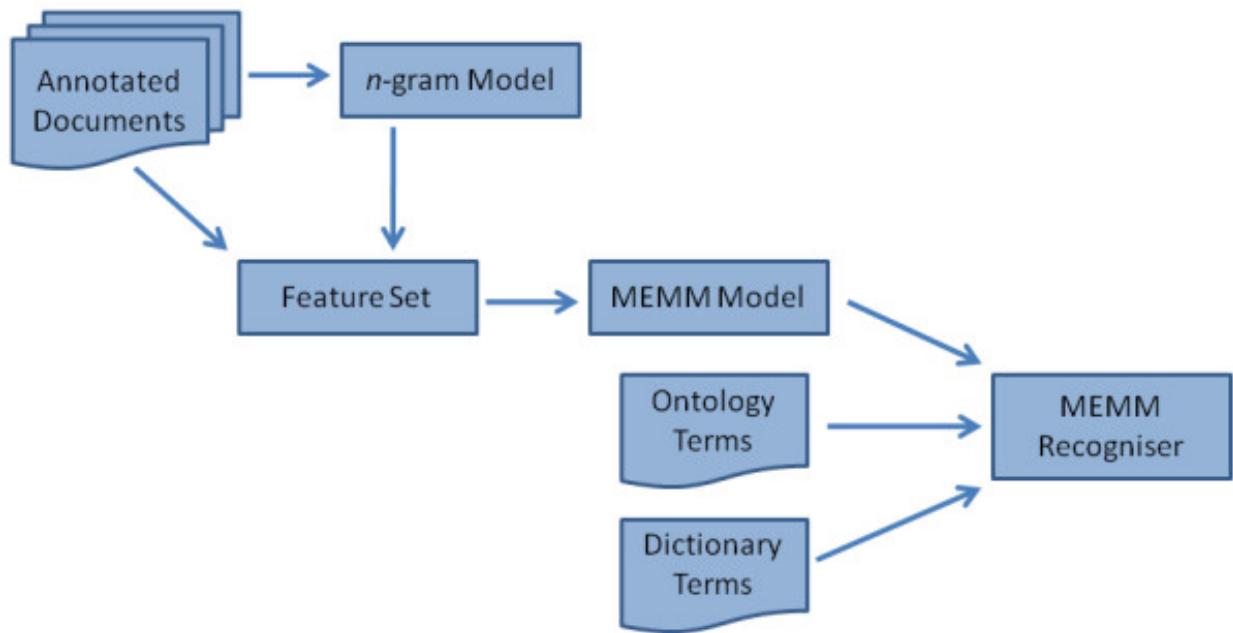


Figure 40.7: Figure 7. MEMMRecogniser architecture  
MEMMRecogniser architecture.

The first method returns the default OSCAR4 OntologyTerms object, which contains an amalgamation of the terms from the ChEBI, FIX and REX ontologies while the second supplies a multimap of ontology terms to IDs. The use of this design pattern throughout the codebase permits, but by no means requires, a user to assume a high degree of control over the functioning of OSCAR.

The use of the properties file and object to control elements of the program execution has been removed. Instead, the required information is either specified as part of a constructor's signature or using a set() method on the object in question. This improves the thread-safety of OSCAR, particularly in a multiuser environment, and contributes to its usability since a user can now trivially see what features may be customised from the outline of the class as opposed to needing to know which and how properties are used by which components.

#### 40.2.5 Input and Output Formats

As previously discussed, OSCAR3 required that input documents be converted into SciXML before processing can occur, using the document formatting as a base against which annotations for identified named entities can be referenced—whether as inline or standoff annotations. XML input turned out to be overly complex as NLP tools require “flat” relatively sequential tokens. The XML markup adds little useful context. OSCAR4 removes this requirement by operating on plain text and producing NamedEntity and DataAnnotation objects to represent recognised sections of text and does not currently produce serialised output, though some support for the serialisation of annotations into XML documents is planned for future releases. It should be realised, however, that there is no single, fool-proof approach to this problem. Different XML schema may use different methods to indicate where in the document section breaks and even text content occur, while it cannot be guaranteed that well-formed inline annotations can be generated for a given input document. Client programmers are therefore recommended to consume NamedEntity objects directly rather than rely upon serialised output, though it is realised that users are likely to want to be able to create serialised, marked-up copies of their documents as well.

## 40.2.6 Non-core functionality

Non-critical code (particularly downstream applications) has been removed from the OSCAR4 codebase to reflect the philosophy that OSCAR4 should act as a library. While some minor supporting code remains, such as that required for generation of key resource files, the majority has been removed entirely as it is envisaged that much of the former functionality could be better implemented by developers with specific use cases.

A number of useful non-core functions are provided in dependent libraries developed at the Unilever Centre in Cambridge. Specifically, subsidiary modules exist to provide the capacity to run OSCAR4 from the command-line, as part of UIMA or Taverna<sup>41</sup> workflows and from the Bioclipse<sup>42</sup> scripting interface, as shown in Figure 8.

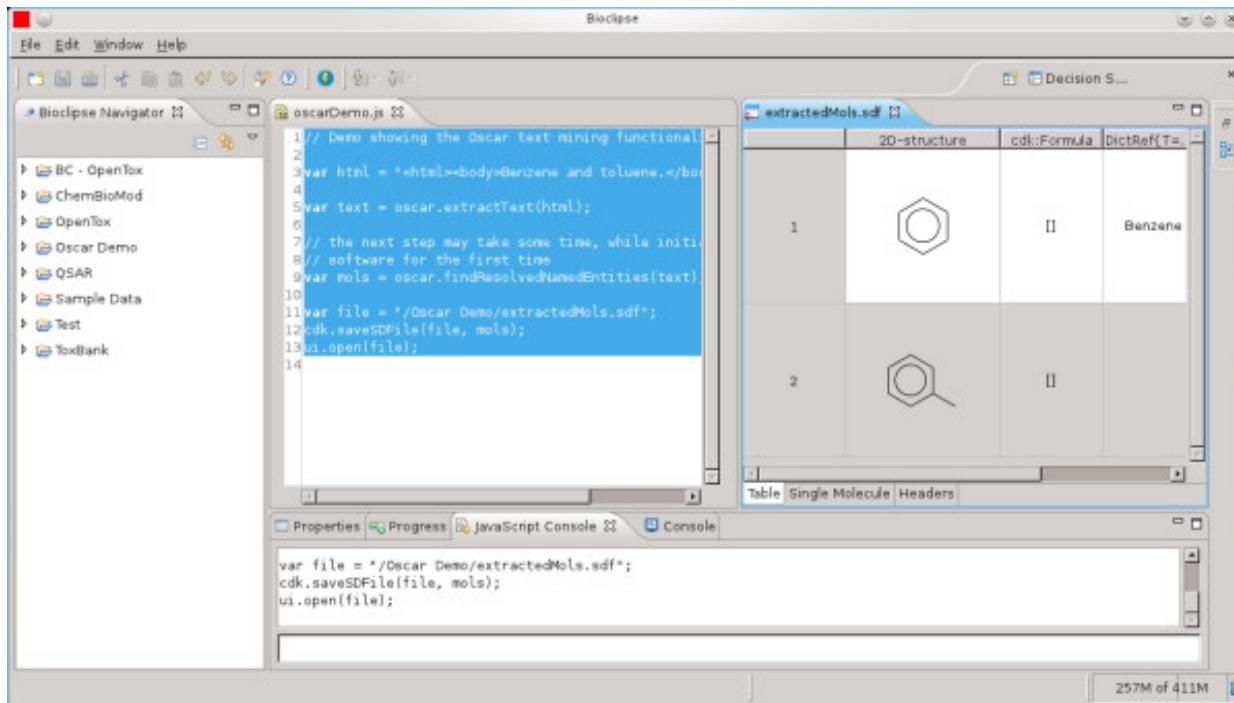


Figure 40.8: Figure 8. OSCAR4 run within Bioclipse's scripting interface (centre pane) identifying named entities in a block of text and saving the connection tables to file (extractedMols.sdf) for viewing (right pane)

**OSCAR4 run within Bioclipse's scripting interface (centre pane) identifying named entities in a block of text and saving the connection tables to file (extractedMols.sdf) for viewing (right pane).**

## 40.2.7 Performance

A number of modifications were introduced to the OSCAR code with the aim of reducing the time required to process documents. Performance hotspots were identified using the YourKit Java profiler and where possible eliminated. Some such improvements focused on the time taken to initialise the various OSCAR components, such as supplying a pre-calculated, serialised copy of the *n*-gram models used for named entity recognition rather than regenerating them each time OSCAR is loaded. Others improved the speed at which OSCAR can process a document by optimising extremely tight loops in the code, such as eliminating unnecessary string declaration while calculating *n*-gram features and avoiding recompilation of regular expressions. Further improvements were made *ad hoc*, as the OSCAR4 developers encountered obvious bottlenecks while working on the code.

<sup>41</sup> Taverna

<sup>42</sup> Bioclipse

In order to quantify the improvement in speed of operation, the time taken by both OSCAR4 version 4.0.1 and OSCAR3 alpha 5 to perform two tasks was measured. The first task measured the time taken to initialise the software to the point that it was ready to begin the task of finding named entities in text; the second task aimed to measure the speed at which the software could process bulk text and consisted of processing the full text of the 68 patents published by the European Patent Office in the week of 2009-05-06—a total of 11468 paragraphs of text. All the tasks were run on a desktop computer equipped with an Intel Pentium 4 (3.00 GHz) CPU and 1 GB of RAM, purchased *c.* 2005, running openSUSE 11.1 and using the Java 1.6.0\_22 32-bit virtual machine with a maximum heap size of 512 MB. The results are summarised in Table 1 and Table 2.

From these data, it can be seen that OSCAR4 performs significantly faster than OSCAR3. Initialisation times for the MEMMRecogniser and PatternRecogniser have been reduced by 17% and 20% respectively, while bulk processing times have been reduced by 18% and 46% respectively. The OSCAR4 MEMMRecogniser and PatternRecogniser processed approximately 26 and 76 paragraphs per second respectively, demonstrating that bulk processing of text is achievable on an acceptable timescale on desktop computers.

## 40.2.8 Deployment

OSCAR4 has generated significant interest in the community, and has been the subject of two meetings at the Unilever Centre for Molecular Science Informatics in Cambridge. The talks from the second of these are available to view online<sup>43</sup>. To our knowledge, the software is in use at the National Centre for Text Mining (NaCTeM), the European Bioinformatics Institute (EBI) and the European Patent Office (EPO) as well as various pharmaceutical companies.

We are aware of successful and straightforward integrations into the Bioclipse and Taverna frameworks, and believe that this is similarly straightforward for other Java environments. We were also pleased to see that at the recent MIOSS meeting at the EBI, OSCAR and OPSIN had been integrated into the .NET environment. For example, OPSIN was demonstrated as running within the JVM in Microsoft Excel, which is acceptable to commercial organisations as the JVM is of proven security.

## 40.2.9 Future Prospects

This is a useful opportunity to reflect on the high cost of producing robust, re-usable software. OSCAR3, and OPSIN, were produced as a continuing activity by a mixture of summer students, PhDs and PDRAs and, until *ca.* 2009, evolved rather than having a top-down software design. When the project became valuable to the world, it was a clear indication that re-factoring was going to be essential, and it is important to realise the necessary but high cost of doing this. In times of lean funding, it will become increasingly difficult to obtain this type of support, and therefore it is always tempting to transfer academic code to commercial entities which can raise revenue.

The downside of this is that we know of very few commercial codes, and certainly none in chemical text analysis, that provide public metrics let alone expose the architecture on which the program is based. Text-mining as an academic subject requires metrics and increasingly requires Openness of the components of the system, as we have done in OSCAR and OPSIN. We are investigating continuing business models where we can continue to re-factor and improve the product while not closing the code and therefore reducing scientific credibility and innovation.

Very recently we have been exploring the use of OSCAR for areas other than organic and biological chemistry. Because OSCAR can be customised by different dictionaries, we have been able to adapt it to process reports of atmospheric chemistry and, more generally, atmospheric science. In conjunction with the European Geosciences Union (EGU, which publishes Open Access papers), we have analysed abstracts and full text for chemical entities and related numerical quantities (*e.g.* amounts, conditions *etc.*) This has led to a design where the domain-independent parts of OSCAR4 can be applied to many physical sciences with bespoke dictionaries and ChemicalTagger rules. We have submitted grants in both the biosciences (“OSCAR-BIO”) and physical sciences (“OSCAR-PHYS”). As part of this work, we will be actively addressing generic tools for metrics and training.

---

<sup>43</sup> OSCAR4 Launch

## 40.3 Competing interests

The authors declare that they have no competing interests.

## 40.4 Authors' contributions

DMJ wrote the manuscript and was lead developer in the OSCAR4 re-factoring.

SEA was the architect and project manager for the OSCAR4 re-factoring and was involved in the original Experimental Data Checker project.

ELW contributed to the OSCAR4 re-factoring and investigated its use in Bioclipse.

LH contributed to the OSCAR4 re-factoring and was involved in the CheTA project.

PMR had the overall vision for, and was involved in, all stages of the various OSCAR projects, and wrote the manuscript.

All authors have read and approved the final version.

## 40.5 Appendixes

### 40.5.1 Appendix A: Additional OSCAR4 resources

The source code, mailing list, tutorials, documentation and support are available at

<https://bitbucket.org/wwmm/oscar4/wiki/Home>

This page also includes instructions for accessing pre-compiled JAR files from the Unilever Centre's Maven repository.

The source code used to measure OSCAR performance is available at <https://bitbucket.org/dmj30/oscar-performance>

The OSCAR4 Javadoc is available at <http://apidoc.ch.cam.ac.uk/oscar4-4.0.1>

### 40.5.2 Appendix B: Building on the OSCAR4 API

The core methods are given in each case. In some cases it will be valuable to extract further information recursively from the results.

1. Searching a given text for Named Entities. These can then be displayed, computed *etc.*

Oscar **oscar** = new Oscar();

List < NamedEntity >\*\*namedEntities\*\*

= **oscar**.\*\*findNamedEntities\*\*(**text**);

2. Where the named entity can be resolved to a chemical structure, extract it:

Oscar **oscar** = newOscar();

List < ResolvedNamedEntity >\*\*entities\*\*

= **oscar**.\*\*findAndResolveNamedEntities\*\*(**s**);

for (ResolvedNamedEntity **entity** : **entities**) {

ChemicalStructure **structure** = **entity**.\*\*getFirstChemicalStructure\*\*(<<http://>>‘\_(FormatType.INCHI));

...

}

3. Find only those entities which are resolvable to structures (*e.g.* “benzene” but not ” the methyl ester”):

```
Oscar oscar = newOscar();
```

```
List < ResolvedNamedEntity >**entities**
```

```
= oscar.findResolvableEntities(s);
```

4. Tailor the system to use different recognizers and dictionaries:

```
ChemicalEntityRecogniser myRecogniser = newPatternRecogniser()
```

```
Oscar oscar = newOscar();
```

```
oscar.setRecogniser(myRecogniser);
```

```
oscar.setDictionaryRegistry(myDictionaryRegistry);
```

```
List < ResolvedNamedEntity >**entities** = oscar.findResolvableEntities(s);
```

## 40.6 Acknowledgements

We gratefully acknowledge OMII-UK, JISC (ChETA project) and EPSRC (Sciborg, Pathways to Impact awards) for funding and Dr Charlotte Bolton for her assistance in the preparation of this article.



# THE SEMANTIC ARCHITECTURE OF THE WORLD-WIDE MOLECULAR MATRIX (WWMM)

## 41.1 Abstract

The World-Wide Molecular Matrix (WWMM) is a ten year project to create a peer-to-peer (P2P) system for the publication and collection of chemical objects, including over 250, 000 molecules. It has now been instantiated in a number of repositories which include data encoded in Chemical Markup Language (CML) and linked by URIs and RDF. The technical specification and implementation is now complete. We discuss the types of architecture required to implement nodes in the WWMM and consider the social issues involved in adoption.

## 41.2 Origins/history/vision

The World-Wide Molecular Matrix (WWMM) was conceived in 2001 in the spirit of the about-to-be launched UK eScience<sup>1</sup> programme and also the rapid and exciting success of peer-to-peer (P2P) systems in the music industry, such as Napster<sup>2</sup>. We interpreted the spirit of the age to be the dawn of a data- and knowledge-rich infosphere which would be self-evidently valuable to science and where every discipline would be actively publishing their data on the web. The vision was also inspired by the cyberpunk of William Gibson<sup>3</sup> and others with his idea of the information matrix where humans and machines would “jack-in” to an essentially infinitely large amount of information resources. This vision was 20 years ahead of its time but besides coining the term “cyberspace”, now has many features of today’s evolving web (“semantic web”) communities. It is from this, and not from the Matrix films<sup>4</sup>, that the word is borrowed with thanks. The concept is sufficiently compelling that others outside this group have set up a Wikipedia article on the WWMM<sup>5</sup>. Inspiration was also provided by the final session at WWW1<sup>6</sup> (1994) where Tim Berners-Lee outlined brilliantly how semantic information would drive and represent events in the real world, and the WWMM has tried to capture this for the domain of chemistry and related sciences.

We have often used the term “chemical semantic web” which is effectively synonymous with the WWMM, the preferred term in this article.

This article describes the evolution of the WWMM. Some of the early ideas (several of which were exposed in the eScience programme) were ahead of implementability but are now linked into general semantic web approaches. The

---

<sup>1</sup> Science & Technology Facilities Council e-Science

<sup>2</sup> Napster

<sup>3</sup> William Gibson

<sup>4</sup> The Matrix franchise

<sup>5</sup> World Wide Molecular Matrix

<sup>6</sup> First International Conference on the World-Wide Web

paper therefore represents an evolving vision of a distributed decentralised system.

The eScience programme held out the vision of a total network (“Grid”) of linked computing resources, with provision for high-speed access and interchange of data. We assumed that this would be a semantic network where many of the resources would not be bytes and CPU but would be structured information. We were grateful to receive early funding from the eScience project (“Molecular Standards for the Grid”<sup>7</sup>) but have been somewhat frustrated by the top-heavy concentration on CPU performance, bulk storage of un-semantic data and almost obsessive concentration on building middleware. The eScience programme, *per se*, contributed little to the semantic web in our fields.

Like many early ideas, it is impossible to predict the requirements for successful autonomous growth and it has taken approximately 10 years for the initial ideas of the WWMM to become an early reality today. The semantic web and, in its wake, the WWMM, have had to wait for the time to be right for them to flourish. This requires a complex mixture of different requirements:

- A widely-distributed toolchain in at least alpha.
- A critical mass of early adopters.
- A general realisation that this was an imperative whose time was bound to come.

In bioscience these ideas have been taken up at an early stage and many semantic resources have been created. There is a large amount of public investment in bioscience information technology driven in part by the Genome publications but also by the realisation that machines were going to be essential for discovery linking and simple inferences from semi-structured knowledge. We believed, optimistically and perhaps naively, that the same philosophy would be taken up in chemical disciplines. Some chemists had led the field of AI in the early 1970s (DENDRAL and CONGEN<sup>8</sup>, LHASA<sup>9</sup>) and it was natural to assume that chemistry would be a growing point for the semantic web.

In fact, there have been relatively few new conceptual developments in mainstream chemical informatics over the last decade or more. Apart from the development of InChI<sup>10</sup> (a semantic identifier system for connection tables), there has been very little central community interest in creating semantic resources. Many businesses and information providers take a 1980s model of capturing data (expensively), packaging it and re-selling it to the community. Similarly almost all publishers of chemistry are closed access and have determinedly remained so. This means that the data deluge expected in 2000 has failed to materialise in chemistry. The consequence of this is that not only is there no data to make semantic, there is little understanding in the community of the value of semantic data.

This situation is now changing. The semantic web is now reaching the high street and powerful commodity tools can be used for managing distributed linked data. Chemistry cannot ignore these developments. The “walled garden”<sup>11</sup> model of data is being shattered in governments, geospatial systems, music, libraries *etc.* where institutions are realising that to fulfil their roles they need to make their data Open and to make it semantic. There are still major cultural social commercial and political barriers; for example, the automated machine extraction of chemistry from electronic articles may result in a legal action by the publisher, and this attitude has held back the development of the WWMM by a considerable period.

In 2000, we envisaged that the technology would be based on P2P systems, where all nodes in the network would be equally able to receive and publish semantic data. The current evolution of the WWMM has been strongly influenced by the technologies in common everyday and business use, and now is much more likely to consist of servers and clients using REST (REpresentational State Transfer)<sup>12</sup> and similar philosophies for information exchange. The original vision however of a community-led process, sharing resources, is still at the heart of the WWMM.

The Napster and similar models worked because of a fortunate combination of circumstances. Almost all nodes were read-and-publish, in that they would consume information they wanted (music tracks) and would install a re-publication server as part of their “bargain” to the community. In addition, the metadata for music is relatively simple and was already widely used. The title of a track or artist generally identifies more or less precisely what is required.

<sup>7</sup> Molecular Standards for the Grid, Cambridge eScience Centre

<sup>8</sup> DENDRAL and CONGEN: Molecular Structure Elucidation in Organic Chemistry

<sup>9</sup> Logic and Heuristics Applied to Synthetic Analysis

<sup>10</sup> The IUPAC International Chemical Identifier (InChI™)

<sup>11</sup> Walled garden model

<sup>12</sup> Architectural Styles and the Design of Network-based Software Architectures

The P2P model survives in systems such as Skype<sup>13</sup> and BitTorrent<sup>14</sup> where owners of clients are prepared to pay for benefits in kind through offering bandwidth and services. A necessary requirement is that software is available which is almost transparent for the client to install and re-use.

The WWMM started with a more complex challenge. The metadata for molecules (and even more, chemical reactions, substances and properties) is not as simple as discovering music on the web. But the biggest challenge was that software would have to be written, which could be trivially distributed and where clients could legitimately and safely offer services without needing to know the details of installation. Nevertheless, the original (2001) concept has lost none of its validity. We envisage an ecology of sites (using a common syntactic and semantic infrastructure) which store a variety of objects in different numbers and with different attributes, and offer them to the world for re-use. Some sites can be expected to provide monocultural collections of certain types of object (*e.g.* molecules) while others might represent the work created in a particular institution. We also expect that there will be specialist sites for aggregating and indexing. This is a potential model for publication of data and metadata. In 2004 we had anticipated that some of the roles of the WWMM would be exemplified by the infrastructure and ecology of university institutional repositories but in reality these are poorly linked and there is no re-use and re-purposing of content.

It has become clear that in science domain-specific repositories are the appropriate model and in several fields there is a critical mass of adoption, support and contribution of content. Many of the bioscience repositories are managed by international data centres such as the NCBI (National Center for Biotechnology Information)<sup>15</sup> and EBI (European Bioinformatics Institute)<sup>16</sup>, but a newer generation of distributed, often university-based domain repositories are emerging. Two examples of these are Dryad<sup>17</sup> (where ecological content is deposited) and Tranche<sup>18</sup> (where proteomics data such as mass spectra are deposited<sup>19</sup>). These models are particularly compelling as it is now a requirement of several journals and publishers that data is committed to them. The WWMM is a technology that can respond to such requirements in chemical publishing.

By contrast, in chemistry, the only mandatory deposition of domain-specific data is in crystallography. Some of this is published openly on publishers' websites (and we use this in CrystalEye<sup>20</sup>), but approximately half of it is deposited directly in the CCDC. This has been a pioneering example of a domain repository but is now hampered by the fact that the data are not Open. While individual crystal structures can be requested by email, a considerable proportion of the raw data (in major journals) are only accessible in bulk by subscription. There are also restrictions on the re-use and re-publication of this data.

In science, repositories seem to work best where there is a central unifying concept found in every entry. For example, Swiss-Prot<sup>21</sup> is based on protein sequences, PDB<sup>22</sup> on protein structures and GenBank<sup>23</sup> on nucleic acid sequences. This may be, in part, because the repositories represent well-accepted concepts in the discipline and in most cases have an organisation or a committed group who oversees the semantics and ontology. It is necessarily a reductionist view and considerable flexibility and detail is lost, but at this stage in scientific information it is vastly better than having nothing at all.

## 41.3 Semantics and Ontologies in Molecular Sciences

The major current repositories of chemical information are generally run outside the community input of chemists and related disciplines. Very few are fully Open (exceptions being bioscience-based collections of molecules *e.g.* PubChem

<sup>13</sup> Skype

<sup>14</sup> BitTorrent

<sup>15</sup> National Center for Biotechnology Information

<sup>16</sup> European Bioinformatics Institut

<sup>17</sup> Dryad

<sup>18</sup> Tranche project

<sup>19</sup> Proteome Commons

<sup>20</sup> CrystalEye

<sup>21</sup> ExPASy Proteomics server (Swiss-Prot database)

<sup>22</sup> RCSB Protein Data Ban

<sup>23</sup> NCBI GenBank<sup>lnonascii\_15l</sup>

<sup>24</sup>, ChEBI <sup>25</sup>, NMRShiftDB <sup>26</sup> and CrystalEye, and the emerging collection in Wikipedia). There is a limited amount of Open data in ChemSpider <sup>27</sup> but Chemical Abstracts asserts copyright over its identifier system, does not publish its ontologies, and charges for lookup of names and identifiers.

For a full semantic implementation we need a variety of identifier systems with ontological mapping between them. We expect that, at some time, chemistry will develop a semantic infrastructure similar to that in current bioscience. In 2011, we note the development of semantic resources between Southampton and ChemSpider, but in general there is conservatism and resistance to the free flow of chemical information, and hence to the development of infrastructure. We have therefore taken a pragmatic view that much of what we implement can be done without formal ontologies and supported by a dictionary concept (see article in this issue).

Identifier systems are essential but very challenging. Semantic identifiers will always fail to represent general concepts, because the decision of which aspects of the concept are important to its identity are fixed, or from a fixed set (*e.g.* InChI). InChI doesn't fall short because of which information it chooses to include, it falls short because it chooses. In contrast, the CAS system is more flexible and can be assigned to a wide range of chemical substances. We recommend the use of relatively short alphanumeric identifiers. Chemists have a long tradition of using numeric identifiers (*e.g.* CAS, in-house compounds, regulatory labels *etc.*) and for most systems sequential numbering seems to be the best way of minting identifiers.

Arbitrary identifiers, however, require a central authority (even if only a server to mint the next in sequence). Without this, name collisions are certain. Moreover, without authorities to maintain identifier systems, they inevitably decay. We hope that this paper may stimulate persistent non-profit organisations (such as international scientific unions and learned societies) to create Open identifier systems. In the absence of this, the most likely solution will be through web persistence as in Wikipedia (though we note that that does not yet have a unique identifier system, being based completely on linguistic approaches).

In chemistry the “molecule” has become a central concept for aggregation. We note that there is much semantic and ontological confusion between substance, compound, connection table, and other concepts describing chemical objects and their composition. Thus, for example, the InChI only formally relates to a connection table, and works where there is a pragmatic correlation between connections tables and the composition of substances. It breaks down where a substance may contain components with different connection tables, where the connection table is dynamic, or where different substances can occur in different macroscopic forms. The technology of WWMM can support concepts such as molecule (connection table) and substance independently.

The WWMM paradigm relies on a unique identifier system for discovering and asserting the identity of objects. This works well where the connection table is a complete description and identification of the substance, but where it fails (*e.g.* “aluminium chloride”, “glucose”, “diamond”) we must rely on an authority to provide a controlled identifier system. The system in commonest use is the Chemical Abstracts registry number (CAS number)<sup>28</sup> but this is not Open and its use outside CAS is restricted to a small percentage of the compounds indexed by CAS. The best candidate for an Open system of substance identifiers is Wikipedia, which at the moment uses textual representations as the public unique identification of pages describing compounds. Until there is a public identifier system, the WWMM concept will be restricted to entries where connection tables suffice.

Figure 1 shows four sites all playing different roles in the WWMM. Site A is an aggregation site which trawls the web, either for other WWMM sites or legacy (white rectangles) and aggregates this in a similar manner to conventional search engines. The objects aggregated in the diagram are molecules with a variable number of properties (physical chemical and metadata). The concept can be extended to other chemical objects such as crystals, spectra, reactions and computational chemistry. In some cases we would have single instances of an object with several different properties (site A, right), while in other cases an object would be observed several times and have different instances of the properties (site A, bottom). Site B represents an archival site (*e.g.* the Internet Archive’s Wayback Machine <sup>29</sup>) where it would mirror for posterity the transient picture on site A. Site C represents data publication at source (*e.g.* our current

---

<sup>24</sup> PubChem

<sup>25</sup> ChEBI

<sup>26</sup> NMRShiftDB

<sup>27</sup> ChemSpider

<sup>28</sup> CAS

<sup>29</sup> Internet Archive’s Wayback Machine

CLaRION project<sup>30</sup> which is designed to publish scientific data from the laboratory to the web). The expectation is that visitors to the site (machines or humans) can then either assess the value of the site itself e.g. for data-oriented peer review, or can aggregate and re-use objects of interest. Site D specialises in one particular facet of objects or properties. This is exemplified by CrystalEye which trawls the web and extracts only crystal structures and collates and systematizes them.

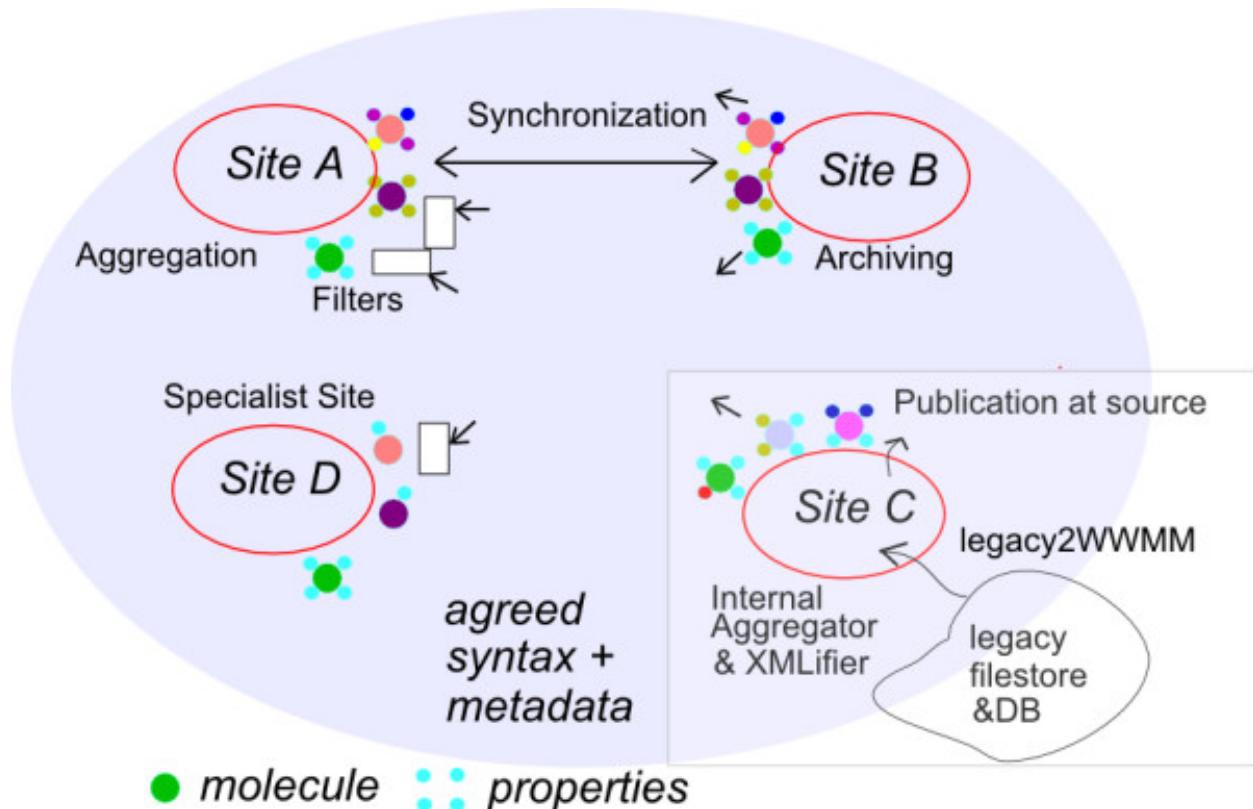


Figure 41.1: Figure 1. A 2004 vision of the variety of functions in the distributed sites of the WWMM concept  
**A 2004 vision of the variety of functions in the distributed sites of the WWMM concept.**

This vision, in 2004, was ahead of the technology to implement it, although we created some early prototypes of parts of the system. In both closed and Open systems, the successes have largely been through centralised sites (e.g. Google, Open StreetMap<sup>31</sup>, ChemSpider, DBpedia<sup>32</sup>). These have the value of coherency and visibility but can run into problems of scale and also potential frustration with central control. The P2P system is more flexible and allows a different type of innovation but is harder to reach to critical mass. It represents a general imperative for the web, a distributed non-hierarchical system of sites collecting and publishing data. This architecture is reflected in both the Quixote project<sup>33</sup> and the OpenBibliography project<sup>34</sup> reported elsewhere in this issue. It is clearly difficult to create off-the-shelf software for these types of system, but we believe that an investment in RDF, a very strong investment in all types of metadata in the system, and, most importantly, a critical mass of a community prepared to explore this will come up with prototypes which show the value.

The WWMM is also designed to hold properties of chemical molecules and substances. In many cases, these concepts are very well defined and managed by community definitions such as the IUPAC Gold Book<sup>35</sup>. However, there

<sup>30</sup> CLaRIO

<sup>31</sup> Open StreetMap

<sup>32</sup> DBpedia

<sup>33</sup> Quixote project on QC databases

<sup>34</sup> Open bibliography and Open Bibliographic Dat

<sup>35</sup> IUPAC Compendium of Chemical Terminology - the Gold Book

is much opportunity for confusion: scientific units of measurement are often omitted and physical constraints (*e.g.* pressure at which a boiling point was measured) are not recorded. In some cases it is unclear what the molar unit is. For example, some programs calculate the extensive properties for a complete unit cell (*e.g.* Na<sub>4</sub>Cl<sub>4</sub>). These properties are supported by a system of dictionaries (see the sibling article in this issue).

Concepts which are relations between objects (*e.g.* chemical reactions and processes, such as chemical syntheses) have been excluded from the initial version of the WWMM until their semantic representation has been more explored within the community.

In a distributed system, there is a major challenge of different versions of the “same” object. Traditionally and currently many systems tackle this by creating a canonical “correct” object by merging different versions into one. Systems such as CrystalEye work well because although there are a variety of sources, there is only one agreed instance of the crystal structure publication. Sites such as ChemSpider normalise chemical names and identities by correcting “wrong” names and structures. Building a more complex system than this is psychologically difficult with the dangers of either oversimplistic representation through normalisation or over-complication of the details of different occurrences of objects. A typical problem is the management of the different versions over time and location of human-authored documents.

Figure 2 illustrates this problem: mol1 exists in Alice but not Bob, and mol5 exists in Bob not Alice. mol2 exists in both, but Alice has more properties (attributes), and mol4 is the reverse. mol3 is identical in both repositories. The arrows show various updating processes so that Bob will need to import mol1 and all its properties to be in sync, and Alice must do this for mol5. For mol2 and mol4 each site would have to import properties and keep them in sync, whilst for mol3 only the values of properties need to be synchronised. In practice it is likely that Alice and Bob will not synchronise at this level and it is up to users of their sites to determine existence of entries and of properties, and the identity relations.

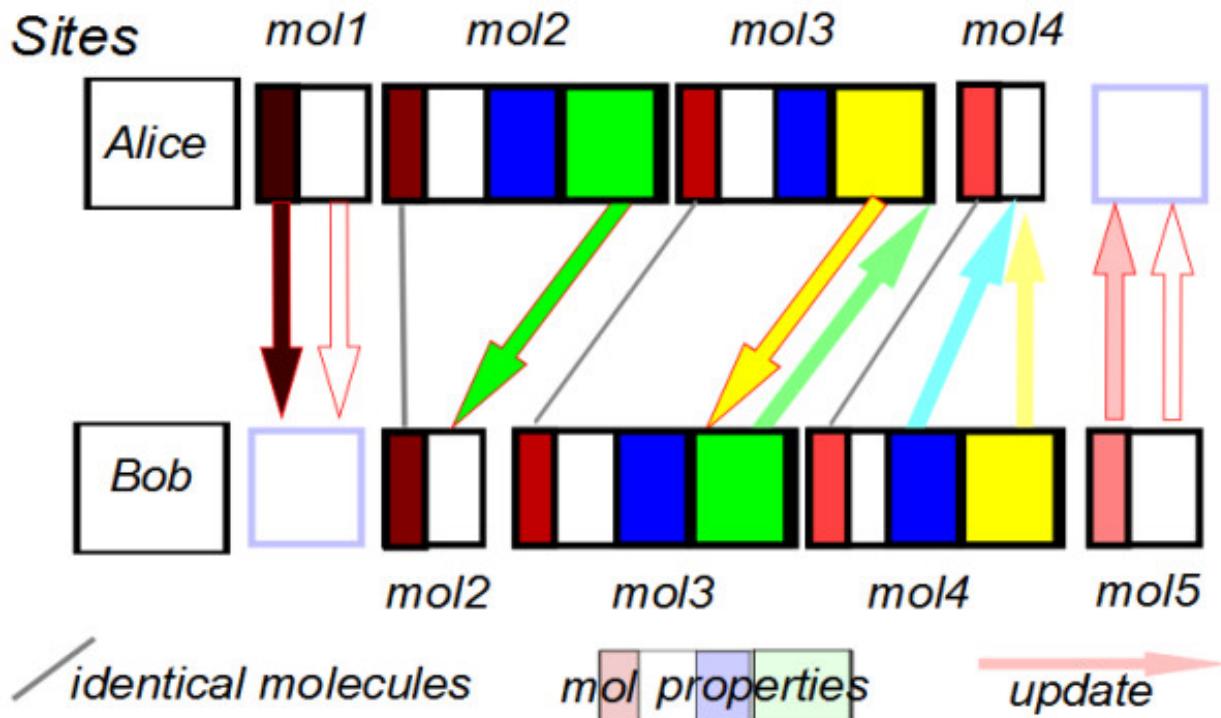


Figure 41.2: Figure 2. An example of the problems of different entries and different versions in two repositories  
**An example of the problems of different entries and different versions in two repositories.**

Because of this, we think that the CrystalEye and Quixote systems are excellent examples of systems that can succeed as distributed WWMM repositories. In CrystalEye the uniqueness is determined by the bibliographic data of the publication (or the metadata from the creators). In Quixote a calculation is the same regardless of which laboratory

carries it out and duplications of calculations have the same canonical representation. We believe that there will be a demand for molecules from CrystalEye and Quixote and that these will be excellent exemplars and workbench for crystallographers, scientists and computational scientists interested in P2P systems and distributed repositories.

The Linked Open Data (LOD) concept and movement has demonstrated the vision of a cloud of interlinked resources, and many of the bioscience databases feature prominently. In 2011, there are still very few Open data resources in chemistry. To be a full member of this graph, a resource has to have a public identifier system (URI) and a license that allows essentially total freedom of access and re-use. “Free resources” (where there is no right of re-use) cannot be included. The following current resources could be transformed into LOD nodes:

- Bioscience databases (PDB, Uniprot<sup>36</sup>, KEGG [#B37]\_\*etc\*)
- NMRShiftDB (a volunteer-driven collection of Open NMR spectra)
- A subset of the ChemSpider resource (a small percentage of items are now labelled as ‘Open Data’ and there is a stable identifier system)
- Chemical entries in Wikipedia
- CrystalEye (semantic crystal structures from Openly published data)
- Computational chemistry from the Quixote project

The data so far has been primarily from current aggregators and voluntary collections. The WWMM concept also included the idea that scientists would publish their data directly onto the web as they carried out experiments or calculations. Although a very small percentage of the community, chemistry has been among the leaders in developing this idea and J-C Bradley<sup>37</sup> and, more recently, Matt Todd in Sydney<sup>38</sup> and the Frey group at Southampton<sup>39</sup> have published tools and data onto the public web. In particular, Todd’s community of collaborative drug design has attracted considerable interest and, assuming it is successful, will be a strong driver to show the value of the semantic web and WWMM approach. The concept of “the contents” of a site can be problematic. At one level, chemists think of collections of molecules as a large defined collection of molecular datafiles which could in principle be distributed on a memory device or published as individual pages on the web. In other circumstances, molecules and their properties are retrieved from a search system. In yet other applications, molecules can be generated “on-the-fly” by web services (an example is our OPSIN server<sup>40</sup> which converts IUPAC names into connection tables and effectively has an infinite number of possible molecules). In many closed systems it is impossible to tell whether particular data is in the system unless the interface allows us to ask this question. The WWMM is conceptually designed as an infrastructure where all of the content can be systematically retrieved and the limitations are technical rather than socio-political.

## 41.4 Design and evolution: technologies

Tim Berners-Lee originally introduced the four principles of linked data<sup>41</sup>:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF\*, SPARQL).
4. Include links to other URIs so that they can discover more things.

The modern WWMM adopts principle 1 completely. All things including not only data but metadata such as dictionaries are completely supported by URIs. Principle 2 brings certain problems. In the initial design of HTML and XML there was a strong architectural differences between URLs (addresses) and URIs (identifiers) and this formal

<sup>36</sup> Universal Protein Resourc

<sup>37</sup> Jean-Claude Bradle

<sup>38</sup> Matthew Tod

<sup>39</sup> Jeremy Fre

<sup>40</sup> OPSI

<sup>41</sup> Tim Berners-Lee’s Four Principles of Linked Data

distinction has to remain in many fields. Tim Berners-Lee simplified this to principle 2 on the basis that everything of interest could have both an address and a URI, and that they could be conflated into the same string. For this to happen, the identified object must be sufficiently stable and conceptually bounded that it is effectively describable as a single persistent object. (There are ontological systems which can describe non-persistent and mutable objects but they are beyond the current scope of chemistry and the WWMM.) The single address requirement can also be problematic. Principle 2 only fails to break when the user or user agent has pervasive access to the web (*e.g.* not in an aeroplane) and where the maintainer of the resource can guarantee 24/7 availability. If this latter condition cannot be met, then either the system breaks (perhaps temporarily) or it has to provide a fall-through mechanism of aliased addresses. CML was originally designed with the clear W3C principle that names and addresses were distinct but we are attracted by the conflated URI vision which we believe will work for much of chemistry. Given at least a partial implementation of principle 2, then we endeavour to satisfy principle 3 by using RDF and SPARQL where appropriate. Principle 4 is a fundamental part of the WWMM and follows practice in, for example, bioscience where most resources have copious links to others. Whether or not resources are normalised is a problem that we have not yet explored in depth.

More recently, discussions on the eGov W3C mailing list<sup>42</sup> refer to Tim Berners-Lee's "five star" model for government data:

- on the web, open license
- \*\* machine-readable data
- \*\*\* non-proprietary formats
- \*\*\*\* RDF standards
- \* Linked RDF

Semantic data requires a minimum of an identifier system. Many published collections of information do not generate identifiers and are only accessible and identifiable through their web addresses. This is a fragile design and it is essential that components of the WWMM have unique permanent identifiers. The traditional use of chemical identifiers has been restricted to large authorities such as CAS, Beilstein<sup>43</sup>, RTECS<sup>44</sup>, and more recently, ChemSpider, PubChem, DrugBank<sup>45</sup>, ChEBI and ChEMBL<sup>46</sup>. Of these, we believe that only the bioscience-oriented systems (ChEBI, ChEMBL, PubChem) are formally Open (*i.e.* that the whole identifier system, with or without the data, can be re-used without permission). There are a small number of spectral identifiers in NMRShiftDB and a small number of reaction identifiers in KEGG, confined to biological transformations. The CrystalEye collection does not have an identifier system yet although the Crystallography Open Database (COD)<sup>47</sup> does. There is no Open system for small molecule crystallographic identifiers (the CCDC<sup>48</sup> codes are for a closed system).

In principle, LOD can be completely held as RDF triples. However, many components of chemistry (molecules, spectra, reactions *etc.*) are more easily understood and processed in XML form (*e.g.* CML). The WWMM, therefore, is a mixture of CML components linked together and annotated by RDF triples. As the semantic web develops new approaches to indexing and describing RDF we can expect the flavour of RDF to evolve. In 2011, it is still unclear exactly what triple-store or other RDF technology is required to support large amounts of RDF, but we believe that for local collections (*e.g.* the output of a laboratory) there are now many good OS RDF engines.

We have built prototype ontologies with formal RDF-based systems such as OWL<sup>49</sup>, and developed an OWL-based system (ChemAxiom<sup>50</sup>) which describes physical properties and aspects of chemical structure and composition. At present, however, we believe the implementation cost (validating the ontology, installing sufficiently powerful servers) not to be cost-effective. This parallels our experience in Open Bibliography (see sibling article in this issue) where the implementation costs were too large to be deployable without additional resource, and we reverted to a simpler model

---

<sup>42</sup> W3C eGov mailing list archive

<sup>43</sup> Beilstein (since 2009, distributed as Reaxys by Elsevier)

<sup>44</sup> RTEC

<sup>45</sup> DrugBank

<sup>46</sup> ChEMBL database

<sup>47</sup> Crystallography Open Database

<sup>48</sup> Cambridge Crystallographic Data Centr

<sup>49</sup> OWL2 Web Ontology Language

<sup>50</sup> ChemAxiom - An Ontological Framework for Chemistry in Science

with some implicit semantics. There is also a psychological barrier in that many scientists working with chemical information need to feel comfortable with the textual representation.

To be semantic, the information must be understandable by machines and humans. In the full semantic web vision, this is (partially) provided by high-level ontological frameworks such as OWL, OBO<sup>51</sup>, Cyc [#B53]\_\*etc\*. In WWMM we take the view that semantics can be provided by a number of inter-operating dictionaries which describe the semantics in human terms and also provide a variety of machine-enforceable constraints and interpretations. These work at a pragmatic rather than a formal level. The success of the WWMM will depend in part on the willingness of the community to create such dictionaries and to make sure that material produced uses the dictionary URIs in its annotation. Unlike all current knowledge bases in chemistry, the WWMM will not have a central repository and service. Like peer-to-peer systems we expect that there will be a federation of repositories adopting common identifier systems and semantics. We do not believe that traditional institutional repositories are the most appropriate place to deposit scientific data, and strongly believe that domain-oriented approaches are required. A scientist wishes to interact with a repository that understands her problem, not with the organisation that happens to employ her. Because chemistry is a multi-disciplinary subject we expect that the WWMM will consist of a considerable number of independent nodes. There is no requirement that any given repository holds “all” the data, nor that data should not be duplicated in different nodes. We expect that the community will evolve systems that make sense in terms of ease of access and robustness.

It will be fundamental to have an indexing and discovery system. Because of the non-textual nature of much chemistry, current search engines such as Bing and Google will not be able to index much of the WWMM material. We therefore need distributed search technologies and in the first instance will rely on RDF and on conventional chemical substructure search. We have designed the system such that it is possible for scientists to add indexers (plug-ins) to a repository to create domain-specific searchable metadata. For example, it is not easy to search on the web for a compound containing between 10-15 carbon atoms, but if a repository exposes a carbon-count field as RDF then it is straightforward to retrieve entries using an RDF query containing combinations of index fields. More complex chemical concepts can also be indexed, such as peaks in NMR spectra, cavities in crystals or HOMO-LUMO gaps in theoretical calculations.

The architecture of the WWMM is built on a number of web standards and protocols, described in detail below:

|nonascii\_10| **SWORD deposit**: publish data to server

|nonascii\_11| **Atom archive feeds**: syndicate published data

|nonascii\_12| **HTTP content negotiation**: retrieve data in human and machine understandable formats

|nonascii\_13| **OAI-ORE/RDF**: machine understandable representation of the data

#### 41.4.1 SWORD/AtomPub

The Atom Publishing Protocol (AtomPub<sup>52</sup>) provides a standardised application-level protocol for publishing and editing Web Resources using HTTP. AtomPub is applicable to many domains, but is particularly widely supported by the Blogosphere, where it enables authoring tools such as Microsoft Word to publish content to different blogging software using a common protocol. The JISC-funded SWORD (Simple Web-service Offering Repository Deposit) project<sup>53</sup> extends the AtomPub protocol to support the deposit of aggregate resources - packages consisting of a number of related files and associated metadata - onto a server. For example, the ‘package’ object needed by WWMM may include crystal structure (CIF and CML formats), picture of the 3D structure, and a 2D representation of the connection table *etc*.

<sup>51</sup> The Open Biological and Biomedical Ontologie

<sup>52</sup> The Atom publishing protocol

<sup>53</sup> Simple Web-service Offering Repository Depositi

## 41.4.2 Atom/RSS Feeds

Web feeds are widely used to provide users with notifications of updated content. Typically a feed document lists recent content - such as active news items, or the list of articles in the current issue of a journal - and by monitoring (“subscribing to”) a feed, users can be alerted when new content is published. Entries in a feed document typically contain the title and summary of an item, along with a link to the full resource. Earlier iterations of the WWMM made use of RSS feeds<sup>54</sup> to alert users of newly published chemical data, but these suffer from the constraint that only the most recent content can be accessed. The Atom Syndication Format offers a solution to this limitation through standardized support for paging, specified by RFC5005. Like an RSS feed, the Atom feed’s document contains a list of recently updated content, however it can also contain a link to a previous page containing entries describing other content. A client application can always access the latest content by retrieving the document at the feed URL, but can ‘walk’ back through the previous pages to discover all the content in the system (Figures 3 and 4).

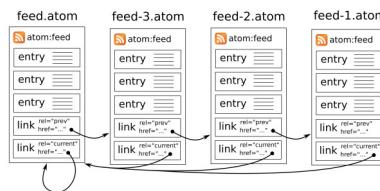


Figure 41.3: Figure 3. An example of “paging” using Atom feeds  
**An example of “paging” using Atom feeds.**

```

http://blog.example.com/feed.atom
<feed ...>
  <!-- entries describing recently published content -->
  <link rel="current" href="http://blog.example.com/feed.atom" />
  <link rel="prev-archive" href="http://blog.example.com/feed-3.atom" />
</feed>

http://blog.example.com/feed-3.atom
<feed xmlns="http://www.w3.org/2005/Atom"
      xmlns:fn="http://purl.org/syndication/history/1.0/">
  <!-- flag indicating that this is an archive document, whose
      list of entries will not change -->
  <fn:archive />
  <!-- entries describing previously updated content -->
  <link rel="current" href="http://blog.example.com/feed.atom" />
  <link rel="prev-archive" href="http://blog.example.com/feed-2.atom" />
</feed>
  
```

Figure 41.4: Figure 4. Atom feed content, based on the example in Figure 3 above  
**Atom feed content, based on the example in Figure 3 above.**

## 41.4.3 HTTP Content Negotiation

The HTTP protocol<sup>55</sup> allows content providers to deliver alternative representations (*e.g.* multiple languages, data formats, size, resolution *etc.*) of a resource (*i.e.* a data object or service identified by a URI) from the same URI, based on the preferences expressed by a client, through a mechanism called content negotiation (Figure 5).

When requesting a resource from a web server, a client may include an ‘Accept’ header in the request, indicating the media types it prefers, and optionally a strength of preference; for example, Mozilla Firefox 4.0.1 uses the following header:

Accept: text/html, application/xhtml+xml, application/xml;q = 0.9, \*/q = 0.8

This says that Firefox prefers HTML (text/html) and XHTML (application/xhtml+xml) content, or less strongly (q = 0.9) XML (application/xml). If none of these are available it will accept anything else (/).

<sup>54</sup> Chemical Markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators

<sup>55</sup> HTTP protocol

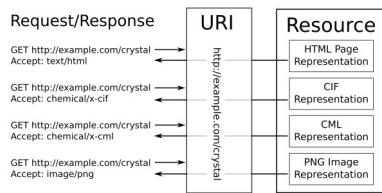


Figure 41.5: Figure 5. HTTP content negotiation delivering different representations of the same URI, based on the content of the ‘Accept’ header

**HTTP content negotiation delivering different representations of the same URI, based on the content of the ‘Accept’ header.**

The WWMM uses content negotiation to publish data in formats that are both human and machine readable. The URI for a resource published on the WWMM can be resolved to alternative representations - an HTML or XHTML ‘splash’ page for humans, or an RDF representation (`application/rdf+xml`) for machines.

This request by a web browser (such as Mozilla Firefox)

<http://crystaleye.ch.cam.ac.uk>

will return a human-friendly HTML page describing the crystal structure, while the following request for the same resource by a machine agent

<http://crystaleye.ch.cam.ac.uk>

can return a machine understandable RDF representation of the data.

#### 41.4.4 OAI-ORE/RDF

Open Archives Initiative Object Reuse and Exchange (OAI-ORE)<sup>56</sup> is a standard for describing aggregations of Web resources, commonly serialized into RDF. The WWMM uses OAI-ORE to describe the resources making up a data item - *e.g.* a crystal structure of NMR spectrum and the aggregate resource (Figure 6).

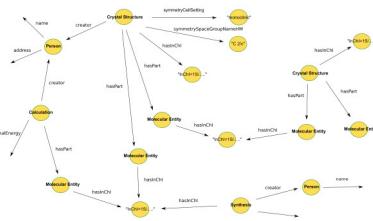


Figure 41.6: Figure 6. Using an RDF representation for data items such as crystal structures and calculations enables them to be connected by shared concepts (InChI, creator) to form a graph of linked data

**Using an RDF representation for data items such as crystal structures and calculations enables them to be connected by shared concepts (InChI, creator) to form a graph of linked data.**

The OAI-ORE model includes three classes of object: Aggregation, Aggregated Resource and Resource Map. Aggregations are an abstract concept, containing one or more Aggregated Resource. An Aggregation may be serialized into a number of different formats, and each of these serializations is termed a Resource Map. Each Resource Map has a unique URI, distinct from the Aggregation’s URI, in order for the different representations of the Aggregation to be resolvable.

As well as describing an Aggregation, a Resource Map may contain additional data about the Aggregation and the individual Aggregated Resources (Figure 7).

<sup>56</sup> Open Archives Initiative Object Reuse and Exchange

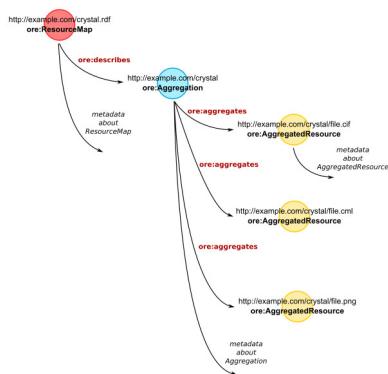


Figure 41.7: Figure 7. The structure of an RDF representation of an ORE resource map describing an aggregation of related resources and associated metadata

**The structure of an RDF representation of an ORE resource map describing an aggregation of related resources and associated metadata.**

## 41.5 Software development environment

We have developed a large number of software components of varying complexity with much inter-dependence between the components. We embrace agile development practices and our software development environment is built upon existing technologies. Substantial use is made of existing Open Source utilities, tools and libraries such as Apache Commons<sup>57</sup>, Restlet<sup>58</sup> and CDK<sup>59</sup>. The majority of the code is written in Java and we use the Apache maven<sup>60</sup> build system (compiles, manages dependencies *etc.*)

We endeavour to write the code with high test coverage (as much unit testing as possible is built-in at the initial stages), and aim for test-driven development. We run a Jenkins continuous integration<sup>61</sup> server and a Nexus maven repository so all the code is developed under source control (a mixture of Subversion (svn)<sup>62</sup> and mercurial<sup>63</sup>). The Jenkins server polls the source repositories at regular intervals and rebuilds and tests any updated projects in a clean environment. If the updated code compiles successfully and passes all the unit tests, it is deployed to the maven repository and any downstream (dependent) projects are then re-compiled/re-tested in the same way. Thus, any modifications which would break compatibility with any other components are flagged, identified and rectified at the earliest possible opportunity (Figure 8).

## 41.6 Virtual communities

The WWMM is predicated on a critical mass of users who are prepared to develop both content and technology. The precise path for the evolution will depend on a mixture of what technologies are available, the familiarity of the community and the resources that are available. It may also depend on the perceived business models and the uptake by significant producers and consumers. The earliest experiment is in our Quixote community where we are producing semantic computational chemistry and disseminating this from RDF-aware servers. It is likely that different members of the community will play different roles. Some may wish to upload their results to a semantifier which deposits them in a given repository (“push”). Others may wish to aggregate legacy data and re-disseminate it (“pull”). For example, a University Department or group might wish to expose its results on its own webpages to enhance the reputation

<sup>57</sup> Apache Commons

<sup>58</sup> Restlet

<sup>59</sup> Chemistry Development Kit

<sup>60</sup> Apache Maven project

<sup>61</sup> Jenkins-ci development environment

<sup>62</sup> Apache™ Subversion<sup>nonascii\_171</sup> project

<sup>63</sup> Mercurial source control management tool



Figure 41.8: Figure 8. Current status page of the ‘new-jumbo-converters’ project on the Jenkins continuous integration server, showing its relationships to other projects

**Current status page of the ‘new-jumbo-converters’ project on the Jenkins continuous integration server, showing its relationships to other projects.**

and provide re-usable material. A national lab might act as an aggregator for a sub-community of scientists (*e.g.* in materials properties prediction).

## 41.7 Future Development of the WWMM

The future of the WWMM will depend on a number of factors which we cannot predict:

1. The change from “walled garden” providers to Open collections.
2. Citizen science. The most dramatic example in science has been the large collaboration involved in GalaxyZoo<sup>64</sup>, and now spreading to other types of activity (Zooniverse<sup>65</sup>). We expect and hope that this philosophy will spread to chemistry and disciplines which require chemistry.
3. The need to link data. Almost all chemical systems at the moment are unsuitable for LOD in both the lack of semantics and the problems of licences. The ChemSpider system is a hybrid in that some of the data are Open and some of the material is exposed in RDF.
4. The realisation that chemistry needs community ontologies.
5. The high and unsustainable cost of closed data collections.
6. The growing dissatisfaction of the upcoming generation of scientists with closed systems.
7. The frustration of the non-academic community in the difficulty of obtaining material published in STM publications.
8. The desire of scientific publishers and editorial boards to publish the semantic data associated with articles.

We understand that the initial collection of 175, 000 molecules in DSpace@Cambridge<sup>66</sup> is regularly used and accounts for ca. 10% of the repository traffic. This further encourages us to believe that decentralised resources are valuable and can be discovered and used by current web technology.

The WWMM is now technically deployable and its critical mass will depend on adopters who need an Open distributed system, and who are prepared to contribute to the infrastructure design, the ontology design and its implementation.

## 41.8 Competing interests

The authors declare that they have no competing interests.

## 41.9 Authors' contributions

PMR had the original vision for the WWMM, has participated in and overseen its development and has written the manuscript. SEA has been responsible for much of the current infrastructure and has written the manuscript. OJD has contributed to the development of the infrastructure, provided valuable advice on technical matters and has written the manuscript. JAT was involved in development of the original infrastructure and content production and has written the manuscript. YZ was involved in the development of the original infrastructure and content production. All authors have seen and approved the final version.

---

<sup>64</sup> GalaxyZoo

<sup>65</sup> Zooniverse

<sup>66</sup> DSpace@Cambridge repository

## 41.10 Acknowledgements

We thank the following for funding over the lifetime of the WWMM: Unilever, DTI eScience, JISC. The assistance of Dr Charlotte Bolton in the production of this manuscript is gratefully acknowledged.



# THE SEMANTICS OF CHEMICAL MARKUP LANGUAGE (CML): DICTIONARIES AND CONVENTIONS

## 42.1 Abstract

The semantic architecture of CML consists of conventions, dictionaries and units. The conventions conform to a top-level specification and each convention can constrain compliant documents through machine-processing (validation). Dictionaries conform to a dictionary specification which also imposes machine validation on the dictionaries. Each dictionary can also be used to validate data in a CML document, and provide human-readable descriptions. An additional set of conventions and dictionaries are used to support scientific units. All conventions, dictionaries and dictionary elements are identifiable and addressable through unique URIs.

## 42.2 Introduction

From an early stage, Chemical Markup Language (CML) was designed so that it could accommodate an indefinitely large amount of chemical and related concepts. This objective has been achieved by developing a dictionary mechanism where many of the semantics are added not through hard-coded elements and attributes but by linking to semantic dictionaries. CML has a number of objects and object containers which are abstract and which can be used to represent the structure and datatype of objects. The meaning of these, both for humans and machines, is then realised by linking an appropriate element in a dictionary.

The dictionary approach was inspired by the CIF dictionaries<sup>1</sup> from the International Union of Crystallography (IUCr) and has a similar (in many places isomorphous) structure to that project. The design allows for an indefinitely large number of dictionaries created by communities within chemistry who recognise a common semantic approach and who are prepared to create the appropriate dictionaries. At an early stage, CML provided for this with the concept of “convention”. This attribute is an indication that the current element and its descendants obey semantics defined by a group of scientists using a particularly unique label (Figure 1).

During the evolution of CML we explored a number of syntactic approaches to representing and imposing semantics through dictionaries. These have ranged from a formally controlled ontology (ChemAxiom<sup>2</sup>) which is consistent with OWL2.0<sup>3</sup> and the biosciences’ Open Biological and Biomedical Ontologies (OBO)<sup>4</sup> framework, to uncontrolled folksonomy-like tagging. Although we have implemented ChemAxiom and it is part of the bioscientists’ description of

---

<sup>1</sup> IUCr CIF dictionaries

<sup>2</sup> ChemAxiom - An Ontological Framework for Chemistry in Science

<sup>3</sup> OWL 2 Web Ontology Language

<sup>4</sup> Open Biological and Biomedical Ontologies (OBO)

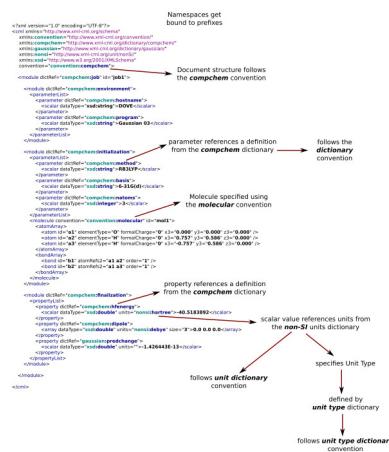


Figure 42.1: Figure 1. The primary semantic components of CML

**The primary semantic components of CML.** Elements in a document link to conventions, dictionaries and units through attributes. The referenced resources are themselves constrained by specification documents (convention spec, dictionary spec, system of units) with unique URIs. Within the dictionaries and the unit collections, every entry has a unique ID and when combined with the dictionary URI produces a globally-unique identifier.

chemistry, we regard it as too challenging for the current practice of chemistry and unnecessary for its communication. This is because chemistry has a well-understood (albeit implicit) ontology and the last 15 years have confirmed that it is highly stable. The power of declaration logic is therefore not required in building semantic structures. The consequence is that some of the mechanics of the semantics must be hard-coded, but this is a relatively small part and primarily consists of the linking mechanism and the treatment of scientific units of measurement. At the other end of the spectrum, we have found that the folksonomy approach is difficult to control without at least some formal semantic labelling. We have also found that there is considerable variation in how sub-communities approach their subject, and we do not wish to be prescriptive (even if we could). For example, the computational solids group (CMLComp) insisted that a molecule should not contain bonds as they did not exist, whereas the chemical informatics community is concerned not only that bonds should exist but that they should be annotated with their formal bond order.

The design of CML has always been based on the need for dictionaries, and has also recognised that there are different conventions within chemical practice. The original design (Figure 2) shows the linked dictionary concept and this has proved resilient and is the basis of the current architecture. However, the precise representation has varied over the years. This article represents a convergence and crystallisation of the semantic environment of CML, and we believe that there are now no immediate requirements for early refinement. This paper can therefore be used, we hope, for several years as a reference in a more robust manner than has been possible up to now. However, the exact practice of the CML community will be primarily governed by public discussions on mailing lists and formal releases of software and specifications.

This practice and principles are general to all the semantic elements in this article, and is best illustrated in the requirements for creating a convention and enforcing it. In the spirit of communal development, any sub-community is at liberty to create their own convention without formal permission from any central governance, subject to the requirement that it must be valid against the (very flexible) CML Schema 3<sup>5</sup>. This is done by associating the convention with a unique namespace identifier and the convention specification shows how this must be done, but does not dictate the contents or scope of any convention. In this way, an indefinite number of sub-communities can develop and ‘do their own thing’ without breaking the CML semantics. The success of a convention is then a social, not technical, phenomenon. If group A develops a convention and groups B, C and D adopt it then there is wide interoperability. If A develops a convention and B develops an alternative then there is fragmentation. It’s not always a bad thing to have “more than one way to do it”<sup>6</sup>, but it can make life very complex for software developers.

