

MedChemica  
CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY

# Improved Maximum Common SubStructure(MCSS) finding coming to RDKit soon



Ed Griffen & Lauren Reid  
MedChemica  
David Cosgrove

# Why Should I care about MCSS?

- MMPA: - WizePairZ
- Automated R group finding and therefore SAR analysis, Free-Wilson, mixed modelling,
- Constrained Docking
- and undoubtedly more...



## WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry

Daniel J. Warner,<sup>\*,†</sup> Edward J. Griffen,<sup>‡</sup> and Stephen A. St-Gallay<sup>†</sup>

Department of Chemistry, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, United Kingdom, and Cancer and Infection Research, AstraZeneca R&D Alderley, Alderley Park, Macclesfield, Cheshire. SK10 4TG, United Kingdom

Received March 1, 2010

Journal of Chemical Information and Modeling 2010, 50 (8), 1350–1357.  
<https://doi.org/10.1021/ci100084s>.

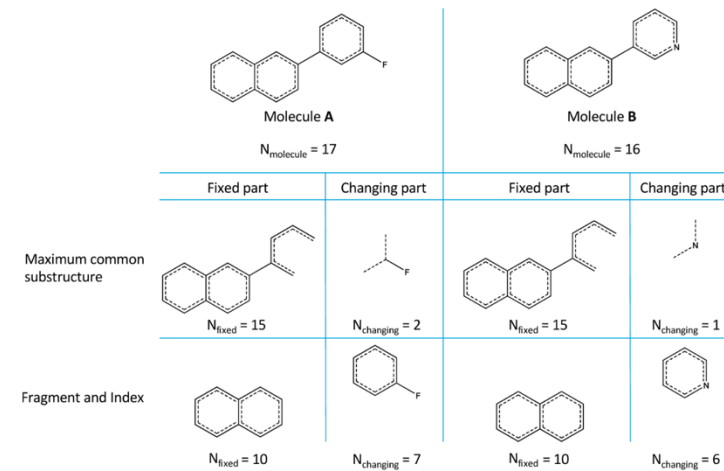


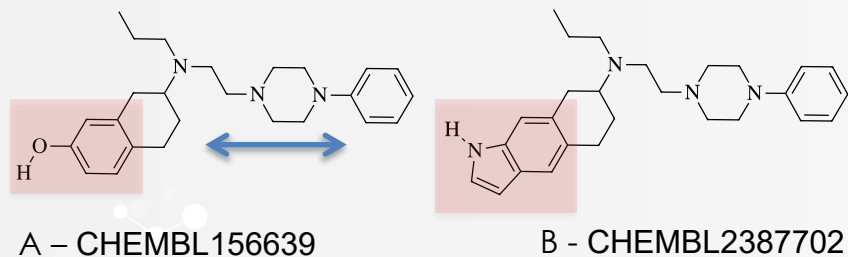
Figure 1. The pair of molecules A and B and the fixed and changing parts identified by the MCS and F+I methods.

Lukac, I.; Zarnecka, J.; Griffen, E. J.; Dossetter, A. G.; St-Gallay, S. A.; Enoch, S. J.; Madden, J. C.; Leach, A. G. Turbocharging Matched Molecular Pair Analysis: Optimizing the Identification and Analysis of Pairs. *J. Chem. Inf. Model.* 2017, 57 (10), 2424–2436.  
<https://doi.org/10.1021/acs.jcim.7b00335>.

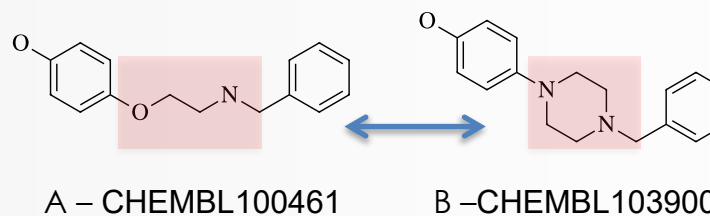
Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *Journal of Chemical Information and Modeling* **2006**, 46 (1), 180–192. <https://doi.org/10.1021/ci0503208>.

Fujita, T.; Adachi, M.; Akamatsu, M.; Asao, M.; Fukami, H.; Inoue, Y.; Iwataki, I.; Kido, M.; Koga, H.; Kobayashi, T.; Kumita, I.; Makino, K.; Oda, K.; Ogino, A.; Ohta, M.; Sakamoto, F.; Sekiya, T.; Shimizu, R.; Takayama, C.; Tada, Y.; Ueda, I.; Umeda, Y.; Yamakawa, M.; Yamaura, Y.; Yoshioka, H.; Yoshida, M.; Yoshimoto, M.; Wakabayashi, K. Background and Features of Emil, a System for Database-Aided Bioanalogous Structural Transformation of Bioactive Compounds. In *Qsar and Drug Design: New Developments and Applications*; Fujita, T., Ed.; Elsevier, 1995; Vol. 23, pp 235–273.

