



Optibrium Consultants' Day  
Cambridge, Nov 2014

# Beyond matched pairs

## Using matched series for activity prediction

**Noel O'Boyle**

NextMove Software

Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity *J. Med. Chem.* 2014, 57, 2704.



# HOW TO CHOOSE WHAT COMPOUND TO MAKE NEXT?

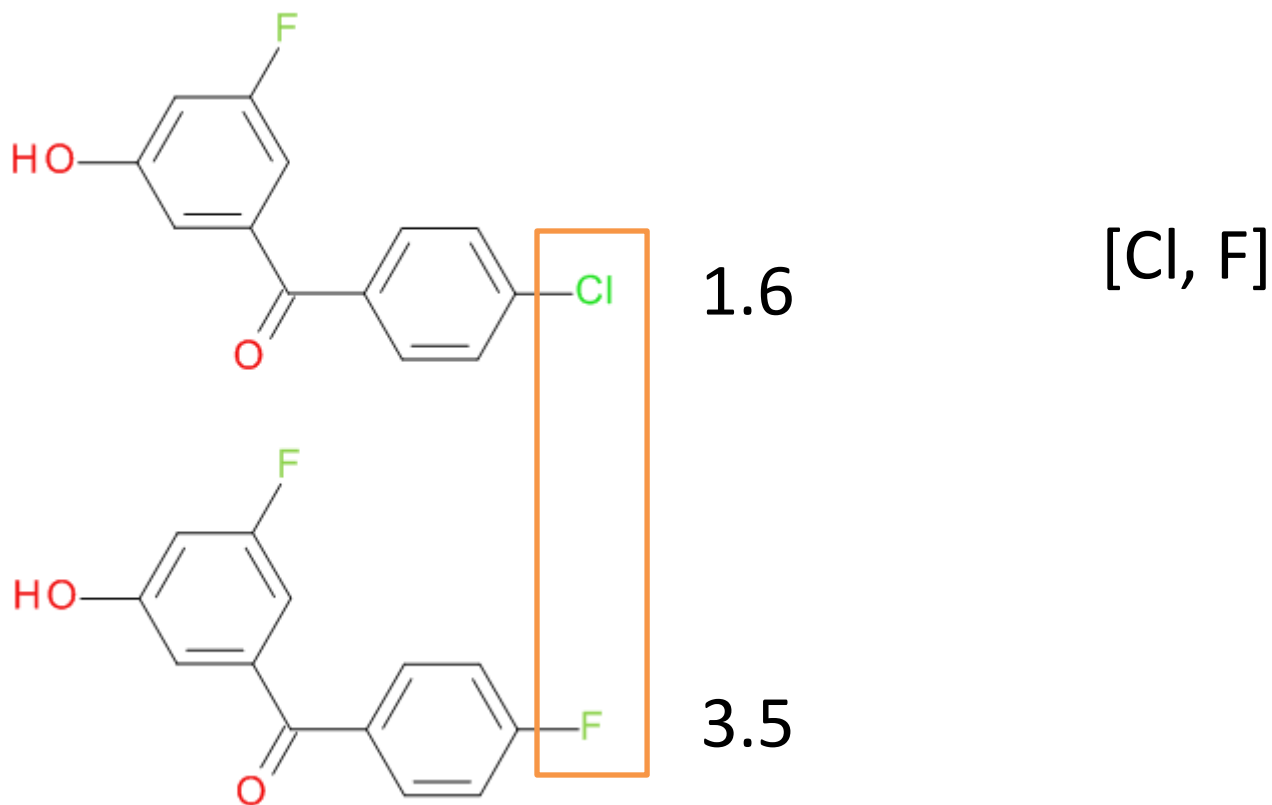
- Based on **experience** on related projects
  - What worked last time?
- By observing an **activity trend**, inferring a SAR relationship, and extrapolating
  - Aka ‘chemical intuition’
- Our additional suggestion:
  - Take advantage of the wealth of experience and trends contained in 57K med chem papers
  - **‘evidence-based medicinal chemistry’**



# MATCHED PAIRS & SERIES



# MATCHED (MOLECULAR) PAIRS



Coined by Kenny and Sadowski in 2005\*

Easier to predict **differences** in the values of a property than it is to predict the value itself

\* Chemoinformatics in drug discovery, Wiley, 271–285.



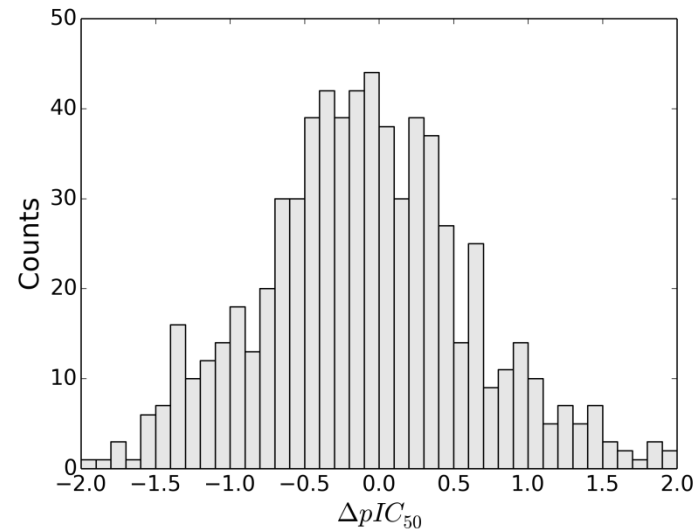
# MATCHED PAIR USAGE

- **Successfully** used for:
  - Predicting physicochemical property changes
  - Finding bioisosteres
- **Not very successful** in improving activity
  - Activity changes dependent on binding environment
  - Need to use matched pair data only for a particular binding pocket for a particular protein
- Hajduk, Sauer. *J. Med. Chem.* **2008**, *51*, 553
  - Data from 30 protein targets at Abbott
  - Most R group transformations led to potency changes normally distributed around 0



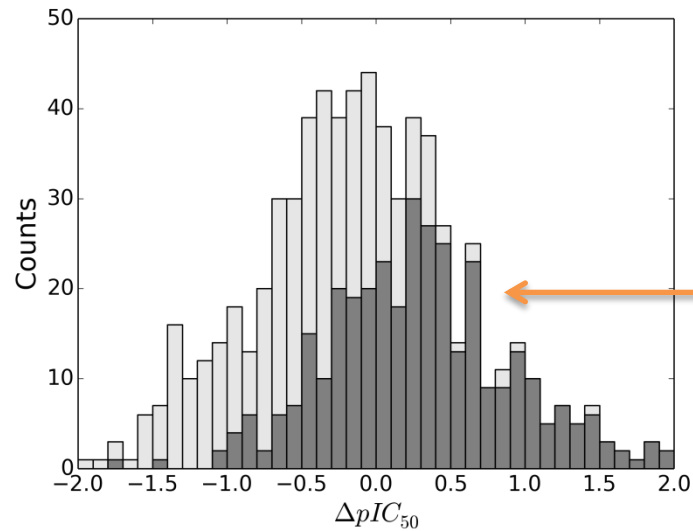
# MATCHED PAIRS AND ACTIVITY

$pIC_{50}(CC) - pIC_{50}(CCCC)$



# MATCHED PAIRS AND ACTIVITY

$$pIC_{50}(CC) - pIC_{50}(CCCC)$$

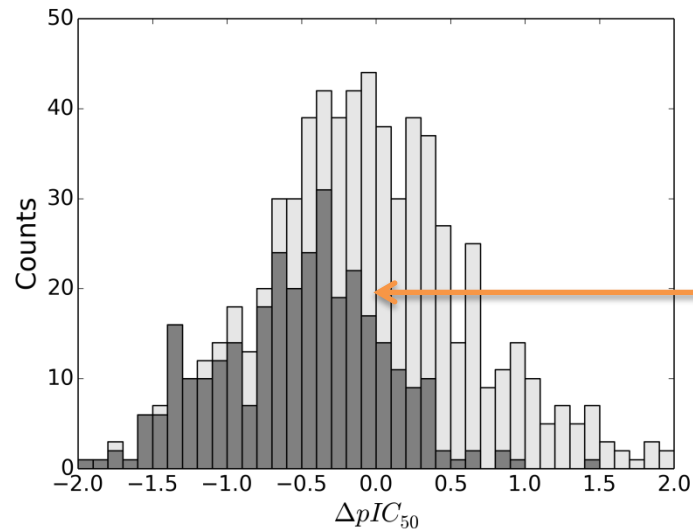


For those cases where:  
[CCC > CCCC]



# MATCHED PAIRS AND ACTIVITY

$pIC_{50}(CC) - pIC_{50}(CCCC)$

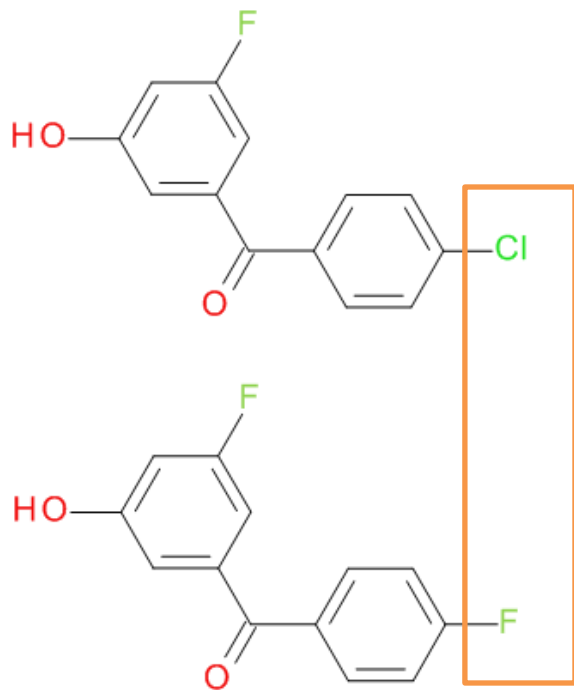


For those cases where:  
[CCC < CCCC]





# MATCHED SERIES OF LENGTH 2 = MATCHED PAIR

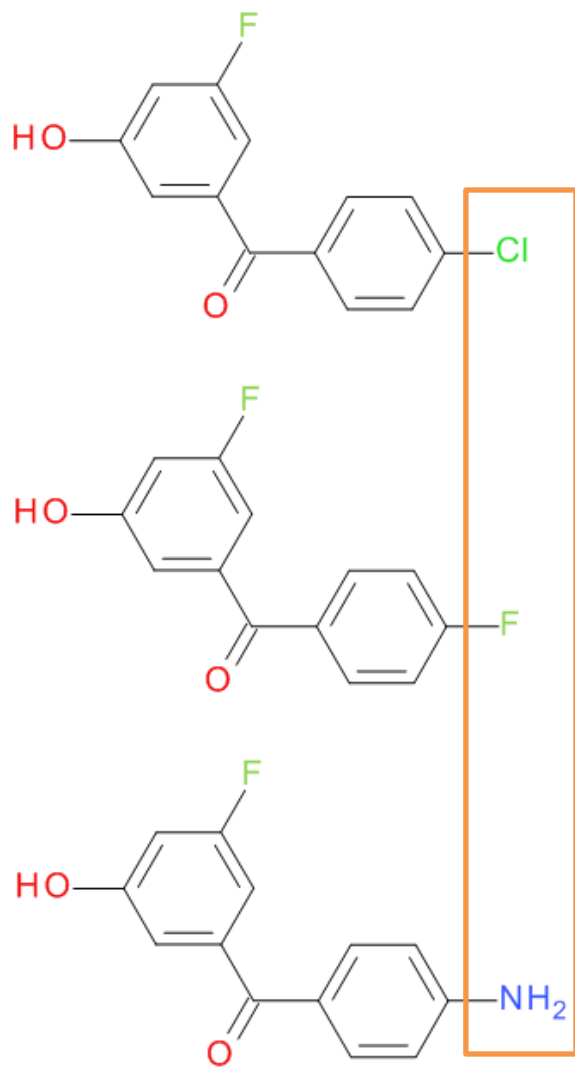


[Cl, F]

