Molecular comparisons by Maximum Common Sub-Structure (MCSS) and application to Matched Molecular Pair Analysis (MMPA)

MedChemica

BIGCHEM Lecture Oct 2016

SlideShare:

Lecture:

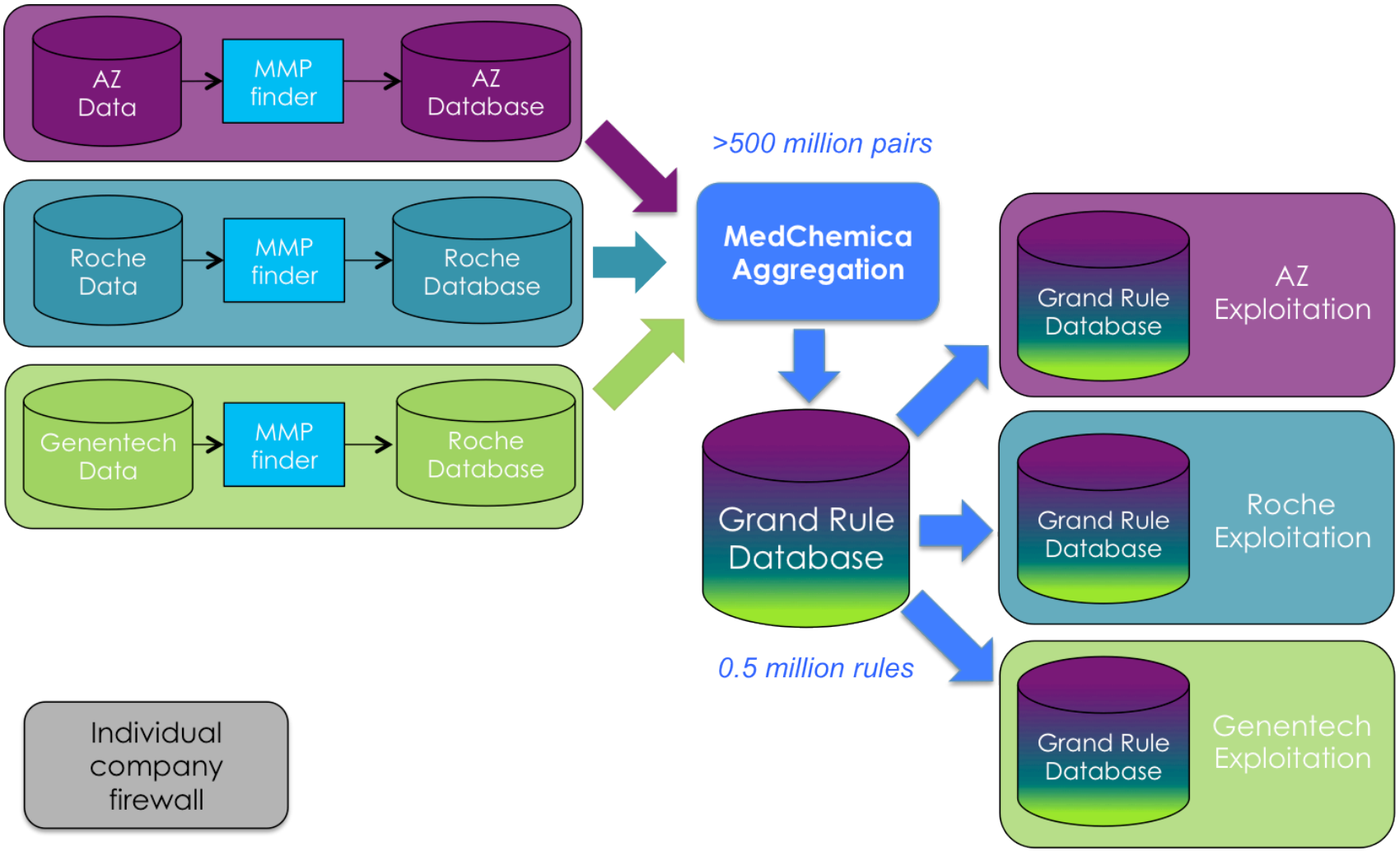
- Brief introduction to MedChemica – MCSS and MMPA at the heart of pharma knowledge sharing
- Comparing molecules:
  - Why do it?
  - What methods, compare contrast?
  - What is your experiment?

- About MCSS:
  - Definitions
  - A simple use case – depicting the common structure of two molecules
  - What tools are out there?
  - A brief bit of code in python (with RDkit and OpenEye toolkits)

- Matched Molecular Pair Analysis:
  - What is it?
  - MCSS is just the start of the problem….chemical encoding
  - Processing the data – generating rules
- What next? MMPA methodology can be extended to extract pharmacophores

MedChemica – Enabling Knowledge sharing

→Better medicinal chemistry



Full Paper, *J.Med.Chem* submitted

Recent Media

THE WALL STREET JOURNAL.

EUROPE BUSINESS NEWS | Updated June 25, 2013, 8:03 p.m. ET

Roche, Astra to Share Drug Research Data

Bloomberg

Swiss Stocks Advance on German Confidence Report

Roche Partnership

Roche advanced 2 percent to 223.30 Swiss francs. The company struck a partnership with AstraZeneca to share data on early-stage drug design, WSJ reported. Under the agreement, both companies will contribute data to a third company, U.K.-based MedChemica Ltd., WSJ said.

"Since the big 10 pharma companies spend $70 billion a year collectively on research without always have a good return, collaborating on research can benefit companies from all the money spent by all companies in the industry. This collaboration will benefit both AstraZeneca and Roche without divulging confidential information" – NASP – 26th June

the guardian

AstraZeneca adds more than 2% on reported R&D tie-up with Roche

The 'Market' approves of pharma collaborating

Lecture:

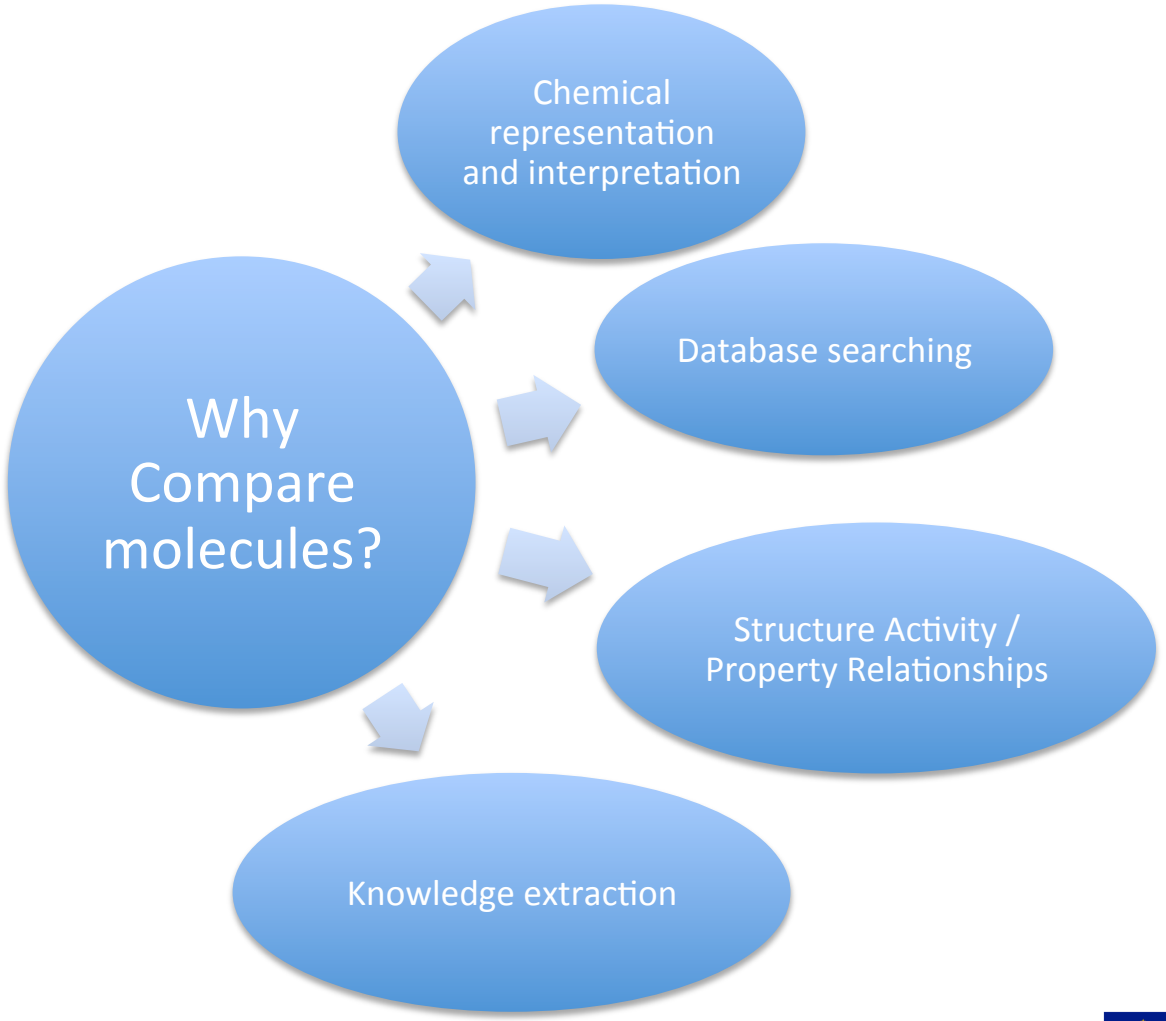
- Brief introduction to MedChemica – MCSS and MMPA at the heart of pharma knowledge sharing
- Comparing molecules:
  - Why do it?
  - What methods, compare contrast?
  - What is your experiment?

- About MCSS:
  - Definitions
  - A simple use case – depicting the common structure of two molecules
  - What tools are out there?
  - A brief bit of code in python (with RDkit and OpenEye toolkits)

- Matched Molecular Pair Analysis:
  - What is it?
  - MCSS is just the start of the problem….chemical encoding
  - Processing the data – generating rules
- What next? MMPA methodology can be extended to extract pharmacophores

Comparing Molecules

- Chemical information processing is the science of representing molecules in computers. Hence the fundamental "object" or data structure within a chemical information system is that of the molecule, its atoms (nodes) and its bonds (edge).

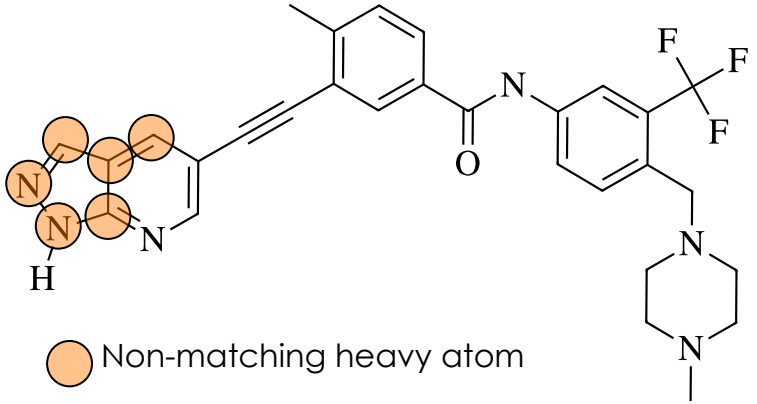


A word about experimental design:

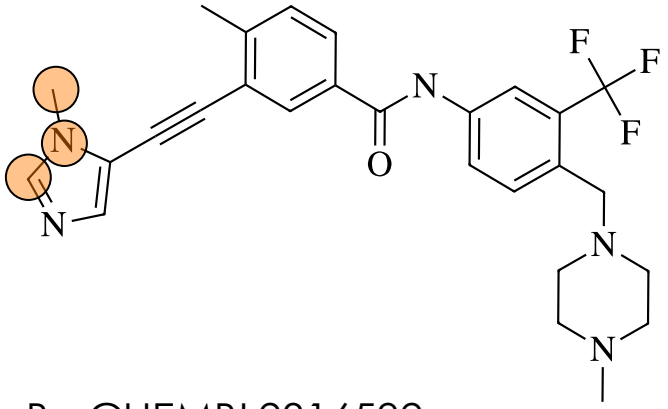
Always consider what question is being asked by the experiment, by considering this we can choose the right technique to use

Compute tasks can take days to run, especially comparing molecules

Maximum Common Sub Structure (or Graph) Method



A - CHEMBL1784632



B - CHEMBL2316582

A/B 36/39 ha = 92% MCS overlap

- Structures matched by overlapping of the matching atoms (nodes) and bonds (edges)
- The compute process uses works by traversing atoms and bonds in a graph representation of the molecules in memory
- The compute problem is NP-complete or NP-hard.

Maximum Common Sub-Structure patterns versus Fingerprints

MCSS

Process

- Convert each molecule into a graph representation in memory. Traverse each molecule finding atoms (nodes) and bonds (edges) that are the same, loop many many times building the largest set of edge linked nodes.

- Larger memory
- **Slow** (many algorithm set a time-out)
- Result is concise atom/ bond structure, so difference in chemistry can be captured

Fingerprints

Process

- Convert each molecule into a 'bit string (0s and 1s) representing parts of the molecule – thus each molecule is a 'number' in a computer – thus comparison between two molecules is easy.

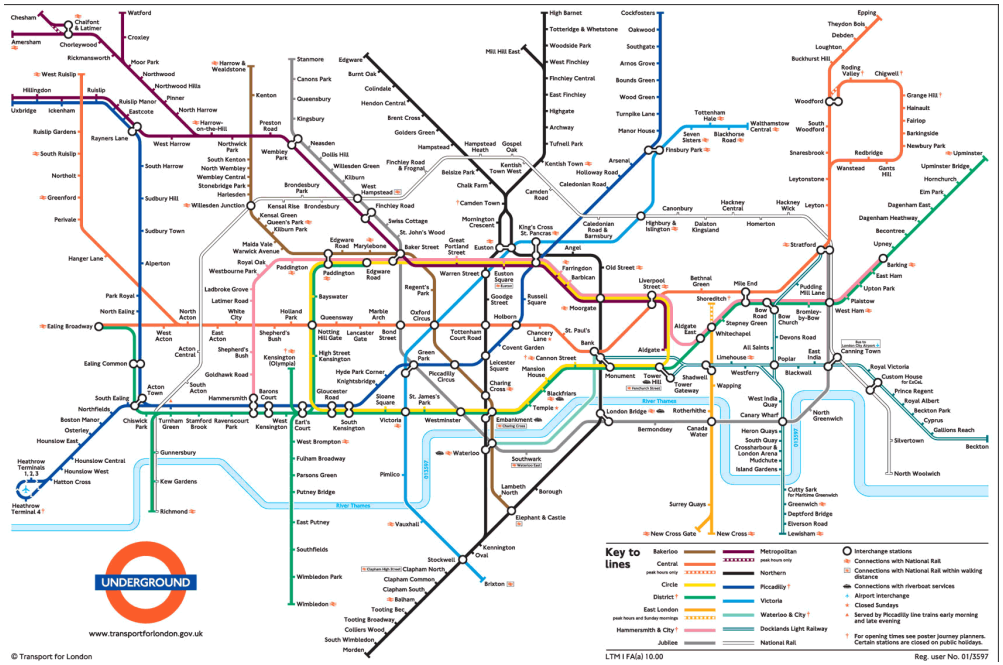
- Low memory
- Very **Fast**
- Result is a numerical difference between the molecules – no exact chemical structural difference

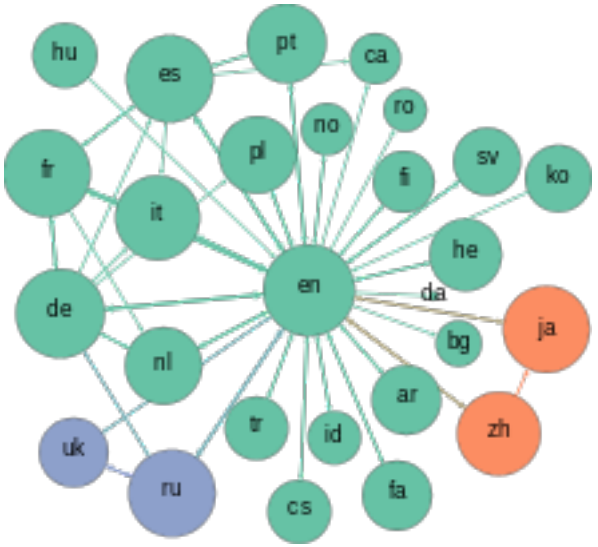
Experimental design is important – how much time and compute resource do you have?

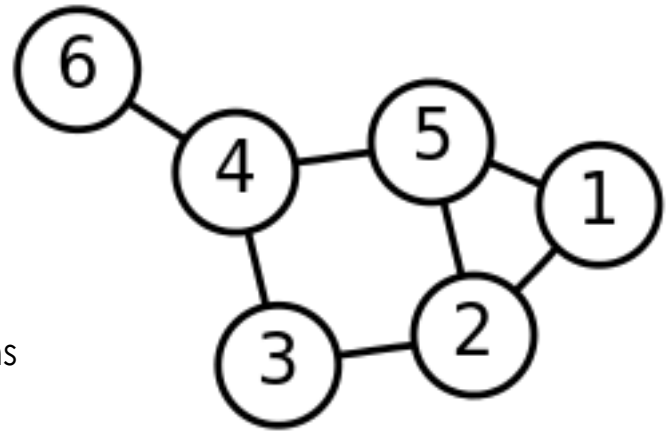
Maggiora, G; Vogt, M.; Stumpfe, D.; Barorath, J; Molecular Similarity in Medicinal Chemistry, *J.Med.Chem*.2013,3186.

Graphs Applied to Molecules

https://en.wikipedia.org/wiki/Graph_theory



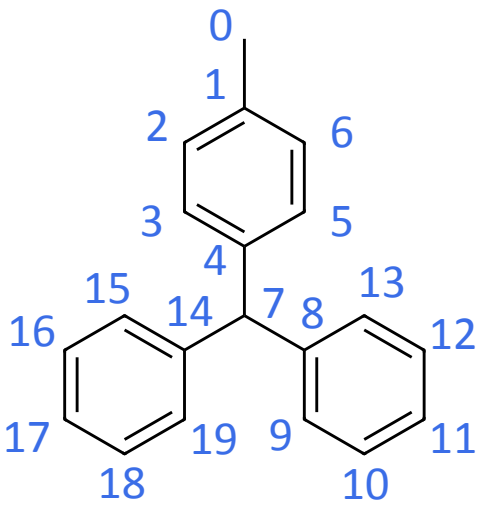




A Graph (G) is a set of Nodes and Vertices G = (N, V)
Each Node has relationship between other nodes by the connections made by Vertices
Each Node is connected to every other node by the Vertices and other Nodes....

....Node 6 is 3 Vertices from Node 2 but there are two paths to get there, via 3 or 5....

Molecules are suited to Graphs



Nodes are atoms

109 atom 'types' – known bonding behaviour

Vertices are bonds

single, double, triple……..

Molecular Graphs are the corner stone for chem-infomatics processing

SMILES → Cc1ccc(cc1)C(c2ccccc2)c3ccccc3 → 'parse' into Molecular Graph

For example: OpenEye Toolkit

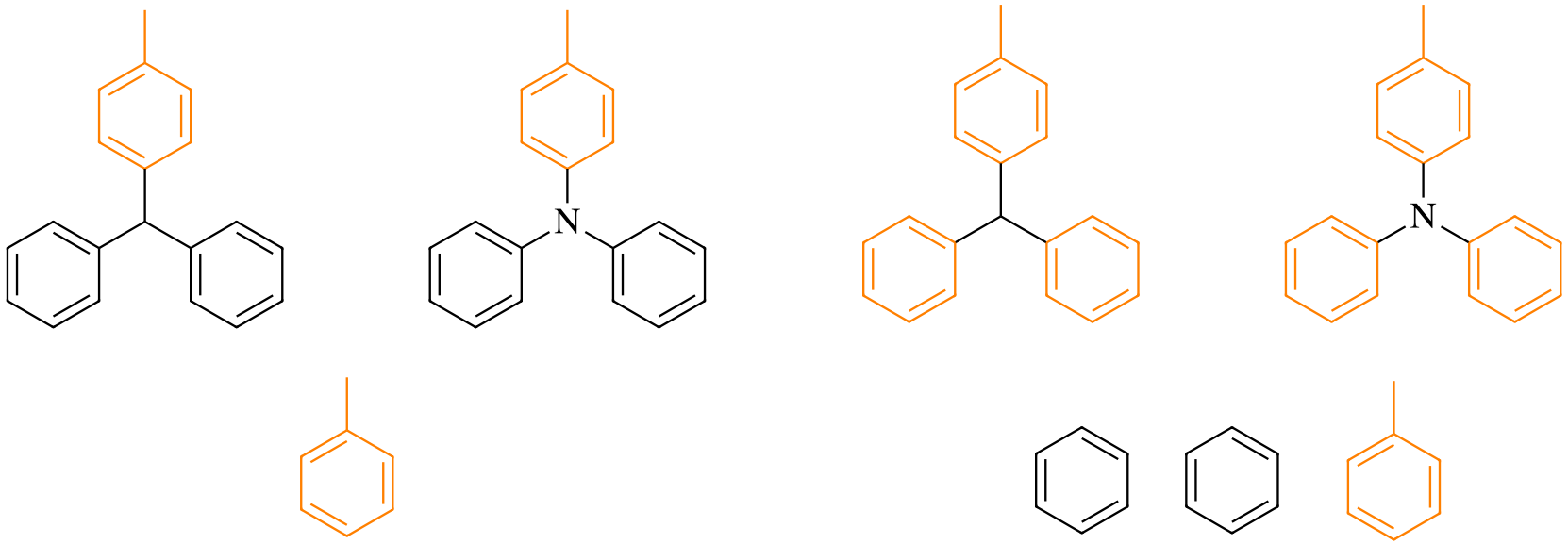
```
>>> from openeye.oechem import OEGraphMol, OESmilesToMol
>>> mol = OEGraphMol()
>>> OESmilesToMol(mol, "Cc1ccc(cc1)C(c2ccccc2)c3ccccc3")
>>> for atom in mol.GetAtoms():
...     print ('Atom Idx :{0} Atomic Num: {1}'.format(atom.GetIdx(),
atom.GetAtomicNum())
Atom Idx :0 Atomic Num: 6
Atom Idx :1 Atomic Num: 6
Atom Idx :2 Atomic Num: 6….
...
for bond in mol.GetBonds():
    print (bond.GetOrder())
```

WARNING

– Do Not be tempted to manipulate SMILES strings with text searches or replace functions

Types of MCSS – lets look at some molecules

MCES – Maximum Common Edge-Induced Substructure – as many matching chemical bonds




```
SMILES → Cc1ccc(cc1)C(c2ccccc2)c3ccccc3  Cc1ccc(cc1)N(c2ccccc2)c3ccccc3
cMCES  → [C][c]1[c][c][c]([c][c]1)
dMCES  → [C][c]1[c][c][c]([c][c]1).[c]2[c][c][c][c][c]2.[c]3[c][c][c][c][c]3
```

cMCES – *connected* MCES
The largest matching fragment of both molecules is itself connected structure

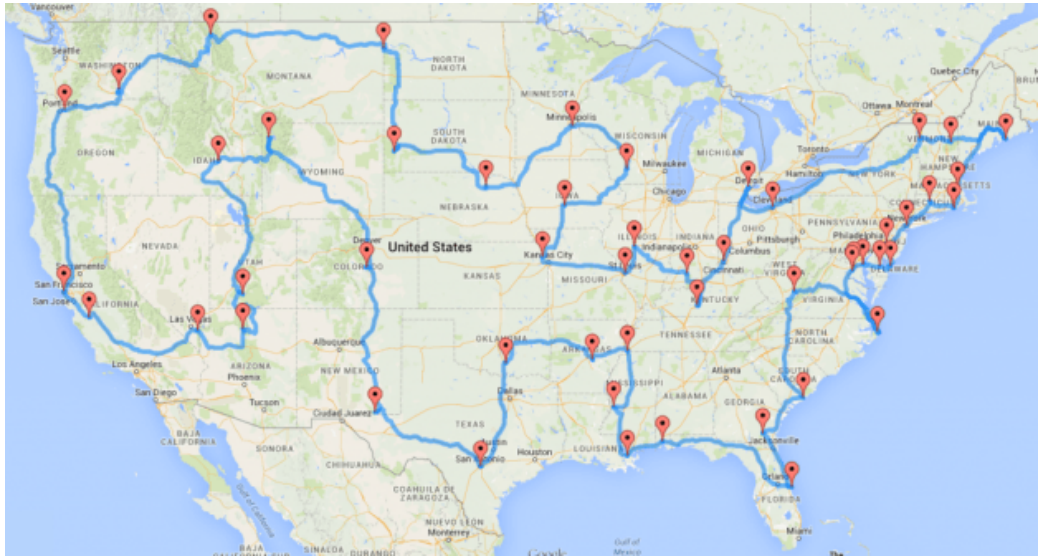
dMCES – *disconnected* MCES
Multiple matching parts of the molecule

