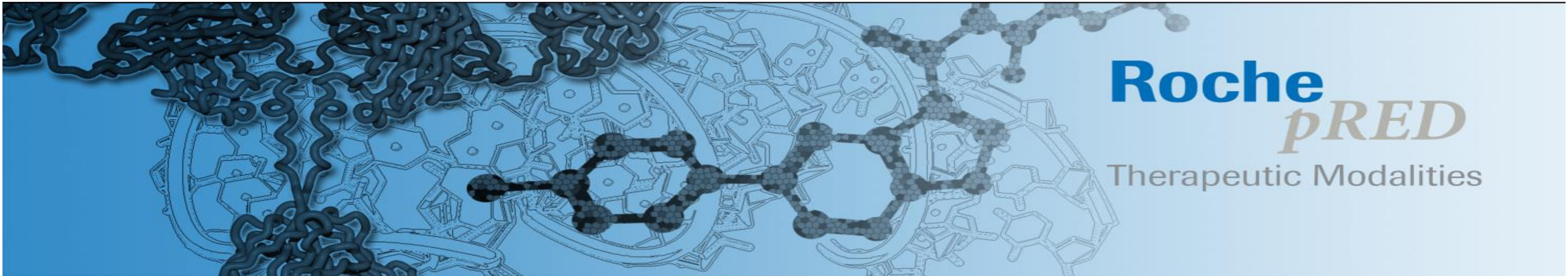
Matched Molecular Series
Measuring SAR transferability

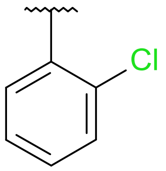
Emanuel Ehmki & Christian Kramer



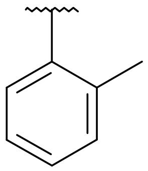
What’s next?

Target: MAP Kinase p38 alpha (ChEMBL260)

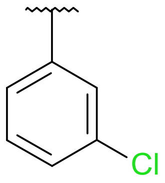
6.76



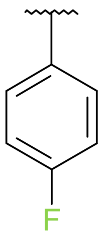
6.84



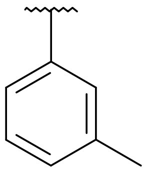
7.14



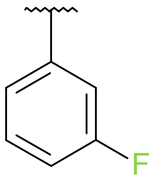
7.32



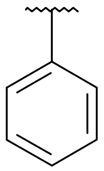
7.48



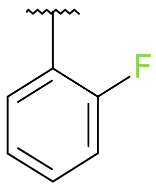
7.48

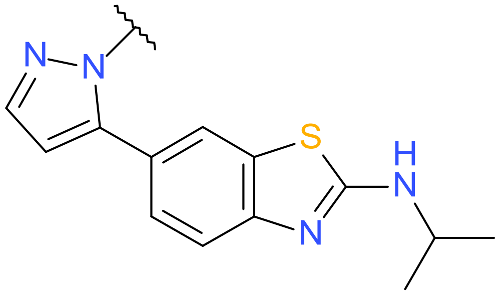


7.68



7.7

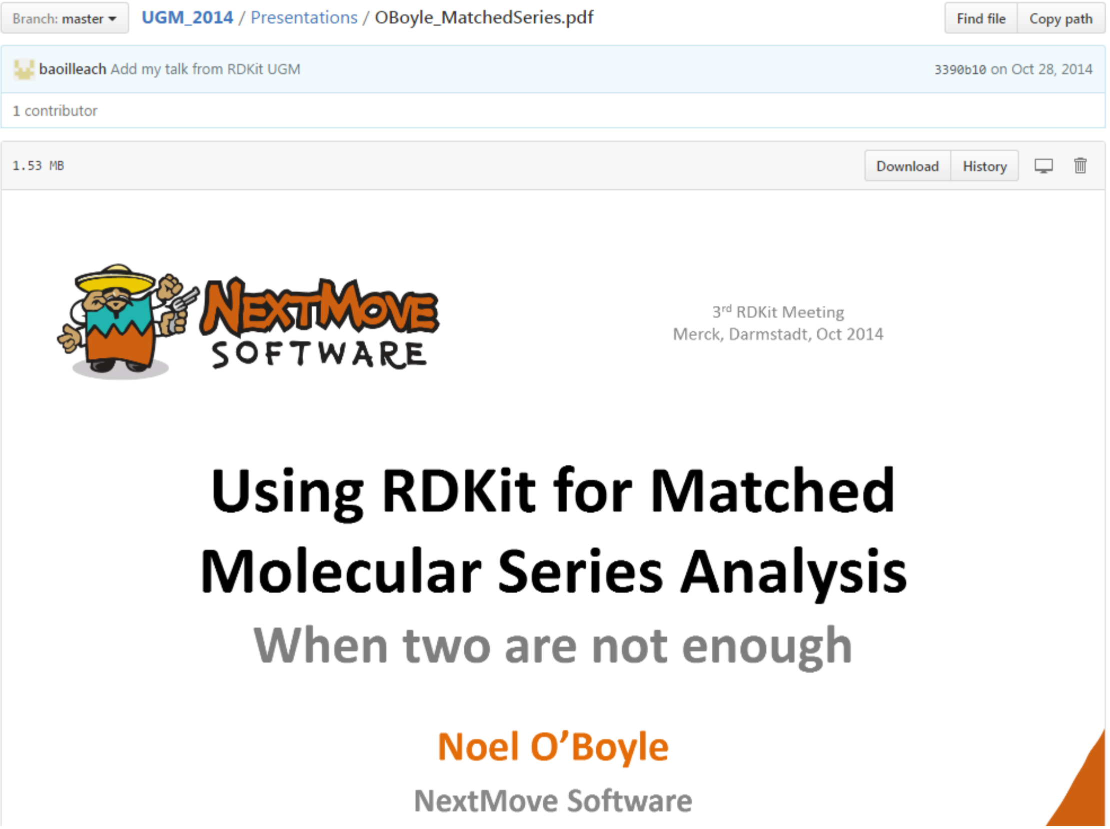




- 8 compounds with measured activity. How to decide what to try next?
  - Randomly select chemical groups (blind guessing)
  - Ask an experienced medicinal chemist
  - Model using 3D structural information
  - Try groups that worked before in similar situations

Typical MedChem Situation

- Trying things that worked before in a similar situation requires being able to measure similarity
  → What is SAR similarity?
  → How do experienced chemists strategize?
- Noel presented on that topic before at the 2014 UGM
- MedChem Intuition aka “I have seen this before”
  → Matched Series



3rd RDKit Meeting
Merck, Darmstadt, Oct 2014

Using RDKit for Matched Molecular Series Analysis

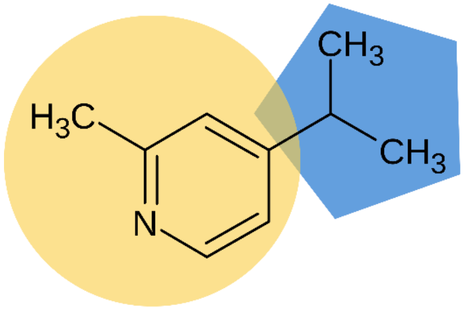
When two are not enough

Noel O’Boyle

NextMove Software

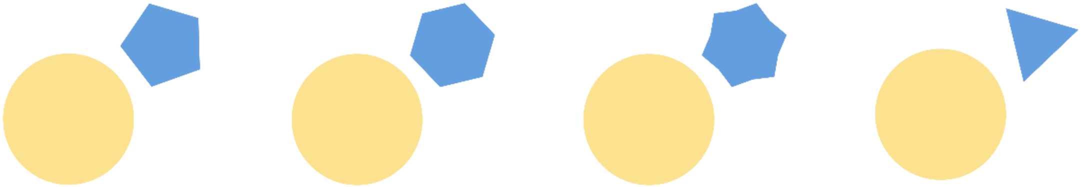
https://github.com/rdkit/UGM_2014/blob/master/Presentations/OBoyle_MatchedSeries.pdf

Matched Molecular Pairs (MMP) vs Matched Molecular Series (MMS)