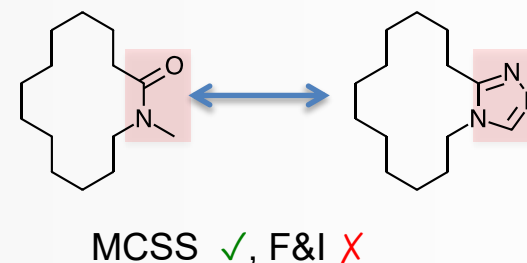
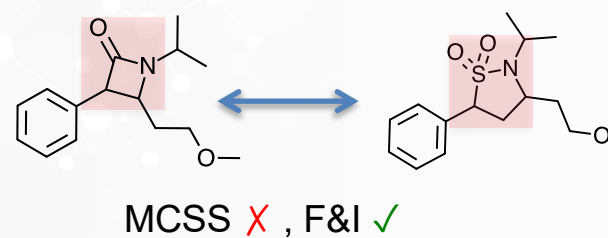
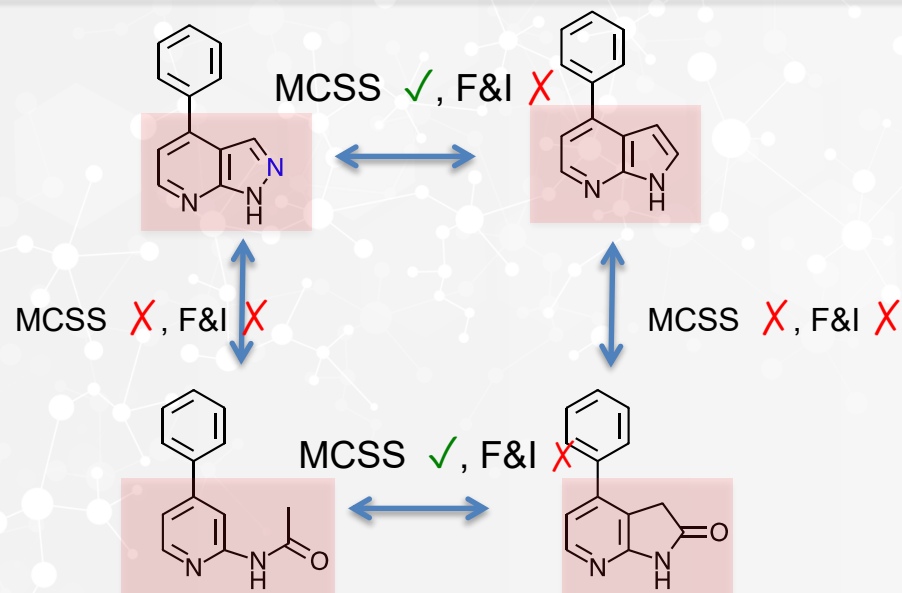
# Two Techniques for finding Matched Pairs



MCSS ✓, F&I ✗



MCSS ✗, F&I ✓



MCSS – Maximum  
Common SubStructure

F&I – Fragment and  
Index

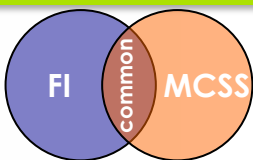
F&I:

Hussain, J.; Rea, C. Computationally  
Efficient Algorithm to Identify  
Matched Molecular Pairs (MMPs) in  
Large Data Sets. Journal of Chemical  
Information and Modeling 2010, 50  
(3), 339–348.

<https://doi.org/10.1021/ci900450m>

# Does the Matched Pair method really matter?

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



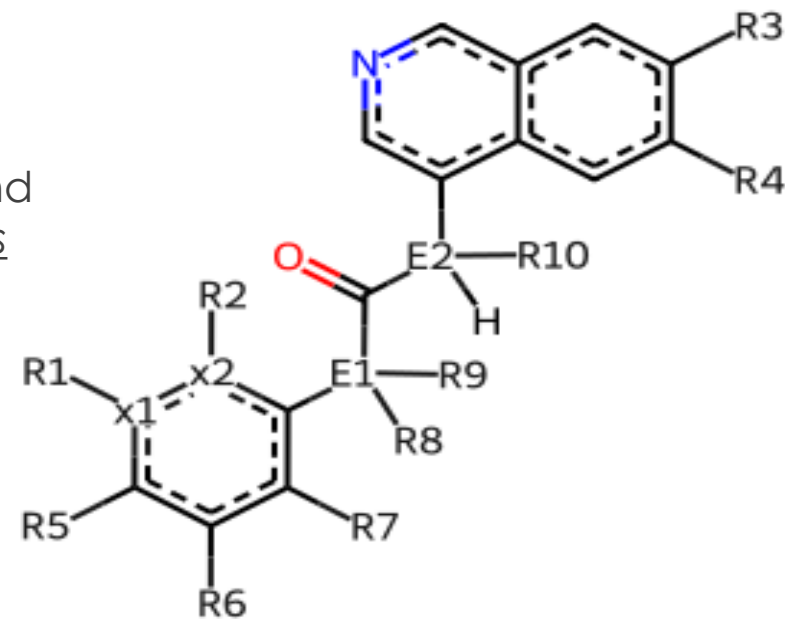
		Counts				Pairings		
	num of cmds	FI only	common	MCSS only	total	FI only %	common %	MCSS only %
VEGF	4466	17172	14631	14823	46626	37	31	32
Dopamine Transporter	1470	8930	4480	3497	16907	53	26	21
GABAA	848	1722	2500	4205	8427	20	30	50
D2 human	3873	13811	12995	13098	39904	35	33	33
D2 rat	1807	6595	5408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



- R Groups analysis – the problem
  - Consistently and rapidly assigning cores and R-groups to sets of molecules,
  - take account of symmetry issues and problematic but common medicinal chemistry core editing like  $C_{Ar}-X \rightarrow N_{Ar}$
  - Create a consistent rigorous algorithmic approach to Rand R group decomposition – without needing core definitions