“Matching molecular series” introduced by Wawer and Bajorath, *J. Med. Chem.* **2011**, 54, 2944



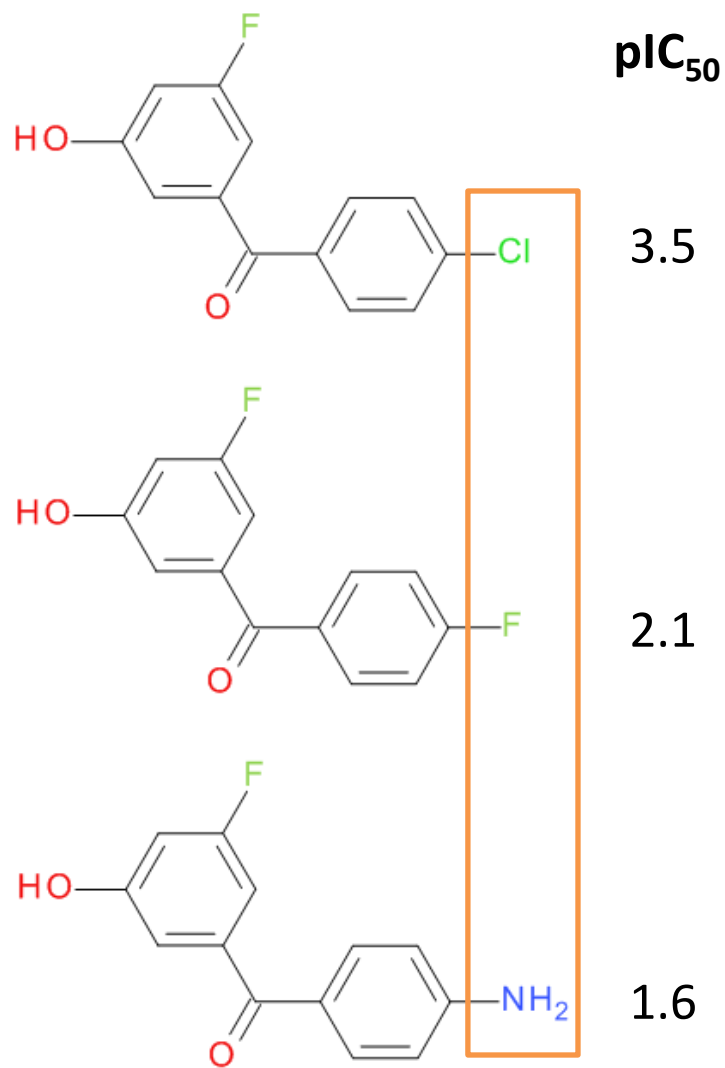
# MATCHED SERIES OF LENGTH 3



[Cl, F, NH<sub>2</sub>]



# ORDERED MATCHED SERIES OF LENGTH 3



[Cl > F > NH<sub>2</sub>]



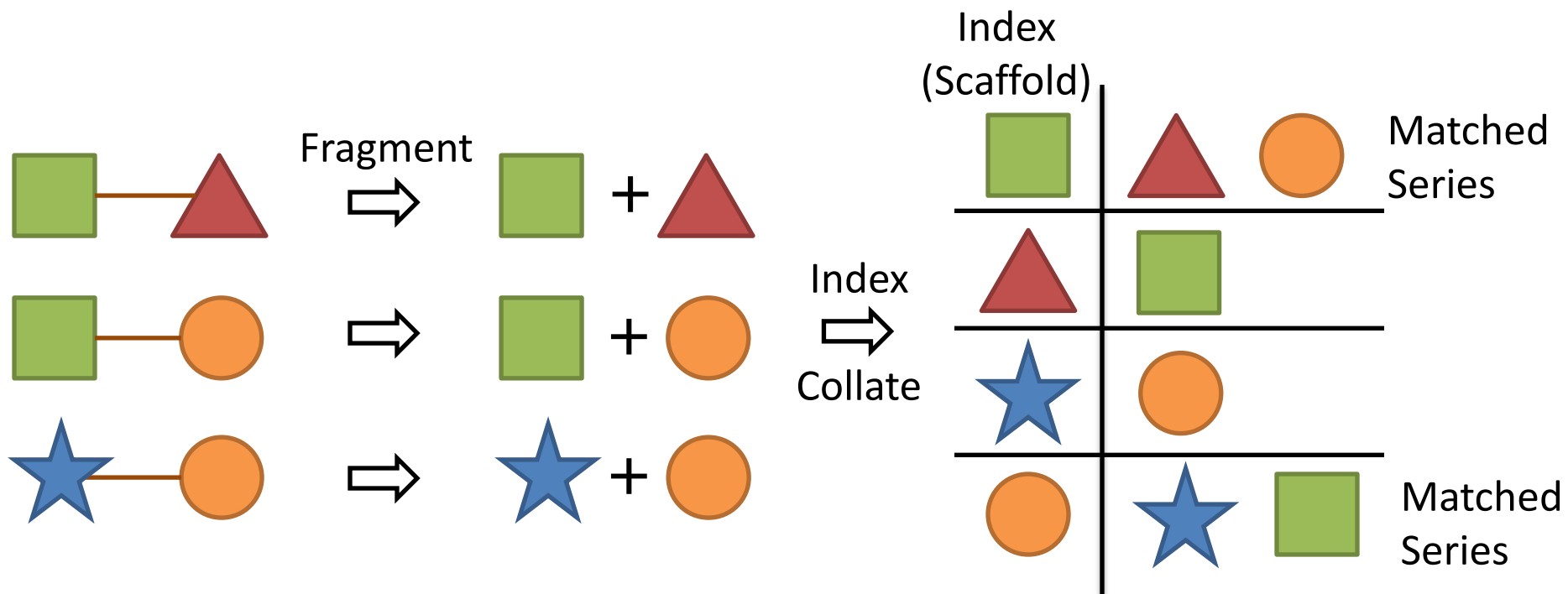
# MATCHED SERIES LITERATURE



- “**Matching molecular series**” introduced by Wawer and Bajorath *JMC* **2011**, 54, 2944
  - Subsequent papers use MMS to investigate SAR transfer, bioisosteres, SAR networks, visualisation of series and networks
- Until ours, only a single other paper on MMS
  - Mills et al *Med Chem Commun* **2012**, 3, 174



# ALGORITHM TO FIND MATCHED SERIES



- **Hussain and Rea** *JCIM* **2010**, 50, 339

- Fragment molecules at acyclic single bonds

- Single-cut only, scaffold  $\geq 5$ , R group  $\leq 12$ , preserve stereochemistry at break point

- Index each fragment based on the other

- A matched series will be indexed together



# CHEMBL BIOACTIVITY DATABASE

- **ChEMBL 19** – July 2014



- 57k papers

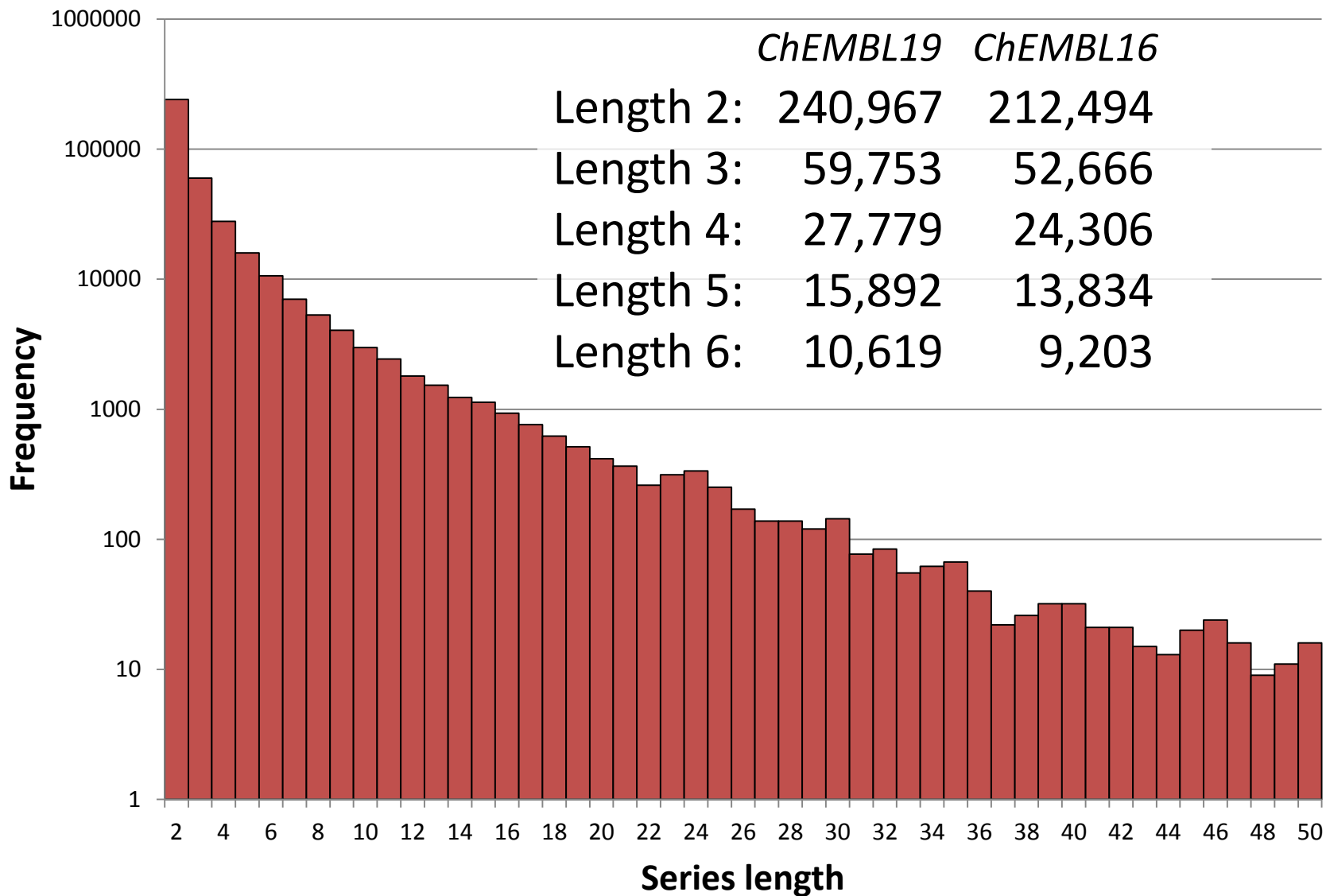
- 94% from *Bioorg. Med. Chem. Lett.*, *J. Med. Chem.*, *J. Nat. Prod.*, *Bioorg. Med. Chem.*, *Eur. J. Med. Chem.*, *Antimicrob. Agents Chemother.*, *Med. Chem. Res.*