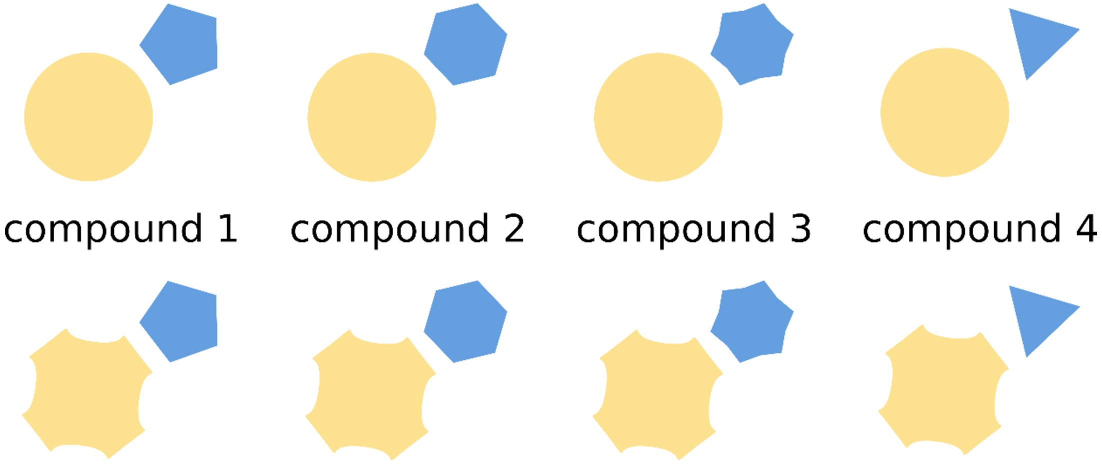
constant variable

Molecular Series



compound 1 compound 2 compound 3 compound 4

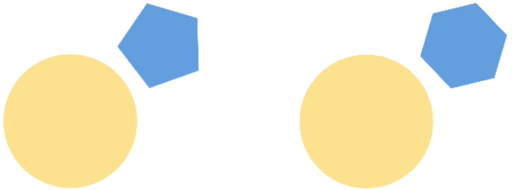
Matched Molecular Series



compound 1 compound 2 compound 3 compound 4

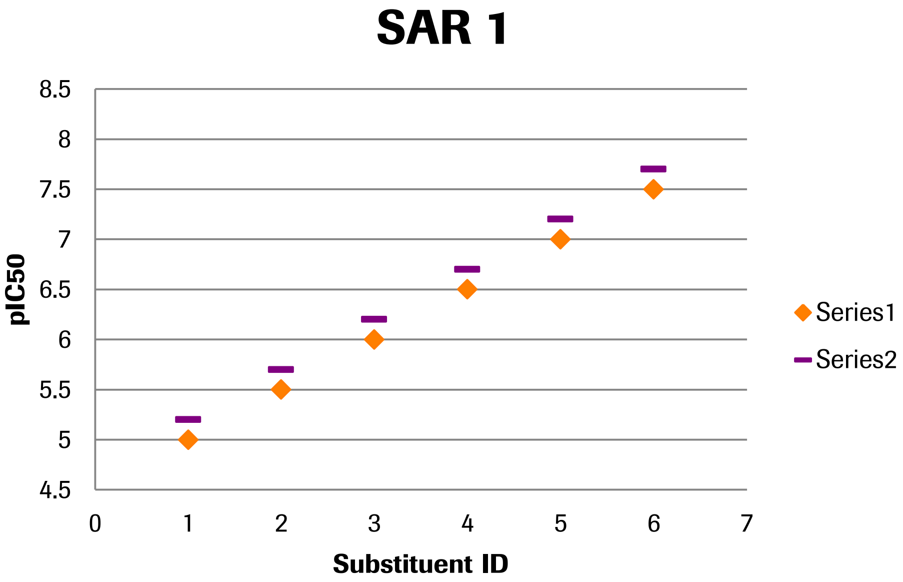
compound 5 compound 6 compound 7 compound 8

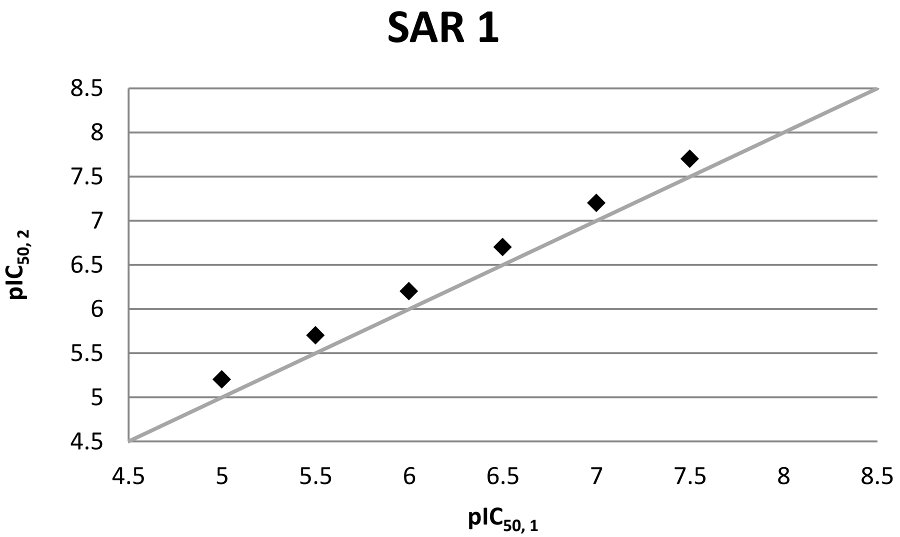
Molecular Matched Pair



compound 1 compound 2

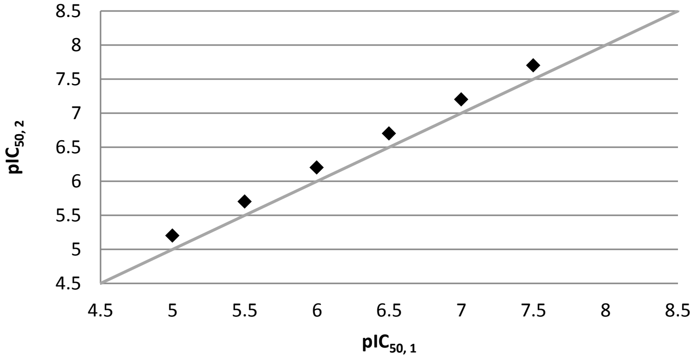
What is SAR similarity?



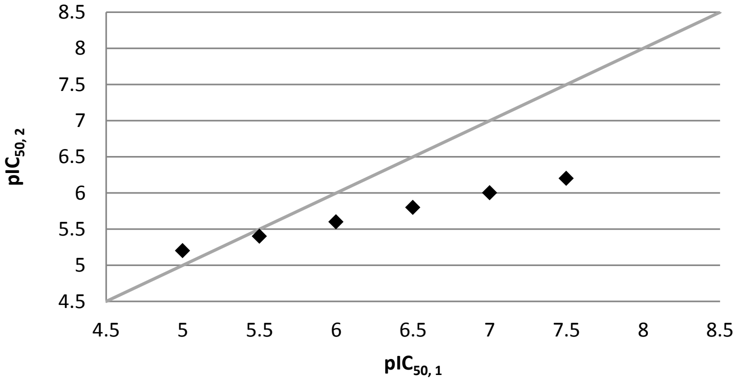


What is SAR Similarity?