The Travelling Sales Man Problem (and other NP-hard problems)

Comparing Graphs to find Maximum Common Sub-Graph is compute intensive
Described as NP

- *Non-deterministic polynomial time*
- not sure how many iterations over the graphs it will take
- Much harder and longer the more complex the graph (lots of atoms and bonds)

What is the shortest possible route for a traveling salesman seeking to visit each city on a list exactly once and return to his city of origin?
It has defied solution to this day

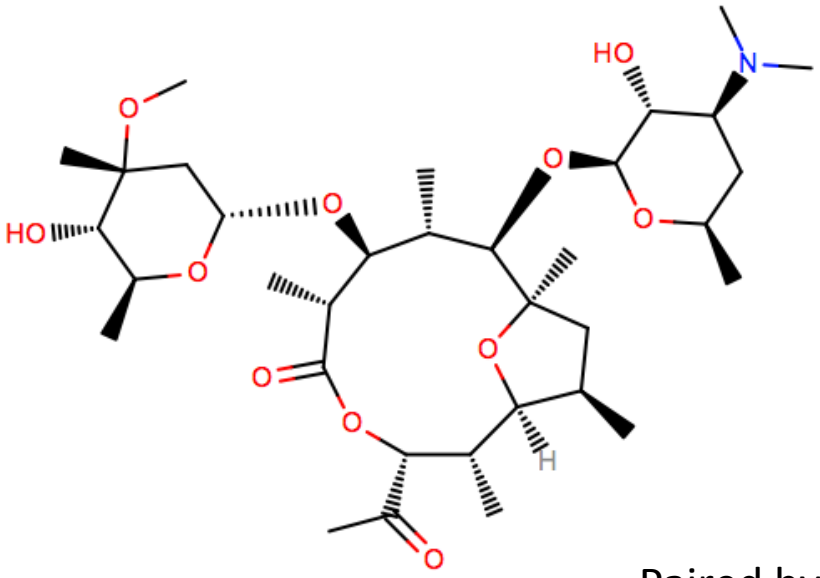


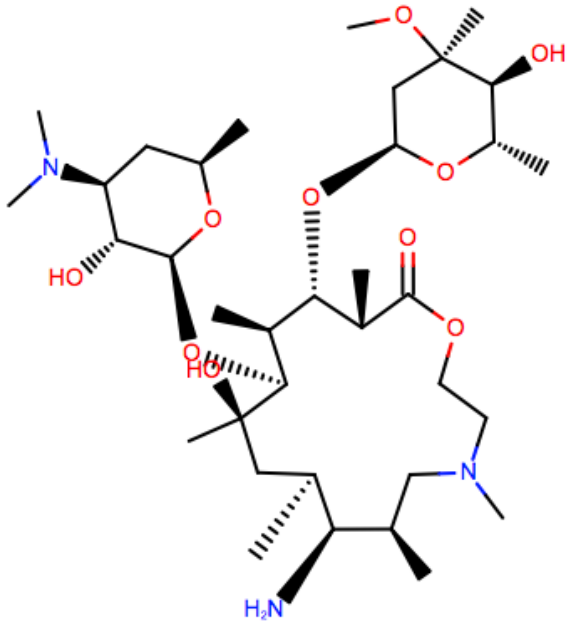
https://en.wikipedia.org/wiki/Travelling_Salesman_(2012_film)

MCSS - References

- Maggiora, G; Vogt, M.; Stumpfe, D.; Barorath, J; *Molecular Similarity in Medicinal Chemistry*, *J.Med.Chem*. **2013**,3186.
- Cao, Y.; Jiang, T.; Girke, T. *A Maximum common substructure-based algorithm for searching and predicting drug-like compounds*. *Bioinfomatics*, **2008**, 366.
- Duesbury, E.; Holliday, J.; Willet, P. *Maximum Common Substructure-Based Data Fusion in Similarity Searching*, *J.Chem.Inf.Model* **2015**, 222.
- Hariharan, R.; Janakiraman, A.; Nilakantan, R.; Singh, B.; Varghese, S.; Landrum, G.; Schuffenhauer, A. MultiMCS: *A Fast Algorithm for the Maximum Common Substructure Problem on Multiple Molecules*, *J.Chem.Inf.Model* **2011**, 786.

A difficult connected MCSS – can you spot it?





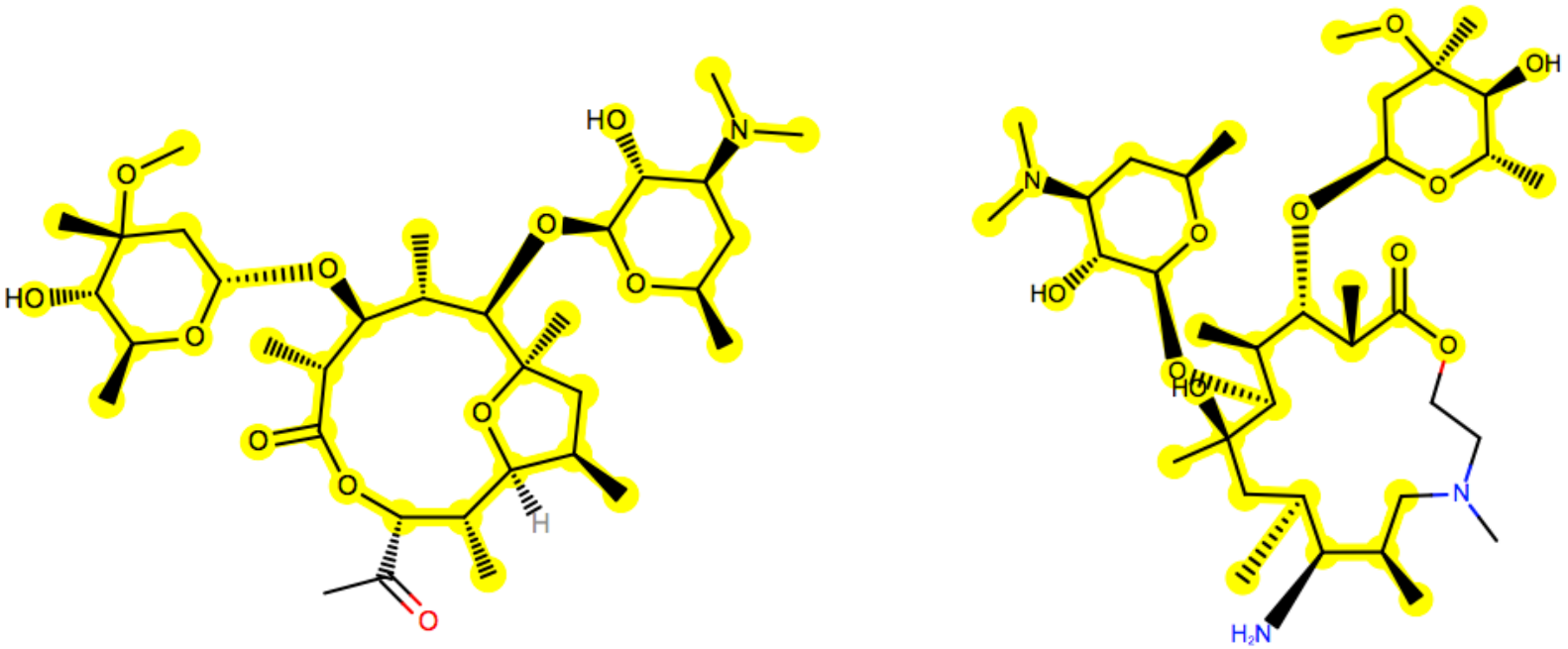
Paired by MCSS

CHEMBL1779022 **CHEMBL2146793**

Hard compute – 54 seconds for just this pair on one CPU
MCSS is compute intensive – lets use this example
400 antibiotic macrocycles - ALL pair comparison to find Structure Activity Relationships (SAR)
Roughly how long?

(400 x 400 x 54 secs) / 2 / 3600 / 24 = 50 days for 1 CPU
….we are going to need a 50 CPUs to run in a day.

A difficult MCSS – can you spot it?



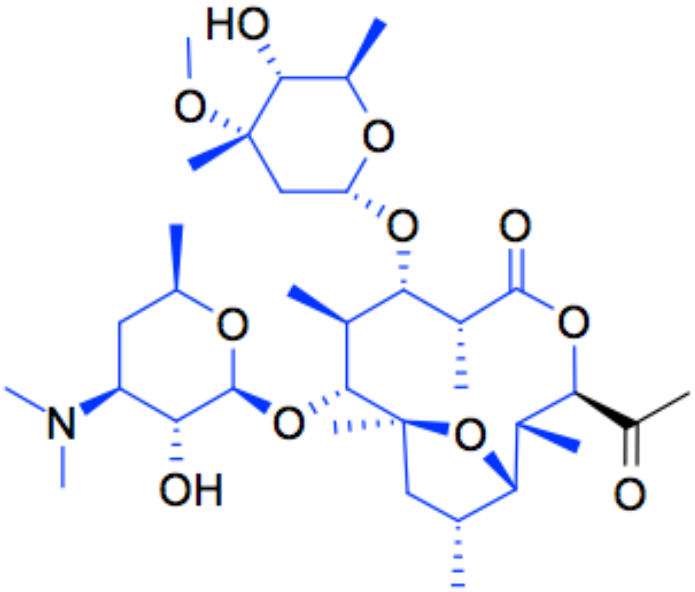
Paired by MCSS

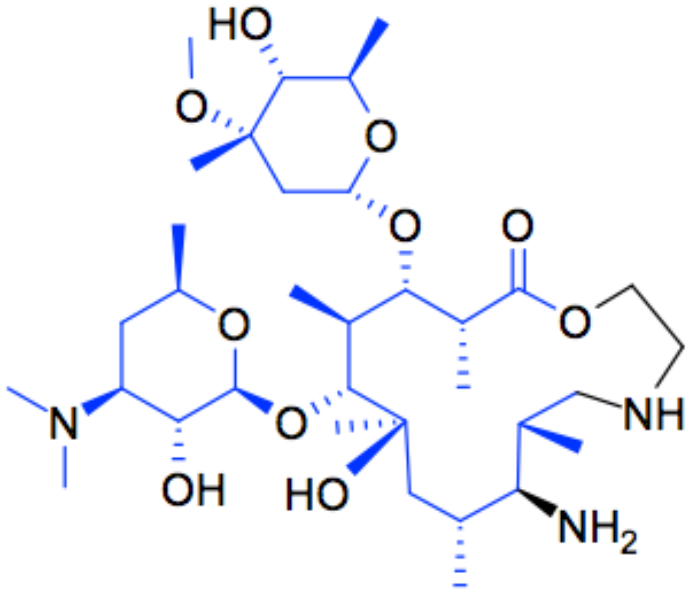
CHEMBL1779022 **CHEMBL2146793**

This is a useful case study:

- MCSS is useful as a sub-substructure that matches both molecules
- Using a depictions tool we can visualise the difference between them
- For compound design / optimisation this is very useful

MCSS on Macrocycles



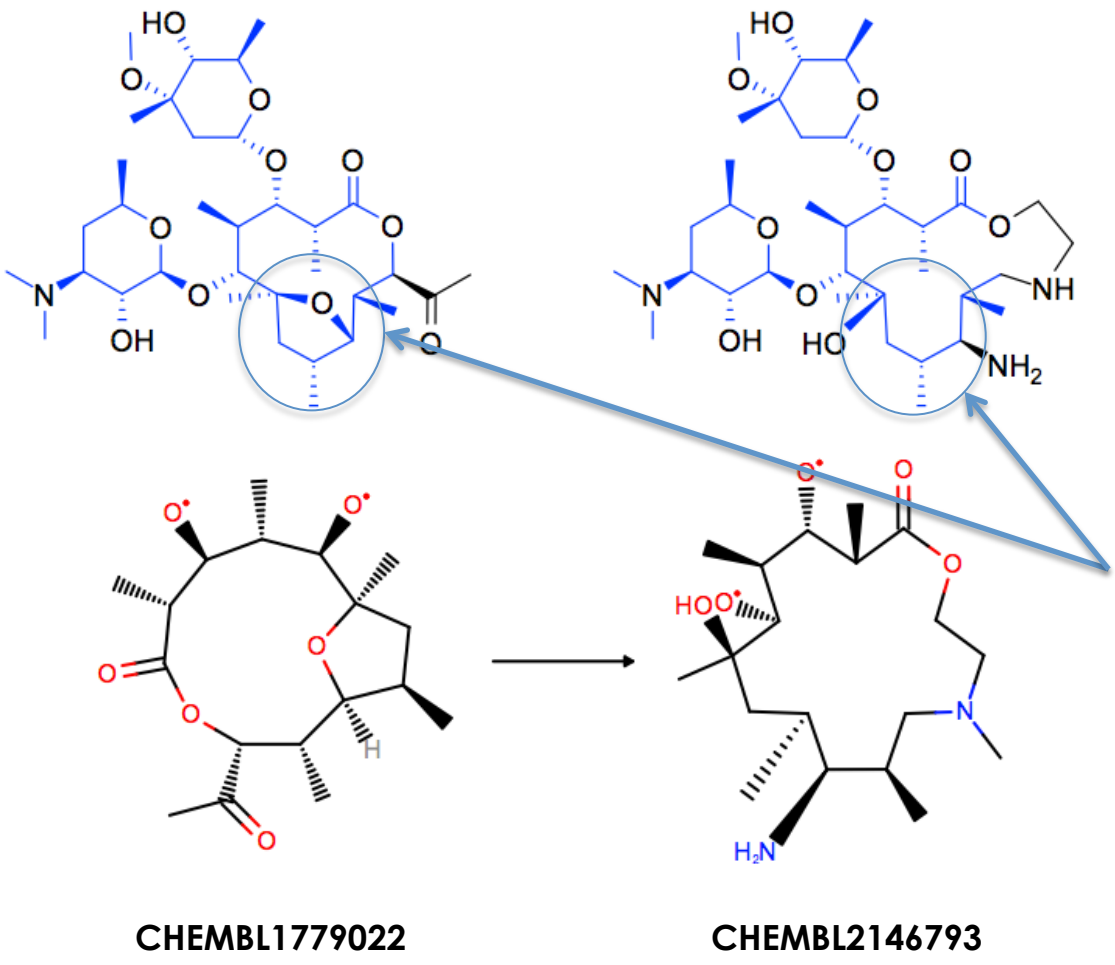


Paired by MCSS
Common substructure and change is clear

CHEMBL1779022 **CHEMBL2146793**

Hard to re-draw by a human – 30 minutes!

MCSS on Macrocycles



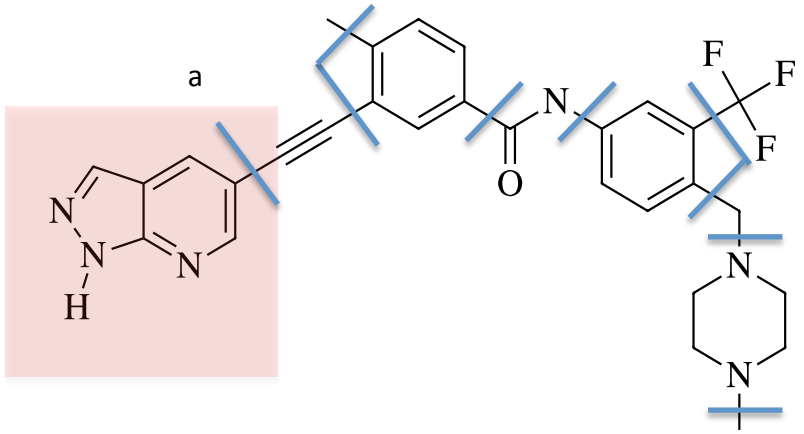
Graph processing algorithm has worked as coded, but did not pick out the difference between cyclised and open chain.

CHEMBL1779022 **CHEMBL2146793**

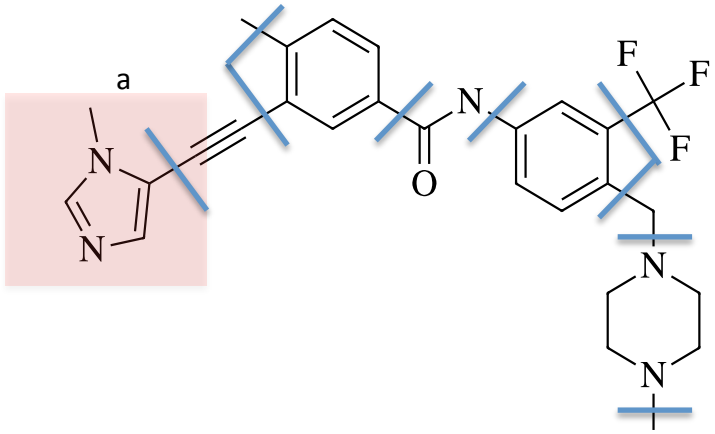
So we need to discover new drugs; as molecules become more complex (anti-biotic macrocycles) using a computer to help analyse and design new molecules becomes hard

Molecule Fragment and Index Method

Break all single rotatable bonds and group on common ‘core’ parts



A - CHEMBL1784632



B - CHEMBL2316582

- Semantically the same matched pair as MCSS but syntactically different in that ‘points of modification’ are not the same

Frag / Index single bond SMARTS

From the Hussain and Rea publication

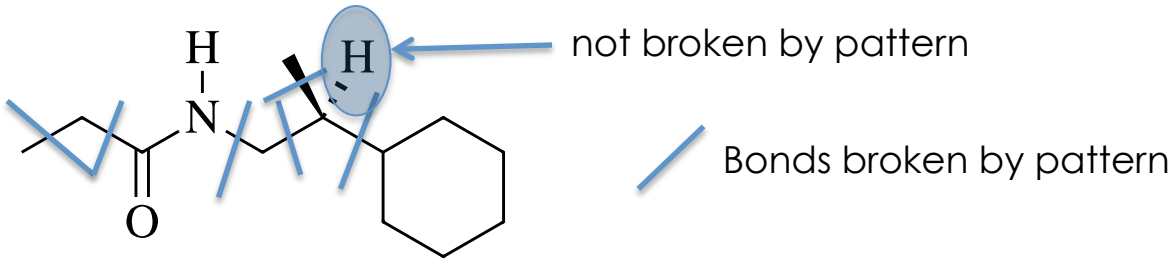
Hussain, J., & Rea, C. (2010). "Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets." Journal of chemical information and modeling, 50(3), 339-348.

A SMARTS pattern is required to identify the single bonds to break

Original `[*]!@!=[*]`
(any atoms, any bond not in ring or double – what about triple bonds?)

Amended in paper (see ref 24) `[#6+0;!$(*=,#[!#6])]!@!=!#[*]`
(carbon atom to any atom [excluding carbon that also has a double or triple] and no triple bond]) Effectively this removes amide bonds as disconnections

MedChemica variation `[#6+0;!$(*=,#[!#6])]!@!=!#[*;!$([H])]`
Carbon atom to any atom excluding hydrogen – stop disconnection to H as part of a explicit chiral centre

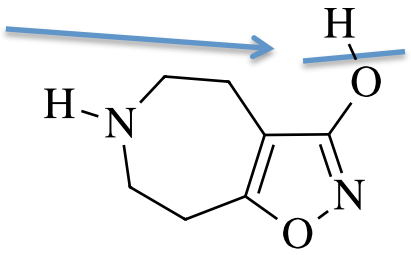


SMARTS pattern can be set in config file or from the command line for direct .smi file processing

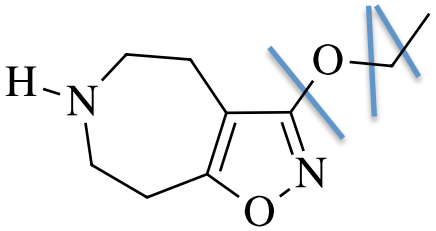
Frag / Index and MCSS pair finding

Consider the following compounds

The bond to H may not be broken but will be added by the H-additions step of the process

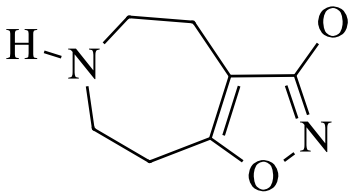


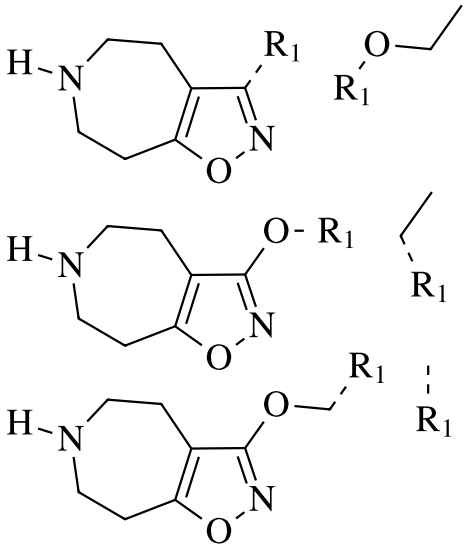
CHEMBL309689



CHEMBL2331793

Strict heavy atom c-MCSS





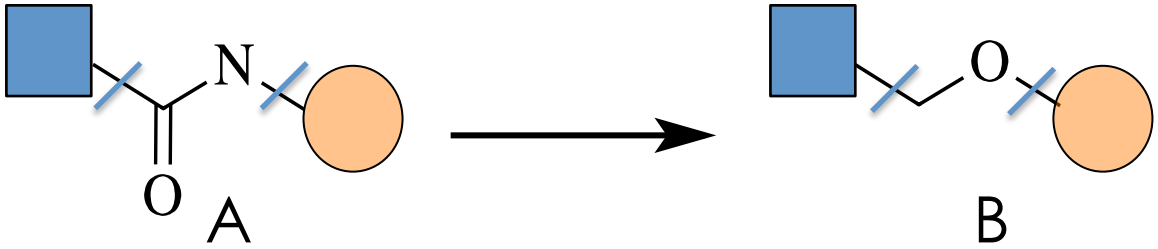
Three ways of Frag / Index finding matches (dependent on settings)

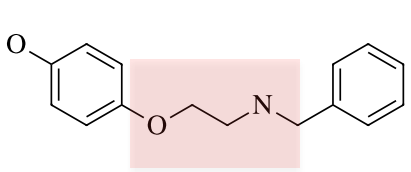
Num. of Trans = 3 x num of Environment atoms

FI and MCSS have found the 'same' pattern for the matched pairs via the same core. FI requires additional compute to find the true MCSS

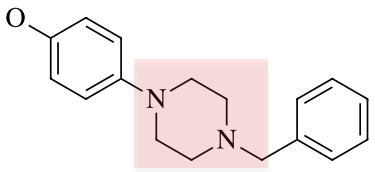
Frag And Index – useful for disconnected MCSS

Molecule Fragment and Index Method: 'Linker' changes found by double cuts

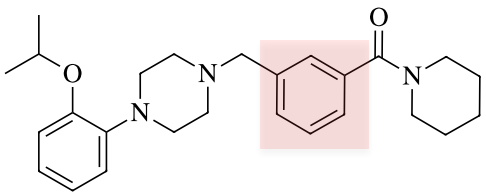




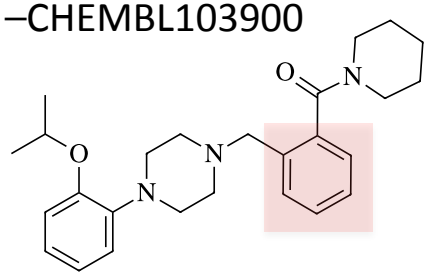
A – CHEMBL100461



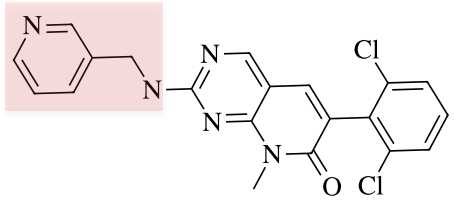
B –CHEMBL103900



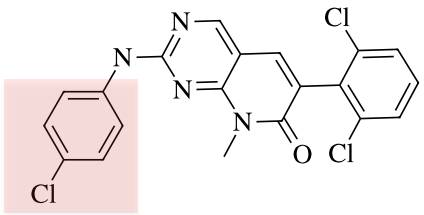
A – CHEMBL10085



B – CHEMBL10327



A – CHEMBL111247



B – CHEMBL110034

A and B are matched by the same context groups (for MCSS these are disconnected graphs so are not currently found) Context can be captured on both ends in the SMIRKS