- 1.4 million compounds with 12 million activities

- 1.1 million assays against 10k targets



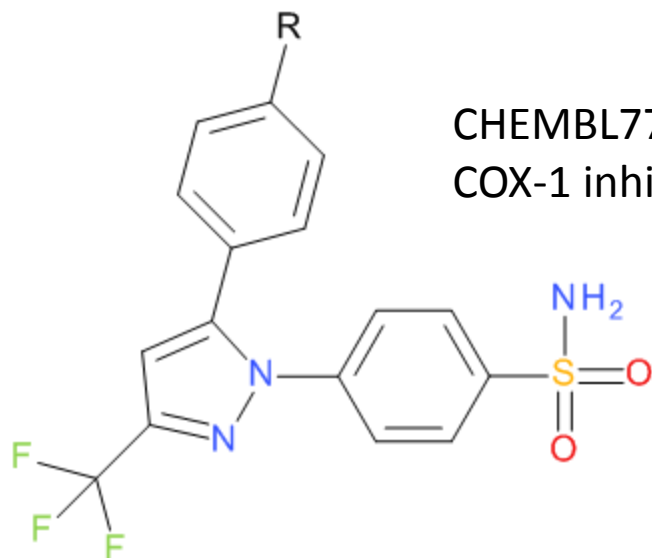
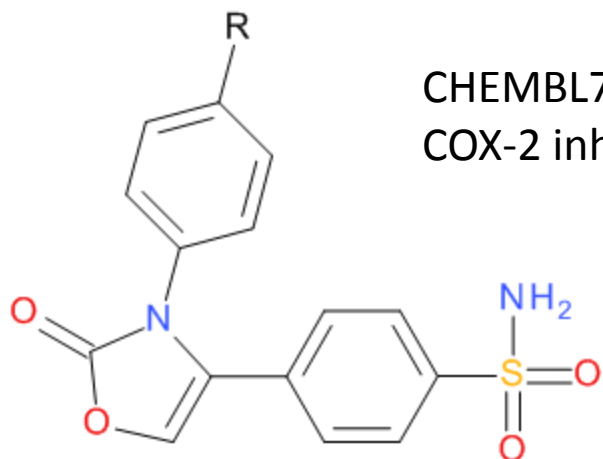
# Matched series in ChEMBL19 IC50 binding assays



SAR TRANSFER







R Group	CHEMBL768956 (pIC <sub>50</sub> )	CHEMBL772766 (pIC <sub>50</sub> )
SMe	??	5.92
NH <sub>2</sub>	??	5.88
OMe	6.68	5.59
Me	6.10	4.82
Cl	5.92	4.75
F	5.82	4.59
Et	5.81	4.54
CF <sub>3</sub>	5.70	<4.00
H	5.62	4.26
COOH	4.23	<3.60

Rank order

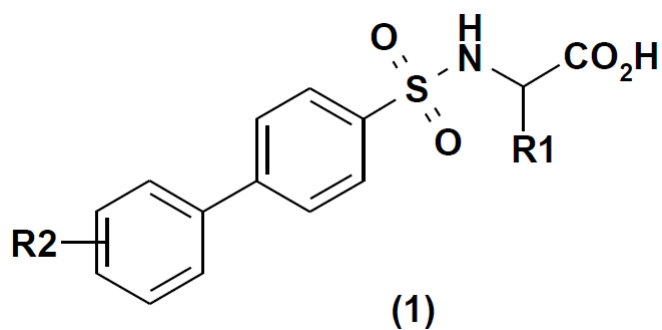
Potential SAR transfer

0.93 rank order correlation

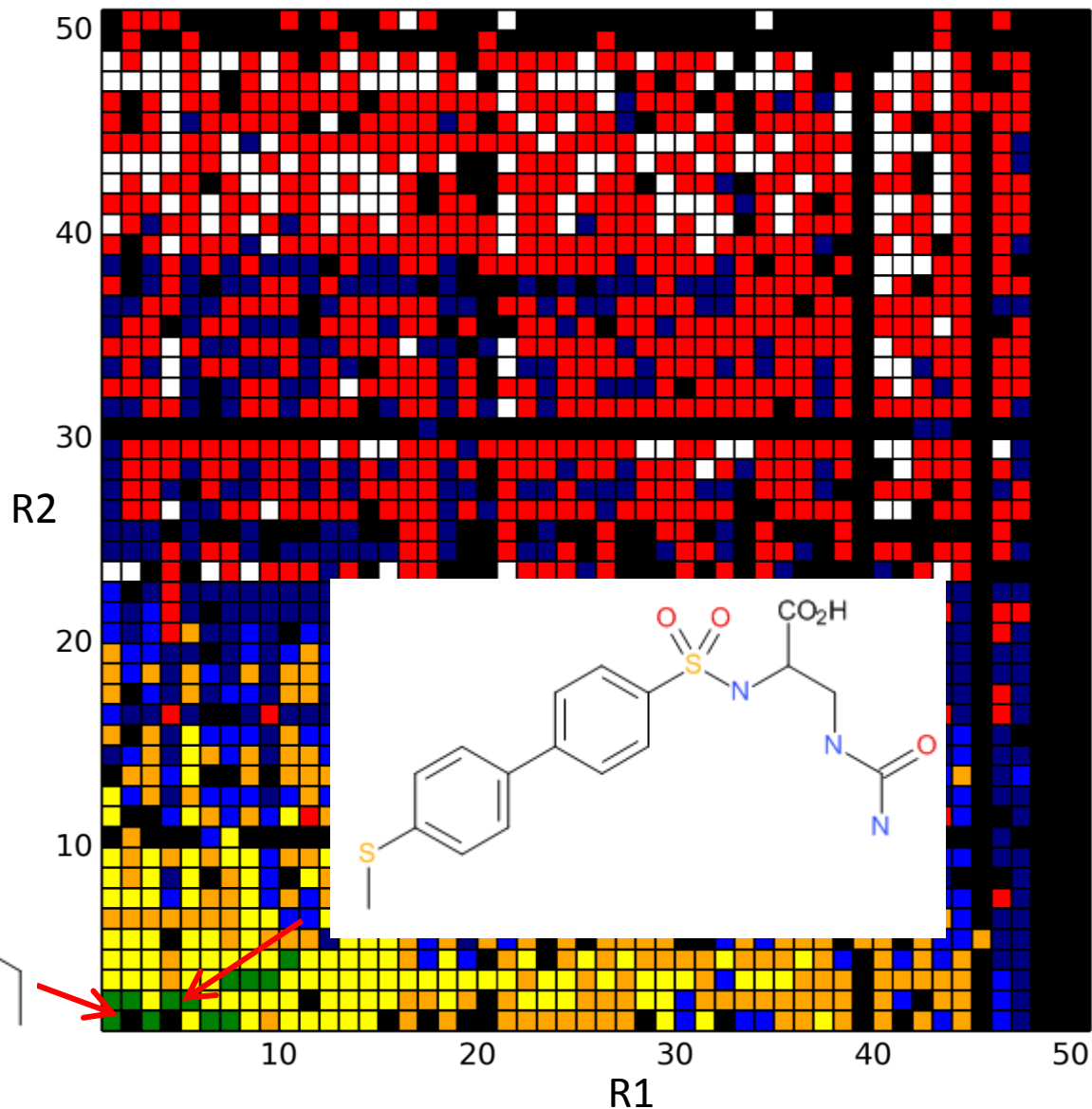
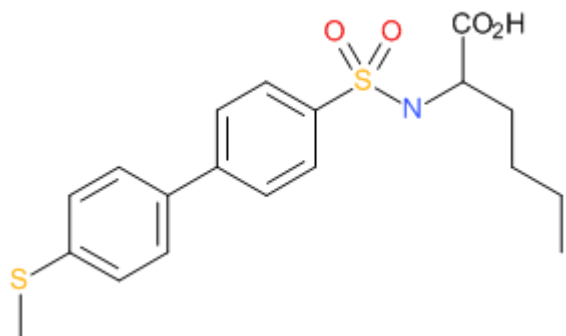
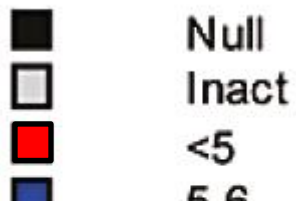


# SOX50 MATRIX FROM PICKETT ET AL.

Pickett, Green, Hunt, Pardoe, Hughes. *ACS Med. Chem. Lett.* **2011**, 2, 28.



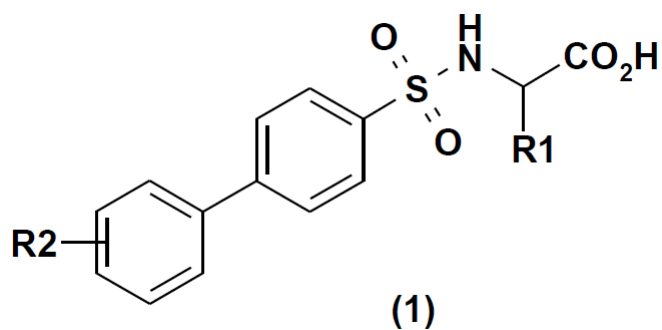
MMP-12 pIC<sub>50</sub>



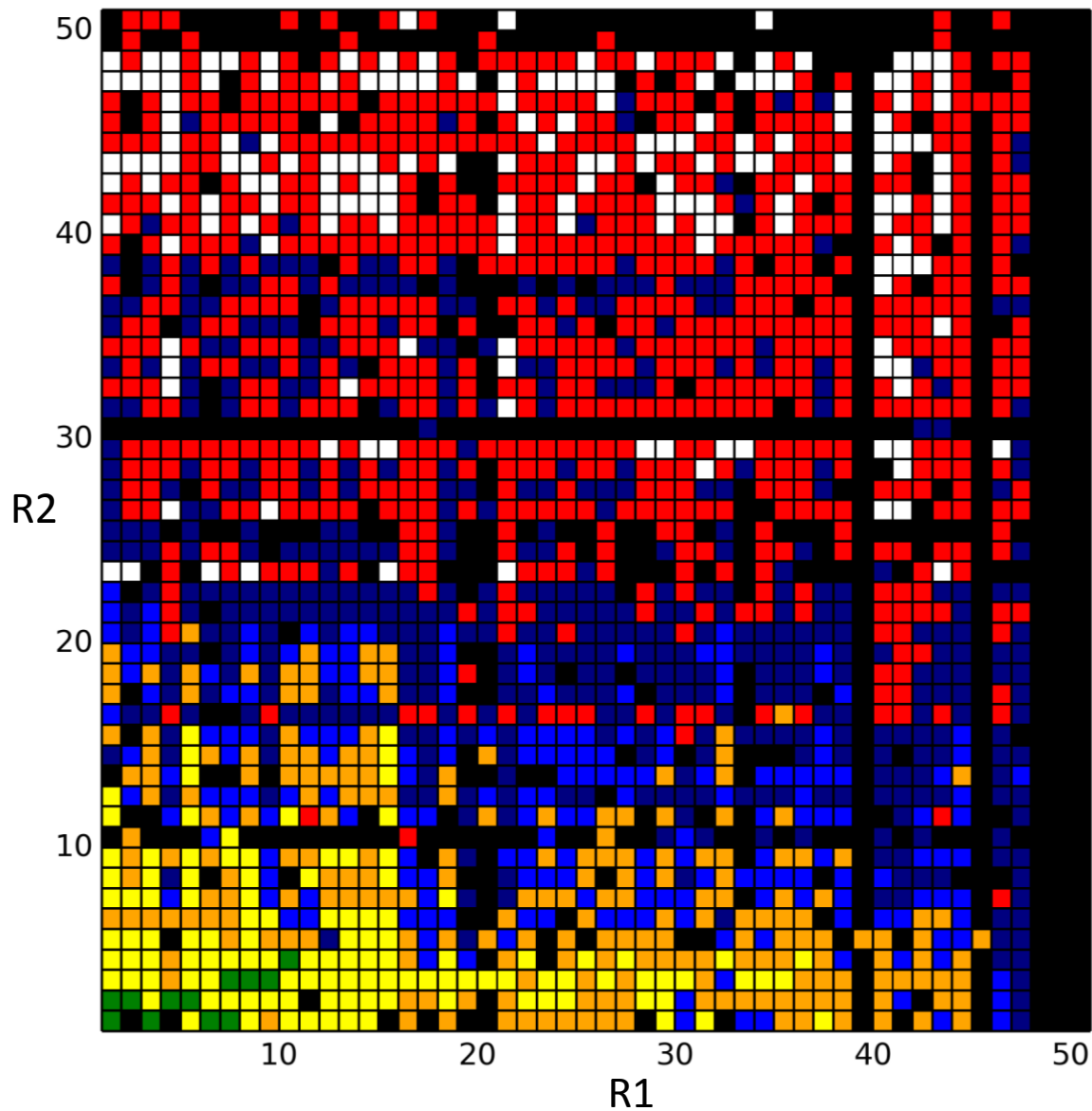
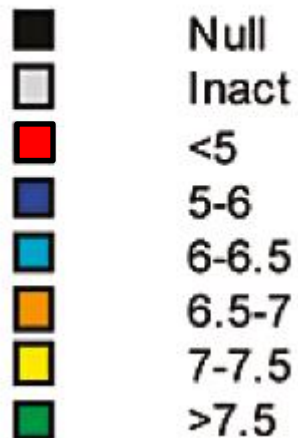