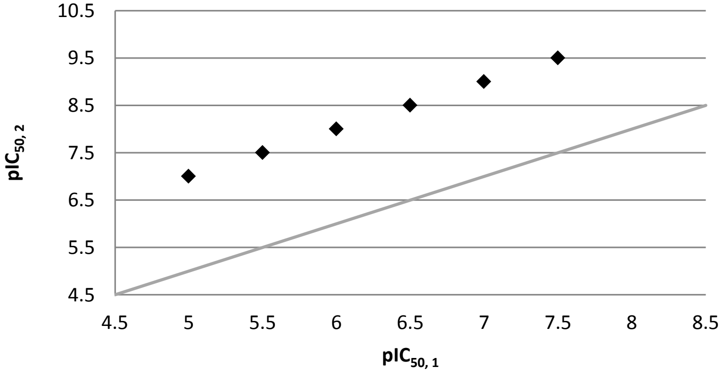
SAR 1



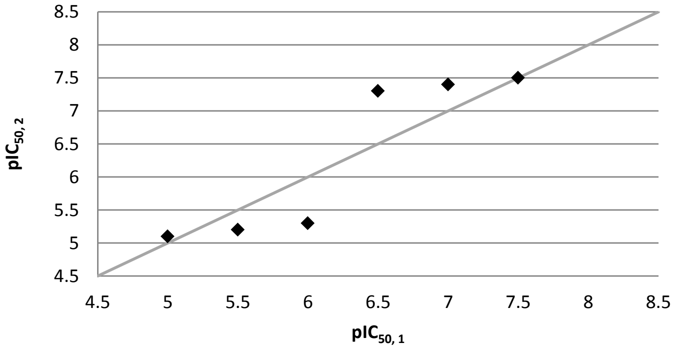
SAR 2



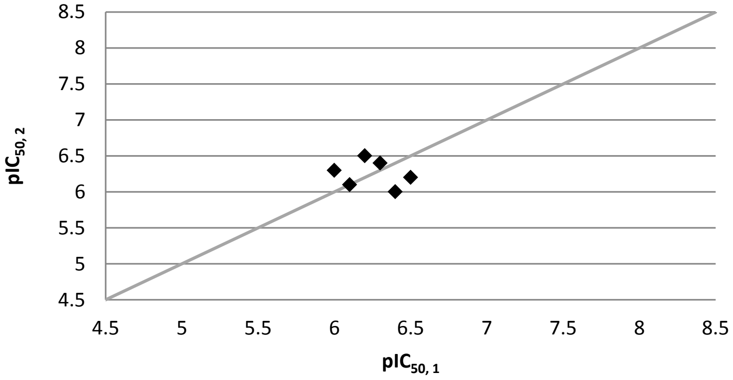
SAR 3



SAR 4



SAR 5



Comparing SAR similarity metrics - Concept

Basic Idea

A good metric is the metric that ranks similar series first.

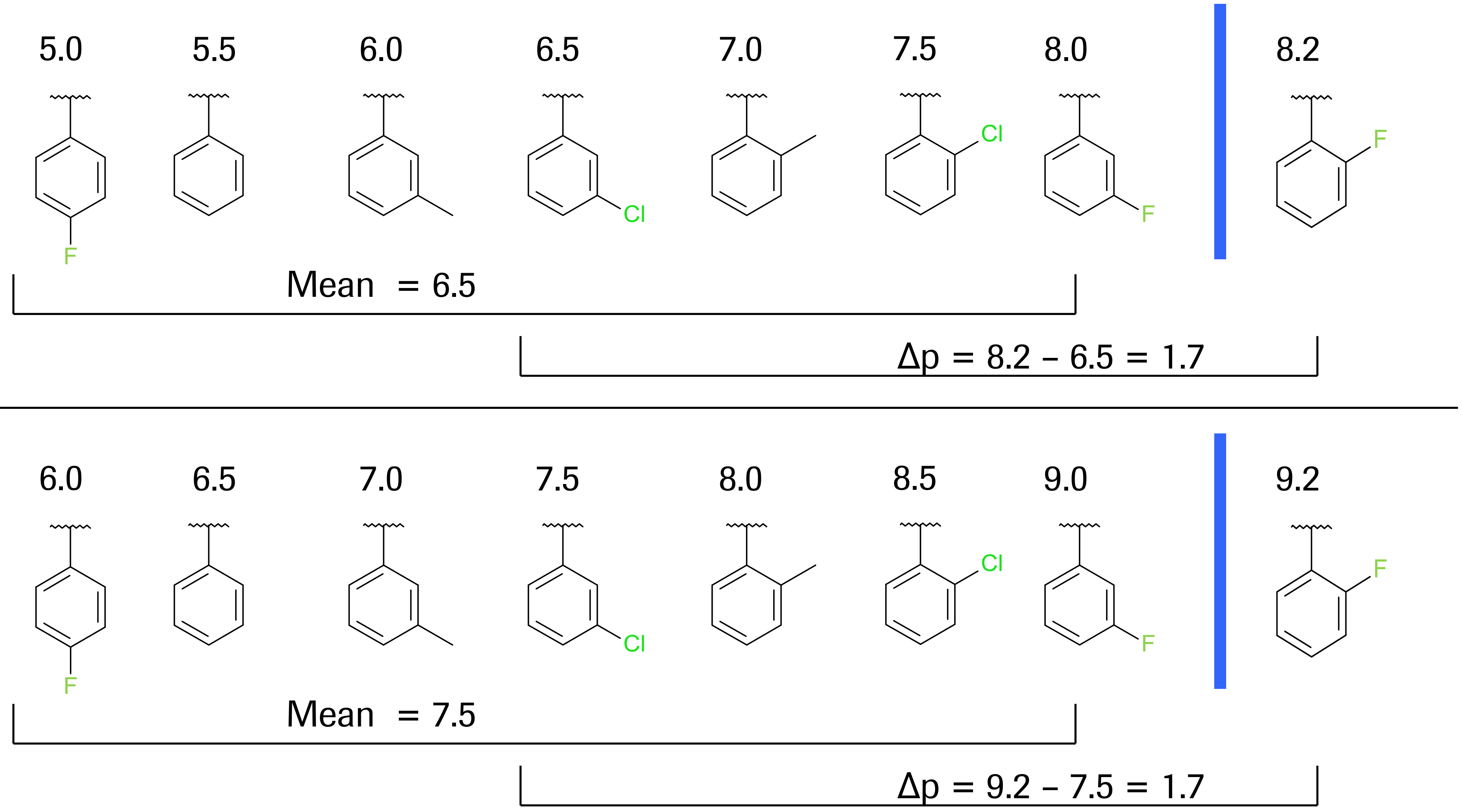
What are similar series?

Series with SAR transferability.

How to measure SAR transferability?

SAR is transferable if next fragment's activity can be predicted based on Δp = distance to series mean.

The Δp value



Query Series

5.0 5.5 6.0 6.5 7.0 7.5 8.0 | 8.2

Mean = 6.5

Δp = 8.2 – 6.5 = 1.7

Reference Series

6.0 6.5 7.0 7.5 8.0 8.5 9.0 | 9.2

Mean = 7.5

Δp = 9.2 – 7.5 = 1.7

ΔΔp = 1.7 – 1.7 = 0.0 → SAR is transferable

* artificial example

When is SAR parallel?

How to measure activity profile similarity?

- Two series with common fragments
- Both series within a similar range of potency
- Both series display a similar potency progression
- Similarity Metrics
  - Pearson correlation
  - Spearman correlation
  - Euclidian distance
  - Manhattan distance (RMSD)
  - centered RMSD (cRMSD)

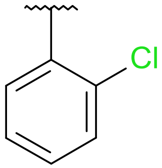
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i \ - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^k \right)^{1/k}$$

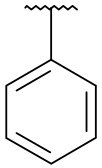
$$cRMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((x_i - \bar{x}) - (y_i - \bar{y})\right)^2}$$