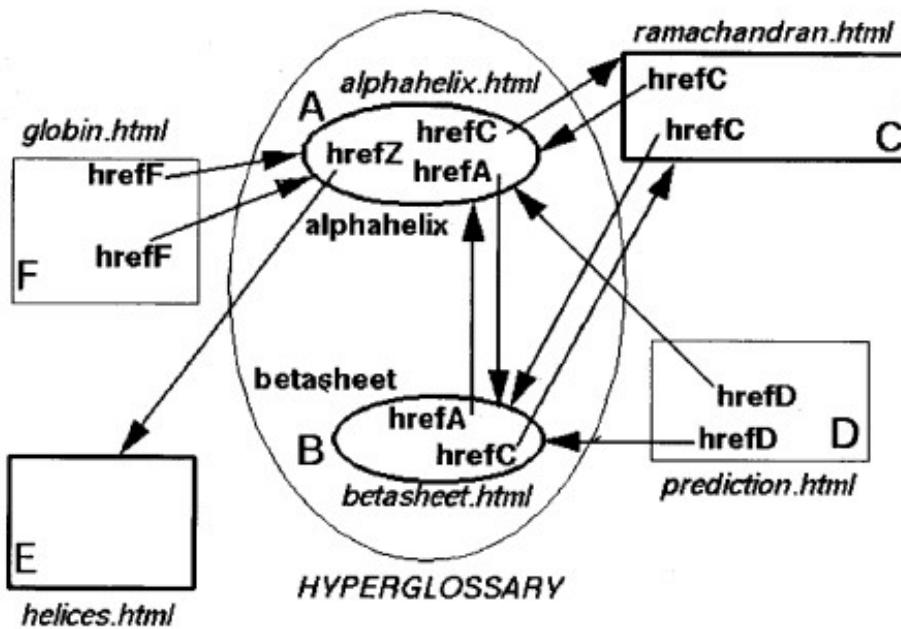
The price for this freedom is that a community cannot by default expect other users of CML to adopt their convention.

<sup>5</sup> CML Schema 3

<sup>6</sup> “There’s more than one way to do it”, Perl motto

## How the HyperGlossary works

Here is a schematic diagram of how the hyperglossary concept works. There might be six authors, A, B, C, D, E, F, and Z (the curator), who need have *no communication* with each other. Here is a hypothetical story.



- A writes a glossary item **alphahelix**
- B writes a glossary item **betasheet**
- E writes her own article on **helices** with no markup
- F writes his own article on **globins** with no markup

Figure 42.2: Figure 2. The original design for CML semantic architecture (1996)

The original design for CML semantic architecture (1996). This shows how different groups can create their own semantics and inter-operate. The concept has been proven over 15 years with appropriate changes to the terminology (*i.e.* we now talk of linked metadata rather than a hyperglossary).

If a community wishes its convention to be used, it needs to educate it in how CML can support it, and almost always to create or re-use software to support the convention. Thus, for example, the CMLSpect convention is supported by the JSpecView <sup>7</sup> software, which has a vigorous community of practice. Similarly, the CMLCryst convention (not yet released) is being driven by the development of the CrystalEye <sup>8</sup> knowledgebase and its adoption by the IUCr.

The dictionary reference mechanism (the *i.e.* it has a prefix as well as a local name. Although this approach is not formally supported by XML, it is widespread in approaches such as XSD Schema. This has turned out to be a valuable design as it is isomorphic to the use of namespaced URIs and indeed the <sup>9</sup> and that both the namespace and the local entry should be resolvable.

The role attribute has been used for a variety of purposes in the past but is now developed as a general “tagging” tool. A typical example is shown in the ‘Roles’ section below.

The semantic tools (dictionary, convention and role) have been fluid over the last decade and there are examples where their use is not compatible with this paper. However, the tools to support them will work with modern CML libraries.

The current tools in CML for adding semantics are therefore:

- **convention**
- **dictRef**
- **role**
- **units\*\*\*\*and unitType**

We now discuss each of these approaches in detail.

## 42.3 Semantic Elements of CML

### 42.3.1 Convention

The initial (1996) use of convention was limited to certain elements such as bond to represent the different values that different communities might use. It has now grown to be a key concept in defining communities of practice, having started to be used *ca.* 2005 when individuals and groups worked to create sub-domains of CML. The leading areas were reactions (mainly enzymes), spectroscopy, crystallography and computational chemistry (compchem). It emerged from these exercises that the elements and attributes of CML were sufficient to support the sub-community but that additional semantics in their use and constraints was necessary. Thus, for example, the CMLSpect <sup>10</sup> community decided that a spectrum must have a child representing the data in the spectrum (it is still possible to have an empty spectrum in CML but it would be used by a different community for a different purpose).

Conventions specify a minimal set of elements and document structure that a community has agreed to. Other elements may be included in a document, but may be transparently ignored by processing software.

Thus, a convention offers the following:

- an announcement that an identified community cares about a sub-domain of chemistry.
- a prose description of the scope and constraints and practice of the convention.
- a validator <sup>11</sup> that determines whether a given document conforms to a convention (and where it deviates).

In addition for software developers it offers:

- a statement as to what the components in a convention are, and how they can be combined.

---

<sup>7</sup> JSpecView software

<sup>8</sup> CrystalEye

<sup>9</sup> Namespaces in XML, QName

<sup>10</sup> Chemical markup, XML, and the world wide web. 7. CMLSpect, an XML vocabulary for spectral data

<sup>11</sup> CML validators

- indications of what constraints may/must/should be imposed on CML documents valid against this convention.
- an indication or a guarantee as to what CML components may be found in a conformant document.
- an indication of their semantics.

CML Schema 3 is less restrictive than Schema 2.4<sup>12</sup> and is designed to be used in conjunction with conventions. The loosening of the restrictions in the schema mean that it is schema-valid to create documents which do not make chemical sense (such as molecules being the children of atoms and bonds being defined in a molecule with no atoms present). The chemical validity and constraints are now imposed through the use of conventions and XSLT/XPath. <http://www.ietf.org/rfc/rfc2119.txt> be a convention document describing a convention.

Currently supported conventions (see Figure 1) are:

- <http://www.xml-cml.org/convention/dictionary>).
- <http://www.xml-cml.org/convention/molecular>).
- <http://www.xml-cml.org/convention/compchem>).
- <http://www.xml-cml.org/convention/unit-dictionary>).
- <http://www.xml-cml.org/convention/unitType-dictionary>).

Examples of constraints implemented in the

- an
- the value of an
- a
- a

### 42.3.2 Dictionaries

In a similar way, a dictionary ecology<sup>13</sup> has developed supporting an extensible set of concepts in CML documents. The dictionaries add semantics to the CML primitives, particularly *e.g.* melting point) and common metadata such as users and dates. Conventions will almost certainly have one or more dictionaries so that compchem has an extended dictionary of concepts such as convergent limits, energies, gradients and so forth. The MACiE<sup>14</sup> dictionary used the IUPAC Gold Book<sup>15</sup> to define terms in reactions and the Atmospheric Chemistry dictionary is again taken from IUPAC<sup>16</sup>.

One important way of creating dictionaries is to extract terms and discourse from CML documents. A particular example is the markup of concepts created in computational chemistry and here we often associate a given program or code with a dictionary specific to that program/code. Thus, for example, a program/code might use a set of keywords found nowhere else; currently around six such dictionaries exist, and the number is increasing. In these cases we often find the need for a hierarchy so that a code might use code-specific dictionary terms in addition to those in the general computational chemistry dictionary. Different programs sometimes produce data with the same label but a different interpretation; does “density” mean electron density or mass density? There can be any number of dictionaries (and we envisage one for each code, or ideally fewer). Each dictionary has a unique namespace so there are no collisions. The entries can be minimal (id, term, definition, *etc.*) but will usually indicate the data structure (

Applying

might specify that this *e.g.* <http://www.ietf.org/rfc/rfc2119.txt> be a dictionary

<sup>12</sup> CML Schema 2.4

<sup>13</sup> CML dictionary ecology

<sup>14</sup> MACiE: a database of enzyme reaction mechanisms

<sup>15</sup> IUPAC Compendium of Chemical Terminology - the Gold Book

<sup>16</sup> IUPAC Project: Glossary of atmospheric chemistry

Example (from <http://www.xml-cml.org/convention/dictionary>):

```
http://www.xml-cml.org/schema“  
http://www.xml-cml.org/convention/“  
http://www.xml-cml.org/unit/nonSI/“  
http://www.xml-cml.org/unit/unitType/“  
http://www.w3.org/1999/xhtml“  
http://www.w3.org/2001/XMLSchema“  
http://www.xml-cml.org/dictionary/dummy/“
```

### 42.3.3 Roles

A third approach to semantics is driven by the need to ‘tag’ information, and for this we provide the role attribute. Roles are less formalised than *ad hoc* semantics. They may, of course, link to formal semantic documents if required, though this cannot be enforced except by convention.

*i.e.* part of a folksonomy, while in other cases

Example showing how role is used in the definition of a fragment within a polymer<sup>17</sup>:

```
<?xml version = “1.0” encoding = “UTF-8”?>  
<fragment id = “cl_nsp2_methyl” convention = “cml:PML-complete” xmlns = “http://www.xml-cml.org/schema”  
xmlns:g = “http://www.xml-cml.org/mols/geom1”>  
<molecule role = “fragment” id = “benzene_1”>  
<atomArray>  
<atom elementType = “C” x3 = “9.526706134000763” y3 = “3.869733600000001” z3 = “5.213518402229052” id =  
“benzene_1_a1”>  
<label dictRef = “cml:torsionEnd”> r6 </label>  
</atom>  
<atom elementType = “C” x3 = “10.243299413197152” y3 = “3.932398500000001” z3 = “6.439022942911609” id =  
“benzene_1_a2”>  
<label dictRef = “cml:torsionEnd”> r1 </label>  
</atom>  
<atom elementType = “C” x3 = “8.713504556428543” y3 = “2.7185301000000006” z3 = “5.01720505576243” id =  
“benzene_1_a6”>  
<label dictRef = “cml:torsionEnd”> r5 </label>  
</atom>  
<atom elementType = “R” x3 = “8.385888936961882” y3 = “2.655387420737078” z3 = “4.323244676535362” id =  
“benzene_1_r6”>  
<atom elementType = “C” x3 = “10.119474056141831” y3 = “2.9008920000000007” z3 = “7.3834992125284815”  
id = “benzene_1_a3”>  
<label dictRef = “cml:torsionEnd”> r2 </label>  
<label dictRef = “cml:torsionEnd”> r2 </label>
```

<sup>17</sup> Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language

```

<label dictRef = "cml:torsionEnd"> r3 </label>
</atom>
<atom elementType = "C" x3 = "9.320371405363035" y3 = "1.8151698000000005" z3 = "7.151684115065878" id =
"benzene_1_a4">
<label dictRef = "cml:torsionEnd"> r3 </label>
</atom>
<atom elementType = "R" x3 = "9.280916015724046" y3 = "1.2657016684721403" z3 = "7.6896692864820775" id =
"benzene_1_r4">
<atom elementType = "C" x3 = "8.610030693701125" y3 = "1.7243409000000007" z3 = "5.934289686115539" id =
"benzene_1_a5">
<label dictRef = "cml:torsionEnd"> r4 </label>
</atom>
<atom elementType = "R" x3 = "10.697234803620145" y3 = "4.543958438540135" z3 = "6.552323882423661" id =
"benzene_1_r2"/>
<atom elementType = "Cl" formalCharge = "0" hydrogenCount = "0" id = "cl_2_a1" x3 = "9.692011995771473" y3 =
"5.151468187879777" z3 = "4.018767207085518"/>
<atom elementType = "N" formalCharge = "0" hydrogenCount = "0" id = "nsp2_3_n1" x3 = "10.889175042006798"
y3 = "2.9930090553818314" z3 = "8.690968080404991"/>
<atom elementType = "R" formalCharge = "0" hydrogenCount = "0" id = "nsp2_3_r3" x3 = "10.937618525919527"
y3 = "3.6140787328234207" z3 = "9.108611091395234"></atom>
<atom elementType = "R" formalCharge = "0" hydrogenCount = "0" id = "nsp2_3_r2" x3 = "11.21097670608745"
y3 = "2.3333885683637092" z3 = "8.845384731388283">
<label dictRef = "cml:torsionEnd"> r1 </label>
</atom>
<atom elementType = "C" id = "me_4_a1" x3 = "7.720415546204135" y3 = "0.5004327093517826" z3 =
"5.6475256189660525"/>
<atom elementType = "H" id = "me_4_a6" x3 = "7.8973212132921615" y3 = "0.15287314794224827" z3 =
"4.629221766363621">
<label dictRef = "cml:torsionEnd"> r1 </label>
</atom>
<atom elementType = "H" id = "me_4_a7" x3 = "7.962448186970064" y3 = "-0.2976542125451189" z3 =
"6.350030754819878"/>
<atom elementType = "H" id = "me_4_a8" x3 = "6.673285676722817" y3 = "0.7788619666709824" z3 =
"5.760051994135236"/>
</atomArray>
<bondArray>
<bond order = "2" id = "benzene_1_a1_benzene_1_a2" atomRefs2 = "benzene_1_a1 benzene_1_a2"/>
<bond order = "1" id = "benzene_1_a1_benzene_1_a6" atomRefs2 = "benzene_1_a1 benzene_1_a6"/>
<bond order = "1" id = "benzene_1_a3_benzene_1_a2" atomRefs2 = "benzene_1_a3 benzene_1_a2"/>
<bond order = "1" id = "benzene_1_a2_benzene_1_r2" atomRefs2 = "benzene_1_a2 benzene_1_r2"/>

```

```
<bond order = "1" id = "benzene_1_r6_benzene_1_a6" atomRefs2 = "benzene_1_r6 benzene_1_a6"/>
<bond order = "2" id = "benzene_1_a5_benzene_1_a6" atomRefs2 = "benzene_1_a5 benzene_1_a6"/>
<bond order = "2" id = "benzene_1_a3_benzene_1_a4" atomRefs2 = "benzene_1_a3 benzene_1_a4"/>
<bond order = "1" id = "benzene_1_a5_benzene_1_a4" atomRefs2 = "benzene_1_a5 benzene_1_a4"/>
<bond order = "1" id = "benzene_1_a4_benzene_1_r4" atomRefs2 = "benzene_1_a4 benzene_1_r4"/>
<bond atomRefs2 = "benzene_1_a1 cl_2_a1" order = "S" id = "benzene_1_a1_cl_2_a1"/>
<bond order = "S" atomRefs2 = "nsp2_3_n1 nsp2_3_r2" id = "nsp2_3_n1_nsp2_3_r2"/>
<bond order = "S" atomRefs2 = "nsp2_3_n1 nsp2_3_r3" id = "nsp2_3_n1_nsp2_3_r3"/>
<bond atomRefs2 = "benzene_1_a3 nsp2_3_n1" order = "S" id = "benzene_1_a3_nsp2_3_n1"/>
<bond order = "1" atomRefs2 = "me_4_a1 me_4_a6" id = "me_4_a1_me_4_a6"/>
<bond order = "1" atomRefs2 = "me_4_a1 me_4_a7" id = "me_4_a1_me_4_a7"/>
<bond order = "1" atomRefs2 = "me_4_a1 me_4_a8" id = "me_4_a1_me_4_a8"/>
<bond atomRefs2 = "benzene_1_a5 me_4_a1" order = "S" id = "benzene_1_a5_me_4_a1"/>
</bondArray>
</molecule>
</fragment>
```

#### 42.3.4 Units

The final component of the semantic framework is scientific units of measurement. In these we specify the type of the unit<sup>[18](#)</sup>. Every

These “essentials” are adapted from NIST Special Publication 811 (SP 811)<sup>[19](#)</sup> and NIST Special Publication 330 (SP 330)<sup>[20](#)</sup>. We use the terminology from NIST, with some variation, and quote verbatim to avoid confusion:

*quantity in the general sense “A*

CML uses the term “

*quantity in the particular sense “A*

CML does not currently use this concept explicitly. Quantities are usually either parameters or properties (but not all parameters and properties (*e.g.* string values) map to quantities).

*physical quantity “A*

CML honours this concept in that

*“A unit is a particular physical quantity, defined and adopted by convention, with which other particular quantities of the same kind are compared to express their value.”*

CML maps onto this concept through the

*value of a physical quantity “The*

CML supports this in the

---

<sup>18</sup> CML unitType dictionary

<sup>19</sup> Guide for the Use of the International System of Units (SI)

<sup>20</sup> The International System of Units (SI)

CML will honour specifications of <sup>21</sup> has been many years in incubation but now seems to be close to production release. CML will continue to use its own semantics for units but may also include interoperability with NIST.

The CML system of units goes somewhat beyond NIST in that it is not limited to physical science and has to support concepts such as mg (drug)/kg (animal) where the semantics of the experiment have to be linked (this is not a simple dimensionless number - “drug” and “animal” do not cancel). CML units allow for dimensions and other concepts to be associated with “dimensionless”, such as ppm). CML software (JUMBO <sup>22</sup>) allows for the values and units to be recomputed (“unit conversion”) and for simple dimensional analysis. Entries in

Users can create their own *e.g.* “The optimum dose of rIL2 was 100-500 units (Jurkat units)/ml, <sup>23</sup> which do not fit easily into the seven primary SI concepts, but are still critical attributes of the experiment. The general structure of the dictionaries is likely to be:

- A single, community-driven and maintained dictionary for *e.g.* fifth virial coefficient units), we see this being gradually and carefully extended.
- A number of local *e.g.* Jurkats).
- A single dictionary for SI <sup>24</sup> (paralleling the
- A small number of core dictionaries for <sup>25</sup> (*e.g.* CGS, atomic units, *etc.*)).
- A larger number of convention-specific

## 42.4 Creating dictionaries

The biosciences have several approaches for creating ontologies, such as the Gene Ontology (GO)<sup>26</sup>. GO was designed as a thesaurus to which individuals and groups could contribute. It has a directed acyclic graph (DAG) structure, where an entry can have several parents and several children. The hierarchy honours the broader/narrower term approach and used three axes (cellular component, molecular function, biological process) but is designed primarily for human navigability rather than machine computability. It and other dictionaries have been transformed to fuller OWL-compliant ontologies using the file format guide provided <sup>27</sup>.

We use the following approaches for creating dictionaries:

**|nonascii\_26| Borrow from established dictionaries** (IUPAC, IUCr, Wikipedia) and convert to CML. The main challenge is that many of the terms are broad concepts and follow human rather than machine conventions. This approach was used for the MaCiE dictionary with terms borrowed from IUPAC where possible and with a hierarchy expressed in CML. We have also translated the IUCr’s CIF dictionary into CML format <sup>28</sup>, and this is used in, for example, the CrystalEye system.

**|nonascii\_27| Observe and collect discourse/practice**, both in program input/output and formulaic text. We create or collect a corpus of documents and extract the common terms. Assuming that they are associated with

These processes lead to a community of dictionaries, with an implied but not necessarily explicit hierarchy.

---

<sup>21</sup> Units markup language (UnitsML)

<sup>22</sup> JUMBO

<sup>23</sup> Antitumor effects of interleukin 2 against renal cell carcinoma: basic study and clinical application

<sup>24</sup> CML SI units dictionary

<sup>25</sup> CML non-SI units dictionary

<sup>26</sup> The Gene Ontology, GO

<sup>27</sup> GO file format guide

<sup>28</sup> CML CIF dictionary

## 42.5 Detailed use cases of dictionary construction

•With the ChemicalTagger<sup>29</sup> system, we have built a natural language framework which recognises parts of speech and phrase. With over 100, 000 patents analysed we have a large corpus representing the current usage in describing chemical synthesis. The automatic analysis<sup>30</sup> of this corpus throws up a variety of abstractions common to many of the texts, in particular for the actions and methods used to describe chemical syntheses. Currently we have extracted 21 types of action phrase from this corpus:

Coupled with these phrases are qualifiers (sometimes English language adverbs) and specific uses of nouns which can be additionally used to label a text. This is an example of a small natural language driven dictionary into which a large number of specific terms can be entered.

•In the Quixote project<sup>31</sup><sup>32</sup> we are creating a semantic infrastructure for compchem. Unlike crystallography, where the community has for many years sat in real and virtual committee to decide on dictionaries and their contents, compchem has very little common practice in this area. There is no commonality of approach to labelling either the input or output of compchem calculations. Our belief is that there is a strong implicit similarity, even isomorphism, between the main computational codes, and that by analysing the discourse (*i.e.* the logfiles), we can collect and systematise the types of object referenced in the logfiles. To do this, we have taken a number of codes (Gaussian<sup>33</sup> (various versions), GAMESS-UK<sup>34</sup>, Jaguar<sup>35</sup>, NWChem<sup>36</sup>, Quantum ESPRESSO<sup>37</sup>) and analysed much of their logfile structure and vocabulary. Although the level of detail varies between programs, there are somewhere between 100-500 concepts in total which can be precisely labelled and which could contribute to a communal dictionary. We are in the process of building a table (spreadsheet) of the terms which occur in codes and their occurrence (or absence) in each code. These normally occur as CML parameters. The concepts currently cover the following areas:

Environment of the calculation. This includes machine configurations, version of code, time constraints, human and institutional metadata and other control parameters.

The method of calculation *e.g.* the functional.

The basis set or pseudo-potential.

Any physical constraints imposed on the system (*e.g.* pressure, temperature or electric field).

Levels of accuracy or cut-off desired in the calculation.

Strategy of calculation and algorithms used (*e.g.* search for a transition state, reaction coordinates, frequencies *etc.*).

The output files normally deal with outcomes of running the job (*e.g.* abnormal termination, level of convergence achieved, elapsed time) and calculated properties.

Most of these concepts are common to all codes and where possible we are creating entries in a single common compchem dictionary<sup>38</sup> (Figure 3). In some cases, however, methods and properties are unique to one code, and many of the intricate details in the logfiles are not directly transferable. For that reason, we are using a hierarchy of dictionaries with the following components:

1. A dictionary common to all or most of computational chemistry (compchem dictionary).
2. A series of dictionaries, one per code, which is initially used to collect defined quantities in the output. At regular stages the community will decide whether these map onto concepts in the main compchem dictionary, and, in those cases, transfer their usage to that dictionary.

---

<sup>29</sup> ChemicalTagger: A tool for Semantic Text-mining in Chemistry

<sup>30</sup> Mining chemical information from Open patents

<sup>31</sup> Quixote project on QC databases

<sup>32</sup> The Quixote project: Collaborative and Open Quantum Chemistry data management in the Internet age

<sup>33</sup> Official Gaussian website

<sup>34</sup> GAMESS-UK software

<sup>35</sup> Jaguar software, Schrödinger Inc

<sup>36</sup> NWChem software

<sup>37</sup> Quantum ESPRESSO software

<sup>38</sup> CML compchem dictionary

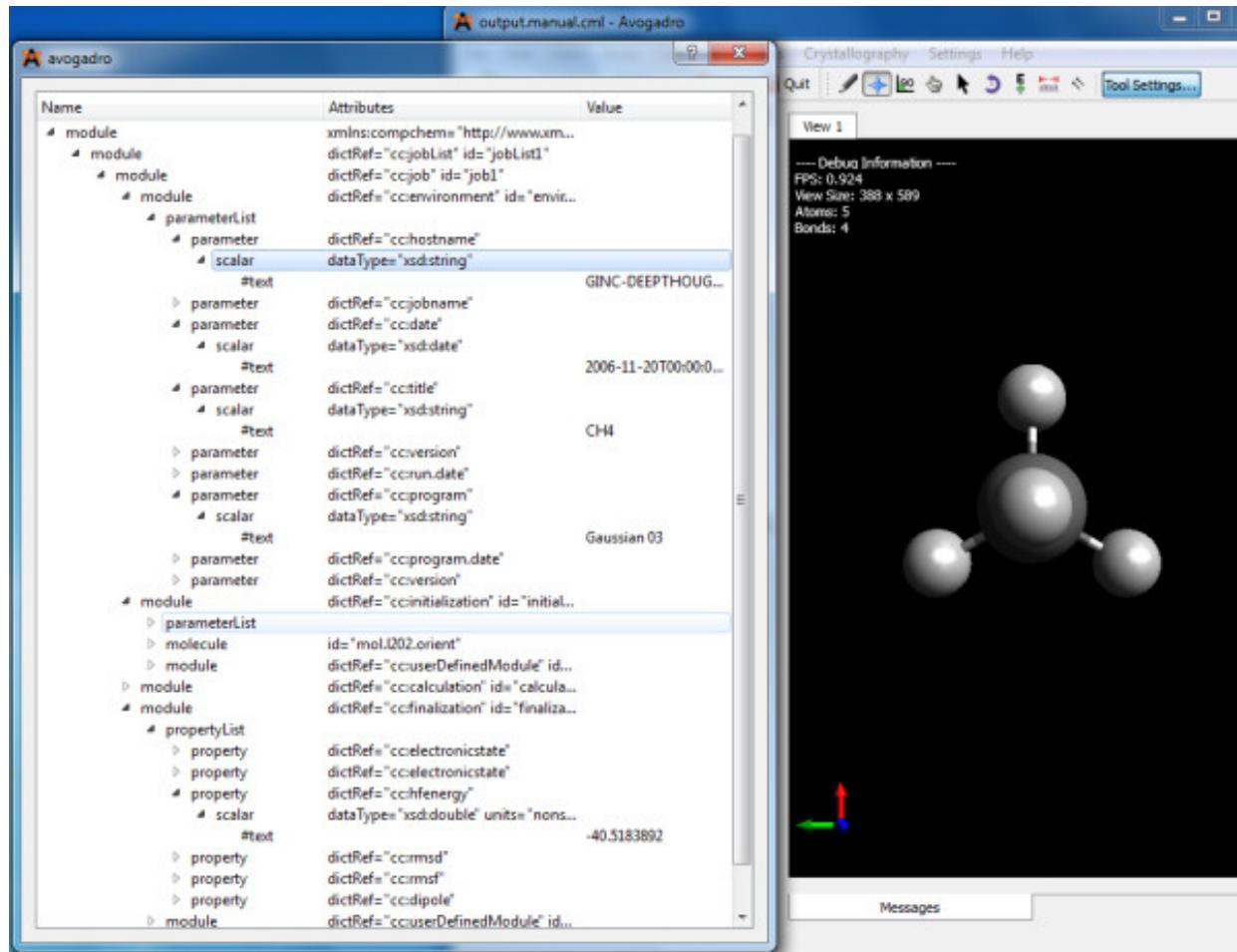


Figure 42.3: Figure 3. A compchem-compliant document read into the Avogadro browser and computational chemistry manager

A **compchem-compliant document read into the Avogadro** [#B39]\_\*\*browser and computational chemistry manager\*\*. The structure of the document is shown with the primary subdivisions. Each piece of information is in a precisely specified position in the hierarchy, so that it may easily be discovered by processing software. For example the

## 42.6 Software support for dictionaries and units

Besides the markup support for dictionaries and units, they are only really useful in chemistry if they are supported by a software system. Some of this can be provided by Web 2.0 tools such as RDF which can be used to lookup whether referenced units are present in appropriate dictionaries. However, it is often important to carry out manipulations on units such as conversion between different systems and multiplier prefixes. For that reason we have developed a suite of software within the JUMBO system for these manipulations.

The following elements are established in CML:

- Dictionary.
- Entry.
- Unit type (and unit type list.)
- Unit (in unit list).

In our recent work with dictionaries (especially in computational chemistry) we use the entries to provide some of the semantics to be applied at “run-time”. For example, a dictionary entry may define a syntactic template for the concept, or an enumeration of allowed values. In using the CIF dictionary, the data type (

## 42.7 Conclusion

The use of conventions and dictionaries has proved of enormous value in the development and robustification of CML. With well-defined protocols, groups can take the formal specifications and build their own systems such that they not only do what they want, but do not break other CML software. We are currently working actively on computational chemistry and, with a wide range of different codes and types of problem, we expect to be able to show that the current architecture is capable of supporting these.

Assuming that semantic computational chemistry becomes widespread, the dictionaries will act as a catalyst to those communities to add more terms and to revise the precise usage of the concepts. It will also act as a demonstration to other areas of chemistry of the value of the convention/dictionary approach.

## 42.8 Competing interests

The authors declare that they have no competing interests.

## 42.9 Authors' contributions

PMR designed dictionaries and their schemas, and wrote the manuscript. JAT implemented conventions and the validator, and wrote the manuscript. SEA built supporting software, and wrote the manuscript. WP implemented conventions and dictionaries. JT designed compchem dictionaries. All authors have read and approved the final manuscript.

# CML: EVOLUTION AND DESIGN

## 43.1 Abstract

A retrospective view of the design and evolution of Chemical Markup Language (CML) is presented by its original authors.

## 43.2 The genesis of CML

The modern online era has brought with it the need to rethink many of the mechanisms by which chemistry as a subject is researched, conducted and disseminated. This retrospective review describes how one infrastructure for doing so, CML or Chemical Markup Language, had its origins, and how the design evolved over a period of around 16 years (see also the accompanying spreadsheet for a breakdown of the historical timeline [Additional file 1]).

Additional file 1

**Timeline.** Spreadsheet of the Chemical Markup Language design evolution timeline.

[Click here for file](#)



Figure 43.1: Figure 1. The PMRz symbiote in a familiar environment  
**The PMRz symbiote in a familiar environment.**

PMR recounts the early background to CML from his point of view: “What are the origins of CML? I think I go back to *ca.* 1980 when I was writing code to extend Sam Motherwell’s great FORTRAN toolkit for the Cambridge database<sup>1</sup> - BIBSER (bibliographic search), CONNSER - the first and greatest chemical substructure algorithm, and GEOM78 - a geometry calculation tool. Between 1977 and 1980 I used to visit Cambridge (from Stirling) and work with Sam on extracting structures from the database and analysing them (Figure 2). There was a rough division of labour and ideas between us. I came with a number of ideas and Sam would modify CONNSER and GEOM to support these - literally within a day or so.

I took the problems back to Stirling and “integrated” Sam’s output with SPSS<sup>3</sup>. I did the analysis on the floor of our living room with an acoustic modem<sup>4</sup> where the handset was plugged into rubber cups. It used to run at 110 baud. The

<sup>1</sup> The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information

<sup>3</sup> SPSS, a computer program for statistical analysis

<sup>4</sup> Acoustic modem

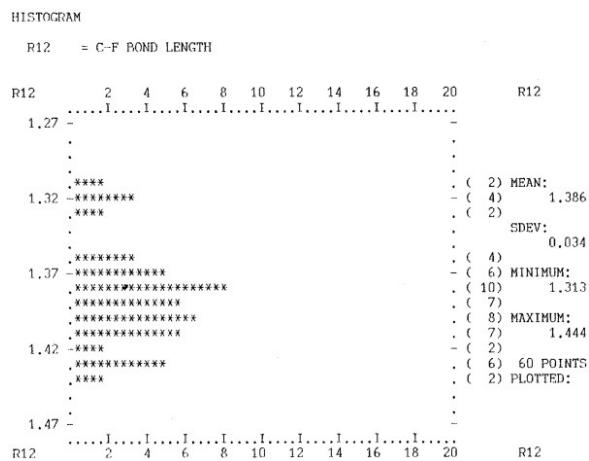


Figure 43.2: Figure 2. Analysis of bond lengths (horizontal axis = frequency) from the Cambridge Crystallographic Database *ca.* 1995

**caAnalysis of bond lengths (horizontal axis = frequency) from the Cambridge Crystallographic Database**. This has evolved into the bond length analysis tool in CrystalEye<sup>2</sup> which allows interactive clicking of points to bring up structures. (Note: Image is a scan of the original line printer output).

sums were originally done at Cambridge (on Phoenix) but I ported the software to UMRCC (the Regional Computing Centre at Manchester) on a CDC 7600. The results were printed out on folding line printer paper on a boustrophedonic ASR33 teletype. I would then extract data by hand and enter them into the statistical programs, but gradually moved to doing the statistics remotely. Remote graphics was always difficult - we could get printer plots posted from Aberdeen but it took a week. So I generally evolved ASCII (line printer output) plots. One consequence is that during these sessions I had a lot of time to think about how to do it better. It was obvious the software had to be modular and I gradually got to thinking about modular data.

In 1981/2 I spent a sabbatical with Jenny Glusker<sup>5</sup> in Philadelphia and there developed a VAX-VMS version of the software. I extended this to plot aggregations of data in two and three dimensions. Again the idea of modular components was clear. I returned to start up molecular modelling/computer graphics in Glaxo and found myself working with a completely different set of data files -ChemX, MDL molfile, etc. I couldn't use these with my analysis code. This burgeoning of portable chemical computational systems had begun in the early 1980s with a number of software products, mainly codes but also datafiles, being developed for the molecular graphics and computational chemistry community. In general, each resource developed its own representation of information, often referred to as a 'file format' or 'file type'. For example, by the mid 1980s there were probably fifty different file formats in chemistry including PDB, CSSR, MDL molfile and more specific program outputs<sup>6</sup> (Endnote 2). It was a major problem to convert between these formats, but despite some initiatives many software producers regarded the formats as proprietary and were resistant to ideas of inter-operability.

It seemed completely wasteful not to have a common format, so I started an activity within the Molecular Graphics Society<sup>7</sup> to systematize file types. In effect this was an attempt to build a chemical ontology. I didn't get much take up and there was active resistance from some software companies who regarded their formats as a commercial weapon.

During this period I had gradually advanced my language skills from FORTRAN to BBC-BASIC and C (part of this was through teaching the MSc in Birkbeck). So when C++ came along (late 1980s) I translated my approach to C++ and started to develop a toolkit/library. That's effectively when CML as a data modelling approach started."

HSR recalls his own early experiences in writing modular code: "In 1977, I had returned from the USA to Imperial College to continue my researches in quantum mechanical molecular modelling. Whilst at Austin in Michael Dewar's

<sup>5</sup> Prof. Jenny Glusker, Fox Chase Cancer Center

Prof. Jenny

<sup>7</sup> Molecular Graphics Society now the Molecular Graphics and Modelling Society

group, where I was learning about this area, I had encountered the famous ORTEP<sup>8</sup> program for displaying images derived from molecular coordinates, overkill of course for what I needed. So in 1977, armed with a Tektronix 4014 vector graphics terminal<sup>9</sup>, I started to write a simplified molecular renderer (STEK) optimized for computational chemistry using FORTRAN code. A number of modules were aggregated, including a much simplified ORTEP-style molecular renderer with an appropriately semi-interactive interface for rotating the molecule into an effective projection, simple XY plotting routines, 2D contour and isometric plotting routines for potential energy surfaces and molecular orbitals, and various labelling and annotating routines. Data was read in from separate files (mostly the output of the MOPAC molecular orbital program) and written out into a (human-readable) single history file which could be used to restore the composite diagram. To separate the various data types, I developed a simple, rather *ad hoc* markup language, with a linear parser which could read back the data objects and associated display attributes to reconstruct the content model for the diagram. The project ground to a halt after about 10 years, largely because I had come to rely on a system graphics library (SIMPLE) targeting solely the Tektronix devices and the CDC computers my institute then operated. Remove this library and the hardware, and my (FORTRAN) program became very difficult to port (especially to the raster devices which were starting to replace the vector displays in the mid 1980s). It did teach me the value of markup (I was already used to word processing using troff<sup>10</sup>), of separating data elements into modular components, and in particular of stateless “round tripping”, the ability to generate the output of a session in either a human- or a machine-readable manner that could be read back without loss, and in a reasonably error tolerant fashion, or with some components re-used for a different context. These absorbed concepts re-emerged some 10 years later when the ideas for CML started circulating.

I think my STEK program lasted perhaps 12 years, since I found that even in 1989, I was still using it<sup>11 12</sup> (good examples can be found in Figure one and Figures three & four respectively of these publications, which are not reproduced here for reasons of copyright). The molecule renderer had one feature unique to MO calculations not found in ORTEP, *i.e.* the ability to display the vibrational displacement vectors for a transition state, which was essential for understanding potential energy surfaces and the stationary points located.

The history of molecular orbital rendering is in itself an interesting one, since it introduces the connection between data, and its most effective representation. Hückel<sup>13 14</sup> was the first to apply MOs to “interesting molecules” such as benzene, but he famously never showed any diagrams. Dewar, starting in the early 1950s, did much to promote the use of molecular orbitals as a conceptual tool in chemistry, but he rarely provided quantitative representations in his articles. The great era of PMO theory in the 1950s was described using largely equations and tabulations of numbers rather than images<sup>15</sup>. Whilst I was a post doc with Dewar from 1974-1977, the group never in my recollection included MO wavefunctions derived using a graphical computer program in its publications! There was no idea to enhance the group’s papers with such from Dewar himself, or the spark from anyone in the group to go find such a program (from *e.g.* QCPE<sup>16</sup> (if indeed any existed)). I introduced routine ORTEP plotting of molecular coordinates to the group’s output, but never myself made the jump to rendering wavefunctions until inspired by an article on “orbital photography” in 1984<sup>17</sup>.

The relevance of MO rendering is that it highlighted in my mind the importance of data (calculated MO coefficients as numbers) and the need to interpret them semantically (which only visual rendering can do), in other words, the importance of both data and its appropriate rendering. Dewar’s group could do the former, but never developed the expertise for the latter. Interestingly, Hoffmann’s group did *both*, and acquired a major reputation in the field, including of course the Nobel prize for interpretations of pericyclic reactions<sup>18</sup>. Dewar always lamented his missing out on the Nobel prize, and arguably it was his inefficient application of data that might have been to blame.

One might argue that the reason the computational and the synthetic chemistry communities rarely mixed at that time

<sup>8</sup> ORTEP: Oak Ridge Thermal Ellipsoid Plot Program for Crystal Structure Illustration

<sup>9</sup> Tektronix 4000 text and graphics computer terminals

<sup>10</sup> AT&T’s document processing system, troff

<sup>11</sup> PM3 Potential Energy Surfaces for Phosphoryl Transfer Reactions

<sup>12</sup> A Comparison of semi-empirical and *ab initio* SCF-MO Potential Energy Surfaces for the Reaction of H<sub>2</sub>C = O with R<sub>3</sub>P = CH<sub>2</sub> and RP = CH<sub>2</sub>

<sup>13</sup> Quantum-theoretical contributions to the benzene problem. I. The electron configuration of benzene and related compounds

<sup>14</sup> Quantum theoretical contributions to the problem of aromatic and non-saturated compounds

<sup>15</sup> A Molecular Orbital Theory of Organic Chemistry. I. General Principles

<sup>16</sup> Quantum Chemistry Program Exchange

<sup>17</sup> Orbital photography

<sup>18</sup> Nobel Laureates in Chemistry

is that the learning curve to understanding the tables of numbers that were being produced in theoretical articles was too steep for synthetic chemists. Perhaps images were also needed for impact? In 1984, the Rubinstein brothers developed ChemDraw (a 2D representational program), followed in 1986 by Chem3D, to be used at that time only on an overtly graphical computer (the Macintosh). This was one of a bevy of commercial programs of that era that started to address both the organic chemistry and modelling communities.”

Back to PMR: “However the crystallographers had a much more unified view of the world. I continue to congratulate the International Union of Crystallography (IUCr) for its efforts in this area. In the mid 1980s the crystallographic community started to formalise its approach to small molecule X-ray diffraction. There was an active group, led by David Brown<sup>19</sup> aiming to create a self-defining format for crystallography- *Standard Crystallographic File Structure*<sup>20</sup><sup>21</sup>. It was essentially a data dictionary where a controlled vocabulary was created and specific semantics were added to items (e.g. data type). This approach was then taken further by the birth of the CIF initiative in 1990 and CIF is now the standard method of exchanging crystallographic information<sup>22</sup>. This was based on data supported by data dictionaries which themselves were constrained to a dictionary definition specification. I started to use the CIF approach to model my scientific world - this was long before XML but it was essentially isomorphic to XML - and it inspired much of the vision of CML.

I started with the most obvious components - geometry and numbers. These are still an integral part of CML (the “ca. 1993 I had a set of objects. But I needed a way to display and manipulate them.

At that stage I met Henry Rzepa. Henry remembers that probably around 1993, the student chemical society at Imperial College invited me to give a talk. I chose the topic of crystallography, but in characteristic fashion, delivered a scintillating talk (Henry’s description!) covering, well, probably almost all of chemistry! Henry chatted to me after my talk, and one of us must have mentioned the Internet. The topic might have been gophers (anyone remember them?) and what their potential was. Henry also visited me at Glaxo in Greenford around that time and we found we had a common interest in the Internet and its power for disseminating chemistry. It must have been about the time of the NCSA Mosaic browser<sup>23</sup> - 1993. This, in both our memories, is now immortalised by our working meeting in the Black Horse pub in Greenford in January 1994 (Figure 3). I had made initial explorations into a common format for chemistry, but it was the major adoption of HTML in 1993 and the announcement of the first World-Wide-Web conference (WWW1)<sup>24</sup> in 1994 that demonstrated that there was by then a critical mass of scientists and informaticians who wished to create semantic frameworks for information.”

In the Black Horse, PMR and HSR agreed they would both attend the WWW1 meeting; HSR ran the session on chemistry and PMR one on biology. PMR: “We had an early version of RasMol<sup>25</sup> which ran on UNIX and Henry had prepared a demo, parts of which were eventually published<sup>26</sup>. We had it running the day before on a Silicon Graphics machine, but when we came back the next day someone had wiped the shared libraries to save space. We got the thing running again 5 minutes after Henry’s talk<sup>27</sup> started.

The theme of WWW1 was, of course, the use of HTML (and HTTP) to create distributed information. Initially the focus was on HTML (with some discussion of the then very new MathML proposal by Raggett<sup>28</sup>), but during breakout sessions (or BoFs, birds-of-a-feather, as they were then called)<sup>29</sup> there emerged a realisation that each discipline would need to create its own approach to information that would be published and consumed in much the same way as HTML. Because it was in CERN and all HTTP sites at that stage were academic, the emphasis was all on science. Was MathML the way to carry maths in HTML? And if so, how could you do the same for chemistry? We didn’t know how.”

The WWW2 conference (Chicago, October 1994) had a focus on the development of the HTML markup language and

<sup>19</sup> Prof. I. David Brown, McMaster University

<sup>20</sup> The standard crystallographic file structure

<sup>21</sup> Standard Crystallographic File Structure-87

<sup>22</sup> CIF, Crystallographic Information Framework

<sup>23</sup> NCSA Mosaic™ web browser

<sup>24</sup> First International Conference on the World-Wide Web

<sup>25</sup> Rasmol, molecular graphics visualization tool

<sup>26</sup> Hyperactive Molecules and the World-Wide-Web Information System

<sup>27</sup> The Web and its Chemistry

<sup>28</sup> Dave Raggett

<sup>29</sup> BoF (Birds of a Feather)



Figure 43.3: Figure 3. The Black Horse pub at Greenford  
The Black Horse pub at Greenford.

was noteworthy for the attendance of non-scientists and many commercial organisations, as well as a representation from chemists<sup>30</sup>. It also served to convince HSR that SGML<sup>31</sup> was the vehicle that would be adopted for non-textual information. As a result, PMR started implementing prototypes of chemical information using SGML and Tcl. At that stage SGML was complex and fragmented, to the extent that relatively few (if any) complete implementations existed. There were very few Open Source implementations and we were grateful to be able to use the nsGMLs parser<sup>32</sup> from James Clark which would take an SGML document and transform it into a structured representation (ESIS). At the same time, a COST project, with help from Joe English, allowed the ESIS to be processed in essentially the way that is now possible with the DOM<sup>33</sup>, and PMR received much online help and guidance from Joe. By 1995, there was a prototype system where it was possible to read an early version of CML into a Tcl/Tk/COST processor and to display the structure of a molecule (Figure 4).

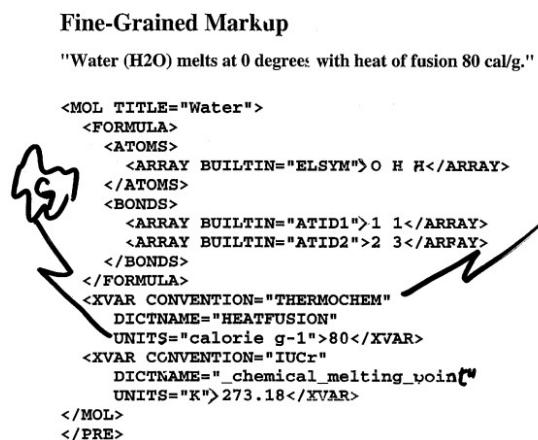


Figure 43.4: Figure 4. Early CML markup, ca. 1995  
Early CML markup. Many of the current concepts were prototyped at this level, such as the

<sup>30</sup> The Web in Chemistry: Hyperactive Molecules

<sup>31</sup> SGML, Standard Generalized Markup Language

<sup>32</sup> NSGMLS SGML system

<sup>33</sup> W3C Document Object Model (DOM)

HSR recalls: “Checking through some ancient (*sic*) files on our web server, I discovered that we first went public with CML on 21 August 1995, in the form of an ACS poster<sup>34</sup>. I notice that it introduced the concept of what I had referred to earlier to as *data round tripping*. The idea was to formalise and normalise both the input and output of a computer program so that the latter could also reliably serve as an input. This was in fact implemented for the MOPAC program, and was a much more formal and structured expression of what I had earlier tried to do with the STEK graphical program. I was in Chicago, and Peter was back in the UK, at a terminal, waiting for comments from the audience on the poster to come flooding in! In fact, when we got to the hotel room that the ACS session occurred in, we discovered no trace of any Internet connection (anywhere) and could not communicate (Internet connections at conference venues only started becoming common from ~2006 onwards). Peter sat in an unrequited silence throughout the entire presentation! The poster was in fact presented as part of a session grandly entitled “Chemistry on the Infobahn”.

In 1995, HSR, PMR and Andrew Payne set up the Open Molecular Foundation (OMF) as a group to support and disseminate the creation of semantic chemistry using CML (SGML) as the infrastructure. This still required the combined operation of SGML processor executable and a wrapper which was being converted to Java 1.02. In late 1996, we became aware of the W3C XML project<sup>35</sup> (SGML on the Web) and joined the early discussion and working groups, one of which was located in central London (the Rembrandt Hotel meeting), which both PMR and HSR attended. There, Jon Bosak<sup>36</sup> declared that Java and XML were the foundations of the web! It became clear at this stage that CML should be based on XML, not SGML, and PMR was now working to an XML/Java representation in 1997 (Figure 5). A forum for XML developers was set up, which went live in February 1997 with a welcome post by PMR, and where many of the important developments involving XML were first discussed during the period of the operation of the list at its initial home (February 1997–December 1998)<sup>37</sup>.

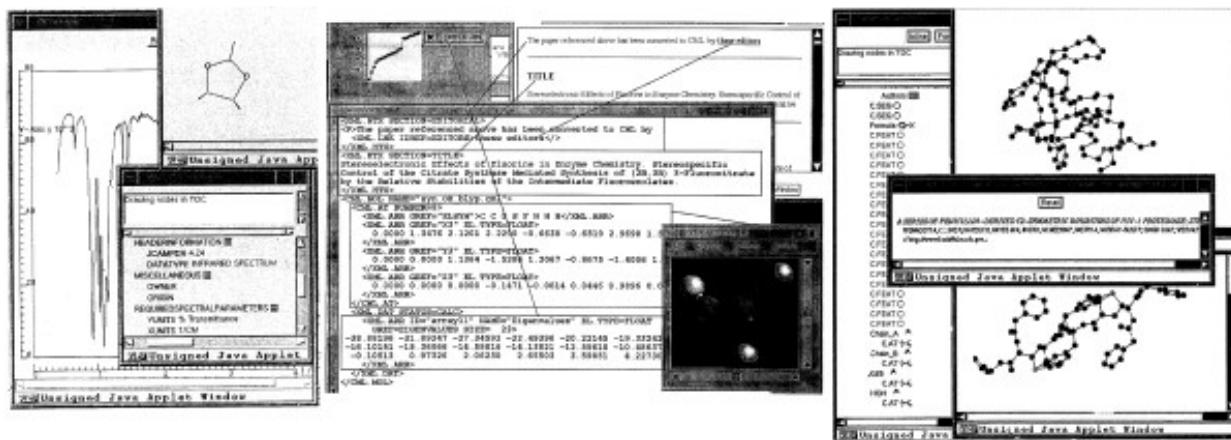


Figure 43.5: Figure 5. JUMBO screenshots (ca. 1997) showing support for spectra, properties, molecular structure in 2D and 3D and a variety of applets and widgets

**caJUMBO screenshots** (. Note the considerable change in syntax from the earlier picture; we have prototyped a namespace approach (e.g.<sup>38</sup>). The molecules and spectra had clickable locations so that peaks and molecules could be linked. (Note: Image is a scan of original overhead transparencies).

### 43.3 The philosophy of CML

The primary purpose of CML has been and is to allow humans and machines to communicate chemical concepts without loss of semantic information. For example, a major role is to allow the output of one program to be converted

<sup>34</sup> CML - Chemical Markup Language

<sup>35</sup> XML Core Working Group Public Page

<sup>36</sup> Jon Bosak

<sup>37</sup> XML-DEVarly archives

into CML and input to another without loss. CML is also designed to create datuments, a combination of semantic text and non-textual information<sup>39</sup>. Our vision is that scientific publications should be represented semantically such that both humans and machines can consume them, again without loss. When CML is universally adopted for both these processes, then the large parts of the current discourse and information interchange in chemistry will be semantic. There is no reason why CML, in combination with other languages such as HTML, MathML, SVG, GML etc. (Figure 6) cannot then be used for at least the following:

The screenshot shows a presentation slide with the title "Text, maths and molecules combined". Below the title is a small logo featuring a stylized molecule and the letters "CML". The main content of the slide is a text block: "The hydrolysis of molecule (I) [click to display] can obey the kinetics:". Below this text is a mathematical equation: "(1)  $dx/dt = -kx$ ". Further down the slide, there is a note: "can be represented by the HTML/XML file text.xml which uses XML-LINK to 2 others". Following this note is some XML code: "<P> The hydrolysis of <A HREF="mol.xml" XML-LINK="SIMPLE"> molecule(I)</A> can obey the kinetics <A HREF="eqn.xml" XML-LINK="SIMPLE" ACTUATE="AUTO" SHOW="EMBED">". Below this is the text "The MathML file: eqn.xml:" followed by MathML code: "<!DOCTYPE MATHML><MATHML> <EQN><EXPR><DIFF/>x<BVAR>t</BVAR></EXPR> <EQ/><EXPR><MINUS/>k<TIMES/>x</EXPR></EQN></MATHML>". Then, the text "The CML file: mol.xml:" followed by CML code: "<!DOCTYPE CML><CML><MOL TITLE="Methyl chloride"><ATOMS><ARRAY BUILTIN="ELSYM">C</ARRAY><ATOMS><BONDS><ARRAY BUILTIN="ATID1">1</ARRAY><ARRAY BUILTIN="ATID2">2</ARRAY><ARRAY BUILTIN="ORDER">1</ARRAY></MOL></CML>". At the bottom of the slide, there is a footer with the text "1 of 2" and "08/31/97 07:39:26".

**Figure 43.6: Figure 6.** A multi-namespace design from 1997 - the first use of the CML alembic logo  
**A multi-namespace design from 1997 - the first use of the CML alembic logo.** This was before the XML Working Group created the current namespace syntax and while CML was still based on DTDs. (Note: Image is a scan of an original overhead transparency).

- Ingestion of data into data- and knowledge-bases
- Extraction of data from knowledge-bases
- Journal articles
- Theses
- Suppliers catalogues
- Textbooks
- Regulatory documents such as patents and new drug applications
- Input to programs
- Output from programs.

<sup>39</sup> The Next Big Thing: From Hypermedia to Datuments

The only technology capable of managing this at present (and probably for some time to come) is XML. It is widely used in publishing and CML can therefore be technically adopted for any of the document-like examples above. XML is also a primary method of marshalling input to databases and there are many tools which allow the construction of db schemas from XML schemas. In addition, however, CML has also shown itself to be valuable in the following areas:

**|nonascii\_11| A semantic infrastructure for physical science.** This is because none of the other scientific disciplines have developed markup support for dictionaries and units and so the CML constructs can support other areas.

**|nonascii\_12| A data structure for computation.** Although not originally intended for this purpose, XML is a very powerful data structure for internal data in computer programs (with some possible sacrifice in performance and memory size). Many of our programs use CML as their complete data representation and operate by adding to, removing from or modifying information on the DOM or infoset. For many applications this is a cleaner approach to passing data as everything is contained within one structure and there is no need for alternative storage of the same information. This, for example, is how all information in Chem4Word<sup>40</sup> and JUMBO<sup>41</sup> is held.

**|nonascii\_13| A computable object in its own right.** Because it is extensible and because computational semantics can be added to some of the elements, it represents a simple functional programming language. This is most developed in PolymerML<sup>42</sup> where a polymer can contain instructions for its own elaboration and the computation is carried out by repeatedly applying polymer extension semantics to the PML representation of the structure.

In designing CML, we have attempted to abstract the current common implicit and explicit concepts in mainstream chemistry. This is done by intensive and repeated analysis of chemical corpus linguistics. A common procedure is to take a recent journal article and to see to what extent CML can support the chemical concepts in that article. Similarly, we take the input and output of chemical programs and abstract new concepts and dictionary entries from those. In this way, we believe that CML is accessible to the chemical world and other scientists who use chemistry without a change in their concept structure.

## 43.4 The evolution of CML

During the evolution of CML, we have been guided by the following factors:

**|nonascii\_14| The evolution of W3C recommendations and web technology and practice.** For example, when W3C introduced the XSD schema<sup>43</sup> recommendation we translated the CML DTD into XSD. When W3C introduced the XSLT<sup>44</sup> specification, we created a library of routines to process most of the CML elements. In similar ways, CML has reacted to incorporate SVG, RDF and OWL. These technologies themselves have had variable amounts of uptake. For example, until recently the main application of SVG was in mobile devices only, but now, 10 years after its launch, it is becoming mainstream in most browsers. We created an early Javascript tool (JUMBO-JS) which ran in the two current browsers in 2000 but which because of the rapid and uncontrolled changes in browser functionality no longer works and has been abandoned. However, it appears that JS is now reborn in a stable implementation and it is possible that we shall shortly create a CML reference. Similarly it has taken at least a decade for the concepts of RDF to mature and for a satisfactory toolset to start to appear. In these cases, CML had had to wait until there are clearly established and widely-used technologies that it can rely upon.

**|nonascii\_15| The interest of chemistry and the wider scientific community in markup languages.** Chemistry is recognised to be one of the more conservative scientific disciplines<sup>45</sup> and a new technology generally requires wide acceptance in other communities first or a powerful and determined commercial implementation. Although our work is well-known in the chemical informatics community, it is often said that “there is no demand for CML from our customers” with the result that a vicious circle ensues. Indeed some of the impetus comes from other subjects such as bioscience.

---

<sup>40</sup> Chem4Word, Chemistry plug-in for Microsoft Word

<sup>41</sup> CML, Chemical Markup Language

<sup>42</sup> Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language

<sup>43</sup> XML Schema

<sup>44</sup> XSL Transformations (XSLT)

<sup>45</sup> Open access in chemistry: information wants to be free?

**|nonascii\_16| The Open Source (OS) community.** The OS community in chemistry is a relatively small part of the volume of software creation but it is highly visible and has a wide variety of offerings. Almost all OS chemical systems can read and/or write CML and there is a general agreement that these systems should converge to inter-operability. With the increasing rise of OS in general, and specifically in chemistry, this will be an important incentive to the adoption of CML.

**|nonascii\_17| Specific market applications.** The publishing industry is universally based on SGML and/or XML and so it is technically straightforward to incorporate CML in publications. The movement away from non-semantic output (such as PDF) is still slow but we believe that this is inevitable and again this will create considerable incentive to use CML. The small proportion of Open Access (OA) in chemistry means that it is very uncommon for scientists to extract information from the literature using machines and indeed many publishers expressly forbid this. As OA increases, we expect that the value of semantic information extracted from the literature will be seen to provide a large amount of additional value.

**|nonascii\_18| Toolset.** The chemistry community is likely to require a range of well-proven tools before it will adopt a new information technology. This takes time and/or financial investment before there is a perceived demand, but we believe that we are close to a situation where the value of this is starting to become apparent.

**|nonascii\_19| Regulatory and archiving.** There are several practices where XML is the recommended approach. In archiving material, XML can represent the semantics of a document and is frequently used by electronic archivists. In regulatory there is a requirement for many regulators to know the precise details of information in the doc and to be able to extract it rapidly. Therefore again XML is frequently required by regulators. Both of these pressures should lead towards a greater acceptance of CML.

## 43.5 JUMBO

It has always been important that CML can be implemented in a reproducible and validatable manner. We are extremely reluctant to allow new elements or attributes in CML unless it can be demonstrated that they can be implemented and deployed in a large number of use cases without problems. It is surprising how often apparently small changes can severely disrupt the greater system. For example, allowing the explicitly declaration *raison d'etre* is to show that information can be reliably processed, including round-tripping, and to provide reference examples of how various constructs should be used. JUMBO deliberately does not add large numbers of chemical methods (*e.g.* substructure search) but consumes these from other OS implementations.

JUMBO has been through six iterations, each more or less re-written from scratch. JUMBO1 used the rather primitive features in AWT1.0 to provide a hierarchical semantic browser of chemical documents (the name stands for Java Universal Molecular Brower for Objects). This was technically successful but not widely deployed because of the relatively small number of CML documents available at the time and the newness of Java. JUMBO2 was a development-only version and mapped the CML elements onto editable widgets such as textboxes, lists and molecules. It used an early version of Swing and because of the difficulties in that was never formally distributed.

JUMBO3 returned to the browser concept and displayed a CML document in a series of windows (rather similar to the current Bioclipse<sup>46</sup> tool). There was a brief flirtation with Java 3D for molecular display but we reverted to including Jmol<sup>47</sup> as a callable window where required. At this stage it became clear that the continued development of the schema made it very difficult to keep the specification and the software in sync. In JUMBO4 we attempted to use the W3C DOM implementation as our data structure. This turned out to be very problematic as the library was not designed for subclassing and all elements had to delegate to the W3C DOM object. There were several other problems with the W3C DOM including the lack of any XPath<sup>48</sup> functionality and towards the end we moved to the much more satisfactory XOM from Elliotte Rusty Harold<sup>49</sup>. This has proved to be an extremely useful and reliable choice as it not only provided a relatively simple view on the DOM but it manages concepts such as namespaces extremely well.

<sup>46</sup> Bioclipse

<sup>47</sup> Jmol: an open-source Java viewer for chemical structures in 3D

<sup>48</sup> XML Path Language (XPath)

<sup>49</sup> XOM, XML Object Model

As more elements were added to the schema, maintenance became a real problem, with 100 elements and 100 attributes admissible in various combinations. Content models which allowed and constrained combinations of child elements became very complex and unmanageable, and it was almost impossible to write consistent code. Therefore, in JUMBO5 we resorted to auto-generation of the basic code from the schema. This meant that when new elements were added the code was regenerated from the incremented schema.

JUMBO6 was primarily a refactoring of the functionality of JUMBO5 resulting in a clean design for the implementation of converter functionality. We have continued to modularise so that now there is a basic CML XOM (where most methods simply represent accessors and mutators). JUMBO6 is a set of tools providing additional chemical functionality, especially those for manipulating the DOM, and a separate large library of JUMBO-Converters which extract the output from programs and documents and convert it into hierarchical CML docs. At this stage, we also developed Chem4Word, which uses a fully validated convention of CML (“CMLLite”). This took a much more lightweight approach to content models and created the concept of validation. In the Chem4Word system all potential input is validated against a rich combination of XSLT expressions which are far more powerful than XML schema (XSD). This is now a continuing philosophy.

## 43.6 Code-driven CML Design

As part of the CML philosophy we have strongly adopted the ‘rough consensus and running code’ stock (originally coined by David Clark in 1992)<sup>50</sup>. An excellent example of the value of developing code in parallel with a specification was given in the early days of XML. The initial design included a nearly-full implementation of SGML parameter entities. At this time two or three prototype XML parsers were being developed (Norbert Mikula, Tim Bray, PMR) and it became clear that the implementation of the full parameter entity model was a major effort for relatively little reward, and it was therefore dropped from the specification. We have found the same in CML, sometimes only surfacing several years after a feature was introduced. The abstraction of data into

Similarly the design of the

It is sometimes impossible to tell what the effect of a schema design will be before deployment. A particularly difficult problem has been white space. In many cases (such as in formatted files like PDB), whitespace is extremely significant (‘

Another major feature in the design was the need to validate combinations of elements and attributes. At one stage, PMR was approached by a group of pharma and related companies, to create a specification of CML for the Object Management Group (OMG)<sup>51</sup>. Hand-coding the combinations of attributes and elements proved impossible (this is effectively a sparse 100\*x\*100 matrix) and it was clear that automated methods of validation and code-generation could be necessary. JUMBO4.6 therefore generated code from the schema model rather than requiring it to be hand-coded. However, the W3C DOM technology was not well-suited to this, leading to the adoption of the simpler XOM model. None of this had an immediate effect on the surface schema of CML but has had deep influences on the subsequent design.

As a result of this, we developed the attributes and elements largely independently; in other words, attributes are generally not context-specific. The commonest attributes are e.g. the 7).

However, with increasing deployment to different areas of chemistry, it became clear that it was going to be impossible to find universal content models for most elements. This is due to not only the diversity of chemistry but also the different ways that chemists might wish to organise information. A molecule might have one or more spectra as children (representing that these are associated analytical data). Alternatively a spectrum could have one or more molecules as children representing that these correspond to different peaks. It was this type of experience that led us to propose an extremely flexible content model, constrained by the use of

The context-free attribute design has been largely successful. In a few cases, common words such as **spectrum type** = “NMR””**reaction type**

---

<sup>50</sup> The Tao of IETF: A Novice’s Guide to the Internet Engineering Task Force

<sup>51</sup> Object Management Group, OMG

The screenshot shows a code editor with Java-like syntax. A tooltip is displayed over the code at line 81, showing a list of methods starting with 'a'. The tooltip includes the following text: 'Did you know that Quick Documentation View (Ctrl+Q) works in completion lookups as well?'.

```

79
80     CMLProperty property = new CMLProperty();
81     property.a
82         m addArray(AbstractArray array)
83         m addAttribute(Attribute att)
84     return m
85     } m addCMLXAttribute(Element element, String attName, String attVa...
86     } m addName(AbstractName name)
87     public C m addNamespaceDeclaration(String prefix, String uri)
88     getI m addScalar(AbstractScalar scalar)
89     getS m addToLog(Severity severity, String message)
90     retu m appendChild(Node child)
91     } m appendChild(String text)
92
93     public L Did you know that Quick Documentation View (Ctrl+Q) works in completion lookups as well?

```

Figure 43.7: Figure 7. The auto-complete functionality in IDEs is underpinned by the content model approach

**The auto-complete functionality in IDEs is underpinned by the content model approach.**

The use of enumerations (*e.g. e.g.* periodic system of the elements); in others (*e.g.*

There is a special case with XSD data types. If we adopted all (*ca. 48*) of these, then there would be a possible commitment that implementers had to implement all of them. If the XML is being used in an XSD system such as entry to a database this is manageable, but for the more flexible requirement of heterogeneous CML it becomes a major burden. Therefore we have arbitrarily selected a small number of data types (

In general, giving multiple options, even apparently simple choices, for values is an extreme burden on implementers. A bond order was originally allowed to have

The requirement that all physical quantities have

It is also difficult to know at the start how many container elements should be used. For common elements such as property, parameter and molecule there are specific container elements (*e.g.*

The increasing availability of XPath-based technology has had a major positive effect on the possible flexibility of the organisation of elements and attributes. For example, in documents as large as several megabytes, it is possible to use XPath expressions to locate, delete, change, add and move (sort) components. A typical computational chemistry logfile or a crystallographic experiment (CIF) in CML can be manipulated with great power and flexibility. This means that content models are almost irrelevant whilst XPath-assisted conventions are a major tool in normalising and re-purposing chemical information.