- SARkush Groups:

- x – variable aromatic atom
- X – variable aliphatic ring atom
- E – variable non-ring atom
- R – variable side chain



Looking at SARS ASAP Weizmann pIC50 -> 64 compounds contained this SARkush

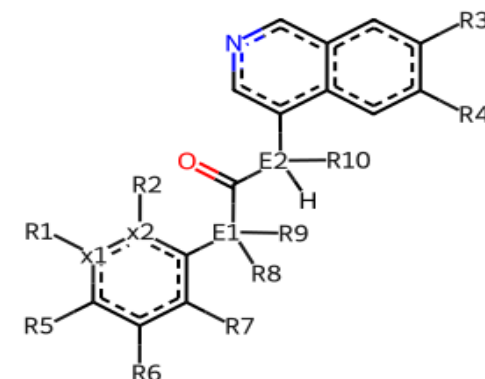
See also:

<https://www.medchemica.com/a-cheminformatics-journey-to-develop-sarkush/>

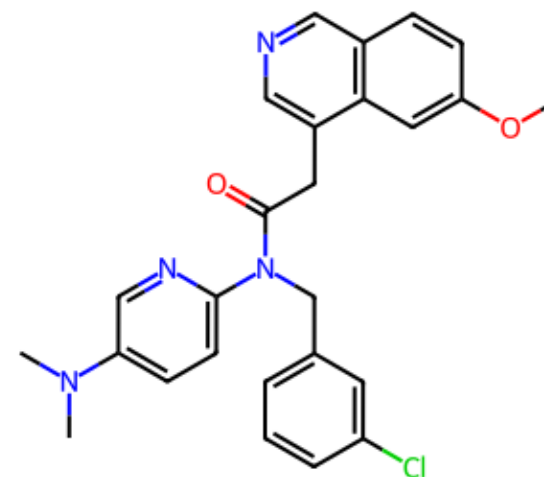
<https://www.youtube.com/watch?v=d7rrVW1mmv0>

Lauren Reid

- SARkush Groups:
  - x – variable aromatic atom
  - X – variable aliphatic ring atom
  - E – variable non-ring atom
  - R – variable side chain



SARkush Group	SMARTS	SARkush Group	SMARTS
x1	c	R4	OC
x2	n	R5	N(C)C
E1	N	R6	H
E2	C	R7	H
R1	H	R8	Cc1cc(Cl)ccc1
R2	None	R9	None
R3	H	R10	H



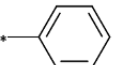
ASAP-0008451

Lauren Reid

# Free Wilson

## How much are different substituents worth?

- SAR technique that aims to assign the contribution that different groups make to a property<sup>1</sup>
- The assumption is that the contribution is additive
- A model is built using the presence (1) or absence (0) of each group
- Predictions can, therefore, be made for molecules that are from the same core
- Previously, if a group has not been seen before, the overall molecule does not generate a prediction. Currently, absent groups are given a zero coefficient.
- We are exploring how to improve this approach
  - (eg use fragment logP or mean estimate from prior datasets)
- Adapted Walter's Free-Wilson Analysis<sup>2</sup> code to accept SARkush structures

	R1			R2	
	*-H	*-Cl	*-F		
Mol1	1	0	0	1	0
Mol2	0	1	0	0	1
Mol3	0	0	0	0	1

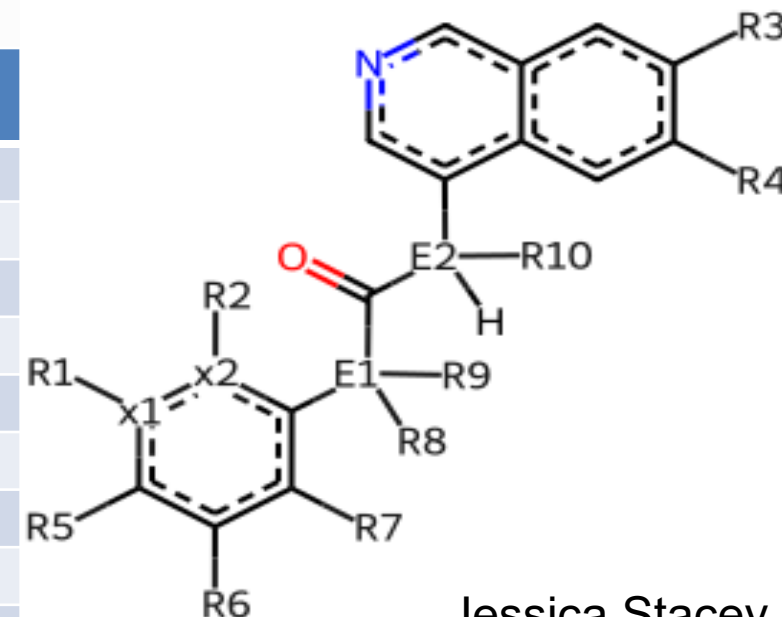
1. Free S.M., Wilson J.W. A Mathematical Contribution to Structure-Activity Studies. J Med Chem. 1964, 7(4), 395-399.