DB example query

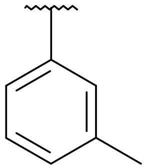
5.46



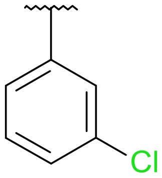
5.47



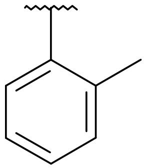
5.31



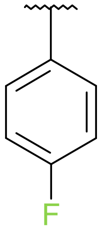
5.45



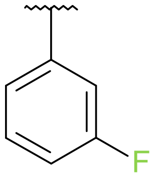
5.32



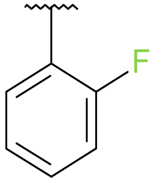
5.34



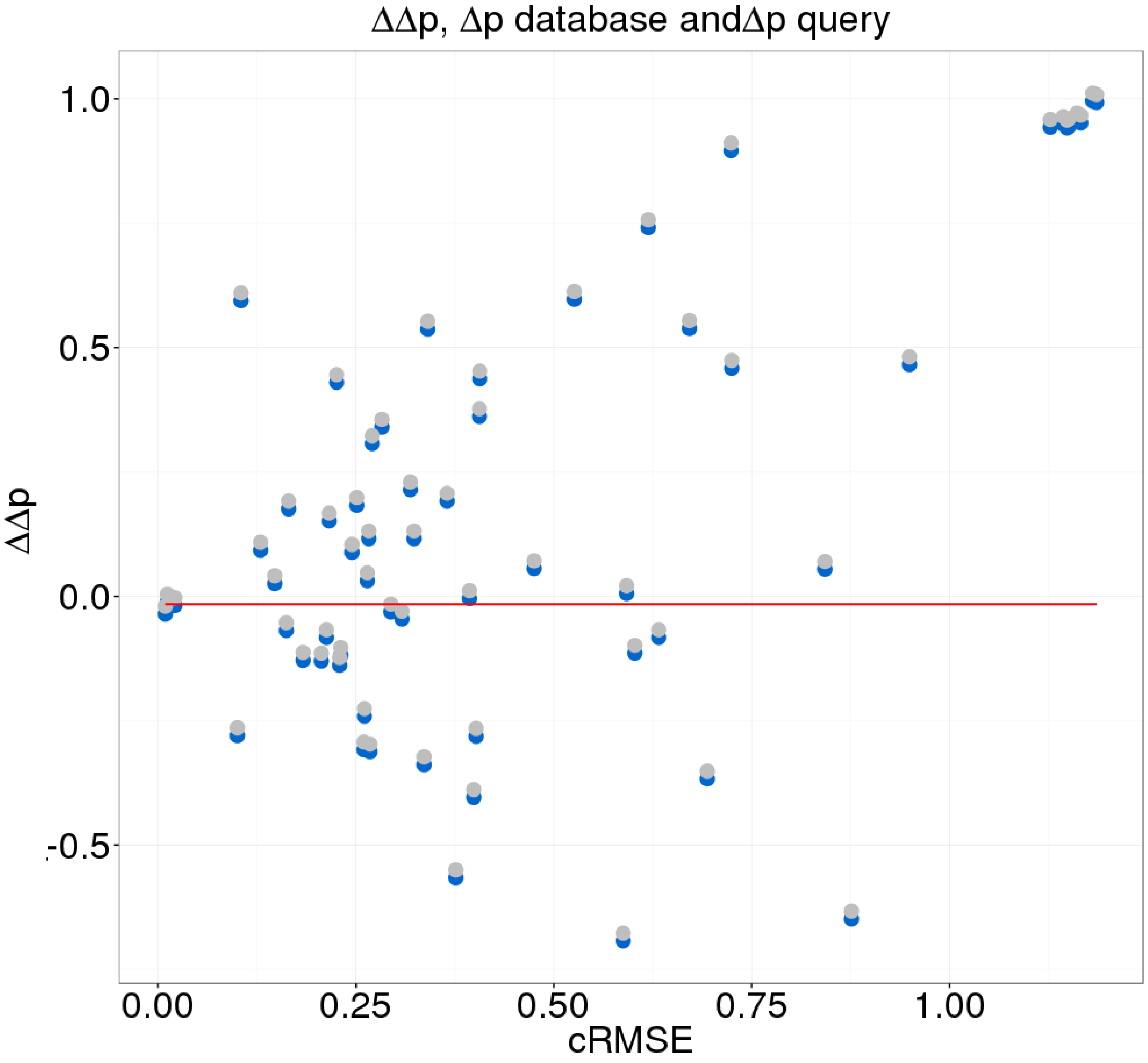
5.35



5.37

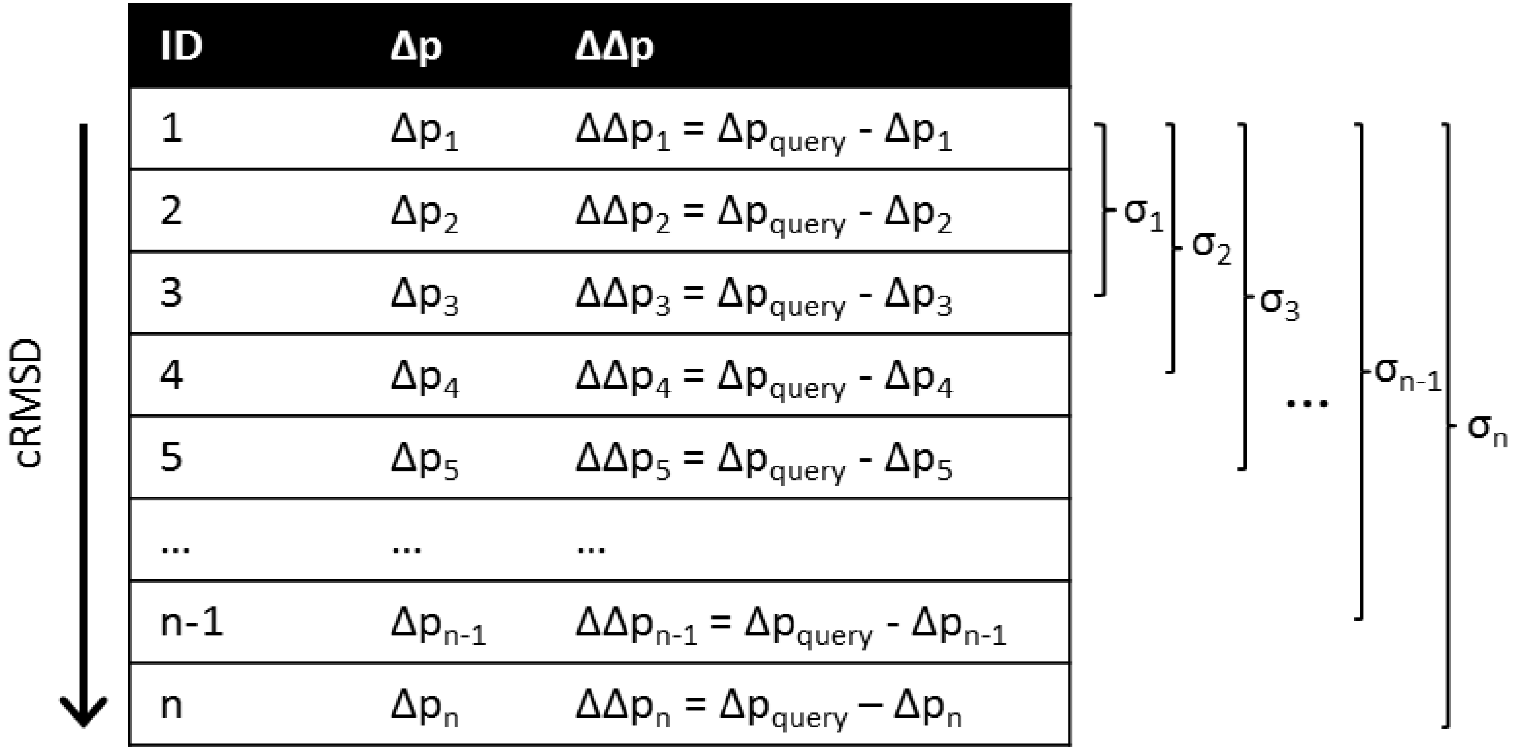


DB example query



How to measure SAR transferrability?