## 43.7 Validation

The original purpose of SGML was to act as a machine-enforceable contract between an author and a typesetter. SGML tools could indicate that a document was valid or invalid and each party would know whether it was their responsibility or that of the other. The DTD therefore also acted as a specification against which compliant software could be written (Figure 8). This idea is very much at the heart of CML and represents the first major infrastructure for validating chemistry (crystallography excepted).

Many of the problems of software and data in chemistry can be traced to the lack of a validation system. Without validation, the author of a program cannot easily write conformant software if the input is variable; similarly the author cannot know whether an input is fit for purpose. Unfortunately the past 30 years have seen a wide variety of formats each with a wide variation in conformance. For example, there is no accepted ‘standard’ for PDB files and many program authors have modified this format for purposes other than managing protein crystal structures (*e.g.* computational chemistry output). As a result, many programs corrupt information because they cannot validate the input, and they make unwarranted assumptions. By ensuring that input and output are both valid or validatable, it becomes possible to link processes and ensure no semantic corruption or loss.

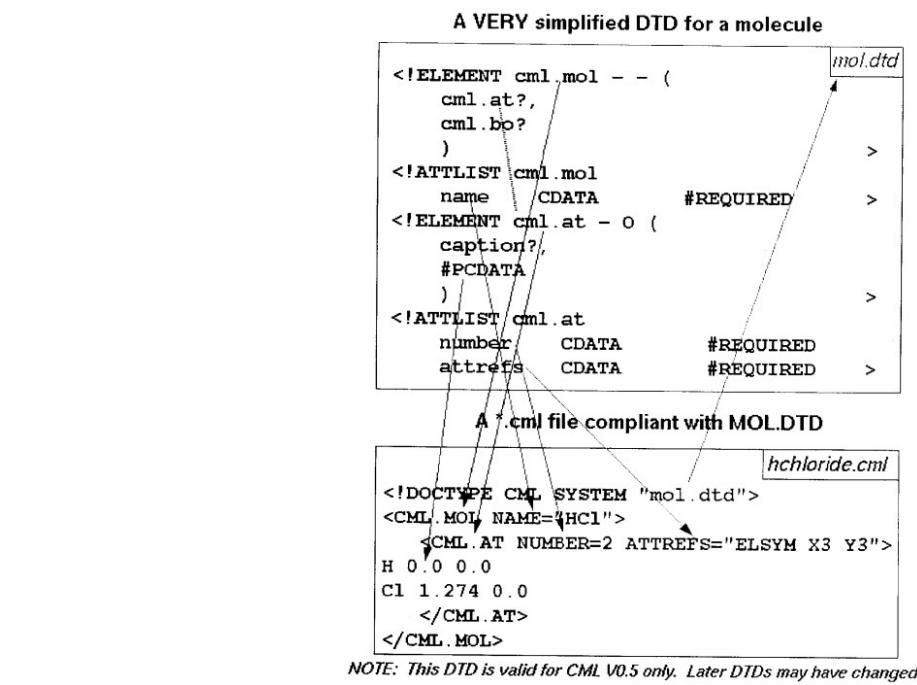


Figure 43.8: Figure 8. A version of the CML DTD (in SGML) from *ca.* 1996

**caA version of the CML DTD (in SGML) from . Note the early development of a namespace philosophy although there was no technology to support it at the time. (Note: Image is a scan of an original overhead transparency with handwritten annotation).**

There are many implicit assumptions about the representation of chemistry that cause semantic problems. For example, very few datafiles state the units of measure of scientific quantities and there are frequent assumptions about the existence of hydrogen atoms. There is much confusion between 2D and 3D coordinates and few systems can hold both at the same time. A major purpose of CML is to make sure that all chemical information is validatable and that the rules for this validation are openly visible. Most recently we have constructed Chem4Word and the CMLLite specification<sup>52</sup> which shows that complete validation of input and output chemistry is possible even in complex systems. In that case, the validation is carried out by stylesheets/XPath which has most of the power that is required.

## 43.8 Community-driven CML Design

CML has always been a community project in that its progress has been visible and it has been possible for anyone to provide feedback. It is not however a community-managed project and is best described by the BDFL model<sup>53</sup> ('Benevolent Dictator For Life'; *cf.* Linux). We feel this is necessary to ensure a consistent vision for the infrastructure and we have also felt that until this achieved stability it was unreasonable to expect others to volunteer contributions when they might be discarded at any stage. There have now been several sub-projects in CML where the community has been actively involved in design, including "CMLReact", "CMLSpect" and some aspects of "CMLComp" ("compchem"). There have been very few major additions to CML in the last three years despite its increasing deployment and we therefore feel that the central architecture is fit for purpose and can be extended by the community in a variety of ways that suit their own needs (mainly through dictionaries and conventions).

<sup>52</sup> CMLLite Schema

<sup>53</sup> Benevolent Dictator For Life, BDFL

## 43.9 Foreseeable evolution of CML

Our explorations in a wide range of chemical documents and documents containing chemistry have shown that CML is capable of managing disciplines as far-ranging as atmospherics, minerals, enzymes, analytical and computational chemistry. The immediate vision is that the world would benefit enormously from having this material available in CML. The barriers are almost all cultural; few chemists see the merit of this and therefore few if any publishers of chemistry make any provision for doing this. We believe that the increasing value of semantic material (*e.g.* Linked Open Data, LOD) will gradually show the community that this is enormously important.

There are four main methods of creating semantic chemistry: a) human authoring (as in conventional articles, reports, laboratory notebooks etc.), b) conversion of chemical data from legacy formats, c) creation of semantic chemistry through computer program output and d) machine extraction of chemistry from unstructured and semi-structured material (*e.g.* electronic chemistry publications). We see the following opportunities and barriers to each of these, listed below. There is a “chicken and egg” aspect to this. We are frequently told that there is “no demand for CML” and as a result people do not create tools that read or produce it. In several cases we have attempted to overcome this by creating believable prototypes in these areas.

**a) Human authoring** is likely to happen when there is a sufficient range of semantic editors. We have created Chem4Word to show that this is possible but it needs an acceptance by the community that semantic authoring is something that is desirable in an editor. Although the default authoring tools are currently Word and LaTeX, we expect that web-based tools such as GoogleDocs and EtherPad will lead to much more attractive environments in which scientists will create documents. Two good examples of this are Southampton’s Blog3 software<sup>54</sup>, where a blogging platform is used to create chemical documents, and Peter Sefton’s Scholarly HTML initiative<sup>55</sup>, showing that modern scholarship, including science, should be managed through HTML and not doc/pdf.

**b) File format conversion.** The problem of converting between different file formats has been largely solved by the Blue Obelisk community<sup>56</sup>. Our own JUMBO-Converters will convert many of the common formats (mol, smi, pdb, cif, cdx etc.), especially the more complex ones, into structured CML without semantic loss. There is no technical barrier to rapid and widespread uptake by the chemical community.

**c) Conversion of legacy to include semantic content.** Many programs such as quantum chemistry and molecular dynamics produce logfiles originally aimed at printing on fan-folded line printer paper. We have shown that it is possible to intercept all output statements and convert them to CML (*e.g.* for SIESTA, CASTEP, MOPAC, DL\_POLY). These outputs have been used in the computational minerals and materials communities and we have created libraries (*e.g.* FoX<sup>57</sup>) to make this process easy. However, at present most codes have not adopted this and we have therefore written a series of converters based on a declarative parsing technology (JUMBO-Parser) which allows for very high (greater than 95%) precision and recall of the structure and semantics of the documents. These are being developed by the Quixote community<sup>58</sup> for computational chemistry programs and are generally adaptable to any program which produces combined text and numeric output.

**d) Machine-extraction of chemistry.** Our OSCAR program<sup>59</sup> has high success (80-90% precision and recall) in extracting chemical entities from unstructured text. The success rate depends on the specific domain but ranges from atmospheric chemistry through biomedical to synthetic chemistry. In all text-mining areas, it is much more difficult to extract processes, relationships and sentiment from documents. However, chemical syntheses are reported in such a formulaic manner that we can extract a very high degree of the underlying semantics of the chemical reactions. The primary barrier to this are the legal prohibitions demanded by mainstream chemical publishers which are generally agreed by subscribing institutions, meaning that it is a contractual violation to undertake text-mining activities. BMC journals are an exception (being published as CC-BY) but there is relatively little mainstream chemistry published in BMC at the moment. We have been able to show that we can extract chemistry from patents, although the quality of many of these is not perfect and there are errors due to transcription. Recently the British Library has argued strongly

<sup>54</sup> blog<sup>3</sup>. A blogging engine for the Semantic Web

<sup>55</sup> Scholarly HTML

<sup>56</sup> The Blue Obelisk

<sup>57</sup> FoX, FORTRAN XML Library

<sup>58</sup> The Quixote project on QC databases

<sup>59</sup> OSCAR4, Open Source Chemistry Analysis Routines

for the reform of intellectual property laws to allow text-mining for scientific and related purposes<sup>60</sup>, and if this were to happen, there are enormous opportunities for CML technology to provide near-universal semantic chemistry.

Assuming that the cultural, political and legal barriers are removed, it is very cost-effective to produce all chemical information using CML. We have shown that, at near-zero cost, the whole of published crystallographic data can be converted into semantic form (250, 000 structures in CrystalEye). Similarly, we have read 100, 000 patents and extracted reactions; others have used OSCAR to add semantics to information from Medline. If the chemical community can agree identifier systems for the components (molecules, reactions, spectra etc.), this would create a huge resource of chemistry to be added to the LOD cloud. We would expect this to be indexed on compounds, substances, reactions, spectra, crystal structures and many aspects of physical chemistry. By creating dictionaries, ideally with the involvement of authorities such as IUPAC and IUCr, we then have a comprehensive semantic framework with a simple computable ontology. It is difficult to predict exactly what the benefits of this will be, but they will be massive. LOD provides for linking between disciplines, *e.g.* the concentration of chemicals in the atmosphere at different times and geographical locations. It allows for systematics within a sub-discipline (*e.g.* comparing all published synthetic procedures and analysing these for the potential value of reaction conditions, catalysts etc.) It allows experimental data (*e.g.* crystal structure) to be used to calibrate computational approaches such as quantum mechanics. With the addition of natural language, this becomes a human-accessible resource where we can ask simple powerful questions to the machine such as “find me all spectra which contain NMR shifts below zero and which do not contain metals” or “find me all solvents involved in reactions above their boiling point”. In computational chemistry, CML can largely automate the process of creating multiple jobs through parameters sweeps and analysing and searching the outputs. Indeed, it is reasonable to see program manuals being replaced by CML dictionaries appropriate to that program, and understandable both by machines and humans.

## 43.10 Sustainability

Every semantic project must address its sustainability. In the past CML has been highly dependent on its two authors. The technical resource to support it in its current form is now relatively modest, so that in principle it could be forked or continued (the “Dr. Who model of OS”<sup>61</sup>) were the current authors to become inactive. It has been implicitly and explicitly endorsed by a range of companies and organisations including the IUCr, Unilever Research, Microsoft Research and the National Cancer Institute of the US National Institutes for Health. There is a wide range of CML-compliant software and a large number of examples. Its technical sustainability therefore seems assured, and its political sustainability is beyond the scope of this article. We are confident that in the not too distant future publishers such as BMC will enthusiastically accept contributions consisting partly or mainly of CML.

## 43.11 Competing interests

The authors declare that they have no competing interests.

## 43.12 Authors' contributions

PMR and HSR both developed CML and wrote the manuscript.

---

<sup>60</sup> British Library press and policy information

<sup>61</sup> ‘The Doctor Who Model of Open Source’ blogpost by Peter Murray-Rust

## 43.13 Endnotes

### 43.13.1 Endnote 1

CML is a joint creation of the two authors over many years. Some of this paper is written in the first person, other sections refer to ‘us’ or ‘we’. Anything that appears to refer to one or other in the singular should be mentally replaced by ‘the PMRz symbiote’.

### 43.13.2 Endnote 2

Open Babel represents an almost comprehensive identification of these formats, which in 2011 has reached 113 formats in chemistry.

## 43.14 Acknowledgements

It is not appropriate to detail every person who has made a contribution (there are 43 listed developers alone in the Sourceforge repository and a similar number were acknowledged at the PMR symposium “Visions of a Semantic Molecular Future” in January 2011). Many people have used marginal resources to contribute to CML and the software. We have been particularly grateful to organizations which have encouraged us to continue developing CML in a climate where most of the conventional chemical data and software creators have been uninterested.



# AMI - THE CHEMIST'S AMANUENSIS

## 44.1 Abstract

The Ami project was a six month Rapid Innovation project sponsored by JISC to explore the Virtual Research Environment space. The project brainstormed with chemists and decided to investigate ways to facilitate monitoring and collection of experimental data.

A frequently encountered use-case was identified of how the chemist reaches the end of an experiment, but finds an unexpected result. The ability to replay events can significantly help make sense of how things progressed. The project therefore concentrated on collecting a variety of dimensions of ancillary data - data that would not normally be collected due to practicality constraints. There were three main areas of investigation: 1) Development of a monitoring tool using infrared and ultrasonic sensors; 2) Time-lapse motion video capture (for example, videoing 5 seconds in every 60); and 3) Activity-driven video monitoring of the fume cupboard environs.

The Ami client application was developed to control these separate logging functions. The application builds up a timeline of the events in the experiment and around the fume cupboard. The videos and data logs can then be reviewed after the experiment in order to help the chemist determine the exact timings and conditions used.

The project experimented with ways in which a Microsoft Kinect could be used in a laboratory setting. Investigations suggest that it would not be an ideal device for controlling a mouse, but it shows promise for usages such as manipulating virtual molecules.

## 44.2 Background

Amanuensis: One employed to take dictation, or copy manuscripts; A clerk, secretary or stenographer, or scribe  
<http://en.wiktionary.org/wiki/amanuensis>

The chemistry laboratory is a difficult environment for using a computer. Space is at a premium; benches and fume cupboards are covered with apparatus and typically have chemicals that are detrimental to computers (Figure 1). The chemist wears protective clothing, and often has gloves on (Figure 2). Lots of little issues add up to make it a challenge to successfully use computers in the lab.

Yet the collection of data has never been more important. Trends in science are to require data in support of experimental results. It is considered that research paid for by public money should have the proceeds visible to anyone who may wish to use them. Recent examples of challenging the conclusions of the scientific community - such as MMR episode <sup>1</sup> or the Climate-Gate emails <sup>2</sup> - plus various examples of scientific fraud, are all events that could have been ameliorated if their data were open for review.

---

<sup>1</sup> MMR vaccine controversy

<sup>2</sup> Climatic Research Unit email controversy



Figure 44.1: Figure 1. Chemist working at a typical fume cupboard

**Chemist working at a typical fume cupboard.** It is common to use the glass front for drawing reactions, jotting notes, etc.



Figure 44.2: Figure 2. A chemist happy in his element, and probably with his elements too  
**A chemist happy in his element, and probably with his elements too.** Note the protective clothing. Space in the lab is at a premium; catering for computer is an after-thought.

In recent years, various projects have highlighted how hardware and software can be used for the collection, management and use of laboratory information. At the University of Cambridge Cavendish Laboratory Baumberg *et al.*<sup>3</sup> illustrate how new hardware forms (desktop “Surfaces” and tablets) facilitate the use of visual sketching techniques to enhance the scientific process, in particular within a group. The Frey group at Southampton have shown<sup>45</sup> how semantic tools can be used to link complex information from the whole experiment lifecycle. The OPSIN chemical name-to-structure program<sup>67</sup> developed by Lowe *et al.* here in the Unilever Centre has been extended to show how complex information can be accessed using smartphones<sup>8</sup>.

The Ami project was created to find improved ways for chemists to use computers in the lab. The goal was to build a prototype next-generation information assistant “natural user interface” for scientists working at the lab bench. The limitations of paper lab notebooks are well recorded, and the Chemistry Department at Cambridge has recently deployed a commercial electronic lab notebook (ELN), a project in which one of the Ami team members was a significant participant. The Ami project aimed to combine and develop existing hardware and software technologies in novel ways to provide an information rich environment for the scientist at the bench.

#### 44.2.1 Methodology

The Ami project used a brainstorming session with chemists from the department plus representatives from the Chemistry Department’s computing service to identify the issues that chemists have to deal with and how computers could be used to address them (see Appendix 1 for further details). A common use-case that emerged was the example of how chemists often reach the end of their experiment and find an unexpected result. Often the suspicion is that the unexpected result could be due to mundane reasons; the conditions used for the reaction may have varied unexpectedly, or reaction components were not added, or timing was critical, or the wrong chemicals were used, etc. What was required was some way of going back over events to see what actually did happen. What was needed was some way of collecting ancillary data - data that is not the primary data that is scientifically obvious to collect - that could be consulted after the experiment is finished if circumstances needed to be investigated further.

The desire to log ancillary data identified three areas to work on. The first was to build some hardware device that could monitor parameters such as the temperature of the reaction vessel, keeping a log over the whole duration of the experiment. The second was a video monitor to provide a close-up visual record of the reaction. The third area was a wide-angle video monitor of the whole fumehood which would log all activity in the vicinity of the reaction.

Windows 7 was chosen as the development platform for Ami. This was because of the wide availability of software tools and utilities available for Windows, and also because of the experience within the group. Where possible code was developed in Java, using the IntelliJ IDEA development tool.

#### 44.2.2 Ami Client Application

The Ami client application - “Ami” - is the central application with which the chemist interacts when in the lab. The chemist logs into Ami using their identity badge which is detected by a Touch-A-Tag RFID reader. A list of experiments is then displayed, plus the option to create a new experiment (Figure 3).

Having selected their desired experiment, Ami then displays the main experiment control screen. Tabs at the top are used to switch between the Event Log, Sensor Control, and Experiment Details screens. All tabs and buttons can be controlled using speech, the keyboard, or the mouse.

The Event Log shows all the events that have occurred in the experiment. The log can be filtered by event type if desired (Figure 4).

---

<sup>3</sup> MetaSurfacing with the Surface

<sup>4</sup> The semantic smart laboratory: a system for supporting the chemical eScientist

<sup>5</sup> The Semantic Grid and chemistry: Experiences with Comb\*e\*Chem

<sup>6</sup> OPSIN, Open Parser for Systematic IUPAC Nomenclature

<sup>7</sup> Chemical Name to Structure: OPSIN, an Open Source Solution

<sup>8</sup> Lowe, D. OPSIN-Android

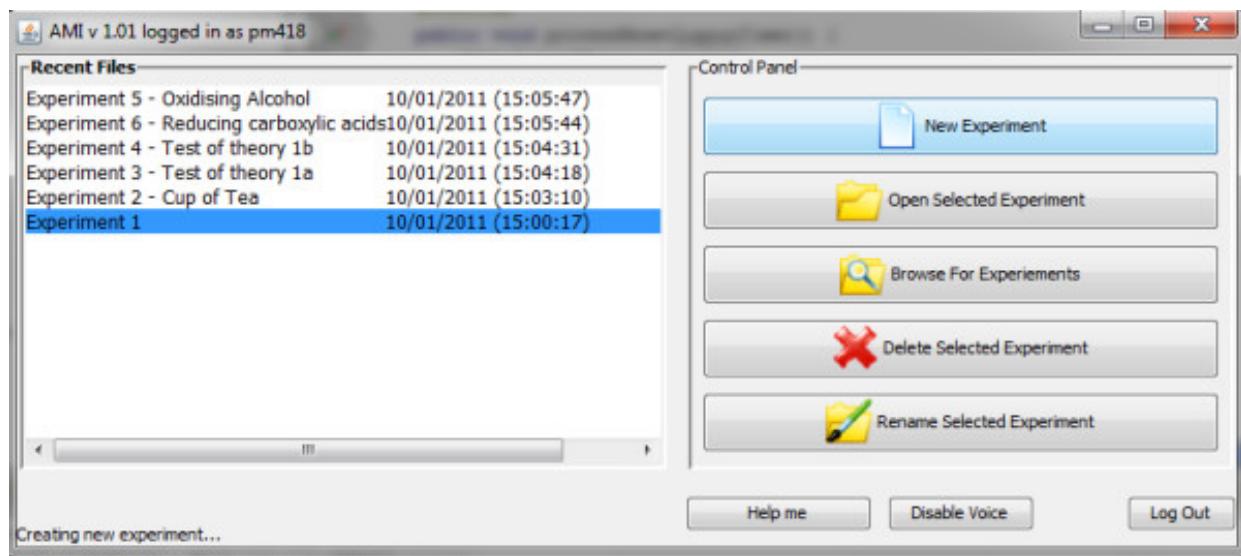


Figure 44.3: Figure 3. The Ami experiment selection screen  
The Ami experiment selection screen.

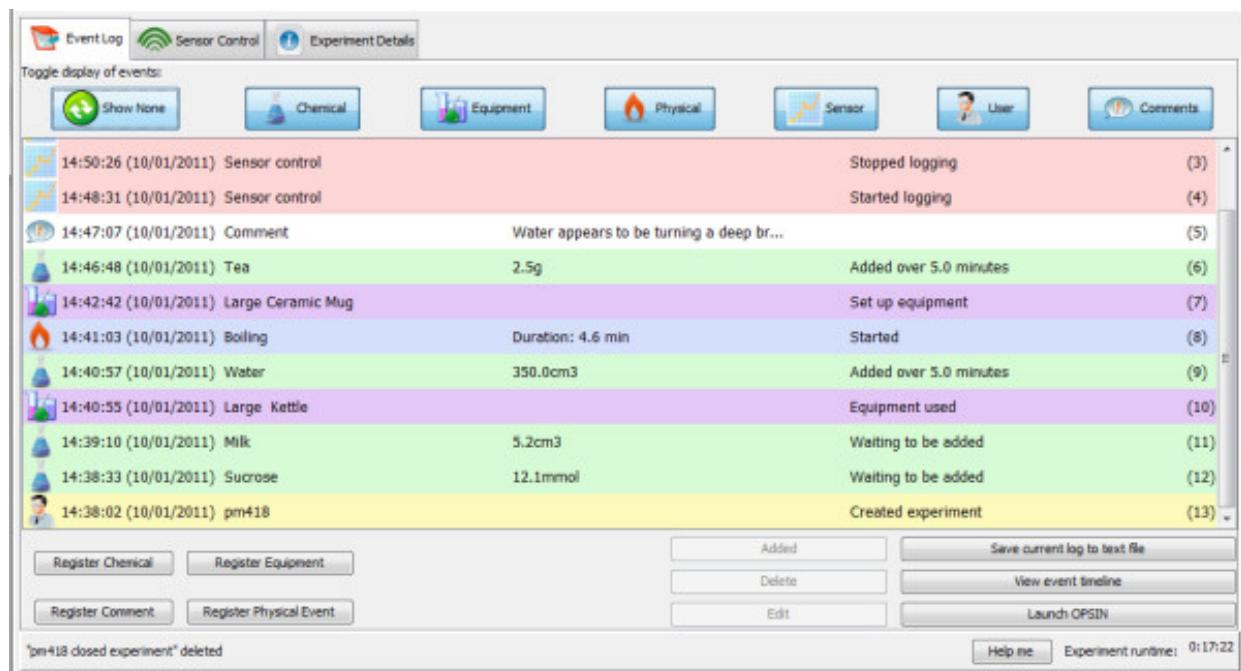


Figure 44.4: Figure 4. The Ami Event Log screen  
**The Ami Event Log screen.** The main body shows events logged. Buttons across the top allow selection of different event types. At bottom left are buttons to log different events. Buttons at bottom right are for saving data and launching the timeline viewer. Tabs at the top switch the display to the Sensor Control and Experiment Details screens.

Ami allows all chemicals and pieces of apparatus used in the experiment to be tagged with an RFID tag. These are easily and cheaply available in a variety of forms so that they are easy to stick to chemical bottles and apparatus. During the experiment, the chemist registers all the components with Ami.

As an experiment proceeds, the chemist logs usage of chemicals and apparatus by simply waving them in front of the RFID reader. The date and time of the event is recorded by Ami, so that a timeline of events is built up showing activity in the experiment. The chemist can also add observations by dictating to the PC's microphone, or simply by using the keyboard.

A refinement of the RFID tagging was an “intelligent labcoat”. This was achieved by using a mini Arduino card with an RFID reader built into the sleeve. The Arduino had a Bluetooth transmitter with it, which was able to transmit readings to the Ami application running on a PC. The RFID detector in the labcoat sleeve automatically registers any tagged chemical or piece of equipment that the chemist’s hand goes near. The events logged by the labcoat are then transmitted to the Ami application for including as part of the experiment timeline (Figure 5).

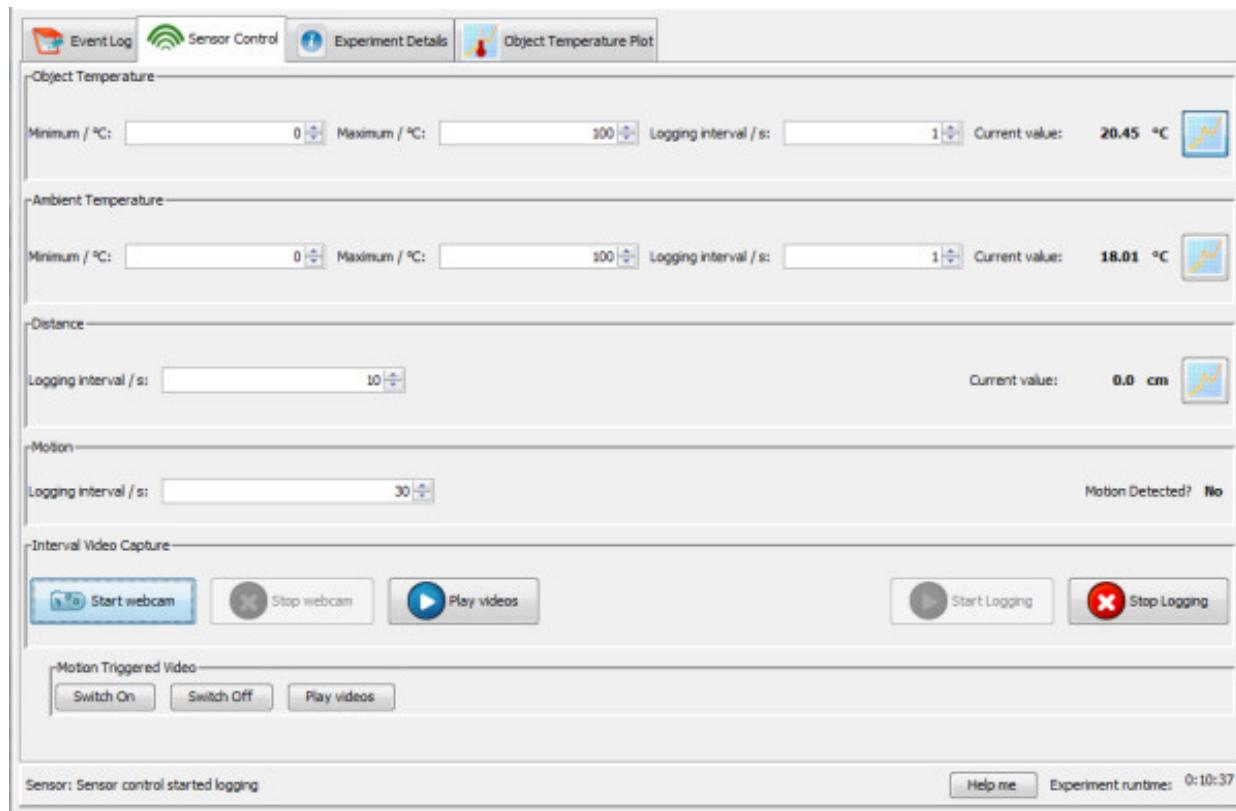
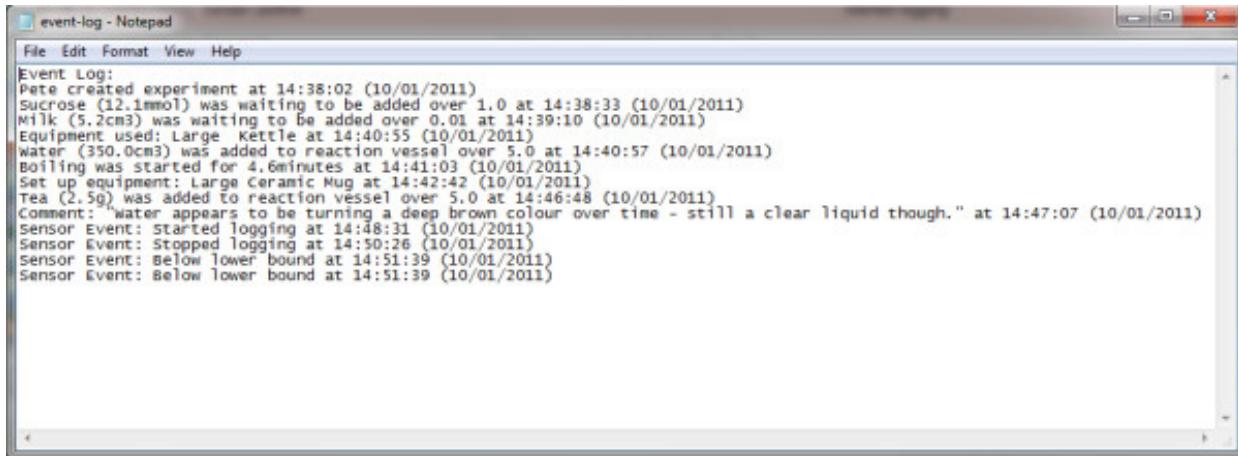


Figure 44.5: Figure 5. Sensor control tab

**Sensor control tab.** Alarm limits can be set for upper and lower temperatures. Time interval can be specified for taking temperature, distance (ultrasonic sensor), and motion detection (infrared motion detector). Buttons at bottom left are used for starting motion-timelapse video and environs-monitoring video. Buttons at top right control display of sensor graphs.

Each sensor monitored by Ami has its own logfile (Figure 6). All output files are stored in one directory for each experiment, making it easy to keep track of all data created, and to transfer it to the electronic lab notebook. The output from all the sensors and events can be displayed as a graphical timeline to facilitate review of the experiment activities (Figure 7).



```

event-log - Notepad
File Edit Format View Help
Event Log:
Pete created experiment at 14:38:02 (10/01/2011)
Sucrose (12.1mmol) was waiting to be added over 1.0 at 14:38:33 (10/01/2011)
Milk (5.2cm3) was waiting to be added over 0.01 at 14:39:10 (10/01/2011)
Equipment used: Large Kettle at 14:40:55 (10/01/2011)
Water (350.0cm3) was added to reaction vessel over 5.0 at 14:40:57 (10/01/2011)
Boiling was started for 4.6minutes at 14:41:03 (10/01/2011)
Set up equipment: Large Ceramic Mug at 14:42:42 (10/01/2011)
Tea (2.5g) was added to reaction vessel over 5.0 at 14:46:48 (10/01/2011)
Comment: "water appears to be turning a deep brown colour over time - still a clear liquid though." at 14:47:07 (10/01/2011)
Sensor Event: started Logging at 14:48:31 (10/01/2011)
Sensor Event: Stopped Logging at 14:50:26 (10/01/2011)
Sensor Event: Below lower bound at 14:51:39 (10/01/2011)
Sensor Event: Below lower bound at 14:51:39 (10/01/2011)

```

Figure 44.6: Figure 6. Example event log file for the creation of a cup of tea, inspired by the Southampton Smart Tea project

**Example event log file for the creation of a cup of tea, inspired by the Southampton Smart Tea project**<sup>9</sup>. New events are appended as they occur.

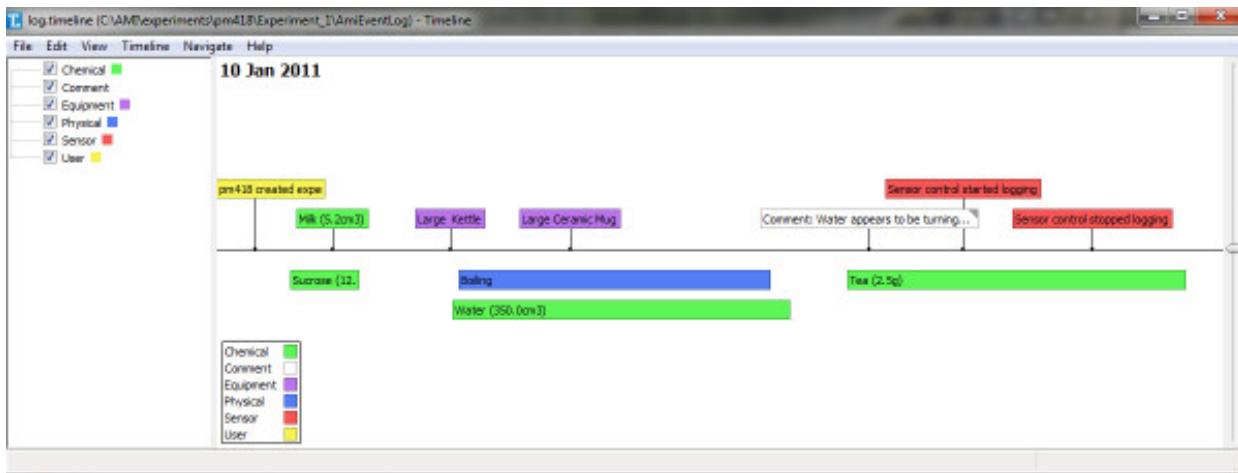


Figure 44.7: Figure 7. Graphical view of experiment timeline

**Graphical view of experiment timeline.** The Timeline application<sup>10</sup> is written in Python and receives an XML event-log file created by Ami. Timeline takes this event log and displays different sorts of events in different colours.

#### 44.2.3 Monitoring Device - Arduino

For close-up monitoring of the reaction, an infrared temperature sensor was used (Figure 8). This was controlled by an Arduino circuit board (Figure 9), which was programmed using the open source Arduino software<sup>11</sup>.



Figure 44.8: Figure 8. The Ami Experiment Monitoring Tool, here monitoring tea temperature  
**The Ami Experiment Monitoring Tool, here monitoring tea temperature...**

The Ami Experiment Monitoring Tool also has an ultrasonic distance sensor and an infrared PIR motion sensor. Output from the sensors is sent to the Ami client application, which is a Java program running on the PC.

#### 44.2.4 Close-up video monitor

The usual way of doing time-lapse photography is to take a still picture at regular intervals, then stitch them together to make a moving picture. We wanted to do something slightly different; instead of a still picture, we wanted to use a few seconds of normal moving video, and then stitch them all together to make a time-lapse video. The advantage of this is that it is then possible to see how a given material is behaving (*e.g.* viscosity) which isn't possible from a still picture.

Recording time-lapse video turned out to be more difficult than expected. We were unable to find an off-the-shelf application that we could use to provide this functionality. Open source Java routines to monitor video had performance

---

<sup>11</sup> Arduino

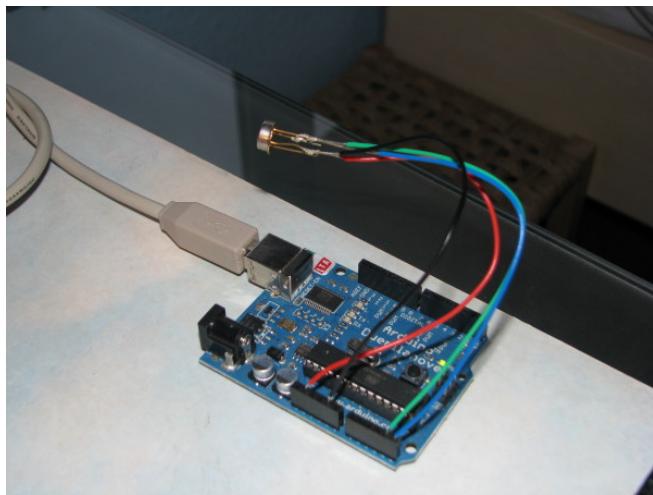


Figure 44.9: Figure 9. The infrared sensor being tested on an Arduino circuit board  
**The infrared sensor being tested on an Arduino circuit board.**

issues, for example very poor frame-rates. The main problem seemed to be that available Java-based open source code was out of date; it was all based on the Java Media Framework (JMF)<sup>12</sup>, the API of which has not changed since 1999, and the last minor modification was in 2004. The open-source FFmpeg<sup>13</sup> utility was also tried, but its video capture functionality is provided through “Video For Windows” (V4W) which is not supported on Windows 7. Eventually we settled on VLC<sup>14</sup>, which is based on Microsoft’s DirectShow framework (better supported and up to date).

Because VLC is an application, the problem arose as to how to start and stop it from the Ami application. Fortunately VLC can be controlled via a telnet connection, so Ami uses telnet to configure the video capture and to start and stop video capture. There is an additional bonus that this separation of video recorder from controller enables multiple cameras to be used and also to start and stop recording on remote systems, without a physical connection to them.

Linking the videos together also turned out to be more difficult than expected. Modern compressed video formats such as AVI, MPEG4 work by encoding differences between successive frames. When concatenating video it is therefore necessary to decompress the videos before combining them together and re-encoding them using a given compression algorithm, such as an MPEG4 based codec.

Fortunately, VLC again has the ability to do this task but due to the nature of video concatenation, this process of stitching together the files is best done at the end of an experiment, rather than repeating the CPU-intensive process for each stage. Compression artefacts are extremely liable to arise from the process of repeatedly decompressing and recompressing as well, degrading the quality of the video. It may be possible to ‘pause’ a recording using the VLC capture, but if any errors arose or power is lost, the video data would have to be recovered manually.

Storing video files alongside the experimental data enables the logs to travel with the data, given a repository such as the ELN that can accept arbitrary files as part of a submission.

#### 44.2.5 Wide-angle video monitor

A common source of error in doing experiments is simply absent-mindedness, forgetting to do something, or using the wrong chemical. The wide-angle video monitor is triggered by activity at the fume cupboard, and records video until activity stops. This gives the ability to replay events over the course of the experiment, hopefully enabling a full picture of what actually was done to be understood. Mounting the webcam high up at the back or side of the fume cupboard gives the best view of activities.

<sup>12</sup> The Java Media Framework

<sup>13</sup> Ffmpeg

<sup>14</sup> VideoLAN, VLC Media Player

Two methods of triggering the recording were identified. The first was to use an infrared movement detector, which was connected to the Arduino. When movement was detected, the event is passed by the Arduino back to the Ami program, which then starts the video monitor. The video simply records for a specified duration after movement is no longer detected. The second method was by monitoring the changes in the image itself, and if a threshold value is reached, to start videoing. Unfortunately time did not permit us to explore this area sufficiently to get a working system going.

#### 44.2.6 Experiments with the Microsoft Kinect

Towards the end of the Ami project, Microsoft released the Kinect (Figure 10). The Kinect is an accessory to Microsoft's Xbox consumer gaming machine, and is an exciting new development in human-computer interactions because it uses purely visual techniques to build a three-dimensional understanding of the space in front of it. The Kinect can recognise when a person is standing in front of it, and automatically determine the positions of the person's head, body, arms, hands, legs and feet. The user does not need such things as a transmitting device or special reflective tags; they just have to stand in front of the Kinect. The Kinect is potentially a disruptive technology and is already showing huge potential in robotics<sup>15</sup>. It has enormous potential for Ami; positioning a Kinect in a fume cupboard could give the user new ways to interact with the computer, and help in monitoring the environment.



Figure 44.10: Figure 10. The Microsoft Kinect

**The Microsoft Kinect.** This connects to a computer via a USB connector, making it a powerful way to communicate with computers. The infrared transmitter is on the left.

The Kinect consists of a relatively small box (about 25 × 12 × 3 cm) which has two video cameras built into it<sup>1617</sup>. One video camera is used for normal videoing using visible-light. The other is an infrared camera which monitors a pattern of infrared dots that the device shines into the room<sup>1819</sup>. The on-board processing built into the Kinect enables it to understand the 3D location of all the objects in front of it (*i.e.* the spatial analysis is done by the Kinect itself, rather than by the computer that it is attached to). The attached computer receives from the Kinect a video feed plus a stream of data points of the 3D locations of all the objects detected by the Kinect. The Kinect also contains four microphones, but using these was not investigated in this project because at the time that this work was done no code had been released which made the sound output available.

<sup>15</sup> The Kinect Sensor in Mobile Robots - Initial Experiments

<sup>16</sup> Kreylos, Oliver. Kinect Hacking

<sup>17</sup> Microsoft Kinect Teardown

<sup>18</sup> Andrewe1. Kinect with Nightshot

<sup>19</sup> RobbeOfficial. Kinect - sensor IR projection

One slight limitation of the Kinect is that its 3D view of the area in front of it is necessarily only seen from one position<sup>20</sup>. This means that it cannot understand a full three dimensional view of an object, because it can only see the side nearest the detector. Anything behind an object, and the back of an object, cannot be seen. This could be improved by using more than one Kinect operating together so that they can pool their individual views, and no doubt the techniques and code necessary to achieve this fuller 3D view will emerge over time<sup>21</sup>.

### Monitoring 3D space

The Kinect returns a three dimensional description of what it can see in front of it as well as a conventional ‘RGB’ view. The resolution of the normal colour image camera is  $640 \times 480$ , whereas the three dimensional camera is  $320 \times 240$ <sup>22</sup>. Whilst this makes it a poor choice for image recognition, logging and so on, the depth camera delivers data that is fundamentally unavailable from other sources. This makes it incredibly exciting in terms of the types of data and interaction it can enable.

The working range of the Kinect is suitable for a large living room, as it was designed with that in mind. It was found that in the cramped confines of a fume cupboard the detector was not far enough away for reliable operation. This rather precludes the Kinect from being used for monitoring the 3D environment within the fume cupboard (the size of a typical fume cupboard is about 1.7 m wide  $\times$  1.2 m high  $\times$  0.7 m deep). So we turned our investigations to using the Kinect for controlling the computer itself; because the Kinect monitors body movements, it might be good for someone who is wearing protective clothing.

### Using the Kinect to control a mouse

One of the intuitive ideas for using a Kinect is to detect human movement and gestures to control a computer, so called ‘natural interaction’. One of the first experiments we tried involved detecting a hand and coupling a mouse pointer to respond to its movements. This was done at first by crudely detecting the closest ‘blobs’ present in the depth, i.e. the hands, and using the motions of these to control the mouse pointer of a computer. At a later stage, we used a much more sophisticated technique to capture the hand motions, which involved ‘skeleton mapping’ that understood and interpreted the depth of field and looking for humans (Figure 11). (Skeleton-mapping, as provided by PrimeSense<sup>23</sup>, required a stack of software to enable, including the <http://www.OpenNI.org> framework and the SensorKinect driver [#B22]\_ ‘<<https://github.com/avin2/SensorKinect>>’ as well as their free closed-source middleware library.) This proved to be much more accurate and not affected unduly by background movement.

However, this mouse-metaphor interface proved to be a poor one in the end. This was not due to technical reasons; simply, the human body is not suited to standing still with an arm held out for periods of time. With a hand resting on a desk, it is easy to have the accuracy needed to click on items. With the arm stretched out, it becomes difficult to hold it in a given position for any amount of time.

It is necessary to build the interface so that the interactions involve periods of relaxation or the ability to ignore actions made unintentionally. One particularly successful form of interaction is selecting items from a menu, where the hand is raised to select an item from a list and a choice is made by swiping the hand across. Swiping the other hand back across is used to cancel that choice. This has the benefit that in between choosing, the arms can be left to relax without worrying that a mouse-pointer would skitter across the screen and select or highlight something unintentionally.

### Using the Kinect to control molecule visualisation

Moving away from the fume cupboard for a moment, one potential use-case for the Kinect is to use it as an appliance in a given meeting room or open space for collaboration or interaction. We implemented this by using the Kinect’s skeletal mapping functions to track hand positions which were then broadcast via multicast to any computer within

<sup>20</sup> Kreylos, Oliver. 3D video capture with Kinect

<sup>21</sup> Kreylos, Oliver. Two Kinects, One Box

<sup>22</sup> Coldewey, Devin. Kinect specifications

<sup>23</sup> PrimeSense

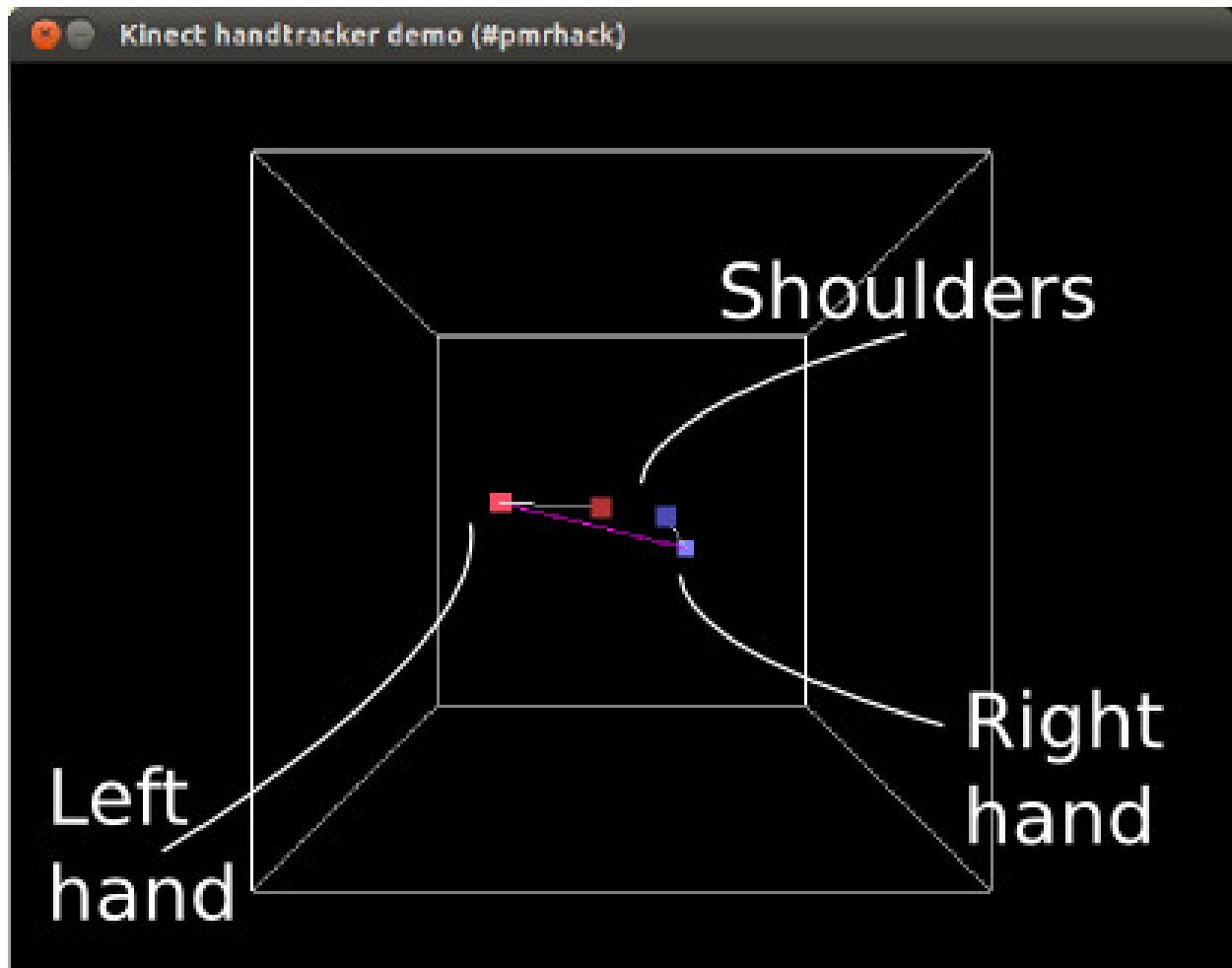


Figure 44.11: Figure 11. Screenshot showing 3D visualisation of the Kinect's interpretation of the volume it can see  
**Screenshot showing 3D visualisation of the Kinect's interpretation of the volume it can see.** Only the user's shoulder and hand joints are shown for clarity.

the same network. This has the advantage that the client computers need no knowledge of how to read and process Kinect data directly, only software that can make use of the hand-position mappings in the same way it might read the location of a mouse pointer<sup>24</sup>. At the end of the project, a symposium was held in honour of Dr Murray-Rust's ideas<sup>25</sup>. Immediately preceding the symposium was a hackfest at which our experiences working with the Kinect were used to control the rotation and zooming of a molecule in Jmol (Figure 12).



Figure 44.12: Figure 12. Controlling rotation of a molecule using hand gestures

**Controlling rotation of a molecule using hand gestures.** The Kinect is on the bench, on top of the silver cylinder. Picture taken at PMR Symposium.

#### 44.2.7 Speech recognition

For a chemist working in the lab, the ability to use speech to communicate with their computer would be a great advantage. Preparative work before the start of Ami showed that Windows Speech Recognition (WSR - the speech recognition facilities build into Windows 7) could be used to control<sup>26</sup> the Chemistry add-in for Microsoft Word, Chem4Word<sup>27</sup>. Dragon Naturally Speaking (DNS) is the leading speech recognition package, so this was also evaluated.

<sup>24</sup> O'Steen, Ben. Code for tracking using the Kinect

<sup>25</sup> Symposium: Visions of a Semantic Molecular Future

<sup>26</sup> Using speech to control input of chemistry in Microsoft Word

<sup>27</sup> Chemistry Add-in for Microsoft Word

Both WSR and DNS have the ability to define macros for navigating around the screen, clicking on buttons, etc. The ability of these tools to use speech to control an application was tested by trying to use only speech to operate the Chemistry Department's ELN, a Java-based application developed by IDBS<sup>28</sup>. However, neither WSR or DNS worked very well with Java applications; WSR in particular is much more functional with Windows-based applications because it has closer ties into the operating system's understanding of what objects are being displayed. Controlling the ELN was best achieved using send-key type instructions; if keyboard shortcuts did not exist for a particular activity, then this limited the possible actions.

Both WSR and DNS can be used to dictate text into Java applications. The demands of a specialist chemistry vocabulary are pretty stringent, however, so for chemistry dictation the transcription accuracy varies significantly. It is possible to train each package to improve voice recognition, but this is a potentially enormous topic and it was not done to any significant extent in this project. DNS has the ability to digest sample documents that the user gives in order to understand their particular vocabularies, but this was not explored beyond initial configuration.

For the Ami application WSR was chosen for doing further development, mainly because of licensing costs; DNS is very expensive. WSR has a free extension, Windows Speech Macros, which enables tailoring of the commands used when speech is recognised. This was fairly successful and it is possible to navigate all of the screens and buttons in the Ami application. WSR listens for key phrases, and then uses send-key instructions (most commonly an Alt- < single-key > code) to send key codes to buttons in the Ami application. Additionally, WSR can be used for dictating comments (experiment observations) directly into Ami, though we did not have the time to investigate its accuracy or ways of enhancing it and it was only used by the development team.

### 44.3 Outcomes & Conclusions

The main outcome from the project was a demonstrator application that shows how experiments and the environment around them can be monitored using various sensors and video monitors. We had the stretch goal of actually having this used by real chemists for real experiments, but unfortunately time prevented us from polishing the system to a sufficient level to allow this.

At the launch meeting for the Dial-A-Molecule EPSRC Grand Challenge<sup>29</sup>, a common theme that emerged was the need to have access to chemical data. Much of the data generated in laboratories does not get collected and made available in a form that other chemists can use. Time pressures mean that very often scientists do not get around to making their data available. The Ami project showed how there is huge potential for computers to help the bench chemist in their activities in the lab, and to make much of this information available for further use. In its six months Ami has investigated many technologies and ideas; an obvious follow-on to the project is to consolidate these ideas into a fully integrated tool that can be used in real laboratories. Additionally, there is much potential for further work on the flow of data from the experiment to electronic lab notebooks to an embargo management tool and thence to open repositories, thus facilitating re-use. Reviewers have pointed out how important this type of data will be for retrospective analysis, especially in cases of unexpected results or experimental reproducibility.

### 44.4 Competing interests

The authors declare that they have no competing interests.

### 44.5 Authors' contributions

BJB was the project leader/manager. He did the project's bureaucracy and organisation, guided the project's direction, and did much of the investigations into the speech. He also wrote the paper.

---

<sup>28</sup> IDBS

<sup>29</sup> Dial-A-Molecule

ALT did most of the programming for the Ami application, and investigated most of the software tools used, including investigations into video capture. He also worked on the Arduino development.

MS, PM and SC worked on Arduino hardware and software development. They also added detail to the project's use-cases, and integrated the Arduino development into the Ami application. They also did demonstrations of the project.

BOS worked on building the video capture for time-lapse video motion-snapshot monitoring. He also did the investigations into the use of the Arduino.

SEA developed software for driving the RFID reader system and advised on application development, testing environments, and troubleshooting. He also helped with the brainstorming session.

JAT configured the code management systems used by the project, and provided feedback, expertise and advice on the design of the application.

PMR was the principal investigator on the project giving advice, guidance, encouragement and enthusiasm. He participated in the brainstorming session, reviewed project reports and both contributed-to and edited this paper.

All authors have seen and approved the final paper.

## 44.6 Appendix 1: Links to documentation, code resources, etc

Brainstorming session:

- Output from the brainstorming session: <https://bitbucket.org/jat45/ami/downloads/Notes%20output%20from%20Ami%20brainstorming%20session.pdf>

Project website & tags:

- Project blog: <http://amiproject.wordpress.com>
- Project Wiki: <http://bitbucket.org/bjb45/ami-project>
- Project code: <http://bitbucket.org/jat45/ami/>

Software used:

- Java development - IntelliJIDEA Community Edition: <http://www.jetbrains.com/idea/download/>
- Speech Macros - Windows: <http://code.msdn.microsoft.com/wsrmacros>
- JFreeChart Java graph package: <http://www.jfree.org/jfreechart/>
- Timeline application: <http://thetimelineproj.sourceforge.net/>
- Natty - Java library for processing data/times: <http://natty.joestelmach.com/>
- Video capture - VLC: <http://www.videolan.org/vlc/>

Project code:

- Data logger - Arduino program: <https://bitbucket.org/jat45/ami/src/096e6df85d58/arduinoControllerWithoutSD/>
- Ami application: <https://bitbucket.org/jat45/ami>
- Experiments with the Kinect: <https://github.com/benosteen/Kinect-tracking-code>

Other tools used:

- Speech recognition - Dragon Naturally Speaking: <http://nuance.co.uk/>
- RFID reader - Touch-A-Tag: <http://www.touchatag.com/>

## 44.7 Acknowledgements

Funding from JISC for the Ami project is gratefully acknowledged, as is funding from Unilever for PMR. Ami was a six month project under the “JISC Rapid Innovation Grants 10/09” programme<sup>30</sup>. Our thanks to Drs Richard Turner, Nadine Bremeyer and Chris Lowe for their scientific advice. The project team was located in the Unilever Centre in the Chemistry Department at the University of Cambridge.

---

<sup>30</sup> Grants for the Virtual Research Environment - Rapid Innovation funding call

# THE PAST, PRESENT AND FUTURE OF SCIENTIFIC DISCOURSE

## 45.1 Abstract

The science journal is 346 years old in 2011, having evolved continuously but largely incrementally over that period. Its reinvention for an online presence has largely preserved its previously printed nature, in the sense that much of the increased functionality which is potentially offered by this new medium has yet to be exploited. In the present article an attempt is made to discuss two previously published papers, one in 1953 and the other in 2010, and to illustrate how additional functionality can be implemented in the form of accessible data sourced from quantum mechanical calculation and how subsequent discourse in the form of blogs may add to the process. In this sense, the reader of this article is invited to try for themselves whether these enhancements improve their scientific understanding, and whether such enhanced journals are good models for the future evolution of the genre.

## 45.2 Introduction

The first journal devoted exclusively to science is generally accepted to have first appeared in 1665 as the Philosophical Transactions (of the Royal Society). The inaugural issue<sup>1</sup>, which famously carries an account by Robert Boyle of “a very odd monstrous calf”, is perhaps not science as we know it nowadays, but it does remind one rather of what one might find in a personal blog, a similarity I will return to later. The structure of the scientific journal and the articles published by this means evolved constantly and mostly incrementally during the next 330 years. One of the more significant, but nevertheless still incremental changes was adding an online presence during the late 1990s. Indeed, some journals founded in the late 1990s offered only an online version<sup>2</sup> and there are signs that some older journals may be preparing to abandon the (relatively expensive) printed form. Access to the online versions in 2011 is predominantly *via* a format which is perhaps best described as digital paper (PDF), although most journals also offer the articles in an alternative hypertext (HTML) format. There is no journal yet that has adopted a format increasingly used for books, the epub/epub3 standard<sup>3</sup>. On the horizon is also the fifth major evolution of hypertext markup language known as HTML5<sup>45</sup>, which strives to offer a richer interactive medium to the reader. Certainly machines (*e.g.* such as those working on behalf of search engines such as Google) now also automatically process the articles published in the modern journal, indexing the full-textual content, adding rich metadata on the topics therein described, and noting (but largely incapable of truly indexing the context of) the images and figures. Software can also usefully replace the old processes of binding the printed journal and the storage of volumes on shelves in a library or an office,

---

<sup>1</sup> An Account of a Very Odd Monstrous Calf

<sup>2</sup> The Internet Journal of Chemistry: What Lessons Have We Learned After Our First Year?

<sup>3</sup> What to expect in EPUB3

<sup>4</sup> Introducing HTML5

<sup>5</sup> HTML5

a process delightfully described (by biologists, not chemists) as *defrosting the digital library*<sup>6</sup>. These processes largely address only the bibliographic issues (*via* rich metadata harvesting) rather than attempting to defrost the scientific or chemical content itself. It is the issues involved in defrosting the latter type of information and data that the present article addresses.

Data has always been the “elephant in the room” of scientific publishing. Because the costs of printing and distributing paper are still significant to this day, print was never really been considered a viable mechanism for distributing the (often very large amounts of) data, whether raw, or partially processed, on which almost all scientific models, theories and their interpretations are based. Instead, starting in the early 1990s and coincident with the first introduction of the Internet, many science journals offered an annex to the main journal in the form of supporting or supplemental information. This was provided in final form by the authors themselves, and the journal itself added little extra value such as indexing to this (often purely visual) content. It was very much up to an interested reader to add their own value to any (visual, textual or numerical) supporting data that might be associated with an article.

The long view over a 350 year period is that these evolutions of the journal could be regarded as largely relating to the production and delivery processes of journals, and arguably have not been matched by similar advances in how scientists *consume* or use journals. In this essay, I will analyze two chemical articles, published in respectively 1953 and 2010, from the point of view of how the original journal presented the scientific discourse, what the limitations of that presentation might have been, and the prospects of how it could evolve into a step-change rather than incremental change in that discourse.

### 45.3 The relationship between a journal article and data

I start the analysis with article that contains (*inter alia*) what has been described as the most famous scientific diagram of the 20th Century, the representation<sup>7</sup> of the double helical structure of the DNA molecule by Watson and Crick (Figure 1).

Indeed, this diagram is the **only** one that actually appears in the article, and one would seek in vain any diagrammatic elaboration of what the molecular structure of DNA is (although components such as deoxyribose or guanine are named as such in the text). Anyone seeking to repeat Watson and Crick’s model building would certainly have to acquire additional molecular data from another sources. Some of that missing information is shown here in Figure 2, although this only describes the connectivity of the various atoms in a single strand of DNA, and not the two or three dimensional relationships of the (125 in this example) individual atoms. Note also that this diagram is presented here for visual consumption by a human, who still has to recover additional semantics such as the stereochemistry at the three stereogenic ribose centres, and note carefully that the unit represented must be accompanied by positively charged counter-ions.

Armed only with the one diagram actually published, curiosity might lead one to pose a scientific question such as “How did Watson and Crick assign the helix as right rather than left handed”? In other words, on what data did they base that conclusion? This does matter! For example, some 733 articles have appeared in the science literature over the last 20 years or so where DNA is represented as having left-handed helicity, in most cases certainly erroneously<sup>8</sup>. Coincidentally, similar issues of left or right-handedness were to be found when Pauling presented his  $\alpha$ -helix models of proteins. In fact, almost all protein helices exhibit right-handedness<sup>9</sup>. A partial answer to that question is actually given in what is called the *full version*<sup>10</sup> of the preliminary article<sup>7</sup> (published as it happens by the Royal Society). Here we are told the following:

- that both chains follow right handed helices ...
- because left handed helices can only be constructed by violating permissible van der Waals contacts.

---

<sup>6</sup> Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web

<sup>7</sup> Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid

<sup>8</sup> The Left Handed DNA Hall of Fame

<sup>9</sup> Pauling’s Left-Handed  $\alpha$ -Helix

<sup>10</sup> The Complementary Structure of Deoxyribonucleic Acid

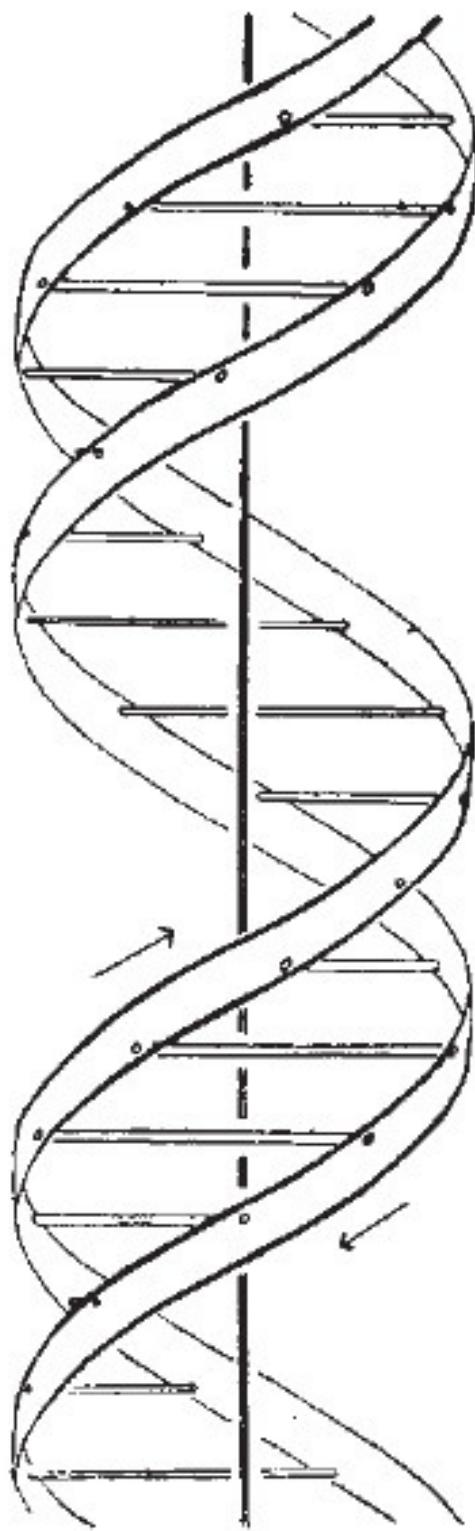


Figure 45.1: Figure 1. The DNA double helix (reproduced with permission), showing a right handed or B-helix  
**The DNA double helix (reproduced with permission [#B8]\_\*\*), showing a right handed or B-helix\*\*.**

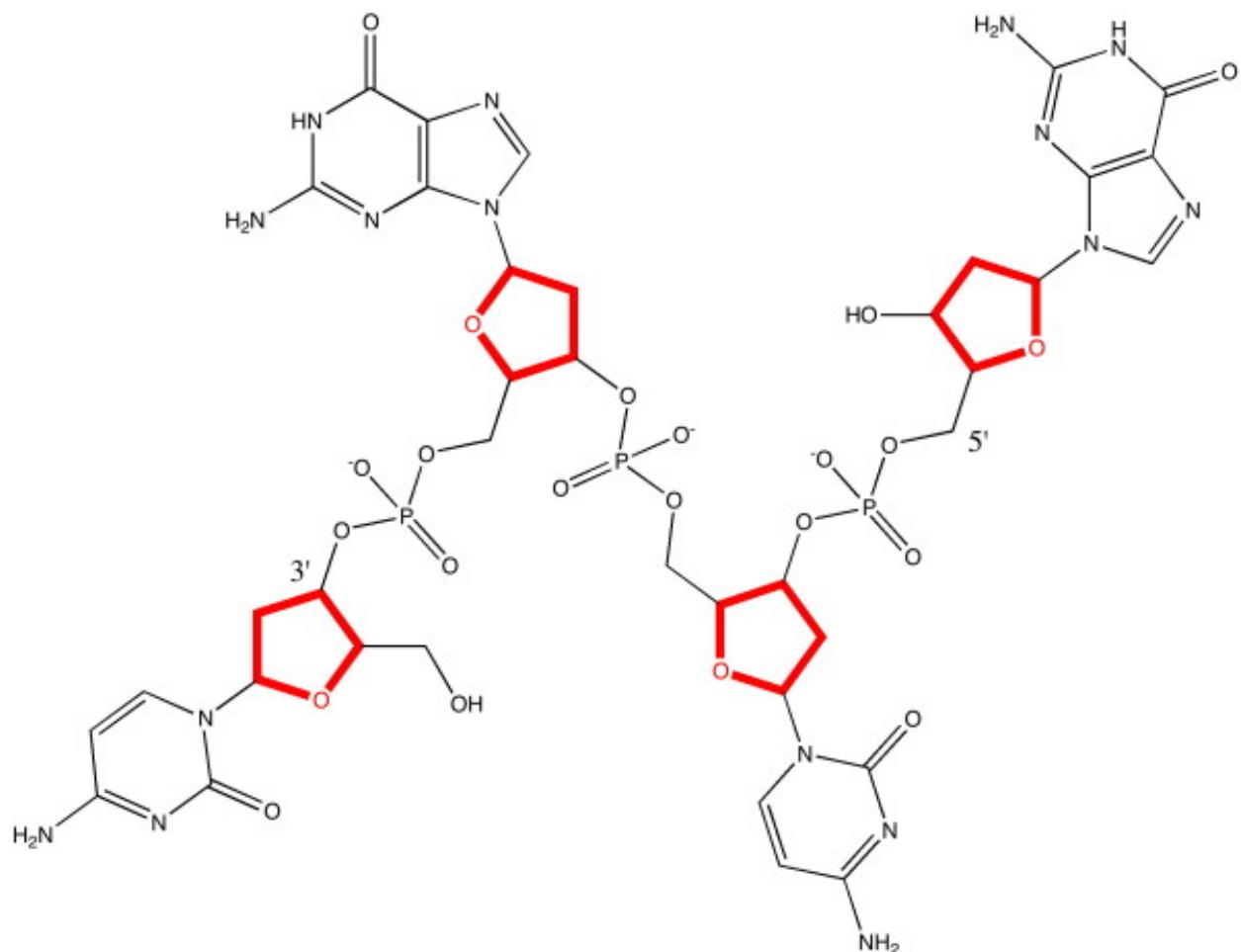


Figure 45.2: Figure 2. The molecular basis of one strand of DNA, based on the CG bases  
**The molecular basis of one strand of DNA, based on the CG bases.**

- We are informed that such permissible contacts include the approach of any two hydrogen atoms in the molecule to a distance of **no less** than 2.1Å.
- We are not however informed what the violations might be in a left handed helix that excludes this model. In other words, just how close can two hydrogen atoms separated (for intramolecular contacts) by at least four bonds approach? In fact, distances of ~1.85Å or less have been observed<sup>11</sup>.

