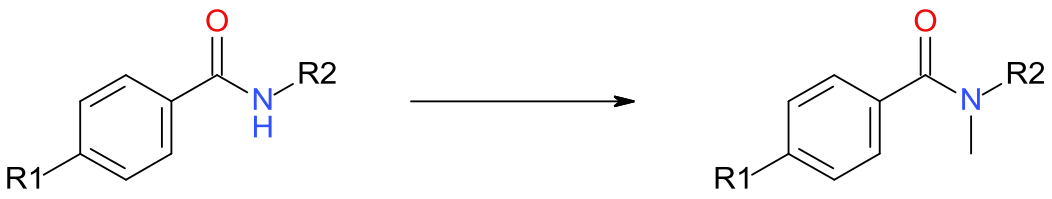
mmpdb
A MMP tool for large ADMET datasets

Christian Kramer

Roche pRED

Purpose

Much of **MedChem Knowledge** can be captured by **Matched Molecular Pairs** (MMPs)



ΔlogD = ?
ΔhERG = ?
ΔClearance = ?
…

An integrated MMP database tool shall allow to:

- Mine inhouse databases for MedChem knowledge in terms of **MMP rules**
- **Apply MMP rules** to new compounds
- Browse rules and pairs **to study MedChem assumptions**
- **Exchange knowledge** without exchanging compounds between companies

MMP – Brief History

Pre 2010: inhouse MMP implementations of various sorts (WiZePairs - AZ, T-ANALYZE – Merck, Lucid – Roche, …)

2010: Jameed Hussain Ceara Rea: Fragment and Index Algorithm published (JCIM)

2010: Papadatos et al.: Local Environment of MMPs is critical for assembling rules (JCIM)

2012: Open source Fragment and Index implementation available in RDKit

2013 - 2016: MedChemica-led SALT consortium (AZ, Genentech, Roche)

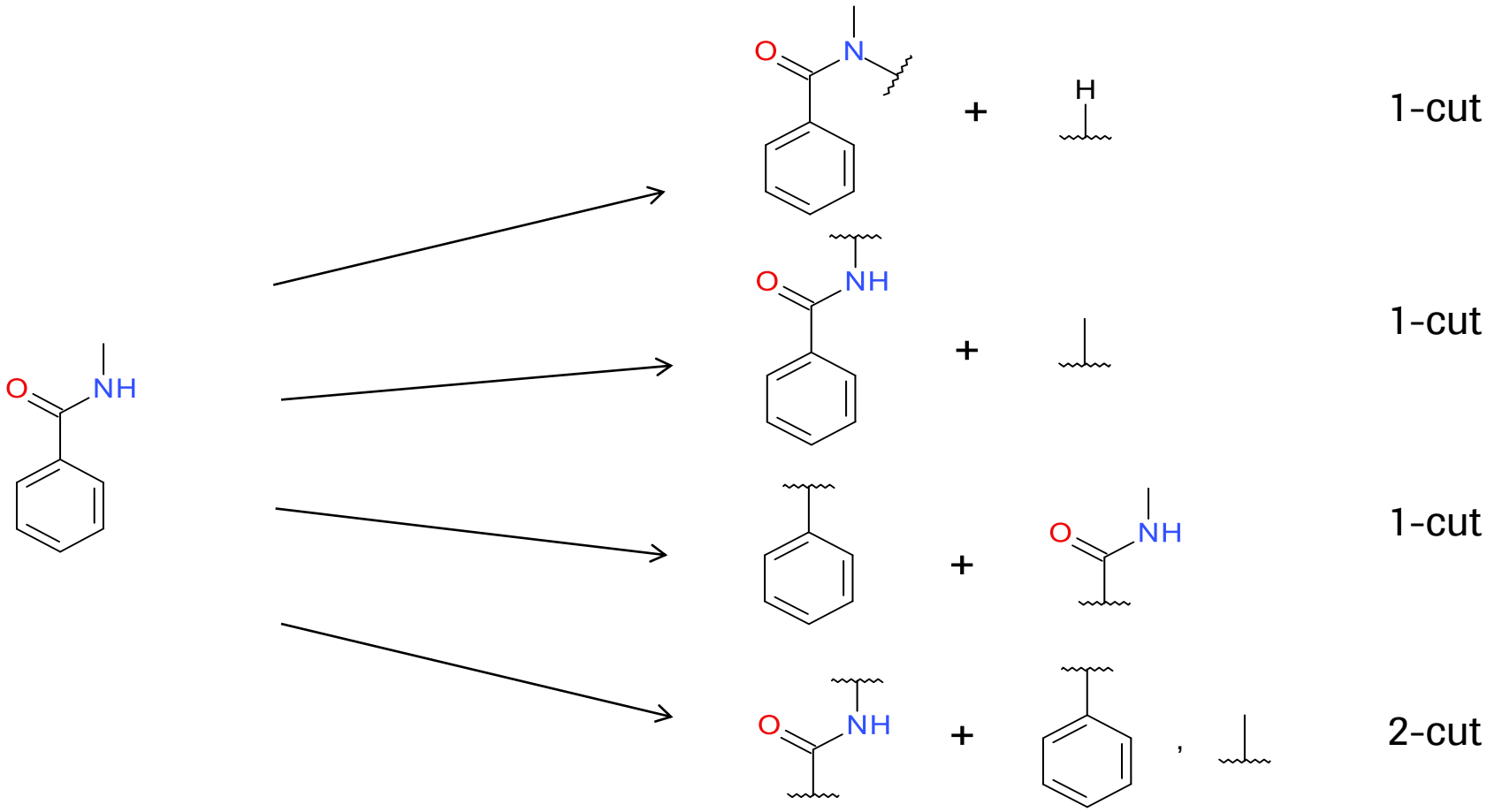
2016: MOEsaic as first commercially available MMP GUI