ΔΔp vs Similarity is too scattered → Expanding sigma method

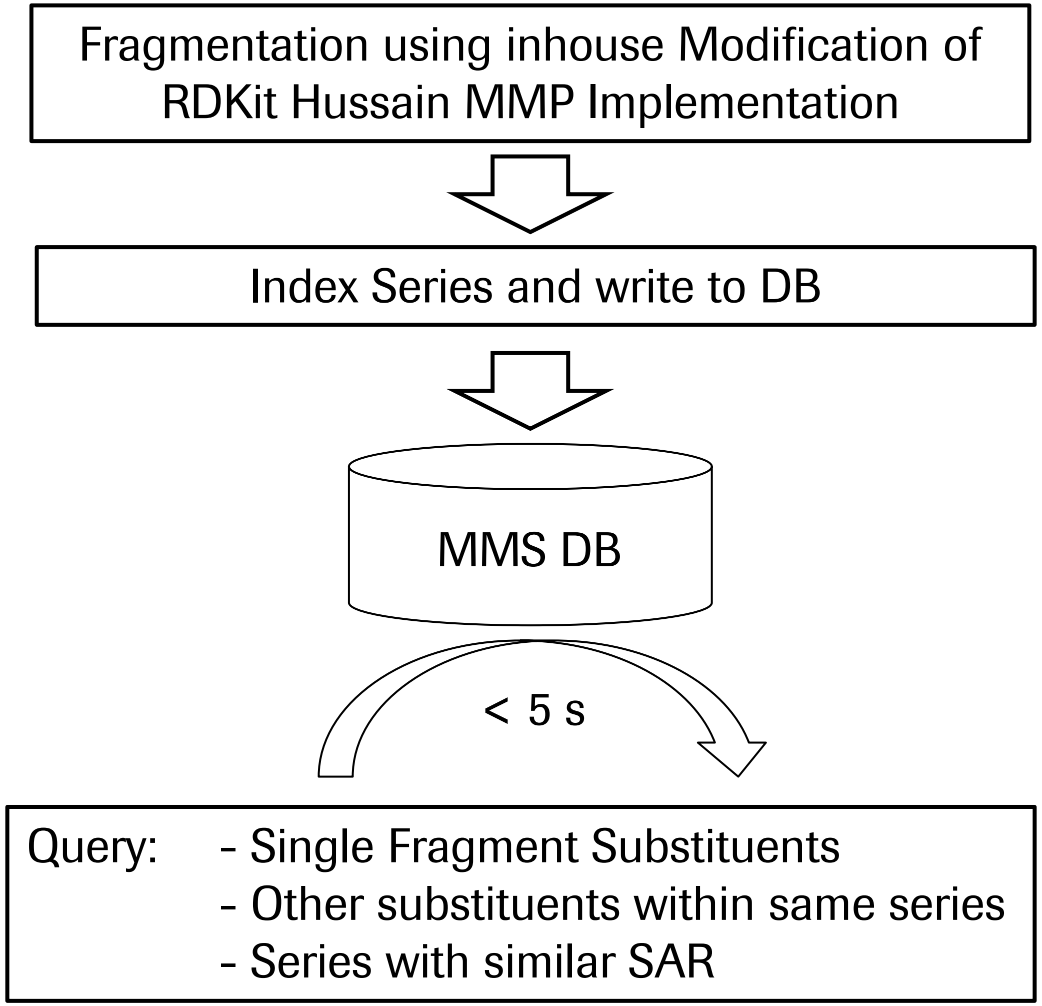


ID	Δp	ΔΔp
1	Δp_1	$\Delta\Delta p_1 = \Delta p_{query} - \Delta p_1$
2	Δp_2	$\Delta\Delta p_2 = \Delta p_{query} - \Delta p_2$
3	Δp_3	$\Delta\Delta p_3 = \Delta p_{query} - \Delta p_3$
4	Δp_4	$\Delta\Delta p_4 = \Delta p_{query} - \Delta p_4$
5	Δp_5	$\Delta\Delta p_5 = \Delta p_{query} - \Delta p_5$
…	…	…
n-1	Δp_{n-1}	$\Delta\Delta p_{n-1} = \Delta p_{query} - \Delta p_{n-1}$
n	Δp_n	$\Delta\Delta p_n = \Delta p_{query} – \Delta p_n$

MMS Implementation

Implementation using:

- RDKit
- Pandas
- MySQL
- ChEMBL21
  & inhouse DB



Fragmentation using inhouse Modification of RDKit Hussain MMP Implementation

Index Series and write to DB

MMS DB

< 5 s

Query: - Single Fragment Substituents
- Other substituents within same series
- Series with similar SAR