# SOX50 MATRIX FROM PICKETT ET AL.

Pickett, Green, Hunt, Pardoe, Hughes. *ACS Med. Chem. Lett.* **2011**, 2, 28.

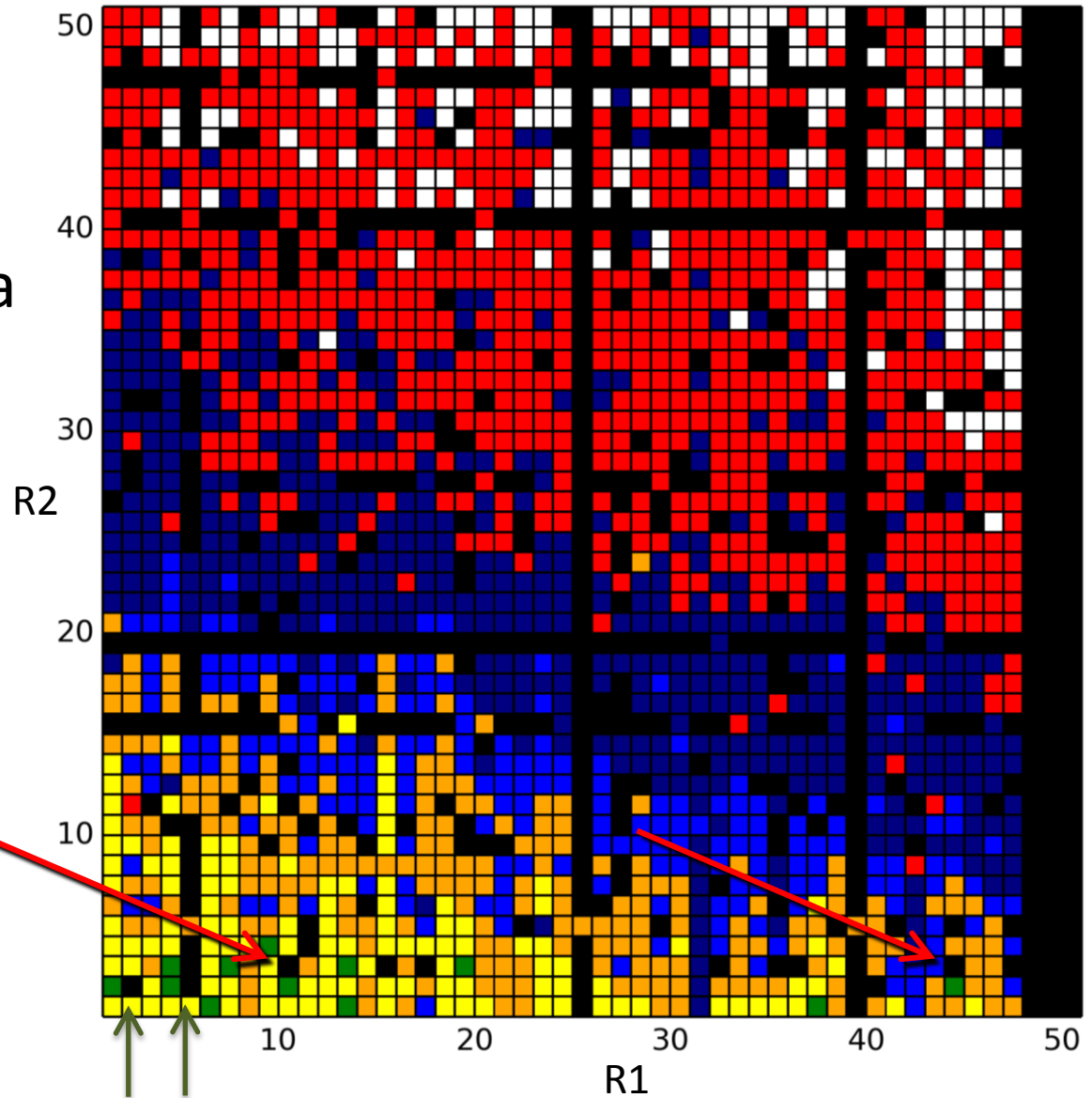


MMP-12 pIC<sub>50</sub>



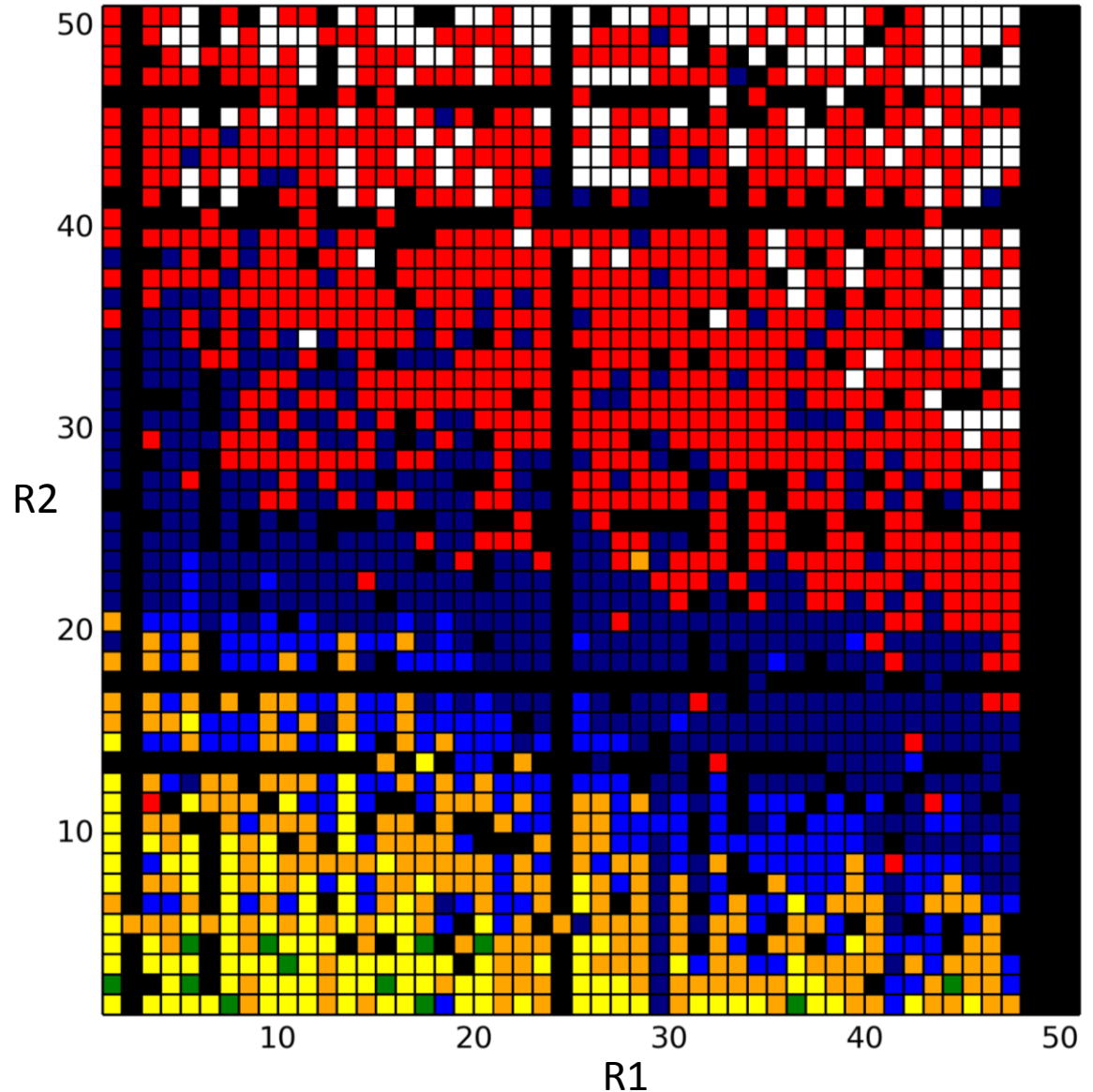
# IT'S A SET OF MATCHED SERIES

- Each row/col is a matched series
- Choose a row and a col with the fewest missing values
- Order other rows/cols by average difference with respect to chosen row/col



# MULTI-DIMENSIONAL SCALING

- Consider the whole pairwise similarity matrix
- Similar results to previous but should be more robust in general

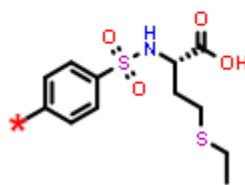
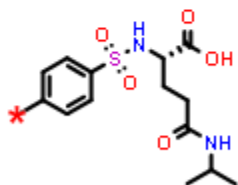
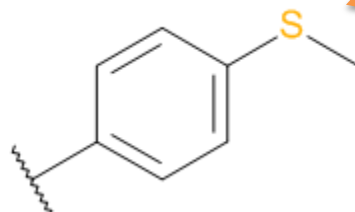


# INTERNAL SAR TRANSFER

Do an all-against-all comparison of the series

**Record\_719**  
**Record\_729**

Corr: 0.82 (p=0.00)  
N1: 39  
N2: 33  
Overlap: 31  
Pearson R<sup>2</sup>: 0.90  
LHS pred err: 0.1  
RHS pred err: 0.1