2016: MMPDB V1.0 (inconsistencies in canonicalization, chirality treatment, no usage of rule statistics)

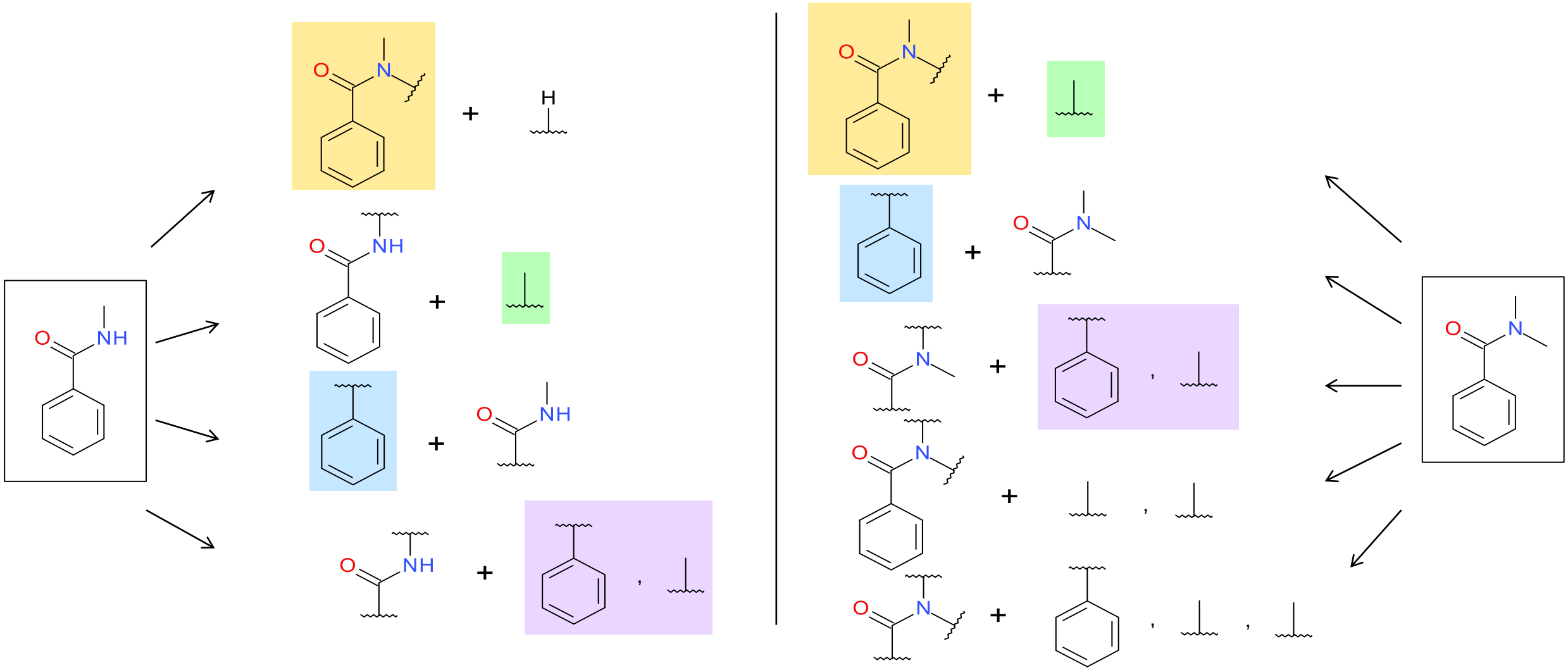
2017: MMPDB V2.0

MMP Fragment & Index – Basic Concept
Fragmentation

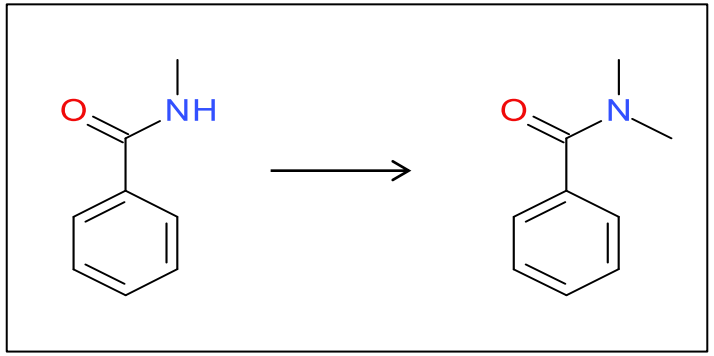
Generate 1,2,3 –cut Fragmentations on single, non-ring, non-amide bonds



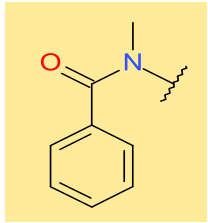
MMP Fragment & Index – Basic Concept Index

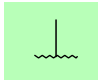


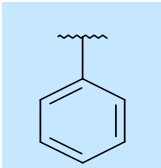
MMP Fragment & Index – Basic Concept
Enumerate transformations