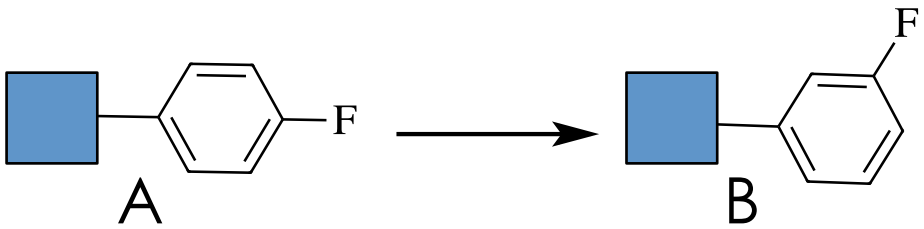
Both acyclic to acyclic and acyclic to cyclic changes are captured

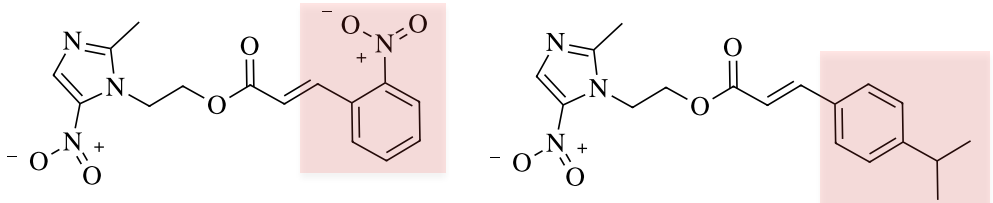
Subtle changes in ring isomers are found in the 'centre' of molecules

For large molecule whole side chain modification can be captured

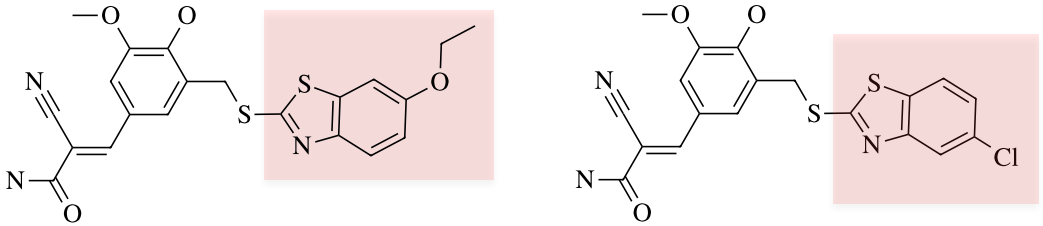
MCSS – subtle differences and ring changes

Maximum Common Sub Structure (or Graph) Method

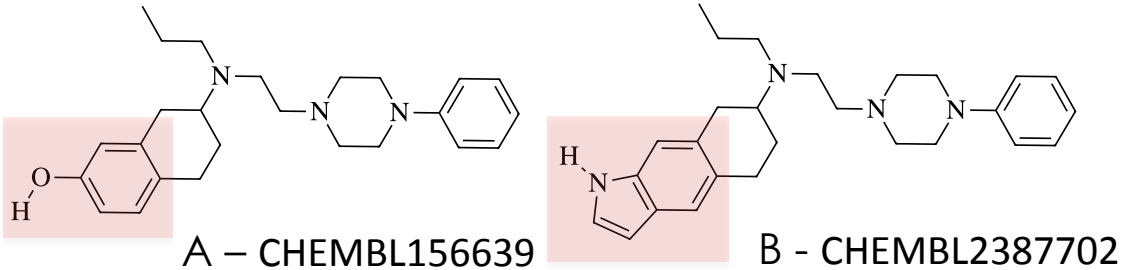




A – CHEMBL1173789 B – CHEMBL1173714



A – CHEMBL117055 B – CHEMBL115519



A – CHEMBL156639 B - CHEMBL2387702

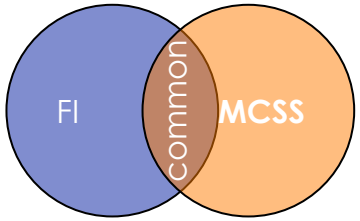
For molecules A and B of the same size (by heavy atom count) that are smaller (<30) transformations are found for positional isomer switches (these can be missed by F/I)

Subtle changes to smaller groups around rings are captured for all examples.

Acyclic to cyclic changes are captured well by MCSS method

Does the comparison method really matter?

Using only one technique will miss between 12% and 56% of pairings

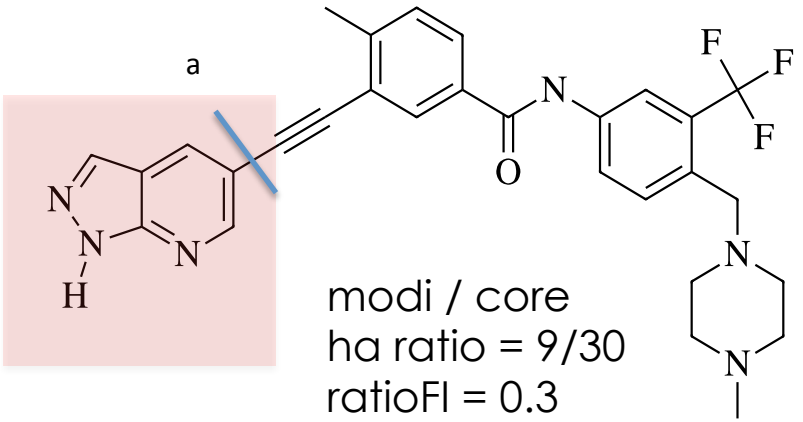


		Pairings				Pairings		
	number of compounds	common	FI only	MCSS only	total	FI only %	common %	MCSS only %
VEGF	4466	14631	17172	14823	46626	37	31	32
Dopamine Transporter	1470	4480	8930	3497	16907	53	26	21
GABAA	848	2500	1722	4205	8427	20	30	50
D2 human	3873	12995	13811	13098	39904	35	33	33
D2 rat	1807	5408	6595	7346	19349	34	28	38
Acetylcholine esterase	383	536	725	1434	2695	27	20	53
Monoamine oxidase	264	653	1156	246	2055	56	32	12
					min	20	20	12
					max	56	33	53

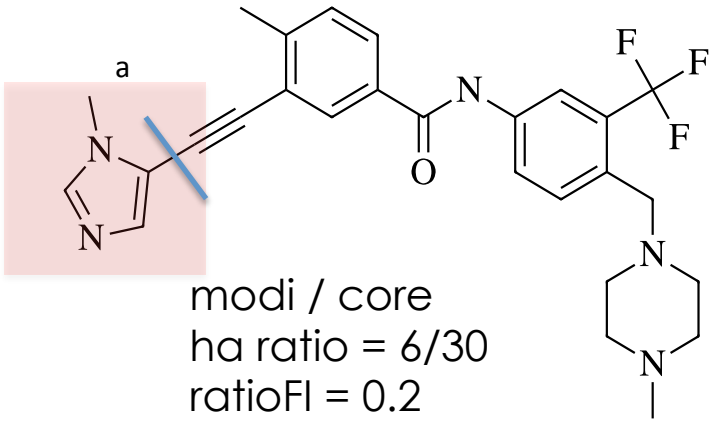
MCPairsv1.x

Settings for FI and MCSS methods

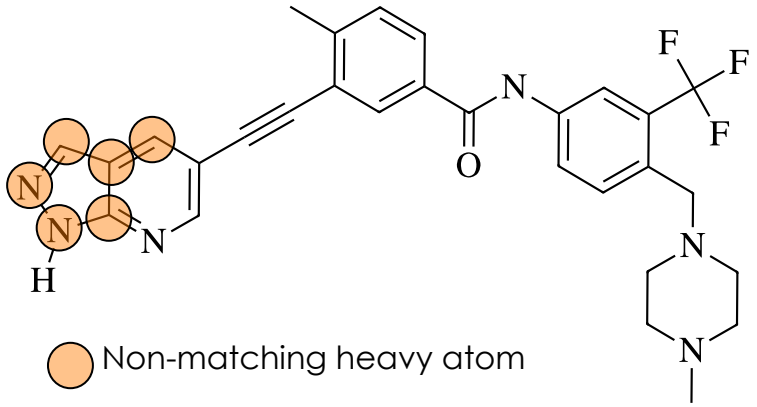
Set FIratio in configs larger inclusion >0.3
lower inclusion <0.3



modi / core
ha ratio = 9/30
ratioFI = 0.3

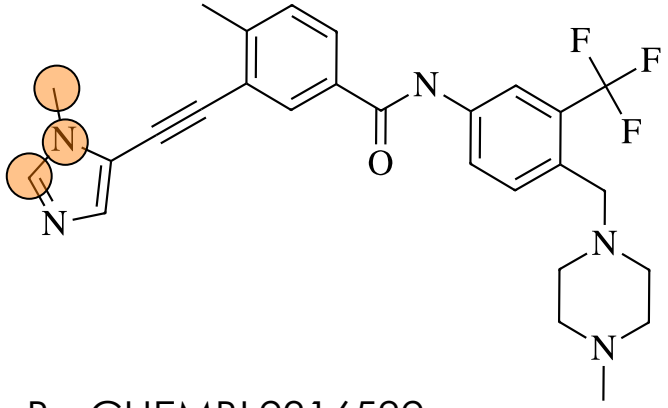


modi / core
ha ratio = 6/30
ratioFI = 0.2



Non-matching heavy atom

A - CHEMBL1784632



B - CHEMBL2316582

A/B 36/39
ha ratio = 0.92
MCS overlap

Set MCSSHAcutoff in configs smaller overlap <0.9
higher overlap >0.9

OpenEye ToolKit MCSS

```
#!/usr/bin/env python
from __future__ import print_function
from openeye.oechem import *

pattern = OEGraphMol()
target = OEGraphMol()
OESmilesToMol(pattern, "c1cc(O)c(O)cc1CCN")
OESmilesToMol(target, "c1c(O)c(O)c(Cl)cc1CCCBr")

atomexpr = OEExprOpts_DefaultAtoms
bondexpr = OEExprOpts_DefaultBonds
# create maximum common substructure object
mcss = OEMCSSearch(pattern, atomexpr, bondexpr, OEMCSType_Exhaustive)
# set scoring function
mcss.SetMCSFunc(OEMCSMaxAtoms())
# ignore matches smaller than 6 atoms
mcss.SetMinAtoms(6)
unique = True
# loop over matches
for count, match in enumerate(mcss.Match(target, unique)):
    print ("Match %d:" % (count + 1))
    print ("pattern atoms:", end=" ")
    for ma in match.GetAtoms():
        print (ma.pattern.GetIdx(), end=" ")
    print ("\ntarget atoms: ", end=" ")
    for ma in match.GetAtoms():
        print (ma.target.GetIdx(), end=" ")

    # create match subgraph
    m = OEGraphMol()
    OESubsetMol(m, match, True)

    print ("\nmatch smiles =", OEMolToSmiles(m))
```

http://docs.eyesopen.com/toolkits/python/oechemtk/patternmatch.html

Rdkit MCSS – Python

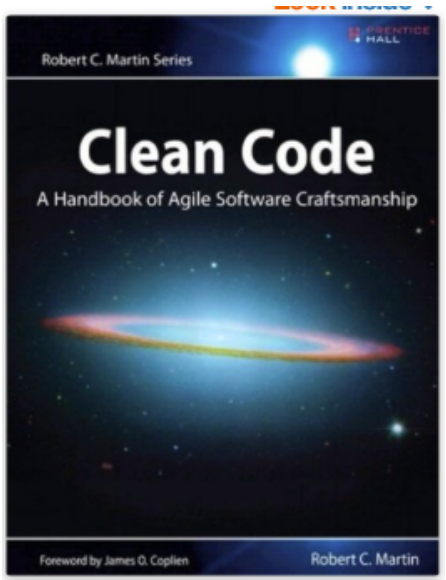
```
>>> from rdkit.Chem import rdFMCS
>>> mol1 = Chem.MolFromSmiles("O=C(NCc1cc(OC)c(O)cc1)CCCC/C=C/C(C)C")
>>> mol2 = Chem.MolFromSmiles("CC(C)CCCCCC(=O)NCC1=CC(=C(C=C1)O)OC")
>>> mol3 = Chem.MolFromSmiles("c1(C=O)cc(OC)c(O)cc1")
>>> mols = [mol1,mol2,mol3]
>>> res=rdFMCS.FindMCS(mols)
>>> res
<rdkit.Chem.rdFMCS.MCSResult object at 0x...>
>>> res.numAtoms
10
>>> res.numBonds
10
>>> res.smartsString
'[#6]1(-[#6]):[#6]:[#6](-[#8]-[#6]):[#6](:[#6]:[#6]:1)-[#8]'
>>> res.canceled
False
```

http://www.rdkit.org/docs/GettingStartedInPython.html#maximum-common-substructure

About Clean Code

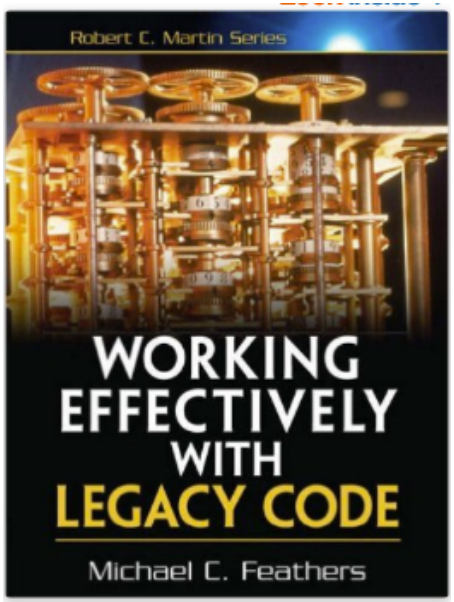
Compute code is written once and read a thousand times
Do:

- write Unit-Tests (especially with processing molecules)
- write functional tests
- write looooooong function, variable names
- write functions that DO ONE THING!
- write modular code – DRY (Do-not Repeat Yourself)



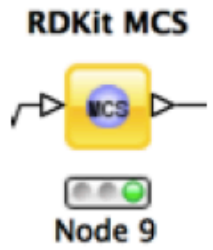
Clean Code: A Handbook of Agile Software Craftsmanship (Robert C. Martin)

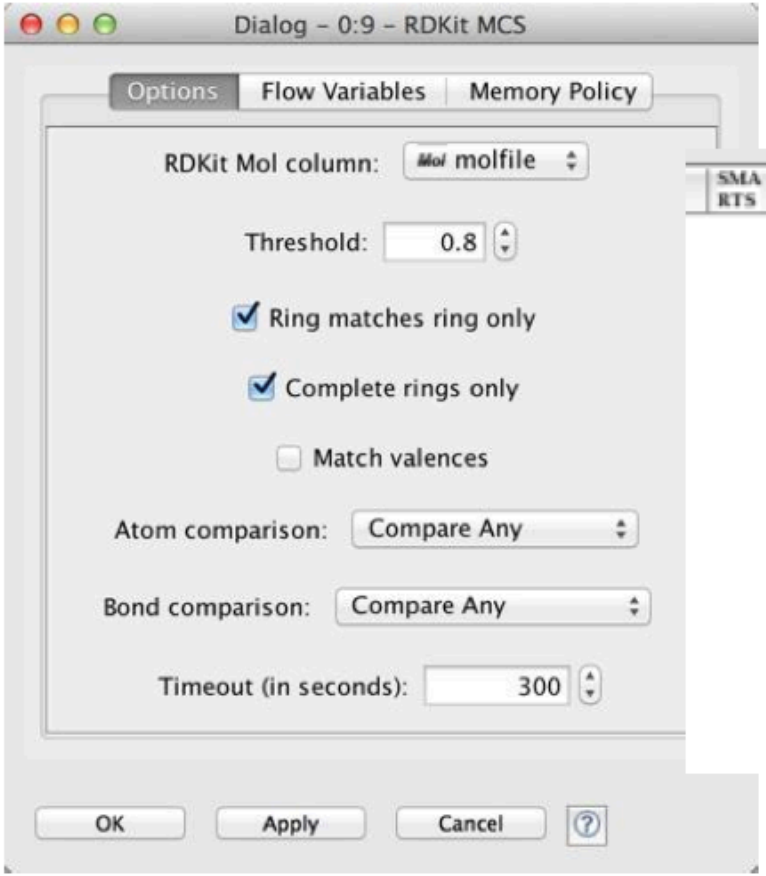
Working Effectively with Legacy Code (Micheal C. Feathers)

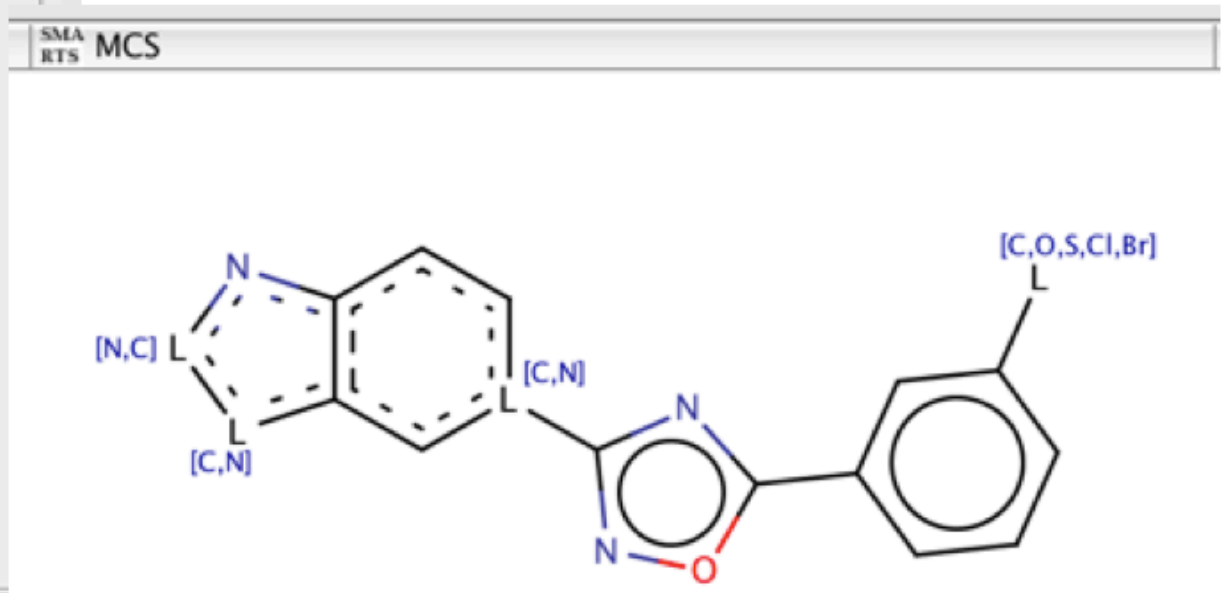


RDKit Maximum Common Substructure (MCS)

- Very flexible MCS implementation
- Includes partial MCS and generic atom/bond types
- Integration with “Group by” node







NOVARTIS

BigChem

Roche

Genentech
A Member of the Roche Group

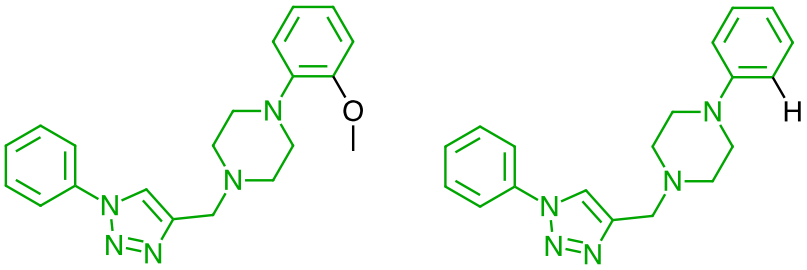
Matched Molecular Pair Analysis

An example of using MCSS (and FI) to compare molecules and extract Medicinal chemistry knowledge

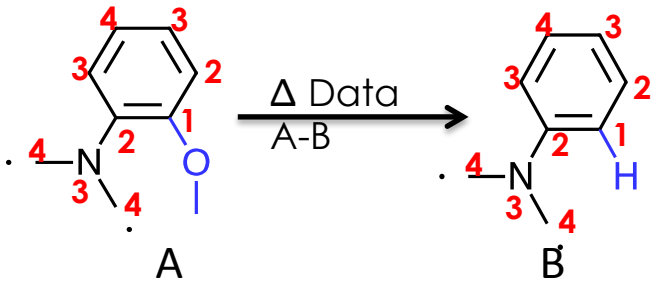
- Example it to show that comparing molecules it just the start of a process
- What experiment are we doing?
- How do we process the chemistry output and the data?
- What things can go wrong?
- Chirality!

Matched Molecular Pair Analysis uses MCSS methods….