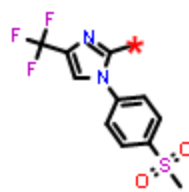
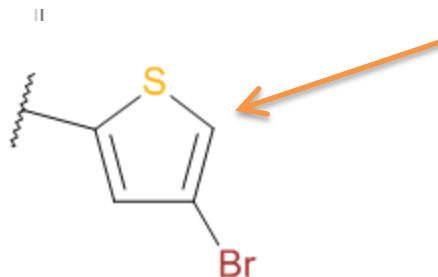
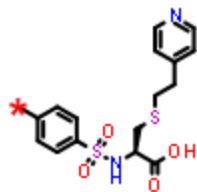
<chem>*c1ccc(cc1)SC</chem>	8.0	7.7
<chem>*c1ccc(cc1)Br</chem>	1	7.5 7.2 3
<chem>*c1ccc(cc1)CC</chem>	2	7.3 7.3 2
<chem>*c1ccc(cc1)OC(F)(F)F</chem>	2	7.3 7.2 3
<chem>*c1ccc(cc1)C(=O)OC</chem>	4	7.2 6.6 8
<chem>*c1ccc(cc1)C</chem>	5	7.1 7 5
<chem>*c1ccc(cc1)CCC</chem>	5	7.1 7.5 1
<chem>*c1ccc(cc1)CO</chem>	5	7.1 6.8 7
<chem>*c1ccc2c(c1)OCO2</chem>	5	7.1 6.2 10
<chem>*c1ccc(c(c1)F)C</chem>	9	7 7 5
<chem>*c1cccc(c1)NC(=O)C</chem>	10	6.7 5.8 13
<chem>*c1cccc1</chem>	10	6.7 6.5 9
<chem>*c1cccc(c1)F</chem>	12	6.6 6.1 11
<chem>*c1ccc(cc1C)F</chem>	13	5.9 5 16
<chem>*c1ccc(c(c1)N(=O)=O)C</chem>	14	5.8 5.3 14
<chem>*c1cccc1F</chem>	14	5.8 5.9 12
<chem>*c1cccc(c1)C#N</chem>	16	5.7 4.8 17
<chem>*c1ccc(cc1OC)OC</chem>	17	5.1 4.5 20
<chem>*c1cccc(c1)/C=C/C(=O)O</chem>	17	5.1 4.4 22
<chem>*c1cccc(c1)C(F)(F)F</chem>	19	5 4.6 19
<chem>*c1cccc(c1F)OC</chem>	20	4.8 4.4 22
<chem>*c1cccc(c1Cl)Cl</chem>	21	4.6 4.1 25
<chem>*c1cccc1Cl</chem>	21	4.6 5.1 15

# EXTERNAL SAR TRANSFER

Do an all-against-ChEMBL comparison

Record\_734  
CHEMBL763870

Corr: 0.74 (p=0.00)  
N1: 38  
N2: 65  
Overlap: 11  
Pearson R<sup>2</sup>: 0.76  
LHS pred err: 0.45  
RHS pred err: 0.06



*c1cc(cs1)Br	8.22	<b>7.59</b>
*c1ccc(c(c1)Cl)C	8.06	7.52
*c1cccc(c1)C	7.34	7.22
*c1cccc(c1)Cl	7.34	7.22
*c1cc(cc(c1)Cl)C	7.05	7.10
*c1cccc(c1)Br	7.05	7.10
*c1cc(c(c(c1)Br)OC)Br	6.93	7.05
*c1ccc(c(c1)C)Cl	6.93	7.05
*c1ccc(cc1)SC	1	7 6.80 4
*c1ccc(c(c1)F)C	2	6.8 6.96 1
*c1ccc(cc1)C	3	6.5 6.80 4
*c1cccc(c1)F	3	6.5 6.92 2
*c1cccc1	5	6.2 6.92 2
*c1ccc2c(c1)OCO2	6	6.1 6.55 7
*c1cccc1F	7	6 6.40 8
*c1cccc(c1)C(F)(F)F	8	5.1 6.68 6
*c1cc(c(c(c1)C)OC)C	9	4.8 6.14 9
*c1cccc1Cl	10	4.4 6.05 11
*c1cccc1OC	10	4.4 6.10 10



# STRENGTHS AND WEAKNESSES

- High confidence in predictions if sufficiently **long series** with correlated activities (or their rank order)
  - Not always able to find such a series
  - For short series will typically find 10s/100s/1000s of matching series with low confidence
- Suited to pairwise comparison within **focused dataset**
  - Dense SAR matrix from target with well-explored SAR



# PREFERRED ORDERS IN MATCHED SERIES



# PREFERRED ORDERS: HALIDES (N=2)

For an ordered matched series (i.e.  $A > B > C > \dots$ ), there are  $N!$  ways of arranging the R Groups:

Series	Observations*
F > H	9761
H > F	8685

Would expect 9223 for each assuming the order is random

– We can calculate **enrichment**

\*Dataset is ChEMBL19 IC<sub>50</sub> data for binding assays (transformed to pIC<sub>50</sub> values)



# PREFERRED ORDERS: HALIDES (N=2)

For an ordered matched series (i.e.  $A > B > C > \dots$ ), there are  $N!$  ways of arranging the R Groups:

Series	Enrichment	Observations
F > H	1.06*	9761
H > F	0.94*	8685

Would expect 9223 for each assuming the order is random

– We can calculate **enrichment**

\*Significant at 0.05 level according to binomial test after correcting for multiple testing (Bonferroni with N-1)



# PREFERRED ORDERS: HALIDES (N=3)

Series	Enrichment	Observations
Cl > F > H	1.90*	1478
H > F > Cl	1.08	838
F > Cl > H	0.86*	673
F > H > Cl	0.78*	607
Cl > H > F	0.76*	589
H > Cl > F	0.63*	490



# PREFERRED ORDERS: HALIDES (N=4)

Series	Enrichment	Observations
Br > Cl > F > H	5.43*	263
Cl > Br > F > H	3.22*	156
<b>H &gt; F &gt; Cl &gt; Br</b>	<b>1.59*</b>	<b>77</b>
Br > Cl > H > F	1.43	69
F > Cl > Br > H	1.40	68
Cl > Br > H > F	0.85	41
...	...	...
<b>H &gt; F &gt; Br &gt; Cl</b>	<b>0.76</b>	<b>37</b>
...	...	...
<b>H &gt; Br &gt; F &gt; Cl</b>	<b>0.50*</b>	<b>24</b>
Cl > H > F > Br	0.48*	23
Cl > F > H > Br	0.45*	22
H > Cl > F > Br	0.43*	21
Br > F > H > Cl	0.41*	20
F > H > Br > Cl	0.41*	20
H > Cl > Br > F	0.41*	20
F > Br > H > Cl	0.35*	17
<b>Br &gt; H &gt; F &gt; Cl</b>	<b>0.23*</b>	<b>11</b>

N=2: Max = 1.06, Min = 0.94

N=3: Max = 1.90, Min = 0.63

N=4: Max = 5.43, Min = 0.232

Longer series exhibit greater preferences

If [H>F>Cl] is observed, will Br increase activity further?  
 149 observations of [H>F>Cl]  
 but only 11 where [Br>H>F>Cl]



MATSY:  
PREDICTION USING  
MATCHED SERIES



# FIND R GROUPS THAT INCREASE ACTIVITY



In-house

Query

**A > B**



**MATSY**

**A > B > C**

**C > A > B**

**D > A > B > C**

**D > A > C > B**

**E > D > A > B**

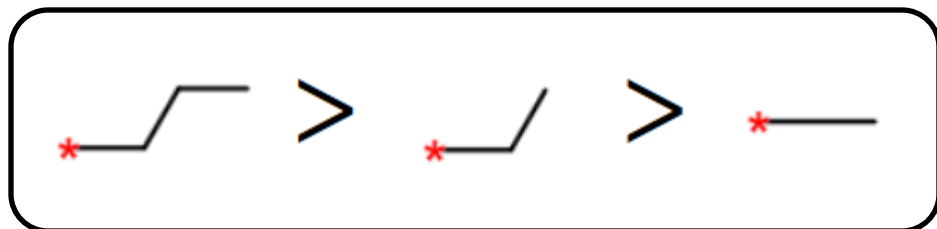
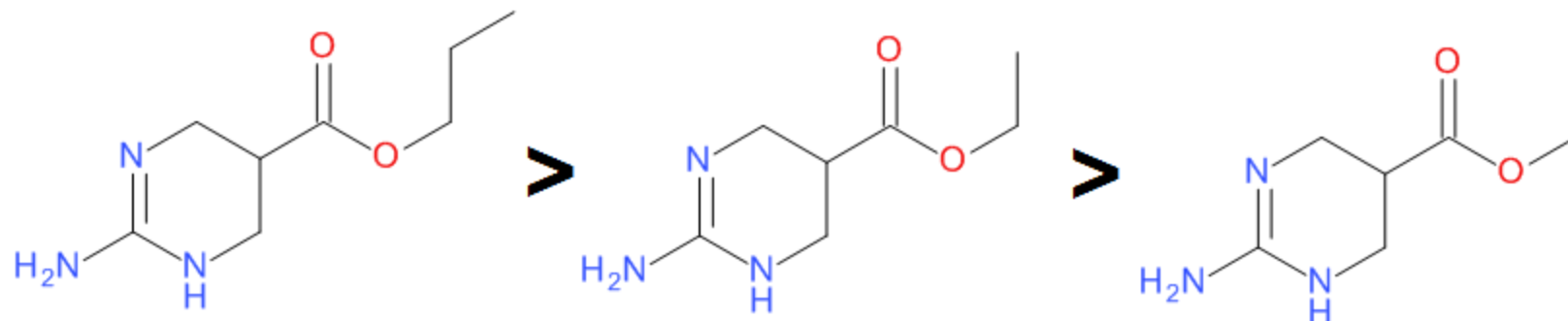
...

R Group	Observations	Obs that increase activity	% that increase activity
D	3	3	100
E	1	1	100
C	4	1	25
...	...	...	...





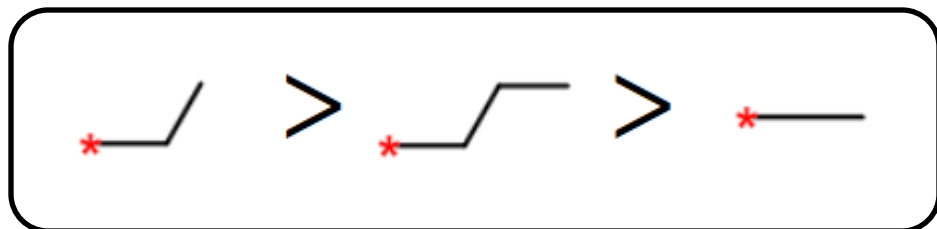
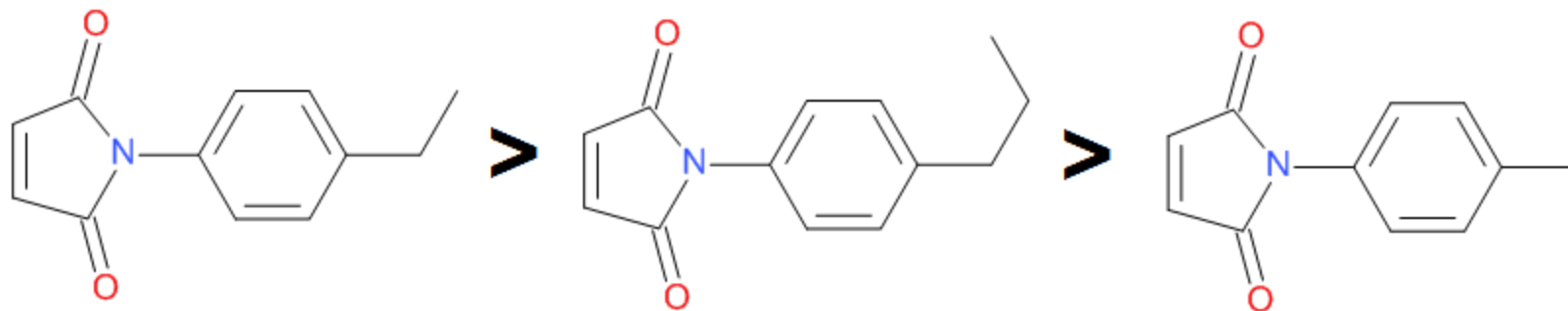
# EXAMPLE