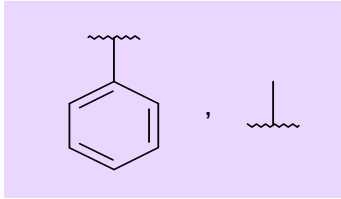


Constant

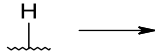


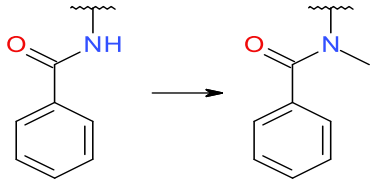


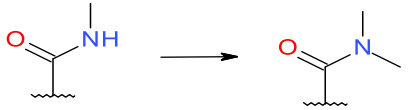


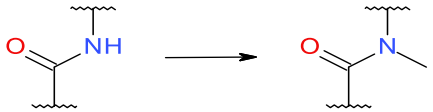


Transformation





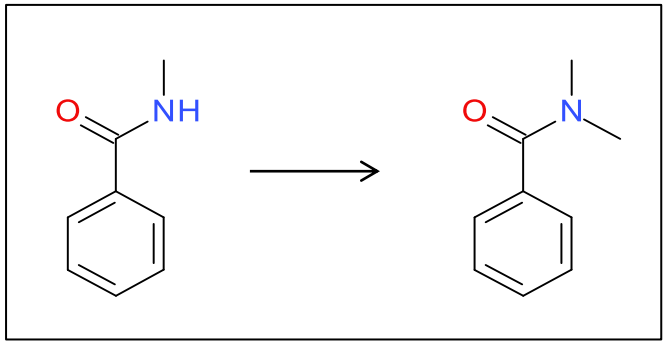




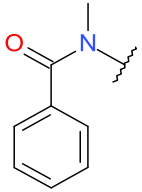
mmpdb – Environment Representation

The chemical environment around an exchanged fragment can have strong influence on the effect. We represent the environment by rooted standard circular fingerprints, encoding atoms up to 5 bond distance from the attachment site.

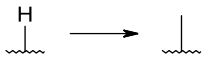
Pair



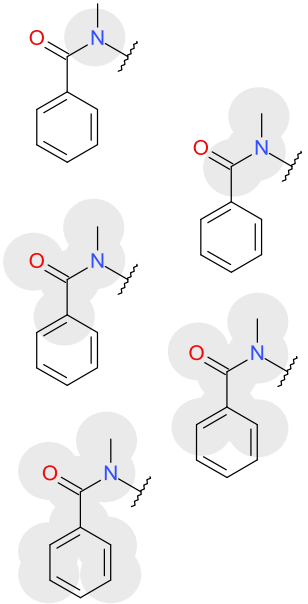
Constant



Transformation



Environment



1

2

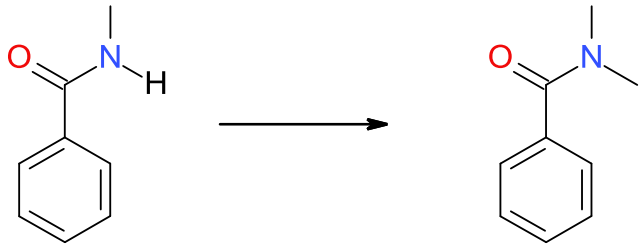
3

4

5

mmpdb - Terminology

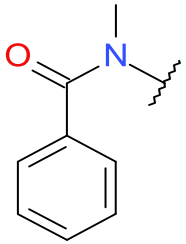
Pair



Compound A
Left Hand Side (LHS)

Compound B
Right Hand Side (RHS)

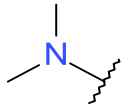
1



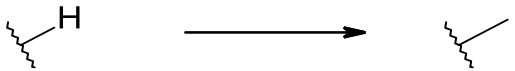
constant

Variable A

Variable B



Environment
at radius 2

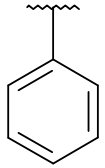


Transformation for A >> B

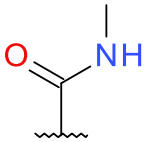
[H][*:1] >> [*:1]C

SMIRKS A>>B

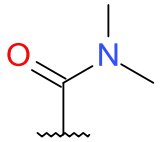
2



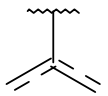
constant



Variable A



Variable B



Environment
at radius 2

Transformation for A >> B

[*:1]C(=O)NC >> [*:1]C(=O)N(C)C

SMIRKS A>>B

...

MMPDB 2.0 can…

… **fragment**, **index**, and **upload** MMP data into a SQLite or Oracle database

… consider **environments up to 5 atoms** away from the cut-atom(s) on the constant part