- Matched Molecular Pairs – Molecules that differ only by a particular, well-defined structural transformation



- Transformation with environment capture – MMPs can be recorded as transformations from A→ B
- Environment is essential to understand chemistry

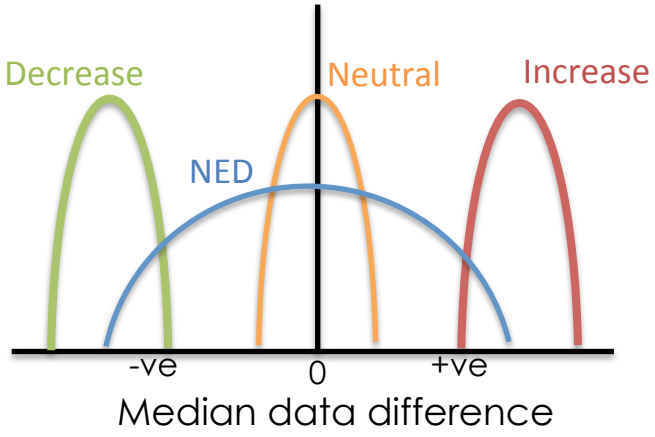


Statistical analysis

- Learn what effect the transformation has had on ADMET properties in the past

*Griffen, E. et al. Matched Molecular Pairs as a Medicinal Chemistry Tool. Journal of Medicinal Chemistry. 2011, **54**(22), pp.7739-7750.*

Rule selection



Identify and group **matching SMIRKS**

Calculate **statistical parameters** for each unique SMIRKS (n, median, sd, se, n_up/n_down)

Is n ≥ 6?

no → Not enough data: ignore transformation

yes → Is the |median| ≤ 0.05 and the intercentile range (10-90%) ≤ 0.3?

no → Perform two-tailed binomial test on the transformation to determine the significance of the up/ down frequency

yes → transformation is classified as '**neutral**'

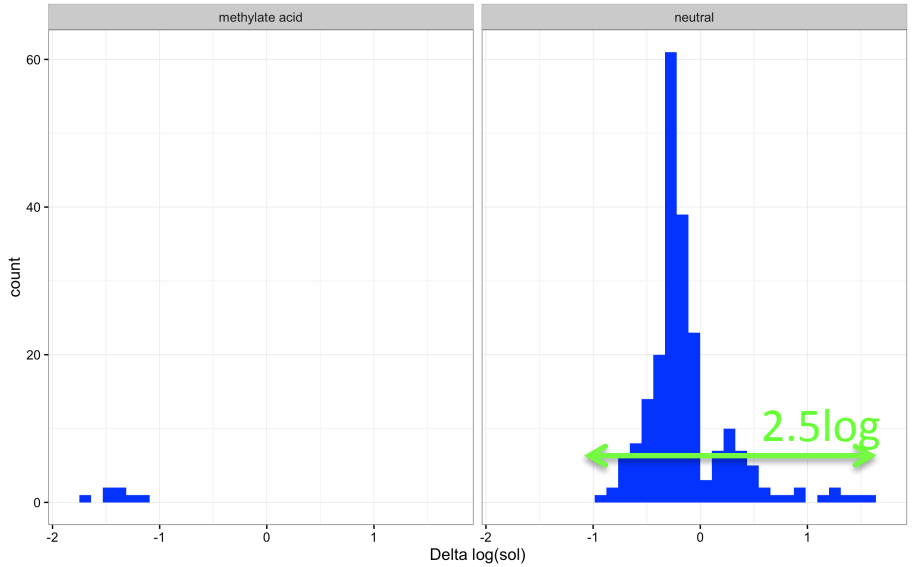
fail → Transformation classified as '**NED**' (No Effect Determined)

pass → Transformation classified as '**increase**' or '**decrease**' depending on which direction the property is changing

Environment really matters

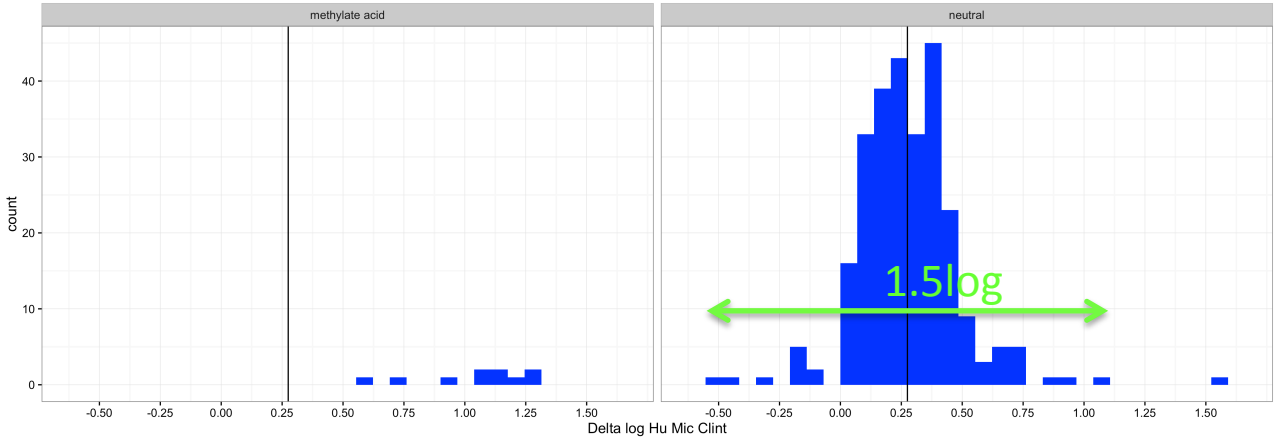
H→Me:

- Median Δlog(Solubility)
- 225 different environments



H→Me:

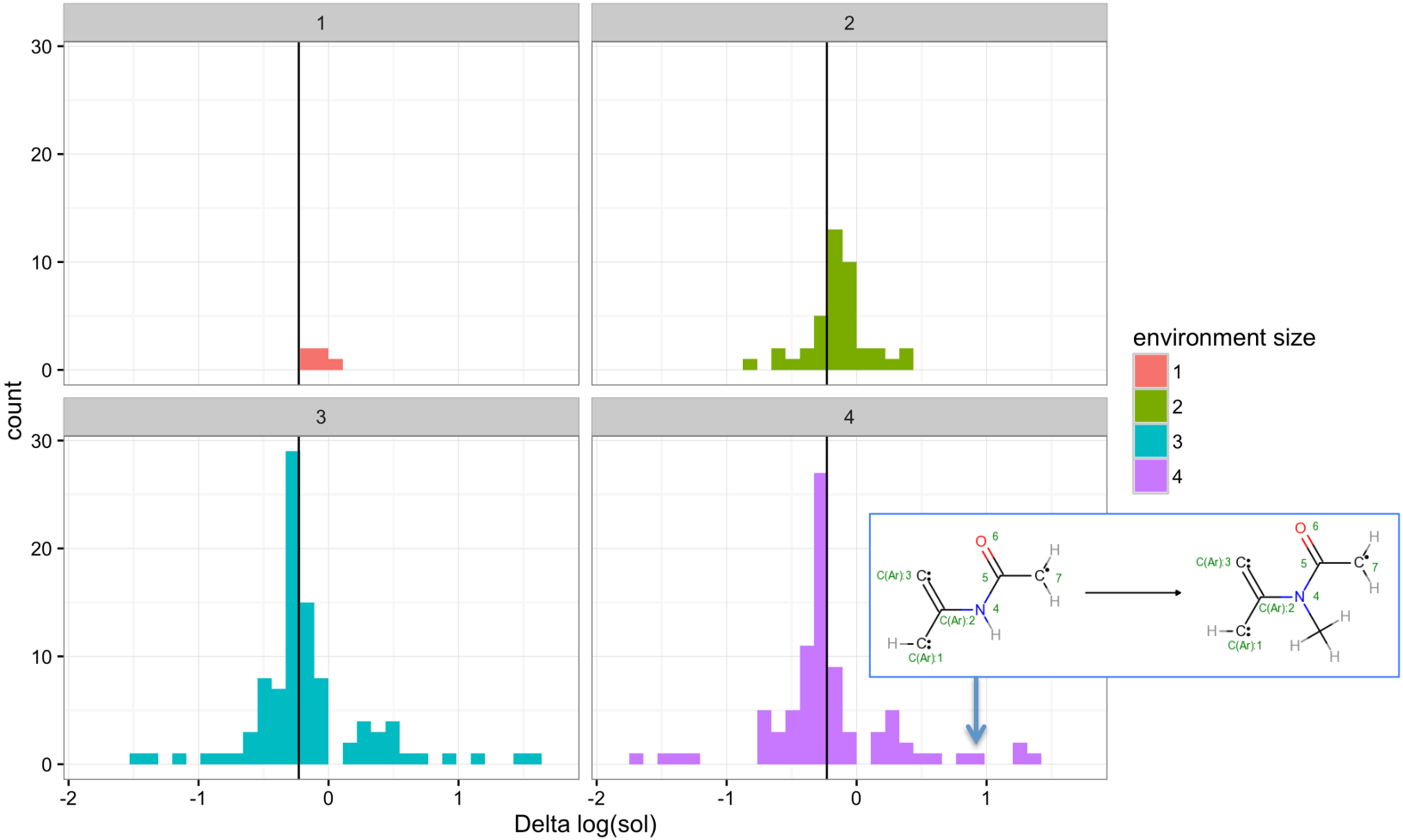
- Median Δlog(Clint) Human microsomal clearance
- 278 different environments



More environment = right detail

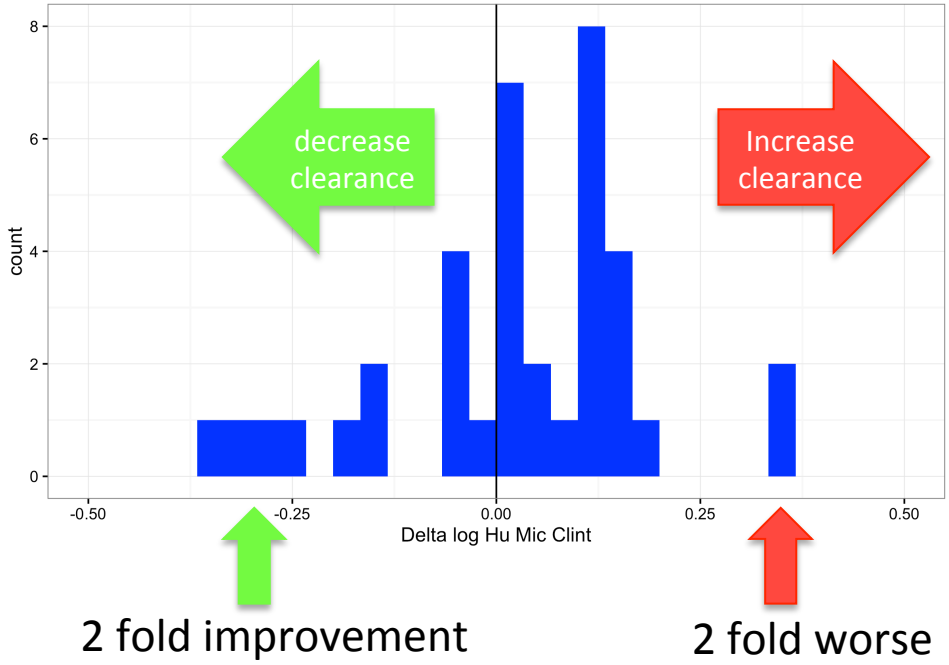
H→Me Solubility:

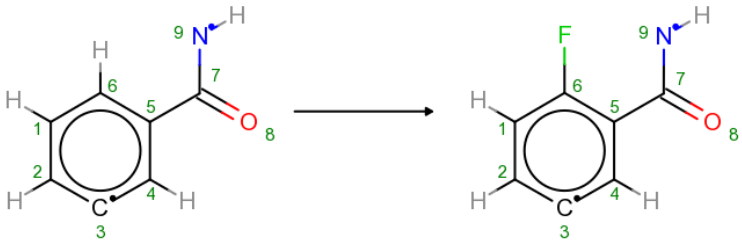
- 225 different environments

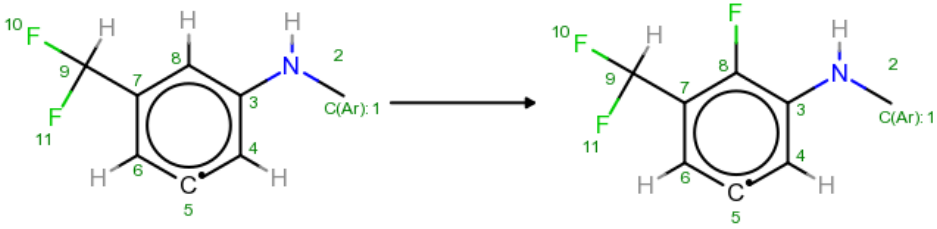


H→F What effect on Clearance?

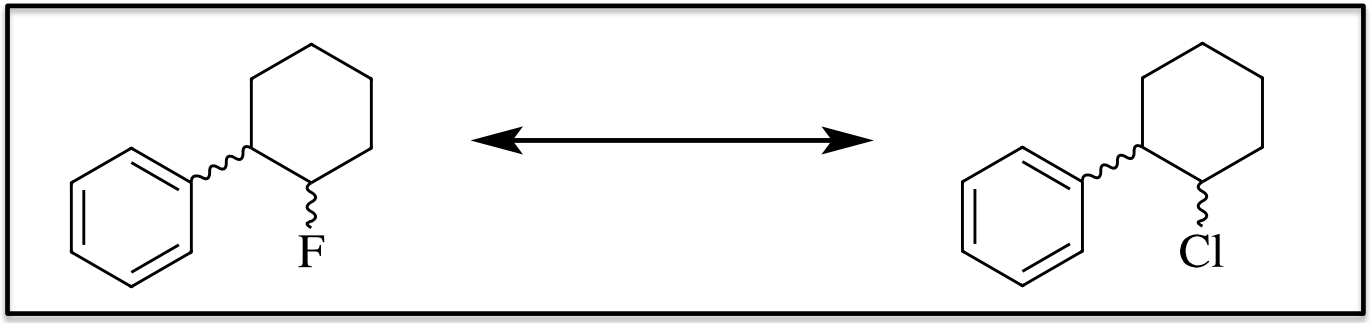
- Median Δlog(Clint) Human microsomal clearance
- 37 different environments

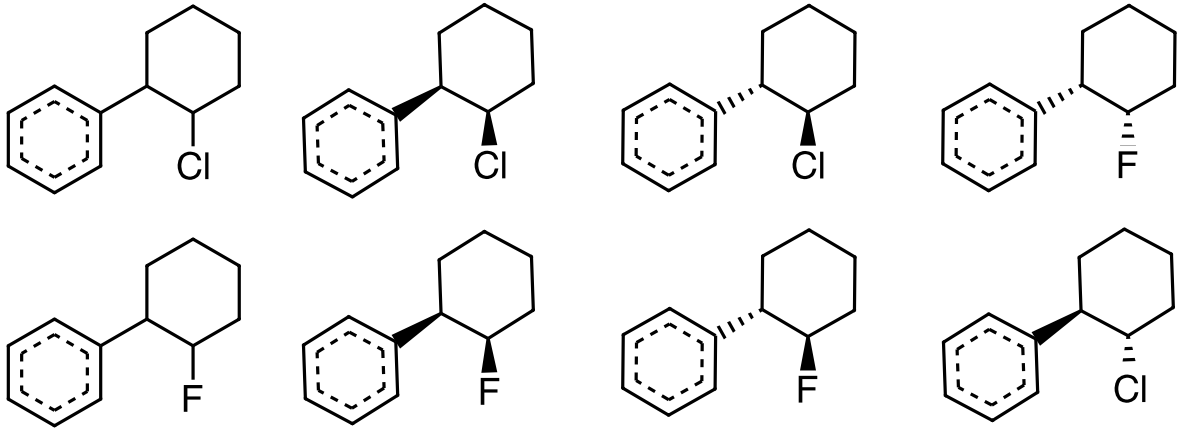






Chiral Test Set example

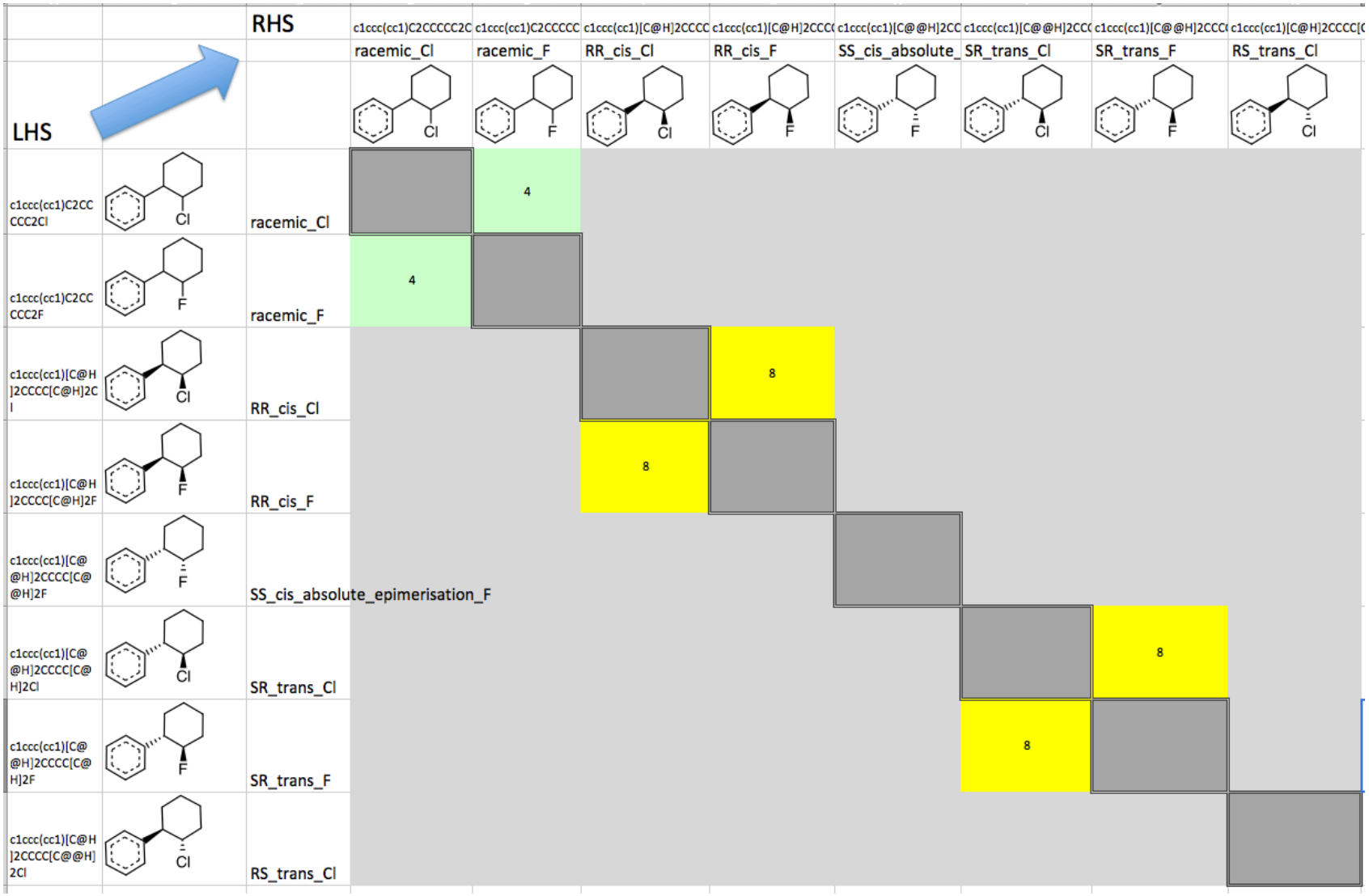




The experiment devised above maps into this .smi file - no duplicate compounds

```
c1ccc(cc1)C2CCCCC2Cl racemic_Cl
c1ccc(cc1)C2CCCCC2F racemic_F
c1ccc(cc1)[C@H]2CCCC[C@H]2Cl RR_cis_Cl
c1ccc(cc1)[C@H]2CCCC[C@H]2F RR_cis_F
c1ccc(cc1)[C@@H]2CCCC[C@@H]2F SS_cis_absolute_epimerisation_F
c1ccc(cc1)[C@@H]2CCCC[C@H]2Cl SR_trans_Cl
c1ccc(cc1)[C@@H]2CCCC[C@H]2F SR_trans_F
c1ccc(cc1)[C@H]2CCCC[C@@H]2Cl RS_trans_Cl
```

Chirality MCPairs --chiralON

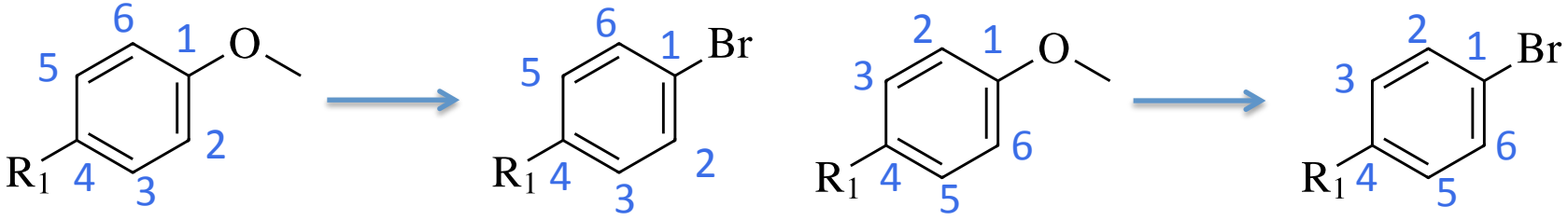


How to ‘store’ chemical information?

MCSS → MMPA → SMIRKS

Storing reactions as canonical SMIRKS – an example of storing chemical information

Canonicalised SMIRKS 1 – Its about Symmetry



4-Atom rule

SMIRKS_A

[C]([H])([H])([H])[O][c:1]1[c:2]([H])[c:3]([H])[c:4][c:5]([H])[c:6]1([H])
>>[c:5]1([H])[c:6]([H])[c:1]([c:2]([H])[c:3]([H])[c:4]1)[Br]

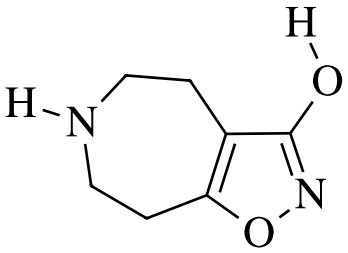
SMIRKS_B

[C]([H])([H])([H])[O][c:1]1[c:2]([H])[c:3]([H])[c:4][c:5]([H])[c:6]1([H])
>>[c:3]1([H])[c:2]([H])[c:1]([c:6]([H])[c:5]([H])[c:4]1)[Br]

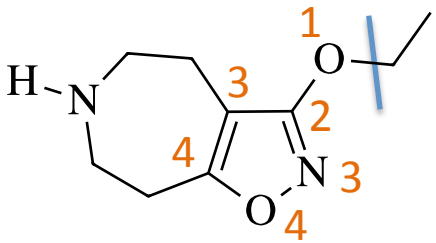
Both SMIRKS are VALID and would transform a molecule correctly
However we want to consistently produce a single SMIRKS for all of these
It is 50:50 how an unchecked algorithm will number around the ring

Therefore we force the numbering in the reactant side by numbering 1,2,3,4..n
Then use symmetry checking to change the numbers to a consistent pattern
Thus SMIRKS_B is produced

Canonicalised SMIRKS 2



CHEMBL309689



CHEMBL2331793

Orange – atom env radius
Blue - atom map index

Highly specific explicit H

1-Atom rule

[H][O:1]>>[C]([H])([H])([H])[C]([H])([H])[O:1]

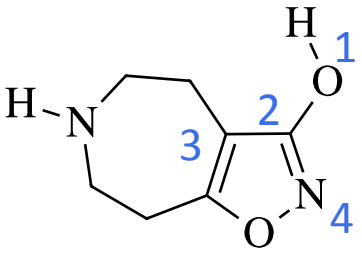
Key mapped atom
With 1 atom rule this is simple – 1 becomes 1

2-Atom rule

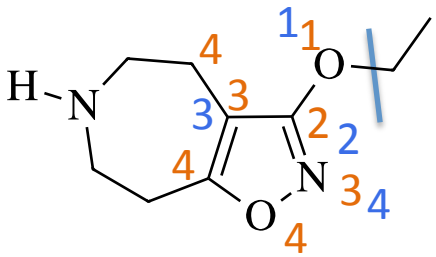
[H][O:1][c:2]>>[c:2][O:1][C]([H])([H])[C]([H])([H])([H])

Mapped atoms
Absolutely key – note the ethyl is right most in SMIRKS

Canonicalised SMIRKS 3



CHEMBL309689



CHEMBL2331793

Orange – atom env radius
Blue - atom map index

Highly specific explicit H

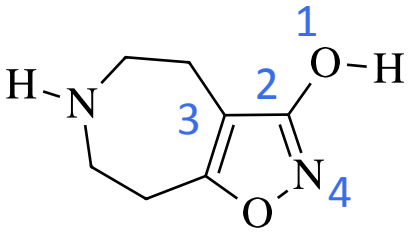
3-Atom rule [H][O:1][c:2]([c:3])[n:4]>>[c:3][c:2]([n:4])[O:1][C]([H])([H])[C]([H])([H])([H])

Key mapped atom
With 3 atom rule environment is more complex

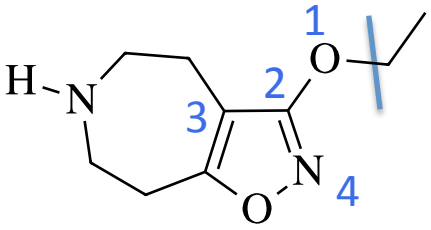
4-Atom rule [H][O:1][c:2]1[c:3]([c:4][o:5][n:6]1)[C:7]([H])([H])>>
[C]([H])([H])([H])[C]([H])([H])[O:1][c:2]1[c:3]([c:4][o:5][n:6]1)[C:7]([H])([H])

Note Mapped atoms run left to right 1,2,3,4…n
Absolutely key – note the ethyl is LEFT most in SMIRKS
Rule change depending on rule size, environment and symmetry

Canonicalised SMIRKS 4



CHEMBL309689



CHEMBL2331793

Orange – atom env radius
Blue - atom map index

3-Atom rule [H][O:1][c:2]([c:3])[n:4]>>[c:3][c:2]([n:4])[O:1][C]([H])([H])[C]([H])([H])([H])

Key mapped atom
Explicit Hydrogen are on the RHS of product side
Thus the mappings become complex and NOT in order 1,2,3,n

Reverse SMIRKS [c:3][c:2]([n:4])[O:1][C]([H])([H])[C]([H])([H])([H])>>[H][O:1][c:2]([c:3])[n:4]

While this SMIRKS is VALID and would transform a molecule correctly
IF we search for the this SMIRKS there is NO MATCH

Actual SMIRKS [c:1][c:2]([n:3])[O:4][C]([H])([H])[C]([H])([H])([H])>>[H][O:4][c:2]([c:1])[n:3]]

Due to Explicit H, Symmetry and Canonicalisation - Map Index must be treated with care – directly swapping the string may not match