In this same full article by Watson and Crick<sup>10</sup>, we are given a table of numerical (polar) coordinates describing the positions of twelve key atoms, but it would have taken a very determined scientist to have used only this combination of information to easily confirm the assertion that a left-handed helix is excluded. Perhaps the lack of a model with which the reader could experiment might account for the relatively slow recognition of the importance of this article in the immediate years following its publication, and the observation that whilst a physical model of DNA had of course been built, it was only available for viewing (but not modifying) by visiting Cambridge!

One tool that modern chemistry now has at its disposal (which Watson and Crick did not have) are accurate molecular models based on quantum mechanical calculations. Such a molecule is quite a challenge to model, since the computation has to take into account subtle interactions such as dispersion (long range correlation) effects, which are more or less equivalent to the van der Waals contacts referred to by Watson and Crick, the ionic phosphate groups, the planar bases and how they stack, so-called anomeric effects at the base-sugar connecting C-N bond, hydrogen bonds between both the obvious NH...N and NH...O atoms and less obvious ones such as C-H...O, and not least the capacity to deal self-consistently and accurately with the optimal positions of (at least) 250-254 atoms. In reality, such models have only very recently become available<sup>12</sup>. To illustrate how this famous article from 1953<sup>2</sup> could now be published in a journal in 2011, I have taken the liberty of updating the original diagram with the one shown in Figure 3 (see additional file<sup>6</sup>). The additional information is made available *via* the figure caption and in Table 1 to conform to established practice in more conventional articles.

Additional file 1

#### **Interactive Jmol-enhanced version of Figure 3.**

[Click here for file](#)

This is a model of a DNA duplex tetramer, built using only the bases CGCG or ATAT in this example, with inclusion of three phosphate groups and calculated for both the left- and right-handed helical form. The first of these was the one deprecated by Watson and Crick on the basis that the model violates permissible van der Waals contacts. The geometry is optimized to high convergence using a recent density functional formalism ( $\omega$ B97XD)<sup>13</sup> which incorporates a correction for the attractive dispersion component of the van der Waals interactions. Justification for the use of this functional in describing hydrogen bonding has recently been published<sup>14</sup>. A 6-311G(d,p) basis set results in the wavefunction being described by up to 3468 basis functions, close to the practical limit using standard computing resources available in 2010. The ionic nature of the system, deriving from the phosphate groups, was treated using a self-consistent-reaction-field continuum solvent (water)<sup>15,16</sup> as implemented in the Gaussian09 package, revisions A.02 and B.01. In such a model, duplex formation by combining two tri-anionic chains is nevertheless exothermic in the computed free energy (Table 1), which suggests the model is not physically unrealistic. Although in principle a full ion-pair resulting from inclusion of a solvated positive counterion (typically Na<sup>+</sup> or NH<sub>4</sub><sup>+</sup> with additional water molecules) could also be treated using this method<sup>17</sup>, the resulting model is too large and complex for the current available computational resources.

The resulting model is presented in this article using suitable software (Jmol in this instance<sup>18</sup>) which itself reads the optimized coordinates of all 250-254 atoms and renders these in suitable form for the reader. Annotation with identified close contacts between pairs of hydrogen atoms or other close contacts can be easily scripted in, and in

<sup>11</sup> Pericyclic reactions of 2-pyrones with nonconjugated dienes. Conformational analysis of the double Diels-Alder adducts by molecular mechanics calculation

<sup>12</sup> Can 1,3-dimethylcyclobutadiene and carbon dioxide co-exist inside a supramolecular cavity?

<sup>13</sup> Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections

<sup>14</sup> Assessment of the Performance of DFT and DFT-D Methods for Describing Distance Dependence of Hydrogen-Bonded Interactions

<sup>15</sup> A Smooth Solvation Potential Based on the Conductor-Like Screening Model

<sup>16</sup> Continuous surface charge polarizable continuum models of solvation. I. General formalism

<sup>17</sup> Successful Computational Modeling of Isobornyl Chloride Ion-Pair Mechanisms

<sup>18</sup> Web-Based Molecular Visualization for Chemistry Education in the 21st Century

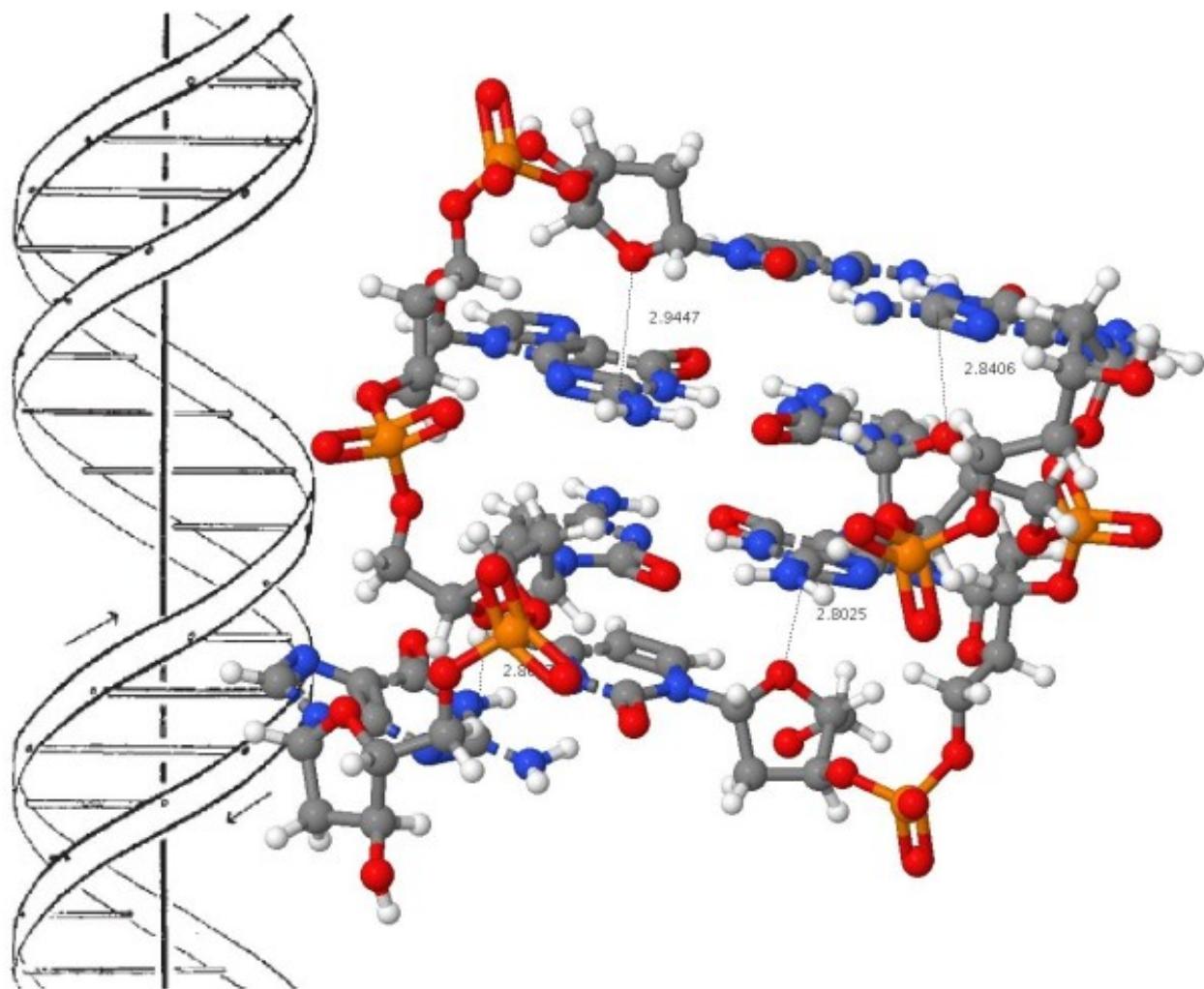


Figure 45.3: Figure 3. A model of the Z-d(CGCG)<sub>2</sub> DNA duplex with a geometry optimized at the ωB97XD/6-311G(d, p) level and embedded in a continuum solvent field for water

:sub:'2' DNA duplex with a geometry optimized at the \nonascii\_9|B97XD/6-311G(d, p) level and embedded in a continuum solvent field for waterA model of the Z-d(CGCG). (a) Load

principle a rich variety of actions and analyses can be built into the figure which are all based on a combination of the underlying data and algorithms implemented by the (Jmol) software. Importantly, the original data used for generating the model can be extracted from the model (the process is described here<sup>19</sup>) and can then be re-applied using alternative software which might provide further analysis, or indeed alternative technologies such as stereoscopic processing. These processes now turn the journal from merely a visual information source into an active scientific instrument. We may also speculate at this point on other forms of rendering data. Jmol was written in Java, and requires the browser to support a Java virtual environment. New generations of mobile information devices, which are primarily designed for long battery life, may not continue with this approach. Instead, one favoured alternative is to interface the browser directly to the graphical hardware using e.g. WebGL, and to implement the functionality of something like Jmol using the emerging HTML5 standard and appropriate scripts<sup>20</sup>. The native ability of a browser to provide such enhanced processing is already apparent in support for SVG, a markup language for vector graphics (examples of which are included in Figure 4 [see additional file]

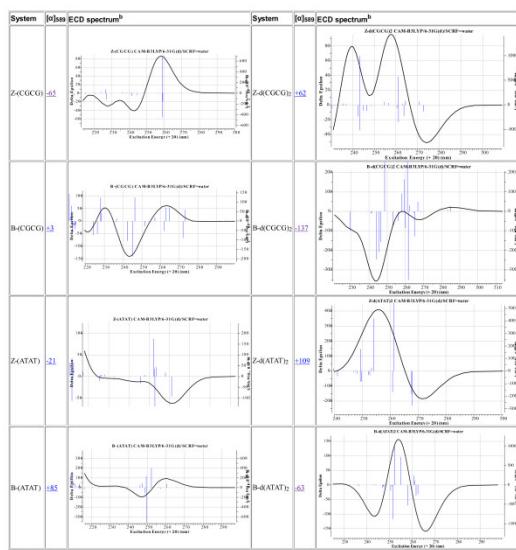


Figure 45.4: Figure 4. Calculated chiro-optical properties for DNA tetramers

**Calculated chiro-optical properties for DNA tetramers.**<sup>a</sup>Computed at geometries optimised at the  $\omega$ B97XD/6-31G(d) level with application of a SCRF solvent continuum field for water. Chiro-optical properties computed at the CAM-B3LYP/6-31G(d,p) level with application of a SCRF solvent continuum field for water. ECD spectra computed at the TD-DFT level, using Nstates = 25 and a linewidth of 0.14 with application of a SCRF solvent continuum field for water. Click on image to expand the view of the ECD spectrum. Click on expanded view of spectrum to access the digital repository entry for that spectrum. <sup>b</sup>ECD spectra are presented as scalable-vector-graphical diagrams (SVG). To view, use an SVG-capable browser.

Additional file 2

**Enhanced version of Figure** :ref:`4<figure\_4>`\*\*containing additional hyper links\*\* (This figure should be viewed with a web browser capable of SVG display, such as Chrome, FireFox, Safari or IE 9).

Please note: The figure is not currently displayed as intended by the author due to technical issues with the BMC site. This will be resolved as soon as possible.

Click here for file

The reader is invited to load the Z-d(CGCG)<sub>2</sub> coordinates in Figure 3. The lengths of the van der Waals H...H attractions and the nucleophilic O...C attractions from the ribose ether oxygen to the electrophilic carbons on the guanine base have been enumerated. Because of the complex 3D nature of this molecule, they can only be truly perceived

<sup>19</sup> (re)Use of data from chemical journals

<sup>20</sup> Learning WebGL

<sup>21</sup> WebGL Specification

if the structure itself can be viewed from any desired angle (something clearly not possible in a conventional journal diagram). It also allows successive layers to be viewed, each perhaps concentrating on a particular aspect, without destroying or overwhelming the initial simple elegance of the overall concept (of a double helix). The identified ~2.8Å O...C interactions to the guanine are unique to the Z or left handed helical form, and since the sum of the vdW radii<sup>22</sup> of these two contact atoms is ~3.22Å, these are presumed to be (electrostatically) attractive. The alternative B-d(CGCG)<sub>2</sub> stereoisomer reveals these O...C contacts are absent, being instead replaced by hydrogen bonds (~1.9-2.1Å) between the ribose ether-oxygen and the NH<sub>2</sub> hydrogens of the guanine. The sum of the O and H vdW radii is ~2.64Å, which suggests these are significantly attractive hydrogen bonds. There are additional C-H...O-P hydrogen bonded contacts of ~2.1-2.2Å. A similar divergence of attractive interactions emerges for chains built of AT bases. The Z-d(ATAT)<sub>2</sub> duplex has only one ~2.8Å O...C interaction to the adenine, with three others having rather longer lengths (~3.0-3.1Å). The B-d(ATAT)<sub>2</sub> duplex instead displays C-H...O contacts of ~2.4-2.5Å.

These differences can be more succinctly summarized as:

1. Z-d(CGCG)<sub>2</sub> is stabilized (*inter alia*) by a short contact between a carbon on the guanine and the ribose ether oxygen, of which there are four per four base pairs.
2. These contacts are replaced in B-d(CGCG)<sub>2</sub> by NH hydrogen bonded contacts to the ribose ether oxygen.
3. In Z-d(ATAT)<sub>2</sub>, the O-contacts to the adenine are much longer, which
4. in B-d(ATAT)<sub>2</sub>, are replaced by short CH...O contacts.

By embedding access to accurate coordinate data within Figure 3, the reader can select whatever level of detail they desire from the diagram. Part of the origins of the relative stability the Z- and B- helical forms is not simply due to the presence (or in this case absence) of “violation of permissible van der Waals contacts”, but also to several types of less common but nevertheless attractive interactions which may not have been inferred by building physical models alone. Such additional insights may in turn impact upon *e. g.* modeling one remarkable property of the DNA polymer, its ability to be stretched to almost twice its normal length without breaking<sup>23</sup>.

It is of course the accumulation of these effects that determines the overall stability of the structure (Table 1). The thermodynamic quantities are computed with inclusion of thermal energies, obtained by solving the appropriate partition functions using calculated vibrational frequencies. Since these require second derivatives of energy with respect to coordinates, a smaller basis set 6-31G(d) was used for the purpose (a calculation time of ~4 days on a 12-core processor is typical). The dispersion corrections were obtained at the slightly higher 6-311G(d,p) basis set level to allow interactions to H to be modelled more realistically.

These energies reveal some surprises. Firstly, the free energy for forming the duplex from the separated chains is significantly exothermic, despite the electrostatic repulsions resulting from each chain carrying a 3- charge. For the resulting helix, the B-d(CGCG)<sub>2</sub> form is 11.9 kcal/mol **less** stable in terms of total free energy than the Z-isomer, but is 4.2 kcal/mol **more** stable for the dispersion/van der Waals term, the criterion suggested by Watson and Crick as generally discriminating against the Z-form (although without a specification of the base type used for the model). The greater stability of the Z-form arises from a contribution of 3.9 from the entropy and 8.0 kcal/mol from the (zero-point energy corrected) enthalpy, which dominates the less favourable dispersion term.

The formation of a B-d(ATAT)<sub>2</sub> duplex is less exothermic than that of the CG duplex. It is now favoured by 5.2 kcal/mol over the Z-isomer in terms of free energy and by 12.8 kcal/mol in terms of dispersion contributions. The assertion often made<sup>24</sup> that the Z-helix is favoured by CG rich oligomers and the B-helix by AT-rich forms is thus confirmed by these calculations.

<sup>a</sup>Thermochemistry computed at geometries optimised at the ωB97XD/6-31G(d) level with application of a SCRF solvent continuum field for water, with thermal corrections derived from computed vibrational frequencies. The dispersion corrections are computed for geometries optimized at the ωB97XD/6-311G(d, p) level with application of a SCRF solvent continuum field for water. The display coordinates are those obtained at this level. <sup>a</sup>Free energy for the dimerisation of a single strand to a duplex.

<sup>22</sup> Consistent van der Waals Radii for the Whole Main Group

<sup>23</sup> Overstretching DNA at 65 pN Does Not Require Peeling from Free Ends or Nicks

<sup>24</sup> On the stability of single- and double-stranded DNA helices: The application of the PPT-MCF method on large fragments of DNA

Armed with optimized coordinates which include the weaker interactions between atoms one can annotate the basic models revealed in Figure 3 with other (computed) properties. For example the optical rotation  $[\alpha]_{589}$  has the value +62° for Z-d(CGCG)<sub>2</sub> and -137° for the B-diastereomer, perhaps surprisingly small values for such an apparently asymmetric molecule. Also surprisingly, the corresponding experimental measurement does not appear to have been reported. Optical rotations are known to be rather fragile, being sensitive to small variations in conformation and solvation, but the electronic circular dichroism spectrum is regarded as rather more robust. Unlike other forms of spectroscopy such as NMR or IR, which can be used to infer structure from simple rules based on the functional groups present (in other words, local properties), these chiro-optical properties tend to be more characteristic of the global features of the molecule. As a result, they can be very difficult to interpret without a reasonably accurate model based on these global properties. A quantum mechanical computation of the molecular wavefunction is one such model, and it is now increasingly routinely used to help interpret optical rotations, electronic (and vibrational) circular dichroism spectra. Theory can now handle molecules containing ~250–254 atoms, such as the DNA tetramers modelled here. Annotation of the models with the calculated ECD spectra is included here in the hope they might prove useful for the interpretation of the experimental spectra.

This example has illustrated how access to accurate data can help provide additional insights into the factors controlling the stability of molecular structures. In this case, the factors controlling the helical stability of DNA duplexes can be teased out. By incorporating these models directly into the journal article (and providing links to digital repositories where a more complete dataset can be acquired if needed) the readers of the journal have an opportunity to discover their own insights within their own spheres of interest.

## 45.4 The crystal structure of 1,3-dimethylcyclobutadiene

The second example chosen for discussion is a more contemporary one. In July 2010, a article appeared<sup>25</sup> reporting the single-crystal X-ray structure of 1,3-dimethylcyclobutadiene achieved by confinement in a crystalline matrix (Figure 5 [see additional file *Mona Lisa of molecules*, and its very instability means that conventional experiments on it are very challenging. The article was however conventional in the sense of being made available in (more or less equivalent) HTML and PDF versions.

Additional file 3

[Interactive Jmol-enhanced version of Figure 5.](#)

[Click here for file](#)

Much of the scientific insight was carried in the form of four colour figures, all presented conventionally as single layered graphics with the viewpoint selected by the authors. Information on acquisition of the data on which these figures were based was given as citation 28 in that article, which lists deposition numbers (CCDC 764864–764868) and a URL that would enable a CIF file for each entry to be downloaded. Whilst this retrieval process is not entirely automatic, it does take only a few minutes to acquire the data. The usefulness of the file is of course predicated on the reader also having access to appropriate software for analysis of a file in this format. It is also important to note that a CIF file allows inspection only of the refined crystallographic model presented in the article and the statistics associated with that model; it does not allow the user access to the underlying (*hkl*) diffraction data which would allow other models to be refined and assessed.

The reaction scheme reported<sup>25</sup> for photochemical generation of trapped 1,3-dimethylcyclobutadiene is shown below in Figure 5. It differs from the original in showing a thermally activated reaction arrow connecting the 1,3-dimethylcyclobutadiene **4** to **2**. This last possibility is not explicitly discussed in the original report<sup>25</sup>, although there is there an implicit assumption that this process is slow at the temperature of the experiment, 175K. The original article therefore seeks to persuade the reader on the basis of crystallographic evidence that the structure of **3** or **4** has been established, with the aid of the four colour figures included in the article, and (optionally for the reader) with acquisition of the CIF files.

<sup>25</sup> Single-Crystal X-ray Structure of 1,3-Dimethylcyclobutadiene by Confinement in a Crystalline Matrix

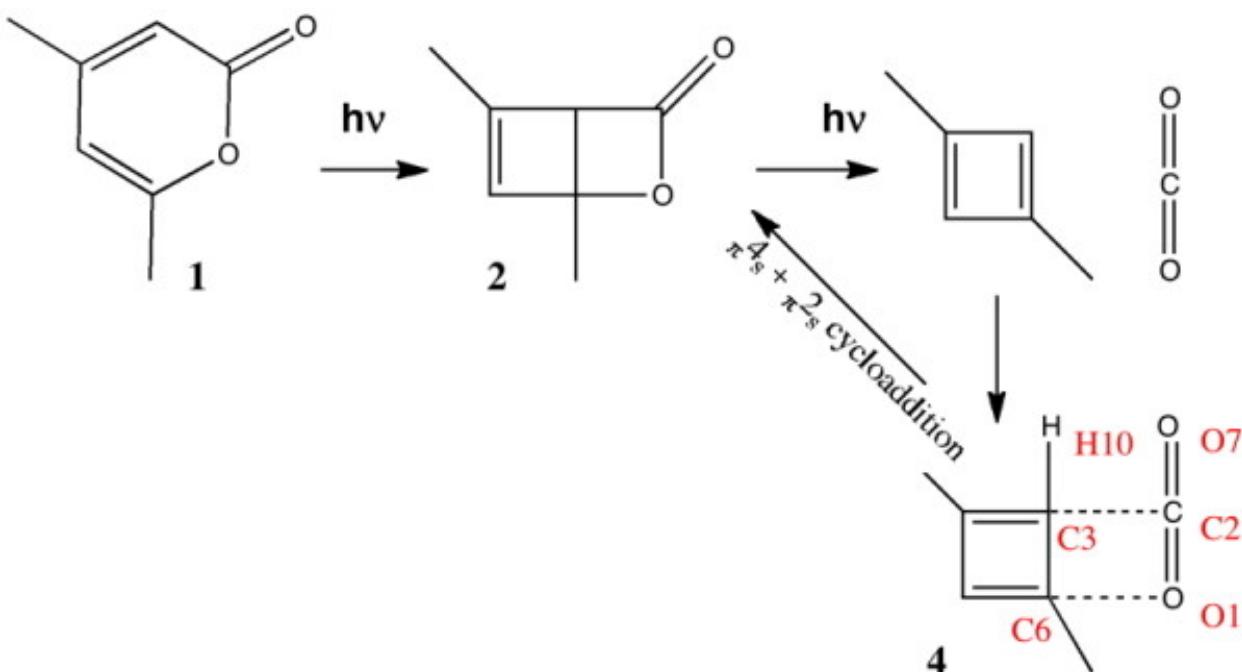


Figure 45.5: Figure 5. The reaction leading to 1,3-dimethylcyclobutadiene<sup>25</sup>

**The reaction leading to 1,3-dimethylcyclobutadiene<sup>25</sup>.** The numbering shown for 4 corresponds to that for the published coordinates. Load coordinates for host-guest structure and just the guest only.

The theme of the present article is to ask how a reader's experience and perception of a scientific article might be enhanced or simply altered by adopting new forms of presentation. In Figure 5, the relevant CIF file can be loaded as a second layer into the reaction diagram. Because of the relatively large number of host atoms involved, the effect can be somewhat overwhelming when this is done and the interpretation may also be made more complex by the presence of disorder in the guest. A further layer of interpretation can be added by annotating the diagram with selected atom-atom distances; the reader can use the display software to add further such annotations of their own if they wish. There are many other actions the reader can perform at this point<sup>19</sup>. A further, this time smaller, alternative layer that contains only the kernel of the scientific problem (as perceived by the present author, which may or may not correspond to the perception of the original<sup>25</sup> authors) has been added here, and again four key measurement annotations made, together with selected bonds highlighted in a different colour.

The scientific problem can now be stated in the form of the following questions.

1. What are the kinetics of the reverse reaction of **4** to give **2** at 175K?
2. Does the crystallographic evidence convince that the guest is best described as 1,3-dimethylcyclobutadiene in close proximity to a detached molecule of carbon dioxide?
3. More specifically, how should the interaction between the labelled atoms C2 and C3 be interpreted? Should it be considered a strong van der Waals contact, as suggested by the original authors<sup>25</sup> or as a covalent bond? The same question might apply to another atom pair, O1 and C6 also connecting carbon dioxide and the cyclobutadiene.
4. Likewise, how should the angles O1-C2-O7 or C2-C3-H10 be interpreted?

The reader may note a common theme emerging between these questions and the origins of helical stability in DNA as discussed above.

The first of these questions was in fact posed in the form of a blog, written by the present author<sup>26</sup> and based on chemical precedent and entropic arguments. It was posted in August 2010, little more than a month after the original report was first published. The precedent for this form of discourse when addressing a scientific issue had already been established<sup>2728</sup>. Questions 2-4 emerged more conventionally and a little later in November 2010 in the same journal as the original article, and took the form of comments submitted by two independent groups<sup>2930</sup>. The original authors have a right of reply to such comments, which they took<sup>31</sup>. These various participants in the debate all had access to the same CIF data as is transcluded into Figure 5. The debate to this point was summarized in a second blog post<sup>32</sup>, and this and the original post themselves attracted ~15 responses in the form of appended comments. These posed further questions, on themes such as the computed structure of 1,3-dimethylcyclobutadiene, a debate on how much energy was required for angular distortion of O1=C2=O7 as an isolated molecule, and whether molecule 2 is transparent to light in the 320 to 500nm excitation range employed by the original experiments. This latter point was followed up by calculations of the UV-visible absorption spectrum of 2 inside the host cavity, also appended to the blog, and finally by calculations of the predicted vibrational spectra. The next stage in the discourse occurred in a conventional journal<sup>12</sup>, taking the form of a set of calculations on the likely barrier preventing 4 and carbon dioxide from recombining inside the host cavity, and addressing question 1 above in more complete detail. This article did have one less conventional aspect; in the “rich HTML” version, an interactive version of the table of data was made available<sup>33</sup> in very much the manner adopted for Figures 3 and 5 in the present article. Additionally, there were links in this table to digital repository entries<sup>34</sup>, which would enable any interested reader to access the complete archived details of all the calculations reported in that article.

Shortly after this last article was published, in December 2010, several publishers chose to highlight this emerging debate with editorial blog posts of their own<sup>35363738</sup>. These posts in turn attracted further comments, including several by one of the original authors. One comment in particular<sup>39</sup> entitled “Request of calculated structure data” highlighted an important aspect concerning the accessibility of previously reported data<sup>12</sup>. This alludes to the “rich HTML” table, and the observation that it is important to provide information on the file formats in which data is held, so that appropriate conversions if needed and concomitant visualization can be performed. This particular query was answered in the form of another blog post<sup>19</sup>, and applies directly to the issue of how to re-use data associated with the current article (Figures 3 and 5).

The scientific discourse described above regarding the nature of the species in a host crystal lattice is still ongoing, and so a final consensus (if ever achieved) cannot be reported at the time of writing. It is noteworthy that the primary (*hkl*) crystallographic data relating to the original measurements has been provided upon request<sup>40</sup> and so further analysis of alternative crystallographic refinement models is now possible.

## 45.5 Conclusions

The two scientific examples discussed in this article span 57 years, a relatively short period in the history of the scientific journal. The first is arguably the most influential scientific article of the 20th century, and clearly the absence of data associated with it has not held back its recognition as such. What is also clear is that addition of such data, albeit 57 years after the original report, may have the potential to reveal further insights into the structure of DNA that may

<sup>26</sup> Reactions in supramolecular cavities - trapping a cyclobutadiene: ! or ?

<sup>27</sup> Portals, blogs and co.: the role of the Internet as a medium of science communication

<sup>28</sup> NaH as an Oxidant - Liveblogging!

<sup>29</sup> Comment on “Single-Crystal X-ray Structure of 1,3-Dimethylcyclobutadiene by Confinement in a Crystalline Matrix”

<sup>30</sup> Comment on “Single-Crystal X-ray Structure of 1,3-Dimethylcyclobutadiene by Confinement in a Crystalline Matrix”

<sup>31</sup> Response to Comments on “Single-Crystal X-ray Structure of 1,3-Dimethylcyclobutadiene by Confinement in a Crystalline Matrix”

<sup>32</sup> Can a cyclobutadiene and carbon dioxide co-exist in a calixarene cavity?

<sup>33</sup> Table 1<sup>a</sup>Calculated relative free energies (kcal mol<sup>-1</sup>) for various models for the reaction between 4 and carbon dioxide

<sup>34</sup> SPECTRa: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories

<sup>35</sup> Crystallographic Confusion

<sup>36</sup> Debating cyclobutadiene

<sup>37</sup> Doubt cast on X-ray structure of trapped reactive species

<sup>38</sup> Has a cyclobutadiene species been isolated?

<sup>39</sup> comment entitled “Request of calculated structure data”

<sup>40</sup> comment entitled “Some experimental details may be useful in this context”

not have hitherto been highlighted. Whether such a data-rich reformulation of the original problem has any measure of impact remains to be established. The second article is only months old, but in that brief period has been subjected to the kind of scrutiny that can only be achieved by having access to rich data sets. One might fairly conclude that the scientific article has evolved to enable that scrutiny. The article that you are now reading I suggest is one model for how such scientific discourse can be both improved and accelerated. It remains to be seen if scientists are prepared to author such articles in the future. There is an early example<sup>41</sup> of an article where both the discourse and the data supporting that discussion were seamlessly integrated into one (XML-based) document, with the presentation being made available to the reader by application of suitable stylesheet-based transformations. The production of an article in this form was however non trivial. Since then tools have appeared to facilitate the process<sup>4243</sup> and the task now much be to reach both the hearts and the minds of scientific authors to encourage them to start adopting this form of enhanced scientific discourse.

## 45.6 Competing interests

The author declares that they have no competing interests.

## 45.7 Acknowledgements

Derived from a presentation given for a symposium celebrating the career of Peter Murray-Rust, in Cambridge UK on January 17, 2011.

---

<sup>41</sup> Development of chemical markup language (CML) as a system for handling complex chemical content

<sup>42</sup> CINF 115 Chem4Word

<sup>43</sup> Chemistry Add-in for Word

# OPEN BIBLIOGRAPHY FOR SCIENCE, TECHNOLOGY, AND MEDICINE

## 46.1 Abstract

The concept of Open Bibliography in science, technology and medicine (STM) is introduced as a combination of Open Source tools, Open specifications and Open bibliographic data. An Openly searchable and navigable network of bibliographic information and associated knowledge representations, a Bibliographic Knowledge Network, across all branches of Science, Technology and Medicine, has been designed and initiated. For this large scale endeavour, the engagement and cooperation of the multiple stakeholders in STM publishing - authors, librarians, publishers and administrators - is sought.

BibJSON, a simple structured text data format (informed by BibTex, Dublin Core, PRISM and JSON) suitable for both serialisation and storage of large quantities of bibliographic data is presented. BibJSON, and companion bibliographic software systems BibServer and OpenBiblio promote the quantity and quality of Openly available bibliographic data, and encourage the development of improved algorithms and services for processing the wealth of information and knowledge embedded in bibliographic data across all fields of scholarship.

Major providers of bibliographic information have joined in promoting the concept of Open Bibliography and in working together to create prototype nodes for the Bibliographic Knowledge Network. These contributions include large-scale content from PubMed and ArXiv, data available from Open Access publishers, and bibliographic collections generated by the members of the project. The concept of a distributed bibliography (BibSoup) is explored.

## 46.2 Technical note

This paper was created using the technologies described in the text. All bibliographic entry references and bibliographic entries were managed in BibJSON then included in the HTML document following the Scholarly HTML convention. The document itself is formally consistent with these specifications and can be read as a normal HTML document. It would alternatively be possible to embed bibliographic records in the document directly from BibJSON via JavaScript. The “flat HTML” should be taken as the definitive version, and can be re-purposed into other formats (Additional file)

Additional file 1

**HTML document.** The complete HTML document of this manuscript.

[Click here for file](#)

## 46.3 Competing interests

The authors declare that they have no competing interests.

## 46.4 Authors' contributions

All authors took equal parts in creating the concepts and tools reported and all authors wrote and revised the manuscript.

# SEMANTIC SCIENCE AND ITS COMMUNICATION - A PERSONAL VIEW

## 47.1 Abstract

The articles in this special issue represent the culmination of about 15 years working with the potential of the web to support chemical and related subjects. The selection of papers arises from a symposium held in January 2011 ('Visions of a Semantic Molecular Future') which gave me an opportunity to invite many people who shared the same vision. I have asked them to contribute their papers and most have been able to do so. They cover a wide range of content, approaches and styles and apart from the selection of the speakers (and hence the authors) I have not exercised any control over the content.

## 47.2 Overview

The articles have a common theme of representing information in a semantic manner - *i.e.* being largely “understandable” by machine. This theme is common across science and many of the articles can and should be read by people outside the chemical sciences, including information scientists, librarians, *etc*. An emergent phenomenon of the last two decades is that information systems can grow without top-down directions. This is disruptive in that it empowers anyone with energy and web-skills, and is most powerful when exercised in communities of people with similar or complementary skills.

It is often possible to move very quickly, and in our hackfests (one was prepended to the symposium) we have shown that it is possible to prototype within a day or two. This creates a new generation of scientist-hackers (I use “hacker” as “A person who enjoys exploring the details of programmable systems and stretching their capabilities”<sup>1</sup>). Several of the authors in this issue would regard themselves as “hackers” and enjoy communicating through software and systems rather than written English. This stretches the boundaries of the possible but also creates tension where the mainstream world cannot react on a hacker timescale and with hacker ethics.

More generally many scientists and information professionals are increasingly frustrated with the conventional means of disseminating science. Most conventional publishers regard scientific articles as “their content” and a very recent article (2011-06-20) from the STM publishers<sup>2</sup> indicates that the publishers believe they have the right to determine how content is, or more often is not, used. As an example most forbid by default indexing, textmining, repurposing, even of factual data to which the scientist has a legitimate subscription. This has an entirely negative effect on information-driven science, preventing even the development of the technology.

Generally, therefore, there is a culture of bottom-up change (“web democracy”) which looks to the modern web and examples of empowerment. (There are also examples of disempowerment such as attacks on Net-neutrality, walled

---

<sup>1</sup> Wikipedia: Hacking (innovation)

<sup>2</sup> Journal Article Mining: A Research study into Practices, Policies, Plans... and Promises

gardens, information monopolies, vendor lock-in, *etc.* and this contrast activates many in the modern informatics world). There are several articles, therefore, whose main theme is the access to Open information.

## 47.3 Openness and the choice of BMC as publisher

I have been critical of many publishers for their stance on closed information, and resolved that the issue reporting the symposium had to be completely Open. This is difficult in chemistry where there are almost no “Open Access” journals (those where by default all articles are Open (“Gold”)). The “Green” approach, where articles may be posted free-as-in-beer but not free-as-in-speech (*e.g.* CC-BY), is useless in science as it is impossible to discover and harvest green articles. Hybrid journals (where articles may be made Open by publication charges) are also of little value as the rights to the contents are usually poorly labelled and a machine cannot discover all “Open chemistry articles”.

While writing this overview and several articles I have become even more convinced that the only way of creating full semantic science is to publish Openly (CC-BY) and to publish completely (*i.e.* all experimental information (CC0/PDDL)). I believe that most funders now recognise this and are pushing, as hard as they can, to create fully Openly published science. I think this has to come, the question is how long it takes and in what form.

I now believe that in many cases it is unethical to restrict access to publicly funded science. Lessig, in his CERN talk (“Scientific Knowledge Should Not Be Reserved For Academic Elite”<sup>3</sup>), showed that it would cost 500 USD for him to read the top 10 papers relating to his child’s condition. These papers are effectively only available to academics in rich universities. A colleague recently told me he had spent a month researching the literature of his child’s condition (to critically effective purpose) and we agreed he could only do this because he was a professor at a University. That is one reason I support the Open Knowledge Foundation and its projects to define and obtain Open information (of which Open Bibliography<sup>4</sup> in this issue is typical).

As part of this effort four of us (including authors in this issue) developed the Panton Principles for Open Scientific Data. These principles are simple and, we hope, self-evidently worth pursuing and would lead to a greatly increased substrate for the Scientific Semantic Web. We were therefore delighted when BioMed Central not only enthusiastically adopted the idea but took positive steps to implement this as part of their publication process, for example by labelling data items with the OKF’s “OPEN DATA” logo. This is valuable not only in making the data repurposable, but also by promoting the concept - many readers will now be familiar with the logo. BMC have also encouraged authors (and editors) to highlight outstanding examples of data publication (and done me the honour of asking me to present their awards).

It is therefore a real pleasure to work with a publisher who understands my, and my co-authors’, intentions and is prepared to work to make them happen. The article explores many new types of publications and BMC have undertaken, as far as technically possible, to implement them as examples of a new generation of publication technologies. I and others have been critical of PDF as a publication format - it destroys semantics and innovation, but we must “eat our own dogfood”<sup>5</sup> and this is shown by several articles. Henry Rzepa creates all his molecules as semantic objects, while in Open Bibliography<sup>4</sup> we use our newly developed BibJSON and ScholarlyHTML to create and publish the article.

I am confident that because of the Openness, the readership of these articles will be much larger than if they were published in a closed access manner, however apparent the prestige of the closed access publisher. It is easy for a mature scientist, such as myself, to publish in an Open Access journal as it is unlikely to affect my career. I’d like to pay credit to all young people who have decided to publish in OA journals despite the possible current (irrational) view that this is detrimental to how they are regarded. I believe that their faith will be justified and that in a very short time the work published here will have higher visibility, and possibly regard, than if it had been published in an apparently more prestigious, closed access journal.

---

<sup>3</sup> Intellectual Property Watch: Lessig At CERN: Scientific Knowledge Should Not Be Reserved For Academic Elite

<sup>4</sup> Open Bibliography for Science, Technology, and Medicine

<sup>5</sup> Wikipedia: Eating your own dog food

## 47.4 Open Data

Five years ago the term “Open Data” was unknown (I started a Wikipedia page<sup>6</sup> to collect instances of usage). Now it is ubiquitous. Most of the public funders (Research Councils UK, Wellcome Trust, NIH, NSF and other national bodies) are now requiring that researchers make their data Openly available.

The first challenge is cultural; researchers have to be persuaded that Open Data is not only inevitable but also beneficial to their activities. Even when an author is convinced of the value of publishing Open Data, it is usually not trivial to do so. Unlike a manuscript where a static, human-readable, webpage can be posted and served for all time, data are frequently much more complex. They may be very large (petabytes), complex in both semantics and organisation, and even distributed over several sites. In bioscience, it is becoming commoner to see data published as Excel and other spreadsheets but in chemistry (apart from crystallography) the tradition is still to publish supplemental data as PDF, which destroys much of its semantics. One simple and achievable goal of these publications is to convince chemists that publishing in semantic form is “almost” no effort, compared to the effort of producing the data in the first place. If we were able to persuade researchers in computational chemistry simply to deposit their logfiles (usually less than 5 MB), or the Word documents for their syntheses, machines would be able to revolutionise the practice and understanding of computational and experimental chemistry. Open Access (CC-BY) implies (but may not explicitly state) that articles can be repurposed by machine extraction of data items, *e.g.* by OSCAR.

We have also addressed the question of what is Open Data and how do we identify it, both to humans and to machines. For many chemists, this may be the first time that they have had to consider this problem, but it is becoming increasingly required in many fields and for that reason, we have in several papers, discussed the question of licenses and contracts.

## 47.5 The semantic vision

I was excited and entranced by chemical informatics in the mid-70s as a result of some of the ground-breaking work done between chemists and computer scientists. The visions of LHASA, CONGEN, DENDRAL and others opened up the prospect of a chemical world where machines were seen as valuable allies of humans. This vision was also held in the world of chess, and indeed many chemical informatics processes are similar to the operations required in ‘artificial intelligence’. Chess has succeeded. Machines can now beat any human on the planet. For whatever reasons, chemistry turned its back on AI and there have been few developments in the last three decades. A necessary condition is the Open availability of semantic data, and if this comes about then there will be a major discontinuity in the way we practice chemistry.

In 1994, Henry Rzepa and I attended the first WWW conference in CERN. It was a remarkable occasion where a number of very early adopters showed what was possible with web technology and gave a vision of how this would change the way that science was not only reported but also done. There was a feeling that we were entering a new frontier where anything was possible and where new rules would evolve to fit the vision of cyberspace. The final session, where Tim Berners-Lee showed how semantic operations altered the real world was one of the seminal events of my last 20 years.

## 47.6 Semantic reality

Not surprisingly, semantic progress has turned out very differently from our original visions. We have stuck to our view that science must adopt semantic technologies including both the formal description of objects and the links between them. Chemistry has been very slow to adopt this, but other subjects have been much more adventurous and in bio- and geo-sciences it is routine to create objects which are derived from, and linked to, other objects.

<sup>6</sup> Wikipedia: Open science data

Many of the problems are cultural and for that reason several of the papers in this issue address the need to change attitudes as much as the technical requirements for the electronic infrastructure. I believe that it is impossible to do modern science unless the key information is completely Open. This applies, for example, to identifier systems, bibliographic data and much factual data. Chemistry, unfortunately in my opinion, has a strong ingrained culture of possession and sale/licensing of data. For this reason, it is often behind other subjects and, in the recent SOAP report<sup>7</sup> chemistry was highlighted as several years behind bioscience in its approach to Openness.

For that reason, some of the things we report are prototypes rather than completely established semantic resources. The biosciences have convinced funders that it is valuable to have completely Open access to sequences, structures, ontologies, *etc*. In chemistry, most of the freely accessible material has been produced by enthusiasts rather than large funded organisations. Indeed, it is the availability of bioscience resources such as ChEBI which to some extent drive the adoption of Open chemical semantics.

It is also an opportunity for our group to summarise formally several of the projects that they have been working on for several years. It is a feature of information projects that there is often no clear point at which a formal publication is immediately relevant and indeed this highlights the disconnect between publishing necessary information and publishing to acquire a community seal of approval ('a publication').

## 47.7 Chemistry as a community

Many disciplines have a close sense of community (I highlight crystallography which has a real sense of communal practice and goals). Many of the ideas in these articles have been inspired by crystallographic practice, its outstanding scientists, and its International Union - probably the leader in driving semantic approaches.

Scientific communities are now common on the web (and even have commercial value) and several of the articles emphasise the role of *ad hoc* and other communities. The web has the great advantage that anyone can, relatively easily, find those people and organizations who share values and goals, amplifying minority or early-adopter initiatives. Their dynamics are unpredictable and most die, but enough survive to provide world-changing mechanisms.

There is no clear community focus for chemistry overall (though sub sections - such as WATOC (World Association of Theoretical Organic Chemists) may provide one). The main drivers (funding, advancement, commerce) have always been present but the modern era has amplified and often dehumanised them. With growing emphasis on publication to generate the income of learned societies there is a decreasing sense that they act as nuclei for community to grow communal goods.

Because of this, chemistry has almost no public ontologies, and we have a vicious circle. Without ontologies, authors cannot reasonably be expected to create semantic information, and without a clear need for semantic information, the community will not take on the considerable load of creating ontologies. Several of the articles argue that the creation of lightweight dictionaries and other semantic metadata is affordable by the community and I believe that if the communal will is present, then it would be possible through bodies such as IUPAC and others, to create a full semantic infrastructure for much of the current published chemistry.

The current legal and contractual restrictions on re-using chemical data are seriously holding chemistry behind other subjects. These articles in this issue are not the place for polemics but we hope that traditional creators of information resources in chemistry will now think carefully about the value of making their data fully Openly available. This will be a considerable act of faith, because it will need a change in business model. Some of those providers have been traditionally held in high esteem by the community and if they use that esteem they have the opportunity to change the practice of chemical informatics.

---

<sup>7</sup> Study of Open Access Publishing: Report from the SOAP Symposium

## 47.8 The value of informatics

A major feature underlying all of the papers is to give an insight into the process of creating an information ecology. Some of them represent scientific discoveries (*e.g.* Rzepa) but most are concerned with building a coherent infrastructure usable by the community. It may be useful to liken this infrastructure to the development of instrumentation in many branches of science. Science depended on the microscope, the telescope, the spectrograph, the Geiger counter and many other types of instrumentation. There is sometimes a modern tendency to discount instrumentation and infrastructure as not being ‘proper science’. We hope that this issue will redress that balance.

As an analogy, Mendeleev required access to other scientists’ work to produce his classification, as did Pauling, Woodward and Hoffmann. I believe that the current chemical and related literature contains considerable amounts of undiscovered science, and that with ‘information telescopes’ we can start to discover this.

The development of infrastructure is a lengthy process. The web has, perhaps, given us an optimistic idea of the speed at which new ways of working can be implemented. We are still often governed by Planck’s observation (“Science progresses one funeral at a time”) and this is equally true for some areas of informatics. Several of the articles reflect the difficulty of catalysing change in what is essentially a mature and therefore conservative discipline.

Henry Rzepa and I were active contributors to the development of XML by running the XML-DEV mailing list (1997). This was a highly successful Open example of true collaboration and for me it culminated in the development of the SAX protocol late that year. XML had been seen as a primarily document- plus typesetting-oriented discipline, but some of us realised its potential for data modelling and transfer, and therefore the need for APIs in XML tools. I nagged continually at the community, and, as a result, Tim Bray, David Megginson and others helped us to develop the SAX protocol, now implemented in every computer on the planet. This protocol was developed in a calendar month and has stood the test of time exceedingly well.

This, perhaps, gave Henry, myself and other early adopters a false vision of how rapidly we would be able to take these new ideas to chemistry. Over the decade 2000-2010, we have developed and published specifications and software which we believe represent a formal but implementable infrastructure for chemical informatics. The uptake of these has been slow, but unlike some new technologies has not gone through the hype and depression syndrome (Gartner curve). In fact, this timescale is not so unusual. HTML itself has been through nearly 20 years of deployment and only now, with HTML5, does it appear that the community is starting to work together rather than fracturing for organisational and personal advantage. Similarly, semantic MathML is taking many years to become established. It is not that these systems, including CML, have been supplanted by ‘better’ ways of doing things, but more that the community as a whole is yet to be enlightened about the value of semantics.

## 47.9 Publishing

Scientific publishing should be a key part of the semantic revolution, but it has so far completely failed to address the vision. This is ironic in that HTML, which catalysed the web, was developed as a way for scientists at CERN to share information, but we have currently regressed to a completely non-semantic (PDF) manner of communication. This has replicated the traditional paper format so well that the only discernable value is to transfer the printing bill from the publishers to the readers. Not only has this held back our imagination, but has actually moulded the new, and I think somewhat unfortunate, values in the publication process. In many cases, authors now publish primarily to attain numerical estimates of worth above communication, validating experiments and other fundamental aspects of the process.

The web can, and, we hope, will, change this. Where you publish should not matter so long as the material is discoverable and the process of reviewing is understood. I believe that the papers in this issue will be read well beyond the cheminformatics community, because their value will be discerned and communicated by methods supplementary to the formal publishing process.

A major challenge in this issue is that the timescales for many of the projects is complex. In many lab experiments (such as chemical synthesis or chemical crystallography) the process is clearly bounded. “make this compound”,

“check success through crystal structure analysis”. Each (normally) has a clear endpoint and can be published as a static document.

In contrast how should we publish software? We use public repositories and these contain a complete record and the current semantic object. If we wish to tell the world about a development we put it on the mailing list. There is no need for a formal publication for those aspects. The motivation is therefore primarily to establish our reputation and there is no simple way to decide when this should be done. JUMBO has had six revisions - should this result in six papers or one or none? (Actually the only JUMBO paper is in 1997<sup>8</sup>). Six papers would confuse - but after 14 active years it’s time for another, I think, which explains the design process. OSCAR3 has its citable publication - a few years back - and we feel it’s useful to publish our current ideas, which have more to do with software engineering than new chemical entity recognition.

Or data? Crystaleye was a spinoff from Nick Day’s thesis - it wasn’t planned as a separate project - but simply a knowledgebase to use for his calculations. It does not have a formal publication other than an archive of a presentation<sup>9</sup>. The system has been running 5 years without serious mishap but the lack of a formal publication makes it difficult to write papers which refer to it. So we shall do this - after the fact. But if we had a semantic publication process it would be “published” by now.

## 47.10 The need to change publication processes

Historically the scientific community has required the following from the publication process:

- Establishment or priority and authorship
- Exposure and preservation of the scientific record
- Communicating the science to one’s peers and the wider world
- Allowing the science to be moderated by peers and others (“reviewing”).

There is perhaps an additional axis in today’s bibliometric-obsessed world: allowing the work to receive an official assessment of merit.

However the publication process is out of sync with the modern web-based world (“Web 2.0”) which allows the publication process to encourage and support:

**|nonascii\_5| Collaborative working** (as seen in many projects such as Wikipedia, Open StreetMap, and in science, Galaxy Zoo). Here each contribution is often an atom in a much larger cloud and the publication process is continuous rather than discrete. Wikipedia articles are “never finished” though there are some efforts to provide frozen versions. This is a strong theme of this “issue”.

**|nonascii\_6| Independence of the source of publication.** Given the ability of search engines, and the social networks, to discover anything of value it matters less *where* something is published. Other than the choice of reviewers the primary issues is whether a piece of information is accessible or limited. History has shown that high quality scholarship on the web will usually surface regardless of where it is published.

**|nonascii\_7| Creation of continuous semantic objects.** By recording everything we do, annotating it, and revising it, we can maintain a current semantic publication object at all times, including a revisitable history. This should be the object of scientific publication, not the current PDF.

**|nonascii\_8| The paper (semantic object) as a driver of research.** The idea of writing a paper before the research is carried out is valuable and not novel (*e.g.* George M. Whitesides<sup>1011</sup>). Here, however, we extend the paper to semantic objects (programs, spreadsheets, molecules, bibliography, *etc.*).

---

<sup>8</sup> JUMBO: An Object-based XML Browser

<sup>9</sup> CrystalEye - From Desktop to Data Repository

<sup>10</sup> Wikipedia: George M. Whitesides

<sup>11</sup> Whitesides’ Group: Writing a Paper

Several of the papers in the article have adopted these later ideas. This has been most obvious in Open Bibliography<sup>4</sup> where effectively the whole concept and technology has been driven during the 6 weeks of “writing the paper”. We started with a blank page and four people (William Waites, Mark MacGillivray, Ben O’Steen, Peter Murray-Rust) and during the writing process brought in new authors (Jim Pitman, Peter Sefton, Richard Jones) and communally created the design, technology and “paper”. The introduction of Scholarly HTML made this paper self-referential. The Quixote paper<sup>12</sup> has also dramatically driven the design of Quixote, particularly the social aspects.

## 47.11 The content of the issue

Several of the articles (CML<sup>13</sup>, OSCAR<sup>14</sup>, OPSIN, dictionaries<sup>15</sup>, WWMM<sup>16</sup>) in this issue cover a decade of work. We hope this will be useful to scientists and scholars who wish to implement new ideas and to give them some idea of what works, and what, more commonly, does not work. Sometimes only the passage of time and persistence achieves some level of success. Again, the short-termism of many infrastructural projects militates against developing a good platform for the future.

The long timescales highlight the difficulty of conventional publication. The world knows of these projects through blogs, online resources, user communities and so on, and a conventional learned paper has little value in communicating or preserving. Its prime merit is to achieve a traditional numeric merit for the work, often delayed by several years through the citation mechanism. I believe that it is important to change the values that we use in our assessment of on-going scientific endeavours, and avoid ritual publication.

Some of the articles (Wilbanks<sup>17</sup>, Neylon<sup>18</sup>) discuss the philosophy and practice of new models of scientific endeavour and communications. Some of the articles have a retrospective look (CML<sup>13</sup>, Zaharevitz<sup>19</sup>) but the fundamental principles are still as important today as when the work was started. A number represent growing points whose development is highly unpredictable. These include the WWMM<sup>16</sup>, where the vision of a distributed peer-to-peer knowledge resource has had to wait a decade until it could be implemented. The Quixote project is only months old but takes this vision and has already built an impressive prototype, which I expect to set the model for computationally-based knowledge repositories. These projects rely heavily on community, and this is most clearly shown in the Blue Obelisk movement<sup>20</sup> which aims to, and has largely succeeded in, creating an Open infrastructure for cheminformatics. A major motivation for this has been not just that software and data should be universally available but also that this is the only manner in which science can be reputably validated both by humans and machines. An example of the need for such validation is shown in Henry Rzepa’s article<sup>21</sup>.

The OpenBibliography project represents a socio-political imperative whose time has come, and for which the technology is appropriate. A year ago the JISC-funded OpenBibliography project could not point to a significant amount of open resources, but in the last year we have helped to catalyse the release of both library data (BL, CUL and several others), and also of scientific bibliography. It is impossible to find Open resources for scientific bibliography but we believe that in a year’s time, readers can look back and see this as a key starting point. It is worth noting that the very process of writing this article has generated a great deal of new formalism and tools in Open bibliography, and effectively given major impetus to the BibJSON approach.

Other articles (OSCAR, Open patents<sup>22</sup>, dictionaries, CML and CMLLite<sup>23</sup>) describe the design and implementation of information systems. In general, there is little funding for developing scientific software, though we have been fortunate to receive some from eScience and from JISC. We have taken this responsibility very seriously and our

<sup>12</sup> The Quixote project: Collaborative and Open Quantum Chemistry data management in the Internet age

<sup>13</sup> CML: Evolution and Design

<sup>14</sup> OSCAR4: a flexible architecture for chemical text-mining

<sup>15</sup> The semantics of Chemical Markup Language (CML): dictionaries and conventions

<sup>16</sup> The semantic architecture of the World-Wide Molecular Matrix (WWMM)

<sup>17</sup> Openness as Infrastructure

<sup>18</sup> Three stories about the conduct of science: Past, future, and present

<sup>19</sup> Adventures in Public Data

<sup>20</sup> Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on

<sup>21</sup> The past, present and future of Scientific discourse

<sup>22</sup> Mining chemical information from Open patent

<sup>23</sup> CMLLite: a design philosophy for CML

group has installed many of the cutting-edge ideas and tools for building high-quality systems. Members of the group collaborate and use common servers for their work (as far as possible on Open sites). Software libraries are used and re-used between group members, and we have developed a culture of communal ownership and responsibility. By using the continuous integration system (Jenkins), a failure in one library can immediately be highlighted and corrected before it impacts on other projects. Where funding is available, and where the culture allows it, we would very strongly recommend these practices in other groups. Again, many of these systems have taken over a decade to evolve from initial concepts to mature libraries, but we believe that almost all the systems reported in this article have been heavily re-factored and, within the academic environment, represent an attainable level of quality.

## 47.12 The future

Several articles are growing points, perhaps none more than AMI<sup>24</sup> where we explore the human-cyber interface in a laboratory, a “memex” which may ultimately replace some (but hopefully not all) of the role of the chemistry laboratory. In the same way Quixote represents a memex for computational chemistry. There is no clear pathway for AMI (and I predict that this will be largely influenced by what happens in the domestic arena).

The relative stagnation of chemical informatics suggests that change is unlikely to happen from within chemistry. As progress occurs in other areas (retail, bioscience *etc.*) chemistry may be dragged into the semantic world regardless. If chemists wish to retain control over their own systems they will be wise to start investing in Open semantic environments, because otherwise the rest of the world will do it for them.

How can chemical informatics survive and prosper? I think the most likely model will be Open publishing, not just of texts but data and other resources, mandated and paid for by funders. Those publishers which are able to adopt an Open model rather than continuing to maintain their own walled gardens, will ultimately triumph, and probably more rapidly than we expect.

---

<sup>24</sup> Ami - The Chemist's Amanuensis

# MOLECULAR DYNAMICS SIMULATIONS AND IN SILICO PEPTIDE LIGAND SCREENING OF THE ELK-1 ETS DOMAIN

## 48.1 Abstract

### 48.1.1 Background

The Elk-1 transcription factor is a member of a group of proteins called ternary complex factors, which serve as a paradigm for gene regulation in response to extracellular signals. Its deregulation has been linked to multiple human diseases including the development of tumours. The work herein aims to inform the design of potential peptidomimetic compounds that can inhibit the formation of the Elk-1 dimer, which is key to Elk-1 stability. We have conducted molecular dynamics simulations of the Elk-1 ETS domain followed by virtual screening.

### 48.1.2 Results

We show the ETS dimerisation site undergoes conformational reorganisation at the  $\text{I}|\text{nonascii\_1}|*\text{I}\beta^*\text{1}$  loop. Through exhaustive screening of di- and tri-peptide libraries against a collection of ETS domain conformations representing the dynamics of the loop, we identified a series of potential binders for the Elk-1 dimer interface. The di-peptides showed no particular preference toward the binding site; however, the tri-peptides made specific interactions with residues: Glu17, Gln18 and Arg49 that are pivotal to the dimer interface.

### 48.1.3 Conclusions

We have shown molecular dynamics simulations can be combined with virtual peptide screening to obtain an exhaustive docking protocol that incorporates dynamic fluctuations in a receptor. Based on our findings, we suggest experimental binding studies to be performed on the 12 SILE ranked tri-peptides as possible compounds for the design of inhibitors of Elk-1 dimerisation. It would also be reasonable to consider the score-ranked tri-peptides as a comparative test to establish whether peptide size is a determinant factor of binding to the ETS domain.

## 48.2 Background

Regulation of gene expression is essential for the development of all living organisms through processes such as cell proliferation, differentiation and morphogenesis. Key to these processes are mitogen activated protein kinases (MAPK), which target nuclear transcription factors, in response to extracellular signals, to elicit the required genetic response. One such transcription factor is Elk-1. Elk-1 (Ets-like protein 1) is a member of a group of proteins called ternary complex factors (TCF), which are targeted by MAPKs for phosphorylation<sup>123</sup> to regulate the transcription of immediate early genes (IEG)<sup>45</sup>. This event involves the formation of a ternary complex, induced by the cooperative binding of TCFs with serum response factor (SRF) dimers<sup>6</sup> on serum response elements found in IEG promoters<sup>789</sup>. TCFs are a subfamily of ETS<sup>10</sup> that binds to a 10-bp ETS binding site containing a 5'-GGA-3' core sequence. Since ETS domains are highly conserved across ETS proteins, ETS binding sites are differentiated by the cooperation of other transcription factors<sup>7112</sup> combined with base-specific interaction with variable bases flanking the central core sequence. Whilst TCFs naturally form a complex with SRF, they are also able to bind to DNA containing high-affinity, autonomous ETS binding motifs independent of a SRF<sup>613</sup>. ETS domain proteins are involved in cellular development, growth and differentiation<sup>141516</sup>. Their deregulation has been linked to multiple human diseases<sup>17</sup>.

The current X-ray crystal structure of the Elk-1 ETS domain is that of a dimer, with each unit bound to an autonomous 13-bp DNA double helix (PDB code<sup>18</sup> composed of a high affinity ETS binding site motif. Like other ETS domain proteins, the structure reveals three *l<sub>nonascii</sub>\_3*\*-helices packed against four anti-parallel \*l<sub>nonascii</sub>\_4\*-\*strands, giving an \*l<sub>nonascii</sub>\_5|l<sub>nonascii</sub>\_6|l<sub>nonascii</sub>\_7|l<sub>nonascii</sub>\_8|l<sub>nonascii</sub>\_9|l<sub>nonascii</sub>\_10|l<sub>nonascii</sub>\_11\* secondary structure (Figure :ref: '1<figure\_1>'). The \*l<sub>nonascii</sub>\_12\*3 helix forms the recognition helix, which slots into the major groove of the DNA target with a GGA core (Figure :ref: '2a<figure\_2a>'). The dimer interface involves the carboxy-end of \*l<sub>nonascii</sub>\_13\*1 and the \*l<sub>nonascii</sub>\_14\*1β1 loop (Figure :ref: '2b<figure\_2b>'). Contrary to the aforementioned structure, unequivocal experimental evidence has indicated that ETS dimers exist only in solution, [#B19]\_[#B20] whilst monomers occur predominantly in the nucleus, where they target DNA [#B21]\_[#B22]. To date, the structure of an unbound ETS domain is yet to be reported. However, Saven *et al.*<sup>19</sup> performed molecular dynamics (MD) simulations of a single Elk-1 ETS domain taken from the dimeric structure. They discerned regions within the simulated monomeric structure which showed large structural deviation with respect to the structure of the domain in the dimeric conformation. These regions include residues at the l<sub>nonascii</sub>\_16\*1β1 loop involved in the ETS dimer interface and residues at the \*l<sub>nonascii</sub>\_18\*2α\*3 loop involved in protein-DNA contacts.