… be **highly customized** in terms of fragmentation and MMP matching options

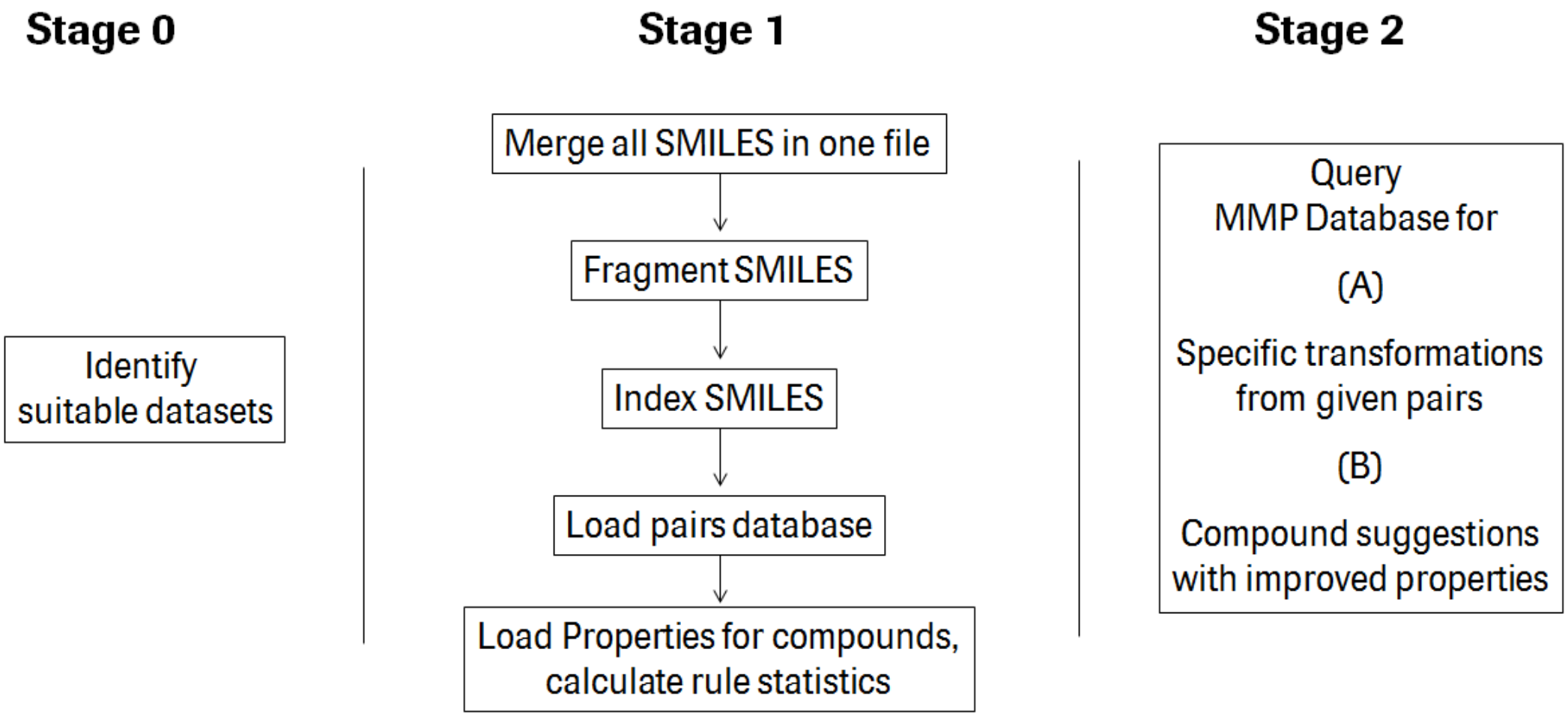
… create, filter, and suggest **hundreds of thousands of novel compounds** for a given input structure **within seconds**

… correctly **handle complicated symmetry** - and **chirality** creation cases

… be accessed through a **web service**

… be extended due to its **modular python architecture**

DB setup - workflow



Fragmentation

Parallelized and simplified Version of RDKit Hussain & Rea fragmentation implementation

For each fragmentation, six different environment fingerprints (radius 0-5) are generated.

Fragment format, chirality issues, and canonicalization solution will be discussed by Andrew.

Options

```
--max-heavies       Maximum number of non-hydrogen atoms, or 'none‘ (default: 100)
--max-rotatable-bonds          Maximum number of rotatable bonds (default: 10)
--rotatable-smarts  SMARTS pattern to detect rotatable bonds (default:
                    '[!$([NH]!@C(=O))&!D1&!$(*#*)]-&!@[!$([NH]!@C(=O))&!D1&!$(*#*)]')
--salt-remover      File containing RDKit SaltRemover definitions. The default ('<default>') uses
                    RDKit's standard salt remover. Use '<none>' to not remove salts.
--cut-smarts        alternate SMARTS pattern to use for cutting (default:
                    [#6+0;!$(*=,#[!#6])]!@!=!#[!#0;!#1;!$([CH2]);!$([CH3][CH2])]'), or use one of:
                    'default', 'cut_AlkylChains', 'cut_Amides', 'cut_all', 'exocyclic',
                    'exocyclic_NoMethyl'
--num-cuts          number of cuts to use (default: 3)
--cache             get fragment parameters and previous fragment information from SOURCE
--num-jobs          number of jobs to process in parallel (default: 4)
-i                  input structure format (one of 'smi', 'smi.gz')
--delimiter         SMILES file delimiter style (one of 'whitespace‘ (default), 'to-eol', 'comma',
                    'tab', or 'space')
--has-header        skip the first line, which is the header line
--output            save the fragment data to FILENAME (default=stdout)
--out               output format. One of 'fragments' or 'fragments.gz'. If not present, guess from
                    the filename, and default to 'fragments'
```

Indexing & DB creation

Lots of Filters for specifying which pairs to form (and which not)

