A *DE FACTO* STANDARD OR A FREE-FOR-ALL?

A BENCHMARK FOR READING SMILES

Noel O'Boyle, John Mayfield and Roger Sayle

NextMove Software

https://github.com/nextmovesoftware/smilesreading

Twitter: @baoilleach

INTRODUCTION

SMILES

- **S**implified **M**olecular **I**dentification and **L**ine **E**ntry **S**ystem
- 30 years since Dave Weininger published:
  - *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,* JCICS, 1988, 28, 31
    - Cited 2389 times, 179 times in 2017 (Google Scholar)
- 31 years since an EPA report on SMILES
  - *SMILES: A line notation and computerized interpreter for chemical structures*, EPA/600/M-87/021, Aug 1987 (available online)

A *DE FACTO* STANDARD...

- Since 1987, Daylight Chemical Information Systems have been responsible for developing SMILES
  - The documentation is freely available online

- Some aspects of the language were clarified by the OpenSMILES specification (drafted 2007)

- One of the most widely-used file formats in cheminformatics
  - Compact, convenient, combineable, can be canonicalized

…OR "FREE-FOR-ALL" ?

- Toolkits can generally read their own SMILES strings
  - But can they read SMILES strings from other programs?
- Typical problems
  - Hydrogens appear and/or disappear when moving from X to Y
  - SMILES written by X are rejected by Y
  - One toolkit says a ring system is aromatic while another says it's not
- Why? Is it a problem? If so, can anything be done?
  - Something fundamentally broken with SMILES?

MY GOAL

- Improve the interoperability of SMILES strings among all tools that support SMILES

1. Identify the most common issues affecting interoperability
   - Differences in interpretation of spec? Errors? Unusual aromaticity models? Counting hydrogens?
2. Work with developers to highlight and resolve issues
3. Provide a resource for future implementations to use to avoid the pitfalls of the past

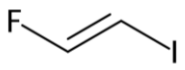
SCOPE OF BENCHMARK

- The benchmark is *not* a SMILES validation suite
  - Does not replace a toolkit's own testing
- Focuses on aspects of SMILES syntax that tend to be incorrectly implemented
  - Stereochemistry
  - Implicit valence
  - (Reading) Aromatic SMILES
- Focuses on SMILES reading
  - Can toolkits agree on the meaning of a SMILES string?

STEREOCHEMISTRY

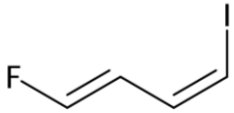
TESTING CISTRANS STEREO

- *cistrans.smi*: 126 SMILES strings



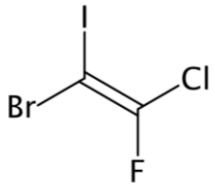
`F/C=C/I`

22 variants



`F/C=C/C=C\I`

24 variants



`I/C(/Br)=C(/F)\Cl`

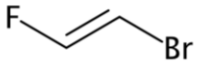
80 variants

- Toolkits must read each SMILES string and write it out again in a canonical form
  - Exception: ChemDraw writes out an IUPAC name which I run through OPSIN to create a SMILES
- Compare the answers with a toolkit of your choice

STEREO RESULTS

- Currently, the benchmark focuses on:
  - Tetrahedral and cis-trans
- To come:
  - Octahedral, extended tetrahedral, square-planar, trigonal bipyramidal

- The main problem observed is handling of stereo at bond closure digits:
  - `F/C=C/1.Br1` should be the same as `F/C=C1.Br\1`



  - `[C@@](O)(Cl)(F)Br` should be the same as `[C@@](O)(Cl)(F)1.Br1` (if supported)

IMPLICIT VALENCE

THE SMILES IMPLICIT VALENCE MODEL

- A SMILES string completely describes the molecular formula of a molecule, including hydrogens

THE SMILES IMPLICIT VALENCE MODEL

- A SMILES string completely describes the molecular formula of a molecule, including hydrogens
- The valence model tells you how to read/write SMILES that leave out the explicit hydrogen count on certain atoms (B, C, N, O, P, S, halogens)
  - co should be read as CH_3OH, i.e. methanol
  - Similarity, methanol, CH_3OH may be written as co
- Same rules must be used for reading as for writing
  - This has consequences for interoperability
- The rules are in the docs…

THE SMILES IMPLICIT VALENCE MODEL

Element	Valence
B	3
C	4
N	3 or 5
O	2
P	3 or 5
S	2, 4 or 6
halogens	1

- Not the same as the MDL valence model
- Not the same as “how many hydrogens is an atom with this valence likely to have?”