2. <https://github.com/PatWalters/Free-Wilson>



# SARkush- how much diversity at each point?

SARkush Group	Number of Examples	Pharmacophores Present	Pharmacophores Missing
E1	2	HD, None	AG, BG, HA
E2	2	HD, None	AG, BG, HA
x1	2	None	HA, HD
x2	2	None	HA, HD
R1	5	Hal, None	AG, AR, BG, HA, HD, LH, RR
R2	8	Hal, HA, HD, HA/HD, None	AG, AR, BG, LH, RR
R3	2	HA, None	AG, AR, BG, Hal, HD, LH, RR
R4	2	HA, None	AG, AR, BG, Hal, HD, LH, RR
R5	10	Hal, HA, None	AG, AR, BG, HD, LH, RR
R6	8	Hal, HA, HA/HD, None	AG, AR, BG, HD, LH, RR
R7	2	Hal, None	AG, AR, BG, HA, HD, LH, RR
R8	20	AG, AR, BG, Hal, HA, HD, HA/HD, LH, None	
R9	3	Hal, None	AG, AR, BG, HA, HD, LH, RR
R10	2	None	AG, AR, BG, Hal, HA, HD, LH, RR



Jessica Stacey

Pharmacophore labels:

AG - acidic group

AR - aromatic attachment

BG - basic group

Hal - halogen

HA - Hydrogen bond acceptor

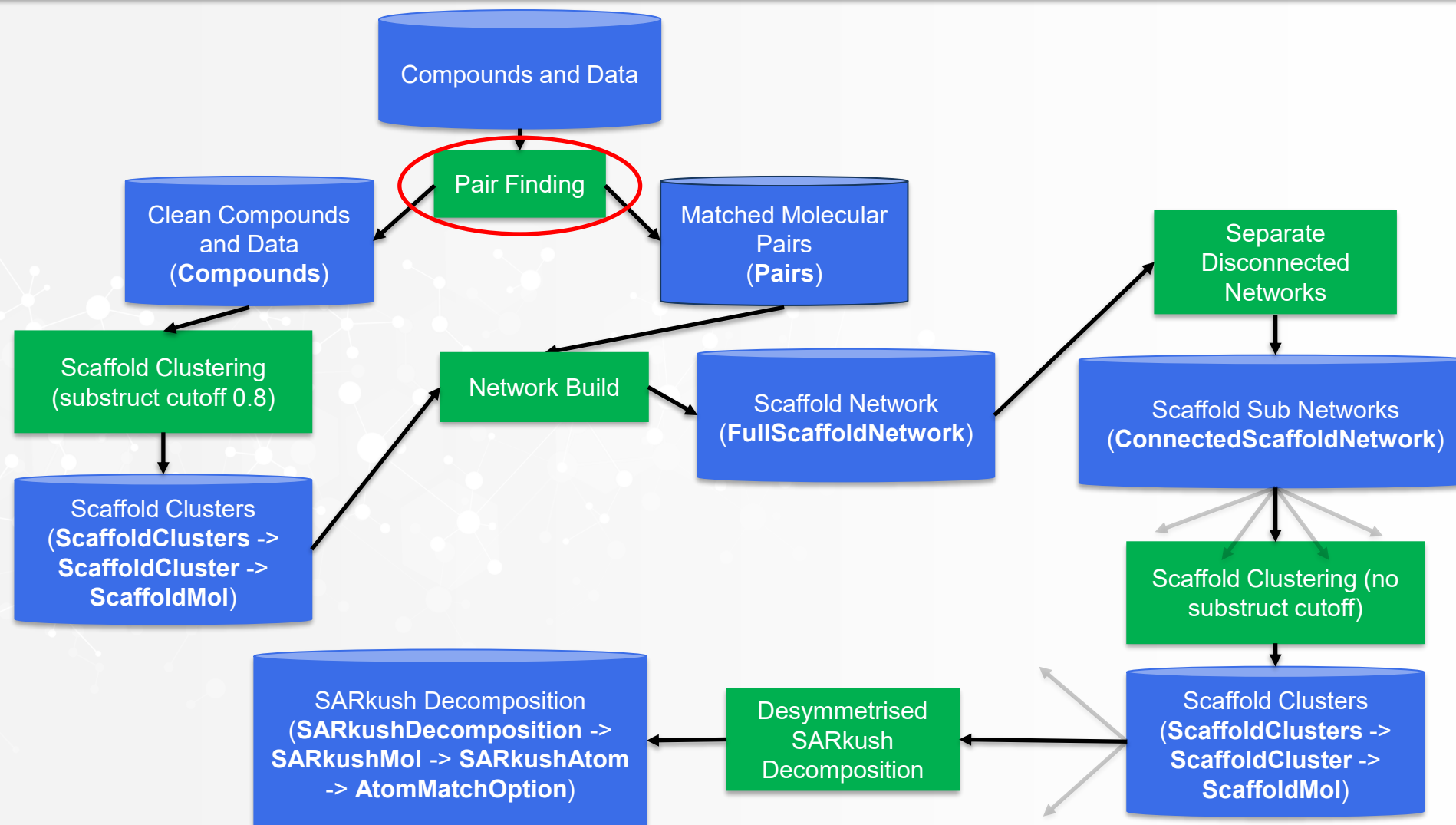
HD - hydrogen bond donor

LH - hydrophobic group

RR - aliphatic attachment

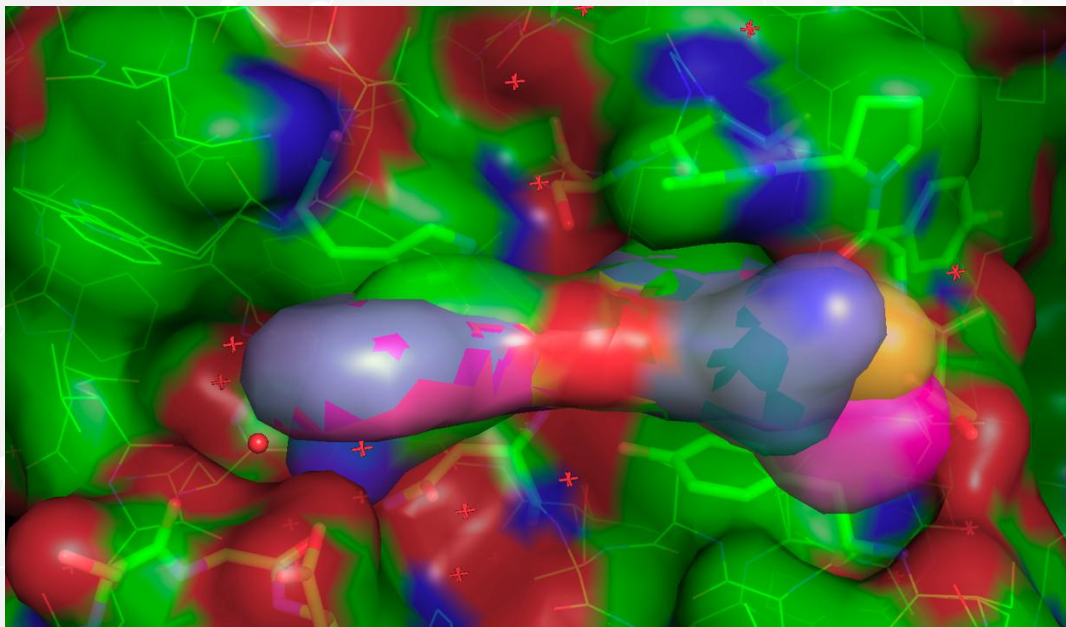
None - no pharmacophore identified

# SARkush Algorithm:



Lauren Reid

- X-ray of ligand in target protein
- generate library of new analogues with conformations
- **determine MCSS between prophetic compound and ligand in protein**
- align each conformation of each prophetic compound in the protein cavity
  - optionally discard conformations or ligands that clash with the protein surface
- Select compounds with preferred interactions



- green ligand is MCSS of original ligand with library,
- purple, orange, pink are potential targets
  - purple looks the best orange and pink clashing with surface on the right