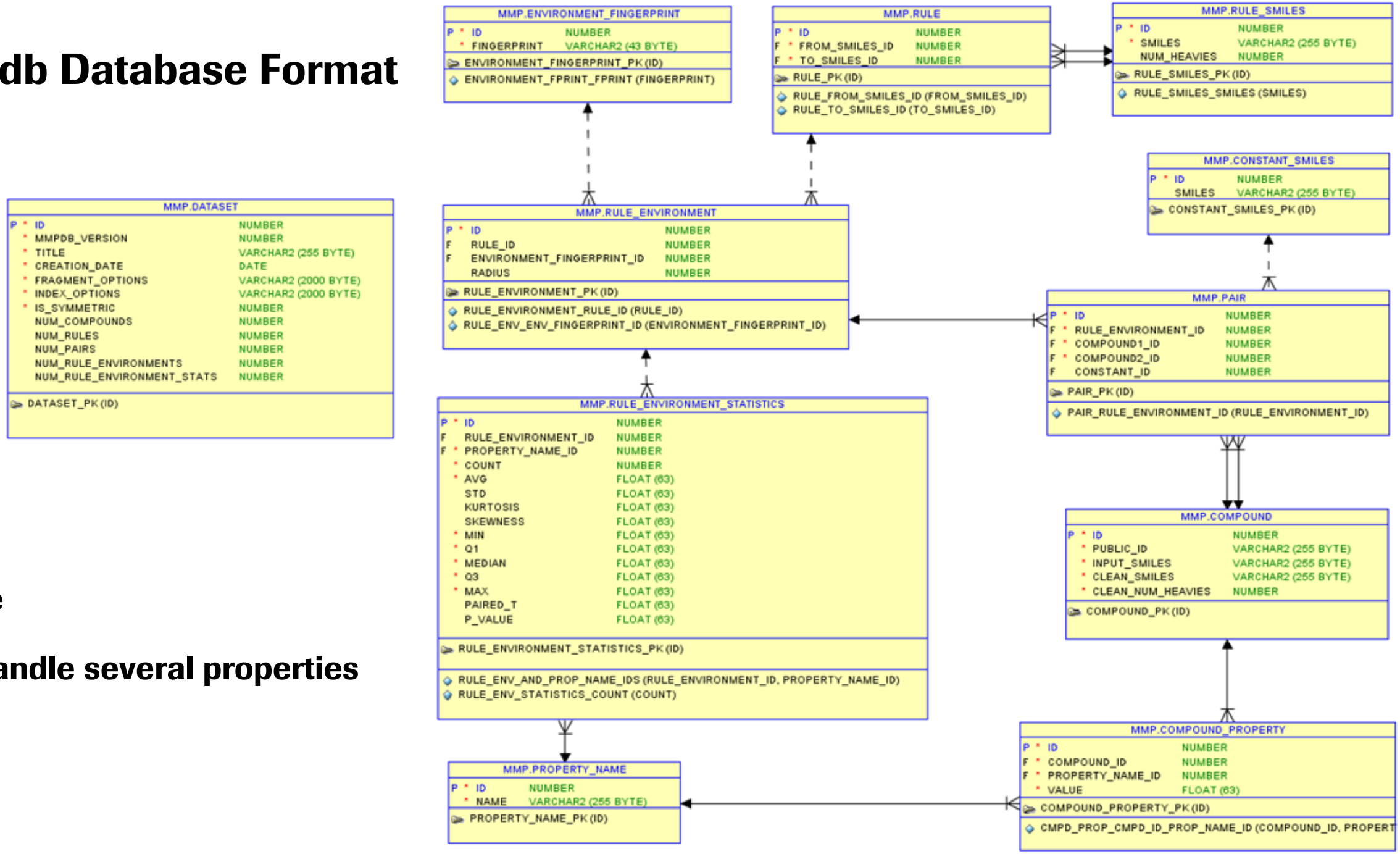
Output options include .csv (for display and browsing in SpotFire/ Vortex) and .mmpdb, i.e. direct DB creation

Properties can be loaded here or in a separate step

Options

--min-variable-heavies	Minimum number of non-hydrogen atoms in the variable fragment. Default ('none')
--max-variable-heavies	Maximum number of non-hydrogen atoms in the variable fragment
--min-variable-ratio	Minimum ratio of variable fragment heavies to heavies in the (cleaned) structure
--max-variable-ratio	Maximum ratio of variable fragment heavies to heavies in the (cleaned) structure
--max-heavies-transf	Maximum difference in the number of heavies transferred in a transformation
--max-frac-trans	Maximum difference in the number of heavies transferred in a transformation
--symmetric	Output symmetrically equivalent MMPs, i.e. output both cmpd1,cmpd2, SMIRKS:A>>B and cmpd2,cmpd1, SMIRKS:B>>A
--properties	File containing the identifiers to use and optional physical properties
--output	save the fragment data to FILENAME (default=stdout)
--out	output format. One of 'mmpdb' (default), 'csv', 'csv.gz', mmpa' or 'mmpa.gz'If not present, guess from the filename, and default to 'mmpdb'
--title	a short description of the dataset. If not given, base the title on the filename
--memory	report a summary of the memory use