Thus far, work on characterising the mechanism for protein-DNA recognition in TCFs has been abundant<sup>182321222324</sup>. However, there has been little on understanding the basis of Elk-1 dimerisation for transcriptional activity. Shaw and colleagues<sup>25</sup> have identified a region of the Elk-1 ETS domain encompassing the l<sub>nonascii</sub>\_26\*1β\*1 loop which

<sup>1</sup> ERK phosphorylation potentiates Elk-1-mediated ternary complex formation and transactivation

<sup>2</sup> Activation of ternary complex factor Elk-1 by MAP kinases

<sup>3</sup> Activation of the Sap-1a transcription factor by the c-Jun N-terminal kinase (JNK) mitogen-activated protein kinase

<sup>4</sup> Ets ternary complex transcription factors

<sup>5</sup> Ternary complex factors: prime nuclear targets for mitogen-activated protein kinases

<sup>6</sup> Elk-1 protein domains required for direct and SRF-assisted DNA-binding

<sup>7</sup> Signalling pathways: Jack of all cascades

<sup>8</sup> The ability of a ternary complex to form over the serum response element correlates with serum inducibility of the human c-fos promoter

<sup>9</sup> Ternary complex factors: growth factor regulated transcriptional activators

<sup>10</sup> The winged-helix DNA-binding motif: Another helix-turn-helix takeoff

<sup>11</sup> When Ets transcription factors meet their partners

<sup>12</sup> junB promoter regulation: Ras mediated transactivation by c-Ets-1 and c-Ets-2

<sup>13</sup> ERK2/p42 MAP kinase stimulates both autonomous and SRF-dependent DNA binding by Elk-1

<sup>14</sup> Molecular biology of the Ets family of transcription factors

<sup>15</sup> The ETS-domain transcription factor family

<sup>16</sup> The Ets family of transcription factors

<sup>17</sup> Ets transcription factors and human disease

<sup>18</sup> Structure of the Elk-1-DNA complex reveals how DNA-distal residues affect ETS domain recognition of DNA

<sup>19</sup> Modulating the DNA affinity of Elk-1 with computationally selected mutations

<sup>20</sup> Characterization of the Elk-1 ETS DNA-binding domain

<sup>22</sup> DNA binding specificity studies of four ETS proteins support an indirect read-out mechanism of protein-DNA recognition

<sup>23</sup> Determinants of DNA-binding specificity of ETS-domain transcription factors

<sup>24</sup> Structures of SAP-1 bound to DNA targets from the E74 and c-fos promoters: Insights into DNA sequence discrimination by Ets proteins

<sup>25</sup> Dimer formation and conformational flexibility ensure cytoplasmic stability and nuclear accumulation of Elk-1

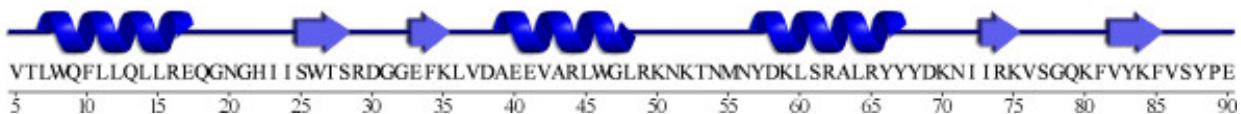


Figure 48.1: Figure 1. ETS domain secondary structure

**ETS domain secondary structure.** Amino acid sequence of the Elk-1 ETS domain, showing the locations of  $\alpha$ -helices and  $\beta$ -strands.

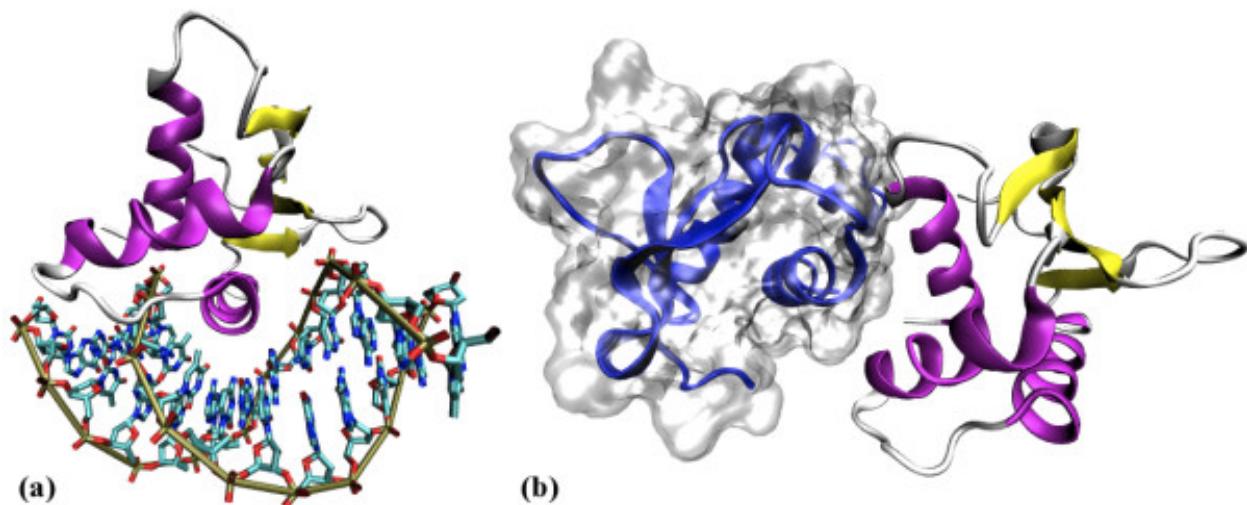


Figure 48.2: Figure 2. ETS domain DNA and dimer complexes

**ETS domain DNA and dimer complexes.** (a) An Elk-1 ETS domain bound to its DNA recognition sequence.  $\alpha$ -helices are in purple and  $\beta$ -strands in yellow. (b) An Elk-1 ETS domain dimer complex, showing the  $\beta$ -loops providing the interface. The images were generated using VMD<sup>20</sup> and PovRay (<http://www.povray.org/>), using coordinates taken from the 1DUX crystal structure<sup>18</sup>.

distinctly contributes to Elk-1 stability in the cytoplasm by directing Elk-1 dimer formation. Also, dimerisation in the cytoplasm appears to prevent rapid degradation and plays a role in translocation of the protein to the nucleus and its subsequent accumulation therein.

In the current work, we identify a series of peptides that can serve as leads for the design of potential peptidomimetic inhibitors of Elk-1 dimerisation. Using a docking-based approach, we screened entire libraries of all possible di- and tri-peptides against the Elk-1 ETS domain, targeting the stability region of the domain identified by Shaw *et al*<sup>21</sup>. Given the findings of Saven *et al.*,<sup>23</sup> it was essential to consider possible structural deviations or fluctuations in the *l<sub>nonascii</sub>\_281\*1β\*1* loop region that may affect binding of such inhibitors. Therefore, we performed MD simulations for an Elk-1 ETS monomer, to generate an ensemble of monomeric ETS conformations to use as docking targets. Herein, we show that tri-peptides appear to be good candidates for the design of inhibitors/binders of the Elk-1 dimer interface, based on size and binding specificity; di-peptides, on the other hand, appeared to behave as generic protein surface binders. We have also identified a set of tri-peptides, which may bind competitively to the ETS dimer interface.

## 48.3 Computational Methods

### 48.3.1 Molecular Dynamics Simulations

All stages of the MD simulations were carried out using CHARMM version 34b1<sup>2627</sup> with the all-atom CHARMM22 force field<sup>28</sup> and CMAP extensions<sup>293031</sup>. Our initial structure of a representative ETS domain monomer was chain C from the 1DUX crystal structure<sup>18</sup>. For residues with alternative positions, the pose with the highest occupancy was retained. Hydrogen atoms were assigned using the HBUILD module<sup>32</sup>. The system underwent three rounds of energy minimisation using the conjugated gradient method to remove any unphysical contacts until the system had converged. During the minimization all non-hydrogen atoms were harmonically restrained with a force constant of 30 kcal mol<sup>-1</sup> Å<sup>-1</sup>, which was reduced by 10 kcal mol<sup>-1</sup> Å<sup>-1</sup> at each successive round. The system was solvated in a cubic solvation box (62.2 Å × 62.2 Å × 62.2 Å), containing 7460 TIP3P water molecules,<sup>33</sup> using periodic boundary conditions. The fully solvated system was minimised using the conjugated gradient method. First, the protein was fixed to allow the water molecules to minimise and then harmonically restrained with a force constant of 30 kcal mol<sup>-1</sup> Å<sup>-1</sup>. A switched cut-off was used at an atom-pair distance of 10 Å for calculations of non-bonded interactions with a 2.0 Å switching region. The Particle Mesh Ewald algorithm was used for calculating long-range electrostatic interactions<sup>34</sup>. The system was gradually heated from 0 K to 300 K and allowed to equilibrate for 100 ps. The SHAKE algorithm<sup>35</sup> was applied to constrain all hydrogen-heavy atom bonds to remove the need to sample the high frequency vibrations. Simulations were performed with a 1 fs timestep with the Leapfrog integrator. Following equilibration, the simulation continued for a further 4 ns in the isobaric-isothermal (constant pressure and temperature, NPT) ensemble for the production run. During this phase, structural coordinates of the system were taken at 0.1 ps intervals to build a trajectory of the system dynamics. Time-dependent properties were calculated from the production trajectory. In preparation for this, the C<sub>l<sub>nonascii</sub>\_40</sub> atoms from each frame of the trajectory were aligned using least-squares fitting to the coordinates of the starting conformation. The root mean square deviation (RMSD) from the initial conformation and radius of gyration were calculated to survey any structural fluctuations over the time-series. To evaluate local structural deviations between the simulated ETS monomer conformations and the initial dimer conformation, a residue-specific RMSD of main-chain atoms (N, C<sub>l<sub>nonascii</sub>\_41</sub>, C, O) was calculated, averaged for the entire conformational ensemble. To complement this, we examined the changes in the backbone dihedral angles for structural fluctuations at residues around the *l<sub>nonascii</sub>\_421\*1β\*1* loop region (16-23).

<sup>26</sup> CHARMM: The biomolecular simulation program

<sup>27</sup> CHARMM: A program for macromolecular energy, minimization, and dynamics calculations

<sup>28</sup> All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins

<sup>29</sup> Importance of the CMAP correction to the CHARMM22 protein force field: Dynamics of hen lysozyme

<sup>30</sup> Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations

<sup>31</sup> Improved treatment of the protein backbone in empirical force fields

<sup>32</sup> Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison

<sup>33</sup> Comparison of simple potential functions for simulating liquid water

<sup>34</sup> A smooth particle mesh Ewald method

<sup>35</sup> Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes

Several snapshots were extracted from the trajectory to represent the various conformations for an Elk-1 ETS monomer. This was done by clustering the trajectory using backbone dihedral angles for residues 20-22 and selecting the conformation closest to the centre of each cluster as a representative conformation. The threshold defining the size of each cluster was the average of the standard deviation for the six chosen angles, over the time-series.

### 48.3.2 Automated Peptide Docking

Libraries of all possible di-and tri-peptide were built using all 20 standard, genetically encoded amino acids (400 di-peptides and 8,000 tri-peptides). The first step was to generate a SMILES string<sup>36</sup> from the raw peptide sequences, using ChemAxon's MolConverter program<sup>37</sup>. For each peptide, tautomers at physiological pH (7.4) were produced using ChemAxon's Calculator Plugins<sup>38</sup>. Any unreasonable peptide structures were removed from each library, including any structures with protonated carbonyl groups, de-protonated amines, structures without formally charged termini, and structures with anionic amides. Each peptide library was docked to the ETS monomer conformations obtained from the clustering. The dockings were carried out using OpenEye's docking program FRED,<sup>39</sup> a rigid docking algorithm, which requires a pre-computed conformer ensemble for screening the conformational space of the ligands. The conformer ensembles were created using Omega version 2.3.2 (OpenEye Scientific Software)<sup>40</sup>. A maximum of 500 low energy conformers were constructed for each peptide, *in vacuo*, using the MMFF94s force field<sup>41</sup><sup>42</sup>. The Coulombic and attractive part of the van der Waals terms were excluded from the force field, to reduce the effects of strong intermolecular interactions (e.g. hydrogen bonds) that can result in folded (peptide) conformations. Conformers with an energy difference greater than 25 kcal mol<sup>-1</sup> from the lowest energy conformer were rejected and conformers in the final ensemble were required to have a heavy atom RMSD greater than the duplicate removal threshold (0.4 Å). These settings were in line with the “high quality screening” settings of Kirchmair *et al*<sup>43</sup>. All remaining parameters were the default values.

The docking site for each receptor was delineated by a grid box encasing residues at the Elk-1 dimer interface site. A protein contact constraint, which all successful dockings were required to satisfy, was defined on Leu45, which is a key pharmacophoric contact for the dimer interface. The di- and tri-peptide libraries (with conformers) were separately docked, using FRED version 2.2.5, to each of the Elk-1 ETS domain conformations. Each multi-conformer peptide-ligand was exhaustively docked to a receptor using default step-sizes and the ChemGauss2 scoring function (a propriety function of OpenEye)<sup>41</sup>.

ChemGauss2 is a chemically aware shape-fitting scoring function, which uses Gaussian functions to describe the shape and chemistry of molecules. The best scoring poses for each compound were optimised in their docked state by half a rotation and translation step in each direction using the OEChemScore scoring function. OEChemScore is an OpenEye variant of the Chemscore<sup>44</sup> scoring function, but lacks a component for an entropy penalty upon complex formation.

On completion of the docking simulations, the single highest-scoring tautomeric state of each peptide was taken to give 400 unique di-peptides and 8,000 unique tri-peptides. Results from both libraries were analysed similarly but separately. The peptides were initially ranked by docking score, where rank 1 corresponded to the highest scoring peptide. Due to the variability in the size of peptides in both libraries, where size is a simple heavy atom count (HAC), and a systematic bias in the scoring functions (including OEChemScore),<sup>45</sup><sup>46</sup> we employed a simple size-independent metric to rank peptides and select the best binders. We used the size-independent ligand efficiency (SILE) metric<sup>47</sup>:

<sup>36</sup> SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules

<sup>37</sup> MolConverter was used for converting peptide sequences to SMILES strings

<sup>38</sup> Calculator Plugins were used for tautomer and protonation state calculations

<sup>39</sup> OpenEye Scientific Software Inc

<sup>40</sup> OpenEye Scientific Software Inc

<sup>41</sup> MMFF VI. MMFF94s option for energy minimization studies

<sup>42</sup> MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries

<sup>43</sup> Comparative performance assessment of the conformational model generators Omega and Catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations

<sup>44</sup> Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes

<sup>45</sup> Docking and scoring in virtual screening for drug discovery: methods and applications

<sup>46</sup> Protein-ligand docking: Current status and future challenges

<sup>47</sup> Simple size-independent measure of ligand efficiency

where *affinity* can be any binding measurement, in our case the docking score; *x* is derived by fitting the maximal ligand efficiency ( $LE_{max}$ ) values from all 12 docking screens against HAC, to a logarithmic function of the form:

Docking data from the di- and tri-peptide sets were fitted and examined separately.

Docked complexes between the highest-ranked peptides and the 12 protein conformations were analysed using HB-PLUS, using default parameters<sup>48</sup>. Only hydrogen bonds between protein and peptide ligands were considered. The number of ETS residues participating in interactions with the top SILE-ranked peptide in each complex was counted. This count was also dissected into the number of specific contacts made, where specificity is defined as interactions between peptide side-chains and ETS residues.

## 48.4 Results and Discussion

### 48.4.1 Analysis of Elk-1 dimer interface

In order to aid the identification of possible peptide binders for the Elk-1 dimer interface, it was important to identify structural features contributing the dimerisation. Interactions between two Elk-1 ETS domains were calculated using the LIGPLOT program<sup>49</sup>. The minimum and maximum interatomic bond distances for non-bonded contacts were 2.90 Å and 3.90 Å, respectively, and for hydrogen bonds: 2.70 Å and 3.35 Å. The LIGPLOT diagram for chains C and F from the X-ray crystal structure of the ETS dimer (Figure 3) reveals a homodimeric interaction between the two ETS domains. Key to the interface were residues 17, 18 and 49, where Gln18 and Arg49 of one domain donate three hydrogen bonds to Glu17 of the partnering domain. Accompanying these hydrogen bond interactions, several residues make large steric contributions to the interface; these are listed in Table 1 together with a percentage accessible surface area of the interface, calculated using NACCESS<sup>50</sup>. The schematic depicting the secondary structure of the ETS domain in Figure 1 shows the relative positions of these residues in the domain.

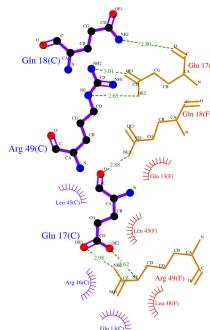


Figure 48.3: Figure 3. ETS domain dimer interface

**ETS domain dimer interface.** LIGPLOT representation of intermolecular interactions between two Elk-1 ETS domains according to the X-ray crystal structure (1DUX) of the dimer complex. Non-bonded interactions are indicated by spokes and hydrogen bonds by dashed green lines, with lengths given in Å. Residues from chain C are shown with purple bonds and chain F in orange.

### 48.4.2 MD simulations of an Elk-1 ETS domain

Over the course of the MD simulation, the radius of gyration (RoG) and the RMSD of the backbone atoms relative to the minimised (initial) structure of each frame in the trajectory remained stable. The mean values for the RMSD and

<sup>48</sup> Satisfying hydrogen bonding potential in proteins

<sup>49</sup> LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions

<sup>50</sup> NOTITLE!

the RoG were  $1.64 \pm 0.24 \text{ \AA}$  and  $12.17 \pm 0.08 \text{ \AA}$ , respectively. The latter was, in fact, identical to the RoG of the initial structure. This indicated that the overall shape and size (packing) of both the monomeric and dimeric conformation of the Elk-1 ETS domain is conserved. To focus on localised structural deviations, we calculated the time-averaged RMSD for each residue, with respect to the main-chain atoms of the initial conformation. This revealed substantial structural deviations for residues 20-22 compared to the dimeric conformation (Figure 4). These residues are situated at the centre of the *Inonascii\_54|\*1 $\beta$ 1 loop*, which was identified by Shaw \*et al.<sup>21</sup> as the region accountable for Elk-1 stability. We also measured the backbone dihedral angles for residues in the loop across the entire trajectory. Residues 16 to 19 and residue 23 showed dihedral angle fluctuations within range of typical thermal fluctuations for proteins, with an average standard deviation about the mean of  $\pm 19^\circ$  across the trajectory for the 10 angles; fluctuations of the backbone dihedrals for residues 21 and 22 were considerably larger, with the lowest standard deviation value of  $\pm 59^\circ$  and the highest of  $\pm 88^\circ$ . The high fluctuation of residues 21 and 22 are consistent with the high RMSD values seen in Figure 4.

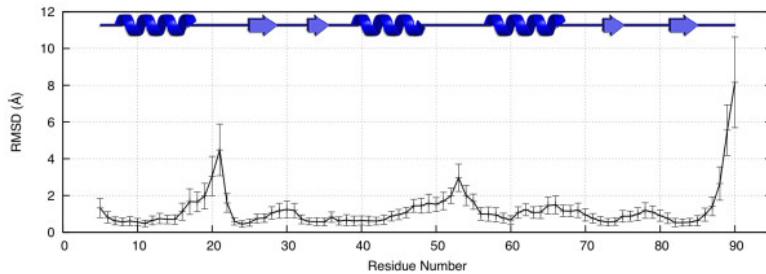


Figure 48.4: Figure 4. Residue specific ETS monomer fluctuations

**Residue specific ETS monomer fluctuations.** Time-averaged RMSD for the main-chain atoms of each residue over 4 ns of simulation of an Elk-1 ETS domain. The bars signify fluctuations about the mean and correspond to one standard deviation.

Since the structure fluctuates in the region coinciding with the *Inonascii\_62|\*1 $\beta$ \*1 loop*, which was our proposed docking binding site, it would have been unreasonable to dock to the single domain conformation taken from the crystal structure of the dimer, or to dock to the averaged structure of the MD trajectory. Instead, we clustered the trajectory, based on the backbone dihedral angles of residues 20-22, to extract several conformations representative of an Elk-1 ETS domain monomer. Using a clustering threshold of  $49.2^\circ$ , which was the average of the standard deviations of the six angles, 12 clusters were obtained. From each cluster, a single conformation was taken (Table 2) and used for the docking study. (see Additional file

Additional file 1

**Superposition of ETS target structures and derivation of maximal ligand efficiency.** Document contains: 1) figures showing alignment of the 12 ETS target structures with the minimised structure and 2) plots showing the maximal ligand efficiency values for the docked di- and tri-peptides.

[Click here for file](#)

### 48.4.3 Peptide Docking

#### Peptide screening

Libraries of all possible di- and tri-peptides, together with possible tautomers of each peptide were constructed. The final libraries (including protonation and tautomeric states) were made up of 1,128 di-peptides and 33,367 tri-peptides. The two peptide libraries were individually screened against the 12 monomer conformations of the Elk-1 ETS domain. Each multi-conformer peptide-ligand was exhaustively docked to the receptors, i.e., all rigid-body translations and rotations of a conformer were enumerated within the docking site, centred on residues for the Elk-1 dimer interface. Although with different affinities, all peptides bound to the docking site with favourable scores.

The docked peptide-ligands for each library were ranked according to docking score, retaining only the highest scoring tautomer of each peptide. Using this simple ranking scheme, particularly for the di-peptides, peptide-ligands with a larger heavy atom count (HAC) were ranked higher than those with a smaller one. Although this effect is apparent in experimental ligand binding data,<sup>51</sup> unfortunately it is unduly amplified in computational docking studies. The problem stems from the additive nature of scoring functions. The scoring function tends to favour larger ligands, as they contribute towards a greater number of intermolecular interactions with the target. This phenomenon is inherent to several docking scoring functions, including OEChemScore, which lack other terms in the function, such as a desolvation penalty term, that can counter-balance the favoured interaction terms. For our peptide ligands, the effect is seen in Figures 5a and 5b for the highest scoring di- and tri-peptides, respectively, taken at each HAC from the 12 docking screens.

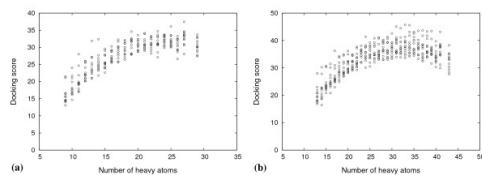


Figure 48.5: Figure 5. Highest scoring peptides by size

**Highest scoring peptides by size.** Highest scoring (a) di-peptides and (b) tri-peptides taken at each HAC from all 12 docking screens. Docking scores have been plotted as non-negative values for convenience.

Because the bias towards larger ligands is counter to the rules for drug bioavailability,<sup>52</sup> a simple metric, called ligand efficiency has been developed to assess the binding of a compound, with respect to the number of atoms, and its potential for lead optimisation<sup>53</sup><sup>54</sup>. Ligand efficiency (LE) is the binding affinity (potency) divided by a measure of the size of a ligand, often the HAC, as defined by Kuntz *et al*<sup>53</sup>. Compounds that can provide the desired binding affinity with fewer atoms are considered efficient. However, in large screening studies of ligands spanning a wide range of molecular sizes, ligand efficiency is non-linearly related to HAC, and appears to fall as size increases<sup>54</sup><sup>55</sup>. This trend can be illustrated by plotting the LE versus HAC (Figure 6), for the peptides used in Figure 5. The trend may be related to the increased complexity of larger compounds. More complex compounds can bind a target with a less than optimal geometry, due to binding constraints and structural compromises<sup>56</sup>. They also offer a smaller surface area per atom to make favourable interactions compared to smaller, less complex compounds<sup>56</sup>. LE over-corrects for the size dependence in docking scores. Therefore, a size-normalised efficiency scale was needed. We used the size-independent ligand efficiency (SILE)<sup>49</sup> scale to rank peptides in the docked libraries. In order to apply a SILE metric for our data, a value for  $x$ , for Equation (1), was obtained by fitting the maximal LE ( $LE_{max}$ ) values taken from Figure 6 to the function in Equation (2) (see Additional file  $max$  is the highest LE value at each HAC. The  $x$  values for di- and tri-peptides were 0.649 and 0.665, respectively, which were close to the generic value of 0.7 suggested by Nissink<sup>49</sup>.

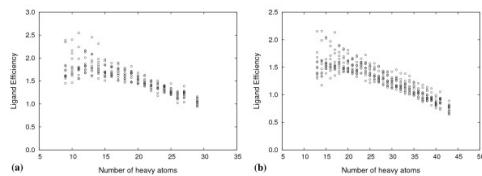


Figure 48.6: Figure 6. Peptide binding efficiencies by size

**Peptide binding efficiencies by size.** Highest ligand efficiency values for (a) di-peptides and (b) tri-peptides taken at each HAC from all 12 docking screens.

<sup>51</sup> The maximal affinity of ligands

<sup>52</sup> Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings

<sup>53</sup> Ligand efficiency: a useful metric for lead selection

<sup>54</sup> Ligand binding efficiency: Trends, physical basis, and implications

<sup>55</sup> The role of molecular size in ligand efficiency

<sup>56</sup> Molecular complexity and its impact on the probability of finding leads for drug discovery

By mapping the two sequence positions for ranked di-peptides on to a  $20 \times 20$  matrix, where each square is graded according to the associated rank, we can see the difference in size-dependence between score- and SILE-ranked results (Figures 7a and 7b). Score-ranked matrices clearly show peptides consisting of heavier amino acid residues such as tryptophan and tyrosine ranked higher, whilst those of smaller residues such as alanine and glycine ranked lower. The SILE-ranked matrices reduce this bias (Figure 7b). Similarly, plots of the distribution of LE and SILE values for the di- and tri-peptide dockings as a function of HAC reveal a reduced size-dependency for SILE values compared to LE values (compare Figure 8b with 8a and 8d with 8c). However, the SILE values for di-peptides maintain some size dependence (Figure 8b) compared to the tri-peptides (Figure 8d). It may be that the binding site readily accommodates the di-peptides, due to their smaller size and low structural complexity, and thus the size bias remains dominant. Therefore, di-peptides with a lower HAC bind and fit the binding site more completely, where a greater number of atoms participate equally in protein-peptide interactions compared to tri-peptides and di-peptides with a higher HAC. A similar result was observed in a peptide docking study to the Fv fragment of a monoclonal IgM cryoglobulin<sup>57</sup>. In that study, docking results were skewed towards di-peptides composed of larger residues. It was suggested that the di-peptides were too small to discriminate between different binding cavities, which is consistent to the hypothesis of ‘a small ball in a large hole’. Thus, the size-independent metric is less effective for compounds of lower complexity. This also suggests that di-peptides are fairly promiscuous protein surface binders and may not offer a specific binding preference for the dimer interface site had the docking site definition been larger.

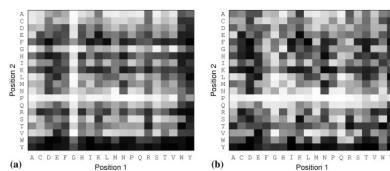


Figure 48.7: Figure 7. Docked di-peptide ranking maps

**Docked di-peptide ranking maps.** 2D-maps representing the peptide rank by (a) docking score and (b) SILE values according to the positions occupied by each residue for di-peptides docked to ETS conformation 9 (ETS9). The ranks are represented as squares shaded from black (highest rank) to white (lowest rank).

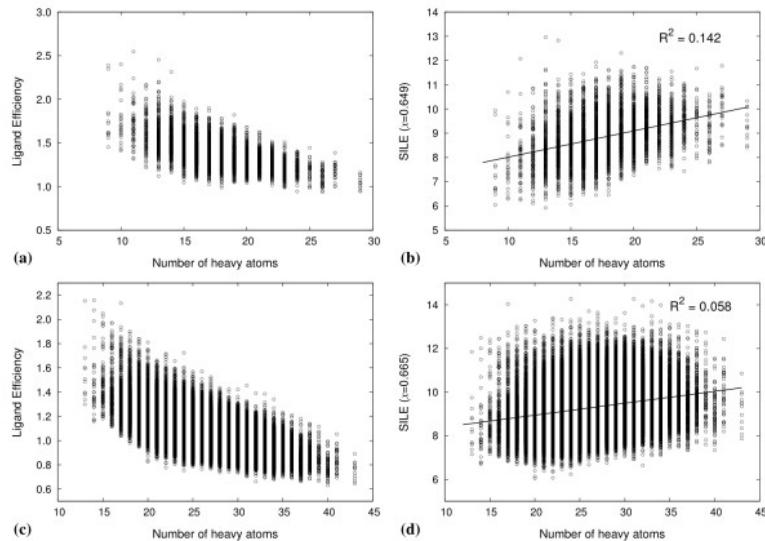


Figure 48.8: Figure 8. Peptide efficiency distribution

**Peptide efficiency distribution.** Distribution of LE ((a) and (c)) and SILE ((b) and (d)) values for docked di- and tri-peptides as a function of the number of heavy atoms. (a) LE and (b) SILE values for di-peptides; (c) LE and (d) SILE values for tri-peptides.

<sup>57</sup> Docking of combinatorial peptide libraries into a broadly cross-reactive human IgM

Tables 3 and 4 list the highest score- and SILE-ranked di- and tri-peptides, respectively. These tables again reveal the preference for peptides consisting of large aromatic residues for the score-ranked results, particularly for the di-peptides. In addition to the factors discussed above, it is possible such residues may behave as anchors to aid the binding of the complete peptides. However, given that a minority (38%) of residues in the dimer interface are non-polar, it is perhaps unlikely that these hydrophobic residues, especially tryptophan and tyrosine, would show particular affinity for the binding site in experimental assays.

### Structural analysis of docked complexes

Interactions between the top SILE-ranked peptide-protein complexes were calculated for each Elk-1 ETS domain conformation. Given the systematic bias in the score-ranked results, they were not considered for interaction analysis. Overall, both sets of peptides interact with residues in the dimer interface, namely the regions at sequence positions 10-20 and 40-50. To investigate the specificity of binding, the number of ETS domain residues hydrogen bonded with a peptide were counted for each of the docked complexes. On average, di-peptide ligands interacted with fewer ETS domain residues compared to tri-peptides (column 2, Tables 5 and 6), although some of these interactions did include those made to ETS domain residues Glu17, Gln18 and Arg49, which were identified as key hydrogen-bond contacts at the dimer interface (see Figure 3). In addition, the highest SILE-ranked tri-peptides make more specific contacts to the protein compared to the highest-ranked di-peptides (column 3, Tables 5 and 6). Here, we measure specificity as interactions between peptide side-chains and ETS residues. Figure 9 shows an “Interaction fingerprint” of the hydrogen bonds between the highest SILE-ranked peptide and the corresponding ETS conformations. The Elk-1 ETS domain dimer fingerprint is given at the top of the figure as a reference.

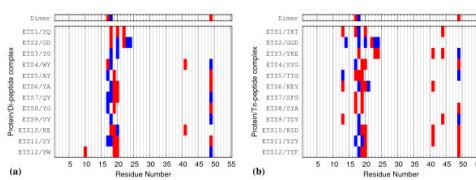


Figure 48.9: Figure 9. Peptide hydrogen bond fingerprints

**Peptide hydrogen bond fingerprints.** Hydrogen bond “Interaction fingerprints” for docked complexes between Elk-1 ETS domain conformations and highest SILE-ranked (a) di- and (b) tri-peptides. Specific contacts, as described in the main text, are given in red and peptide main-chain contacts in blue.

Perhaps naively we may have expected a peptide corresponding to a contiguous sequence of residues involved in the Elk-1 dimer interface would have been ranked high in the docking simulation, but this was not the case. The most obvious such peptides were, the tri-peptide Arg-Glu-Gln, which corresponds to residues 16-18 in the ETS domain, and the di-peptide Glu-Gln corresponding to residues 17-18 (as seen in Figure 3). The best SILE-ranked Glu-Gln di-peptide was ranked 52 out of 400 in complex with ETS7 and had an average ranking of 133 for all 12 docked complexes. Whilst the best SILE-ranked Arg-Glu-Gln tri-peptide was ranked 1423 out of 8000 in complex with ETS8 and an average ranking of 3729 (Table 7). This is largely because these peptides, although capable of providing some of the hydrogen bond interactions found at the dimer interface, are unable to mimic the interactions of other residues involved in dimer interface (see Table 1 and Figure 3), particularly van der Waals contacts. This has been recognised in other efforts to discover small molecules that disrupt protein-protein interactions<sup>58</sup>. For this reason, the binding of the two aforementioned peptides may be weaker than the higher ranked peptides, which satisfy more of the pharmacophoric constraints of the dimer.

## 48.5 Conclusions

It is well-established that TCFs, such as Elk-1, play a critical role in transcriptional activation in response to extracellular signals and a consequent role in the growth and development of cells. Using MD simulations we have identified

<sup>58</sup> Reaching for high-hanging fruit in drug discovery at protein-protein interfaces

possible conformations for an Elk-1 ETS domain monomer and observed a structural variation from the dimeric form at the  $\text{\textit{Inonascii\_66}}*\beta^*\text{1}$  loop, where two Elk-1 proteins dimerise. Against these monomeric conformations we screened all possible di- and tri-peptides and have identified several peptides with potential to mimic and possibly inhibit Elk-1 dimerisation. The size and binding specificity of the tri-peptides make them ideal candidates for the design of peptidomimetics of the Elk-1 dimer interface. The di-peptides, on the other hand, appear to be a generic set of protein surface binders and are unlikely to produce experimental binding affinity for the ETS dimer interface site that would correlate with the docking data. The notion of using tri-peptides as potential candidates for peptidomimetic design has also been supported in a recent review by Ung and Winkler<sup>59</sup>.

Since docking scoring functions are based on a number of simplifications and assumptions, their predictions for binding free energies for a protein-ligand complex are not quantitative. This also makes it very difficult to discriminate between strong/weak binders and non-binders, particularly for a relatively at and exposed binding site, as investigated here. Although this is a major limitation in a docking protocol, the exhaustive search algorithm of docking programs has been successful in predicting correct binding geometries of known hits<sup>48</sup>. As with all docking protocols, true validation can only be achieved through experimental binding measurements correlating with the docking results. For an experimental binding study, it would be reasonable to test the binding affinity of the top SILE-ranked tri-peptides listed in Table 4. The score-ranked tri-peptides may also be worth considering as a comparative test to establish whether size of the peptide is a determinant factor of binding to the ETS domain or if it is, indeed, just an artefact of docking. Binding data for the Arg-Glu-Gln peptide may also be useful in explaining the poor predicted binding by the docking simulations.

It is quite clear that complex formation of a protein and ligand is a dynamic mechanism. Here we have shown a combination of MD and docking simulations can be used to provide an understanding of the effects on ligand binding to a dynamic representation of the receptor, which a single configuration crystal structure would fail to reveal. Thus, computer simulations on protein-ligand complexes can enhance crystal structure data in this respect. We plan to extend the current work by performing all-atom MD simulations of selected peptides complexed with an Elk-1 ETS domain to assess the stability of the complexes, whilst incorporating any induced fitting of the peptides and obtain accurate binding data for use in designing future docking studies of optimised peptides. We also plan to apply free energy perturbation methods<sup>60</sup> to a set of the best peptides to calculate relative binding free energy of alchemical transformations of the peptides in complex with the Elk-1 ETS domain. This may also go as far as identifying tetra-peptides with potentially superior binding affinities compared to the tri-peptides we have considered here.

## 48.6 Competing interests

The authors declare that they have no competing interests.

## 48.7 Authors' contributions

PES and JDH together conceived the study. AH and JDH devised the strategy and analyses undertaken. AH performed the calculations, analysis and drafted the manuscript. JDH supervised the study and with PES participated in the discussion of the results. All authors have read and approved the final manuscript.

## 48.8 Acknowledgements

We thank OpenEye Scientific Software for provision of academic licences for FRED and Omega. We also thank ChemAxon for providing a licence for Marvin, which was used to build the peptide libraries.

<sup>59</sup> Tripeptide motifs in biology: Targets for peptidomimetic design

<sup>60</sup> Free energy calculations: Applications to chemical and biochemical phenomena



# 2D-QSAR FOR 450 TYPES OF AMINO ACID INDUCTION PEPTIDES WITH A NOVEL SUBSTRUCTURE PAIR DESCRIPTOR HAVING WIDER SCOPE

## 49.1 Abstract

### 49.1.1 Background

Quantitative structure-activity relationships (QSAR) analysis of peptides is helpful for designing various types of drugs such as kinase inhibitor or antigen. Capturing various properties of peptides is essential for analyzing two-dimensional QSAR. A descriptor of peptides is an important element for capturing properties. The atom pair holographic (APH) code is designed for the description of peptides and it represents peptides as the combination of thirty-six types of key atoms and their intermediate binding between two key atoms.

### 49.1.2 Results

The substructure pair descriptor (SPAD) represents peptides as the combination of forty-nine types of key substructures and the sequence of amino acid residues between two substructures. The size of the key substructures is larger and the length of the sequence is longer than traditional descriptors. Similarity searches on C5a inhibitor data set and kinase inhibitor data set showed that order of inhibitors become three times higher by representing peptides with SPAD, respectively. Comparing scope of each descriptor shows that SPAD captures different properties from APH.

### 49.1.3 Conclusion

QSAR/QSPR for peptides is helpful for designing various types of drugs such as kinase inhibitor and antigen. SPAD is a novel and powerful descriptor for various types of peptides. Accuracy of QSAR/QSPR becomes higher by describing peptides with SPAD.

## 49.2 Background

Research on the classification of small molecules using computers was popular in the 1990s<sup>12345</sup>, with similarity analysis of compounds being a major objective. At the time, there were mainly two methods for similarity analysis: the fingerprint description approach<sup>46</sup> and the inductive logic programming approach<sup>789</sup>. In the fingerprint description approach, a molecule is described as a sequence of bits, each of which corresponds to the existence of a chemical substructure. Atom-pair descriptor<sup>4</sup> or substructure type fingerprints are popular descriptors.

Research on the classification of peptides became popular in the year 2000<sup>101112</sup>. The hidden Markov model (HMM) approach<sup>12</sup> and physical data description of peptide approach<sup>11</sup> were the major approaches. The main subject of these papers is the natural twenty amino acids, such as isoleucine, valine, and so on. For example, the subject of immunity concerns peptides whose components are one of 20 natural amino acids. In traditional research for the classification of peptides, an amino acid residue was described as an alphabet or a set of physical or chemical values<sup>11</sup>.

However, in practical virtual screening, describing other amino acid inductions such as cyclohexyl alanine or F5 phenylalanine is necessary. The traditional description of peptides is not sufficiently powerful because the common characteristics among amino acid residues cannot be described sufficiently. For example, tyrosine and phenylalanine have an aromatic ring substructure in common. In the alphabetic description, tyrosine and phenylalanine are described as ‘Y’ and ‘F’ respectively. However, understanding that symbols ‘Y’ and ‘F’ have a common substructure on a machine learning algorithm is impossible. Research of two-dimensional QSAR has been undertaken for various types of peptides. In the atom-pair holographic code (APH)<sup>13</sup>, each peptide is described with the method similar to atom-pair descriptor<sup>3</sup>. Our novel descriptor, substructure-pair descriptor (SPAD), captures different characteristics of peptides from APH and has greater descriptive power than APH. The combination of APH and SPAD may lead to better QSAR for peptides with many types of amino acid inductions<sup>14</sup>.

Tanimoto coefficient<sup>15</sup> is a popular indicator for measuring similarity between two compounds<sup>16</sup>. In binary case, Tanimoto coefficient  $T^*(X, Y)$  between vectors  $X$  and  $Y$  is defined as following expression.

Tanimoto coefficient becomes large when two vectors have more similar bit-pattern. When the structure of two compounds is similar, Tanimoto coefficient is also high.

In machine learning, excessive features degrade the performance of machine learning algorithms due to over-fitting problems<sup>17</sup>. Under excessive feature space, predictive models lose robustness. Feature selection is necessary for building more accurate predictive models. Kohavia proposed the relevance of features instead of maximizing accuracy of an algorithm<sup>18</sup>. Discussions about relevance of features are popular in various types of algorithm<sup>19</sup>. Relevance is defined as the difference between probability density function  $P^*(Y = y)$  and conditional probability density function  $P^*(Y = y|X = x)$ . When  $P^*(Y = y|X = x) > P^*(Y = y)$ ,  $x$  is relevant. Otherwise,  $x$  is irrelevant.

<sup>1</sup> Compass: A shape-based machine learning tool for drug design

<sup>2</sup> Machine learning approaches for the prediction of signal peptides and other protein sorting signals

<sup>3</sup> Atom pairs as molecular features in structure-activity studies: definition and applications

<sup>4</sup> Chemical Similarity Using Geometric Atom Pair Descriptors

<sup>5</sup> Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors

<sup>6</sup> Applications of 2D Descriptors in Drug Design: A DRAGON Tale

<sup>7</sup> Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming

<sup>8</sup> Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming

<sup>9</sup> Pharmacophore Discovery Using the Inductive Logic Programming System PROGOL

<sup>10</sup> Predicting Protein-Peptide Binding Affinity by Learning Peptide-Peptide Distance Functions

<sup>11</sup> Prediction of MHC Class I Binding Peptides Using an Ensemble Learning Approach

<sup>12</sup> Prediction of MHC Class I Binding Peptides by a Query Learning Algorithm Based on Hidden Markov Models

<sup>13</sup> A novel atom-pair hologram (APH) and its application in peptide QSARs

<sup>14</sup> Design and Evaluation of Bonded Atom Pair Descriptors

<sup>15</sup> A Computer Program for Classifying Plants

<sup>16</sup> Similarity-based virtual screening using 2D fingerprints

<sup>17</sup> Information theory and an extension of the maximum likelihood principle

<sup>18</sup> Wrappers for feature subset selection

<sup>19</sup> Spectral feature selection for supervised and unsupervised learning

In information theory<sup>20</sup>, entropy is an indicator for measuring the amount of information. We denote probability of  $i$ -th substructure  $x$  as  $P^*(\text{:sub: } i|x)$ . Entropy  $*E$  is defined as next function.

## 49.3 Methods

### 49.3.1 Definition of several terms

In this paper, we define several terms as follows.

- Substructure: a part of structure of peptides
- Descriptor: The function for mapping a structure of amino acid residues or peptides to a bit according to substructure.
- Feature: A bit as the result of a descriptor.

A target protein binds some amino acid residues of peptides by some kinds of chemical or physical interactions. For example, hydrogen bonds and hydrophobic effect are representative interactions. In our QSAR approach, we describe the two-dimensional structure of peptides with a sequence of bits and analyze the relationship between peptides structure and its activity statistically. When we analyze this relationship with a data mining algorithm, QSAR rules are extracted automatically from dataset annotated with peptides' activity. From a chemical viewpoint, describing various types of amino acid inductions properly is important for improving QSAR analysis.

From a statistical viewpoint, features which maximize the accuracy of an algorithm for analyzing QSAR are the best. Kohavi proposed the relevance of features instead of maximizing accuracy of an algorithm. Discussions about relevance of features are popular in various types of algorithm<sup>19</sup>. Relevance is defined as the difference between probability density function  $P^*(\text{*Y} = y)$  and conditional probability density function  $P^*(\text{*Y} = y|\text{:sub: } i|X = \text{:sub: } i|x)$ . When  $|P^*(\text{*Y} = y|\text{:sub: } i|X = \text{:sub: } i|x) - P^*(\text{*Y} = y)| > \text{nonascii\_5}$ ,  $i$  is relevant. Otherwise,  $i$  is irrelevant.

We define each symbol as Figure 1. The SPAD is defined with these symbols.

### 49.3.2 Definition of the base substructure set for amino acid inductions

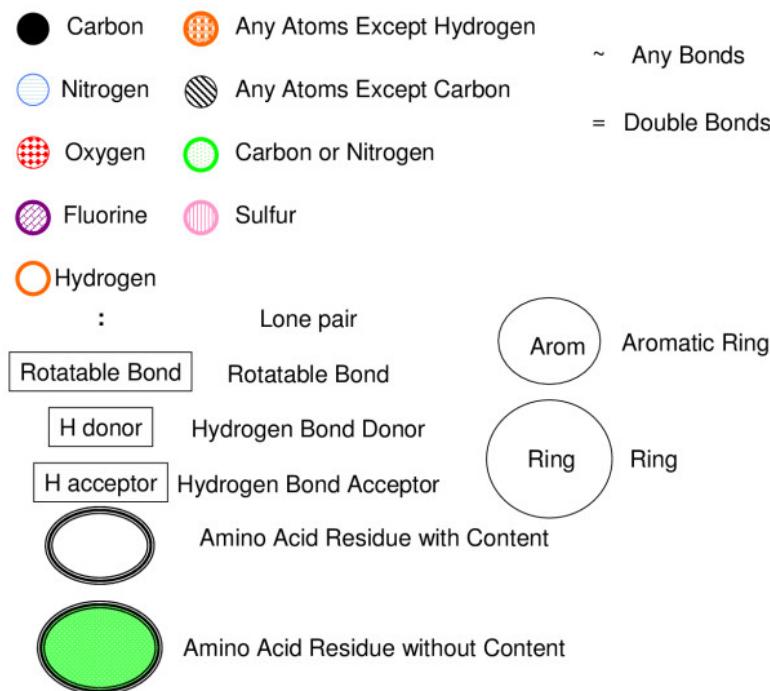
The aim of defining the base substructure (Figure 2) set is the description of important interactions between a target protein and a peptide such as hydrogen bonds, the hydrophobic effect, and so on. However statistically redundant or specific descriptor may degrade the accuracy of an algorithm for QSAR analysis. We defined the base substructure set under next three conditions.

- Describe potential factors for interactions such as hydrogen bond acceptor.
- Features of amino acid residues should be weak relevant to each other mathematically. This is the condition for avoiding strong relevant features. Abandon features with strong relevance.
- A feature should have high entropy (in information theory) after mapping structures of 450 types amino acids to a sequence of bits. This is the condition for avoiding too specific descriptor. Abandon descriptors with low entropy.

The first item is essential for QSAR analysis because key substructures such as hydrogen bond acceptor may cause the activity of peptide for target protein. Under the condition lack of description of them, most of algorithms analyzing QSAR become powerless. The second and third items are necessary for efficient analysis from a statistical viewpoint. The second item prohibits the redundancy of features. Even if the structures of two amino acid inductions are chemically different, two features may be relevant to each other. Then, these two features are redundant statistically. The third item is necessary for generating robust QSAR rules. Features with low entropy (in information theory) lose generality.

The set of substructures  $Z$  includes the forty-nine substructures shown in Figure 2. These substructures are roughly categorized into three parts. Three categories are “the number of atoms”, “Substructures” and “Properties”. The

<sup>20</sup> A mathematical theory of communication



**Definition of symbols.**

number of atoms indicates how many atoms there are in an amino acid residue. “Substructures” indicates whether an amino acid residue has a specific substructure or not. “Properties” indicates whether an amino acid residue has some character from a viewpoint. For example, the first item of “Properties” describes the structure that is the methylene group and a hydrogen bond acceptor are connected via any atom.

An element  $z \in Z$  denotes each substructure shown in Figure 2. Then, we can define any substructures except  $z$  as  $z^*$ . In other word, each element  $z^*$  is defined corresponding to each  $z$ . The substructure  $z^*$  is complement of the substructure of  $z$  because  $z \cap z^* = \emptyset$ ,  $\cup z \cap z^* = All$ . Then, we define the set  $Z^*$  as all elements  $z^*$ . Finally, we define the base substructure set  $X$  as  $X = Z \cup Z^*$ .

### 49.3.3 Definition of a set of intermediate bindings between any two base substructures

The activity of a peptide is determined not only by the structure of each amino acid residue but also by the relationship among amino acid residues. Here, we define an intermediate binding between two amino acid inductions as the distance between any two base substructures.

The definition of intermediate bindings among base substructures is arbitrary. For example, we can define an intermediate binding among three base substructures. When we describe the relationship among  $m$  substructures, the number of combinations is  $O^*(\sup{m}n)$ . Here,  $n$  is the number of substructures. The number of combinations increases by exponential order. To avoid the exponential order, we limited the number of substructures to 2.

Structures of peptides are more flexible than small compounds because peptides have many rotatable bonds. Descriptors for peptides should have a potential for describing the flexibility to obtain high accuracy.

We defined the intermediate bindings shown in Figure 3. To increase flexibility of descriptors, we added a set of bindings within some length to the definition. In Figure 3, '\*' denotes an amino acid residue and '~' denotes a peptide binding. '{ }' denotes 'or' condition. For example, ' $\{\sim, \sim, \sim, \sim^*\sim^*\}\}$ ' represents the peptide consisting of amino acid

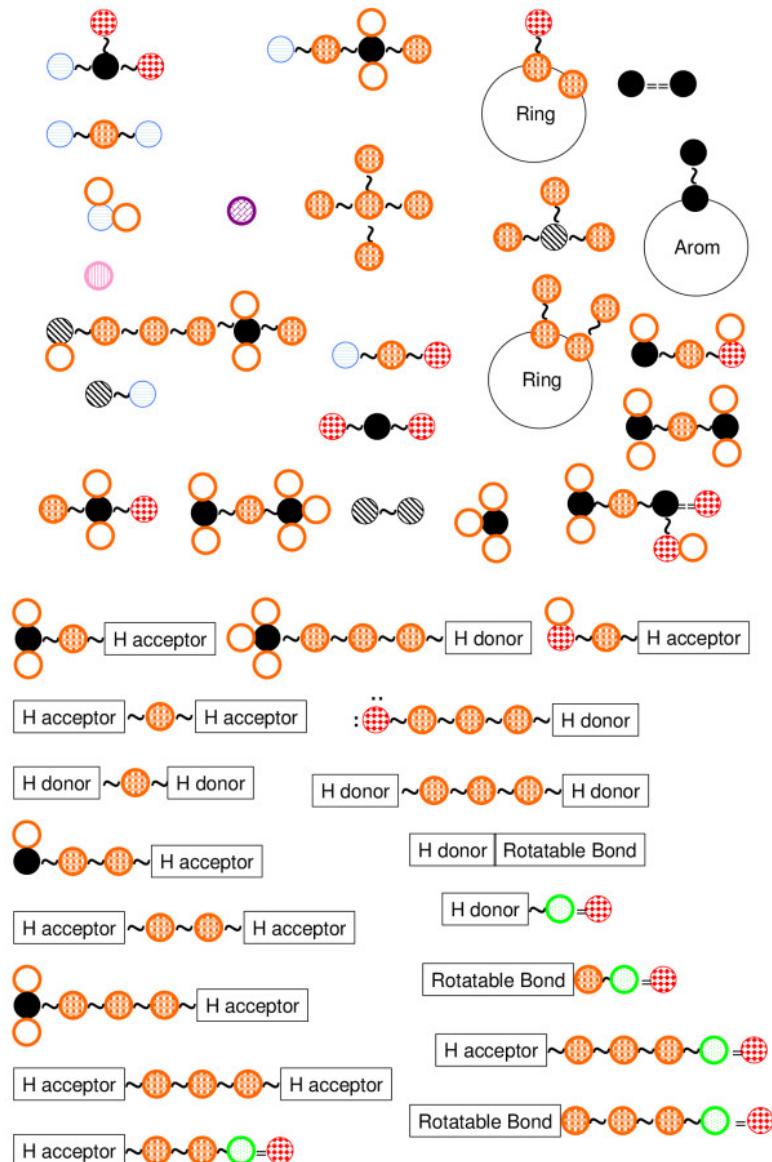


Figure 49.2: Figure 2. Definition of a set of base substructures in SPAD, which roughly has three categories, i.e., number of atoms, substructures (above), and peptide properties (below). Number of atoms includes 'Cl', 'F', 'N', 'O', 'C', 'C in aromatic ring', 'S', 'N in aromatic ring' and 'Sum of left atoms'.

residues from 0 to 2. We represent a set of intermediate bindings as set  $*Y$ .

Figure 49.3: Figure 3. Definition of a set of intermediate bindings in SPAD

**Definition of a set of intermediate bindings in SPAD.** Intermediate bindings between two substructures are shown.

#### 49.3.4 Definition of substructure-pair descriptor

Then, SPAD is defined as next function. We suppose that the number of  $X$  is  $N$  and that the number of  $Y$  is  $M$ .

When  $i$ ,  $x$ ,  $j$ ,  $y$  and  $k$  are given, a peptide  $a$  is converted to a bit with function  $F(i, x, k, y, j, x, a_p)$ . Here, we denote the suffix set  $(i, j, k)$  as  $b$ . Then, we obtained the matrix  $(_{ab}M) = (F(i, x, k, y, j, x, a_p))$  for the input of QSAR analysis algorithm. The vector  $(^{*}M^{**}a^1:\text{sub}: \ , \ ^{*}M^{**}a^2:\text{sub}: \ )$  is corresponding to the features of the peptide  $a_p$ .

## 49.4 Results and Discussion

#### 49.4.1 Definition of Datasets

We use two types of datasets for evaluation of the proposed descriptors. One is C5a inhibitors<sup>21</sup> and the other is kinase inhibitors<sup>22</sup>. Positive data are defined as peptides with high inhibitory potential, and negative data are defined as other peptides and peptides with random arrays. Content of dataset is as follows.

- C5a Inhibitors:
    - The number of positive peptides: 116
    - The number of negative peptides: 451
  - kinase inhibitors:
    - The number of positive peptides: 24
    - The number of negative peptides: 325

#### 49.4.2 Difference between SPAD and APH definition

SPAD is different from APH in defining whether any two substructures are connected directly to an intermediate binding. For example, when the main chain is connected to an aromatic ring of a side chain via a carbon chain and two amino acid residues have carbon chains which are different to each other in its length, APH classifies two amino acid residues. However, SPAD does not. The structures of amino acid residues are very similar so it is natural to consider

<sup>21</sup> C5a inhibitors [WO/2006/074964]

<sup>22</sup> Kinase inhibitors [WO/2003/059942]

that their properties are approximately similar. In this case, the descriptor that ignores the difference is better. The second different point between SPAD and APH is whether the information about properties is included in descriptors. It may be unnecessary to distinguish amino acid residues from a viewpoint of some property.

### Comparison of descriptors correlated highly with peptides' activity

By comparing each descriptor, we know that the range of the substructures of SPAD (Figure 4) is wider than that of APH (Figure 5). The range of APH is from 3 to 7 atoms. On the other hand, the range of SPAD is from 3 to 6 amino acid residues, which usually comprises 6-12 atoms. SPAD captures a wider range of characteristics than APH. Therefore, the range of SPAD is more appropriate for capturing properties of peptides than that of APH.

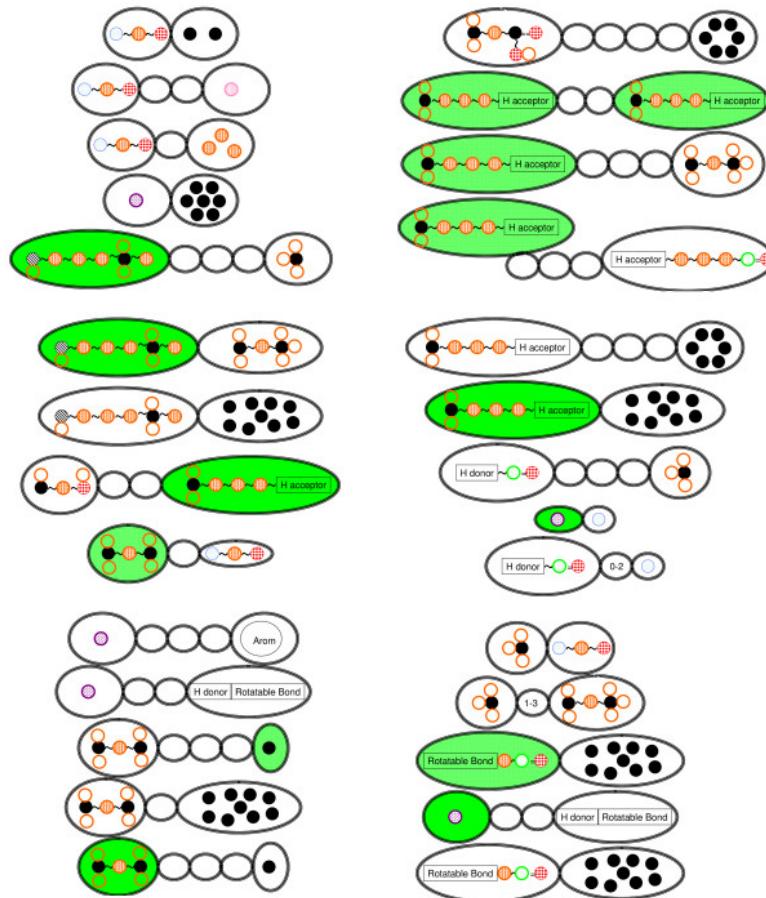


Figure 49.4: Figure 4. Descriptors with high correlation to peptides' activity in SPAD  
**Descriptors with high correlation to peptides' activity in SPAD.** The range of them is from 3 to 6 amino acids.

### Capturing Area of APH and SPAD in active peptides

In the case of SPAD (curve in Figure 6),  $x \ Z$  or  $x \ Z^*$  where  $x$  denotes a substructure. We show substructures  $x \ Z$  with high correlation to peptides' activity. In case of APH (dotted curve in Figure 6), we show substructures with high correlation to peptides' activity. There are few overlapped regions between SPAD and APH. SPAD and APH capture different regions complementarily. APH inclines to capturing a component of a peptide. On the other hand, SPAD descriptor inclines to capturing a relationship of side chains between two amino acid residues.

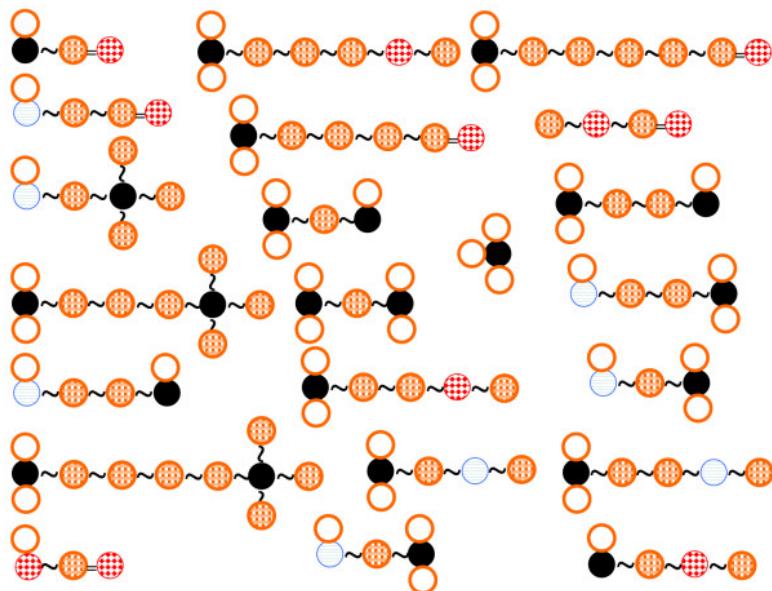


Figure 49.5: Figure 5. Descriptors with high correlation to peptides' activity in APH

**Descriptors with high correlation to peptides' activity in APH.** The range of them is from 3 to 6 atoms. Its length is shorter than that of SPAD.

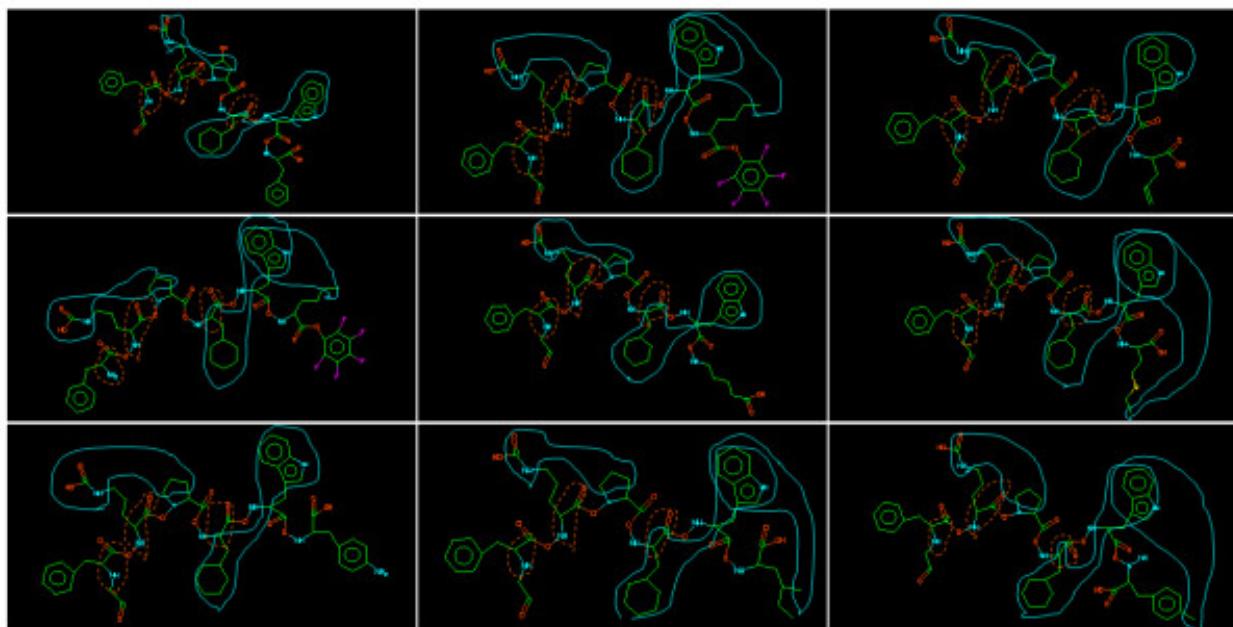


Figure 49.6: Figure 6. Mapping of representative descriptors with high entropy of SPAD and APH to C5a active peptide

**Mapping of representative descriptors with high entropy of SPAD and APH to C5a active peptide.** Curve indicates SPAD and dotted curve indicates APH. There are few overwrapped regions between two descriptors.

## Definition of dataset for similarity search with Tanimoto coefficient

Peptides are classified in three categories:

- non-active: negative peptides.
- active reference: positive peptides which are the basis of similarity search with Tanimoto coefficient.
- active: positive peptides except for active reference.

All peptides were ordered by descendent ordering with Tanimoto coefficient.

## Comparison of the performance of SPAD with APH

When the structure of two peptides is similar and a descriptor captures a whole structure or property of peptides, these two features have similar sequences of bits. As a result, Tanimoto coefficient between these peptides becomes large. Structures of active peptides for a target protein are usually similar to each other because the pocket of target protein is same. When we describe peptides with a descriptor capturing whole peptides' structures or properties, Tanimoto coefficient between any two active peptides is larger.

Oppositely, Tanimoto coefficient between an active peptide and a non-active peptide is smaller because these two features are different to each other. However, if we describe peptides with a poor descriptor, we cannot always measure the similarity of peptides with Tanimoto coefficient. Poor descriptors break the similarity of structures at mapping to features. Therefore, Tanimoto coefficient is an indicator of the descriptor's performance.

All peptides are ordered by descendent ordering with Tanimoto coefficient. Then, we count the number of active peptides with this ordering. Figure 7 shows the enrichment factor with Tanimoto coefficient. The horizontal-axis and the vertical-axis is defined as follows.

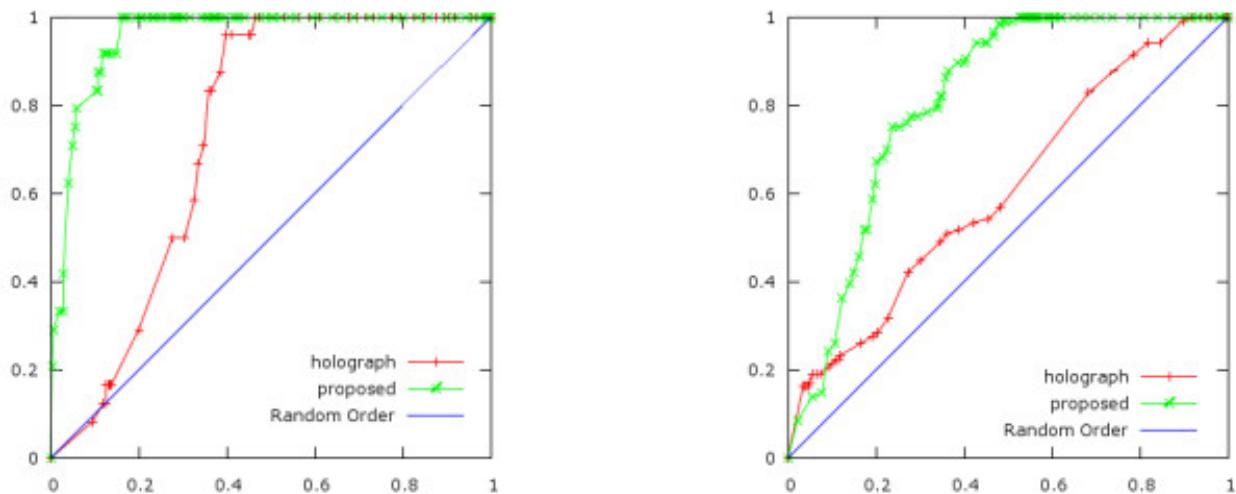


Figure 49.7: Figure 7. Enrichment factor with Tanimoto coefficient

**Enrichment factor with Tanimoto coefficient.** C5a case (Left) and kinase inhibitor case (Right). The horizontal axis indicates the percentage of peptides ordered by descendent ordering with Tanimoto coefficient. The vertical axis indicates the percentage of active peptides in this ordering. The random line (diagonal line) indicates theoretically obtained curve in case of random ordering. ‘x’ dotted line shows the performance of SPAD and ‘+’ dotted line shows the performance of APH. In both case, the enrichment factor of SPAD is much higher than that of APH.