Lauren Reid



# Why bother?, after all fMCS is good...

fMCS is specifically designed to solve the problem of finding the MCSS for sets of more than 2 molecules. And is unique in being able to do that!

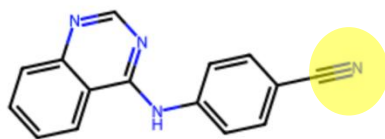
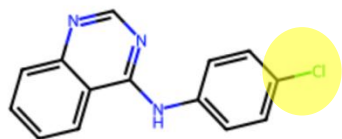
Basic algorithm:

1. Take the smallest molecule in the set and create a set of fragments (called seeds) from it by deleting bonds.
2. Match the seeds against the other molecules in the set and discard any that don't match.
3. 'Grow' the seeds by bonded atoms from the first molecule.
4. Goto 2. If there are no matches, the previous seed set comprises the MCSS for the full molecule set.

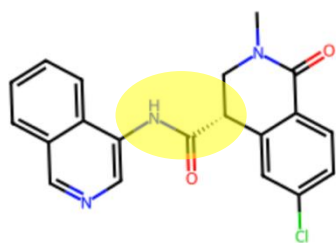
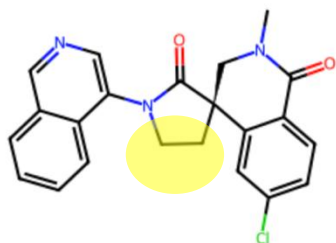
Issue for larger molecules - it does **an awful lot** of substructure matching as it builds up the seeds.

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**, 51 (4), 788–806. <https://doi.org/10.1021/ci100297y>.

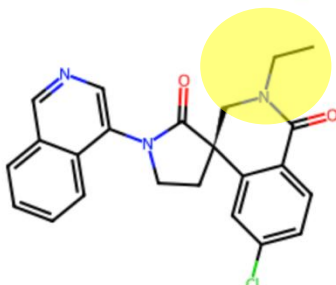
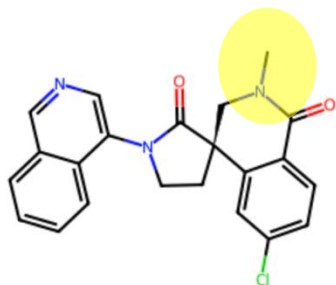
- Some standard molecules to see how fast fMCS can be:



original mcss : 0.00015s



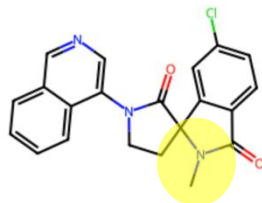
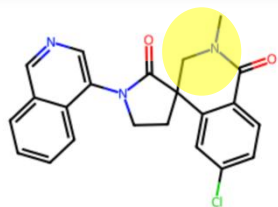
original mcss : 0.00016s



original mcss : 0.00025s

# So what's the problem?

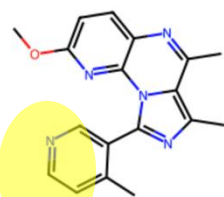
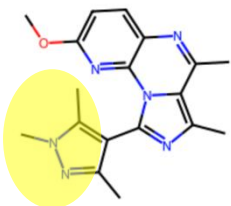
- We have a number of **real world** (not toy or deliberately engineered) difficult cases,
  - where finding the MCSS is very, **very** slow...
- we're in the business of doing  $10^{6-7} \times 10^{6-7}$  comparisons....