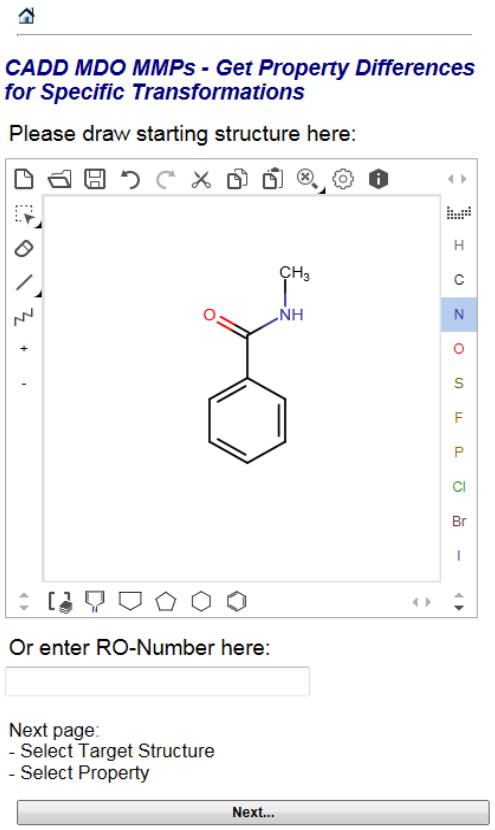
mmpdb Database Format

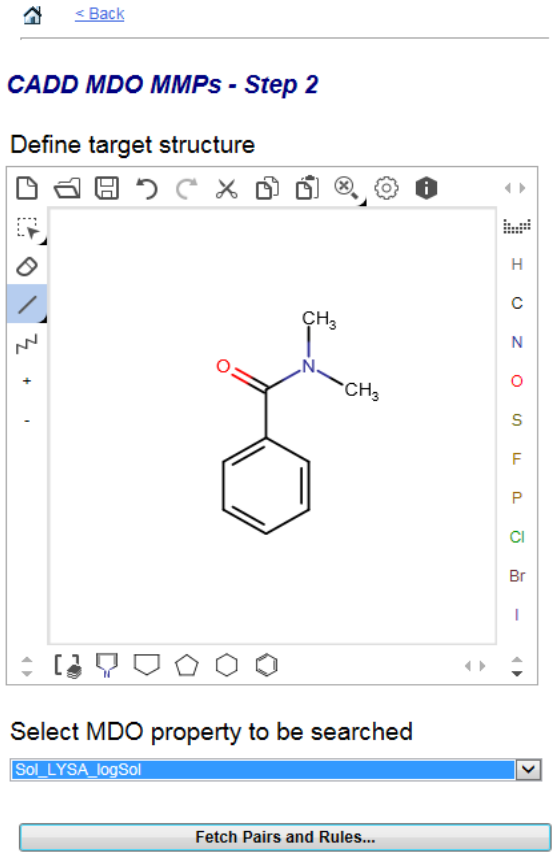


SQLite

Can handle several properties

Finding the change in a MDO property for a given pair: “Predict Differences”





→ Analyze output

1.) Draw starting molecule or enter RO number
→ click “Next”

2.) Draw target molecule

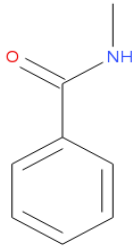
3.) Select MDO property
→ click “Fetch Pairs and Rules”

mmpdb – Use Case 1: predict

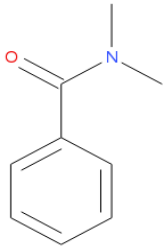
Discngine Web Panel

Your Query was

Starting Structure



Target Structure



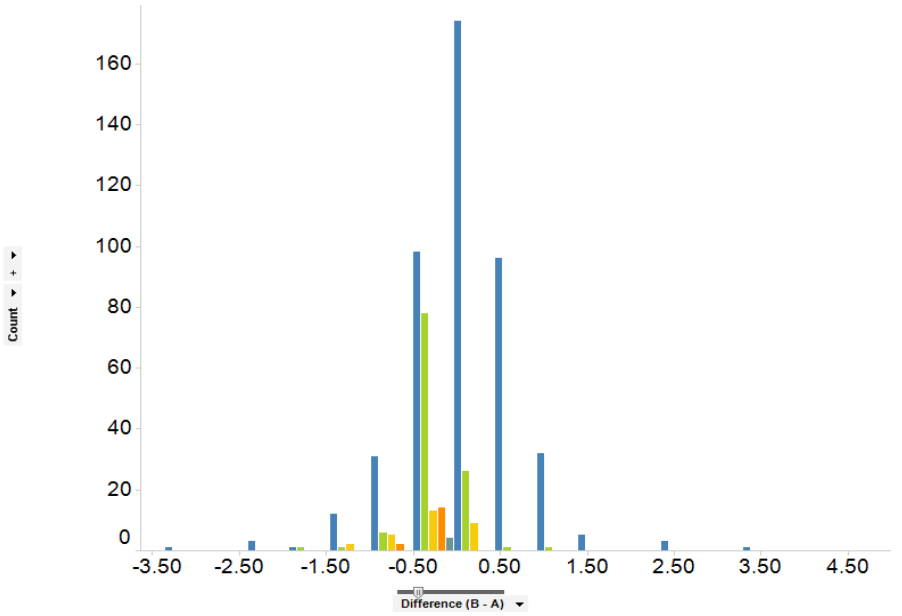
Query Property: logD

< Perform another search

Rules

SMIRKS	Radius	count	avg	std
H–A ⟶ A—	1	538	0.13	0.60
H–A ⟶ A—	2	123	-0.17	0.32
H–A ⟶ A—	3	41	-0.30	0.29
H–A ⟶ A—	4	20	-0.31	0.16
H–A ⟶ A—	5	4	-0.19	0.11