NOTE – The same applies to Fragment SMILES

Take home messages

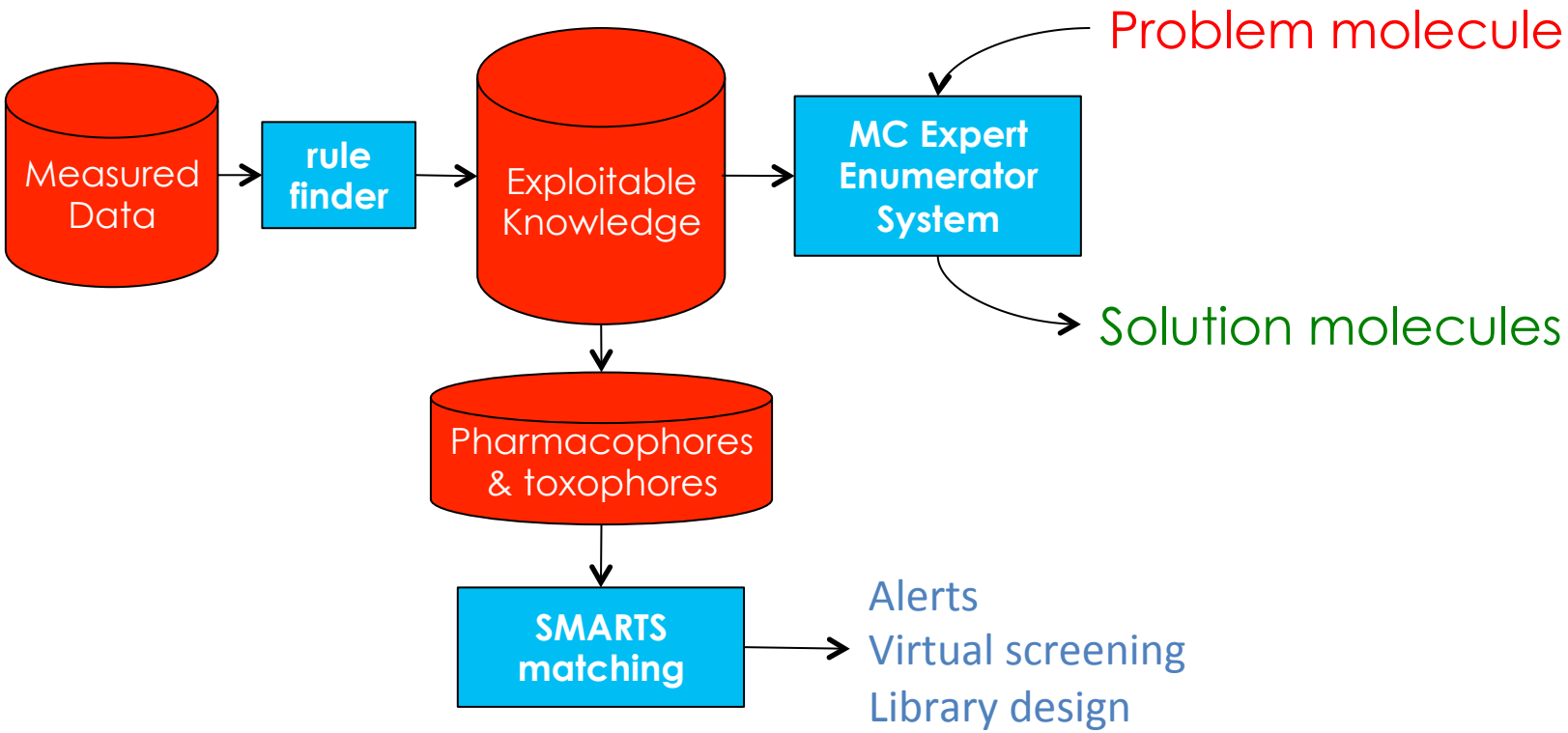
- Molecular Graphs are the corner stone of chem-infomatics
  - Allow traversal of atoms and bonds
  - Allow graph theory techniques to be applied to comparing molecules
- Maximum Common Sub-Structure is found by NP-complete graph techniques and is compute intensive
- The Fragment and Index method can find MCSS in many circumstances but not all
- Further compute processes are required to deal with chirality and encoding of the common parts of molecules and the difference between molecules
- Matched Molecular Pair analysis is a data analysis technique, based on MCSS, to extract chemical design knowledge

Appendix

IP security

- Contributing company structures are NOT shared
- Contributing company identifiers are NOT shared
- Each contributing company receives it's OWN custom GRD copy to enable drill back to OWN example data
- Minimum of n=6 example pairs required for inclusion in the GRD
- Enforced limits of shared substructure sizes
- Option for "time-blocking" sharing of data to withhold data < 6 months old for IP submission

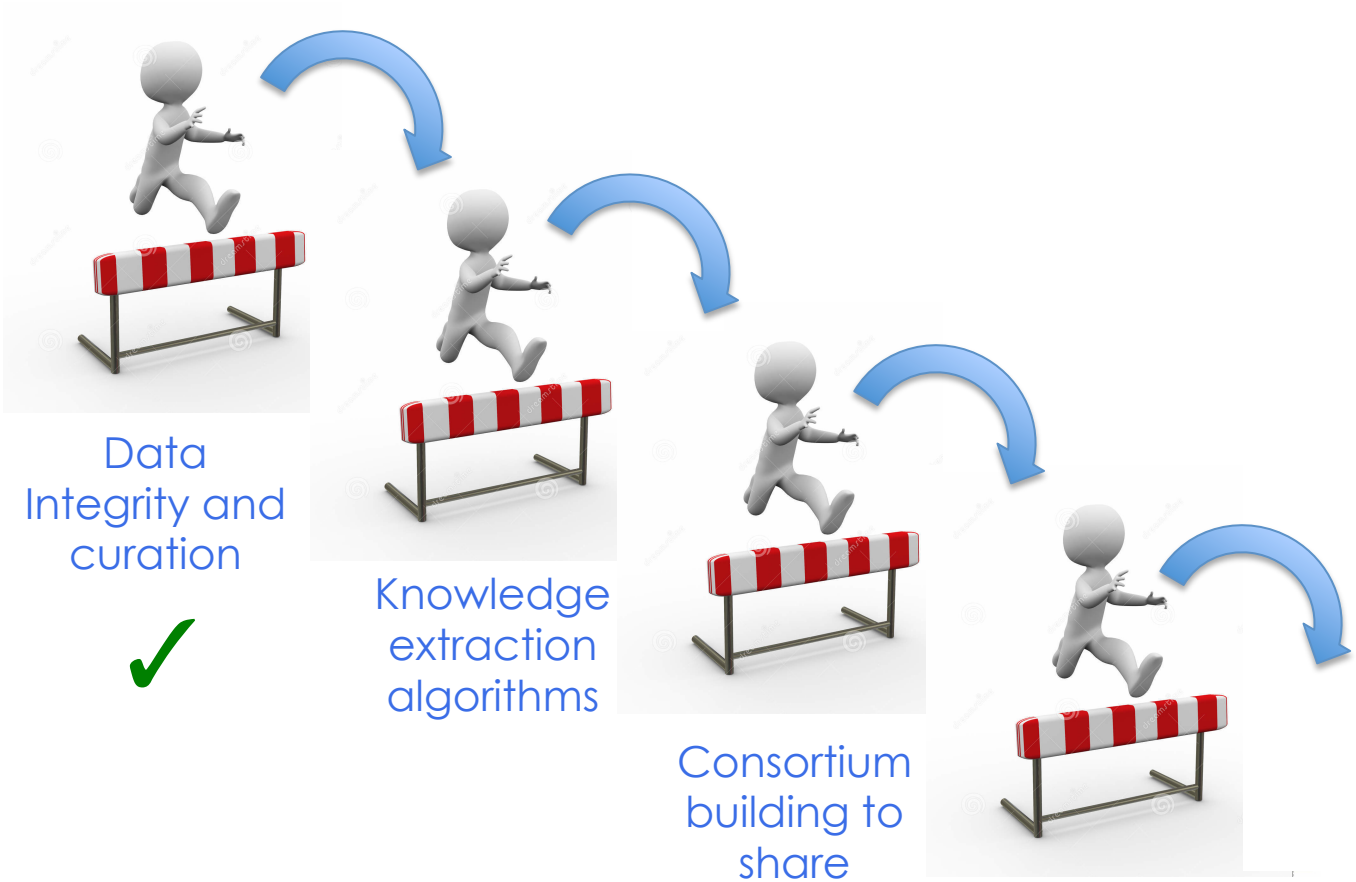
MCPairs Platform



- Extract rules using *Advanced* Matched Molecular Pair Analysis
- Knowledge is captured as transformations
  - **divorced from structures => *sharable***

Protect the IP jewels

Barriers Broken to Sharing Knowledge



Data Integrity and curation

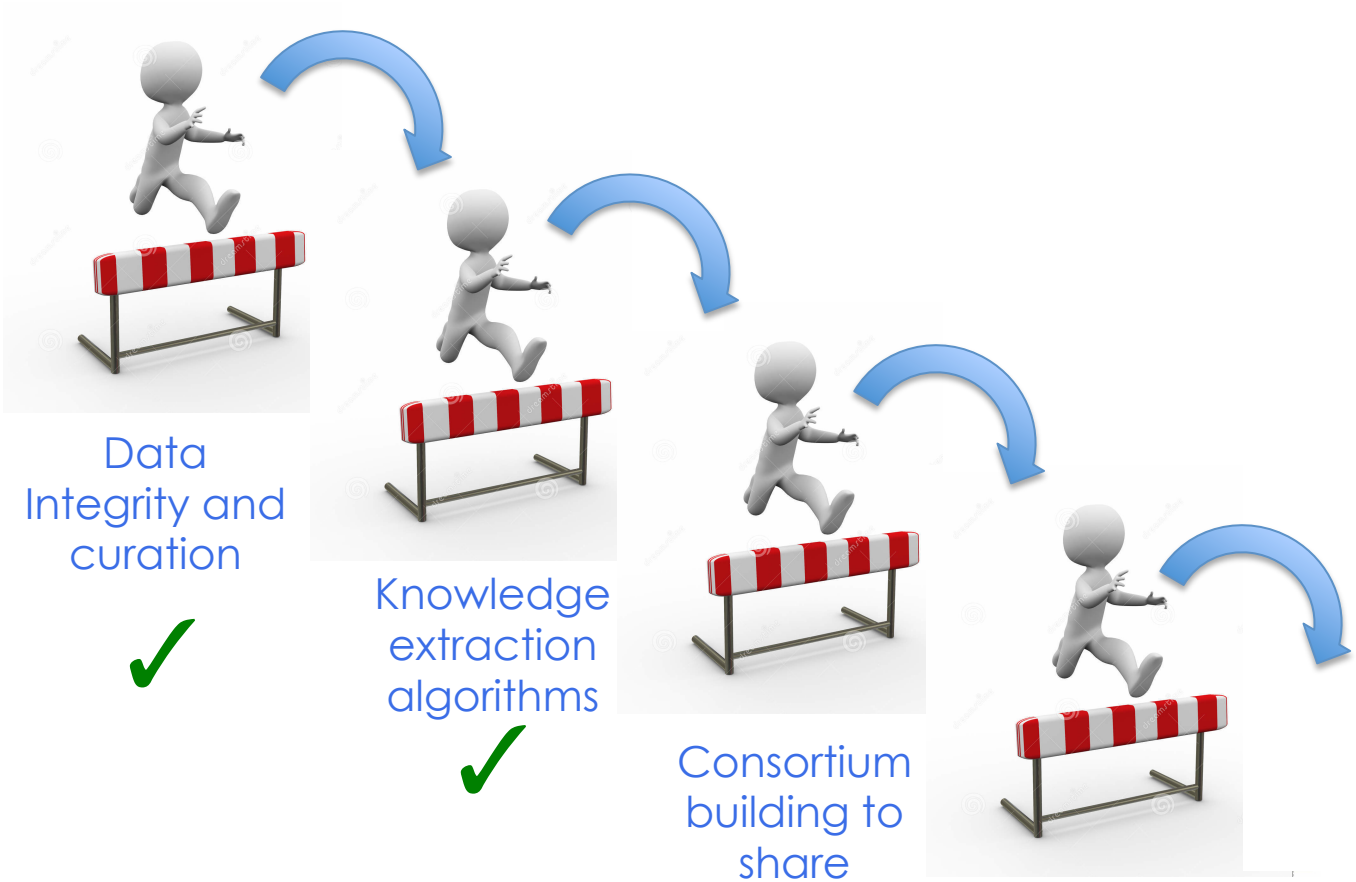
✓

Knowledge extraction algorithms

Consortium building to share knowledge

Into the minds of chemists

Barriers Broken to Sharing Knowledge



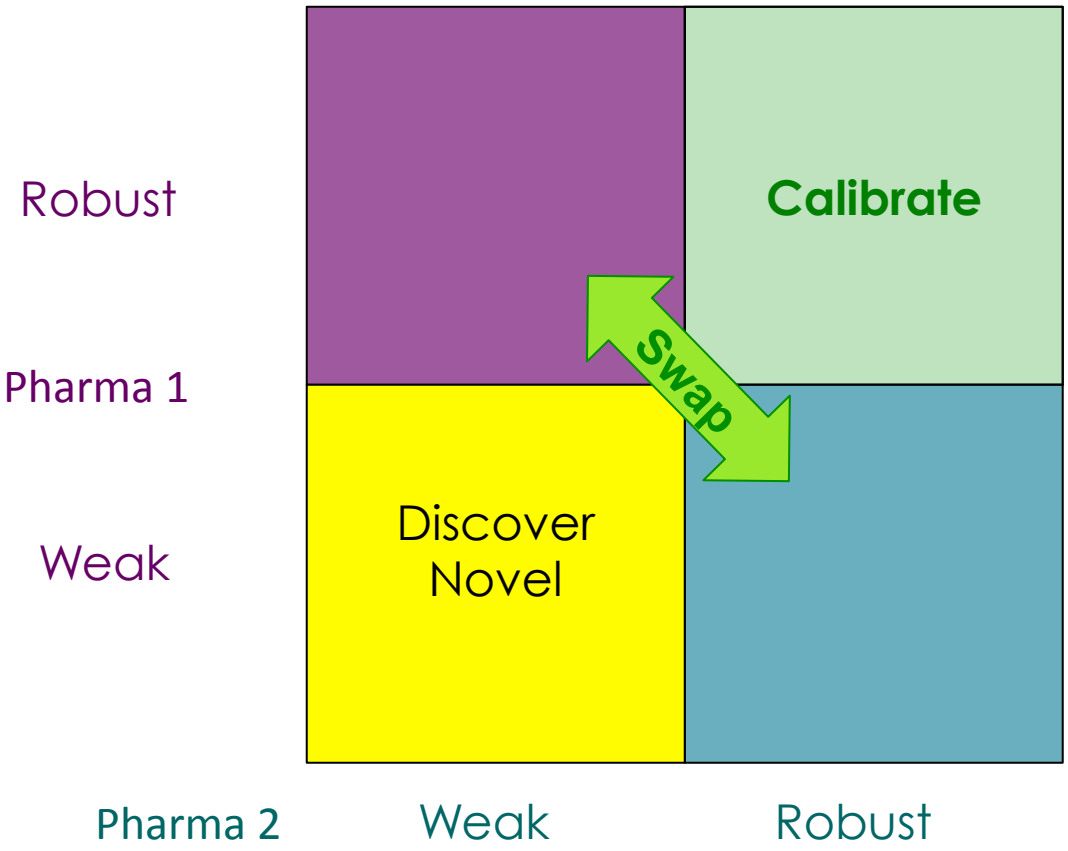
Data Integrity and curation ✓

Knowledge extraction algorithms ✓

Consortium building to share knowledge

Into the minds of chemists

Merging knowledge



- Use the transforms that are robust in both companies to calibrate assays.
- Once the assays are calibrated against each other the transform data can be combined to build support in poorly exemplified transforms
- Methodology precedented in other fields

Merging Datasets

- Datasets are standardized by comparison of transformations shared by contributing companies
- Transformations are examined at the "pair example" level
- Minimum of 6 transformations, each with a minimum of 6 pairs (42 compounds bare minimum) required to standardise
- "calibration factors" extracted to standardize the datasets to a common value – mean of calibration factors 0.94, typical range 0.8-1.2.
- Datasets with too few common transformations have standard compound measurements shared for calibration.

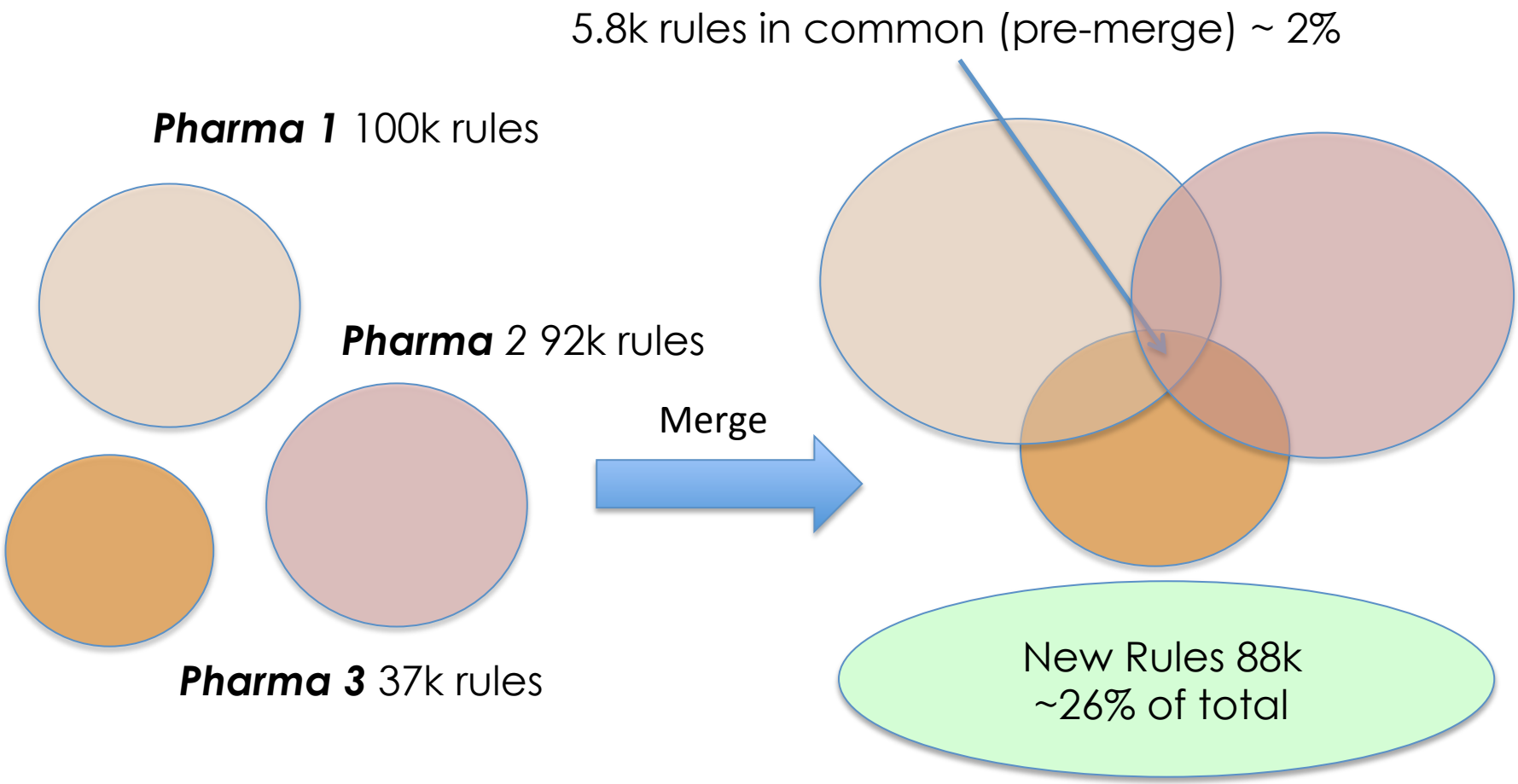
"Blinded" source of transformations

Current Knowledge sets – GRDv3

Numbers of *statistically valid* transforms

Grouped Datasets	Number of Rules
$logD_{7.4}$	153449
Merged solubility	46655
In vitro microsomal clearance: Human, rat ,mouse, cyno, dog	88423
In vitro hepatocyte clearance : Human, rat ,mouse, cyno, dog	26627
MCDK permeability A-B / B – A efflux	1852
Cytochrome P450 inhibition: 2C9, 2D6 , 3A4 , 2C19 , 1A2	40605
Cardiac ion channels NaV 1.5 , hERG ion channel inhibition	15636
Glutathione Stability	116
plasma protein or albumin binding Human, rat ,mouse, cyno, dog	64622

Merging knowledge – GRDv1

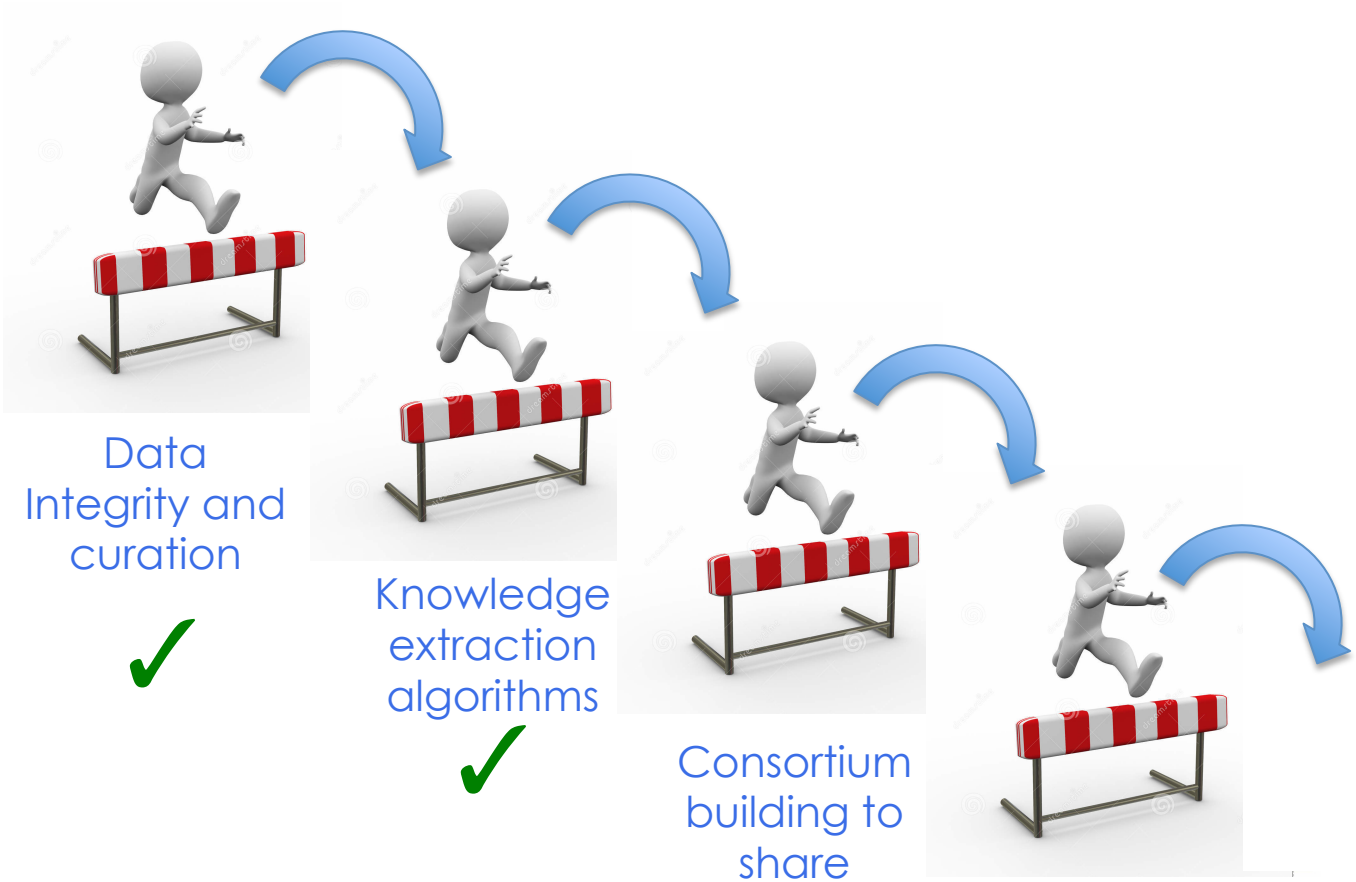


Combining data yields brand new rules
Gains: 300 - 900%

Key findings:

- Secure sharing of large scale ADMET knowledge between large Pharma is possible
- The collaboration generated great synergy
- Many findings are highly significant.
- MMP is a great tool for idea generation.
- The rules have been used in drug-discovery projects and generated meaningful results
- MMPA methodology can be extended to extract pharmacophores

Barriers Broken to Sharing Knowledge



Data Integrity and curation ✓

Knowledge extraction algorithms ✓

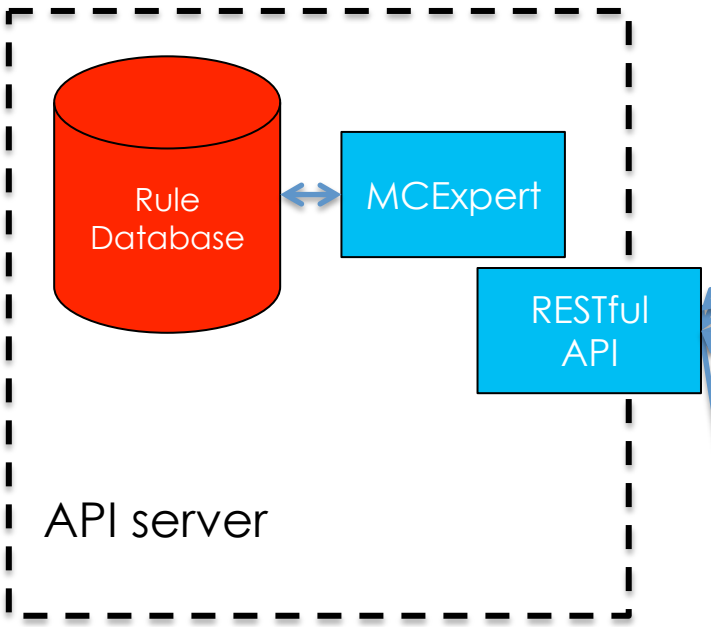
Consortium building to share knowledge ✓

Into the minds of chemists

Integrating knowledge exploitation

..many possible connections

..high Stability and flexibility**

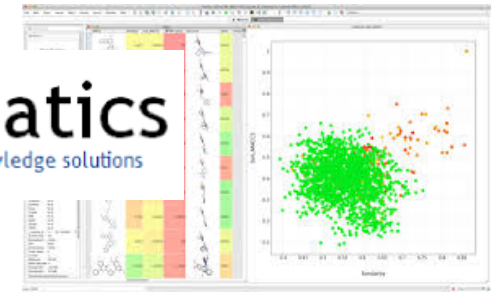


elixir | insight through communication

chemTrax

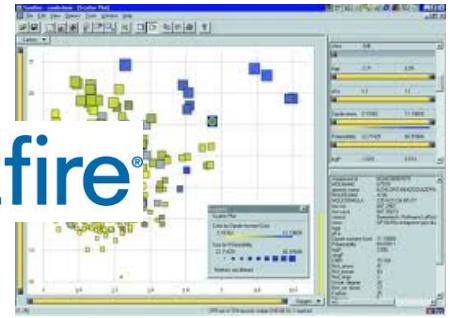
Web tool

dotmatics knowledge solutions



visualise

Spotfire®



visualise

Chemistry Shape and electrostatics

Enhanced decisions

KNIME

workflow

** api.medchemica.com has delivered 18 months of uptime and drives SaltTraX (Elixir)
Approx 50 chemists use the tool

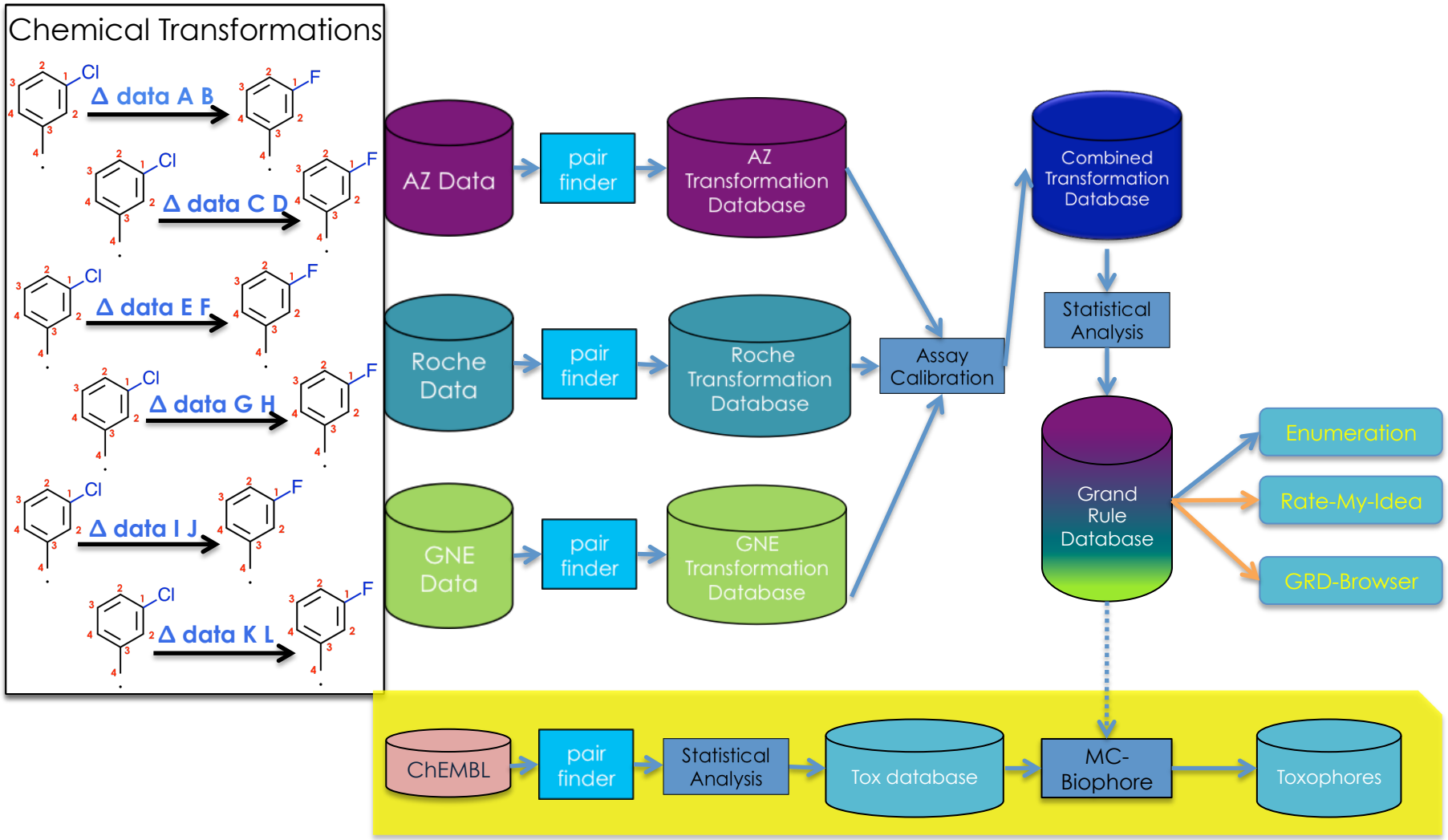
BigChem

Roche

Genentech
A Member of the Roche Group

Matched Molecular Pair Analysis (MMPA) enables SAR sharing

Without sharing underlying structures and data



BigChem

Roche

Genentech
A Member of the Roche Group

Collaborators and Users - experience

Roche

Genentech
A Member of the Roche Group

elixir | insight through communication

ROYAL SOCIETY OF CHEMISTRY

ICR The Institute of Cancer Research

UNICAMP

OpenEye SCIENTIFIC

Galápagos

Public Health England

SYNthesis med chem

BigChem

Medicines for Malaria Venture

BLUEBERRY THERAPEUTICS

SGC

THE UNIVERSITY of LIVERPOOL

THE UNIVERSITY OF WARWICK

UNIVERSITY OF DUNDEE

Liverpool John Moores University

UNIVERSITY OF LEEDS

The University Of Sheffield.

CANCER RESEARCH UK

MANCHESTER 1824 The University of Manchester

BigChem Roche Genentech A Member of the Roche Group

A Collaboration of the willing

Craig Bruce	OE
John Cumming	Roche
David Cosgrove	C4XD
Andy Grant★	
Martin Harrison	Elixir
Huw Jones	Base360
Al Rabow	Consulting
David Riley	AZ
Graeme Robb	AZ
Attilla Ting	AZ
Howard Tucker	retired
Dan Warner	Myjar
Steve St-Galley	Syngenta
David Wood	JDR
Lauren Reid	MedChemica
Shane Monague	MedChemica
Jessica Stacey	MedChemica
Andy Barker	Consulting
Pat Barton	AZ
Andy Davis	AZ
Andrew Griffin	Elixir
Phil Jewsbury	AZ
Mike Snowden	AZ
Peter Sjo	AZ
Martin Packer	AZ
Manos Perros	Entasis Therapeutics
Nick Tomkinson	AZ
Martin Stahl	Roche
Jerome Hert	Roche
Martin Blapp	Roche
Torsten Schindler	Roche
Paula Petrone	Roche
Christian Kramer	Roche
Jeff Blaney	Genentech
Hao Zheng	Genentech
Slaton Lipscomb	Genentech
Alberto Gobbi	Genentech

ACS Philadelphia 2016

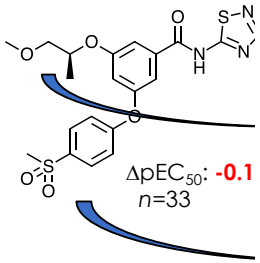
Glucokinase Activators

- Fix hERG problem whilst maintaining potency

MMPA

ΔpEC_{50}: **-0.1** n=27 $\Delta logD$: **-0.6** n=27 $\Delta hERG\ pIC_{50}$:**-0.5** n=7

pEC_{50}: 7.0
logD: 2.9
hERG pIC_{50}: 5.1



ΔpEC_{50}: **-0.1** n=33

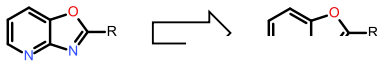
ACS Philadelphia 2016

A Less Simple Example

Increase logD and gain solubility

Roche

Question:



What is the effect on lipophilicity and solubility?

Roche data is inconclusive! (2 pairs

Waring *et al*, Med. Chem. Commun

Available Statistics:

Property	Number of Observations
logD	8
Log(Solubility)	14

MedChemica | 2016

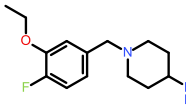
ACS Philadelphia 2016

Solving a ^{t}Bu metabolism issue

CANCER RESEARCH UK | MANCHESTER INSTITUTE

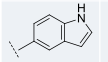
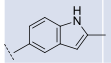
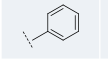
Roger Butlin
Rebecca Newton
Allan Jordan

Roche Example:



logD = 2.65
Kinetic solubility = 84 µg/n
IC50 SST5 = 0.8 µM

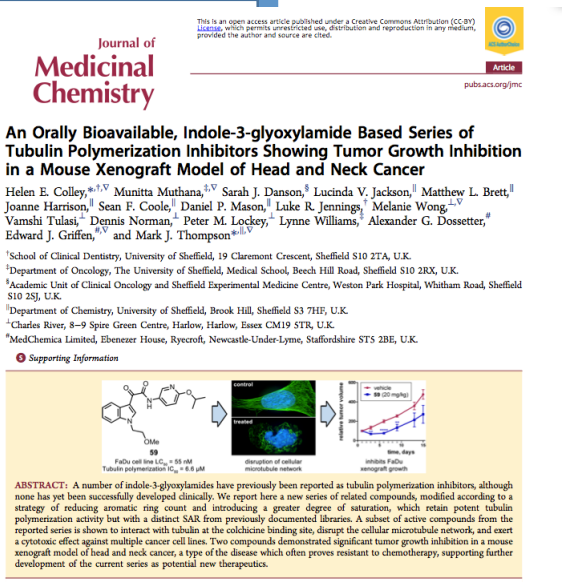
MedChemica | 2016

R1 \ R2	Benchmark compound: tBu				Me	
	99 392	16 64	78 410	53 550	99 288	78 515
	35 128				24 62	
	39 445	3 21			20 27	

Predicted to offer most improvement in microsomal stability (in at least 1 species / assay)

- Data shown are Cl_{int} for HLM and MLM (top an

MedChemica | 2016



Journal of Medicinal Chemistry

This is an open access article published under a Creative Commons Attribution (CC-BY) License, which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

Article

pubs.acs.org/jmc

An Orally Bioavailable, Indole-3-glyoxylamide Based Series of Tubulin Polymerization Inhibitors Showing Tumor Growth Inhibition in a Mouse Xenograft Model of Head and Neck Cancer

Helen E. Colley,[*,†,▽] Munitta Muthana,[‡,▽] Sarah J. Danson,[§] Lucinda V. Jackson,[∥] Matthew L. Brett,[∥] Joanne Harrison,[∥] Sean F. Coole,[∥] Daniel P. Mason,[∥] Luke R. Jennings,[†] Melanie Wong,[⊥,▽] Vamshi Tulasi,[⊥] Dennis Norman,[⊥] Peter M. Lockey,[⊥] Lynne Williams,[†] Alexander G. Dossetter,[#] Edward J. Griffen,[#,▽] and Mark J. Thompson[*,∥,▽]

[†]School of Clinical Dentistry, University of Sheffield, 19 Claremont Crescent, Sheffield S10 2TA, U.K.
[‡]Department of Oncology, The University of Sheffield, Medical School, Beech Hill Road, Sheffield S10 2RX, U.K.
[§]Academic Unit of Clinical Oncology and Sheffield Experimental Medicine Centre, Weston Park Hospital, Whitham Road, Sheffield S10 2SJ, U.K.
[∥]Department of Chemistry, University of Sheffield, Brook Hill, Sheffield S3 7HF, U.K.
[⊥]Charles River, 8–9 Spire Green Centre, Harlow, Harlow, Essex CM19 5TR, U.K.
[#]MedChemica Limited, Ebenezer House, Ryecroft, Newcastle-Under-Lyme, Staffordshire ST5 2BE, U.K.

Supporting Information

ABSTRACT: A number of indole-3-glyoxylamides have previously been reported as tubulin polymerization inhibitors, although none has yet been successfully developed clinically. We report here a new series of related compounds, modified according to a strategy of reducing aromatic ring count and introducing a greater degree of saturation, which retain potent tubulin polymerization activity but with a distinct SAR from previously documented libraries. A subset of active compounds from the reported series is shown to interact with tubulin at the colchicine binding site, disrupt the cellular microtubule network, and exert a cytotoxic effect against multiple cancer cell lines. Two compounds demonstrated significant tumor growth inhibition in a mouse xenograft model of head and neck cancer, a type of the disease which often proves resistant to chemotherapy, supporting further development of the current series as potential new therapeutics.

BigChem | Roche | Genentech A Member of the Roche Group

Cathepsin K – Di-methoxy surprise – Man and Machine

pIC50	7.95
LogD	0.67
HLM	<2.0
Solubility	280 μ M
DTM	~1.0 mg/kg UID

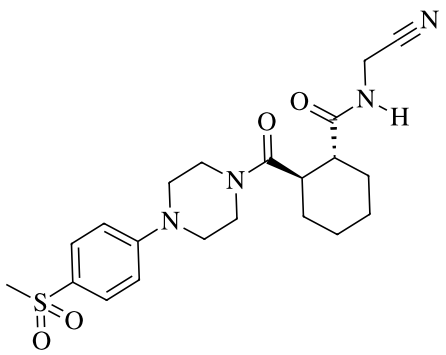
Potent

Too polar / Renal Cl

Increase in LogP,
Properties improved

pIC50	8.2
LogD	2.8
HLM	<1.0
Solubility	>1400 μ M
DTM	0.01 mg/kg UID

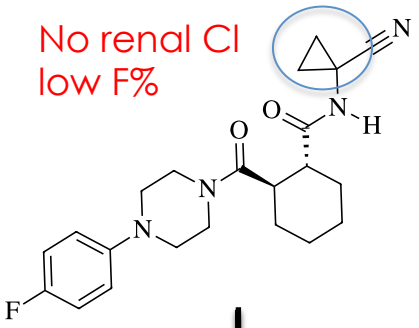
High F% / stability maximised



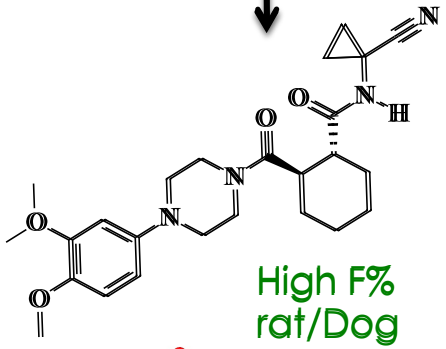
ΔpIC_{50}	- 0.1
$\Delta LogD$	+1.4
$\Delta pSol$	+1.2
ΔHLM	+ 0.25

Solubility

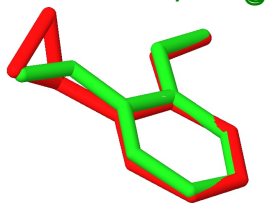
No renal Cl
low F%



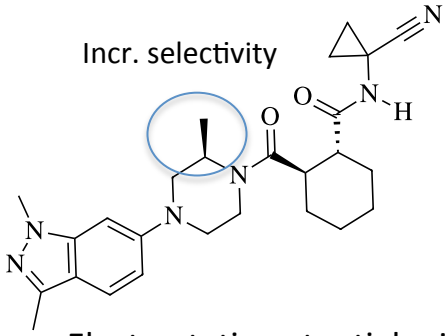
ΔpIC_{50}	+0.1
$\Delta LogD$	- 0.7
$\Delta pSol$	~0.0
ΔHLM	- 0.25



High F%
rat/Dog



Incr. selectivity



Electrostatic potential minima between oxygens
Approx like N from 5-het, new compound can not form a quinoline

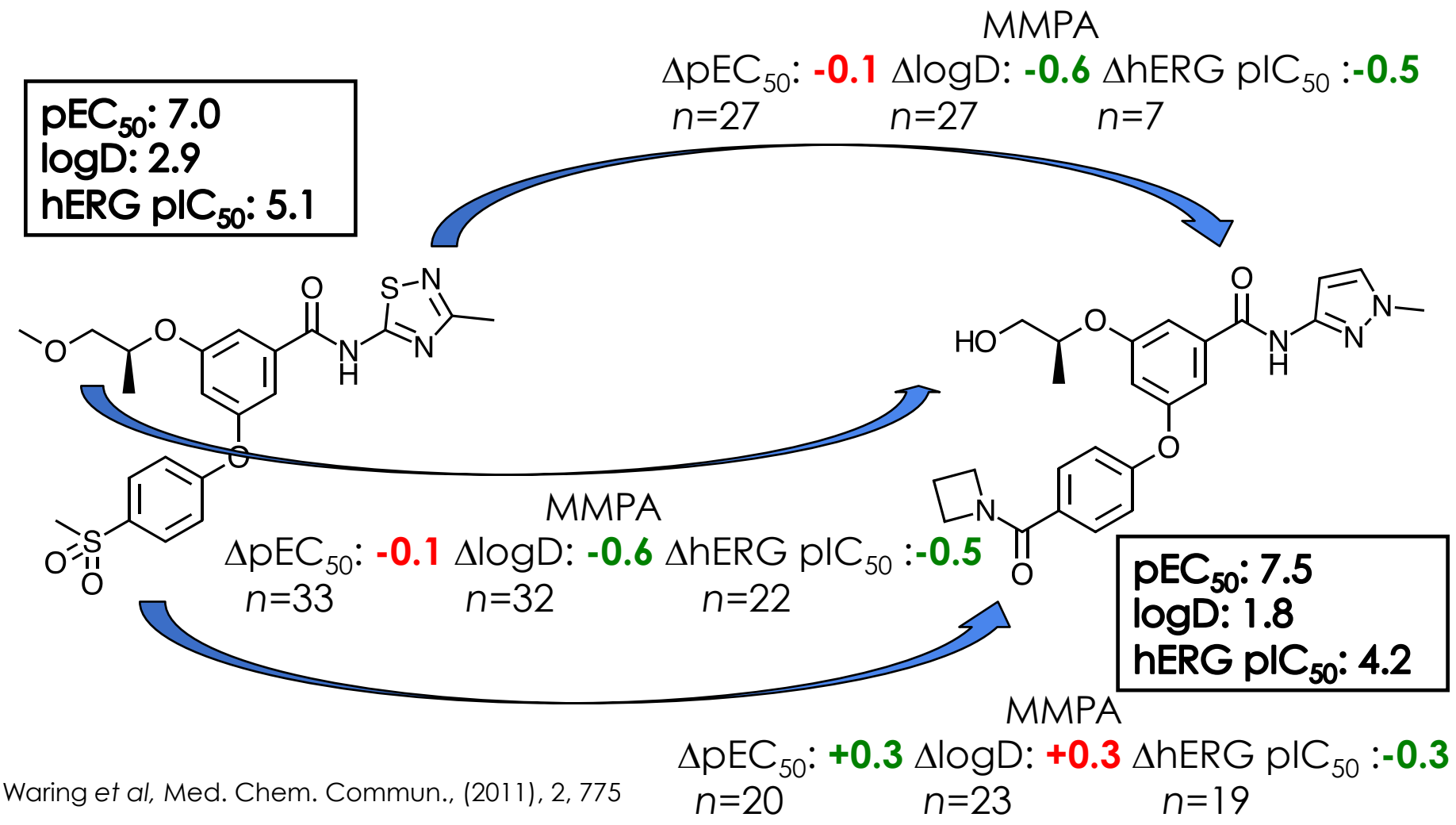
Crawford, J.J.; Dossetter, A.G *J Med Chem. 2012, 55, 8827.*
Dossetter, A. G. *Bioorg. Med. Chem.* **2010**, 4405
Lewis et al, J Comput Aided Mol Des, **2009**, *23*, 97–103

PDB - 97% of structures

Roche Genentech A Member of the Roche Group

Glucokinase Activators

- Fix hERG problem whilst maintaining potency



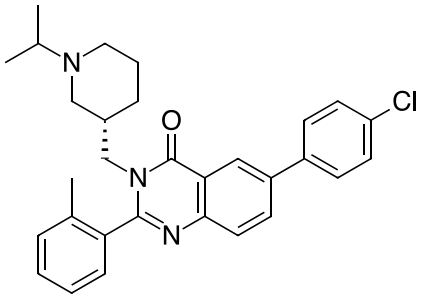
Waring *et al,* Med. Chem. Commun., (2011), 2, 775

Knowledge Based Design – MPO

Ghrelin Inverse agonists

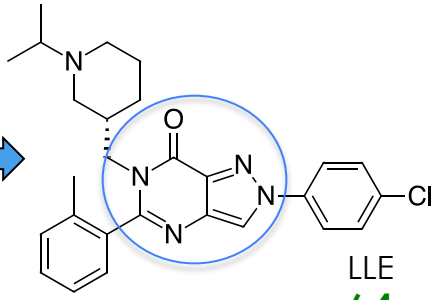
- Novel more efficient core required, improve hERG for CD
- CNS penetration, good potency and deliver tool for in vivo testing

pIC50	9.9
logD	5.0
hERG pIC5	5.0
LLE	**4.9**
very potent	
very lipophilic	



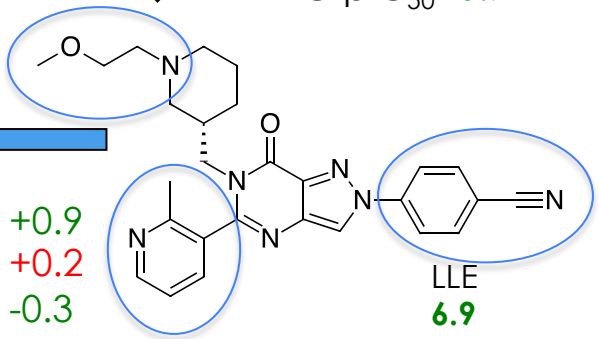
MMPA Cores

ΔpIC_{50}	-0.4
$\Delta logD$	-1.8
$\Delta hERG\ pIC_{50}$	+0.4



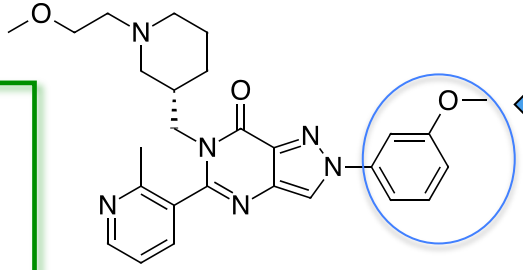
LLE **6.4**

ΔpIC_{50}	-2.2
$\Delta logD$	-2.2
$\Delta hERG\ pIC_{50}$	-0.7



LLE **6.9**

ΔpIC_{50}	+0.9
$\Delta logD$	+0.2
$\Delta hERG\ pIC_{50}$	-0.3



100 compounds made

pIC50	8.2
logD	1.3
hERG pIC50	4.4
LLE	**6.9**

McCoull, Dossetter *et al,* Med. Chem. Commun., (2013), 4, 456

LLE = lipophilic ligand efficiency:
LLE=pIC50-logD

BigChem Roche Genentech A Member of the Roche Group

Early successes

From GRDv1 May 2014

Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

An Orally Bioavailable, Indole-3-glyoxylamide Based Series of Tubulin Polymerization Inhibitors Showing Tumor Growth Inhibition in a Mouse Xenograft Model of Head and Neck Cancer

Helen E. Colley,[*,†,▽] Munitta Muthana,[‡,▽] Sarah J. Danson,[§] Lucinda V. Jackson,[∥] Matthew L. Brett,[∥] Joanne Harrison,[∥] Sean F. Coole,[∥] Daniel P. Mason,[∥] Luke R. Jennings,[†] Melanie Wong,[⊥,▽] Vamshi Tulasi,[⊥] Dennis Norman,[⊥] Peter M. Lockey,[⊥] Lynne Williams,[‡] Alexander G. Dossetter,[#] Edward J. Griffen,[#,▽] and Mark J. Thompson[*,∥,▽]

[†]School of Clinical Dentistry, University of Sheffield, 19 Claremont Crescent, Sheffield S10 2TA, U.K.