%   
 >   
 Counts   
 ΔLogP

	%	Counts	ΔLogP
	90	21	+3.3
	72	60	+1.7
	69	32	+2.8
	63	27	+1.6
	60	40	-0.1

# EXAMPLE II

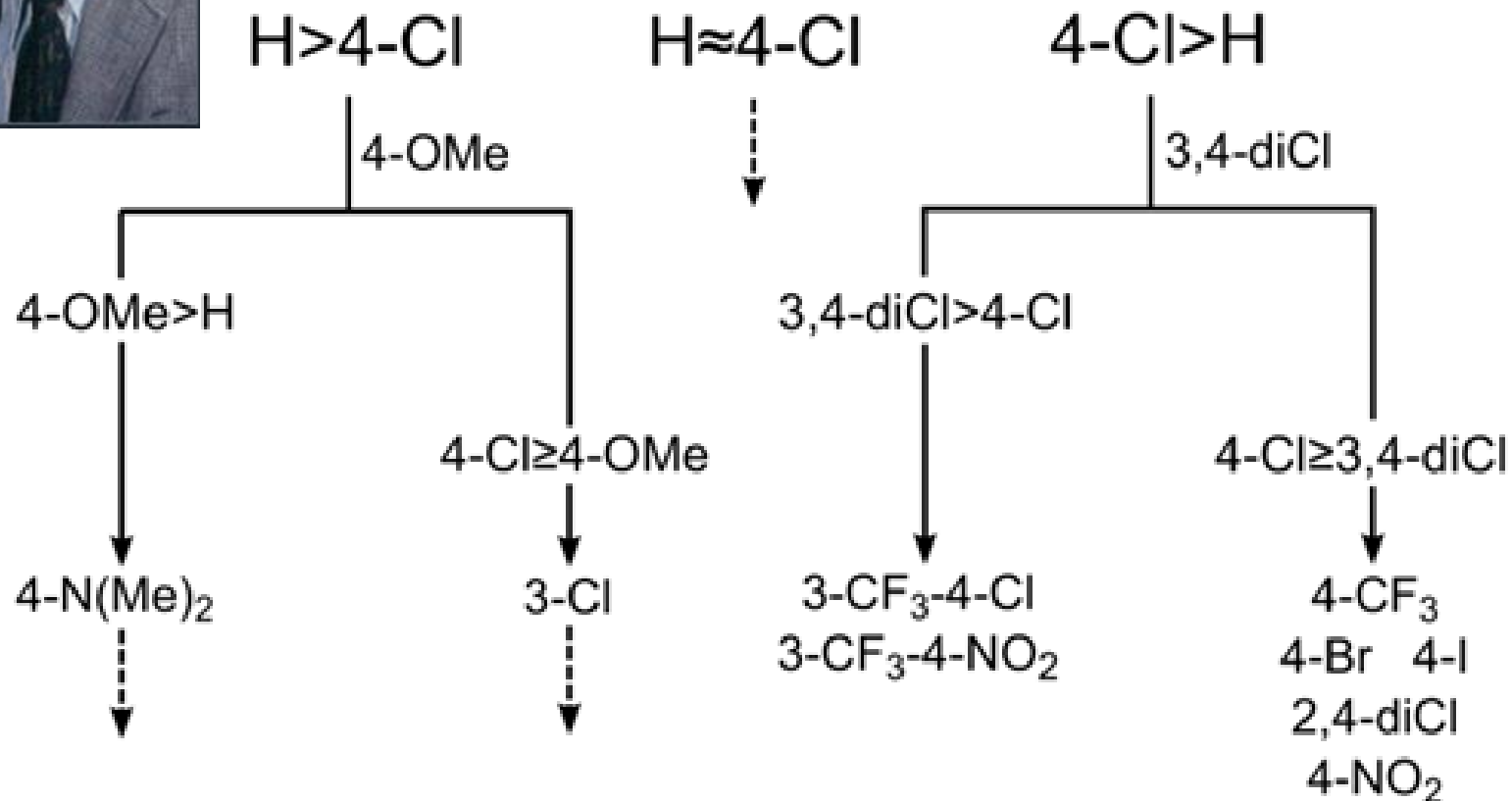


⬆ % ⬇ Counts ⬇  $\Delta\text{LogP}$  ⬆  
>

	38	21	<b>-0.8</b>
	37	27	+0.9
	33	<b>111</b>	+0.3
	33	27	+1.0
	33	21	<b>-1.6</b>



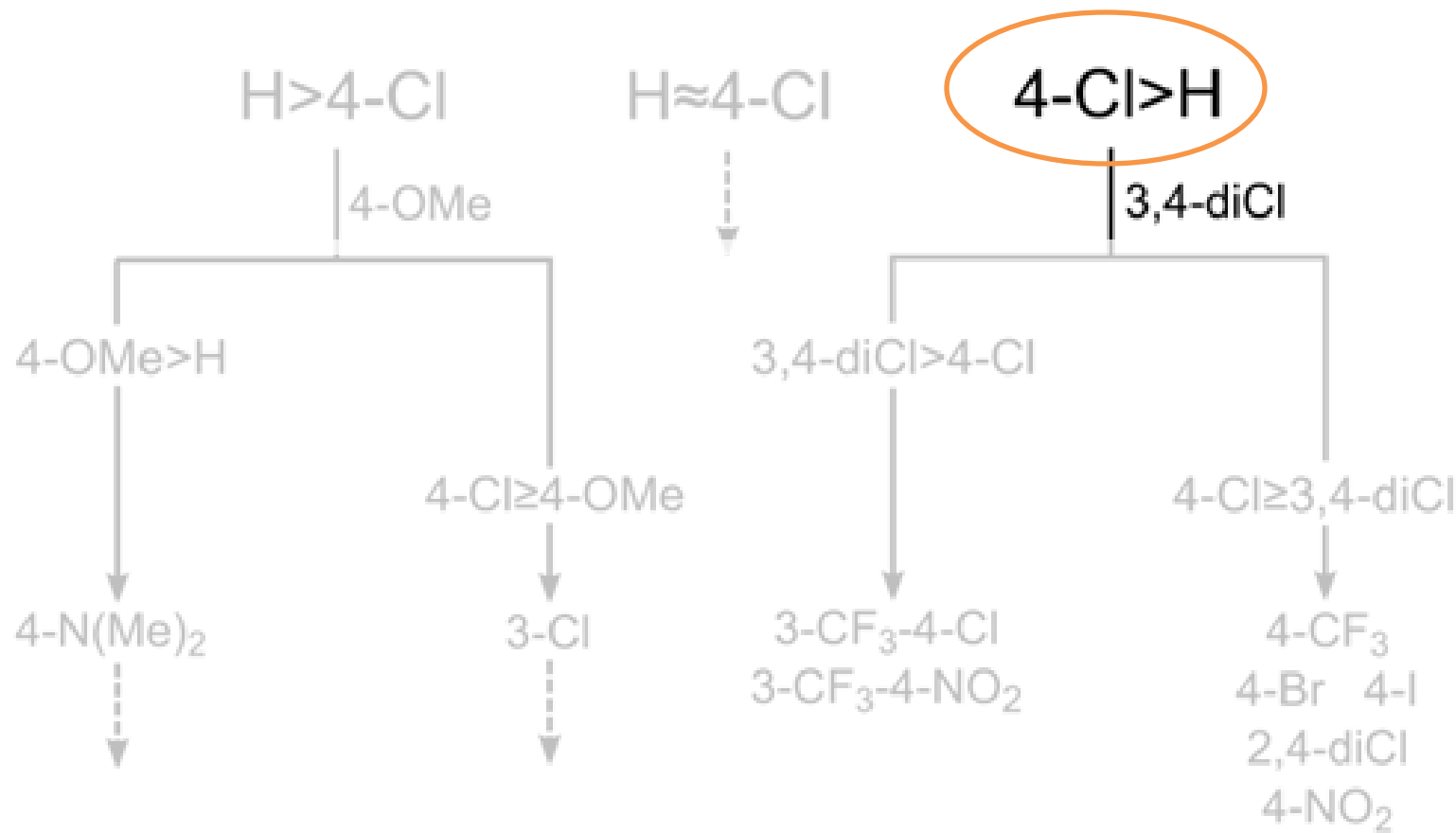
# TOPLISS DECISION TREE



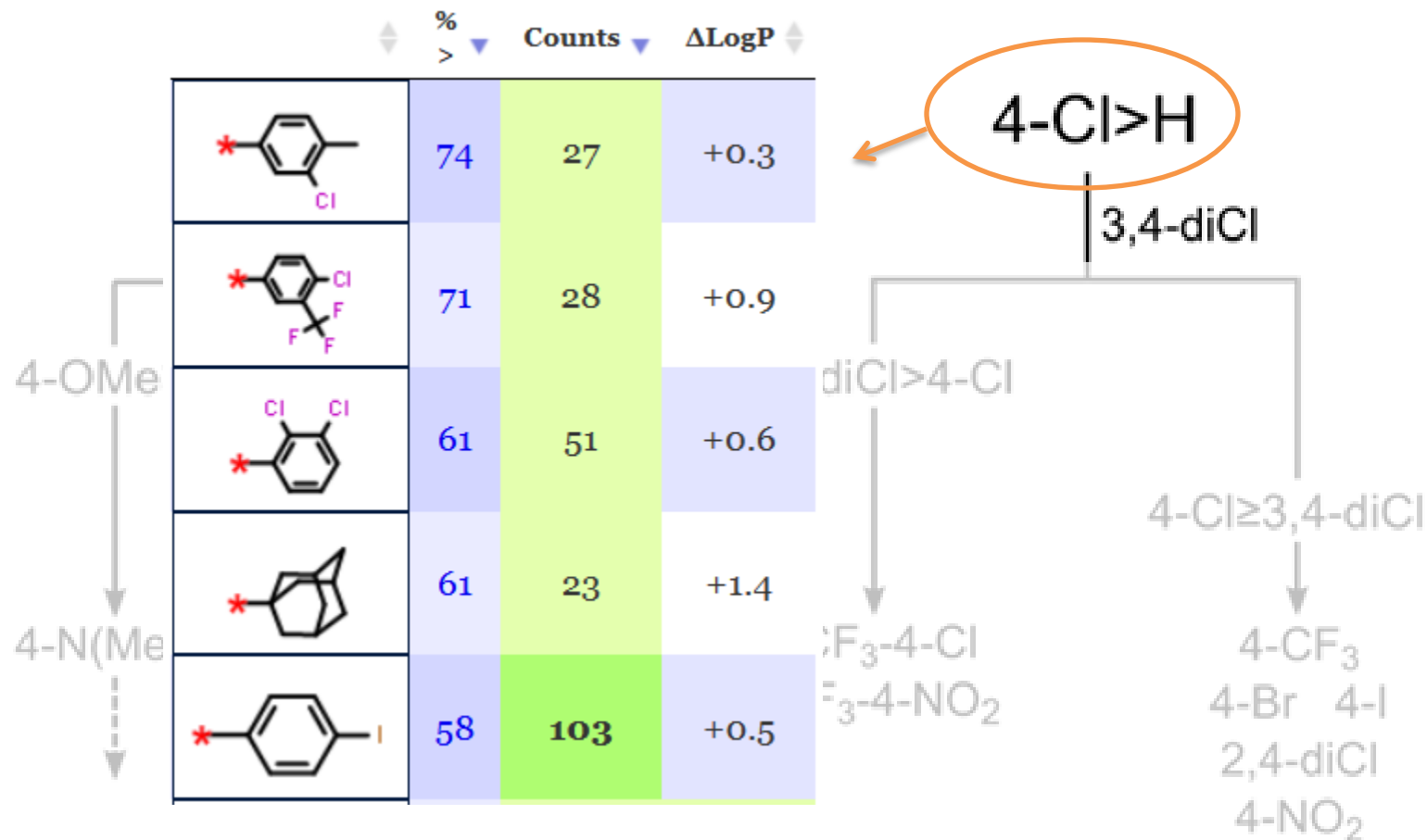
Topliss, J. G. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Chem.* **1972**, *15*, 1006–1011.