- The horizontal-axis
- The vertical-axis

The graph increases more rapidly as active peptides have larger Tanimoto coefficient than non-active peptides.

In both cases, C5a (left figure at Figure 7) and kinase inhibitors (right figure in Figure 7), the graph in case of SPAD is higher than the graph in case of APH. The enrichment factor with the SPAD is higher than with APH at any percentage of active peptides. Therefore, the SPAD translates similar structures to similar features more precisely than the APH. This fact means that the performance of the SPAD is higher than the performance of APH in the case of analyzing peptides' activity.

## 49.5 Conclusions

It is necessary for two-dimensional QSAR of peptides that are sequences of 450 types of amino acid inductions to capture various properties with descriptors. The atom pair holographic code and substructure pair descriptor that we proposed are such descriptors. APH captures internal characters of an amino acid induction. On the other hand, SPAD captures the relationship between two amino acid inductions. SPAD captures much more information for QSAR of peptides than APH and distinguishes active peptides from non-active peptides more accurately.

## 49.6 Competing interests

The authors declare that they have no competing interests.

## 49.7 Authors' contributions

TO conceived the method, evaluated this method and described this manuscript. SM discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

# AN INVESTIGATION INTO PHARMACEUTICALLY RELEVANT MUTAGENICITY DATA AND THE INFLUENCE ON AMES PREDICTIVE POTENTIAL

## 50.1 Abstract

### 50.1.1 Background

In drug discovery, a positive Ames test for bacterial mutation presents a significant hurdle to advancing a drug to clinical trials. In a previous paper, we discussed success in predicting the genotoxicity of reagent-sized aryl-amines ( $\text{ArNH}_2$ ), a structure frequently found in marketed drugs and in drug discovery, using quantum mechanics calculations of the energy required to generate the DNA-reactive nitrenium intermediate ( $\text{ArNH}^+$ ). In this paper we approach the question of what molecular descriptors could improve these predictions and whether external data sets are appropriate for further training.

### 50.1.2 Results

In trying to extend and improve this model beyond this quantum mechanical reaction energy, we faced considerable difficulty, which was surprising considering the long history and success of QSAR model development for this test. Other quantum mechanics descriptors were compared to this reaction energy including AM1 semi-empirical orbital energies, nitrenium formation with alternative leaving groups, nitrenium charge, and aryl-amine anion formation energy. Nitrenium formation energy, regardless of the starting species, was found to be the most useful single descriptor. External sets used in other QSAR investigations did not present the same difficulty using the same methods and descriptors. When considering all substructures rather than just aryl-amines, we also noted a significantly lower performance for the Novartis set. The performance gap between Novartis and external sets persists across different descriptors and learning methods. The profiles of the Novartis and external data are significantly different both in aryl-amines and considering all substructures. The Novartis and external data sets are easily separated in an unsupervised clustering using chemical fingerprints. The chemical differences are discussed and visualized using Kohonen Self-Organizing Maps trained on chemical fingerprints, mutagenic substructure prevalence, and molecular weight.

### 50.1.3 Conclusions

Despite extensive work in the area of predicting this particular toxicity, work in designing and publishing more relevant test sets for compounds relevant to drug discovery is still necessary. This work also shows that great care must be taken in using QSAR models to replace experimental evidence. When considering all substructures, a random forest model, which can inherently cover distinct neighborhoods, built on Novartis data and previously reported external data provided a suitable model.

## 50.2 1. Introduction

### 50.2.1 1.1 Aims

In the field of drug-discovery, a positive Ames test can halt development of a particular chemotype and possibly work on an entire drug target because genotoxicity of a potential therapeutic would be a serious issue that needs to be avoided. Sufficiently nuanced rules do not exist to fix such a problem while maintaining the careful balance of potency and properties. Compounding this problem is that impurities or metabolites that could be generated in parts-per-million quantities ( $10 \mu\text{g/day}$ ) are just as serious from a regulatory standpoint, which could eliminate an essential core structure. Thus, prediction of whether a starting material, degradation product, or drug will be mutagenic in the Ames genotoxicity test is our primary goal. More specifically, our initial focus was on aryl-amines, which are commonly used reagent building blocks in many small molecule drug-discovery projects and appear as a substructure in at least 13% of currently marketed drugs<sup>1</sup>. Aryl-amines also have a known mechanism for genotoxicity. In a previous article, we have shown that an *in silico* assessment of aryl-amines using quantum mechanics reaction energy calculations can provide excellent detection of mutagenic aryl-amines<sup>2</sup>. However, we were surprised that statistical models incorporating additional descriptors did not improve the performance of the single nitrenium formation energy parameter given the wealth of QSAR literature showing accuracy approaching or exceeding the known experimental error. Additionally, we found that the set of Novartis aryl-amines was surprisingly challenging to model compared to those in the literature.

Our ultimate goal is to provide medicinal chemists with usable models to improve the chances of avoiding a toxicity trap that is often visible only after low-throughput tests come back. The aryl-amines can be predicted reliably with the nitrenium formation energy calculation but comparing all-substructure external Ames results to our Novartis results, we found that these were also much harder. Other groups in pharmaceutical companies have noted difficulties in predicting mutagenicity in aryl-amines<sup>3</sup>, and in internal all-substructure data sets using commercial software<sup>45</sup>.

Previous to this article, differences between data sets typically used in the literature for building mutagenicity predictive methods and the data at pharmaceutical companies have not been compared. This is key to the disconnect from literature studies and pharmaceutical studies. The high level of performance of statistical models in this arena with constructed test sets is misleading and does not reflect performance in pharmaceutically relevant sets. Here we show the relative difficulty in predicting the Ames test result in the Novartis aryl-amines and other substructures, in contrast to literature sets.

### 50.2.2 1.2 The Significance of the Ames Test

Many compounds in the environment released from industrial pollution and production are known to cause cancer<sup>6</sup>. Regulatory agencies around the world in cooperation with industry experts have adopted stringent test methods to identify and regulate the use of chemical mutagens that might be exposed to the environment or administered

---

<sup>1</sup> Known drug space as a metric in exploring the boundaries of drug-like chemical space

<sup>2</sup> Avoidance of the Ames test liability for arylamines via computation

<sup>3</sup> Reaction energies computed with density functional theory correspond with a whole organism effect; modelling the Ames test for mutagenicity

<sup>4</sup> Comparative Evaluation of *in Silico* Systems for Ames Test Mutagenicity Prediction: Scope and Limitations

<sup>5</sup> The computational prediction of genotoxicity

<sup>6</sup> A brief history of scrotal cancer

to humans directly as pharmaceuticals<sup>7</sup>. Carcinogenicity is usually determined by an array of in-vivo and in-vitro surrogate tests, which are specified by regulatory authorities before administration to man. The Ames bacterial test is a simple experiment to perform and it is a mandatory regulatory test that has been in use for almost 40 years and correlates with life-time rodent carcinogenicity studies that require 2 years to complete<sup>89</sup>.

At the molecular level, this test for mutagenicity<sup>1011</sup> detects a substance's ability to cause mutations in engineered strains of *Salmonella typhimurium* by observing return of function by point mutations in an altered His operon gene. The mutations in the His operon strains prevents histidine biosynthesis, thus random mutations or mutations due to an external agent must occur for colony growth on histidine-deficient medium. Many compounds are converted to mutagenic compounds after metabolism, so the test is performed with and without pre-incubation of the compound with rat liver enzymes. The bacterial strains used in the test have been further engineered to have permeable cell membranes, a reasonably high spontaneous mutation rate, and diminished DNA repair capacity<sup>12</sup>.

Although the result of the Ames test can be reported as a standardized quantity of the number of colonies formed, in most recent studies and databases, including the Novartis internal test results, are reported as categorical results: "Ames-positive" (Ames+) or "Ames-negative" (Ames-). Additionally, it has been shown that the qualitative carcinogenicity result is not improved by quantitative mutagenicity potency data<sup>8</sup>. An increase in number of colonies over control by at least a factor of 2 and a clear dose dependence in the mini-Ames screening test<sup>13</sup> is classified as a positive result. Although high-throughput screening assays exist, they do not faithfully predict the result of the Ames test and at the same time require a significant investment<sup>1415</sup>. Consequently, the volume of data available for the Ames test is fairly limited. The turnaround time and cost for Ames testing makes accurate in silico models quite useful.

There are some limitations of the Ames test that present a challenge to building accurate in silico models. The exact sensitivity of the test for carcinogenicity is somewhat controversial<sup>8</sup>, but in a recent retrospective analysis by FDA and EPA researchers of carcinogenicity and surrogate test results showed the Ames test is positive for 49% (275Ames+/557 rodent carcinogens) of carcinogenic compounds but only 19% of the Ames+ compounds are not carcinogenic to rats (85 Ames+/431 rodent non-carcinogens)<sup>16</sup>. Reproducibility both across and inside one laboratory conducting the test is another serious issue. Both literature and internal intra-laboratory assessments of the test, at least in a 2-strain screening version of the test, have found discrepancies on the order of 15-20%<sup>9</sup>. Based on a retrospective analysis of 237 compounds at Novartis with multiple Ames screen test results, this is a realistic estimate; there were 49 (21%) with discrepant results. Among aryl amines, 13 out of 57 compounds with multiple test results were discordant (23%). The test is sensitive and uses high concentrations of the test chemical, which can increase the effect of impurities including metals<sup>17</sup>, degradation products, or reagents<sup>1819</sup>. The chemical can also be toxic to the bacterial system, most notably antibacterials or cytotoxic compounds, but must still be tested to the maximum possible concentration<sup>7</sup>.

### 50.2.3 1.3 Substructure alert and QSAR methods

The cause of cancer through the action of chemicals has been studied extensively, and the process typically begins with the chemical, or one of its metabolites, interacting with DNA, which subsequently leads to mutations<sup>20</sup>. The principle of mutagenicity through reaction of DNA with electrophiles has been especially useful in rationalizing and deriving

<sup>7</sup> S2(R1) Guidance on genotoxicity testing and data interpretation for pharmaceuticals intended for human use

<sup>8</sup> Predicting rodent carcinogenicity from mutagenic potency measured in the Ames *Salmonella* assay

<sup>9</sup> Alternatives to the carcinogenicity bioassay: in silico methods, and the in vitro and in vivo mutagenicity assays

<sup>10</sup> The Ames *Salmonella*/microsome mutagenicity assay

<sup>11</sup> Detection of carcinogens as mutagens in the *Salmonella*/microsome test: assay of 300 chemicals

<sup>12</sup> Isolation of plasmid pKM101 in the Stocker laboratory

<sup>13</sup> Comparison of the Results of a Modified Miniscreen and the Standard Bacterial Reverse Mutation Assays

<sup>14</sup> Evaluation of high-throughput genotoxicity assays used in profiling the US EPA ToxCast (TM) chemicals

<sup>15</sup> Evaluation of the Vitotox(TM) and RadarScreen assays for the rapid assessment of genotoxicity in the early research phase of drug development

<sup>16</sup> An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: I. Identification of carcinogens using surrogate endpoints

<sup>17</sup> A survey of metal-induced Mutagenicity in vitro and in vivo

<sup>18</sup> An evaluation of the sensitivity of the Ames assay to discern low-level mutagenic impurities

<sup>19</sup> Risk Assessment of Potentially Genotoxic Impurities within the Framework of Quality by Design

<sup>20</sup> Advances in chemical carcinogenesis: a historical review and prospective

“toxicophores,” substructures that are strongly associated with mutagenicity<sup>2122</sup>. Some of these mechanisms have been studied carefully in vitro and in vivo<sup>23</sup>, and DNA or protein adducts can be measured and observed experimentally<sup>2425</sup>. The first line of defense in avoiding carcinogenicity in drug design is through the use of alerts to chemicals commonly associated with carcinogenicity, mostly derived from environmental testing<sup>2226</sup>. Kazius et al. provided an analysis of mutagenicity data correlating chemical substructures to mutagenicity<sup>212728</sup>. Most of these toxicophores are associated with Michael acceptors, electrophiles, or enophiles including  $\alpha,\beta$ -unsaturated carbonyl systems, aziridines and epoxides, aliphatic halides, azides, and acid halides. Others such as aryl-amines and nitroaromatics are known to be converted to more reactive species through oxidation, reduction, and conjugation metabolism reactions<sup>29</sup>. The simplest of prediction systems search a chemical for these substructures and uses rules to correctly predict the Ames+ compounds for known mutagenic substructures. However, despite the inclusion of detoxifying rules, these methods misclassify many of the Ames- compounds as positive. For chemical sets containing many unknown classes of mutagens, structural alert systems like DEREK correctly predict only around 50% of the Ames+ compounds<sup>53031</sup>. A similar result (55% sensitivity) was found for compounds tested at Novartis (all mini-Ames screening results, August 2009) using an internally modified DEREK rule set<sup>32</sup>. There is a long history of modeling mutagenicity on chemicals expected to be encountered from environmental and food exposure<sup>93334353637</sup>. Recent reviews on statistical models of mutagenicity<sup>9333839</sup> and a recent collaborative head-to-head mutagenicity prediction challenge summarize the current state of the art for external sets<sup>40</sup>. A summary of some recent models is included in the Supporting Information (Additional file<sup>4041</sup>.

Additional file 1

**Supplemental figures and tables.** Supplemental figures and tables referred to in the text of the article.

Click here for file

## 50.3 2. Methods

### 50.3.1 2.1 Data set preparation

Ames test data is available from a number of sources including literature reviews, regulatory agencies, and funding agencies<sup>42</sup>. For our analysis, we focused on an internal Novartis set and two literature sets (combined into one) for aryl-amines and four datasets covering all substructures as detailed in Table 1.

<sup>21</sup> Derivation and Validation of Toxicophores for Mutagenicity Prediction

<sup>22</sup> Computer Prediction of Possible Toxic Action from Chemical Structure; The DEREK System

<sup>23</sup> Carcinogenesis by chemicals: an overview-GHA Clowes memorial lecture

<sup>24</sup> Monocyclic aromatic amines as potential human carcinogens: old is new again

<sup>25</sup> DNA adducts formed by a novel antitumor agent 11 $\beta$ -dichloro in vitro and in vivo

<sup>26</sup> Computer prediction of possible toxic action from chemical structure: an update on the DEREK system

<sup>27</sup> Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity

<sup>28</sup> Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP

<sup>29</sup> A Comprehensive Listing of Bioactivation Pathways of Organic Functional Groups

<sup>30</sup> Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules

<sup>31</sup> Computational prediction of genotoxicity: room for improvement

<sup>32</sup> S18: In silico assessment of safety concerns esp. of carcino-genic potential

<sup>33</sup> Predictivity and Reliability of QSAR Models: The Case of Mutagens and Carcinogens

<sup>34</sup> A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100

<sup>35</sup> Quantitative structure-activity relationships of mutagenic aromatic and heterocyclic amines

<sup>36</sup> Structure-activity relationships of chemical mutagens and carcinogens

<sup>37</sup> Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors

<sup>38</sup> Benchmark Data Set for in Silico Prediction of Ames Mutagenicity

<sup>39</sup> Collection and Evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity

<sup>40</sup> Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set

<sup>41</sup> Predicting Mutagenicity of Aromatic Amines by Various Machine Learning Approaches

<sup>42</sup> Collaborative development of predictive toxicology applications

For the aryl-amine sets, molecules with other substructures associated with mutagenicity, such as nitroaromatic, nitrile oxide, N-nitroso substructures, were removed from the analysis. Set A was from internal Novartis Ames screening test results tested in one laboratory up to 2009. Set B is the aryl-amine subset from compilations published by Hansen et al<sup>38</sup><sup>43</sup> and Kazius et al<sup>21</sup>. All Ames screening results at Novartis excluding those with discrepant values comprised Set C. The complete set of Hansen et al. was used as Set D. Set E represents a second pharmaceutically relevant set of marketed pharmaceuticals extracted from a recent review by Brambilla and Martelli<sup>44</sup>. The complete Kazius set, Set F, was included in the analysis to give a combined collection of 9423 molecules. A basic summary of the sets is shown in Table 1.

### 50.3.2 2.2 Computational studies

For all PLS and random forest models, a set of 185 2D descriptors available in the MOE software program<sup>45</sup> and a circular Morgan fingerprint<sup>46</sup><sup>47</sup> generated with a radius of 3 bonds (ECFP6) hashed to 1024 count variables using RDKit<sup>48</sup>. Quantum mechanics reaction energies for nitrenium formation from the primary amine were calculated as described in a previous publication considering conformation, tautomer, and spin state<sup>2</sup> using B3LYP hybrid density functional theory energies with a 6-31G\* basis set for all C, F, H, O, S, N, and P atoms and the LANL2DZ basis set and ECP for Cl, Br, and I in Gaussian03<sup>49</sup>. The nitrenium formation energy is equal to the energy of the lowest energy amine conformation subtracted from the energy of the lowest energy nitrenium ion plus hydride anion ( $\text{ArNH}_2 \rightarrow \text{ArNH}^+ + \text{H}^-$ ). The anion formation energy and radical formation energy were similarly calculated and generated and also used the 6-31G\* basis set, though improvement in the energy could be expected by adding a diffuse function for the anion. The AM1<sup>50</sup> HOMO and LUMO orbital energies were calculated using the MOPAC<sup>51</sup> module implemented in MOE. Nitrenium ion charges were determined by using the lowest energy B3LYP/6-31G\* nitrenium ion conformation and calculating the NBO population analysis<sup>52</sup><sup>53</sup> using B3LYP and a 6-311G\* basis set in Gaussian03.

The random forest classification models used in this article were constructed using the randomForest package<sup>54</sup> for R<sup>55</sup> using the approach developed by Breiman<sup>54</sup><sup>56</sup>. The method was used by constructing 500 unpruned trees using a random sample of  $\sqrt{N}$  of the available predictors for each tree and a 0.632 bootstrap sample of the data for each tree. The remaining data was predicted using the tree and averaged to create the combined out-of-bag (OOB) predictions depicted in the receiver operator characteristic (ROC) plots.

The PLS classifications were done using a PLS regression implemented using the kernel algorithm available in the PLS<sup>57</sup> package in R<sup>55</sup>. Variables showing little variance among cases were removed using the nearZeroVar function in the caret<sup>58</sup> package and all variables were centered by the mean and divided by the standard deviation using the preProcess function in the caret package. The response variable was 0 for Ames- or 1 for Ames+ in these models and the predicted value found from the regression was used as a cutoff in constructing a classification model. All ROC plots and area-under-the-curve (AUC) metrics used the ROCR package<sup>59</sup> in R. Averaging of model performances in the ROC plots was done with vertical averaging of performance at a given false-positive rate, and error bars give the standard deviation. A random sample of 70% of the data was used for training and the process was repeated 100 times

<sup>43</sup> Benchmark Data Set for In Silico Prediction of Ames Mutagenicity

<sup>44</sup> Update on genotoxicity and carcinogenicity testing of 472 marketed pharmaceuticals

<sup>45</sup> Molecular Operating Environment

<sup>46</sup> The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service

<sup>47</sup> Extended-Connectivity Fingerprints

<sup>48</sup> NOTITLE!

<sup>49</sup> Gaussian 03

<sup>50</sup> Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model

<sup>51</sup> MOPAC

<sup>52</sup> Natural hybrid orbitals

<sup>53</sup> Natural population analysis

<sup>54</sup> Classification and Regression by randomForest

<sup>55</sup> R: A Language and Environment for Statistical Computing

<sup>56</sup> Random Forests

<sup>57</sup> pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)

<sup>58</sup> caret: Classification and Regression Training

<sup>59</sup> ROCR: visualizing classifier performance in R

representing in part how small batches of Ames results might perform. Variables with zero variance were removed prior to training thus removing 906 variables for the Novartis set and 956 for Set B, and variables were mean-centered and variance-scaled at each training step.

The aryl-amine data sets were constructed as previously described<sup>2</sup>. The all-substructure sets were combined using Pipeline Pilot<sup>60</sup> ignoring chirality due to a lack of chirality in our 2D descriptors and after generating a canonical tautomer. It is also worth noting that absolute chirality determination cannot be done for all compounds and inevitable data entry errors can make this another source of error. A consensus Ames result was used in these all-substructure data with the definition that any Ames+ result in any of the sources was an Ames+ result. Substructure counts were calculated using a Pipeline Pilot<sup>60</sup> protocol with substructure queries that were able to closely reproduce the counts generated in the work of Kazius et al.<sup>21</sup> for their data set (see Additional file

Additional file 2

**Ames toxicophore queries.** Compressed library of Ames toxicophore queries as.mol files.

[Click here for file](#)

Additional file 3

**Public structures used in the research.** SDF file with 6812 structures and data in public sets B, D, E, and F of the paper. Fields BrambillaMarketedDrug, Hansen, and Kazius indicate which set the structure is in and if this is 1, the corresponding fields BrambillaExpAmes, HansenExpAmes, or KaziusExpAmes respectively will indicate the Ames result for that set. Sets D, E, and F are unfiltered, all-substructure sets: Set D is from Hansen et al.,<sup>38</sup> Set E is the marketed drug set taken from Brambilla et al.<sup>44</sup>, and Set F is from Kazius et al.<sup>21</sup> Set B is an aryl amine set which includes the union of Set D and F filtered by the aryl amine substructure. Other fields are results of counting the query substructures in Additional file

[Click here for file](#)

The Self-Organizing Map<sup>61</sup> for the combined all-substructure set was generated in Schrodinger Canvas version 1.4<sup>62</sup> with a 30 cell by 30 hexagonal cell output grid. The program uses Euclidean distance to measure similarity between compounds, and the internal Morgan<sup>46</sup>-type circular fingerprints<sup>47,63</sup> generated with radius 2 and functional atom types were used as descriptors (ECFP4). The TopKat mutagenicity prediction was centered and scaled to give results from 0 to 1 and the random forest model provided probabilities between 0 and 1 for the Ames+ class. The deviation was then the difference between either 1 for an experimentally Ames+ or 0 for Ames- result and the model output. For the aryl-amine set, the ‘kohonen’ package<sup>64</sup> in R was used instead due to a discovered problem in Canvas with applying trained maps to new compounds. In this case, RDKit was used to generate circular Morgan fingerprints hashed to 1024 count variables as described for the statistical modeling.

## 50.4 3. Results and Discussion

In the following results, the differences in the sets are examined in terms of their properties, presence of previously identified mutagenic substructures, and structural similarity and clustering visualized using Kohonen self-organized maps. The difference in predictivity of multiple statistical methods and descriptors between pharmaceutically relevant data and literature compilations is analyzed firstly for aryl-amines and then for sets containing all substructures. For aryl-amines, the quantum mechanically derived reaction energy for forming a known reactive intermediate was shown to be a more stable and accurate predictor than statistical models with more descriptors.

---

<sup>60</sup> Pipeline Pilot

<sup>61</sup> NOTITLE!

<sup>62</sup> Canvas

<sup>63</sup> Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods

<sup>64</sup> Self- and Super-organising Maps in R: the kohonen package

### 50.4.1 3.1 Comparison of molecules in external Ames data sets and pharmaceutically relevant sets

As can be seen in Table 1, the Novartis Set A of aryl-amines and Set C of all substructures tested have a low number of Ames+ compounds compared to their literature counterparts (Sets B and D). In aryl-amines (Set A), only 22% of the molecules are Ames+, and in the entire set of test results at Novartis (Set D), only 15% of the molecules are Ames+. This low percentage is quite similar to other recent reports on Ames results at other pharmaceutical companies such as the recent report from Hillebrecht et al. from Roche<sup>4</sup> where  $300/2335 = 13\%$  of the internal compounds were Ames+. A paper by Leach et al.<sup>3</sup> on aryl-amines from AstraZeneca had a slightly higher percentage ( $109/312 = 35\%$ ) of Ames+ aryl-amines less than 250 g/mol. However, in the literature sets (Sets B and D): 71% of aryl-amines and 54% of the entire chemical space are Ames+. Perhaps surprisingly, the marketed pharmaceutical set, set E, has a non-zero incidence of Ames+ test results but it is fairly low-around 12%. An Ames+ test result is only part of a potential drug's profile but the risk of carcinogenicity in later stage animal testing and added regulatory scrutiny present a significant hurdle to drug development in a competitive space.

Another major difference between the sets is the number of compounds of intermediate molecular weight (200-500 g/mol). This range was nearly absent in the benchmark sets shown in the left plot of Figure 1, but for the Novartis and marketed drugs sets in the right plot, there is a large percentage of the compounds. The bias towards larger molecules likely reflects that the Ames test has often been considered later in drug development, when molecules and their precursors have more complex structures. For all Novartis compounds tested, the median weight was 415 with a fairly wide distribution from 80 to 600 g/mol as shown in Figure 1 (right). In contrast, the median weight for Set D is about 229, with a slightly sharper distribution as shown in the left plot in Figure 1. In quantitative terms, the range of 400-600 g/mol in the literature set, Set D, contains just 372 molecules (6% of the set) compared to 1380 compounds (50% of the set) found in the Novartis Set D. The set of marketed pharmaceuticals with Ames test results is shown in green in the right plot of Figure 1. It has a molecular weight distribution more like the Novartis set, with a median weight of 309 g/mol.

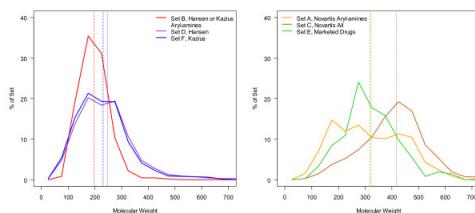


Figure 50.1: Figure 1. Molecular weight distributions of Ames test data sets

**Molecular weight distributions of Ames test data sets.** Molecular weight distributions of all six sets plotted up to molecular weight of 700 g/mol with mean molecular weight marked with the dotted line of the same color. The literature sets are shown in the left plot, and the Novartis (Sets A and C) and the marketed drugs compilation (Set E) are shown on the right.

For the aryl-amine sets, Sets A and B, the situation is similar: the Novartis set, Set A, has a higher average molecular weight but there is an even distribution of weights from about 150-500 g/mol. In Set B, there are only 3 aryl-amine data in the range of 400-600 g/mol. In Set A, there are 93, which is almost 30% of the set. The fact that there is such an even distribution, including a large fraction of lower molecular weight compounds, in the Novartis set may reflect the importance of this class and the response to the issue of genotoxicity. When an issue is identified, the typical medicinal chemistry approach is to synthesize dozens of molecules and test all of them. Building blocks that are components of larger molecules are often tested in case of trace genotoxic impurities and for internal guidelines are tested if used for a final clinical candidate. Also drugs for different disease areas such as neuroscience may require smaller molecules.

The “toxicophores” described in Kazius were used to construct a further comparison of two of the all-substructure sets, Set C (Novartis) and Set D (Hansen). Figure 2 portrays the overall count of these functional groups in the two sets and Table 2 summarizes the percentage of Ames+ compounds in each class, done in a filtered manner where nitroaromatic is the first class. The labels “[OH, NH2][O, N]” and “ArN(CH2C)2”, denote an alcohol or amine bonded to an oxygen or nitrogen, such as a hydrazine or a hydroxylamine, and a di-alkylarylamine respectively. Naturally, a number of these functional groups are less common in drug design because of their reactivity or under-represented in test results

or in the compounds synthesized due to concerns for toxicity in the Ames test. Aryl-amines, aryl-amine-amides, and dimethylarylamines are quite-well represented and have a lower Ames+ rate. Nitroaromatics were not nearly as represented in this set and are well-established as having a high probability of being responsible for genotoxicity. Building statistical models in the other data sets may benefit greatly from having a feature so strongly associated with genotoxicity. The structures in our set with a nitroaromatic group were Ames+ 40% of the time and in Set D, 84% were Ames+.

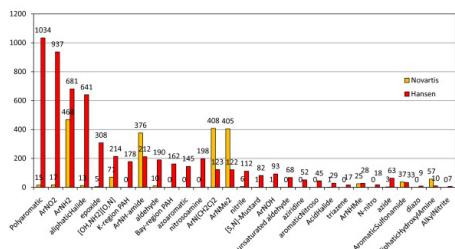


Figure 50.2: Figure 2. Mutagenic substructure distributions of Ames test data sets (non-aryl-amine)

**Mutagenic substructure distributions of Ames test data sets (non-aryl-amine).** Comparison of Novartis Set C (orange bars) and the Hansen et al. set, Set D (red bars) by mutagenic substructure counts taken from Kazius et al.

Even within a distinct substructure, aryl-amines, the pharmaceutically relevant set is much different from the Ames test results typically presented in the literature. The use of Kohonen, or Self-Organizing, Maps<sup>61</sup> (SOMs) was helpful for visualizing the differences between the sets using distances between molecular fingerprints of the molecules. This technique clusters molecules with similar substructure with each other in the best matching cell while also maintaining a 2-dimensional grid of cells such that similar molecules appear in adjacent cells. Multidimensional scaling and simple clustering was also investigated for visualization but yielded unsatisfactory neighbors in the first case, and a less useful visualization tool in the second. A SOM map built with the aryl-amines found in all sets is shown in Figure 3 but colored by property. The left plot is colored by where the aryl-amine is from: whether the molecule is a Novartis aryl-amine (orange) or from the external sets (blue). The center plot is colored by whether the compound is Ames+ (red) or Ames- (green). Finally, some representative structures are shown in the approximate locations of the map in the right plot. Cells with some of each class are colored as pie charts depicting the relative fraction of each class present. The approach knows nothing of the set membership of each compound, yet it shows a striking separation of the 1005 aryl-amines both by whether they are part of a drug company's tested compounds or from a literature Ames compilation. Due to the clustering by substructure and that Set D is so largely Ames+, this method also provides excellent separation of the Ames+ aryl-amines. Polyaromatic amines such as aminoacridines, aminophenazines, or aminochrysenes are not highly common in medicinal chemistry. However, they are quite common in the available literature sets. These structures are in Ames+ cells and can be easily represented using molecular fingerprints found in many QSAR models. This makes these sets easier to model.

In Figure 3, we also show where commercial aryl-amines that have been calculated by our model lie in the map. A significant population exists near CF<sub>3</sub>-substituted anilines in the top right, which have historically been Ames- (2<sup>nd</sup> plot) and have higher nitrenium formation energies. The top left of the map contains mostly larger and more polar aryl-amines, which were purposely left out of the calculations because of the goal of identifying safer starting materials and the better performance of the predictor for lower molecular weight aryl-amines. The center-right area of the map is where a large proportion of the commercially available aryl-amines are located avoiding some of the larger polyaromatic and triphenyl systems. It is also an area that has cells that contain Ames+ and Ames- amines. The nitrenium formation energy predictor can clarify which compounds in this area are safer bets as discussed in the next section.

The set of 9423 unique compounds included in sets C, D, E, and F are depicted in the SOM in Figure 4. The left-most, top plot colors the SOM by whether it is from Novartis (orange) or an external set (red), the center-top plot shows the distribution of Ames+ (red) and Ames- (green) compounds, and the upper-right plot shows the population of the cells. For the aryl-amine SOM, the population was somewhat uniform, but in the all-substructure plot, the number of molecules per cell varies from 1 to 66. This is natural due to the more extensive differences in the set. The bottom

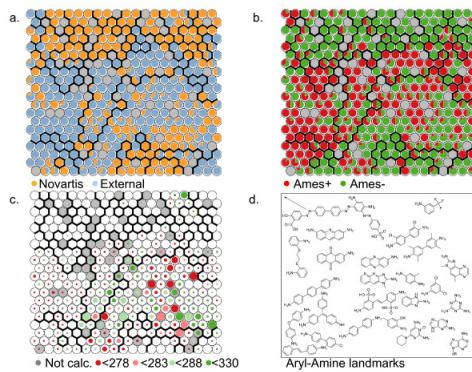


Figure 50.3: Figure 3. Self-organizing map of aryl-amine chemical space

**Self-organizing map of aryl-amine chemical space.** Comparison of aryl-amines in Set C and D using a self-organizing map (SOM) based on circular Morgan fingerprints, the SOM cells are shown in the top two plots with coloring applied based on a.) whether the compound is from Novartis, or b.) whether an Ames+ result (red) exists for the molecules in the cells. In c.) commercial aryl-amines have been mapped to the SOM trained on known aryl-amines and colored by their predicted Ames test result based on nitrenium formation energy in kcal/mol. Size of the marker conveys the number of compounds in the cell. In d.), an approximate location of some Set D aryl-amines is given.

three plots then further characterize where certain substructures are distributed in the SOM. The blue cells show the presence of a polyaromatic substructure in the bottom-left. The aryl-amines are distributed throughout the area and depicted in shades of red. Those molecules with multiple aryl-amine substructures have an increasingly pink hue sector of the pie marker. Finally in the bottom-right plot, the nitroaromatics are highlighted in shades of green. As in the case of the aryl-amines, multiple substructures are given as separate pie-chart sectors of increasing brightness. These are seen almost solely in the external set and in regions of high mutagenicity.

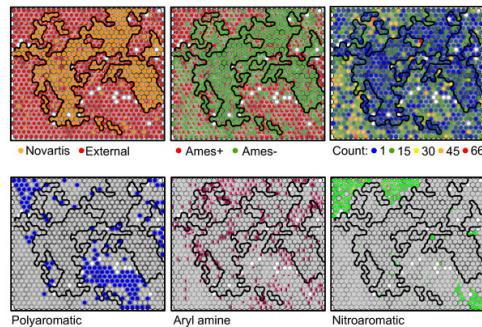


Figure 50.4: Figure 4. Self organizing map of the chemical space of compounds considered colored by properties

**Self organizing map of the chemical space of compounds considered colored by properties.** SOM for all compounds in Sets C, D, E, and F colored according to property (with pie charts to represent the percentage of molecules in the cell matching a property). The properties labeled from top-left clockwise: whether the compound is in the Novartis deck, whether it has an Ames+ result, the number of molecules in each cell, whether the polyaromatic substructure occurs in a cell, how many compounds have an aryl-amine substructure present, or how many compounds have a nitroaromatic substructure.

## 50.4.2 3.2 Predicting aryl-amine mutagenicity using quantum mechanically derived descriptors alone

In a previous report, we determined that for aryl-amines, a case in which reactivity is a principle mode of toxicity, a quantum mechanics reaction energy provides an excellent classifier of Ames+ and Ames- compounds across multiple

aryl-amine data sets<sup>2</sup>. The principle of reactivity has been included in models using energies of the HOMO and LUMO orbitals calculated with a fast semi-empirical quantum method such as AM1 or PM3<sup>346566</sup>. The HOMO energy correlates with the ionization potential, or the energetic cost of losing an electron, while the LUMO correlates to electron affinity, or the gain of an electron. Good performance using these descriptors has been achieved for small sets of aryl-amines with only a few terms in linear classification and regression models<sup>35</sup>. The HOMO energy also does surprisingly well in discriminating Ames+ amines (Figure 5) despite the fact that it does not represent a reaction.

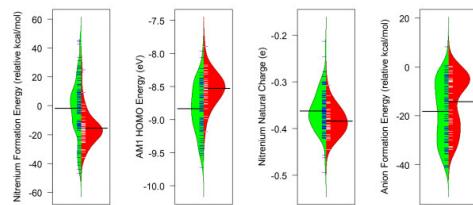


Figure 50.5: Figure 5. Beanplots of four QM descriptors considered in our study

**Beanplots of four QM descriptors considered in our study.** The beanplot is a way to show all data while also conveying a sense of the distribution. In this case, the Ames- points are plotted to the left of the center of each of the four plots with a green distribution and dark points; Ames+ data points are plotted to the right of the center in each plot and shown as white whiskers on a red distribution plot. The mean of each distribution is given as a long dark line. Reaction energies are given relative to aniline.

A number of groups have also studied the utility of studying the reactions of aryl-amines to understand mutagenicity<sup>23356768</sup>. It was determined that the most statistically significant factor for predicting Ames toxicity was the reaction energy for forming the reactive intermediate, the nitrenium ion, from the aryl-amine<sup>23</sup>. This simple descriptor alone can provide a useful prediction of mutagenicity<sup>36768</sup>. These energies are dependent on 3D conformation and the electronic spin state of the reactive intermediate and thus require care to ensure the calculated value is accurate. Using this reaction energy for all Novartis aryl-amines was initially disappointing since good to excellent performance was observed in previous reports for other datasets, in addition to our prediction of external sets gathered for our testing. Upon closer examination, it was clear that most of the sets did not have a uniform distribution up to the range of molecular weight of final pharmaceutical compounds and natural products that comprise a significant portion of the Novartis set. As shown in Figure 6, the performance was much lower for molecules with higher molecular weight in Set A (orange dotted line). Considering that the principle toxicity mechanism of aryl-amines requires metabolic activation, one possible explanation is that larger molecules have more selectivity in metabolic enzymes. Anecdotally, smaller molecular fragments that present themselves as impurities, degradation products or metabolic products were the most common aryl-amine Ames problem at Novartis. Therefore, prediction of lower molecular weight, reagent-like aryl-amines were the principal interest.

Other groups have introduced other quantum mechanics descriptors for aryl-amines in addition to nitrenium forming reactions<sup>2367686970</sup>, including the charge on the nitrenium ion nitrogen<sup>68</sup>, relative energy of anion formation and relative iron complexation energy in a CYP1A2 binding site model<sup>70</sup>, and finally reaction energy for aminyl radical formation, another species that could be produced in the cytochrome systems and has been associated with DNA damage<sup>71</sup>. Some of the reactions that have been used are summarized below in Equations 1-5 and these have been compared for Set A. The Pearson correlation matrix in Table 3 shows that all of the nitrenium forming processes represented by Equations 1-3 are closely correlated and all provide good discrimination. The area-under-the-curve (AUC) for the ROC plot for each of these parameters is given in Table 3 and shown graphically in Additional file

<sup>65</sup> Mechanistic QSAR of aromatic amines: New models for discriminating between homocyclic mutagens and nonmutagens, and validation of models for carcinogens

<sup>66</sup> QSAR models for discriminating between mutagenic and nonmutagenic aromatic and heteroaromatic amines

<sup>67</sup> An in Silico Method for Predicting Ames Activities of Primary Aromatic Amines by Calculating the Stabilities of Nitrenium Ions

<sup>68</sup> Ultimate Carcinogenic Metabolites from Aromatic and Heterocyclic Aromatic Amines: A Computational Study in Relation to Their Mutagenic Potency

<sup>69</sup> Deprotonation and hydride shifts in nitrenium and iminium forms of aminoimidazole-azaarene mutagens

<sup>70</sup> Explanation for Main Features of Structure-Genotoxicity Relationships of Aromatic Amines by Theoretical Studies of Their Activation Pathways in CYP1A2

<sup>71</sup> Purified prostaglandin synthase activates aromatic amines to derivatives that are mutagenic to *Salmonella typhimurium*

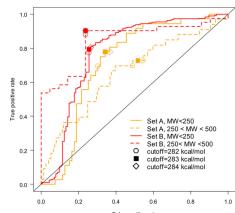


Figure 50.6: Figure 6. ROC curve for using the single parameter nitrenium formation energy for aryl-amine sets A and B

**ROC curve for using the single parameter nitrenium formation energy for aryl-amine sets A and B.** Using the reaction energy for nitrenium intermediate formation ( $\text{RNH}_2 \rightarrow \text{RNH}^+ + \text{H}^-$ ) as a cutoff (labeled by shapes as indicated in the legend) for Novartis and External aryl-amines for the subsets with  $\text{MW} < 250 \text{ g/mol}$  (solid line) and  $250 < \text{MW} < 500 \text{ g/mol}$  (dashed line).

The best single parameters include the nitrenium formation energy reactions and the AM1 HOMO orbital energy. The Ames+ compounds have values tightly clustered in these two parameters as shown in the beanplots in Figure 5 and show the expected relationship to the barrier of forming the nitrenium intermediate. Larger HOMO orbital energies of the amine and lower reaction energies for forming the reactive nitrenium ion would make it easier to form the intermediate. Ames+ amines tend to have a more negative charge on the nitrenium nitrogen, which has been presented previously<sup>68</sup>, but the relationship is clearly not as strong. As suggested in a recent article<sup>70</sup>, we looked at the anion formation energy (Equation 5) and though on its own it has little discrimination as shown in Figure 5 and its AUC in Table 3, it appears to provide a useful complement to the nitrenium formation energy. Higher sensitivity at equivalent false-positive ratios in the 80-87% sensitivity region of the ROC curve (Additional file

Out of all of the QM parameters, the most useful parameter by PLS loadings and Random Forest variable importance (top ranked in all runs) using all of the data was the nitrenium formation energy (Equation 1). This particular reaction is also the easiest to calculate out of Equations 1-3 since it reduces the number of atoms in the system compared to losing -OH or -OAc as the leaving group (Equations 2 and 3). While HOMO energy has a high correlation with nitrenium formation energy (0.84), the nitrenium formation energy provided better overall performance in the 70-84% sensitivity range.

### 50.4.3 3.3 Predicting aryl-amine Ames test results using multivariate statistics

Multi-dimensional statistical models improving upon the performance of the nitrenium formation energy parameter alone were difficult to construct. A number of available approaches including  $k$ -Nearest-Neighbors (\*k>NN), random forest, partial least squares (PLS), support vector machines (SVM), and PLS with discriminant analysis provided similar performance for Set A. A comparison of these methods and other approaches to modeling Ames toxicity when all mutagens are included have already been presented in other studies<sup>36</sup>. We have chosen to focus on PLS and random forest analysis of the aryl-amine data for further discussion because of the interpretability of PLS, the ability to include a large number of correlated variables, and the straightforward assessment of the importance of variables.

Figure 7 shows the receiver operator characteristic (ROC) curve averaged over true-positive rates for 100 PLS models at identical false-positive rates for Set A (left) and Set B (right) using 1, 2, and 3 components of a PLS model. As summarized in Table 4, the performance of the method on the Novartis set, Set A, was highly variable and significantly poorer than for Set B. The performance on the test set decreased dramatically when adding a second component leading to a decrease in average AUC of 0.12. The same approach for the external set, Set B, resulted in a significantly higher AUC performance in the test set of 0.79, 0.80, and 0.81 for a 1-, 2-, or 3-component PLS model respectively. This was higher than the test set performance of Set A by 0.17 for the PLS models in the test set. After 2 components, the test set accuracy at a cutoff chosen in the training set to produce 80% sensitivity began to decrease for Set B. The random forest out-of-bag model performance on the test set for Set B averaged over 100 runs was significantly better than the 2-component PLS model. The performance of a random forest model for Set A was almost identical to the 1-component PLS model, so the AUC performance difference between Set A and Set B random forest models was an even higher 0.24.

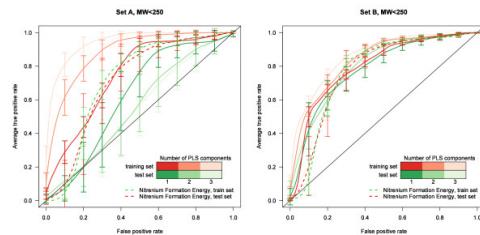


Figure 50.7: Figure 7. Averaged ROC curves of PLS models for aryl-amine sets ( $MW < 250 \text{ g/mol}$ )

**Averaged ROC curves of PLS models for aryl-amine sets ( $MW < 250 \text{ g/mol}$ ).** Performance using 1, 2, or 3 components using 100 randomly sampled test and training sets is shown with Set A on the left and Set B on right. The darker shades of lines have fewer components (see scale) and the green ROC curves are for the 100 random test sets (30% sample) and the red for 100 training sets (remaining 70% of set) both averaged over identical false positive rates. Error bars represent standard deviation.

For both Set A and Set B, the multiple-variable PLS models offered an improved prediction over using nitrenium formation energy alone (dashed line) in the training set but not in the test set. The performance of the Set A PLS model on the test set was much worse on average than using this single parameter. The model in Set B was slightly better but unfortunately, most of the performance increase over the nitrenium formation energy (0.03 in AUC) was in a low-sensitivity region of the ROC curve (< 50% true positive rate). The prediction in this range was not considered to be useful for excluding Ames+ fragments. These results are frustrating but provoked thought about why the molecules commonly used in the literature are different and easier to model.

In an attempt to address the problem of overfitting in this PLS model, a smaller selection of variables was chosen guided by the PLS loading weights, Pearson correlation between variables, and variable importances from a random forest model of the set. The weights were averaged over the 100 models and the largest 30 mean loading weights were used. Table 5 shows the variable loading and jack-knife significance testing run in the PLS cross-validation as well as the mean decrease in Gini coefficient over all trees for the random forest model built with the widest selection of parameters. Two additional descriptors (the Balaban j index<sup>72</sup> and density) are given, which were suggested by random forest importance measures and their low correlation with the other descriptors. The Balaban index was also identified as a discriminating variable in a previous investigation of aryl-amines and depends partly on the number of rings<sup>3</sup>.

The first principal component included the nitrenium formation energy and other descriptors relating to electrostatics, hydrophobicity, and indirect properties such as the number of atoms. The variable  $\chi_{\text{C}}$  ( $^0\chi_{\text{C}}$ ),<sup>73</sup> is a valence-modified carbon atom connectivity index which depends on the number of carbon atoms in the structure and how many non-hydrogen atoms are connected to them.  $a_{\text{count}}$  and  $a_{\text{nH}}$  are simply the number of atoms and hydrogens respectively. GCUT\_SLOGP calculates log P based on atomic contributions and a modified graph distance<sup>74</sup>, while BCUT\_SMR calculates the molar refractivity based on atomic contributions and bond order<sup>75,76,77</sup>. Q\_VSA\_POS is the sum of atomic contributions to van der Waals surface area where the sum of partial charges of the atoms are positive<sup>78</sup>, and density is the molecular weight divided by total van der Waals volume.

The most interpretable variables in the second component for Set B related to flexibility and included the number of rotatable bonds and number of rotatable single-bonds,  $b_{\text{rotN}}$  and  $b_{\text{1rotN}}$  respectively, the Kier flexibility parameter<sup>79</sup>, abbreviated here as KierFlex. The number of oxygen atoms and a fingerprint bit associated with an aryl-amine substructure was also significant.

Using just the first component parameters shown in bold in Table 5 resulted in less decrease in performance between training and test sets and decreased performance by less than 0.03 AUC. Fitting all data led to an intermediate perfor-

<sup>72</sup> Highly discriminating distance-based topological index

<sup>73</sup> Nature of structure-activity-relationships and their relation to molecular connectivity

<sup>74</sup> Prediction of Physicochemical Parameters by Atomic Contributions

<sup>75</sup> Metric Validation and the Receptor-Relevant Subspace Concept

<sup>76</sup> Molecular identification number for substructure searches

<sup>77</sup> A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix

<sup>78</sup> Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges

<sup>79</sup> The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling

mance between the training and test sets as would be expected. A random forest model using only these descriptors performed much better than one using all of the potential descriptors for Set A, and for Set B this approach had similar but slightly lower performance. The likely overfitting in the random forest model was quite surprising and indicates a tendency for many of the parameters to introduce conflicting results. These results are summarized in Table 6 and the full ROC curves are shown in Figures 8 and 9. The single parameter nitrenium formation energy can met or exceeded the performance of PLS models that were given far more information. It was also able to perform well on the challenging Novartis set.

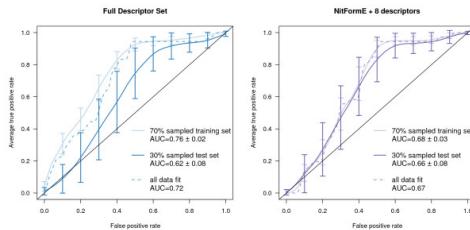


Figure 50.8: Figure 8. Averaged ROC curves for Set A aryl-amine PLS models

**Averaged ROC curves for Set A aryl-amine PLS models.** The plot on the left is for the model built with a full descriptor set and the plot on the right is for a limited descriptor set. The left plot shows the averaged (identical false positive values) ROC curves for the test (dark line) and training (lighter line) sets of 100 PLS models built on a random 70% sample of the Set A aryl-amine data (MW < 250 g/mol) and the performance of a PLS model built on all of the data as a dashed line with all non-zero-variance descriptors. The right plot uses only 9 descriptors including nitrenium formation energy.

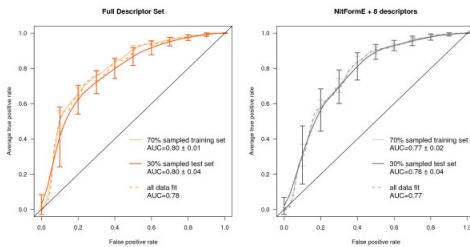


Figure 50.9: Figure 9. Averaged ROC curves for Set B aryl-amine PLS models

**Averaged ROC curves for Set B aryl-amine PLS models.** The plot on the left uses a full descriptor set while the plot on the right is for a model using a limited descriptor set. The left plot shows the averaged (identical false positive values) ROC curves for the test (dark line) and training (lighter line) sets of 100 PLS models built on a random 70% sample of the Set B aryl-amine data (MW < 250 g/mol) and the performance of a PLS model built on all of the data as a dashed line with all non-zero-variance descriptors. The right plot uses only 9 descriptors including nitrenium formation energy.

#### 50.4.4 3.4 Cross set performance-training with Set A and testing Set B and vice versa

In a further attempt to characterize the differences in Set A and Set B, the sets were used as a test set for a model built from the other set. The results of this experiment are shown in Table 7 and Figure 10. The difference in performance was quite instructive and shows that the performance of Set B is less able to extrapolate to the aryl-amines in Set A than vice versa. The performance of the Set A model was actually better for Set B data than for the data used to train it while a model based on Set B had clear difficulty in predicting Set A. This can be quickly seen in Figure 10 which shows the ROC curve for Set B predicted by a PLS model built on Set A (solid red line, left graph) and the ROC curve for Set A predicted by Set B (solid orange line, right graph). The performance of the Set A model on Set B (0.73) was almost identical by AUC to the Set A performance (0.72), representing a decrease in AUC of 0.07 from the Set B model performance. In fact even the 9-descriptor Set A model gave a performance of 0.73 for Set B. PLS models

built on Set B performed extremely well on Set B with AUCs around 0.8. However, when these models were applied to Set A, the performance was markedly worse and the 9-descriptor model performed much better than the model with all of the descriptors. The unscaled PLS scores are shown in Figure 11 in the form of a boxplot for each model. The Set A scores are broadly and almost normally distributed in the Set A model encompassing all of the scores of the Set B data, mostly in the middle 50% of the data (interquartile range). However, the Set B data is not as centered in the model and most of the Set A data is outside the middle 50% of the Set B scores.

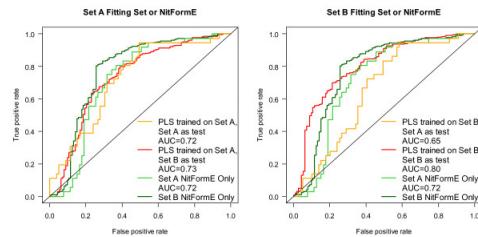


Figure 50.10: Figure 10. Extrapolation from one aryl-amine data set to another: cross-set performance

**Extrapolation from one aryl-amine data set to another: cross-set performance.** Comparison of the performance obtained when a PLS (1-component) model is fitted using all of the data in Set A and used to predict Set B (left plot, solid red line) to that obtained when fitting to the data in Set B and using that model to predict the data in Set A (right plot, solid orange line). The performance for the training set is also shown in dashed lines and the performance of using only the nitrenium formation energy is shown with a dot-dashed line.

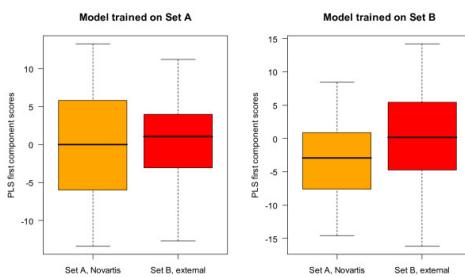


Figure 50.11: Figure 11. Extrapolation from one aryl-amine data set to another: cross-set Tukey Boxplot of the distribution of scores

**Extrapolation from one aryl-amine data set to another: cross-set Tukey Boxplot of the distribution of scores.** Distribution of scores obtained for Set A (orange box) and Set B (red box) when applying a model trained on all of the data in Set A (left) or a model trained on all of the data in Set B (right) as a measure of outliers and domain.

### 50.4.5 3.5 Performance of a commercial model on aryl-amine data-TOPKAT

The pre-built mutagenicity prediction model available to us in TopKat<sup>80</sup> was explored as a possible prediction method. The performance of the prebuilt model on the 327 aryl-amines in Set A was quite poor with an AUC of the ROC curve of 0.59 for the molecular weight < 250 g/mol subset and 0.61 for the molecular weight 250 g/mol subset. The model provides the Tanimoto similarity with the most similar compound used to construct the model as one way to assess model applicability. Set A has an average closest Tanimoto distance of  $0.41 \pm 0.14$  for aryl-amines less than 250 g/mol molecular weight and  $0.57 \pm 0.07$  for those between 250 and 500 g/mol. At least a large portion of the aryl-amine data in Set B was used to build the TOPKAT model and the aryl-amines in this set have an average Tanimoto distance close to zero and a fantastic AUC performance of 0.92 for MW < 250 g/mol ( $N = 398$ ) and 0.997 for MW 250 ( $N = 62$ ). Although it could be argued that these models require retraining when applied to data far from the training set, such data are often not available. A simple retraining using a three-fold cross-validation experiment, resulted in

<sup>80</sup> TOPKAT

only marginal improvement in performance for the Novartis set with AUCs 0.60, 0.64, and 0.69 achieved in the three test sets over the AUCs of 0.58, 0.63, and 0.57 obtained for the default model for the respective sets. Most of the improvement in the ROC curve was in the range of more than 50% false positive performance, which would not be considered a useful range. The ROC curves for these investigations are shown in Figure 12. The good performance for the aryl-amines in Set B suggests that the aryl-amine substructure alone is not problematic in developing these models. Previous publications have not separated the performance by substructure, so it was unclear that this would be true.

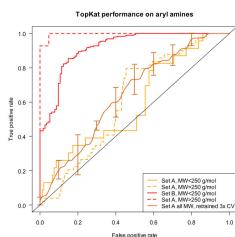


Figure 50.12: Figure 12. Performance of the TopKat Ames mutagenicity prediction module on aryl-amines

**Performance of the TopKat Ames mutagenicity prediction module on aryl-amines.** Performance of the default, prebuilt TopKat Ames mutagenicity model on Set A (orange) and Set B (red) for MW < 250 g/mol (solid line) or for 250 < MW < 500 g/mol (dashed). Additionally, the vertically averaged performance of a 3-fold random cross-validated retraining of the TopKat model using Set A is shown in brown with standard deviation error bars.

#### 50.4.6 3.6 Modeling Ames test results for all substructures

Given the difficulty of addressing aryl-amines, we began to search for reasons the set would be more difficult and if the result would be true for more than just this subspace. Literature reports have provided excellent results for benchmark sets containing all mutagens and small collections of aryl-amines or nitroaromatics. Even better performance could be obtained using multiple models based on the applicability domain of a mutagen under consideration such as Sushko et al.<sup>40</sup> for multiple substructures and Leong et al.<sup>41</sup> for just the aryl-amine substructure. Though surveys of the poor performance of pre-built commercial model performance on proprietary sets has been presented, reports on models of large proprietary sets and delineation of substructure seemed to be lacking. A classification model given a collection of distinct features strongly associated with mutagenicity would be expected to perform better than a model missing such clear-cut mutagenic features such as nitroaromatics mentioned previously.

Table 8 and Figure 13 describes the performance of 2 global models, the TopKat pre-built commercial model and a random forest model built from all data in Sets C, D, E, and F. For clarity, substructure ROC plot performance is shown only for the TopKat in the left plot. Removing molecules with the typically mutagenic polycyclic aromatic, aryl-amine, and nitroaromatic substructures resulted in significant performance decreases in both models in both Set C (Novartis, orange, solid line to orange, dashed line) and Set D (Hansen et al., red, solid line to red, dashed line). The decreases in performance were greater for the TopKat model and for Set D. The global random forest model contained more training data which improved the performance on Set D compared to TopKat, and Set C had fewer of these mutagenic substructures as was presented in Figure 2. The nitroaromatic mutagenic substructure had much better performance in the TopKat model and accounts for over 10% of Set D. However, the nitroaromatic subset in the random forest global model and the aryl-amine and polycyclic aromatic mutagenic substructure performance in both models were equivalent or slightly worse than the overall performance.

The performance of the random forest model built on the global set was similar to the performance of local random forest models built on the individual sets for Sets C and D. It is important to note the extremely high performance for the Kazius set which is a large portion of the training data in the TopKat method. The random forest model constructed from all of the data also did well on this set which again indicates that it is inherently a simpler set to model using the commonly used descriptors. This all-data model has good performance across the entire chemical space map as detailed in Figure 14 where greens indicate a successful prediction (over 500 trees) while yellow indicates an equivocal prediction and oranges and reds would be expected to give the wrong prediction. In fact, a random forest model built with just Set D provided a fairly good prediction but gave more equivocal results in the regions occupied mainly by

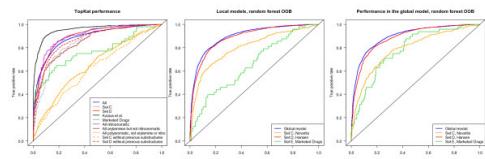


Figure 50.13: Figure 13. ROC curve performance for models on all-substructure data sets

#### ROC curve performance for models on all-substructure data sets.

Ames mutagenicity model on Sets C (orange), D (red), E (green), and F (black) as well as the performance for particular substructures in all sets (blue): nitroaromatics (purple), aryl-amines (brown), or polyaromatic (gray). Dotted lines for Sets C and D show performance after removing these substructures. The center plot shows the out-of-bag performance of random forest models built on Sets C (orange), D (red), and E (green) and the global model (blue) when they are used as the training set. The right plot shows the Set C, D, or E subsets of the global out-of-bag performance (blue) of a random forest model built on all of the data.

Novartis compounds. As might be expected from the good performance of TopKat on set D and poor performance on the Novartis set, Figure 14 shows that most of the cells with Novartis compounds would be misclassified by the TopKat model (red cells) but also gets other regions wrong such as the lower left-hand corner. Unlike the case of the aryl-amines, the Novartis all-substructure model does not perform well on regions occupied mostly by the other sets. The performance map provides almost a perfect opposite to TopKat, though the lower-left-hand corner is still difficult to predict with many equivocal cells. Looking at the Ames+/Ames- coloring of the cells in the left-hand plot gives an idea of why this might be possible. This region of substructure space has many cells that contain a mix of Ames+ and Ames- compounds.

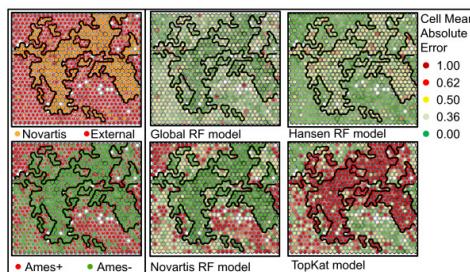


Figure 50.14: Figure 14. Performance of global models mapped to chemical space

**Performance of global models mapped to chemical space.** Self-organizing map of Sets C, D, E, and F (from Figure 4) again colored by whether the molecule is in the Novartis set (Set C, orange) or is Ames+ (red) in the left box. In the right box, the SOM

is colored by the mean absolute error of the predictions in the cell for the indicated model. Dark green indicates correct classification, yellow an equivocal prediction, and red an incorrect prediction. Cell mean absolute error is defined as the difference between the predicted probability of being mutagenic and the experimental class (0 or 1).

## 50.5 4. Conclusions

In this article we have shown that there are significant differences in the physicochemical and biological properties of compounds used in drug discovery and those in compiled Ames test results from the literature. This includes molecular weight, substructure distribution, and the percentage of mutagenic compounds in the data set. This is important to communicate, as much of the literature data is being used to test prediction methods as well as playing a role in current testing strategy debates. The compounds in the Novartis test results are mostly drug precursor molecules, while literature mutagenicity results are often petrochemicals and pesticides of primary concern as environmental pollutants. The size and complexity of the molecules tested at Novartis was significantly larger on average than that of molecules included in external sets, as visualized by distributions in molecular weight. Chemical functional groups or substructures that have a high association with mutagenicity determined from the literature data are largely

absent in the Novartis set taking away a valuable discrimination feature. Additionally, the proportion of mutagens in external sets is higher and disturbingly close to 50% as might occur from successive culling for balanced model development. As a result of these factors, many drug discovery molecules are outside the applicability domain of pre-built commercial models. The data is also more difficult due to lack of strongly associated structural features and would lead to worse performance of these statistical models if they were included in the training set. Therefore these models cannot provide adequate performance to predict, let alone, avoid a positive Ames test. The Ames test, as well as other genotoxicity tests, continue to be a significant problem in drug discovery, and companies should work together to share data with the wider community of scientists and organizations. The best-validated and best-performing prediction available for low molecular weight aryl-amines is still a quantum-mechanics reaction energy representing the formation of the nitrenium ion. Effective predictive models could be built for all-substructure sets using the random forest methodology and commonly available 2D descriptors and chemical fingerprints. Performance was still significantly lower for molecules from Novartis and marketed pharmaceuticals. Despite extensive work in the area of predicting this particular toxicity, work in designing more difficult test sets and more adaptable models is still necessary.

## 50.6 Abbreviations

Ames+: positive Ames test result; Ames-: negative Ames test result; PLS: partial least squares; ROC: receiver operator characteristic; AUC: area-under-the-curve; N: number of points; NitFormE: nitrenium formation energy; SOM: self-organizing map; OOB: out-of-bag; HOMO: highest occupied molecular orbital; LUMO: lowest occupied molecular orbital; IQR: interquartile range

## 50.7 Competing interests

The authors are employees of a pharmaceutical company and must comply with regulatory guidelines for genotoxicity; however, all analysis is based on the 2-strain non-GLP screening version of the test. Novartis licenses software for a fee from Schrodinger, Accelrys, CCG, and Gaussian, Inc.

## 50.8 Authors' contributions

PM wrote, developed the idea behind the work, and performed the statistical analysis in the paper. CS substantially contributed to the idea for the paper, discussion, statistical experiments, editing, and technique suggestions. LW contributed significantly to the idea for the paper, guidance, editing, and discussion. All authors have reviewed and approved the final manuscript.

## 50.9 Acknowledgements

We would like to thank the devoted help and computational resources provided by the NIBR IT Scientific Computing group and especially Steve Litster, Michael Derby. P.M. is a NIBR postdoctoral fellow and thanks the NIBR Education Office for funding.



# AUTOMATED ANNOTATION OF CHEMICAL NAMES IN THE LITERATURE WITH TUNABLE ACCURACY

## 51.1 Abstract

### 51.1.1 Background

A significant portion of the biomedical and chemical literature refers to small molecules. The accurate identification and annotation of compound name that are relevant to the topic of the given literature can establish links between scientific publications and various chemical and life science databases. Manual annotation is the preferred method for these works because well-trained indexers can understand the paper topics as well as recognize key terms. However, considering the hundreds of thousands of new papers published annually, an automatic annotation system with high precision and relevance can be a useful complement to manual annotation.

### 51.1.2 Results

An automated chemical name annotation system, MeSH Automated Annotations (MAA), was developed to annotate small molecule names in scientific abstracts with tunable accuracy. This system aims to reproduce the MeSH term annotations on biomedical and chemical literature that would be created by indexers. When comparing automated free text matching to those indexed manually of 26 thousand MEDLINE abstracts, more than 40% of the annotations were false-positive (FP) cases. To reduce the FP rate, MAA incorporated several filters to remove “incorrect” annotations caused by nonspecific, partial, and low relevance chemical names. In part, relevance was measured by the position of the chemical name in the text. Tunable accuracy was obtained by adding or restricting the sections of the text scanned for chemical names. The best precision obtained was 96% with a 28% recall rate. The best performance of MAA, as measured with the F statistic was 66%, which favorably compares to other chemical name annotation systems.