[‡]Department of Oncology, The University of Sheffield, Medical School, Beech Hill Road, Sheffield S10 2RX, U.K.

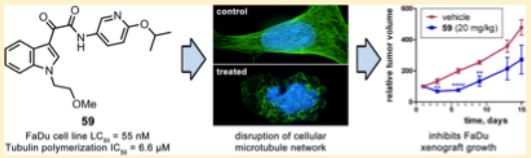
[§]Academic Unit of Clinical Oncology and Sheffield Experimental Medicine Centre, Weston Park Hospital, Whitham Road, Sheffield S10 2SJ, U.K.

[∥]Department of Chemistry, University of Sheffield, Brook Hill, Sheffield S3 7HF, U.K.

[⊥]Charles River, 8–9 Spire Green Centre, Harlow, Harlow, Essex CM19 5TR, U.K.

[#]MedChemica Limited, Ebenezer House, Ryecroft, Newcastle-Under-Lyme, Staffordshire ST5 2BE, U.K.

Supporting Information



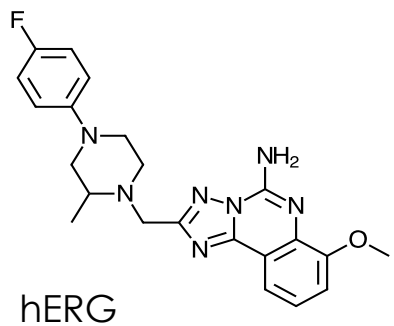
ABSTRACT: A number of indole-3-glyoxylamides have previously been reported as tubulin polymerization inhibitors, although none has yet been successfully developed clinically. We report here a new series of related compounds, modified according to a strategy of reducing aromatic ring count and introducing a greater degree of saturation, which retain potent tubulin polymerization activity but with a distinct SAR from previously documented libraries. A subset of active compounds from the reported series is shown to interact with tubulin at the colchicine binding site, disrupt the cellular microtubule network, and exert a cytotoxic effect against multiple cancer cell lines. Two compounds demonstrated significant tumor growth inhibition in a mouse xenograft model of head and neck cancer, a type of the disease which often proves resistant to chemotherapy, supporting further development of the current series as potential new therapeutics.

J. Med. Chem., 2015, 58 (23), pp 9309–9333
DOI: 10.1021/acs.jmedchem.5b01312

BigChem Roche Genentech A Member of the Roche Group

Comparison of Merck in-house MMPA with SALTMiner™

Structure:



ADMET Issue: hERG

Lead A_{2A} receptor antagonist compound in Merck Parkinson's project

- Also shows potent binding to the hERG ion channel
- Deng *et al* performed in-house MMPA on hERG binding compound data and have published 18 resulting fluorobenzene transformations, which they have synthesized and tested for hERG activity

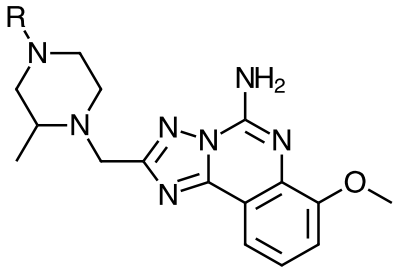
Deng et al,
Bioorg. & Med Chem Let (2015),
doi: http://dx.doi.org/10.1016/j.bmcl.2015.05.036

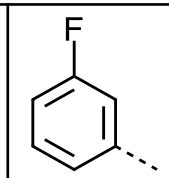
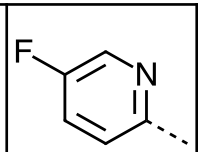
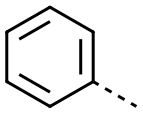
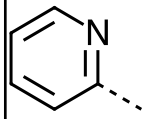
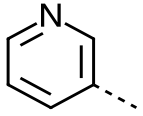
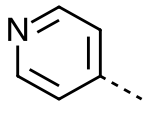
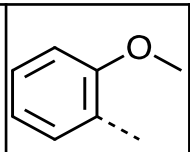
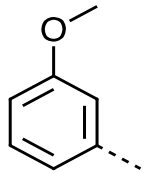
138 suggestion molecules with predicted improvement in hERG binding

How many match the results of Merck?

Results:

8 out of the 18 fluorobenzene transformations produced by Merck were also suggested by MCExpert to decrease hERG binding:



R group:								
Measured hERG pIC50 change	-1.187	-1.149	-1.038	-1.215	-1.157	-0.149	-1.487	-1.133
GRD median historic pIC50 change	0	-0.171	-0.1	-0.283	-0.219	-0.318	-0.159	-0.103

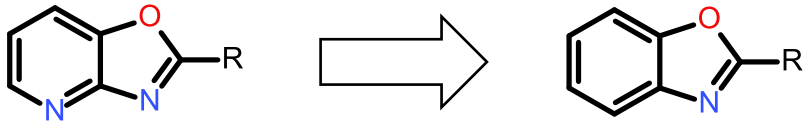
Searching the GRD for transformations that *increase* hERG there were ***none*** that matched the remaining 10 of 18 transformations in the paper.

MCExpert also suggested an additional 50 fluorobenzene replacements to decrease hERG binding NOT mentioned in the publication.

A Less Simple Example

Increase logD and gain solubility

Question:

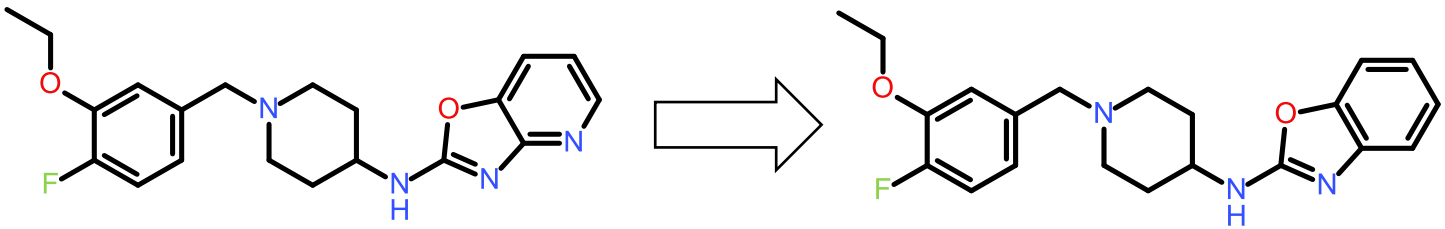


What is the effect on lipophilicity and solubility?
Roche data is inconclusive! (2 pairs for logD, 1 pair for solubility)

Available Statistics:

Property	Number of Observations	Direction	Mean Change	Probability
logD	8	Increase	1.2	100%
Log(Solubility)	14	Increase	1.4	92%

Roche Example:



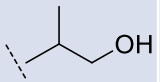
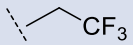
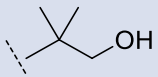
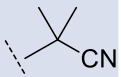
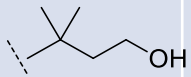
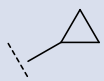
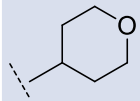
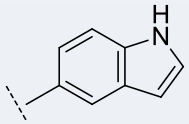
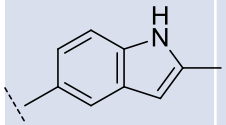
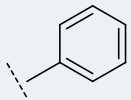
logD = 2.65
Kinetic solubility = 84 µg/ml
IC50 SST5 = 0.8 µM

logD = 3.63
Kinetic solubility = >452 µg/ml
IC50 SST5 = 0.19 µM

Solving a ^{t}Bu metabolism issue

CANCER RESEARCH UK | MANCHESTER INSTITUTE

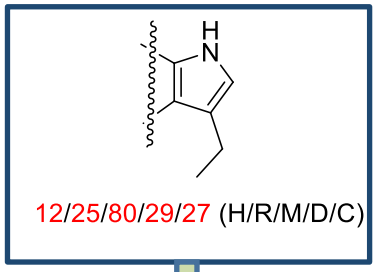
Roger Butlin
Rebecca Newton
Allan Jordan

	Benchmark compound	Predicted to offer most improvement in microsomal stability (in at least 1 species / assay) ←									
R1 \ R2	tBu				Me				Et	iPr	
	99 392	16 64	78 410	53 550	99 288	78 515	41 35	98 327		92 372	24 247
	35 128				24 62				60 395		
	39 445	3 21			20 27			57 89		54 89	

- Data shown are Cl_{int} for HLM and MLM (top and bottom, respectively)

BigChem | Roche | Genentech A Member of the Roche Group

The application helped lead optimization in project



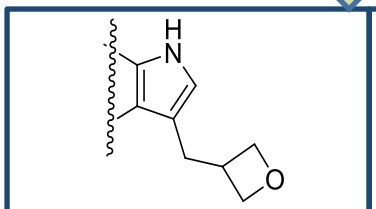
Objective: improve metabolic stability

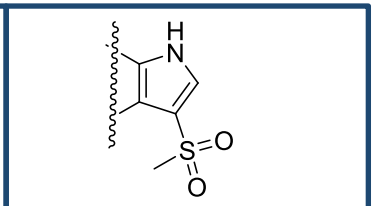
MMP Enumeration

- 193 compounds
- Enumerated

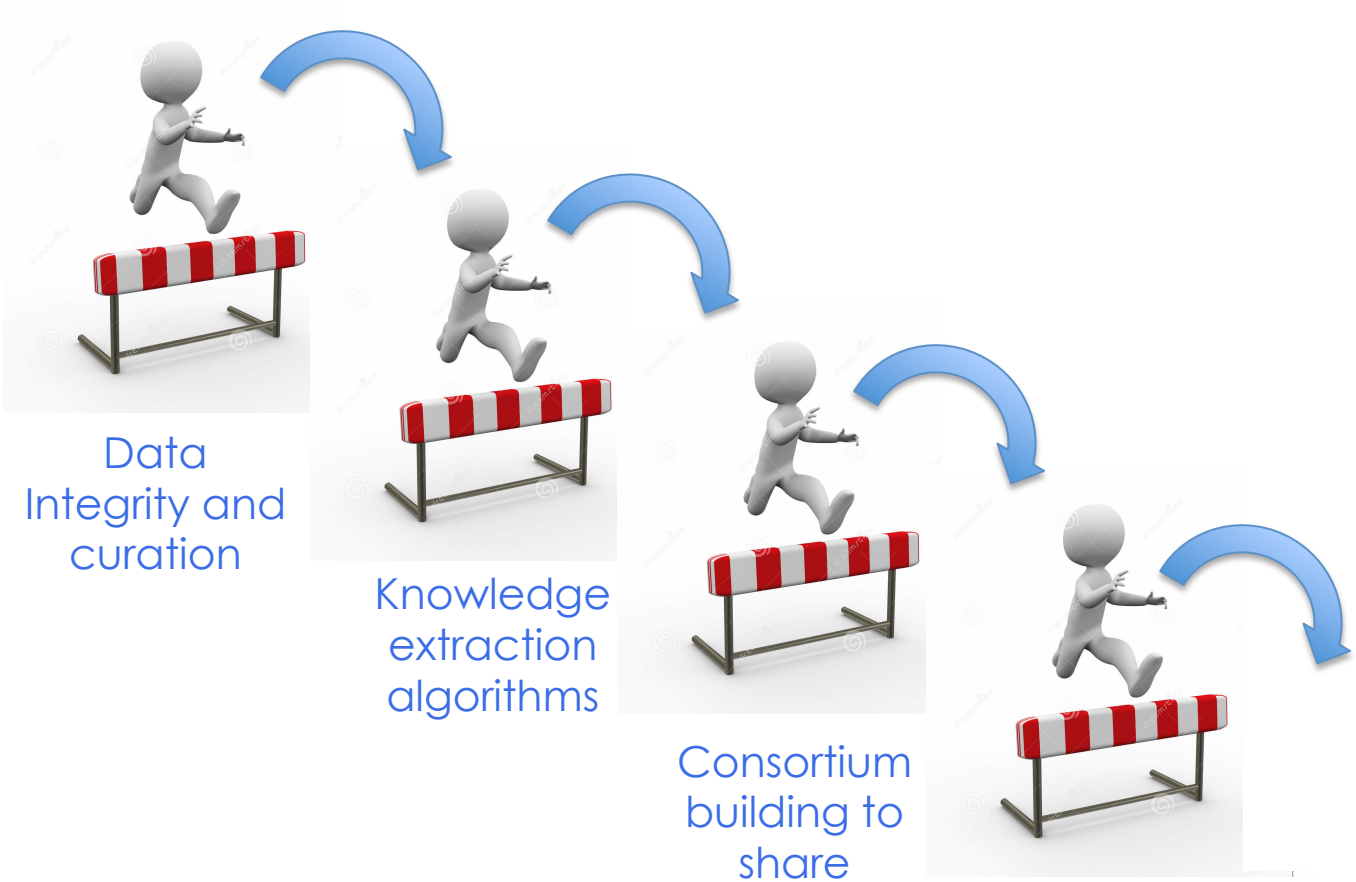
Calculated Property Docking

8 compounds synthesized





Barriers Broken to Sharing Knowledge



Data Integrity and curation

Knowledge extraction algorithms

Consortium building to share knowledge

Into the minds of chemists

Data Integrity and Curation

Structural

- Extensive standards for inclusion of mixtures, chiral compounds, salt forms
- Tautomer and charge state canonicalisation client side
- Automated validation of structures run client side

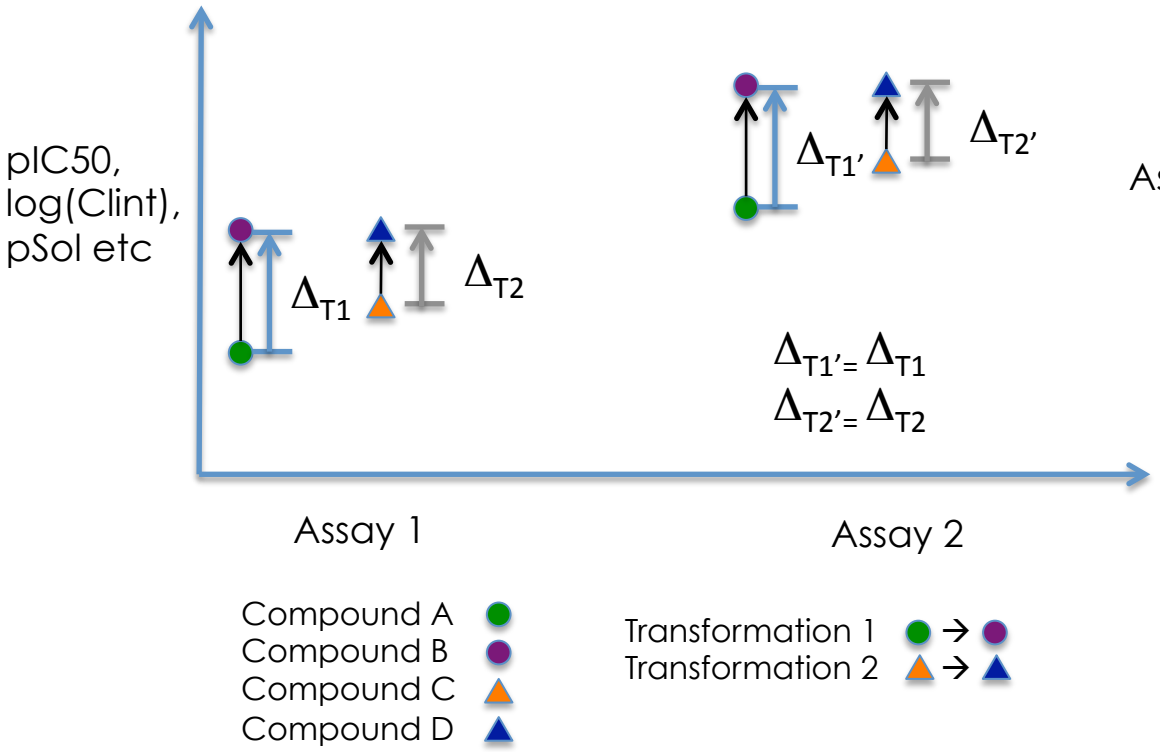
  = "clean" comparable structures submitted to pair finding

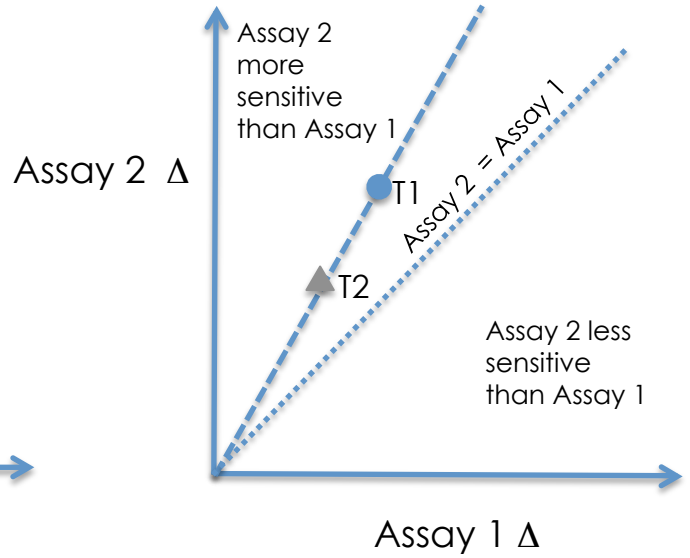
Measured Data

- Assay protocols reviewed prior to merging
- Precise documentation on unit definitions and data reporting standards
- Option to share standard compound measured values
- Automated extensive data validation checks prior to merging data

"client side" = behind Pharma firewall

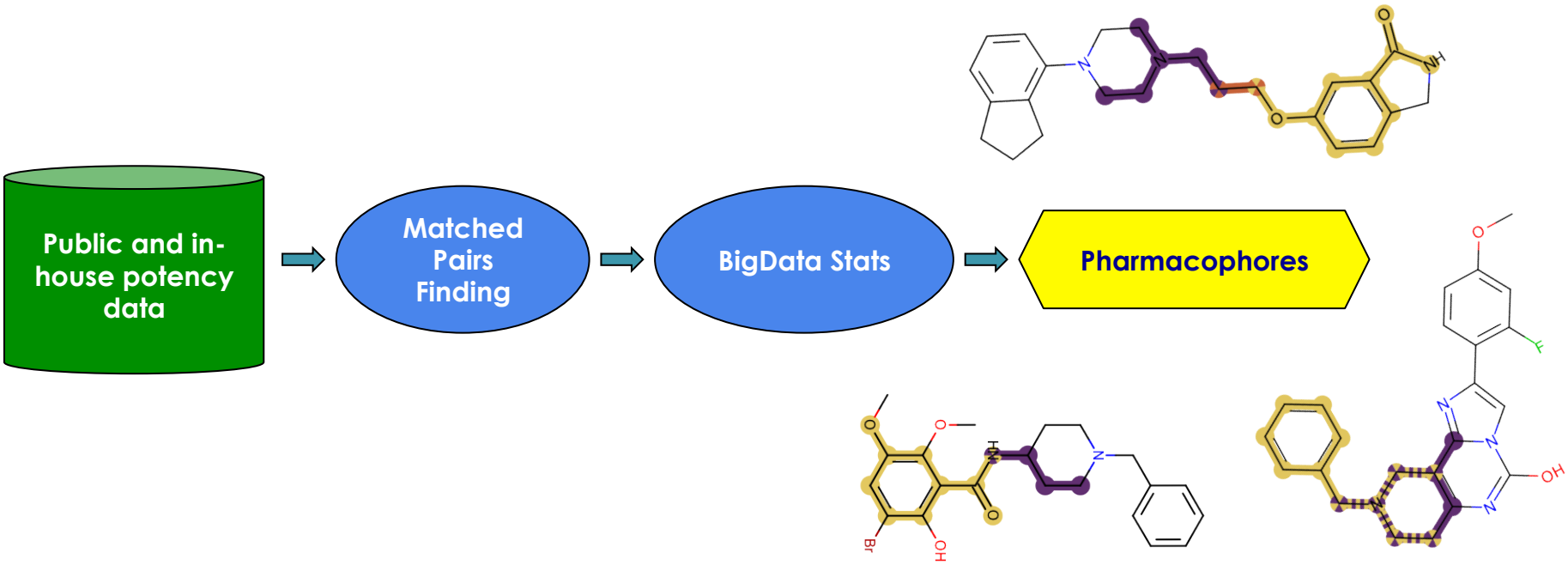
Calibrating Assays



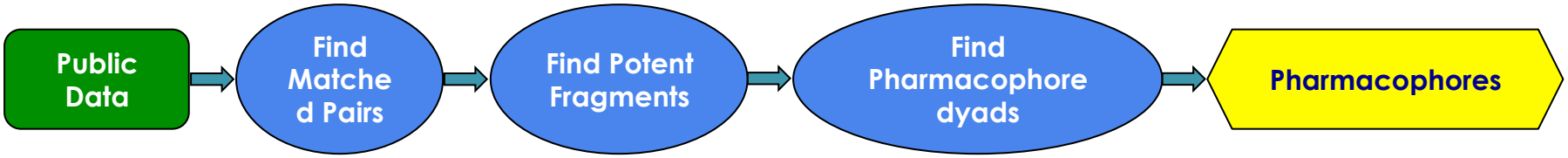


- Sets of transformations can be calibrated against each other as we are comparing Δ values in assays not absolute values
- Assays are usually linearly displaced against each other
- Data analysis equivalent of FEP

Pharmacophores and Toxophores by extended analysis from the MMPA



Critical safety target analysis



- Build models using 10-fold cross validated PLS
- Assess using ROC / BEDROC, R^2 vs 100 fold y-scrambled R^2 and geomean odds ratio

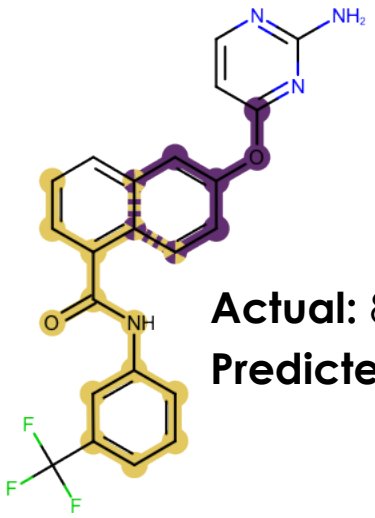
Target	Number of compounds	Number of compound pairs	Number of Fragments	Number of Pharmacophore dyads after filtering	R^2	RMSEP	ROC	odds_ratio (geomean)
Acetylcholine esterase - human	383	27755	44	10	0.43	1.57	0.80	4
β 1 adrenergic receptor	505	145447	276	313	0.64	0.70	0.96	833
Androgen receptor	1064	113163	186	46	0.47	0.77	0.86	140
CB1 canabinnoid receptor	1104	88091	165	90	0.61	1.02	0.87	96
CB2 canabinnoid receptor	1112	82130	194	158	0.19	0.85	0.64	5.7
Dopamine D2 receptor - human	3873	230962	483	602	0.42	0.88	0.69	110
Dopamine D2 receptor - rat	1807	118736	267	377	0.29	0.85	0.78	125
Dopamine Transporter	1470	106969	282	336	0.58	0.73	0.88	141
GABA A receptor	848	39494	106	167	0.70	0.76	0.97	560
hERG ion channel	4189	242261	392	76	0.61	0.96	0.92	55
5HT2a receptor	642	50870	197	267	0.61	0.59	0.83	600
Monoamine oxidase	264	15439	44	11	0.12	1.25	0.48	181
Muscarinic acetylcholine receptor M1	628	48200	97	510	0.62	0.94	0.89	48
μ opioid receptor	1128	37184	33	11	0.69	1.30	0.87	81

J. Bowes, *et al Nat. Rev. Drug Discov.*, vol. 11, no. 12, pp. 909–922, Nov. 2012

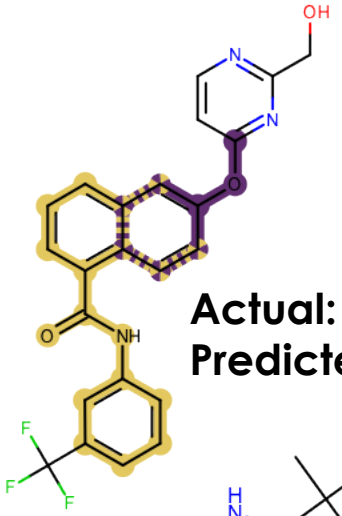
Novartis Predictions From Our Model

Domain of Applicability….