APPLY THE VALENCE MODEL

Element	Valence
B	3
C	4
N	3 or 5
O	2
P	3 or 5
S	2, 4 or 6
halogens	1

CCl

- The carbon has explicit valence of 1
  - Round up to **4** with hydrogens
- The molecule is CH_3Cl

- 15 programs tested (with default options):
  - 15 say CH_3Cl

APPLY THE VALENCE MODEL

Element	Valence
B	3
C	4
N	3 or 5
O	2
P	3 or 5
S	2, 4 or 6
halogens	1

Cl(C)C

- The chlorine has explicit valence of 2
  - Round up to 1 with hydrogens
- The molecule is $Cl(CH_3)_2$

- 15 programs tested (with default options):
  - 2 reject the molecule (bad valence)
  - 9 say $Cl(CH_3)_2$
  - 3 say $HCl(CH_3)_2$
  - 1 says $H_5Cl(CH_3)_2$

APPLY THE VALENCE MODEL

Element	Valence
B	3
C	4
N	3 or 5
O	2
P	3 or 5
S	2, 4 or 6
halogens	1

N(C)(C)(C)C

- The nitrogen has explicit valence of 4
  - Round up to **5** with hydrogens
- The molecule is $HN(CH_3)_4$

- 15 programs tested (with default options):
  - 2 reject the molecule (bad valence)
  - 10 say $HN(CH_3)_4$
  - 3 say $N(CH_3)_4$

BENCHMARK DATASET

- 61 SMILES strings covering the organic subset
  - Here are the testcases for nitrogen

Atom Type	SMILES	SMILES valence	H Count
N0	N	3	3
N1	NC	3	2
N2	N(C)C	3	1
N3	N(C)(C)C	3	0
N4	N(C)(C)(C)C	5	1
N5	N(C)(C)(C)(C)C	5	0
N6	N(C)(C)(C)(C)(C)C	5	0

Element	Valence
N	3 or 5

DISAGREEMENTS WITH SMILES VALENCE MODEL

Avalon	Cl2 Cl4 Br2 Br4 I2 I4
BIOVIA Draw	Cl2 Cl4 Br2 Br4 I2 I4
Cactvs	N4 P4 S3 S5 (or none*)
CDK	
CEX (Weininger)	
ChemDoodle	
ChemDraw	
Indigo†	
iwtoolkit	N4 Cl2 Cl3 Cl4 Cl5 Br2 Br3 Br4 I2 I4 (or P4 S3 S5*)
JChem	
KnowItAll	
OEChem	
Open Babel	
OpenChemLib	N4 Cl2 Cl4 Br2 Br4 I2 I4
RDKit†	P6 I3 I4

"Happy valence models are all alike; every unhappy valence model is unhappy in its own way."

...with apologies to Tolstoy

'9.5'/15 correct now.
When I started, it was 6/15.

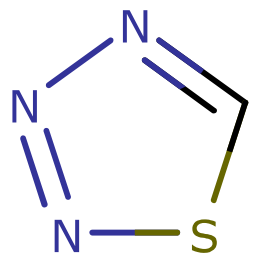
** If the default options are modified*

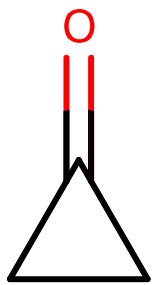
† Results exclude 17 atom types rejected by Indigo, and 19 rejected by RDKit

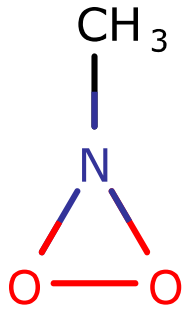
READING AROMATIC SMILES

ChEMBL23

47464 ring systems







O1ON1C
C1C(=O)C1
C1=NN=NS1

12 benchmark datasets

CDK

o1on1C
C1C(=O)C1
c1nnns1

OpenChemLib

CN1OO1
O=C1CC1
c1nnns1

RDKit

o1on1C
C1C(=O)C1
c1nnns1

+9 others

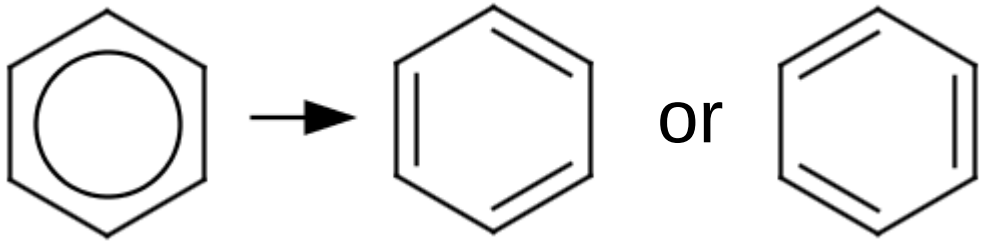
TEST ABILITY TO READ AROMATIC SMILES

- How to read an aromatic SMILES? Here's my way
  - https://baoilleach.blogspot.com/2017/08/my-acs-talk-on-kekulization-and.html