### 51.1.3 Conclusions

Accurate chemical name annotation can help researchers not only identify important chemical names in abstracts, but also match unindexed and unstructured abstracts to chemical records. The current work is tested against MEDLINE, but the algorithm is not specific to this corpus and it is possible that the algorithm can be applied to papers from chemical physics, material, polymer and environmental science, as well as patents, biological assay descriptions and other textual data.

## 51.2 Background

Significant portions of the biomedical literature refer to chemical structures. For example, metabolites and small signaling molecules are crucial to life and well-studied, while many natural and synthetic products are examined in the context of drug discovery. The accurate identification and annotation of chemical names that are topically relevant to literature is a critical first step to establish links between scientific publications and the databases containing information about the chemical structure the name represents (*e.g.*, molecular structures, measured biological activities, and drug information). Currently, manual identification is the preferred method for these chemical annotations as well-trained indexers can semantically understand and rank paper topics as well as recognize key terms; however, when considering the hundreds of thousands of new scientific articles published annually, an automatic annotation algorithm with high precision and relevance is a useful adjunct to manual annotation.

Current studies <sup>12345678910111213</sup> on the text mining of small molecule names focus on the named entity recognition (NER) of chemical descriptors, including systematic chemical names such as IUPAC names and common names. Dictionary and rules (DR) based methods and statistical machine learning (ML) methods are two major approaches in this area. In 1992, Chowdhury and Lynch <sup>12</sup> developed a dictionary and rule based semiautomatic method to convert chemical texts into structure representation by morphological analysis and dictionary lookup. In 1999, Wilbur, *et. al.* <sup>3</sup> compared three NER methods (one rule and dictionary based method and two Naïve Bayes statistical methods) to recognize chemical terms in biological text, and concluded that an integrated method might perform best. Hettne and co-workers <sup>45</sup> generated dictionaries identifying small molecules and drugs in text, and found that a dictionary generated from a reliable single source, ChemIDplus <sup>14</sup> performs as well as a dictionary from combined multiple sources. Wren <sup>6</sup> evaluated a first order Markov Model for its ability to distinguish chemical names from words. Klinger <sup>7</sup> implemented a new machine learning approach based on Conditional Random Fields (CRF) to detect IUPAC and IUPAC-like chemical names in the scientific literature and obtained good performance: an F measure of 85.6% on a MEDLINE <sup>15</sup> corpus. Corbett, Jessop and co-workers <sup>891011</sup> performed studies on chemical name mining on text and developed OSCAR4 <sup>11</sup>, an open source system to identify chemical names in scientific articles. Kolarik and her co-workers <sup>1213</sup> analyzed chemical terminology resources and generated an annotated text corpus for evaluation of dictionaries. Recently, Zhou and his co-workers <sup>16</sup> designed and implemented a chemistry text hybrid search engine to combine both chemistry text and structure searching in literature. Generally speaking, the ML based methods perform extremely well in recognition on IUPAC or IUPAC-like chemical names, but not as well on common names. On the other hand, the DR based methods can identify both IUPAC and trivial names, but it is not possible to identify names not in the dictionary. Nevertheless, both approaches concentrate on the identification of chemical names but focus less on ranking the annotations for relevance, which is a key goal of this study. For example, a chemical mentioned in a metabolic pathway paper may not be the molecule that can trigger the pathway, rather it might be an inactive chemical compound, related to the methodology, or a substrate mentioned in a longer protein name or gene name. Banville <sup>17</sup> addressed a similar problem: how do you find documents of relevance to a chemical instead of simply finding the chemical name present in a document?

Medical Subject Headings <sup>18</sup> (MeSH) annotation, which is performed by trained curators of the National Library

---

<sup>1</sup> Automatic interpretation of the texts of chemical patent abstracts. 1. lexical analysis and categorization

<sup>2</sup> Automatic interpretation of the texts of chemical patent abstracts. 2. processing and results

<sup>3</sup> Analysis of biomedical text for chemical names: A comparison of three methods

<sup>4</sup> A dictionary to identify small molecules and drugs in free text

<sup>5</sup> Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining

<sup>6</sup> A scalable machine-learning approach to recognize chemical names within large text databases

<sup>7</sup> Detection of IUPAC and IUPAC-like chemical names

<sup>8</sup> An Architecture for language technology for processing Scientific texts

<sup>9</sup> High-throughput identification of chemistry in life science texts

<sup>10</sup> Annotation of chemical named entities

<sup>11</sup> OSCAR4: A flexible architecture for chemical text-mining

<sup>12</sup> Chemical names: Terminological resources and corpora annotation. In: European Language Resources Association

<sup>13</sup> Identification of new drug classification terms in textual resources

<sup>14</sup> ChemIDplus

<sup>15</sup> MEDLINE

<sup>16</sup> Chemical-Text Hybrid Search Engines

<sup>17</sup> Mining chemical and biological information from the drug literature

<sup>18</sup> Medical subject headings (MeSH)

of Medicine (NLM) to index and categorize articles in the MEDLINE databases, is a reliable source for users of MEDLINE and PubMed<sup>19</sup> to obtain relevant and accurate scientific term annotations. To aid human indexing of the MEDLINE database, NLM developed an automatic indexing system<sup>20 21 22</sup>, the Indexing Initiative system (IIS), to identify candidate MeSH concepts in papers being indexed, helping to speed the manual annotation of the biomedical literature.

In recent years as the volume of literature has grown, the accuracy and relevance of retrieved information have become key performance indicators of on-line chemical databases. A single query can retrieve many thousands of records, making it essential that the top ranked results are highly relevant to the user. Additionally, it is very useful to link records from one database to those in another. In the NCBI Entrez query system<sup>23</sup>, MeSH plays a vital role in both improving query performance and for making links between databases. For example, the MeSH vocabulary allows for synonym expansion in PubMed queries, precise querying of chemical names, and allows the linking of abstracts to the small molecule records in the PubChem<sup>24</sup> database; however, manual annotation of MeSH onto PubMed abstracts can have a time lag of a few months and other sources of scientific literature may not have MeSH annotation at all. In these situations, it may be useful to have an algorithm for automatic annotation of MeSH terms.

In this article, we present an implementation of an automated chemical name annotation system based on the MeSH controlled vocabulary called MeSH Automated Annotations (MAA). The primary aim of MAA is to reproduce the MeSH term annotations created by curators.

## 51.3 Methods

### 51.3.1 1. Corpus generation

The annotated text corpus is generated directly from MEDLINE, with PubMed identifier (PMID) ranging from 16200042 to 17342794. In order to increase the recall of automated annotation, we only select entries with both title and full abstract available, giving a total of 261,227 MEDLINE abstracts inside the corpus. Each paper in the corpus has been annotated by the NLM indexers. These human annotations of chemical names are used as the “gold standard” for comparison with the various versions of MAA described in this paper. We performed spot checking of randomly selected manual annotations and found they are reliable to be used as standards. However, the NLM indexers’ aim is to annotate topic-related chemical entities, thus the selections depend on the indexers’ understanding of the topic of a paper. It is nontrivial to tell if the unselected chemical entities are valid or not. Nevertheless, an automated annotation system should provide improvements on the possible errors of manual indexing. A randomly selected data set, which contains 26,123 abstracts, was selected to test our MAA program. The remaining abstracts were used as a training set to obtain statistics used to set thresholds for various filters used in the algorithm.

### 51.3.2 2. MeSH chemical dictionary generation

MeSH is a controlled vocabulary thesaurus from the NLM used to help index the biomedical literature. MeSH is organized in a hierarchical tree where each scientific concept is either a node or leaf of the tree. Scientific concepts include a MeSH heading (being the most common name used to refer to the concept), synonyms, and inflectional MeSH term variants. The parts of the MeSH tree associated with chemicals is composed of two parts, the ‘Chemicals and Drugs’ branch of the MeSH hierarchy and an independent set of supplementary concept records (denoted as MeSH substances). Each MeSH substance is mapped to at least one MeSH term. Chemical compounds of relatively recent biomedical interest are either appended to the MeSH tree or added to the MeSH substances. The MeSH chemical vocabularies are used as the basis of our dictionary. In the following text, we will use the phrase ‘MeSH term’ to

<sup>19</sup> PUBMED

<sup>20</sup> The NLM indexing initiative

<sup>21</sup> The NLM indexing initiative’s medical text indexer

<sup>22</sup> A strategy for assigning new concepts in the MEDLINE database

<sup>23</sup> Entrez

<sup>24</sup> PubChem: Integrated Platform of Small Molecules and Biological Activities

refer to any MeSH heading, MeSH term, or MeSH substance under or mapped to the Chemicals and Drugs branch of MeSH.

### 51.3.3 3. Statistical terminology for evaluation of MAA

The objective of our MAA system is to find relevant MeSH terms in abstracts. As mentioned previously, in an abstract there may be many MeSH terms found in the text, but not all of these are related to the topic of the document. The human MeSH indexer annotates these relevant MeSH terms by reading and understanding the subject material. Thus, we compare our MAA system to the manual annotations of the MeSH indexers. In our approach, we intend to reproduce the MeSH indexers' annotation by extracting relevant terms and filtering out unimportant MeSH terms. Using this manual indexing as the standard, the terminologies used for evaluation of MAA are:

True positive (TP) match – A MeSH term found by both MAA and manual indexing.

True negative (TN) match – A MeSH term not found by either MAA or manual indexing.

False positive (FP) match – A MeSH term found by MAA but not by manual indexing.

False negative (FN) match – A MeSH term found by manual indexing but not by MAA.

Note that the false negatives include terms that are not in the title or abstract of the documents as the MeSH indexers have access to the complete document. These terms cannot be found by MAA as the algorithm does not have access to the complete document. In the following figures and discussions, the total FN matches were separated into two groups: the group “In Text, Not Found” refers to MeSH annotations where the MeSH terms are in the abstract but are not found by MAA and “Not in Text” to refer to instances where the MeSH term is not present in the text. In the latter case, terms are typically found in the body of the paper.

Precision and recall are calculated as following:

### 51.3.4 4. Chemical Tokens and Rules

A Chemical token is a string used to build chemical names. In this study, a chemical token dictionary was created to generate chemical morphemes. These chemical tokens are made by dissecting chemical names at white space and other separators. The chemical names are taken from MeSH terms and PubChem Compound synonyms, encompassing over 31 million chemical records. We chose the PubChem as a source of chemical names as it is a large database of small molecule structures (> 30 million), including depositions from many popular chemical databases, such as ChemIDPlus, ChEBI, ZINC, etc. MeSH was selected as it is a comprehensive controlled vocabulary that has been applied extensively in biomedical literature indexing, including the indexing of most PubMed records, making it likely to contain a significant subset of biomedically interesting chemical names. However, the MAA algorithm is not limited to these sources - in particular, a more detailed controlled vocabulary may improve the results of the algorithm. After the tokens are generated, two English novels “*Jane Eyre*” and “*Pride and Prejudice*” are used to filter out common English words from the tokens. Numbers, numerical identifiers, single characters and special characters are removed. Overall, there are total 326,610 chemical tokens stored in our token dictionary. These chemical tokens, along with name decision rules, were used to check if a MeSH term embedded in text is a full name or a sub-string of another name. In MAA, a MeSH term and two tokens before and after the MeSH term are analyzed. If the combination of the MeSH term and the tokens fulfill one of the name decision rules, the MeSH term will be marked as a likely substring of a complete chemical name.

## 51.4 Results and Discussion

There are several steps in the MAA algorithm. The first step is free-text matching of the MeSH vocabulary to the MEDLINE abstracts. To measure the performance of this step and subsequent steps, we compare the results to manual annotations of these abstracts done by the MeSH indexers. In this comparison, we take into account that in some cases the indexer used a more general term (aka a “relative node”) than the precise name of the chemical, such as

“Benzodiazepines” instead of “Diazepam.” Note that it is not possible for the algorithm to find all terms annotated by the indexers as the indexers have access to the complete paper and the algorithm does not.

Subsequent steps in the algorithm attempt to reduce the number of false positives, which are the matches found by the algorithm but not indexers. The first step, the “MeSH Filter” eliminates MeSH records that do not have an associated chemical structure. The second step, “Tokens and Rules”, discards partial matches to terms that follow chemical nomenclature rules or have additional chemical name tokens. The third step, “Protein and Gene Names”, screens out protein and gene names as these names can contain the names of chemicals. Finally, the “TP filter” eliminates matches using MeSH terms that are also common English terms, such as “lead.”

The comparison between the various steps in MAA and manual indexing is depicted in Figure 1 and will be discussed in detail below. From left to right, each group of bars indicates the matching results for different steps in the algorithm. Within each group there are 4 bars, starting from the left:

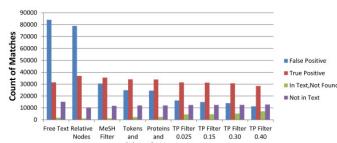


Figure 51.1: Figure 1. Comparison of MeSH automated (MAA) and manual indexing annotations with a series of algorithmic filters added

**Comparison of MeSH automated (MAA) and manual indexing annotations with a series of algorithmic filters added.** From left to right, filters are applied cumulatively. “False Positives” are matches found by the algorithm but not in manual indexing. “True Positives” are found by the algorithm and in manual indexing. “In text, not found” are found in manual indexing, but not by the algorithm. “Not in text” are manually indexed terms not found in the abstract.

1. False positive (blue): MeSH terms found in an abstract by MAA but not found by manual indexing.
2. True positive (red): MeSH terms found in an abstract by MAA and also found by manual indexing.
3. In Text, Not found (green): MeSH terms present in the abstract and found by manual indexing but not found by MAA.
4. Not in text (purple): MeSH terms not present in the abstract but found by manual indexing. As mentioned earlier, some terms are found in the body of a paper and not in the abstract. Since the MAA algorithm does not have access to the body of the paper, it is unable to find these terms. The value shown in the figure is likely an upper bound as it is possible that the algorithm may not find a term due to various potential issues (e.g. punctuation, spelling, unknown synonyms, etc.).

### 51.4.1 1. Free-Text MeSH matching

As displayed on Figure 1, the free-text MeSH string matching, if used to annotate small molecules directly, generates a significant number of false-positive cases compared to the MeSH indexers’ annotations. More than 70% of the MeSH terms appearing in the text were not annotated by indexers (see Table 1 for values and Figure 2 for graphs of the precision, recall and F measure). By inspection, it appears that in most cases the MeSH term may either be a text fragment of another scientific concept (e.g., many protein names include aspects of a chemical name) or the MeSH term is simply present but not relevant to the paper context (e.g., used in a descriptive or comparative sense, as a reagent in an experiment, etc.). Thus, in order to annotate a chemical term accurately, algorithmic filters need to be applied to the free-text matches. In some cases, the indexer missed annotating a valid chemical name. However, this category was not examined as it would have required another standard of truth, which was unavailable for the large set of abstracts considered in this study.

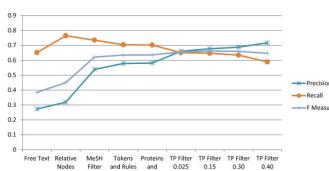


Figure 51.2: Figure 2. The recall, precision and F measure of MeSH automated annotations (MAA) on the titles and abstracts of the test corpus with a series of algorithmic filters added cumulatively

**The recall, precision and F measure of MeSH automated annotations (MAA) on the titles and abstracts of the test corpus with a series of algorithmic filters added cumulatively.**

### 51.4.2 2. Using relative nodes in the comparison

In comparing results between the MAA and the manual indexers, tree-node expansion is applied to the results from the MAA system. The term “tree-node” comes from the hierarchical tree structure of the MeSH thesaurus. Examination of the MeSH annotations in MEDLINE finds that the indexers will sometimes select a higher level node in the MeSH tree than the nodes that correspond exactly to the chemicals mentioned in an abstract. For example, they may select a more generic term “Penicillins” instead of the “Penicillin G” and “Penicillin V” mentioned in a paper. This use of higher level (“super-concept”) nodes also happens for MeSH substances as they are manually mapped to nodes in the MeSH tree. Therefore, for a given MeSH substance or MeSH term, we include its assigned MeSH tree node and/or super-concept node, respectively. This tree-node expansion significantly increases the number of matches between manual indexing and our MAA algorithm (Figure 1), while also increasing the recall and precision due to the increase in the number of true positives (Table 1 and Figure 2).

### 51.4.3 3. Improving free-text MAA by adding filters

#### 3.1. MeSH filter: using terms with associated chemical structure

From the entire set of chemical MeSH terms, MeSH terms representing small molecules are extracted to generate a chemical dictionary. This constraint is implemented by selecting MeSH terms which can map to PubChem (<http://pubchem.ncbi.nlm.nih.gov/>) CIDs, which identify unique small molecule structures. This filter removes almost 65% of the total MeSH terms (see Table 2, the numbers of MeSH terms change from 521 k to 178 k). Filtered terms include protein names, category names, non-specific names and chemical names which cannot be represented by molecule structures. After filtering out terms, we expand the MeSH dictionary by adding 541 chemical formulas of inorganic compounds as synonyms of associated MeSH concepts. All of these chemical formulas are from Wikipedia<sup>25</sup> and were manually verified. This addition increases the ability of MAA to recognize chemicals such as ‘KOH’ and ‘NaCl’ if no compound common name is mentioned in the text. After updating our dictionary, we performed another free-text match and compared the results with manual indexing. The results are shown in the second group of bars labeled “MeSH Filter” in Figure 1. Compared to “Relative Nodes”, the number of FP cases after adding the MeSH Filter drops from 79 K to 30 K, and the number of TP cases decreases less than 1 K. The performance of MeSH filter, indicated by precision, recall and F measure is displayed in Table 1. The precision jumps from 0.32 to 0.54, and loses 0.03 recall. As a result, the F measure has an increase from 0.45 to 0.62. The MeSH filter assures that all extracted MeSH terms by MAA are entries in PubChem. It thus implicitly links chemical names in literature to variant features of the PubChem database, such as chemical structures, properties and bioactivities etc.

#### 3.2. Tokens and Rules: removing false positive annotations by syntactic analysis

MeSH terms that are sub-strings of another entity name is one reason for false-positive annotation. The chemical tokens and chemical name decision rules (introduced in Methods part 4) were used to decide if matched MeSH terms are full names or substrings.

<sup>25</sup> Wikipedia: List of inorganic compounds

Some of the applied rules are listed below:

[1]. If two words in front of a matched MeSH term are both chemical tokens, the MeSH term is treated as a FP annotation. This rule by itself yields a 0.25% increase in precision, a 0.34% decrease in recall and a 0.04% increase in F measure;

[2]. If one word in front of a matched MeSH term is a chemical token, the MeSH term is treated as a FP annotation. This rule by itself yields a 1.3% increase in precision, a 1.6% decrease in recall and a 0.24% increase in F measure;

[3]. If one word behind a matched MeSH term is a chemical token, the MeSH term is treated as a FP annotation. This rule by itself yields a 2.2% increase in precision, a 1.9% decrease in recall and a 0.68% increase in F measure. Note that the F measure is the harmonic average of precision and recall, which is why the change in F measure is not exactly the difference between the change in precision and recall.

Using more than two tokens before and after the MeSH term did not yield any improvements. For example, if the algorithm checks 3 tokens before and after the matching MeSH term, the recall decreases 0.72% and precision decreases 0.87%.

In addition to these name decision rules, we also created several prefix and suffix rules to check whether a matched term is FP annotation. For example, if the token ‘poly’ is the prefix of a MeSH term, this MeSH term is treated as a FP annotation, yielding a 0.12% increase in precision, 0.03% decrease in recall and 0.07% increase in F measure; if ‘ase’ is the suffix of a MeSH term (except ‘release’ and ‘base’) the MeSH term is treated as a FP annotation, yielding a 1.2% increase in precision, 0.15% decrease in recall and 0.75% increase in F measure. These rules were primarily heuristic in nature and were developed by manual examination of the annotations.

It is possible to apply hundreds of rules to increase the precision of MAA. However the recall decreases as each rule applied. It is nontrivial to decide which rules should be used. In the MAA system, we select rules according to the computed F measure. If we obtained a relatively significant positive increment of F measure by applying a rule, the rule was kept.

The following is an actual annotation of a PubMed abstract (PMID 16704345) to show how this filter works:

*benzoyl-CoA 4-hydroxybenzoyl-CoA adenine 4-hydroxybutyryl-CoA ...Related enzymes are the ATP-dependent .*

The bold words are mapped MeSH terms, and the words underlined are chemical tokens found before or after MeSH terms. For example, according to the rules, “adenine” is the complete name and MeSH term “adenine” is just part of this name. Thus, the MeSH term “adenine” is regarded as a false-positive by our MAA program.

In Figure 1, the fourth group of bars indicates the change after adding the “Token and Rules” filter. Compared to previous group of bars (MeSH Filter), the blue bar (false-positive annotation) dropped more than 5000 and red bar (true-positive) only lost 1300 annotations. Please see additional file

Additional file 1

**Details of token generation and rules.** Detailed description of chemical token generation and chemical name decision rules.

[Click here for file](#)

### 3.3. Protein and gene names: removing MeSH terms that are sub-strings of protein, gene and non-chemical MeSH terms

Chemical terms are a common part of protein names, such as “benzoyl-CoA reductase” and “4-hydroxybutyryl-CoA dehydratase” shown above. When these protein names are mentioned in text, it is likely that the topic of the paper is the protein instead of the prefix chemicals. To address this issue, we created a group of “negative vocabularies” to collect names that contain MeSH terms as sub-strings. In the MAA algorithm, if a term in the negative vocabularies is found in text, then its sub-string will not be annotated if this sub-string is a MeSH term. The protein and gene names are collected from MeSH and the NCBI Entrez Gene database. The performance of this method depends on the completeness of the negative vocabularies. It is not possible to construct a complete dictionary, as new names are generated every day. In Table 1, we can see that this filter results in only a small increase of the F measure at best.

This is because the “token and rules” filter and the “protein and gene” filter are not mutually exclusive: some rules in section 3.2 already remove many protein names. If “tokens and rule” and “protein and gene name” filters are applied independently on the same corpus, the former will yield 2.6% more precision and 0.5% more F measure. This result is possibly due to the fact that the “token and rule” filter attempts to be a superset of the “protein and gene name” filter. Nevertheless, the protein and gene name rule is still useful in removing false positive matches for certain protein names.

### 3.4. TP filter: removing MeSH terms with low TP ratios

Some MeSH terms, such as the dental sealant “Conclude” (also known as “Concise”), have a high false positive rate due to nonspecific matching. These terms are filtered out to improve match statistics. To do this, we pre-calculated the true positive ratios of each MeSH term using free-text string matching on the training set. A binary value (1 or 0) was assigned to each MeSH term to indicate if it exists or not in the MEDLINE abstract. If a term was mentioned multiple times in an abstract, it was still counted as 1. The ratio of TP annotation for a specific MeSH term was calculated by the number of times the term was applied during manual indexing divided by the count of abstracts with free text matches. This ratio is used to measure the propensity of a MeSH term to be correctly annotated in text. Some MeSH terms with their TP ratios are listed in Table 3. Common chemicals such as ‘water’ and ‘glucose’ tend to have a less than 50% TP ratio. The term ‘lead’ has only 11% TP ratio, which indicates in only 11 out 100 papers, ‘lead’ is indexed as a chemical element. Additional term types with low TP ratios include homonyms of common English words, such as ‘link’ and ‘monitor’, or acronyms such as ‘CI-2’ that have only a few characters. Using the TP ratio, one may set up a tunable threshold to eliminate non-specific MeSH terms in automatic annotation.

Once a threshold ratio is selected, MeSH terms with a ratio lower than the threshold will not be annotated on the testing data set. Selecting a reasonable threshold will remove false-positive annotations and increase the precision of MAA while not significantly reducing the recall. For example, if the threshold is set to 0.025, there are only 401 total MeSH terms eliminated, but nearly 8297 FP annotations are removed (in Figure 1, this difference is shown by the blue bar when going from ‘Proteins and Genes name’ to ‘TP ratio 0.0025’), while 2466 TP annotations are lost (In Figure 1, this difference is shown by the red bar when going from ‘Proteins and Genes name’ to ‘TP ratio 0.025’). In our study, the thresholds are adjusted from 0.025 to 0.4 to show the trade off in recall as precision increases. Thresholds larger than 0.5 were not evaluated, since the MAA will lose more TP annotations than FP annotations. The best threshold ratio by F measure is between 0.1 and 0.2 (see Figure 2). This TP ratio filter provides a degree of tunable accuracy for the MAA system.

#### 51.4.4 4. Term position in the text

The title is often a summary of a paper. In the abstract, the author often mentions objectives in the first sentence (FT) and conclusions in the last sentence (LT). The appearance of a chemical name in these parts of an abstract is likely to indicate a high degree of relevance. We performed MAA on the title, FT and LT of the abstracts and then again just on the title of the abstract to see if we could obtain higher precision. The results are presented in Figure 3 and Table 1. In the title-only annotation, MAA could provide the 96% precision if the TP filter threshold is set to 0.4. At this filter level, there is greater than 27% recall on the corpus. Eliminating the TP filter for title-only annotation yields 91% precision with 33% recall. This may be because all words in the title are relatively important, reducing the necessity of the TP filter to remove non-specific MeSH terms. For an information retrieval task that requires a high degree of specificity, MAA on the title-only is a reasonable selection. Including the FT and LT in MAA yields less precision than title alone MAA, but with better precision than MAA on the entire abstract.

#### 51.4.5 5. Comparison with other studies

In Table 4, the performance of MAA is compared to similar studies that matched MeSH to text. The top two rows in Table 4 list results from Hettne<sup>4</sup> and Kolarik<sup>13</sup>. As a gold standard, Hettne and Kolarik used a text corpus with Kolarik’s manual annotations of chemical terms, which were not restricted to the MeSH vocabulary. We applied MAA to Kolarik’s testing corpus (2009 version, containing 100 full abstracts). The dictionary used in MAA is a

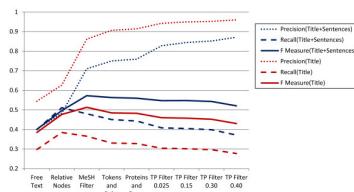


Figure 51.3: Figure 3. The recall (dashed line), precision (dotted line) and F measure (solid line) of MeSH Automated Annotation (MAA) on titles only and on title and selected abstract sentences of the test corpus with a series of algorithmic filters added cumulatively

**The recall (dashed line), precision (dotted line) and F measure (solid line) of MeSH Automated Annotation (MAA) on titles only and on title and selected abstract sentences of the test corpus with a series of algorithmic filters added cumulatively.**

Sentences = first and last sentences of abstract.

combination of MeSH tree chemical terms (MeSH C) and MeSH substances (MeSH S), but both Hettne and Kolarik separated these vocabularies. We first perform free-text matching of the MeSH dictionary to the Kolarik’s corpus. The performance is very similar to Kolarik’s, which were also performed using free-text matching. Then we applied each filter cumulatively as we did in our testing set in Section 3 of this paper. Once the TP filter threshold was set to 0.4, the performance we obtained is quite similar to those of Hettne’s work, in which he used a “term disambiguation” pipeline to filter out some MeSH terms.

MAA has a precision range from of 0.44 ~ 0.79 with different filters, which is better than Kolarik’s precision range of 0.34~0.44 (for MeSH C and MeSH S, respectively). However, the MAA gives a higher recall range (0.23 ~ 0.37) than Hettne’s (0.22~0.07) or Kolarik’s (0.27~0.10). The best F measure which MAA generated is 0.43, which is better than the F-measure taken from either work (maximum of 0.34). Overall, the MAA results are closer to Kolarik’s results if no filters applied and Hettne’s results if TP threshold set to 0.4. However, as shown in Section 3.4, the higher TP threshold doesn’t necessarily produce the better performance as ranked by F measure. When examining our 26123 abstracts testing set, the best performance of MAA was obtained when TP threshold was set to 0.15 and, when examining Kolaik’s corpus, it was without applying the TP filter. This is consistent with results on our test corpus; while the TP filter significantly increases precision, it does so at the cost of recall.

The bottom rows of Table 4 show results from our MAA system and the Medical Text Indexer (MTI), which was developed for NLM’s Indexing Initiative system (IIS) and whose goal was to provide suggested annotations to MeSH indexers. The results of MTI are not restricted to chemical names, so we cannot directly compare the results of MTI to MAA, but we include the results for reference. When MTI lists up to 25 recommendations for each article from a 273 articles corpus, it provided a recall of 0.55 and a precision of 0.29.

## 51.5 Conclusion and Future Application

In this article, we present the design and implementation of an automated chemical name annotation system (MAA). This annotation system uses the MeSH controlled vocabulary applied to biomedical abstracts from MEDLINE. To avoid false positive annotations, we implemented filters to allow for tunable accuracy. The maximum precision obtained was 96% with 28% recall when performing MAA on titles of the abstracts. The best performance of MAA as measured with the F statistic was 66%, which required applying all filters (including the FP filter with a threshold of 0.15). The MAA system compared favorably to other chemical name retrieval studies. The current work is tested against MEDLINE, but the algorithm is not specific to this corpus and it is possible that the algorithm can be applied to papers from chemical physics, material, polymer and environmental science, as well as patents, bioassay descriptions and other textual data. Accurate MeSH annotation and text mining can help researchers not only to identify important chemical names in abstracts, but also match unindexed and unstructured texts to chemical records.

## 51.6 Competing interests

The authors declare that they have no competing interests.

## 51.7 Authors' contributions

JDZ carried out the study and wrote the manuscript. LYG helped conceive, design and coordinate the study. LYG also revised the manuscript. EEB helped conceive and design the study. EEB also revised the manuscript. SHB supervised and helped conceive and design the study. All authors read and approved the final manuscript.

## 51.8 Acknowledgements

We are indebted to John Wilbur and Natalie Xie of NCBI for their helpful comments and assistance in obtaining data. This research was supported (in-part) by the Intramural Research Program of the NIH, National Library of Medicine.

# MYCHEMISE: A 2D DRAWING PROGRAM THAT USES MORPHING FOR VISUALISATION PURPOSES

## 52.1 Abstract

MyChemise (My Chemical Structure Editor) is a new 2D structure editor. It is designed as a Java applet that enables the direct creation of structures in the Internet using a web browser. MyChemise saves files in a digital format (.cse) and the import and export of .mol files using the appropriate connection tables is also possible.

MyChemise is available as a free online version in English and German. The MyChemise GUI is designed to be user friendly and can be used intuitively. There is also an English and German program description available as a PDF file.

In addition to the known ways of drawing chemical structure formulas, there are also parts implemented in the program that allow the creation of different types of presentation. The morphing module uses this technology as a component for dynamic visualisation. For example, it enables a clear and simple illustration of molecule vibrations and reaction sequences.

## 52.2 Introduction

2D drawing programs account for some of the first computer applications in chemistry and are widely distributed. Many publications, especially in recent times<sup>123</sup>, show that this sector is developing continually. The incorporation of new programming languages or improvements in methods for targeted structure searching in the Web continues to present a challenge for chemists who are interested in programming.

Having been given the option at work of setting up a new database with conventional structural images, the author asked himself the question how he could get hold of these images. Four options were considered:

1. To anchor the images using a link into the Web (e.g. using CAS-numbers).
2. Copy the structural images from a source in the Web.
3. Use a program already available on the market for drawing.
4. Develop a program oneself.

---

<sup>1</sup> Molecular structure input on the web

<sup>2</sup> FlaME: Flash Molecular Editor - a 2D structure input tool for the web

<sup>3</sup> Open Source: Strukturen zeichnen

The first two options can be ruled out because not all the compounds entered were available as finished structures. Additionally, the structural images should look standardised if they are to be used as a marketing instrument for customer relations. Obtaining the images from different sources would have been unsatisfactory because the sizes and types of representations vary greatly. Known drawing programs would have been useable; however the drawings would have still required subsequent work in order to achieve a company specific layout. Therefore the fourth option seemed most sensible and the most interesting.

MyChemise was written with the intention of producing a stand-alone approach in this field and because of the fun in programming chemistry software. The Version 11.01 presented here was created as sideline work in the period between January 2008 and March 2011.

## 52.3 Implementation

MyChemise is a modern scientific online 2D drawing program for chemical structural image. It was programmed in Java and runs as an applet in any browser that has an up-to-date Java plug-in (see <http://www.java.com/en/download/install.jsp> to check if one is available). MyChemise can be opened as the English version [http://www.knalltundstinkt.de/MyChemise\\_englisch/ChemiseZert\\_Home.html](http://www.knalltundstinkt.de/MyChemise_englisch/ChemiseZert_Home.html). If MyChemise is not running it is possibly necessary to activate javascript and/or to reduce the security settings of the browser. A program description is available as a PDF file [http://www.knalltundstinkt.de/MyChemise\\_englisch/Description.pdf](http://www.knalltundstinkt.de/MyChemise_englisch/Description.pdf) and additional file<sup>4</sup> including appropriate instructions. From here you can always open the latest version. One advantage of this online technology is that any subsequent program enhancements do not need to be downloaded and there is even no need for a new installation to be carried out because the latest MyChemise version is always available online.

Additional file 1

**Description.** The software description of MyChemise describes and presents the menu items. Well-known commands from standard-software (save, open etc.) or self-explanatory commands are not included.

Click here for file

The readers of this article can also download a zip-file (additional file)

Additional file 2

**mychemise.** It contains two files (MyChemise.html and ChemJar.jar). It can be downloaded and installed for running MyChemise in the off-line mode, too.

Click here for file

MyChemise is a signed applet, which means that it does not have to obey the applet sandbox principle intended for security purposes. It is possible to work online in the browser and once the work is finished to save the files on your own PC. This means that drawings can be directly exported as image files into other applications using the clipboard (Note for Linux users: see program description).

The (theoretical) possible number of atoms that can be drawn in a file in MyChemise was set to 100000.

MyChemise is optimized for Windows platforms with Firefox as browser. The minimum system requirements are a 1.6 GHz processor and 2 GB of RAM.

## 52.4 Results and discussion

The input screen (Figure 1) shows the most important commands, relevant menu items and toolbars in a clear way. This means that the screen is not overloaded with symbols belonging to windows program technology, only those toolbars that belong to the menu items that have just been opened are made visible.

---

<sup>4</sup> Knalltundstinkt

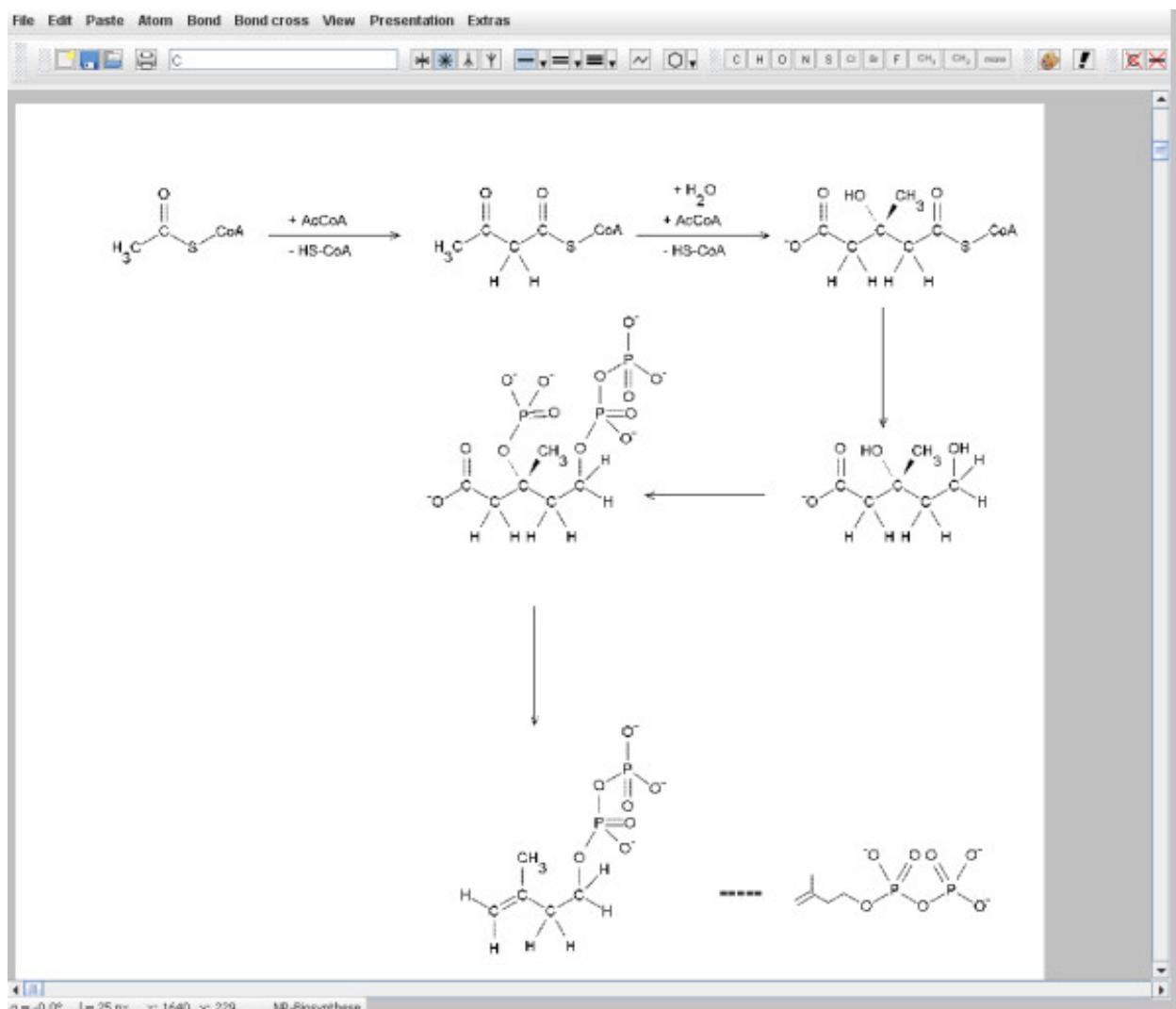


Figure 52.1: Figure 1. Input screen  
**Input screen.** It shows the menu items and the file-toolbar.

All atoms and bonds can be coloured. This allows aesthetically pleasing structural images to be easily created. Formulae can be represented in long form (with C- and H atoms labelled) and in short form. Many different bond types are available for selection. Cyclic hydrocarbons and aromatics can be quickly designed from a pull-down menu. Different 6-ring conformations can be constructed in the same way. Heterocyclics can be created by simply replacing the C atoms. Bond angles and lengths can be continuously adjusted by dragging the mouse or fit to the grid. Preset molecular geometries (tetrahedral, octahedral, etc.) make drawing easier. Newman projections are also possible.

Atoms can be shown with symbolised atomic shells (Figure 2), which can be separately shaped.

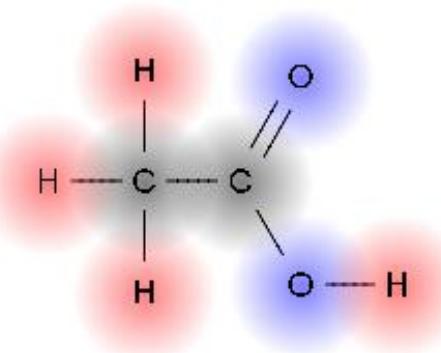


Figure 52.2: Figure 2. Structural image with atomic shells

**Structural image with atomic shells.** Atomic shells can be rendered with colour gradients.

A diverse range of drawing components (arrows, brackets, shapes) can be created using a dialog window (Figure 3). Using club and banana basic shapes orbitals can be symbolically represented. The switching on of colour gradients increases this impression.

If you want to give the structural images an individual style, then you can insert a background image into the drawing area which can be modified using a dialog window (Figure 4 and 5). A company logo (example of use see: <sup>5</sup>) can be used as a marketing instrument to increase corporate identity (Figure 6).

Single rows of text can be entered using the atom input box. A special editor (Figure 7) is available for multiple rows of text. You can also use the editor to directly help you create short chemical-specific text in MyChemise, without the need for an extra writing program. This makes work easier and can help to save time, and therefore costs.

A selection of special characters can be opened using a dialog window (Figure 8). Individual characters can be inserted into the editor or the atom input box by copying and pasting.

Atomic symbols are also shown with their atomic and mass numbers (Figure 9).

### 52.4.1 File import and export

The mol file format <sup>67</sup>, version V2000 was chosen as an interface so that MyChemise can also exchange files with similar programs. Mol files can be opened in the MyChemise screen and can be saved in mol format. In addition, mol files can be inserted just by a single click of the mouse into cse files and can also be attached onto existing structures. When doing this, the atom farthest to the left is always used as the coupling atom (atom with the smallest x-coordinate).

The export function does not automatically limit the number of atoms and bonds that are to be exported. MyChemise allows the input of a large number of atoms, whilst mol files are setup for a maximum of 999 atoms and 999 bonds,

<sup>5</sup> WST-Winkel GmbH

<sup>6</sup> NOTITLE!

<sup>7</sup> CTfile Formats

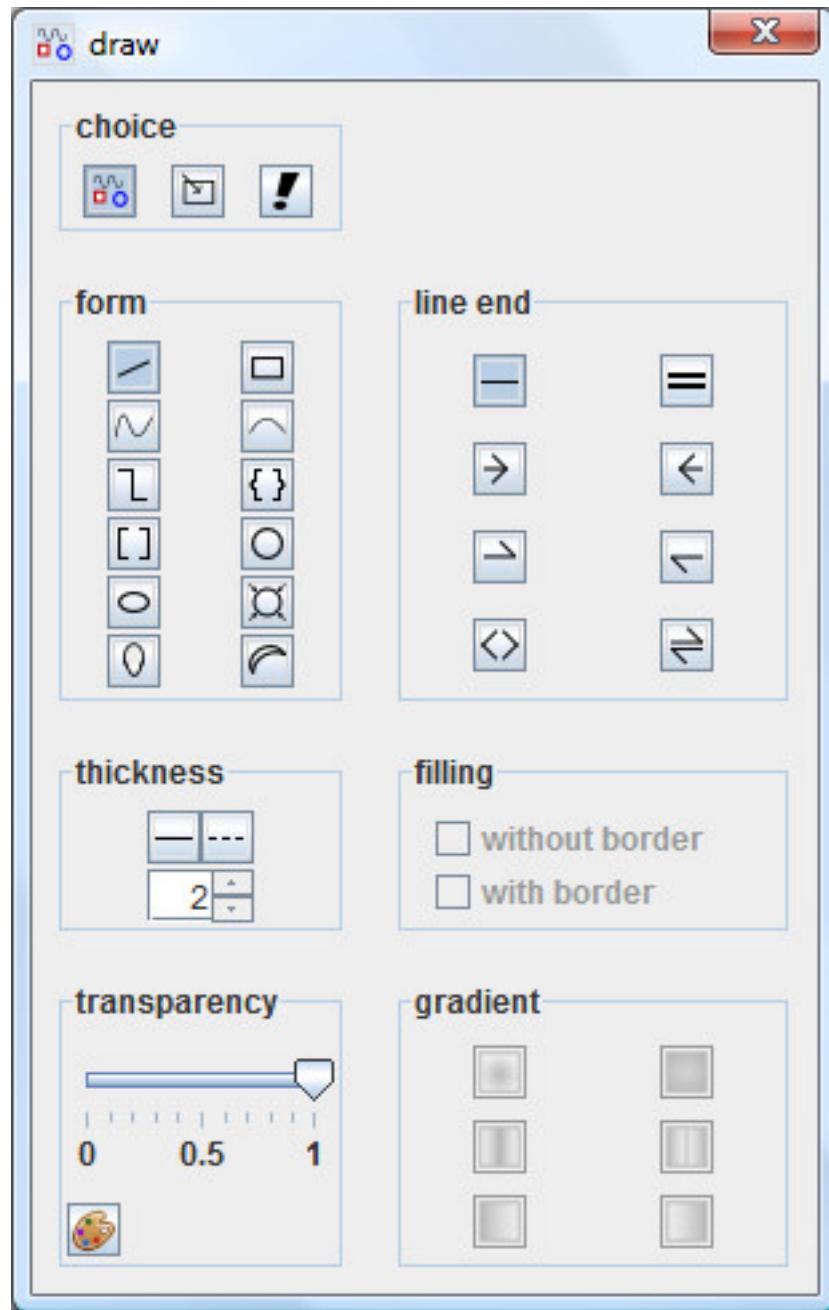


Figure 52.3: Figure 3. Dialog window of drawing components

**Dialog window of drawing components.** A choice of shapes can be added to the sketch area. Some of them are useful to symbolize orbitals.

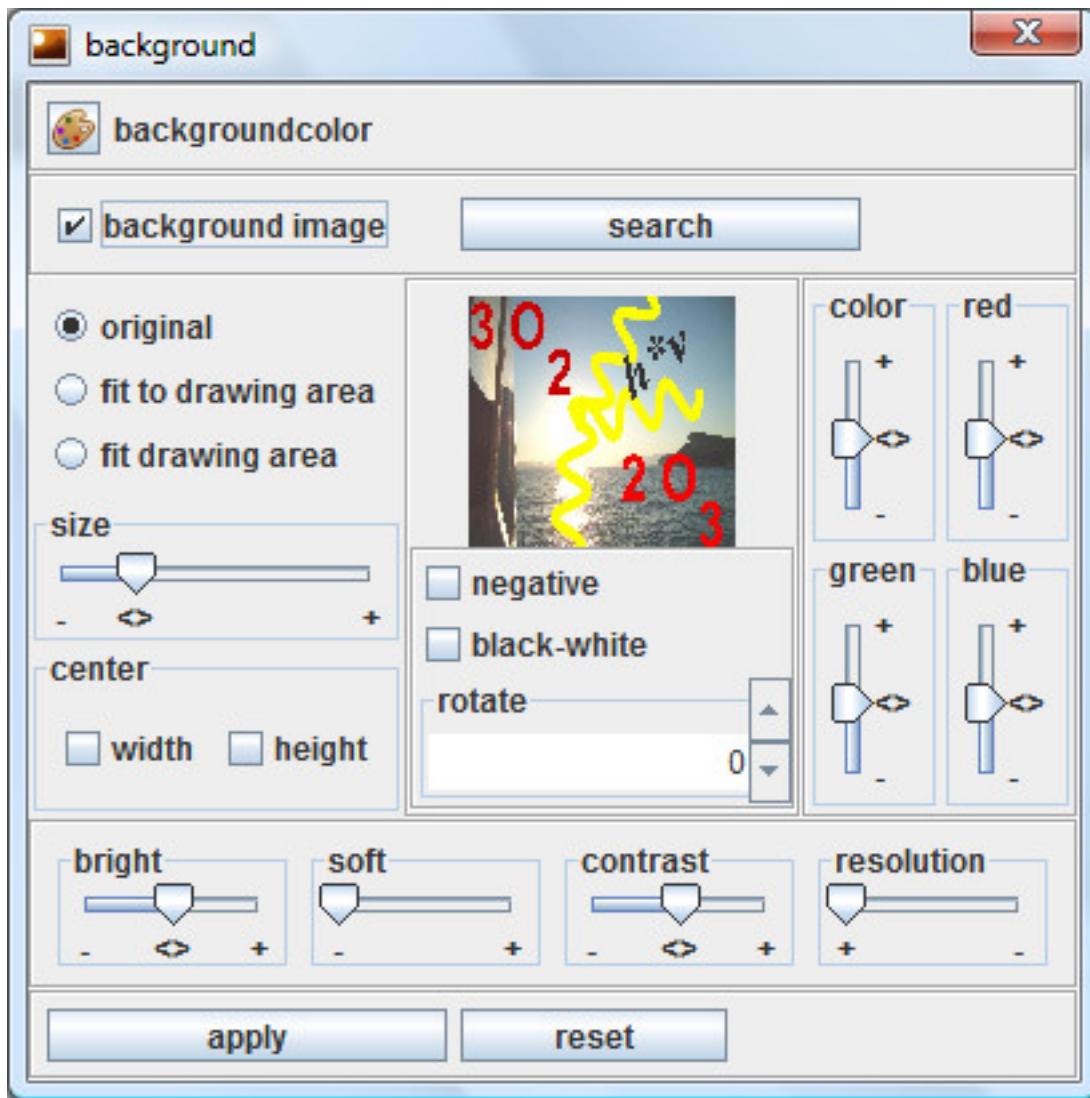


Figure 52.4: Figure 4. Dialog window for the background

**Dialog window for the background.** It allows you to modify background images in several ways. Watermarks can be generated by increasing the brightness.

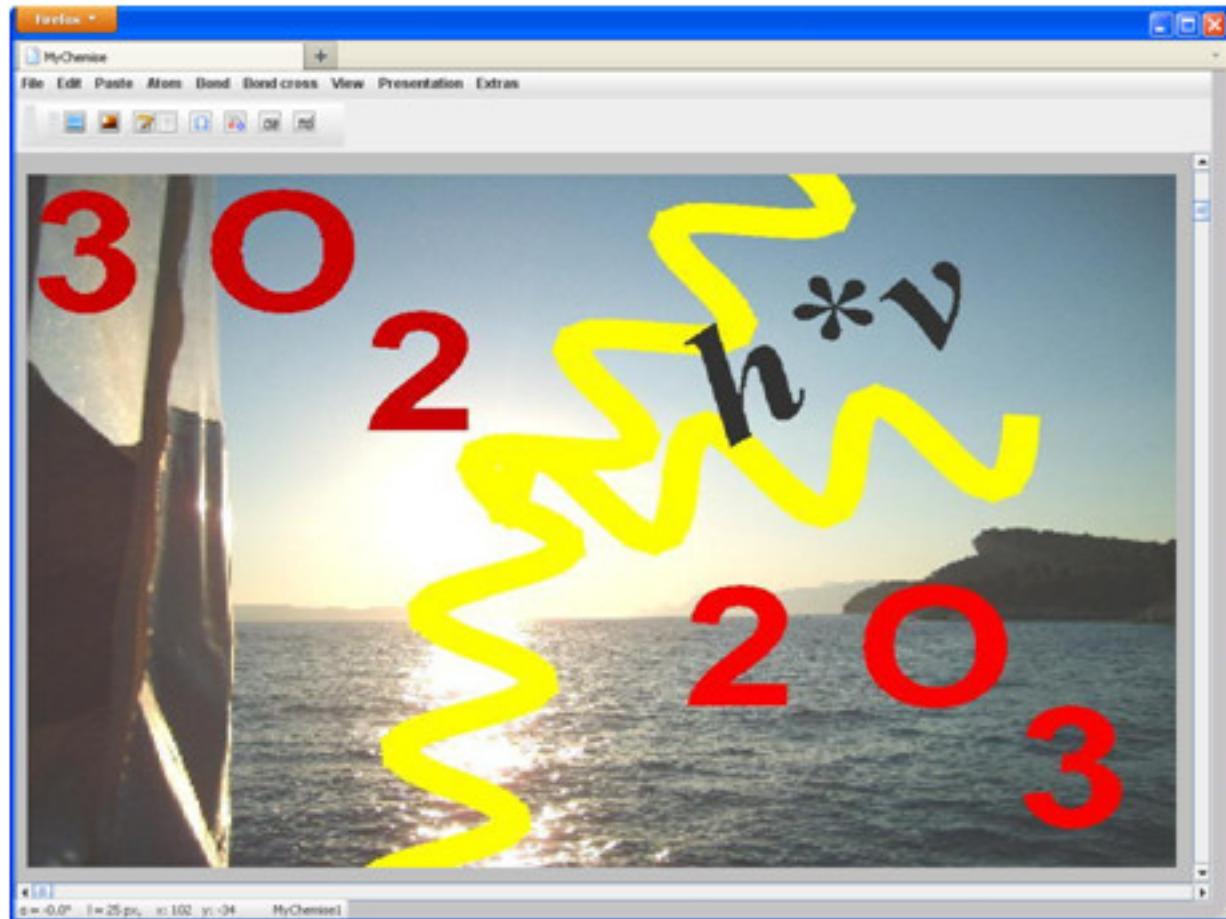


Figure 52.5: Figure 5. Example for a chemical depiction with a background image  
**Example for a chemical depiction with a background image.** The background image can be used as drawing surface.

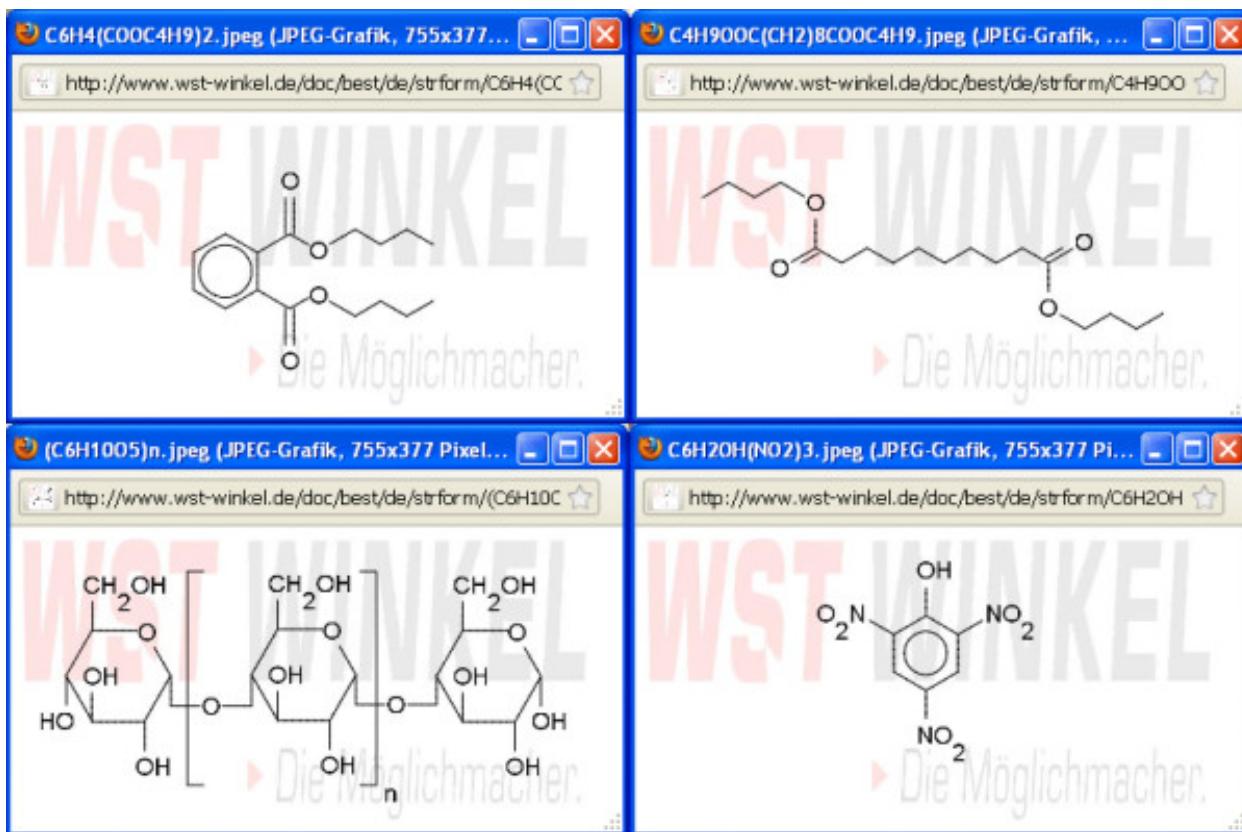


Figure 52.6: Figure 6. Structural images with a company logo

**Structural images with a company logo.** Chemical depictions with equal company logo can be used as a marketing instrument to increase corporate identity.

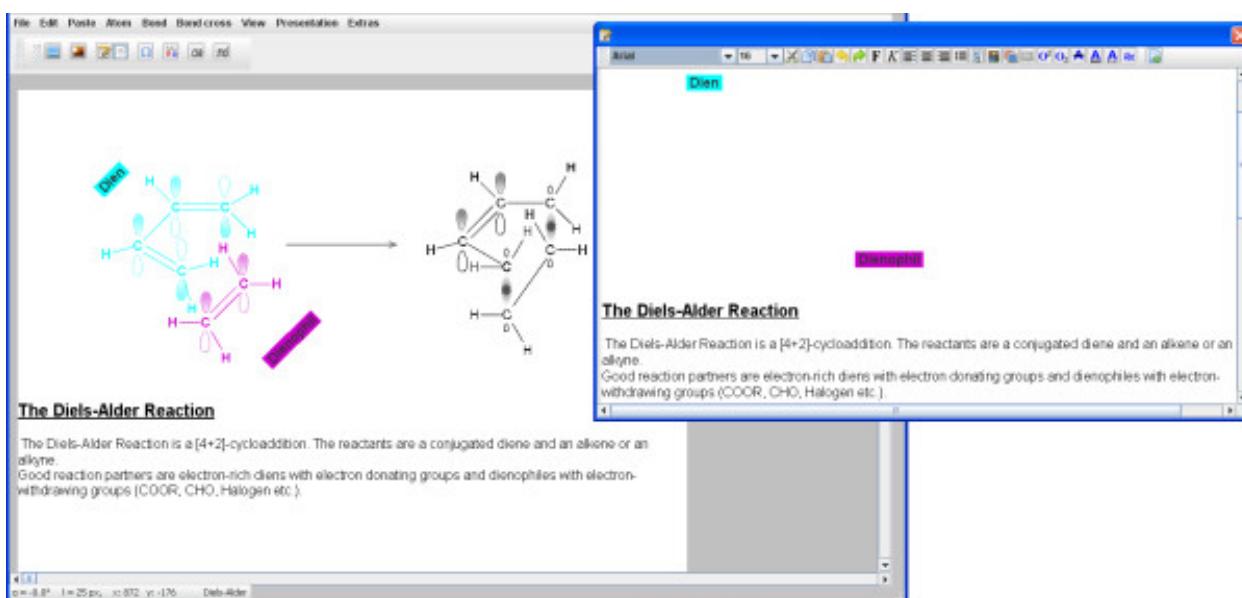


Figure 52.7: Figure 7. The text editor

**The text editor.** With the text editor multiple rows of text can be formatted and added to the sketch area. Rotated text is only shown there.

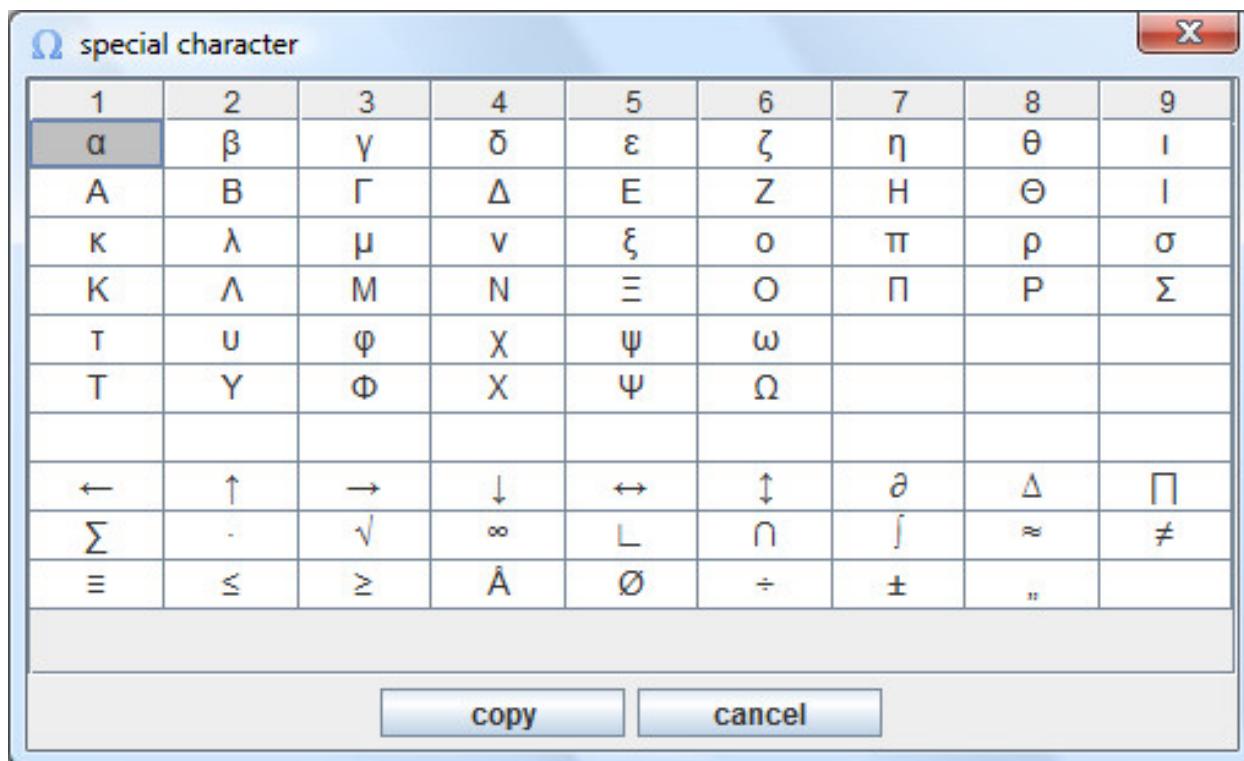


Figure 52.8: Figure 8. Dialog window for special characters  
**Dialog window for special characters.** A selection of often used symbols.

elements																	
1A	2A	3B	4B	5B	6B	7B	8B	8B	8B	1B	2B	3A	4A	5A	6A	7A	8A
H																	He
Li	Be									B	C	N	O	F	Ne		
Na	Mg										Al	Si	P	S	Cl	Ar	
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac	Rf	Db	Sg	Bh	Hs	Mt									
			Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	
			Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	
<b>C</b>				<b>6C</b>				<b>12,011 C</b>				<b>12,011 6C</b>					

Figure 52.9: Figure 9. Atomic symbols  
**Atomic symbols.** Atomic symbols can be pasted with their atomic and mass numbers or without.

this means that every user must ensure for themselves that when creating files they make them compatible with other programs.

### 52.4.2 Specials

A special highlight in MyChemise is the option of presenting any created drawings, structural images and texts in different ways; this can be done directly in the program itself. One of the options available allows you to put together different documents into a script in order to present a slide show. Another option exists that, for example, can arrange the fluctuating border structures of chemical depictions into an animation. Such animations can of course be integrated into a script. Morphing as a means of teaching chemistry/science has up until now been used very little. MyChemise offers you morphing as method of presentation.

### 52.4.3 The morphing module

When morphing is carried out, two images are brought together. This involves allocating those areas of the images with each other that are to be transformed. Changes are made in steps and apply to both shape and colour. Intermediate steps are interpolated from the starting images, whereby the share of one image reduces in dimensions while the dimensions of the other image increase<sup>8</sup>. Using MyChemise, images can be morphed using affine or three-point mapping (i.e. division into triangular sections, affine mapping) and by dividing up into square areas. Four-point mapping (projective mapping) is mathematically solved using the unit squares method<sup>8</sup>.

When illustrating chemical states, it is sometimes more useful to transform only specific areas into each other. MyChemise achieves this by automatically recognising only the bonding and atom areas in the structural images used as being areas to be morphed. As soon as only two drawings have been made, dynamic representations can then be quickly produced by calculating the intermediate steps. These allow movements (e.g. molecular vibrations) and sequences (e.g. reaction mechanisms) to be graphically simulated.

The menu item Extras enables you to upload various, simple morphing examples in the on-line mode (Figures 10, 11, 12). The Morphing window then opens and they can be started from within the Morphing menu item. The process behaviour can be changed in the morphing set-up dialog box. Several morphing steps can be combined to a sequence (Figure 13).

## 52.5 Conclusions

MyChemise is a new 2D drawing program that places special importance on simple operation and versatile ways of creating structural images. Continual advancements in processors have led to increasingly faster desktop PCs. Greater amounts of RAM have also enabled the inclusion of methods for displaying dynamic processes in such programs, in this case the morphing module.

An enhancement for MyChemise is currently being worked on, which will, amongst other things, be able to export SMILES strings<sup>9</sup> that can be used for structure searching in databases.

The continuation of MyChemise as an open source project has been planned for a later date.

## 52.6 Competing interests

The author declares that they have no competing interests.

---

<sup>8</sup> NOTITLE!