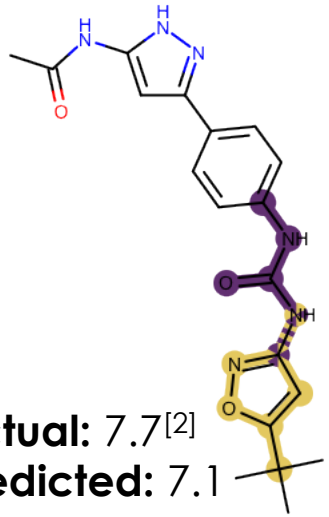
1. J MedChem(2016), Bold et al.
2. MedChem Lett (2016), Mainolfi et al.



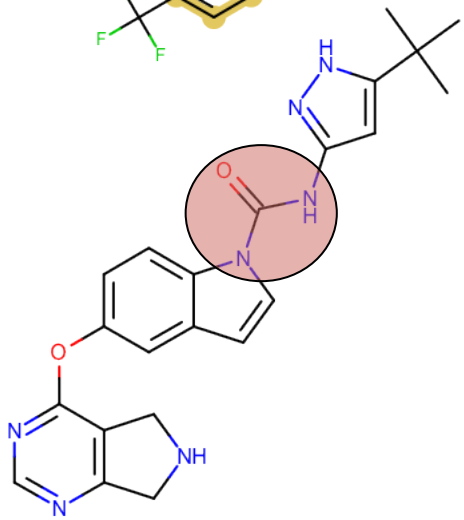
Actual: 8.4[1]
Predicted: 7.5



Actual: 7.6[1]
Predicted: 7.5

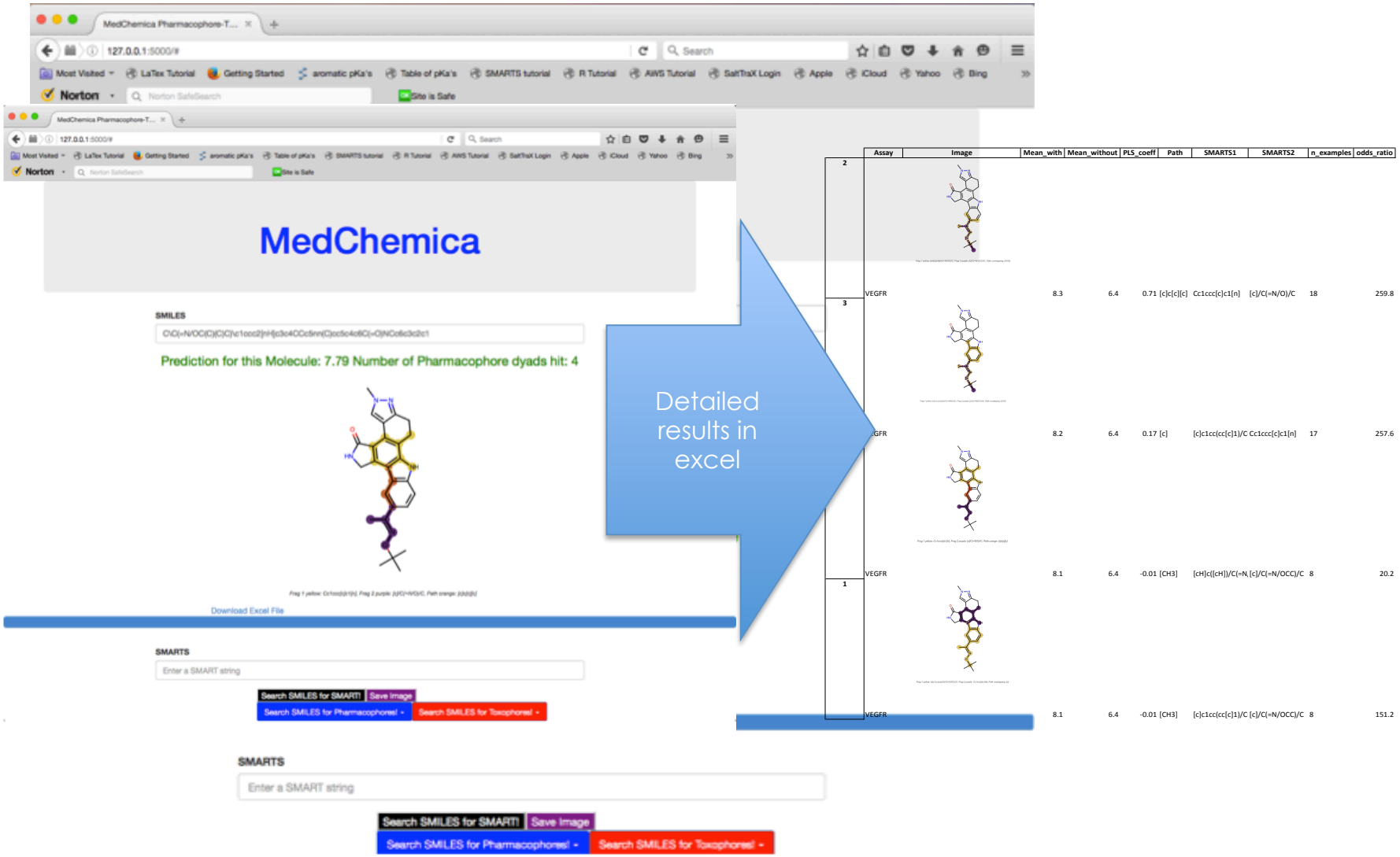


Actual: 7.7[2]
Predicted: 7.1



Actual: 9.0[2]
Predicted: *Out of Domain*

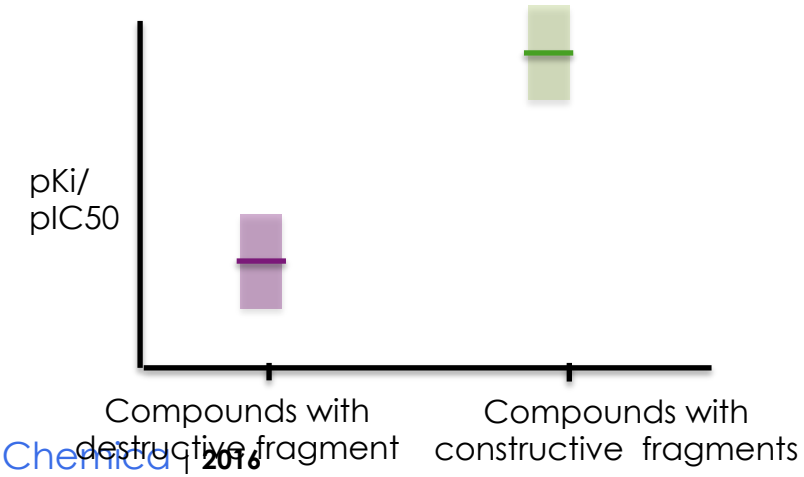
MCBiophore GUI screenshot

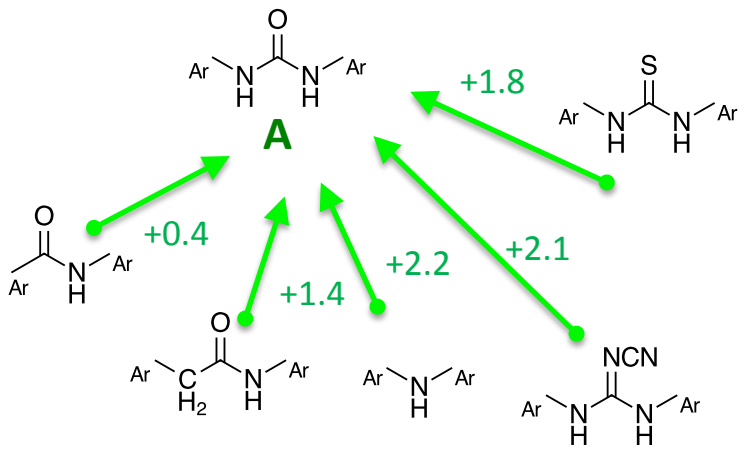


Mining transform sets to find influential fragments

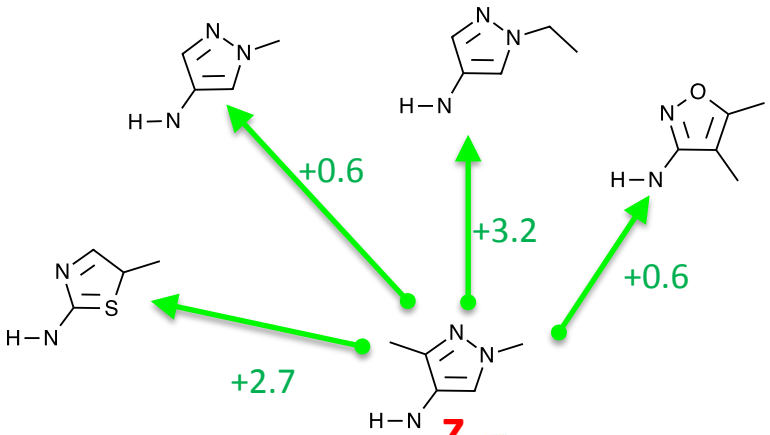
Identify the 'A' fragments associated with a *significant* number of potency decreasing changes – irrespective of what they are replaced with
'A' is 'better than anything you replace it with'

Identify the 'Z' fragments associated with a *significant* number of potency increasing changes – irrespective of what they are replaced with
'Z' is 'worse than anything you replace it with'

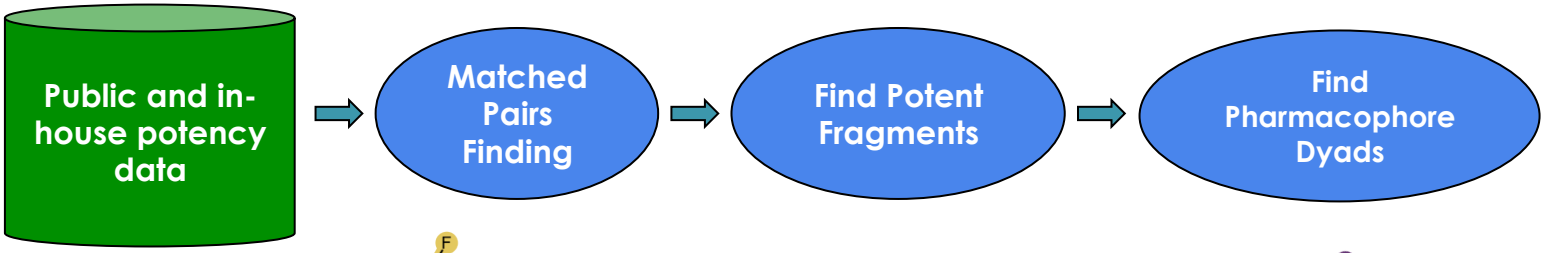


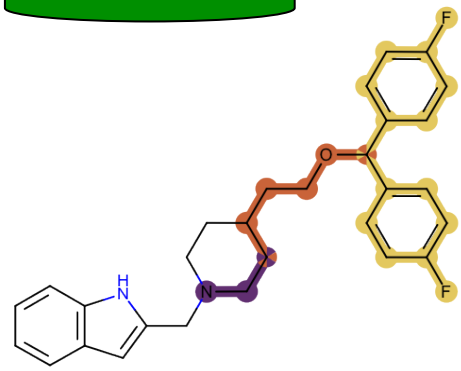


Generate Pharmacophore dyads by permutating all the fragments with the shortest path between them



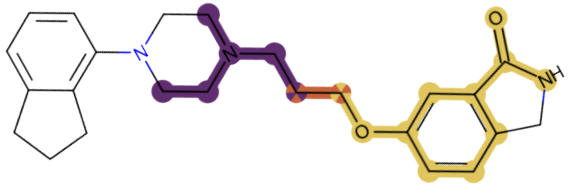
Toxophores - Detailed, specific & transparent





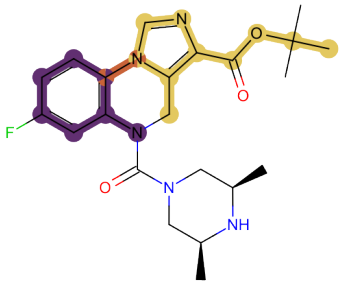
Dopamine Transporter

Actual:	9.1
Predicted:	8.6
Mean with:	8.3
Mean without:	6.7
Odds Ratio:	407



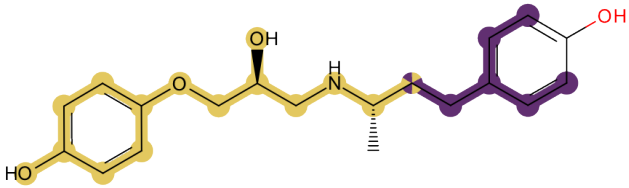
Dopamine D2 receptor human

Actual:	9.5
Predicted:	9.1
Mean with:	8.0
Mean without:	6.6
Odds Ratio:	340



GABA-A

Actual:	9.0
Predicted:	8.7
Mean with:	8.0
Mean without:	6.8
Odds Ratio:	1506



β1 adrenergic receptor

Actual:	7.8
Predicted:	7.7
Mean with:	6.5
Mean without:	5.7
Odds Ratio:	1501

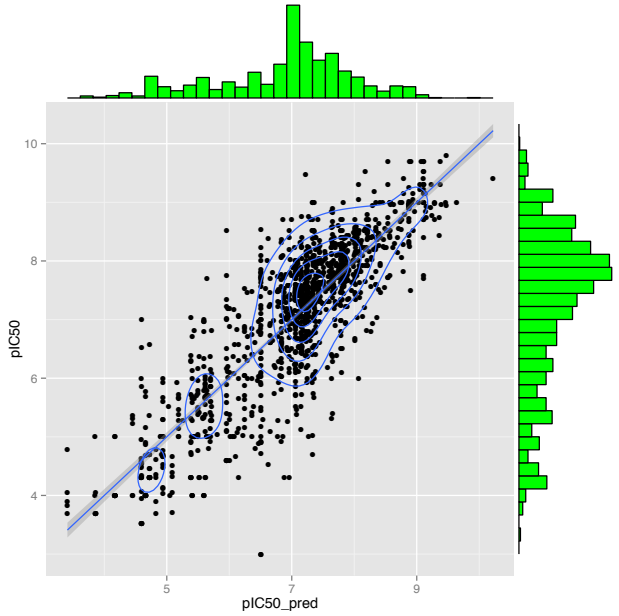
Prediction of unseen new molecules

The acid test…

- Vascular endothelial growth factor receptor 2 tyrosine kinase (KDR)
- Inhibitors have oncology and ophthalmic indications
- Large dataset in CHEMBL
- 10 fold cross validated PLS model
- Selected model by minimised RMSEP

Compounds	4466
Matched Pairs	288100
Fragments	678

Pharmacophore dyads	787
RMSEP	0.8
R^2	0.64
Y-scrambled R^2	0.0
ROC	0.95
Geomean odds ratio	80



Future developments

Methodology

- "Metabolophore" extraction
- Rule partition by charge
- Enhanced statistical rule selection methods
- Inferred rule extraction (A→B + B→ C = A→C) / matched series / matched networks
- Meta rule identification (eg halogen>>alkyl)
- Rule partition by shape
- Fuzzier atom typing (eg matching indole NH with ArNHC(=O)Me)

Technology

- Transform searching and clustering
- Graph database
- Distributed compute (eg Apache Spark)

} Bigger, faster, bigger & faster!

Science

- Explaining counter dogma transformations
  - Single crystal x-ray for solubility
  - Route of metabolism studies
  - Serum protein albumin co-crydtallisation for PPB
- Cardiac and liver tox screening panel development

Essentials of the collaboration

- Roche, Genentech and AZ all have ADMET data processed inside their firewalls to generate transformations (matched pairs transformations with change data)
- The transformations (fragments of molecules only) are shared with MedChemica
- MedChemica combines the transformation data and returns the aggregated knowledge
- Therefore NO party can drill back to anyone else's structures or original data
- There is no reach-through by any party
- MedChemica facilitates the science coordinating group to make suggestions on improvements and enhancements to the data sets, methods for extraction, analysis and exploitation

How Specific are Pharmacophore dyads?

- How selective is the pharmacophore?
- What are *the odds* of it hitting a molecule in the test set vs CHEMBL?

- Odds of finding in potency set =

$$\frac{n(\text{pharmacophore hits in potency set})}{n(\text{in potency set})}$$

$$\frac{27}{1470}$$

- Odds of finding in CHEMBL =

$$\frac{n(\text{pharmacophore hits in CHEMBL not in potency set})}{n(\text{in CHEMBL})}$$

$$\frac{62}{1351211}$$

- Odds ratio = selectivity =

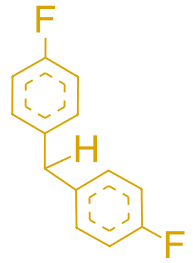
$$\frac{\textit{Odds of finding in potency set}}{\textit{Odds of finding in CHEMBL(not potency set)}}$$

$$\frac{27/1470}{62/1351211}$$

=407

(95% confidence limits: 259-642)

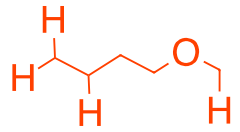
Fragment 1

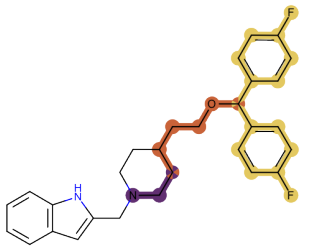


Fragment 2

$[CH_2]CN$

Path



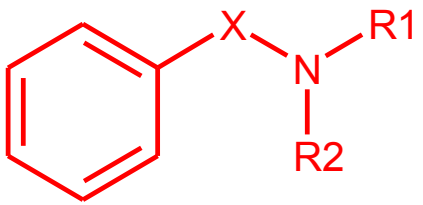


Odds of hitting a potent compound are 407 times greater than a random compound in CHEMBL

Benchmarking Specificity

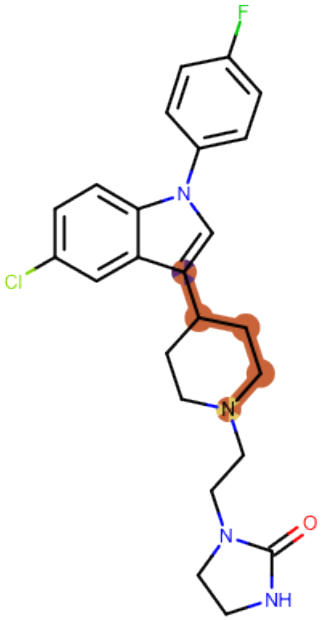
What does a *bad* odds ratio look like?

Early *simple* hERG model



Lipophilic base, usually a tertiary amine
X = 2-5 atom chain, may include rings, heteroatoms or polar groups

[$([NX3;H2,H1,H0;!$(N[C,S]=[O,N])]~*~*~*~c),$([NX3;H2,H1,H0;!$(N[C,S]=[O,N])]~*~*~c),$([NX3;H2,H1,H0;!$(N[C,S]=[O,N])]~*~*~*~*~c),$([NX3;H2,H1,H0;!$(N[C,S]=[O,N])]~*~*~*~*~*~c)]



What is the odds ratio?

Found in CHEMBL **565658/1352681**

Found in CHEMBL240 – hERG where pIC50 >=5 **1985/2451**

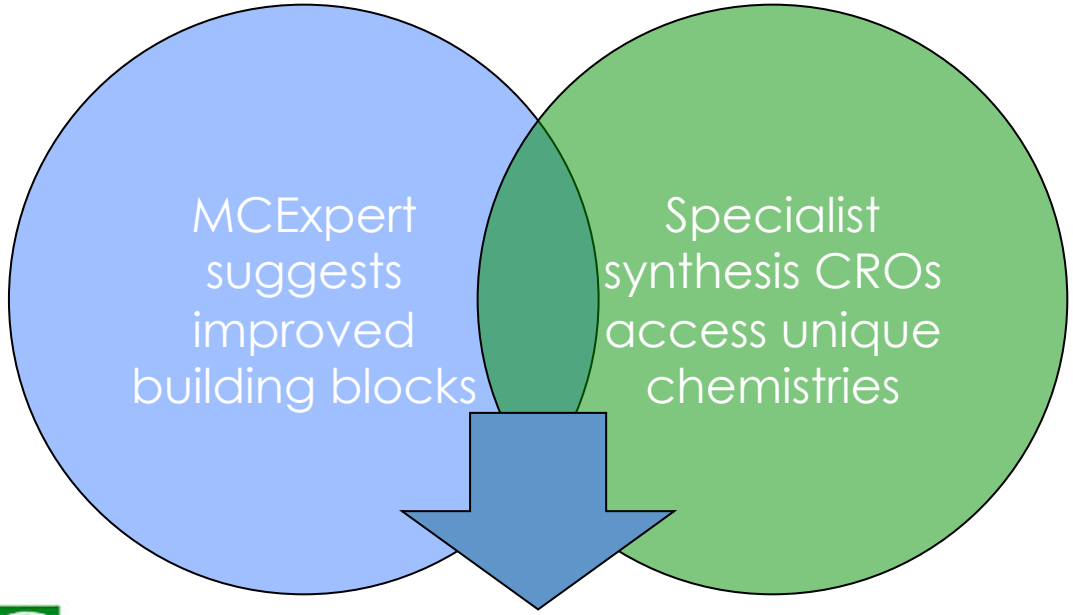
$$\text{OR} = \frac{1985/2451}{565658/1352681} = \frac{0.81}{0.42}$$

=1.94 (95% conf 1.83 – 2.05)

Ar-linker-base has only been found *1.9x* more often in hERG inhibitors than at random in ChEMBL

e. g. sertindole: 14nM vs hERG

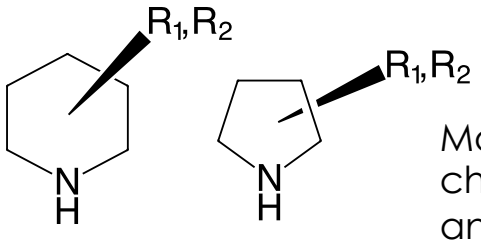
Fast building block access from CRO collaboration



LCC LIVERPOOL CHIROCHEM

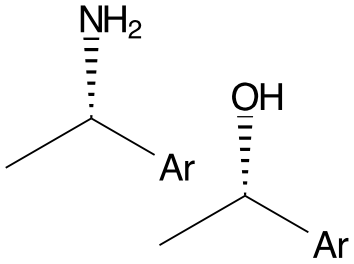
Rapid access to building blocks that address metabolism and solubility issues

yprotech



Mono & disubstituted chiral piperidines and pyrollidines

Chiral α methyl aryl amines and alcohols



BigChem Roche Genentech A Member of the Roche Group

Better compounds designed from Data