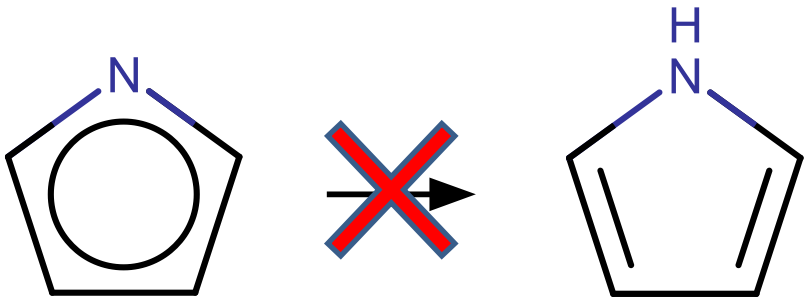
- How to test whether two programs have interpreted an aromatic SMILES the same way?
  1. Do they agree on whether it is kekulizable?
     - Either it is, or it isn't
  2. If kekulizable, do they agree on the hydrogen count on each atom?
     - The hydrogen count is independent of the specific Kekulé form

KEKULIZATION

c1ccccc1



c1cncc1



- Given a molecule where some atoms and bonds have been marked as aromatic
  - Assign bond orders of either one or two to the aromatic bonds such that the valencies of all of the aromatic atoms are satisfied (i.e. are consistent with sp^2)

13 toolkits tested reading aromatic SMILES		12 benchmark datasets as input											
		1	2	3	4	6	7	8	9	10	11	12	13
Avalon	1												
BIOVIA Draw	2												
CDK	3												
ChemDoodle	4												
ChemDraw	5												
Indigo	6			**13 x 12 x 47464**									
iwtoolkit	7												
JChem	8					**results**							
KnowItAll	9												
OEChem	10												
Open Babel	11												
OpenChemLib	12												
RDKit	13												

13 toolkits tested reading aromatic SMILES		12 benchmark datasets as input											
		1	2	3	4	6	7	8	9	10	11	12	13
Avalon	1												
BIOVIA Draw	2												
CDK	3												
ChemDoodle	4												
ChemDraw	5												
Indigo	6												
iwtoolkit	7												
JChem	8												
KnowItAll	9												
OEChem	10												
Open Babel	11												
OpenChemLib	12												
RDKit	13												

	Different H Count	Kekulization Failure
Avalon	0	1
BIOVIA Draw	0	0
CDK	Reference	0
ChemDoodle	13*	
ChemDraw	37	26
Indigo†	456	23
iwtoolkit	91	69
JChem	0	4
KnowItAll	0	N/A
OEChem	0	0
Open Babel	0	0
OpenChemLib	9	136
RDKit†	7	1

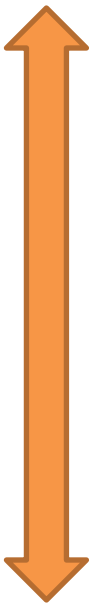
* It is not possible to distinguish between kekulization failures and differences in hydrogen count

† Results exclude 8 structures rejected by Indigo, and 15 by RDKit

13 toolkits tested reading aromatic SMILES

12 benchmark datasets as input

		1	2	3	4	6	7	8	9	10	11	12	13
Avalon	1												
BIOVIA Draw	2												
CDK	3												
ChemDoodle	4												
ChemDraw	5												
Indigo	6												
iwtoolkit	7												
JChem	8												
KnowItAll	9												
OEChem	10												
Open Babel	11												
OpenChemLib	12												
RDKit	13												



CORNER CASES FROM CDK VS RDKIT

- Avalon SMILES C1CCCCN2/C(=N\1)\CN=C2
  - Parse error with CDK, warning with RDKit
- Avalon SMILES O=s1(cccc1)=O
  - Kekulization failure with CDK, read by RDKit
- BIOVIA Draw SMILES CN1CCCNp12np(S)(S)np(Cl)(Cl)n2
  - Kekulization failure with RDKit, read by CDK
- Indigo SMILES c12:c:c:[Te]:c1cccc2
  - RDKit warns about aromatic bond to non-aromatic atom, CDK treats as aromatic
- iwtoolkit SMILES [Te++++]1[O-]CC[O-]1
  - Parse error with RDKit, read by CDK