original mcss : 0.026s



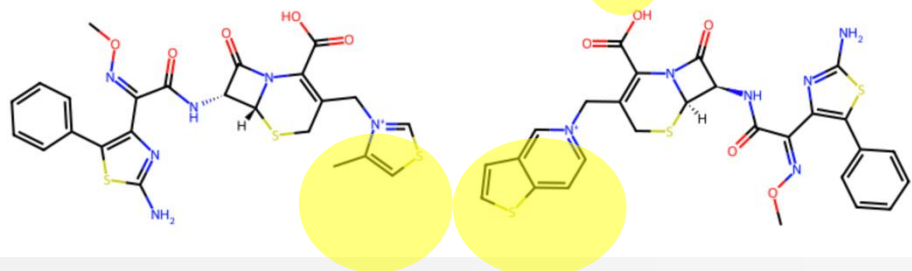
100x slower



original mcss : 0.28s



1000x slower



original mcss : 16.6375s



100000x slower



We are interested in :

1. the 2 molecule case,
  2. just connected fragments – F&I is excellent for the other cases,
  3. constraining the MCSS to be larger than a given size,
    - all organic molecules have a carbon in common
    - but even “reduction to the common benzene” may be true but is not useful...
- Do these constraints allow for some more optimisation?

# What about something different - RASCAL

<https://greglandrum.github.io/rdkit-blog/posts/2023-11-08-introducinggrascalmces.html>

<https://eprints.whiterose.ac.uk/id/eprint/3568/1/willets3.pdf>

RASCAL finds the **Maximum Common Edge Subgraph** – ie the set of bonds in common,

it's:

- *usually* very fast,
- *usually* the MCES = MCSS
- *sometimes* finds disconnected substructures

key to speed:

- runs 2 filters on the pair of molecules first to estimate the Johnson Similarity

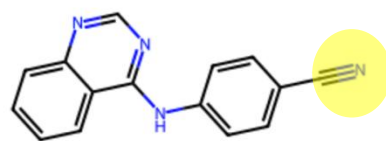
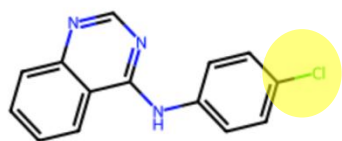
$$sim = \frac{(E(MCES) + V(MCES))^2}{(E(Mol1) + V(Mol1)) * (E(Mol2) + V(Mol2))}$$

$E(MCES)$  and  $V(MCES)$  are the number of edges (bonds) and vertices (atoms) in the MCES, and similarly for molecule 1 and molecule 2.

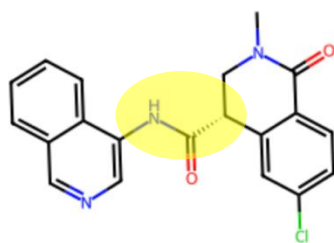
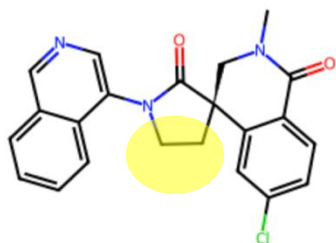
- then uses clique detection



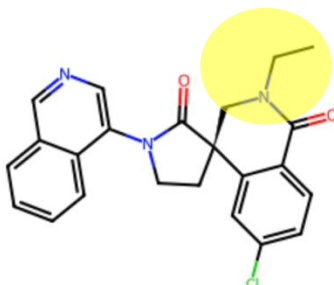
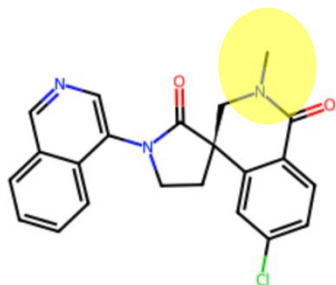
- Some standard molecules to see how fast RASCAL can be:



original mcss : 0.00015s  
RASCAL: 0.00021s



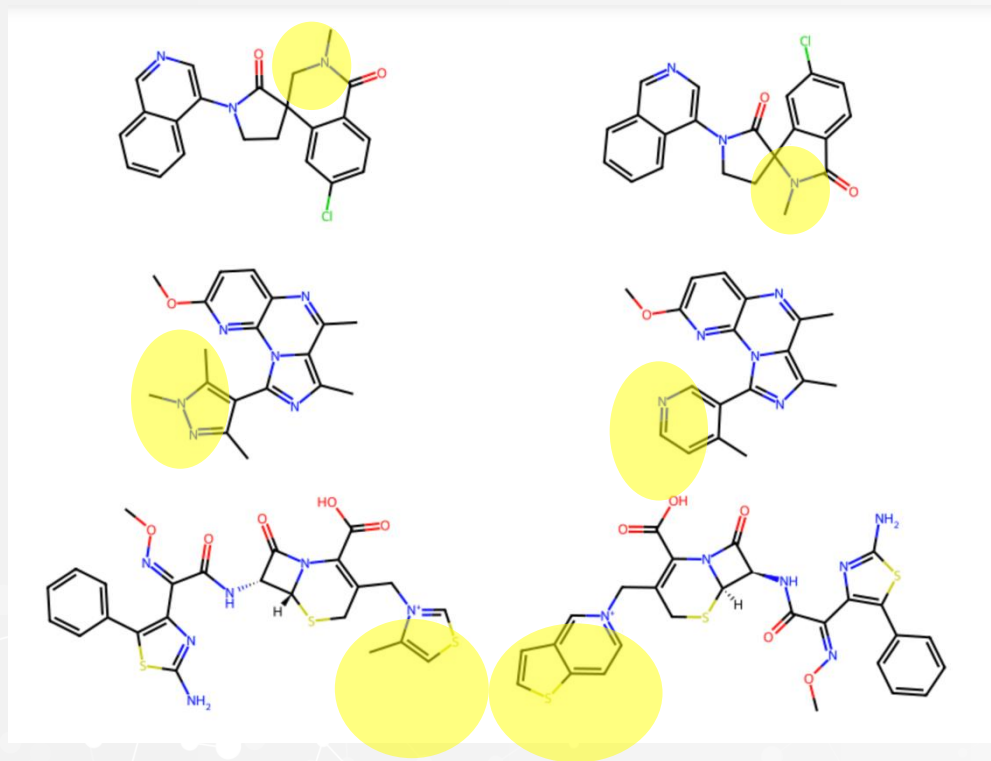
original mcss : 0.00016s  
RASCAL: 0.00072s



original mcss : 0.00025s  
RASCAL: 0.00059s



# RASCAL on Slow Pairs



original mcss : 0.026s 🤖  
RASCAL: 0.00047s

original mcss : 0.28s 🤖 🤖 🤖  
RASCAL: 0.0092s

original mcss : 16.6375s 🤖 🤖 🤖 🤖  
RASCAL: 0.0013s 🤖

RASCAL – much, much faster on slow pairs, but a bit slower on “easier” pairs – and gives a subtly different result(it’s a different algorithm with a different objective)



Insert better  
seeds here

Basic algorithm:

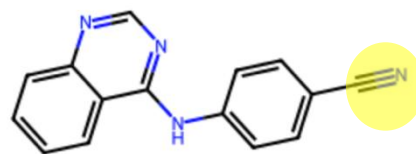
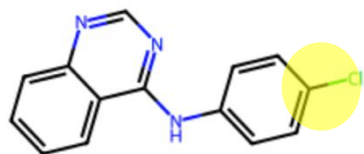
1. Take the smallest molecule in the set and create a set of fragments (called seeds) from it by deleting bonds.
2. Match the seeds against the other molecules in the set and discard any that don't match.
3. 'Grow' the seeds by bonded atoms from the first molecule.
4. Goto 2. If there are no matches, the previous seed set comprises the MCSS for the full molecule set..

1. Add an option so that for 2 molecule input sets it uses a clique detection algorithm to make a large initial seed that is then fed into the existing fMCS algorithm at point 1. More often than not, this is the final answer, but occasionally an extra atom or bond is added.
  1. Clique detection is the standard method of finding the MCS of 2 molecules. The classic algorithm for MCSS is the Bron-Kerbosch algorithm. It has the same feature as Rascal - it will return multi-fragment MCSSs. Koch has produced a modification of BK that solves that. That is what we propose:
    1. For every atom in molecule 1, make a list of matching atoms in molecule 2, and hence a 'pairs list' comprising all combinations of an atom in molecule 1 and a matching atom in molecule 2
    2. Form a correspondence graph. The nodes of this graph are the atom pairs from 1. Make an edge in the correspondence graph for all (A11, A21), (A21, A22) if A11, A21 are bonded, and A21, A22 are also bonded and the bonds match, or if A11, A21 are not bonded and A21, A22 are not bonded.
    3. A clique in the correspondence graph corresponds to an MCSS.
    4. The clique detection algorithm tries to find a clique starting with each node in the correspondence graph in turn. **There's a new parameter MinMCSSSize**. This can improve the search times quite a lot. If mol1 and mol2 both have 25 atoms, and the minimum clique size required is 20, then once one has started the clique detection from atom pairs from the first 6 atoms in molecule 1, you must already have found any cliques of 20 or more atom pairs if there is one so you can stop. If you find a larger clique earlier than that, you can stop even sooner. The start points are sorted so that atom pairs where molecule 1 atoms had the fewest number of matches to atoms in molecule 2 are tried first. This also makes things faster.

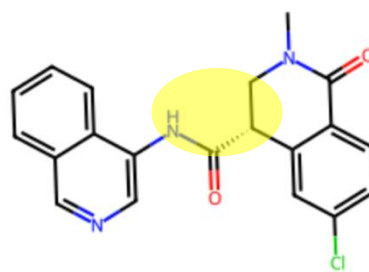
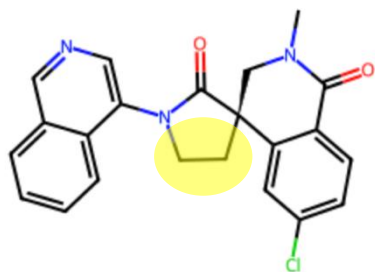
Future possibility would be to graft on the Johnson similarity estimator from Rascal for an even quicker early stop.