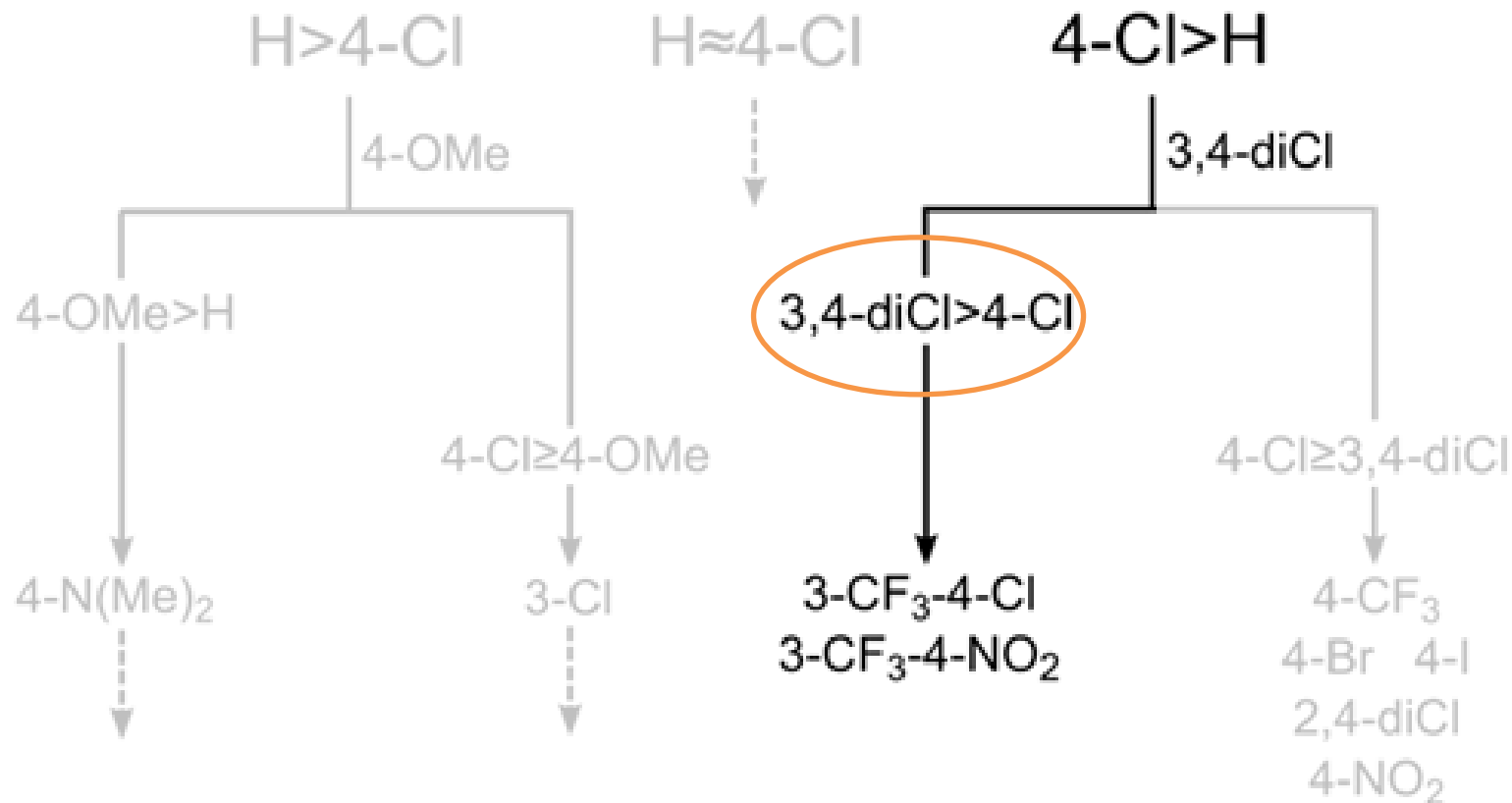
# TOPLISS DECISION TREE



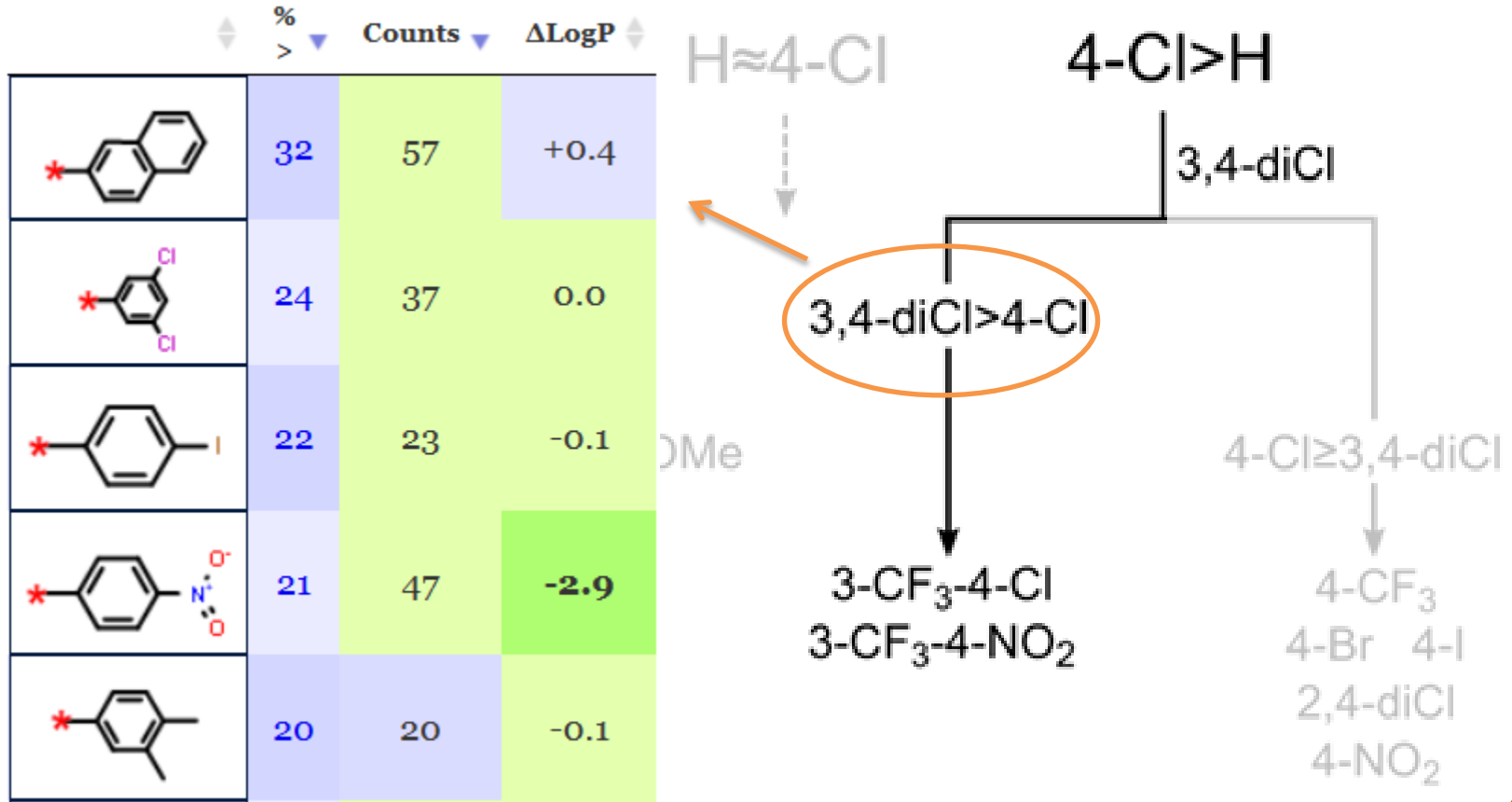
# TOPLISS DECISION TREE



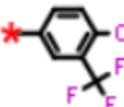
# TOPLISS DECISION TREE



# TOPLISS DECISION TREE

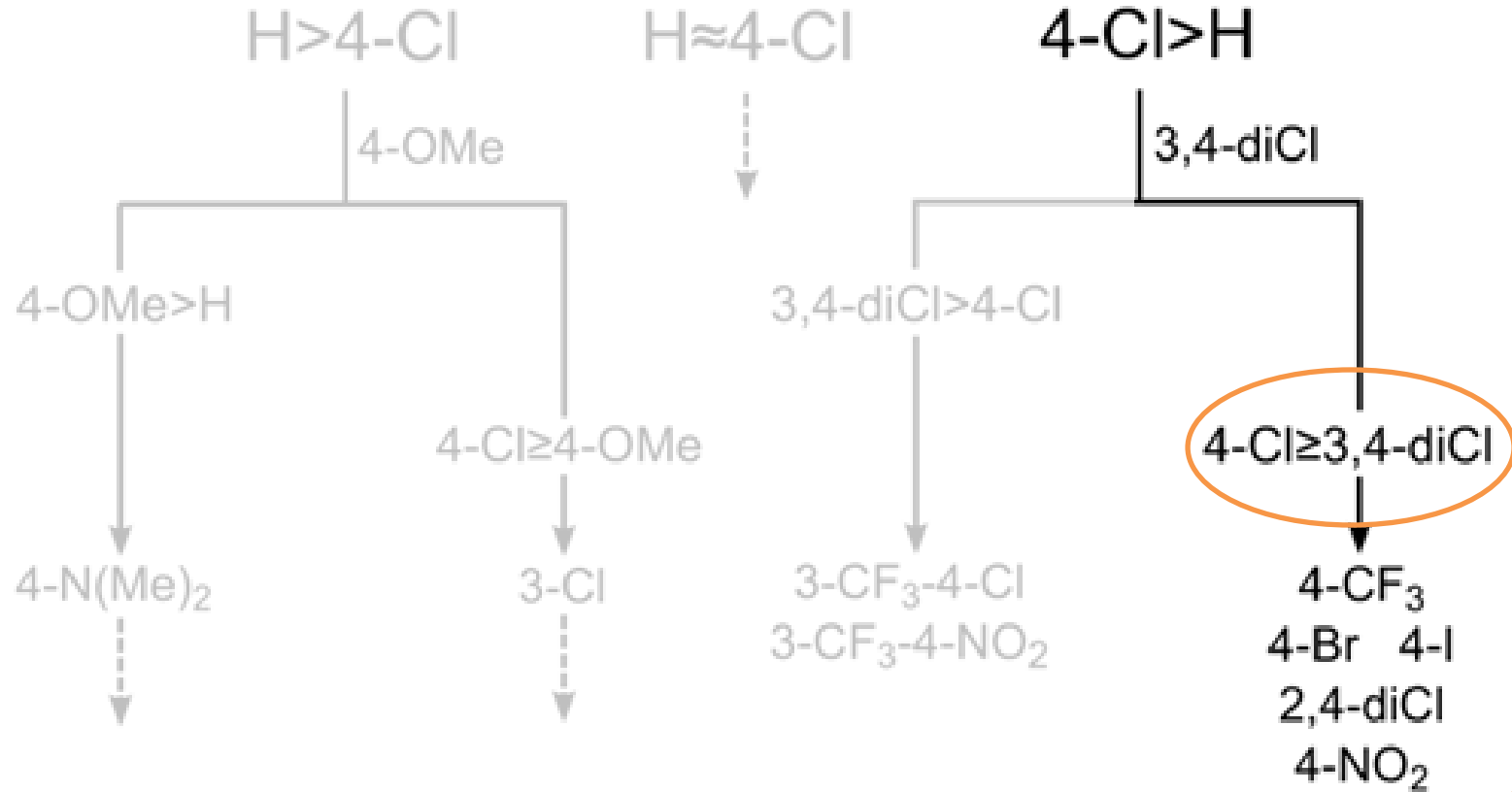


(1<sup>st</sup> if lower cutoff)