DB Schema

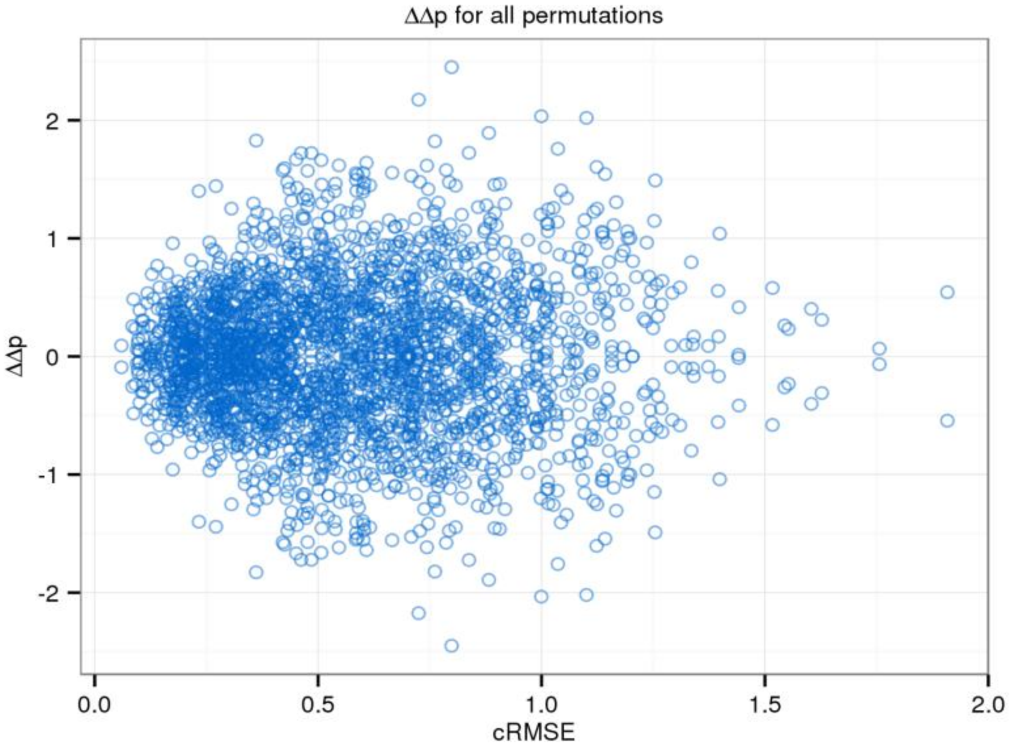
Fragments

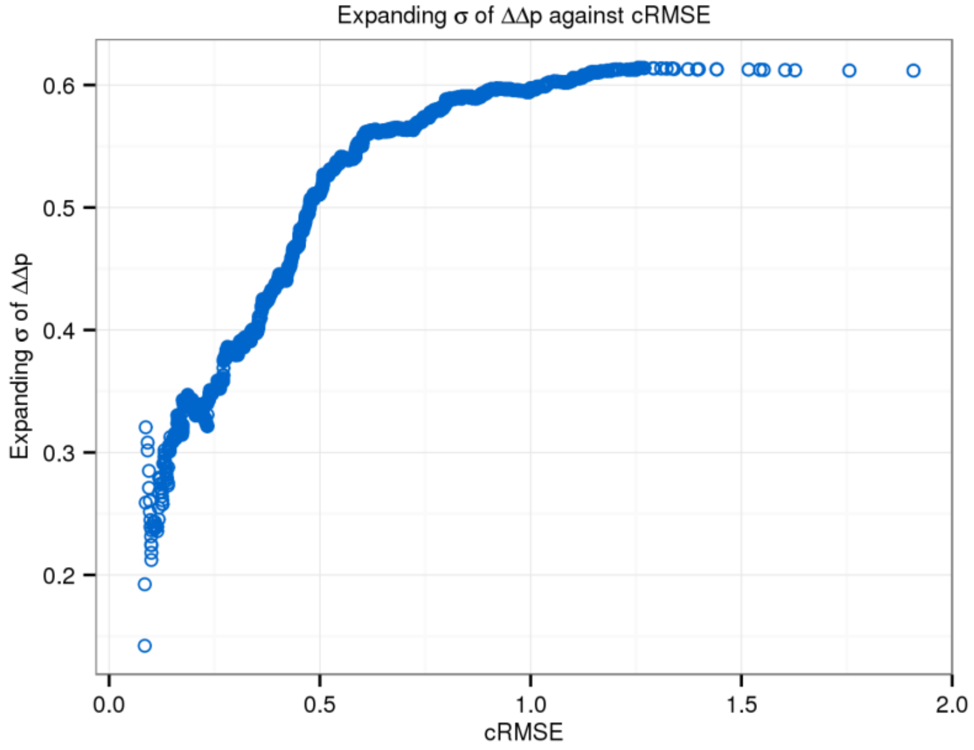
SMILES	ID	Series ID
…		
…		
…		

Series

Series ID	SMILES	Values	Other annotation
…			
…			
…			

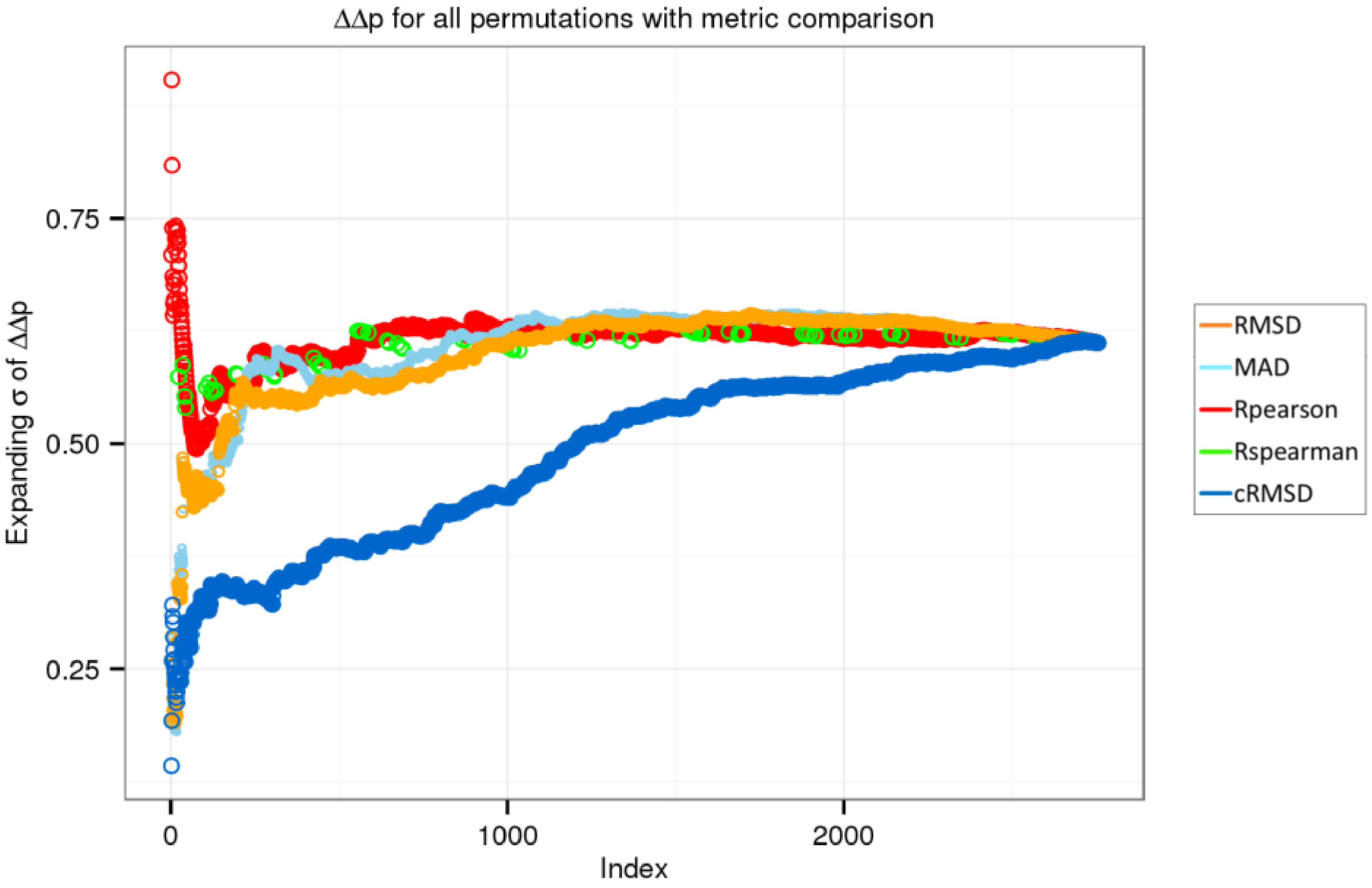
Results with single query series





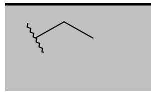
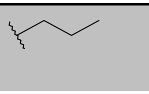
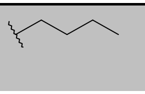
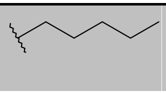
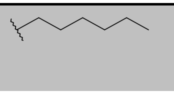
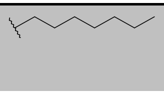
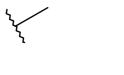
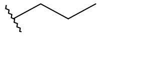
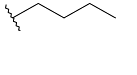
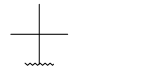
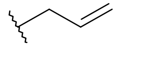
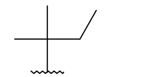
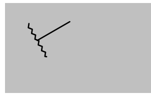
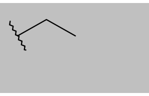
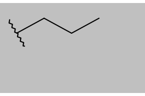
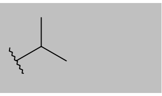
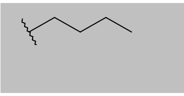
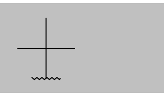
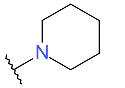
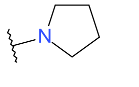
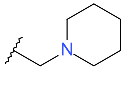
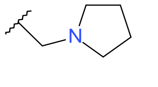
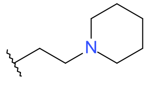
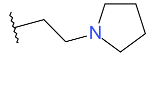
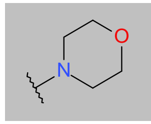
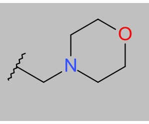
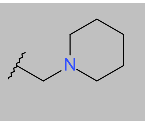
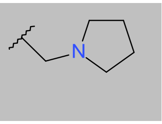
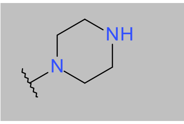
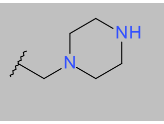
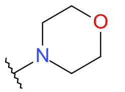
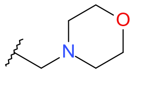
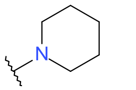
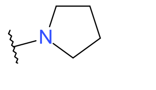
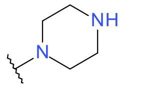
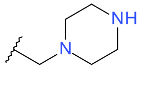
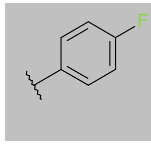
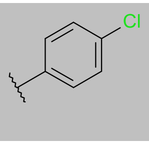
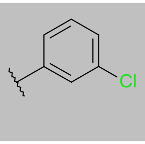
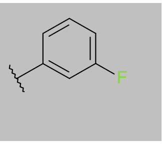
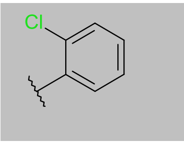
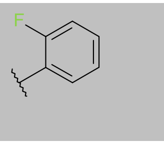
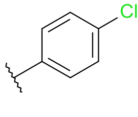
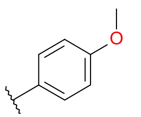
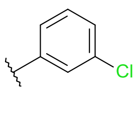
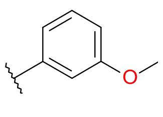
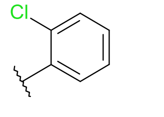
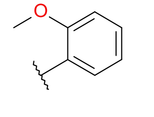
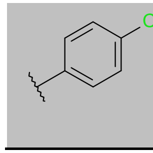
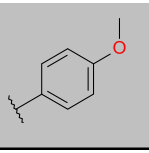
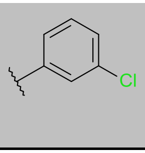
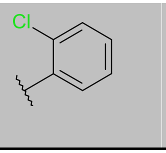
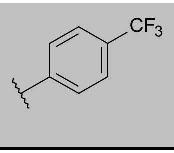
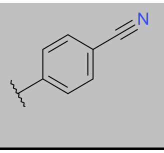
Query Series: All DB series with [*]-H, [*]-C, [*]-Cl, [*]-CN, [*]-NO_2 substituents, predicting [*]-OC (53 series)