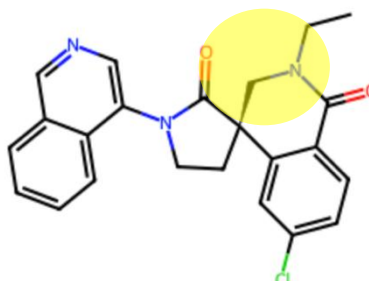
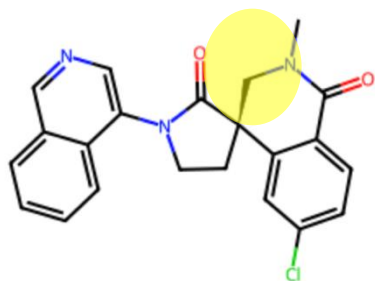
# New MCSS (fMCS with a turbo for 1 pair)



original mcss :	0.00015
RASCAL:	0.00021
new MCSS:	0.00095



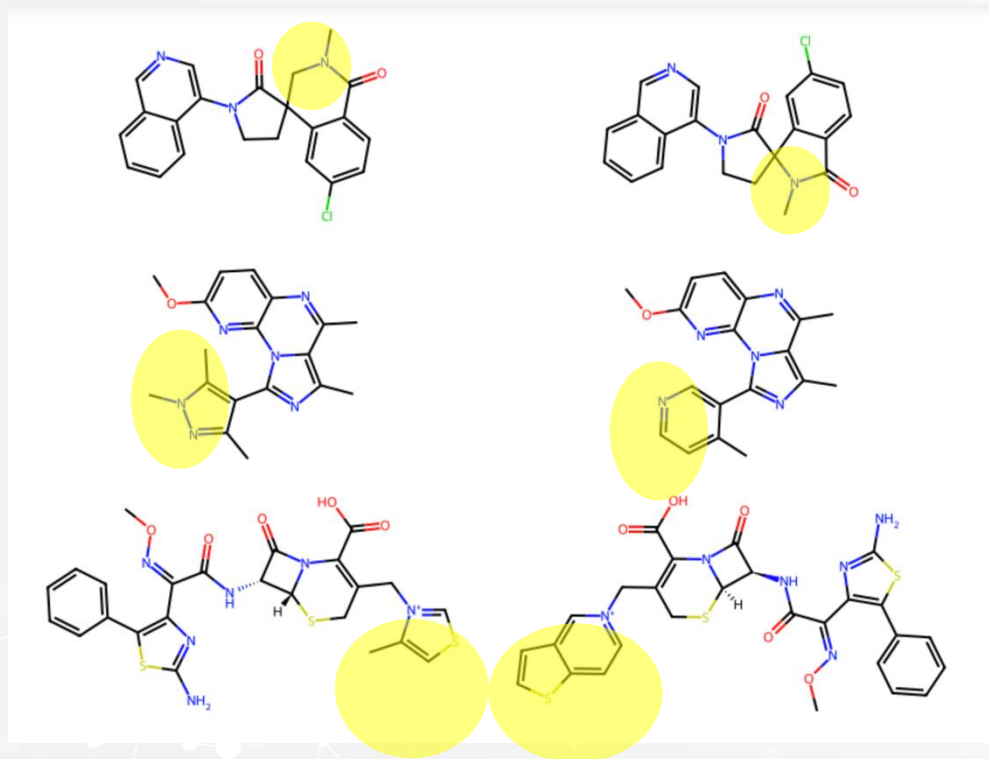
original mcss :	0.00016
RASCAL:	0.00072
new MCSS:	0.0067



original mcss :	0.00025
RASCAL:	0.00059
new MCSS:	0.015

*frustratingly not an improvement on the fast molecules....*

# New MCSS (fMCS with a turbo for 1 pair)



original mcss : 0.026  
RASCAL: 0.00047  
new MCSS: 0.0091

original mcss : 0.28  
RASCAL: 0.0091  
new MCSS: 0.0092

original mcss : 16.6375  
RASCAL: 0.0013  
new MCSS: 0.0085

New MCSS **huge improvements for the slowest molecules** –how does it do on a bulk set rather than just a limited set?

# New MCSS Scale test

1000 ChEMBL compounds – subset of those below:

1000  $n(n-1)/2$  triangle, 120s timeout for pair finding

Old time to run : 12891.71s      min= 0.000072 max=120.932067 median= 0.003598

New time to run : 3005.92s      min= 0.000056 max= 47.017234 median= 0.002352

5998 ChEMBL compounds (ChEMBL EGFR data ChEMBL203 – cleaned & tautomer standardized))

5998  $n(n-1)/2$  triangle

Old time to run : 1620336.84s      min= 0.000027 max=122.145290 median= 0.006705

New time to run : 73892.90s      min= 0.000020 max= 62.112001 median= 0.003705

At scale ~ 20x speed up – median and maximum run times both improved



We have

- a PR ready to submit for the these improvements,
- a jupyter notebook with all the examples and tests and files shown to share,

What we would additionally do – provide a blog post to document these changes,

MedChemica would like to contribute this to RDKit – any conceptual objections?

- the average user sees no difference,
  - one additional benefit is you can specify the smallest acceptable MCSS
- 
- there is the possibility for further speed up – Dave just needs to have space in his diary to work on it...