Difference Distribution



Select Environment Radius

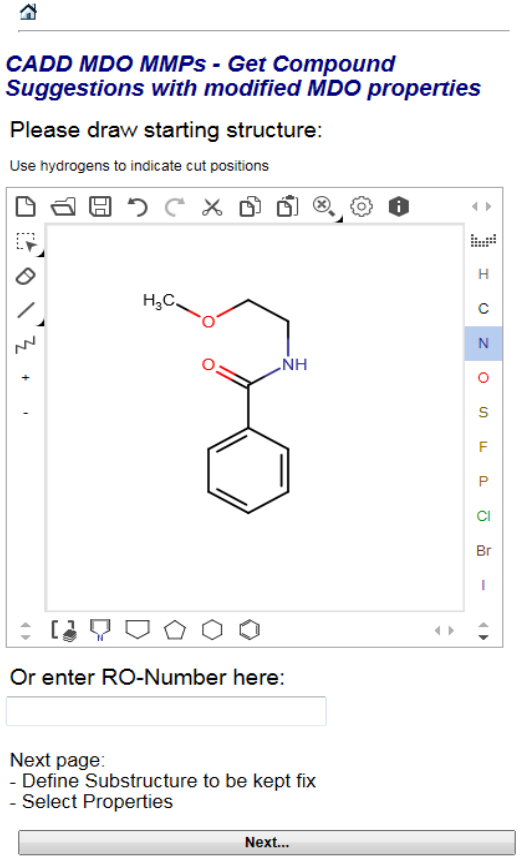
- [ ] 0
- [x] 1
- [x] 2
- [x] 3
- [x] 4
- [x] 5

Pairs

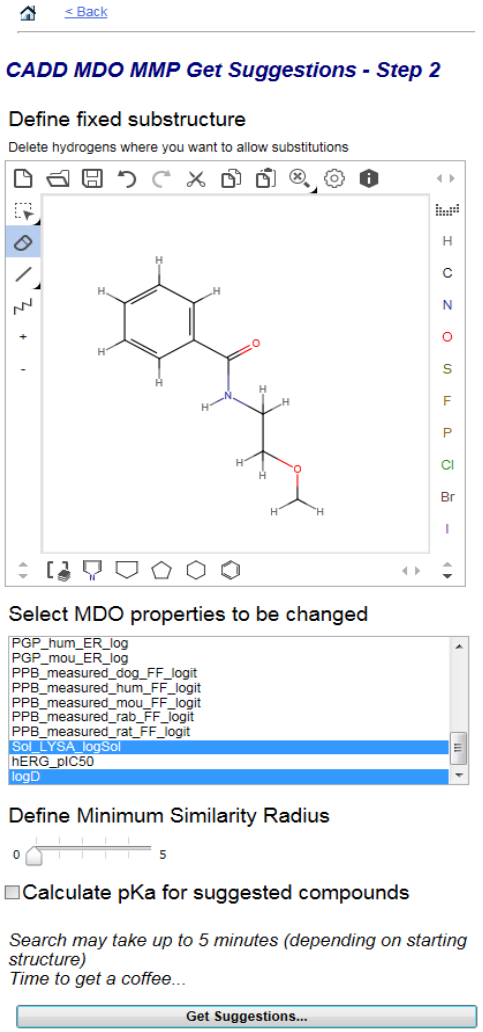
SMIRKS	ID1	ID2	SMILES1	SMILES2	max_R...	prop_1	prop_2	Delta
H–A ⟶ A—	[redacted]				1	2.39	2.26	-0.13
H–A ⟶ A—					1	3.09	3.50	0.41
H–A ⟶ A—					3	1.49	1.47	-0.02
					1	1.49	1.47	-0.02

Online | 14 of 14 rows | 0 marked | 15 columns | pred_detail_rules

Looking for transformations applied to your compound + MDO changes?



1.) Draw starting molecule or enter RO number
→ click “Next”



2.) Define substructure that shall exist in suggested molecules (remove hydrogens and heavy atoms to allow changes)

3.) Select MDO properties of interest and pKa if desired

4.) Higher Similarity radius: Less suggestions, used rules are based on more similar pairs
→ click “Get Suggestions”

→ Analyze results

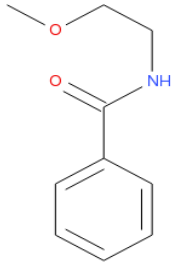
mmpdb – Use case 2: Transform

Page | MMP_suggestions

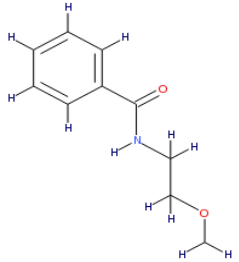
Discngine Web Panel

Your Query was

Query Structure



Substructure to be kept fix



Query Properties:
Sol_LYSA_logSol,logD

Minimum Radius of Similarity from anchoring Atoms: 1

< Perform another search

Switch to MMP suggestions table

Ideas from MMP search

ID	SMILES	transformation	avg_Sol_...	p_value...	count_Sol_...	std_Sol_LYSA...	avg_logD	p_value...	count_l...	std_logD	min_Sol...	q1_Sol_
1			0.425	0.348	2	0.367	0.20	0.376	8	0.603	0.166	0.166
2							0.89		1			
3			-1.67		1		1.21	0.148	2	0.403	-1.67	-1.67
4							-0.43		1			
5			-0.393	0.441	2	0.462	0.72	2.25e-07	9	0.134	-0.72	-0.72
6			-1.1	0.0193	6	0.793	0.84	0.000207	8	0.338	-2.18	-1.58
7							-0.83		1			
8							-0.40		1			
9			-0.902		1		-0.22	0.131	2	0.0636	-0.902	-0.902
10			0.629		1		-1.18		1		0.629	0.629
11			-0.119		1						-0.119	-0.119
12							1.64		1			