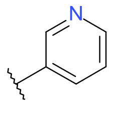
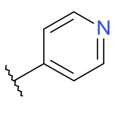
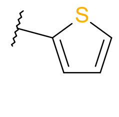
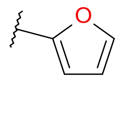
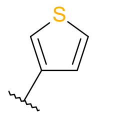
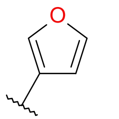
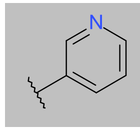
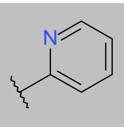
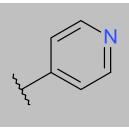
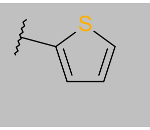
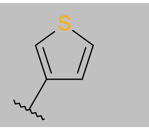
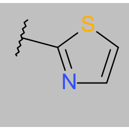
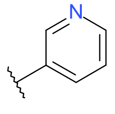
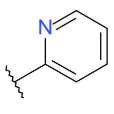
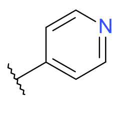
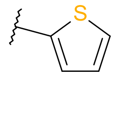
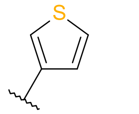
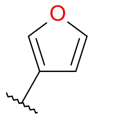
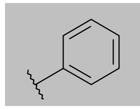
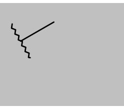
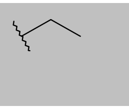
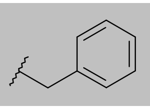
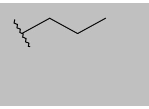
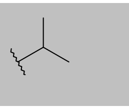
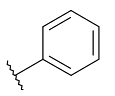
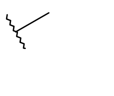
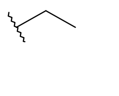
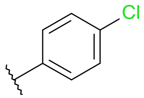
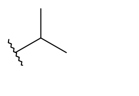
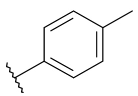
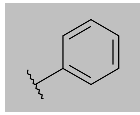
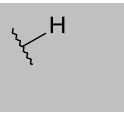
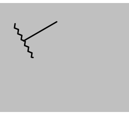
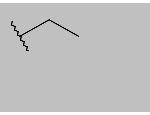
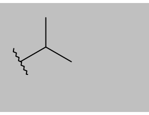
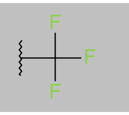
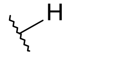
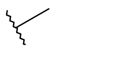
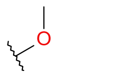
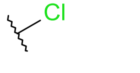
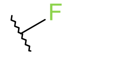
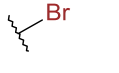
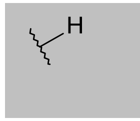
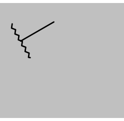
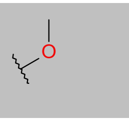
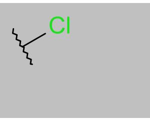
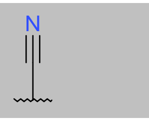
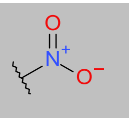
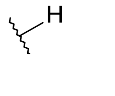
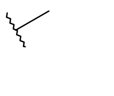
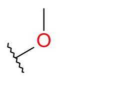
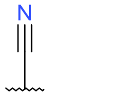
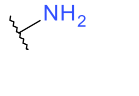
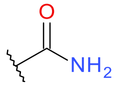
Comparing different similarity metrics on single query series



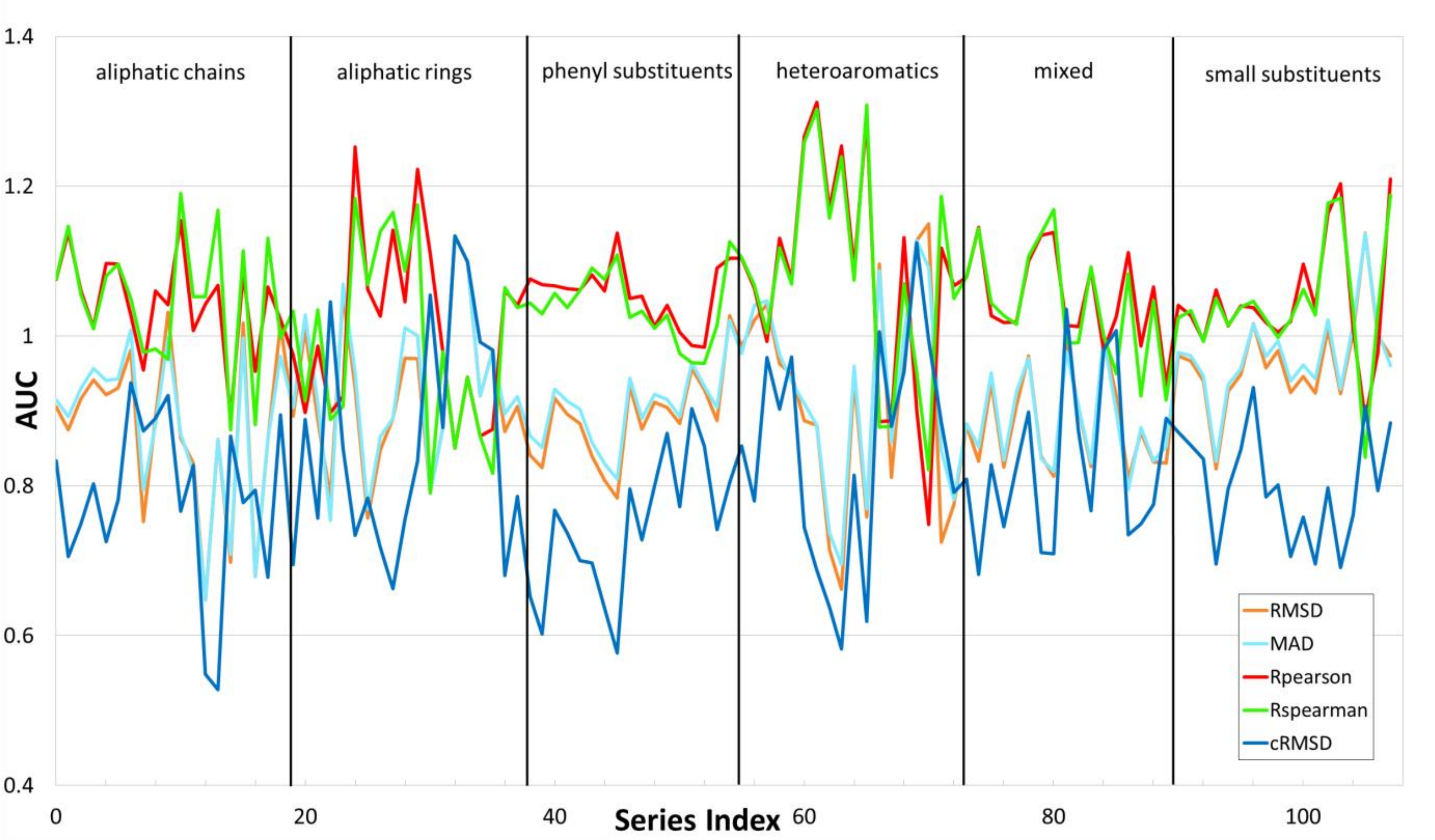
Query Series: All DB series with [*]-H, [*]-C, [*]-Cl, [*]-CN, [*]-NO_2 substituents, predicting [*]-OC (53 series)

Single query is not representative - selection of 108 test series

Substituent 1	Substituent 2	Substituent 3	Substituent 4	Substituent 5	Substituent 6	DB Count
						113
						11
						52
						8
						8
						8
						132
						122
						41

Substituent 1	Substituent 2	Substituent 3	Substituent 4	Substituent 5	Substituent 6	DB Count
						32
						10
						10
						32
						17
						14
						235
						53
						11

Results: Test on 108 query series

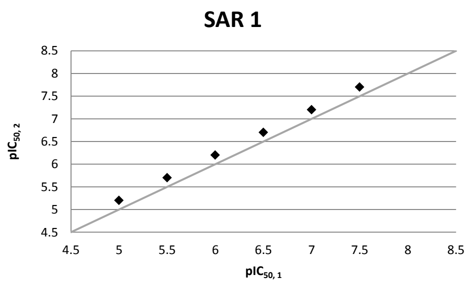


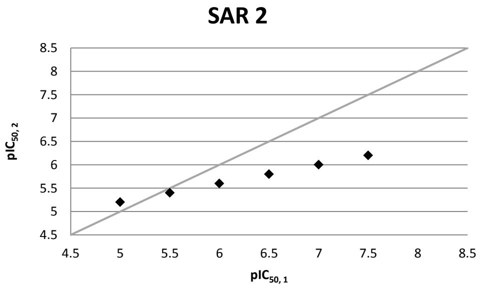
→ cRMSD performs best on 81 out of 108 test series (p-value 1.34e-34)

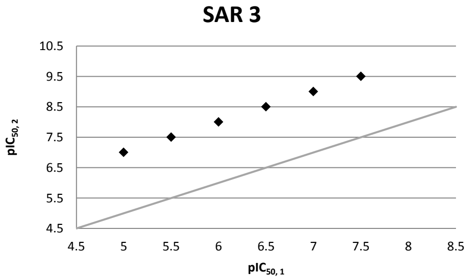
Why does cRMSD work best?