CORNER CASES FROM CDK VS RDKIT

- iwtoolkit SMILES
  c12c3c4c5C6c3c3c7c8C6c6c9c8c8c%10c%11c%12c8c8c9c9c%13c6=c5c5c%13c6c%13c%14c%15C(c4c5%14)C1c1c4c2c3c(c7%10)c2c4c3c4c1=c%15c1c%13C57c6c9c8c6C5([N+]5(C)CCN7CC5)c(c14)c(c6%12)c3c2%11
  - Kekulization failure with RDKit, read by CDK
- Open Babel SMILES n1c2c3cccc4cccc(c34)c2n(=N)c2ccccc12
  - Kekulization failure with RDKit, read by CDK
- OpenChemLib SMILES CN(c1c(cc[nH]2)c2n2(CCCCCC2)cc1N1)C1=O
  - Kekulization failure with RDKit, read by CDK
- OpenChemLib SMILES
  C(/C(/c1/c2cccc1)=C1)/C2=C\C(C/C2)=N/C2=C/C(C/C2)=N/C2=C/c2ccc1[nH]2
  - Different hydrogen count for atoms 3 and 4 between RDKit and CDK

A similar analysis could be done for any pair of toolkits

FINAL THOUGHTS

WORK IN PROGRESS

- These results represent a snapshot in time
  - Much better than six months ago
- Software that has been (or will be) changed in response to this benchmark

CACTVS	ChemDoodle	Open Babel
CCDC Toolkit	iwtoolkit	OpenChemLib
CDK	JChem	Pipeline Pilot
Ceres (LHASA)	KnowItAll	RDKit

- RDKit issues #1972, #1928 (fixed), #1569, #1900 (WIP)
- Encourage your favourite tools/toolkits to take part!

CONCLUSIONS

- While stereochemistry is well-handled, adherence to the SMILES valence model and ability to read aromatic SMILES tend to be problem areas
  - Checking for agreement in hydrogen count is a surprisingly powerful way of identifying errors
- While disagreement exists, all is not lost
  - On inspection, it has always been clear what the correct answer is
- Developers are (mostly) open to addressing issues
  - The only area of push-back from developers is implementing the SMILES valence model
- Overall, much more successful than expected!

ACKNOWLEDGEMENTS

- Greg Landrum and developers of other toolkits
- Matt Swain for providing JChem results
- Roger Sayle, John Mayfield

Datasets, results and scripts are available at:
https://github.com/nextmovesoftware/smilesreading

noel@nextmovesoftware.com

How many hydrogens are on the nitrogen in the molecule represented by this SMILES string?

`N(C)(C)(C)C`

1. None
2. One
3. It depends
4. Cannot say as no such molecule

3.2.1 Atoms

Atoms are represented by their atomic symbols: this is the only required use of letters in SMILES. Each non-hydrogen atom is specified independently by its atomic symbol enclosed in square brackets, **[ ]**. The second letter of two-character symbols must be entered in lower case. Elements in the "organic subset" **B, C, N, O, P, S, F, Cl, Br,** and **I** may be written without brackets if the *number of attached hydrogens conforms to the lowest normal valence consistent with explicit bonds*. "Lowest normal valences" are B (3), C (4), N (3,5), O (2), P (3,5), S (2,4,6), and 1 for the halogens. Atoms in aromatic rings are specified by lower case letters, e.g., aliphatic carbon is represented by the capital letter **C**, aromatic carbon by lower case **c**. Since attached hydrogens are implied in the absence of brackets, the following atomic symbols are valid SMILES notations.

C	methane	(CH4)
P	phosphine	(PH3)
N	ammonia	(NH3)
S	hydrogen sulfide	(H2S)
O	water	(H2O)
Cl	hydrochloric acid	(HCl)

Atoms with valences other than "normal" and elements not in the "organic subset" must be described in brackets.

[**S**]	elemental sulfur
[**Au**]	elemental gold

Within brackets, any attached hydrogens and formal charges must always be specified. The number of attached hydrogens is shown by the symbol **H** followed by an optional digit. Similarly, a formal charge is shown by one of the symbols **+** or **-**, followed by an optional digit. If unspecified, the number of attached hydrogens and charge are assumed to be zero for an atom inside brackets. Constructions of the form