mmpdb transform – Rule Selection

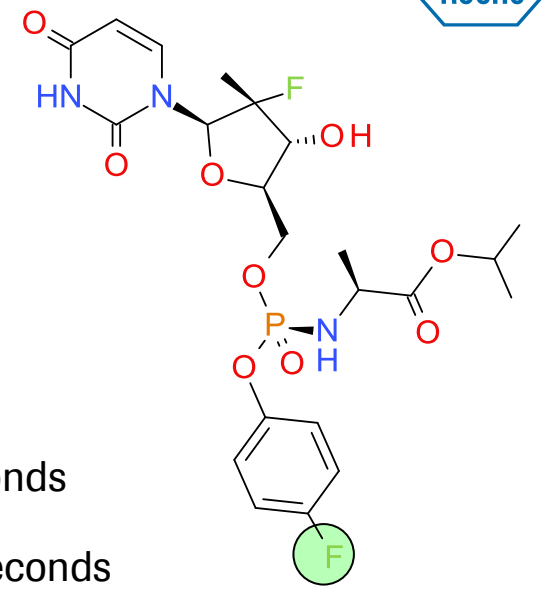
What if several rules (different transformations/ environments) lead to the same compound?

<u>Current implementation:</u>

```
If #pairs in one or more rules >= 10:
        out of those, select rule with lowest SD
Else:
        for i in (5,3,2,1):
                if #pairs in one or more rules >= i:
                        out of those, select rule with lowest SD
                        break
```

Rule selection is crucial. There might be other scientifically better ways to do it -> ideas?

mmpdb – Timings



Setup:

- Ki, IC50, AC50, CYP3A4 and hERG data from ChEMBL23.
- Remove duplicates:
  → 14377 compounds for CYP3A4
  → 6192 compounds for hERG
  → 302 compounds overlap
  → 20267 compounds overall
  (real use needs better data preparation)
- Workstation running RedHat 7 and python 3,
  32 GB of RAM and 10 Intel Xeon CPUs with 2.3 GHZ
- Remove compounds with more than 70 HA, 20 rotatable bonds.

Timing:

Fragmentation: 777 seconds

Indexing & create DB: 80 seconds

Load Properties into DB: 62 seconds

Transform query with Sofosbuvir (+ substructure = Sofosbuvir) against hERG and CYP3A4 (1620 suggested compounds):

Standard call: 51 s
Second call: 39 s
On web service: 1.7 s

Predict query with Sofosbuvir (Target: F-Sofosbuvir, green above) against hERG:

Standard call: ~17 s
On web service: 1.4 s

Outlook

- GUI
- Support for qualitative data
- Rule pruning (Environment is great, but leads to a crazy number of rules)
- Research on how to best do rule selection
- Research on how to share data with mmpdb
- Other use cases (to come in the future)

Availability

GitHub - rdkit/mmpdb: A

GitHub, Inc. [US] | https://github.com/rdkit/mmpdb

README.md

mmpdb 2.0 - matched molecular pair database generation and analysis

Synopsis

A package to identify matched molecular pairs and use them to predict property changes.

Requirements

The package has been tested on both Python 2.7 and Python 3.6.

You will need a copy of the RDKit cheminformatics toolkit, available from http://rdkit.org/ . Apart from other standard scientific python libraries like scipy and numpy, this is the only required third-party dependency for normal operation, though several optional third-party packages may be used if available.

3-clause BSD License

Request

Please use mmpdb and feedback,
either privately or on the RDKit Mailing List

A lot of extensions to mmpdb are possible. If you want to contribute, do not hesitate to contact us. We might know somebody else already working on your topic…

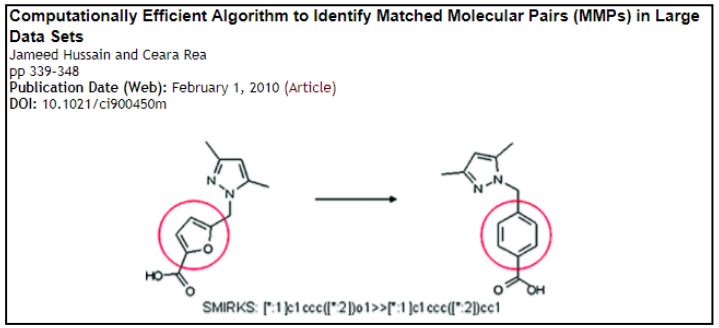
Acknowledgments

Andrew Dalke

Jerome Hert

Jameed Hussain (mmpa)

RDKit (hosting)



Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets

Jameed Hussain and Ceara Rea

pp 339-348

Publication Date (Web): February 1, 2010 (Article)

DOI: 10.1021/ci900450m

SMIRKS: [*:1]c1ccc([*:2])o1>>[*:1]c1ccc([*:2])cc1

Doing now what patients need next