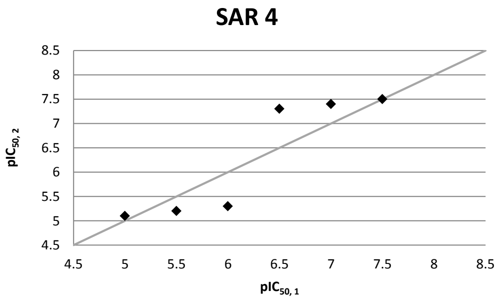
$$cRMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((x_i-\bar{x})-(y_i-\bar{y})\right)^2}$$

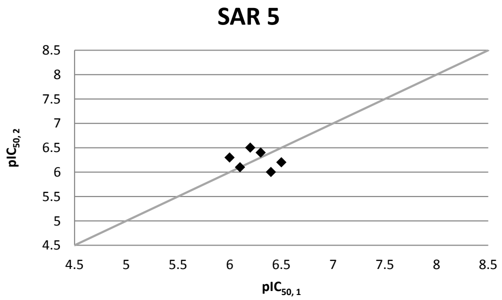
- cRMSD centers values – removes effect of remaining scaffold
- cRMSD evaluates absolute differences, assuming additivity
- Other metrics do not represent activity cliffs/ shifts, scrambling due to experimental uncertainty







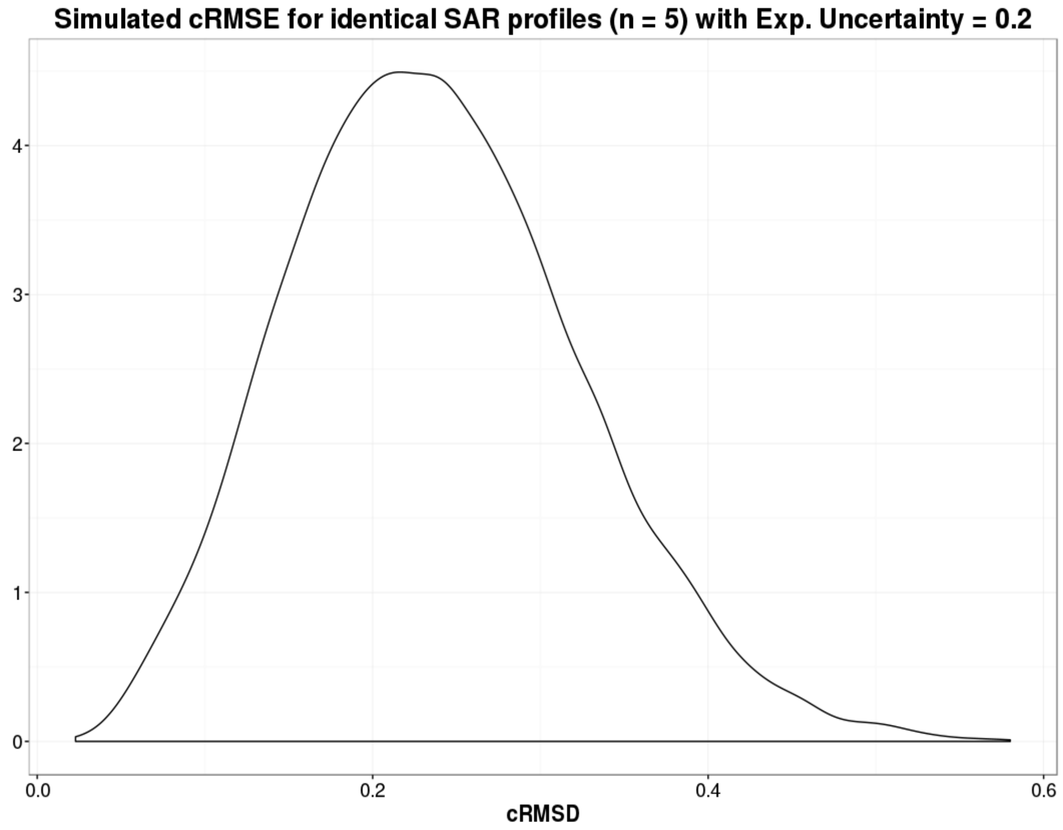




In Practice: How to set the cRMSD threshold?

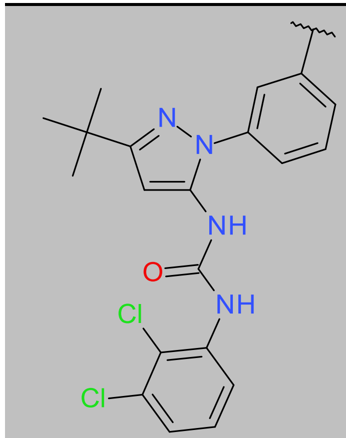
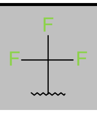
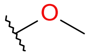
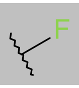
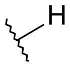
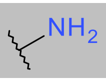
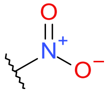
- Data has experimental uncertainty
- Uncertainty model:
  isotropic Normal Distribution, σ= 0.2 log units
- Simulated lower uncertainty threshold:
  cRMSD ~0.25

→ Below cRMSD = 0.25, cRMSD is likely entirely explained by experimental uncertainty



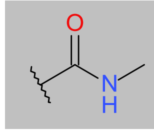
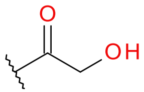
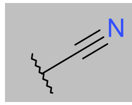
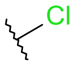
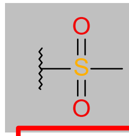
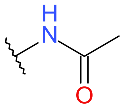
Simulating Matched Series use in real life application

Query Series: p38 Kinase inhibitors

Core	Substituent	IC_{50} [nM]	pIC_{50} [logM]
		56	7.25
		53	7.46
		33	7.48
		30	7.52
		13	7.89
		11	7.96

Dumas, J.; Hatoum-Mokdad, H.; Sibley, R.; Riedl, B.; Scott, W. J.; Monahan, M. K.; Lowinger, T. B.; Brennan, C.; Natero, R.; Turner, T.; Johnson, J. S.; Schoenleber, R.; Bhargava, A.; Wilhelm, S. M.; Housley, T. J.; Ranges, G. E.; Shrikhande, A. 1-Phenyl-5-Pyrazolyl Ureas: Potent and Selective p38 Kinase Inhibitors. *Bioorg. Med. Chem. Lett.* **2000**, *10* (18), 2051–2054.

Suggested Fragments

Substituent	Frequency	cRMSD	$\Delta\Delta p$
	2	0.33	1.06
	1	0.33	1.0
	14	0.33	0.74
	27	0.23	0.15
	9	0.29	-0.64
	1	0.58	-3.07

This compound has been made and it is inactive (> 500 nM)

Summary & Outlook

- **cRMSD clearly outperforms** all other metrics tested for predicting SAR similarity.
- MMS implementation based on RDKit, Python, Pandas, MySQL, & ChEMBL21 is possible with **query times ~5 s.**
- SAR Similarity threshold for cRMSD below 0.25 does not make sense due to **experimental uncertainty**.

- MMS analysis could be used as a **lightweight baseline model for MM-GBSA/FEP** approaches.
- MMS can be the starting point to **decipher MedChem Intuition**.

Doing now what patients need next