<sup>9</sup> SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules

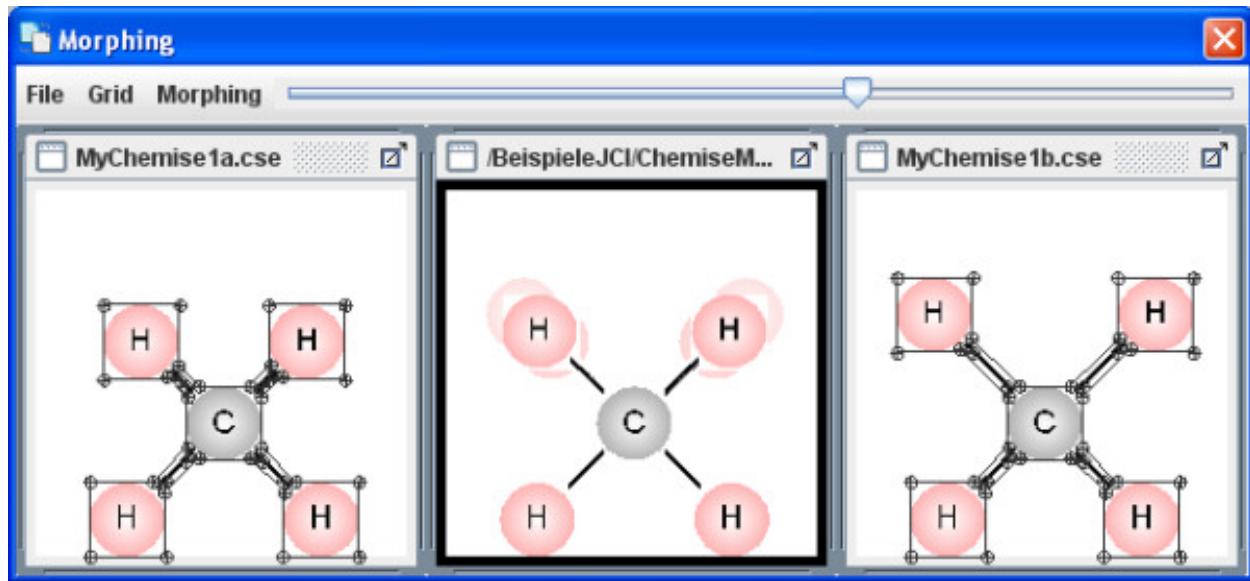


Figure 52.10: Figure 10. Example Morph 1

**Example Morph 1.** The morphing window shows an example of molecule vibrations (symmetrical stretching).

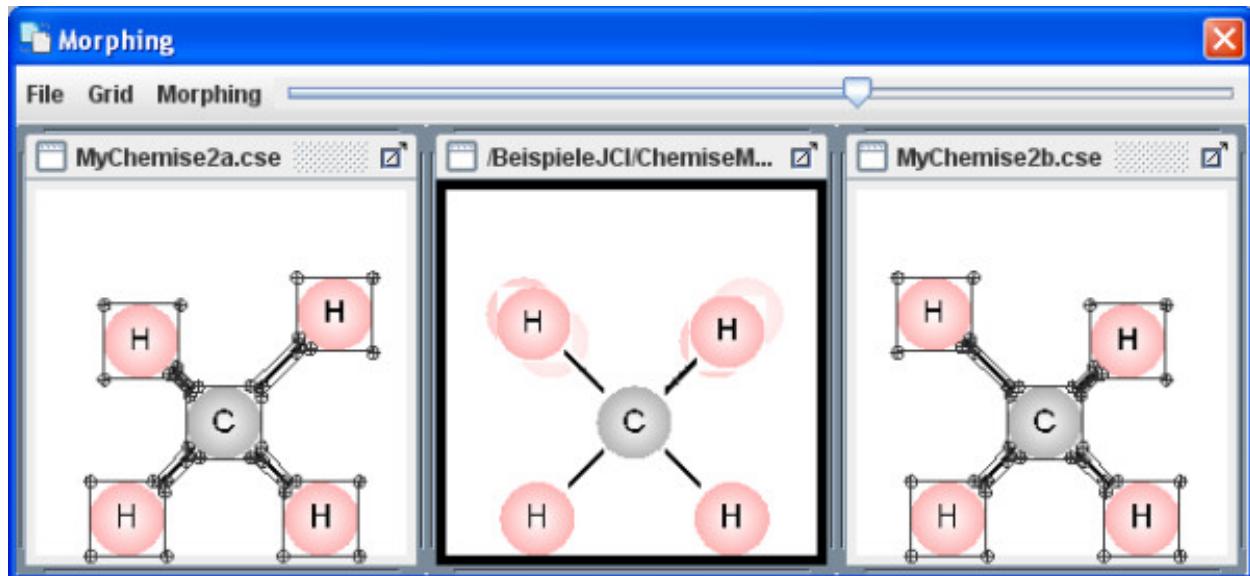


Figure 52.11: Figure 11. Example Morph 2

**Example Morph 2.** The morphing window shows an example of molecule vibrations (asymmetrical stretching).

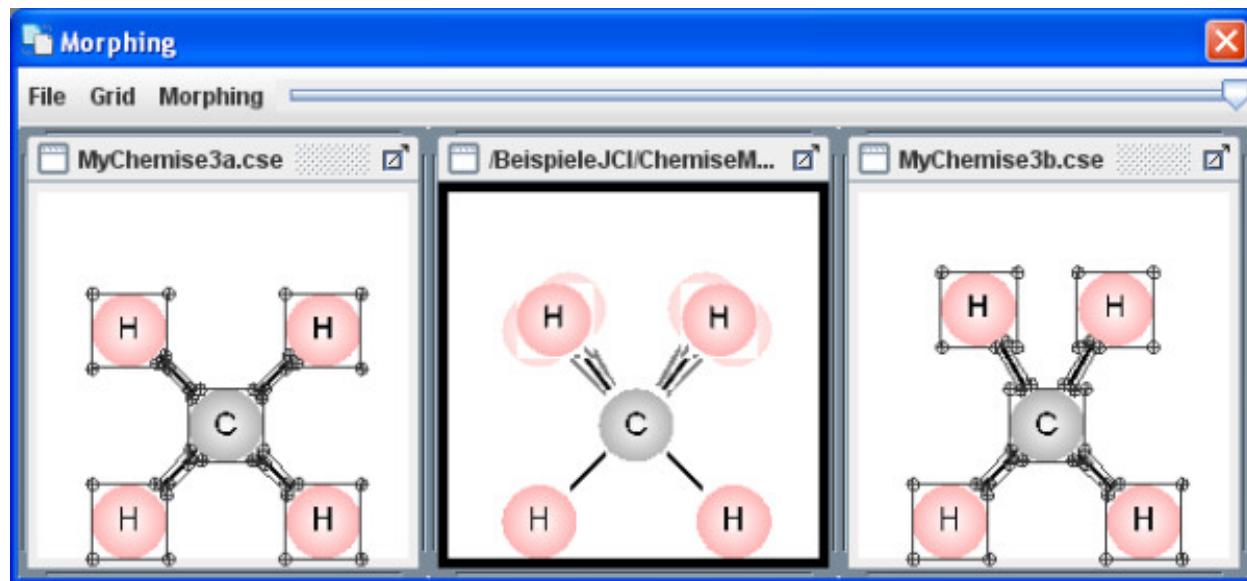


Figure 52.12: Figure 12. Example Morph 3

**Example Morph 3.** The morphing window shows an example of molecule vibrations (scissoring).

## 52.7 Acknowledgements

Devoted to Andreas Schumann and Hans Wilhelm.

I would like to thank Ingrid for providing specialised literature (and not only this). Daniel receives thanks for his support in setting up the homepage. MyChemise was written using free available java editor JOE<sup>10</sup>. SignTool<sup>11</sup> was very useful for signing the applets.

### 52.7.1 List of sources

Parts of the following program codes were adapted for and extended for MyChemise:

For the preview window in the File- > Open menu: PreviewPanel.java <http://www.quignon.de> accessed 11/22/2011

For the morphing module: TriangulatedImage.java and MorphingCandS.java from [#B12]\_<<http://public.rz.fh-wolfsburg.de/~Klawonn/computergrafik/>>\_ accessed 11/22/2011

For displaying round colour gradients: RoundGradientContext.java from [#B13]\_<<ftp://ftp.oreilly.com/pub/examples/java/2d/>>\_ accessed 11/22/2011

<sup>10</sup> JOE

<sup>11</sup> SignTool

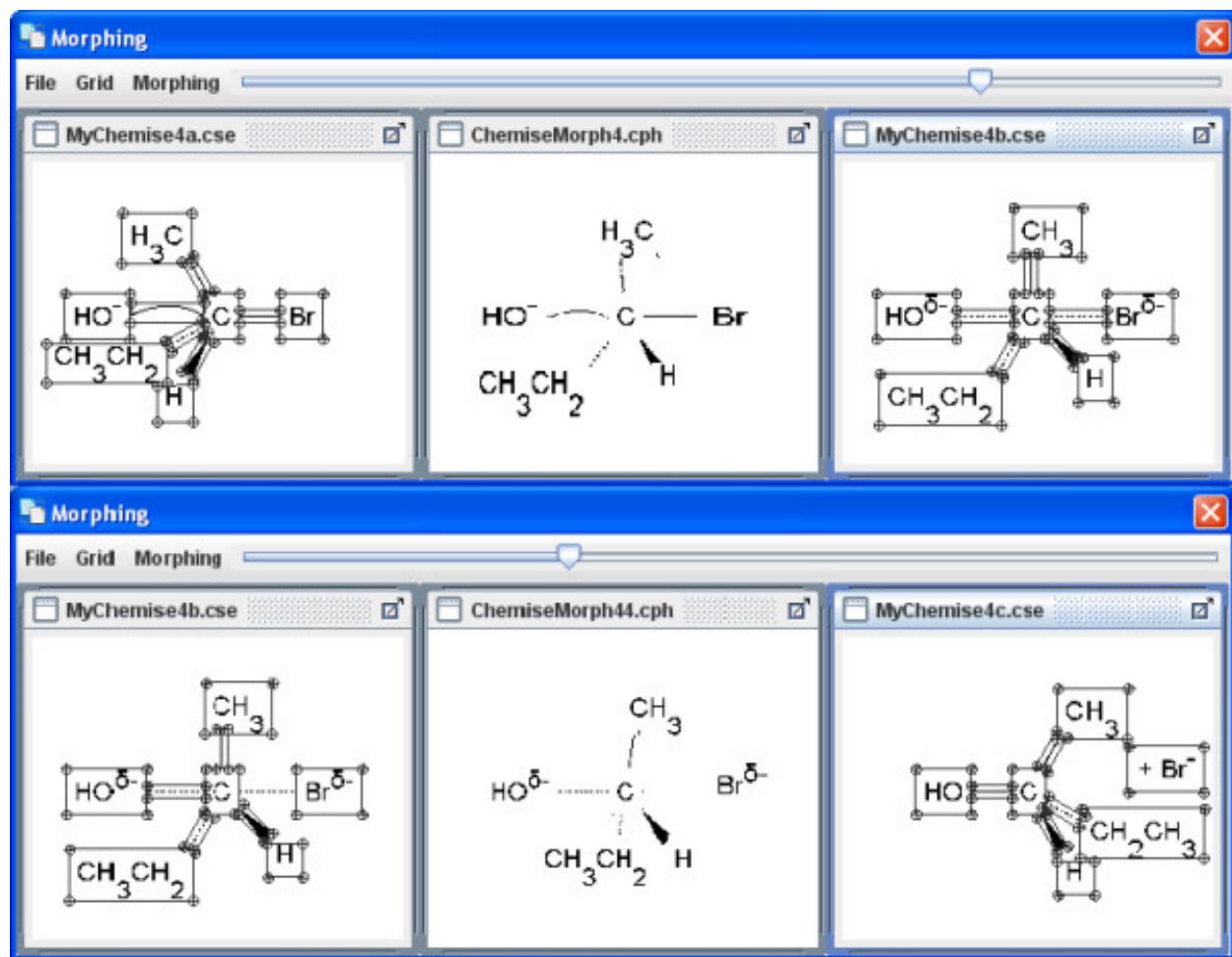


Figure 52.13: Figure 13. A SN<sub>2</sub>-reaction as an example for a morphing sequence  
**A SN<sub>2</sub>-reaction as an example for a morphing sequence.** Reaction sequences can be visualized by combining two or more morphing steps.



# NEW DEVELOPMENTS ON THE CHEMINFORMATICS OPEN WORKFLOW ENVIRONMENT CDK-TAVERNA

## 53.1 Abstract

### 53.1.1 Background

The computational processing and analysis of small molecules is at heart of cheminformatics and structural bioinformatics and their application in e.g. metabolomics or drug discovery. Pipelining or workflow tools allow for the Lego™-like, graphical assembly of I/O modules and algorithms into a complex workflow which can be easily deployed, modified and tested without the hassle of implementing it into a monolithic application. The CDK-Taverna project aims at building a free open-source cheminformatics pipelining solution through combination of different open-source projects such as Taverna, the Chemistry Development Kit (CDK) or the Waikato Environment for Knowledge Analysis (WEKA). A first integrated version 1.0 of CDK-Taverna was recently released to the public.

### 53.1.2 Results

The CDK-Taverna project was migrated to the most up-to-date versions of its foundational software libraries with a complete re-engineering of its worker's architecture (version 2.0). 64-bit computing and multi-core usage by paralleled threads are now supported to allow for fast in-memory processing and analysis of large sets of molecules. Earlier deficiencies like workarounds for iterative data reading are removed. The combinatorial chemistry related reaction enumeration features are considerably enhanced. Additional functionality for calculating a natural product likeness score for small molecules is implemented to identify possible drug candidates. Finally the data analysis capabilities are extended with new workers that provide access to the open-source WEKA library for clustering and machine learning as well as training and test set partitioning. The new features are outlined with usage scenarios.

### 53.1.3 Conclusions

CDK-Taverna 2.0 as an open-source cheminformatics workflow solution matured to become a freely available and increasingly powerful tool for the biosciences. The combination of the new CDK-Taverna worker family with the already available workflows developed by a lively Taverna community and published on myexperiment.org enables molecular scientists to quickly calculate, process and analyse molecular data as typically found in e.g. today's systems biology scenarios.

## 53.2 Background

Current problems in the biosciences typically involve several domains of research. They require a scientist to work with different and diverse sets of data. The reconstruction of a metabolic network from sequencing data, for example, employs many of the data types found along the axis of the central dogma, including reconstruction of genome sequences, gene prediction, determination of encoded protein families, and from there to the substrates of enzymes, which then form the metabolic network. In order to work with such a processing pipeline, a scientist has to copy/paste and often transform the data between several bioinformatics web portals by hand. The manual approach involves repetitive tasks and cannot be considered effective or scalable.

Especially the processing and analysis of small molecules comprises tasks like filtering, transformation, curation or migration of chemical data, information retrieval with substructures, reactions, or pharmacophores as well as the analysis of molecular data with statistics, clustering or machine learning to support chemical diversity requirements or to generate quantitative structure activity/property relationships (QSAR/QSPR models). These processing and analysis procedures itself are of increasing importance for research areas like metabolomics or drug discovery. The power and flexibility of the corresponding computational tools become essential success factors for the whole research process.

The workflow paradigm addresses the above issues with the supply of sets of elementary workers (activities) that can be flexibly assembled in a graphical manner to allow complex procedures to be performed in an effective manner - without the need of specific code development or software programming skills. Scientific workflows allow the combination of a wide spectrum of algorithms and resources in a single workspace<sup>123</sup>. Earlier problems with iterations over large data sets<sup>4</sup> are completely resolved in version 2.0 due to new implementations in Taverna. Taverna 2 allows control structures such as “while” loops or “if-then-else” constructs. Termination criteria for loops may now be evaluated by listening to a state port<sup>5</sup>. In addition the user interface of the Taverna 2 workbench has clearly improved: The design and manipulation of workflows in a graphical workflow editor is now supported. Features like copy/paste and undo/redo simplify workflow creation and maintenance<sup>6</sup>.

The CDK-Taverna project aims at building a free open-source cheminformatics pipelining solution through combination of different open-source projects such as Taverna<sup>7</sup>, the Chemistry Development Kit (CDK)<sup>89</sup>, or the Waikato Environment for Knowledge Analysis (WEKA)<sup>10</sup>. A first integrated version 1.0 of CDK-Taverna was recently released to the public<sup>4</sup>. To extend usability and power of CDK-Taverna for different molecular research purposes the development of version 2.0 was motivated.

### 53.2.1 Implementation

The CDK-Taverna 2.0 plug-in makes use of the Taverna plug-in manager for its installation. The manager fetches all necessary information about the plug-in from a XML file which is located at <http://www.ts-concepts.de/cdk-taverna2/plugin/>. The information provided therein contains the name of the plug-in, its version, the repository location and the required Taverna version. Upon submitting the URL to the plug-in manager it downloads all necessary dependencies automatically from the web. After a subsequent restart the plug-in is enabled and the workers are visible in the services. The plug-in uses Taverna version 2.2.1<sup>6</sup>, CDK version 1.3.8<sup>11</sup> and WEKA version 3.6.4<sup>12</sup>. Like its predecessor it uses the Maven 2 build system<sup>13</sup> as well as the Taverna workbench for automated dependency management.

---

<sup>1</sup> Cheminformatics analysis and learning in a data pipelining environment

<sup>2</sup> Scientific workflows as productivity tools for drug discovery

<sup>3</sup> Taverna/my Grid: Aligning a Workflow System with the Life Sciences Community

<sup>4</sup> CDK-Taverna: an open workflow environment for cheminformatics

<sup>5</sup> Taverna, Reloaded

<sup>6</sup> Taverna 2

<sup>7</sup> Taverna: a tool for the composition and enactment of bioinformatics workflows

<sup>8</sup> The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics

<sup>9</sup> Recent Developments of The Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics

<sup>10</sup> The WEKA Data Mining Software: An Update

<sup>11</sup> The Chemistry Development Kit(CDK)

<sup>12</sup> Waikato Environment for Knowledge Analysis (WEKA)

<sup>13</sup> Apache Maven

### 53.2.2 CDK-Taverna 2.0 worker implementation

The CDK-Taverna 2.0 plug-in is designed to be easily extendible: The implementation allows to create new workers by simply inheriting from the single abstract class

- 
- 
- 
- 
- 

Finally a new worker has to be registered to be available in the Taverna workbench. For this purpose Taverna offers the class

Besides the basic implementation it is possible to define a configuration panel for a worker which allows the specification of parameters. A configuration panel has to inherit from the abstract class

- 
- 
- 
- 
- 

The configuration panel has to be registered in the [http://cdk-taverna-2.ts-concepts.de/wiki/index.php?title=Main\\_Page](http://cdk-taverna-2.ts-concepts.de/wiki/index.php?title=Main_Page).

### 53.2.3 Requirements

CDK-Taverna 2.0 supports 64-bit computing by use with a Java 64-bit virtual machine. The CDK-Taverna 2.0 plug-in is written in Java and requires Java 6 or higher. The latest Java version is available at <http://www.java.com/de/download/>. The CDK-Taverna 2.0 plug-in is developed and tested on Microsoft Windows 7 as well as Linux and Mac OS/X (32 and 64-bit).

## 53.3 Results and Discussion

The CDK-Taverna 2.0 plug-in provides 192 workers for input and output (I/O) of various chemical file and line notation formats, substructure filtering, aromaticity detection, atom typing, reaction enumeration, molecular descriptor calculation and data analysis. Parallel computing with multi-core processors by use of multiple concurrent threads is flexibly implemented for many workers where operations scale nearly linear with the number of cores. Especially the machine learning and the molecular descriptor calculation workers benefit from parallel computation. An overview is given in Tables 1 and 2. Many workers are described by example workflows available at [http://cdk-taverna-2.ts-concepts.de/wiki/index.php?title=Main\\_Page](http://cdk-taverna-2.ts-concepts.de/wiki/index.php?title=Main_Page). Additionally, the workflows can be found at <http://www.myexperiment.org/>.

CDK-Taverna 1.0 was confined to 32-bit Java virtual machine and thus was restricted to in-memory processing of data volumes of at most 2 gigabyte in practice. Version 2.0 also supports 64-bit computing by use of a 64-bit Java virtual machine so that the processable data volume is only limited by hardware constraints (memory, speed): 64-bit in-memory workflows were successfully performed with data sets of about 1 million small molecules. Since the memory restrictions of version 1.0 were a main reason to use Pgchem::tigress as a molecular database backend<sup>4</sup> the corresponding version 1.0 workers were not migrated to the current version 2.0 yet.

### 53.3.1 Advanced reaction enumeration

CDK-Taverna 1.0 provided basic functions for combinatorial chemistry related reaction enumeration: They supported the use of two reactants, a single product and one generic group per reactant. The new enumeration options used by CDK-Taverna 2.0 offer major enhancements like multi-match detection, any number of reactants, products or generic groups as well as variable R-groups, ring sizes and atom definitions. The extended functionality was developed and applied in industrial cooperation projects. Advanced reaction enumeration features are illustrated in Figure 1. The *Variable RGroup* feature allows the definition of chemical groups which can be flexibly attached to predefined atoms with syntax  $[A:B,B,B...-RC]$  where  $A$  is a freely selectable identifier,  $B$  are numbers from an *Atom-to-Atom-Mapping* defining the atoms to which the generic group can be attached and  $C$  is the chemical group identifier which can be any number. The *Atom Alias* feature offers the possibility to define a wild card for preconfigured elements. The syntax is  $[A:B,B,B...]$  where  $A$  is a freely selectable identifier and  $B$  are the string representations of the possible elements. The *Expandable Atom* feature enables the definition of freely sizeable rings or aliphatic chains with syntax  $[A:[]B]$  where  $A$  is a freely selectable identifier and  $B$  is the maximum number of atoms to insert. Figure 2 depicts a workflow for reaction enumeration. The capabilities of the advanced reaction enumerator implementation are summarized in Figure 3 which also demonstrates multi-match detection, i.e. multiple reaction centers within one molecule.

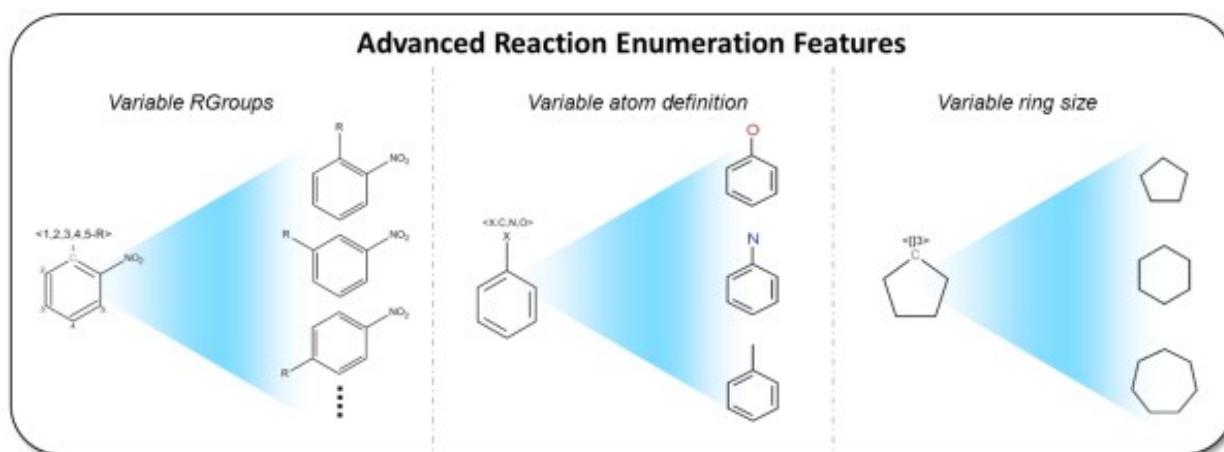


Figure 53.1: Figure 1. Advanced reaction enumeration features: (left) The *Variable RGroup* feature allows the definition of chemical groups which can be flexibly attached to predefined atoms

**Variable RGroup Advanced reaction enumeration features:** (left) The . (middle) The Atom Alias feature offers the possibility to define a wild card for preconfigured elements. (right) The Expandable Atom feature enables the definition of freely sizeable rings or aliphatic chains.

### 53.3.2 Evaluation of small molecules for natural product likeness

In recent years, computer assisted drug design studies use natural product (NP) likeness as a criterion to screen compound libraries for potential drug candidates<sup>1415</sup>. The reason to estimate NP likeness during candidate screening is to facilitate the selection of those compounds that mimic structural features that are naturally evolved to best interact with biological targets.

Version 2.0 of CDK-Taverna provides two groups of workers that re-implement the work of *Ertl et al* to score small molecules for NP-likeness<sup>14</sup>. The workers in the Molecule Curation folder are dedicated to the pre-processing of chemical structures: The Molecule Connectivity Checker worker removes counter ions and disconnects fragments, the Remove Sugar Groups worker removes all sugar rings and linear sugars from structures and the Curate Strange Elements worker discards structures that are composed of elements other than non-metals. This set of curation workers

<sup>14</sup> Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries

<sup>15</sup> Metabolite-likeness as a criterion in the design and selection of pharmaceutical drug libraries

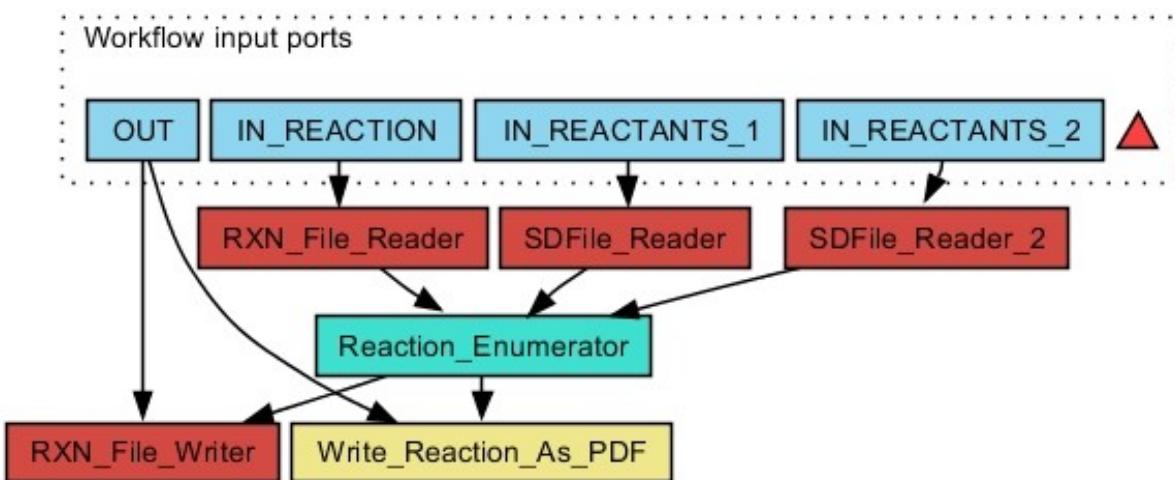


Figure 53.2: Figure 2. Workflow for reaction enumeration: After loading a generic reaction (

**Workflow for reaction enumeration:** After loading a generic reaction (\*\*, from a MDL RXN file) and two educt lists (\*\*, from MDL SD files) the worker performs the enumeration with the results stored as MDL RXN files. An additional PDF file is created which shows all enumerated reactions in a tabular manner. The results are stored in the output folder determined by the OUT input port.

finally creates scaffolds olds and sub structures. From these structures atom signatures<sup>16</sup> are generated using the Generate Atom Signatures worker and exploited as structural descriptors in charting the compound's region in the chemical structure space. The combined workflow of curation and atom signature generation workers is illustrated in Figure 4. Using this workflow, atom signatures can be generated for user-defined training (Natural products and synthetics) and testing (compound libraries) structural dataset. Workers of the Signature Scoring folder use atom signatures generated from compound libraries and rank them for NP-likeness based on the statistics suggested by Ertl *et al*<sup>14</sup>. This scoring workflow is illustrated in Figure 5. The whole package of workflows is available for free download at <http://www.myexperiment.org/users/10069/packs>. The curation and signature scoring workers may not only be applied in evaluating the NP-likeness of compound libraries but also in evaluating the metabolite-likeness of theoretical metabolites for predicting whole metabolomes. The latter application was the original purpose for the worker development and corresponding results will be presented in a subsequent publication.

### 53.3.3 Clustering and machine learning applications

Unsupervised clustering tries to partition input data into a number of groups smaller than the number of data whereas supervised machine learning tries to construct model functions that map the input data onto their corresponding output data. If the output codes continuous quantities a regression task is defined. Alternatively the output may code classes so that a classification task is addressed. Molecular data sets for clustering consist of input vectors where each vector represents a molecular entity and consists of a set of molecular descriptors itself. Molecular data sets for machine learning add to each input vector a corresponding output vector with features to be learned - thus they consist of I/O pairs of input and output vectors.

The clustering and machine learning workers of CDK-Taverna 2.0 allow the use of distinct WEKA functionality. As far as clustering is concerned the ART-2a worker of version 1.0 is supplemented with five additional WEKA-based workers which offer

•<sup>17</sup>.

•<sup>18</sup>.

<sup>16</sup> The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

<sup>17</sup> Maximum likelihood from incomplete data via the EM algorithm

<sup>18</sup> A best possible heuristic for the k-center problem

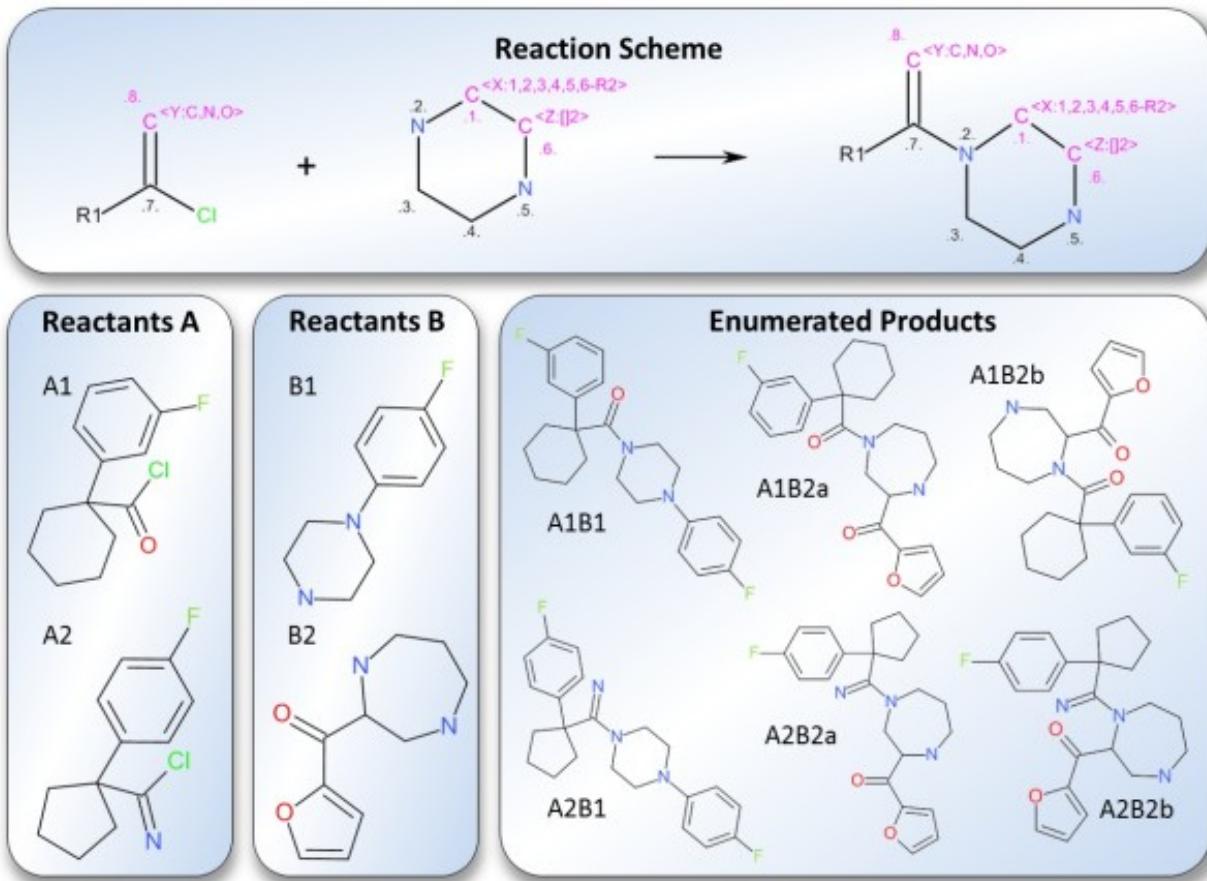


Figure 53.3: Figure 3. Capabilities of the advanced reaction enumerator: The sketched generic reaction contains three different generic groups labelled X, Y and Z.

**Capabilities of the advanced reaction enumerator:** The sketched generic reaction contains three different generic groups labelled X, Y and Z. Group x defines a Variable RGroup which can freely attach to all atoms of the ring. The Atom Alias group labelled Y is a wild card for the elements carbon, oxygen and nitrogen. The Expandable Atom group Z defines a variable ring size: The ring can be expanded by up to two additional carbon atoms. The enumerated products with the small letters a and b originate from multi-match detection.

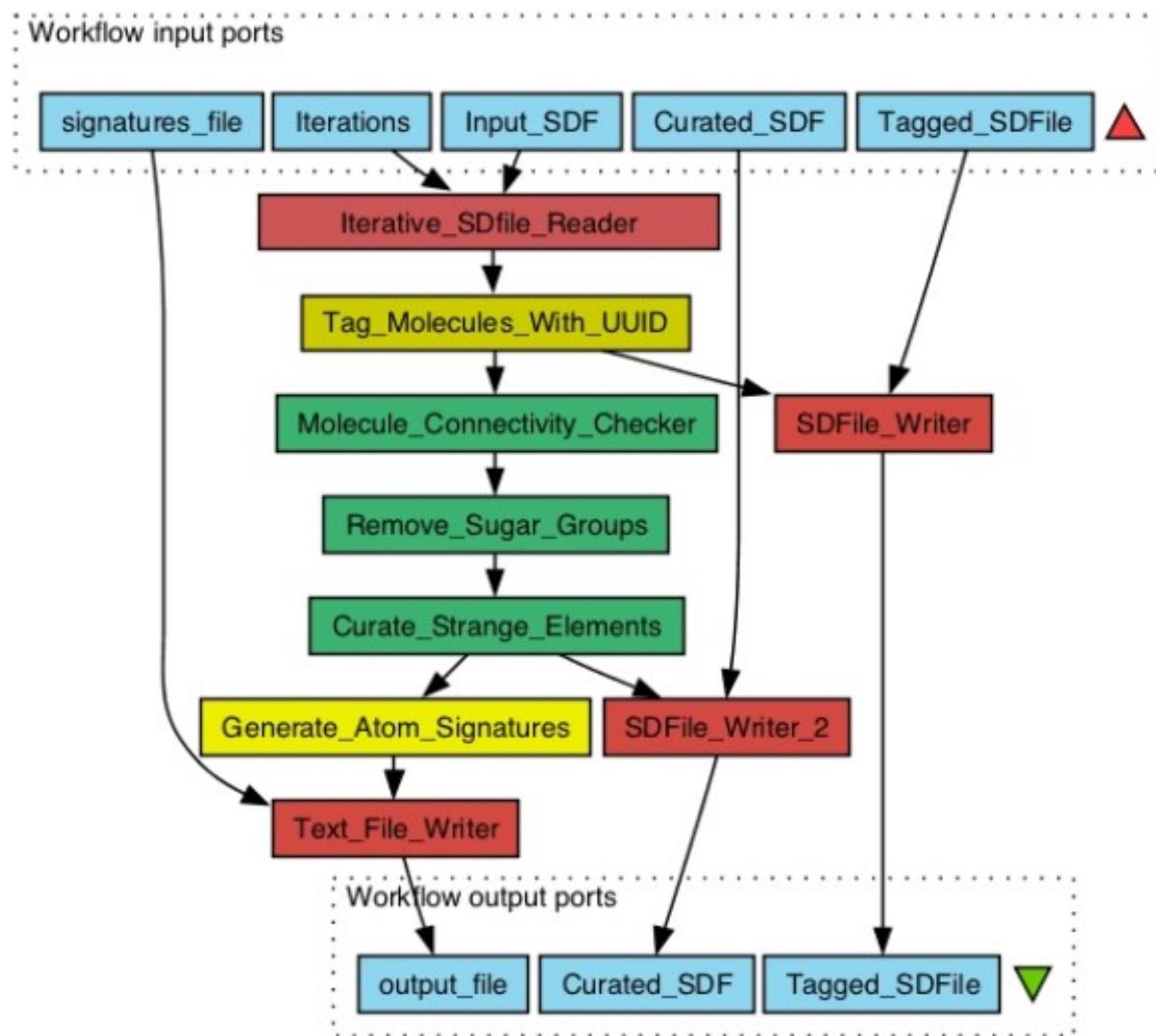


Figure 53.4: Figure 4. Molecule curation and atom signature descriptor generation workflow: The **Molecule curation and atom signature descriptor generation workflow**: The takes the Structure-Data File (SDF) of compounds (\*\*\*) as input and pass the structures down the workflow for molecule curation and atom signature generation\*\*. The number of structures to be read, and pumped down the workflow can be configured  
[\(<http://www.myexperiment.org/workflows/2120.html>\).](http://www.myexperiment.org/workflows/2120.html)

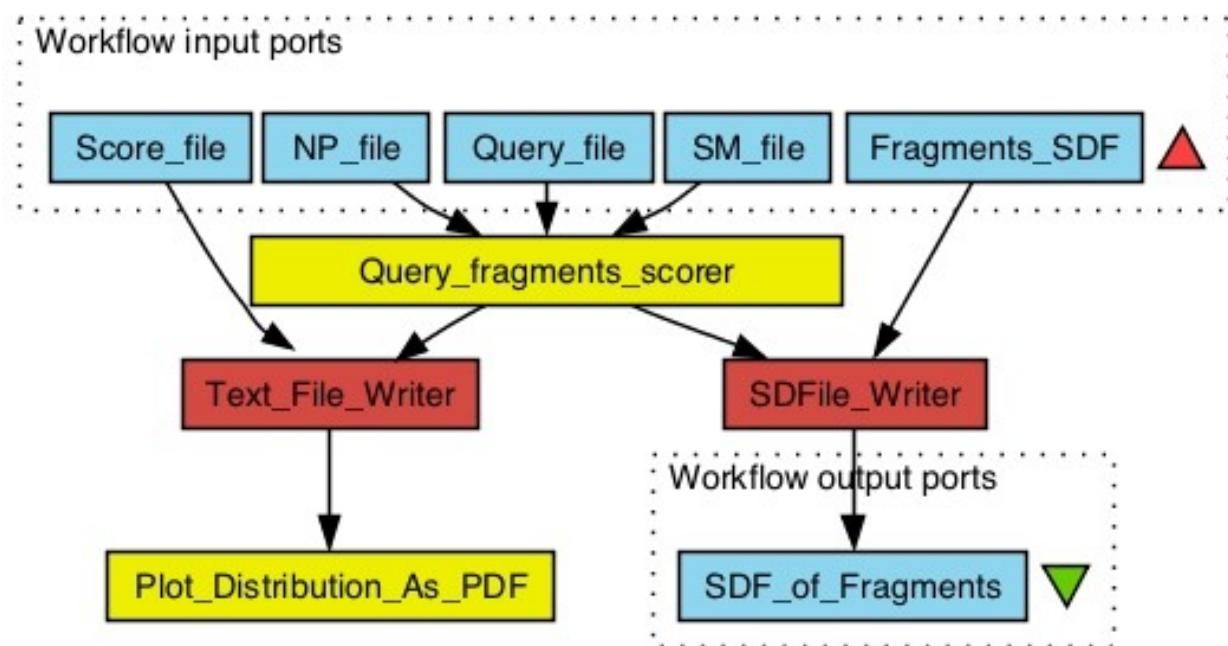


Figure 53.5: Figure 5. NP-likeness scoring workflow: This workflow take inputs of atom signatures file generated from the user defined natural products library (

**NP-likeness scoring workflow:** This workflow take inputs of atom signatures file generated from the user defined natural products library (\*\*) as well as synthetics (\*\*) and compound libraries (\*\*) and score the compound libraries (\*\*) for

**NP-likeness.** The higher the score the more is the NP-likeness of a molecule. The

<http://www.myexperiment.org/workflows/2121.html>.

- <sup>19</sup>.
- <sup>20</sup>.
- <sup>21</sup>.

Machine learning workers support the significance analysis of single components (i.e. features) of an input vector to obtain smaller inputs with a reduced set of components/features, the partitioning of machine learning data into training and test sets, the construction of input/output mapping model functions and model based predictions as well as result visualization. There is a total of six WEKA-based machine learning methods available: Two workers allow regression as well as classification procedures...

- <sup>22</sup>.
- <sup>23</sup>.
- ... two workers do only support regression...
- 
- <sup>24</sup><sup>25</sup>.

... and two workers are restricted to classification tasks:

- <sup>26</sup>.
- <sup>27</sup>.

For selection of an optimum reduced set of input vector components there are two workers available. The [6](#) illustrates the procedure. The [28](#). In each iteration the single component is discarded that has the smallest influence on the RMSE - up to a last “most significant” component. Figure [7](#) shows a result of a “leave-one-out” analysis and Figure [8](#) depicts the related workflow.

For training and test set partitioning the [28](#):

- 
- 
- 

Figure [9](#) shows a workflow using the [10](#). Classification workers may be used in an equivalent manner. Figure [11](#) depicts diagrams and output of a QSPR analysis to predict HPLC retention times for small molecules: The experimental dataset consists of 183 I/O pairs with a set of molecular descriptors for each small molecule as an input and the corresponding retention time as an output. The molecular descriptors were calculated with the

### 53.3.4 CDK-Taverna 2.0 Wiki

Based on the free MediaWiki framework a Wiki was developed for the CDK-Taverna 2.0 project [29](#). The web page provides general information about the project, documentation about available workers/workflows and on how to create them as well as about installation procedures. The Wiki can be found at [http://cdk-taverna-2.ts-concepts.de/wiki/index.php?title=Main\\_Page](http://cdk-taverna-2.ts-concepts.de/wiki/index.php?title=Main_Page).

---

<sup>19</sup> WEKA API Documentation

<sup>20</sup> Some methods for classification and analysis of multivariate observations

<sup>21</sup> X-means: Extending K-means with an efficient Estimation of the Number of Clusters

<sup>22</sup> NOTITLE!

<sup>23</sup> LIBSVM: a library for support vector machines

<sup>24</sup> Learning with continuous classes

<sup>25</sup> Induction of model trees for predicting continuous classes

<sup>26</sup> Estimating continuous distributions in Bayesian classifiers

<sup>27</sup> NOTITLE!

<sup>28</sup> NOTITLE!

<sup>29</sup> MediaWiki

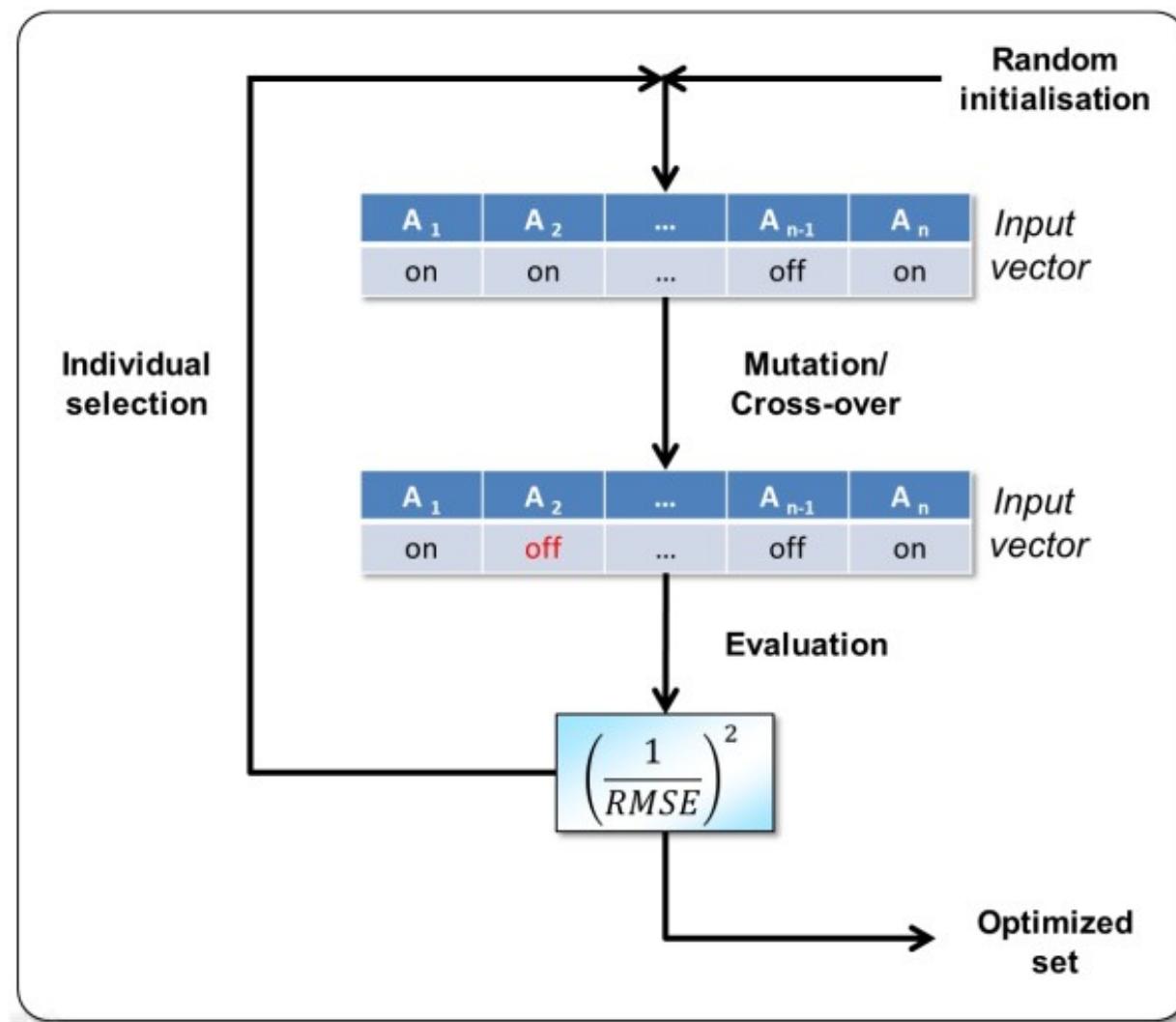


Figure 53.6: Figure 6. Genetic algorithm for selection of an optimum reduced set of input vector components: The algorithm starts with a random population in which each chromosome consists of a random distribution of enabled/disabled (on/off) input vector components denoted  $A_1$  to  $n$   $A$  (where the number of components with “on” status remains fixed during evolution)

A:sub:‘1’ to :sub:‘n’ AGenetic algorithm for selection of an optimum reduced set of input vector components: The algorithm starts with a random population in which each chromosome consists of a random distribution of enabled/disabled (on/off) input vector components denoted . This distribution is changed by mutation and cross-over. The fitness of each chromosome is evaluated by the inverse square RMSE. The selection process for each generation is performed by Roulette wheel selection where chromosomes are inherited with probabilities that correspond to their particular fitness.

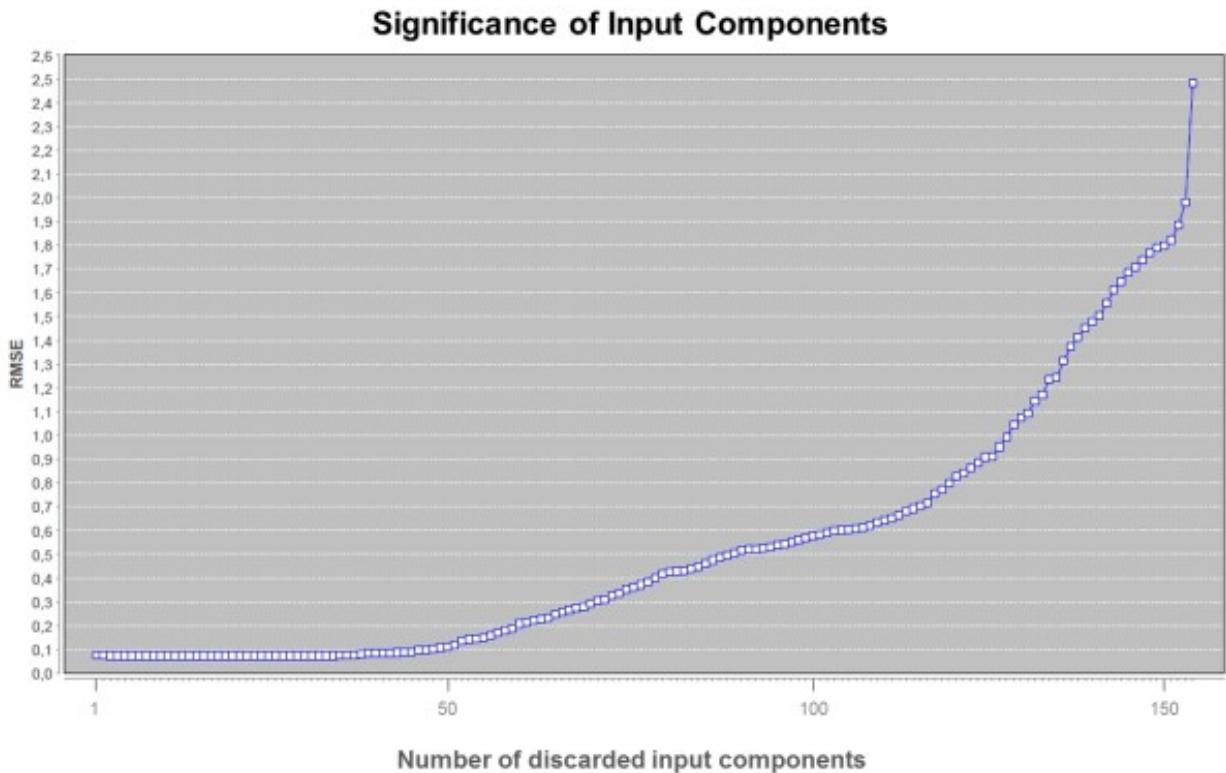


Figure 53.7: Figure 7. “Leave-One-Out” analysis to estimate the significance of input vector components: The root mean square error (RMSE) rises with an increasing number of discarded components (i.e. a decreasing number of input vector components used for the machine filearning procedure)

**“Leave-One-Out” analysis to estimate the significance of input vector components: The root mean square error (RMSE) rises with an increasing number of discarded components (i.e. a decreasing number of input vector components used for the machine filearning procedure).** The relative RMSE shift from step to step may be correlated with the significance of the discarded component. In this case it is shown that the first fifty components do only have a negligible influence on the machine learning result and thus may be excluded from further analysis.

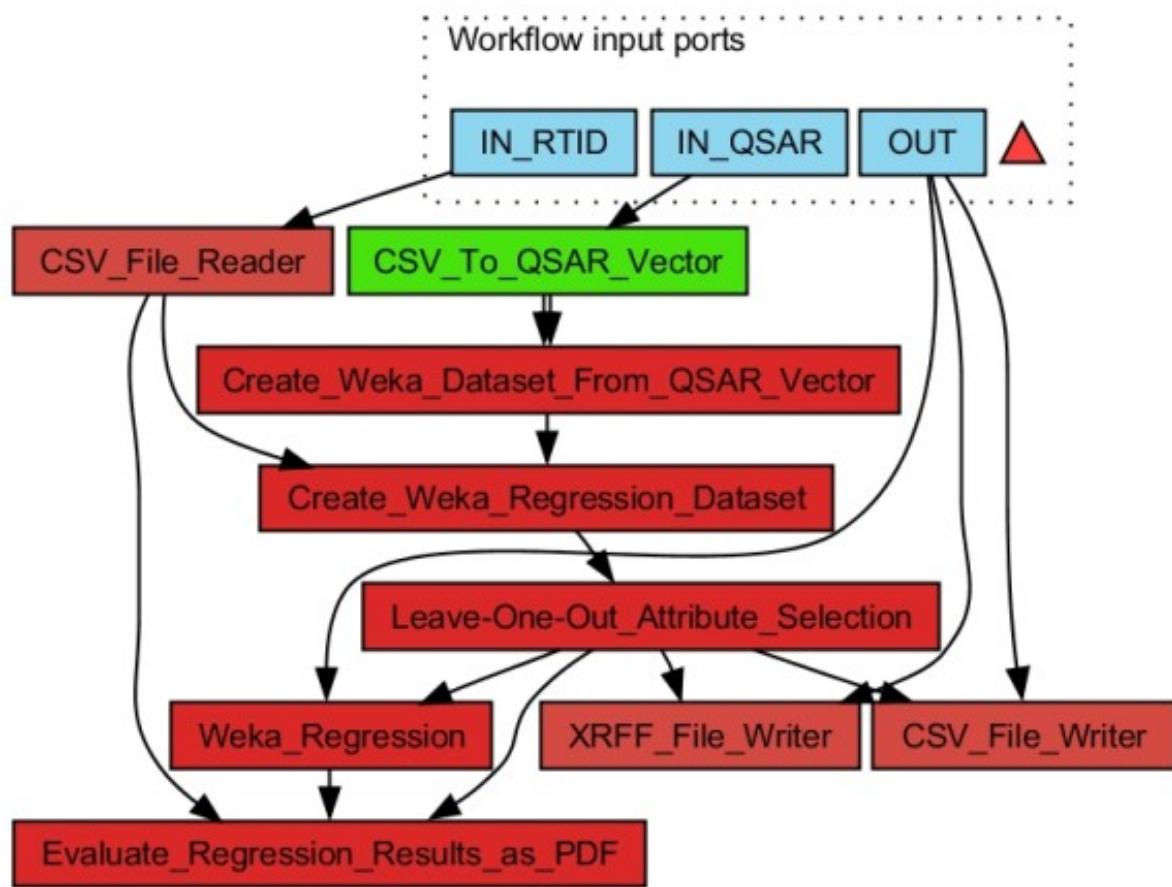


Figure 53.8: Figure 8. Workflow for “Leave-One-Out” analysis: First a regression dataset is generated from a CSV file with UUID and molecular descriptor input data for each molecule (

**Workflow for “Leave-One-Out” analysis:** First a regression dataset is generated from a CSV file with UUID and molecular descriptor input data for each molecule (\*\*\*) and a CSV file containing the UUID of the molecule and the corresponding output (regression) value (\*\*). Then the

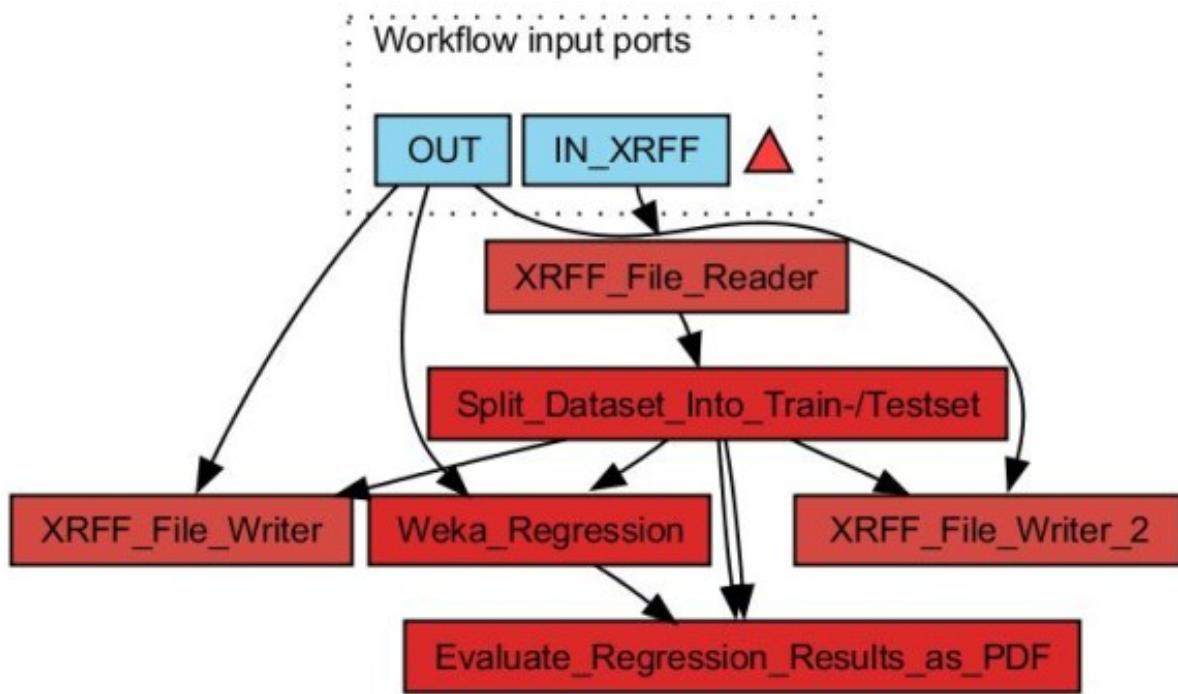


Figure 53.9: Figure 9. Partitioning into training and test set: A regression dataset is split into a training and a test set which is performed by the

**Partitioning into training and test set: A regression dataset is split into a training and a test set which is performed by the**

## 53.4 Conclusions

CDK-Taverna 2.0 provides an enhanced and matured free open cheminformatics workflow solution for the biosciences. It was successfully applied and tested in academic and industrial environments with data volumes of hundreds of thousands of small molecules. Combined with available workers and workflows from bioinformatics, image analysis or statistics CDK-Taverna supports the construction of complex systems biology oriented workflows for processing diverse sets of biological data.

## 53.5 Competing interests

The authors declare that they have no competing interests.

## 53.6 Authors' contributions

EW initiated the integration of Taverna and CDK and supported deployment and architecture. CS and AZ conceived the project and lead the further development. SN supported the reaction enumeration enhancements. KV provided workers for molecular fragmentation. AT did the majority of CDK-Taverna re-engineering and enhancements and developed the project to its current state. All co-authors contributed to the manuscript. All authors read and approved the final manuscript.

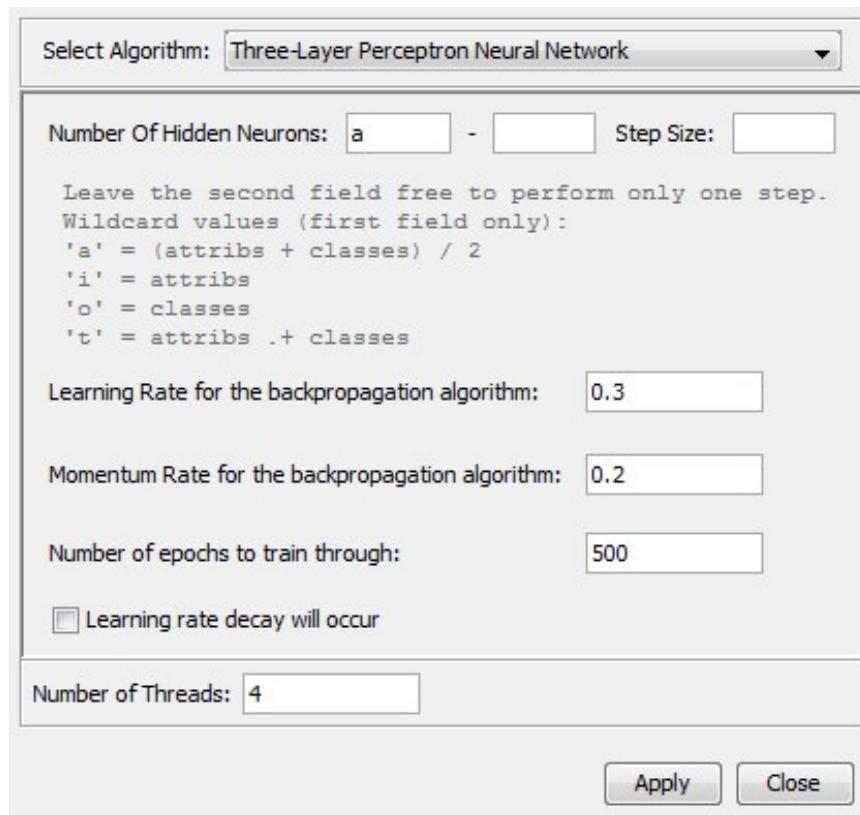


Figure 53.10: Figure 10. Configuration panel for the Weka Regression worker: The configuration for a three-layer perceptron neural networks is selected. Each machine learning method consists of a parameter panel for individual configuration.

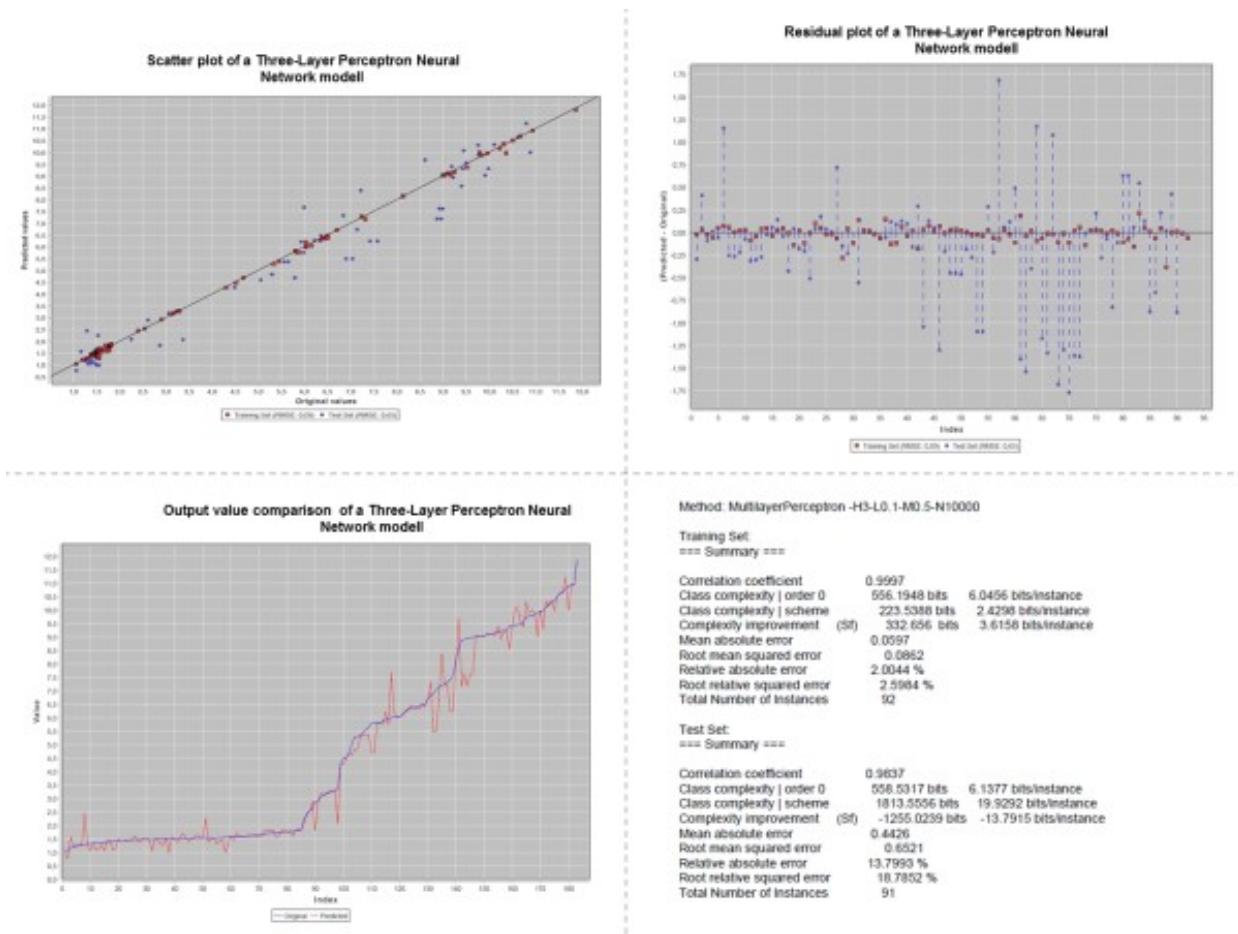


Figure 53.11: Figure 11. Diagrams for machine learning results

**Diagrams for machine learning results:** (upper left) Scatter plot with experimental versus predicted output values. (upper right) Residuals plot with differences between the predicted and experimental output values. (lower left) Experimental output data are plotted over corresponding sorted predicted output data. (lower right) Characteristic quantities of the predicted model.

## 53.7 Acknowledgements

The authors express their gratitude to the teams and communities of Taverna, CDK and WEKA for creating and developing these open tools.