	33	5.00	15	+0.3
---	----	------	----	------

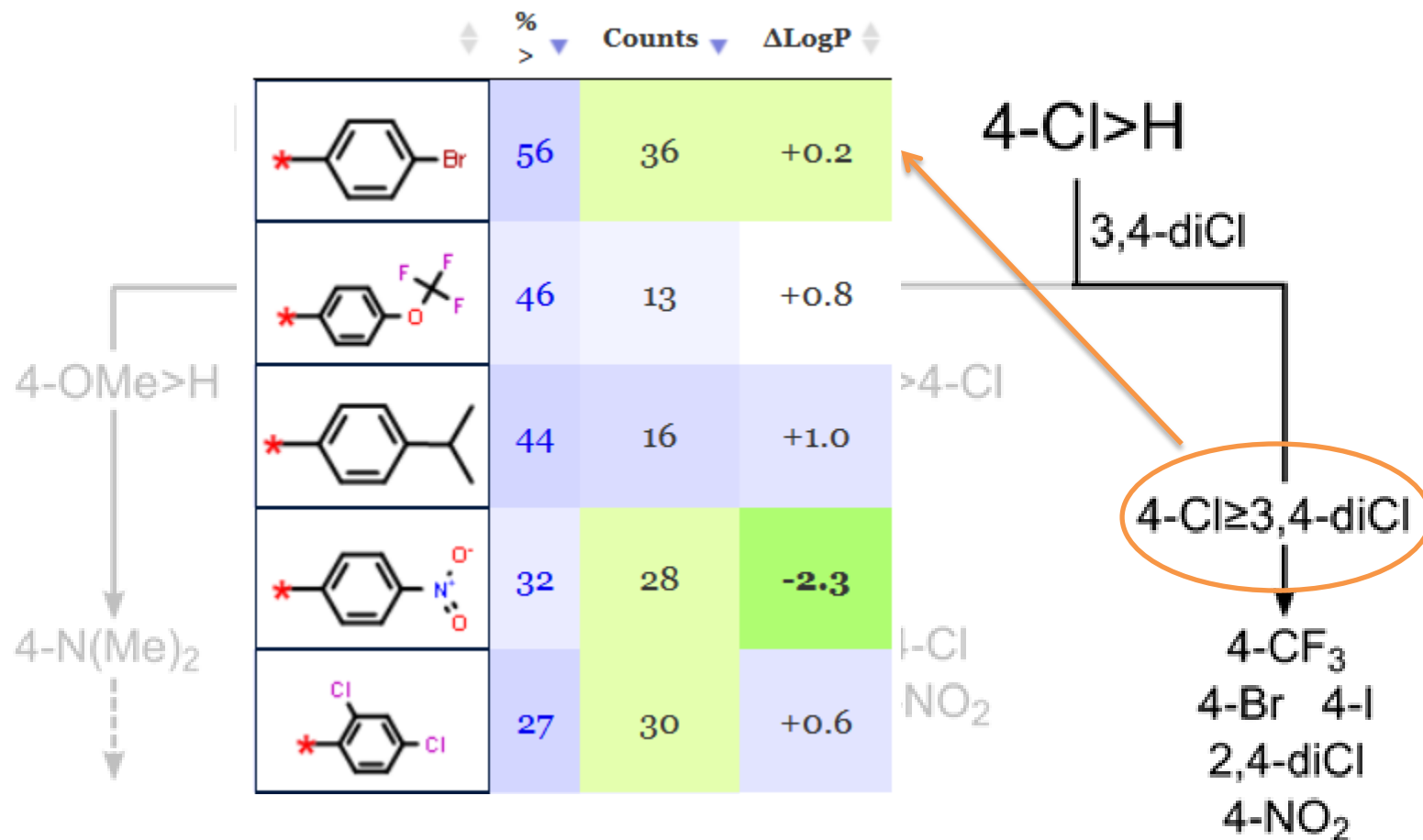


# TOPLISS DECISION TREE





# TOPLISS DECISION TREE

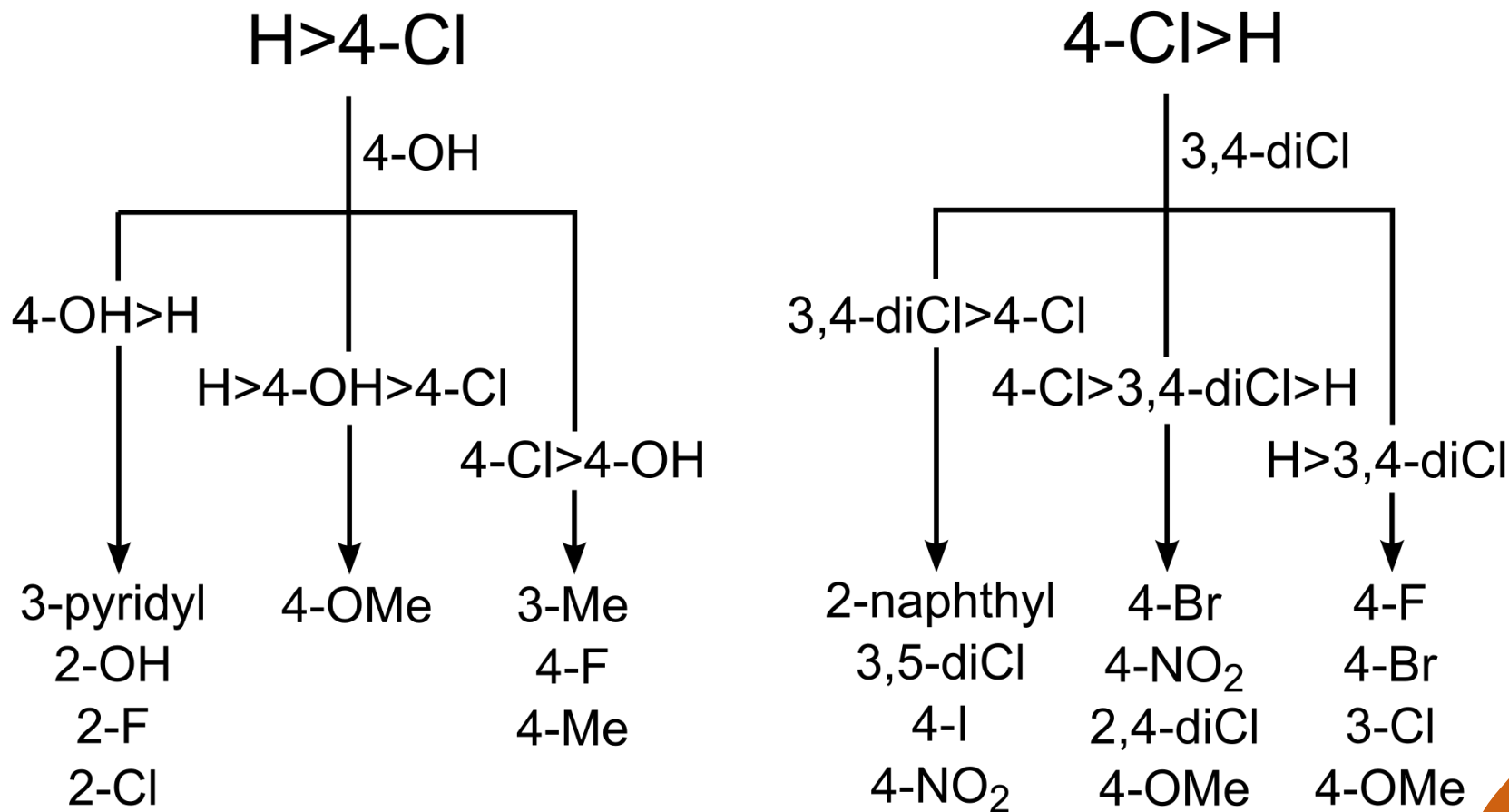


(20<sup>th</sup>)

	11	28	+0.3
---	----	----	------

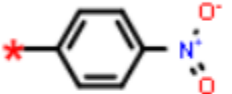
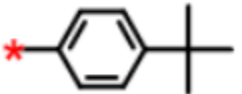


# MATSY DECISION TREE (ONE OF MANY)



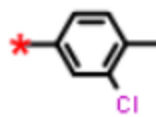
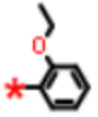
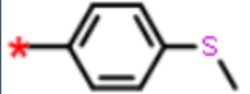
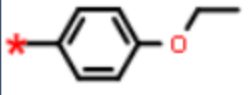
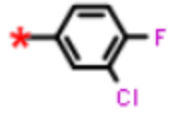
# MODIFYING THE PREDICTIONS FOR

# 4-Cl > H

	% >	Counts
	67	30
	47	30
	46	24
	44	25
	42	77

**Kinases**

Target-specific

	% >	Counts	$\Delta\text{LogP}$
	63	27	+0.3
	55	20	-0.4
	49	63	0.0
	48	46	-0.4
	48	46	+0.1

**$\Delta\text{LiPE} > 0$**

Incorporate metrics

# DRAG-AND-DROP INTERFACE TO MATSY

1/2 3 4 5 6 7 8 9 10 11 12 **Ph 1** Ph 2 Ph 3 Ph 4 Ph 5/6 Custom

Stronger binding < ChEMBL19 pIC50 > Weaker binding

? > [ ] > Clc1ccc(Cl)cc1 > Clc1ccc(cc1) > [ ]

2277

	Counts	$\Delta\text{LogP}$
<chem>Clc1ccc(Cl)cc1</chem>	54 391	+0.6
<chem>c1ccc2ccccc2c1</chem>	54 37	+2.0
<chem>CCCC1=CC=CC=C1</chem>	53 30	+1.1
<chem>Fc1ccc(Cl)cc1</chem>	52 46	+0.1
<chem>BrC1=CC=CC=C1</chem>	50 521	+0.2
<chem>Fc1cc(Cl)ccc1</chem>	50 32	+0.1
<chem>S1=CC=CC=C1</chem>	49 63	0.0
<chem>COCC1=CC=CC=C1</chem>	48 46	-0.4
<chem>Clc1ccc(Cl)cc1</chem>	48 25	+0.3
<chem>Oc1ccc(O)cc1</chem>	48 21	-2.2

Showing 11 to 20 of 111 entries

Previous Next

IS IT JUST LOGP?

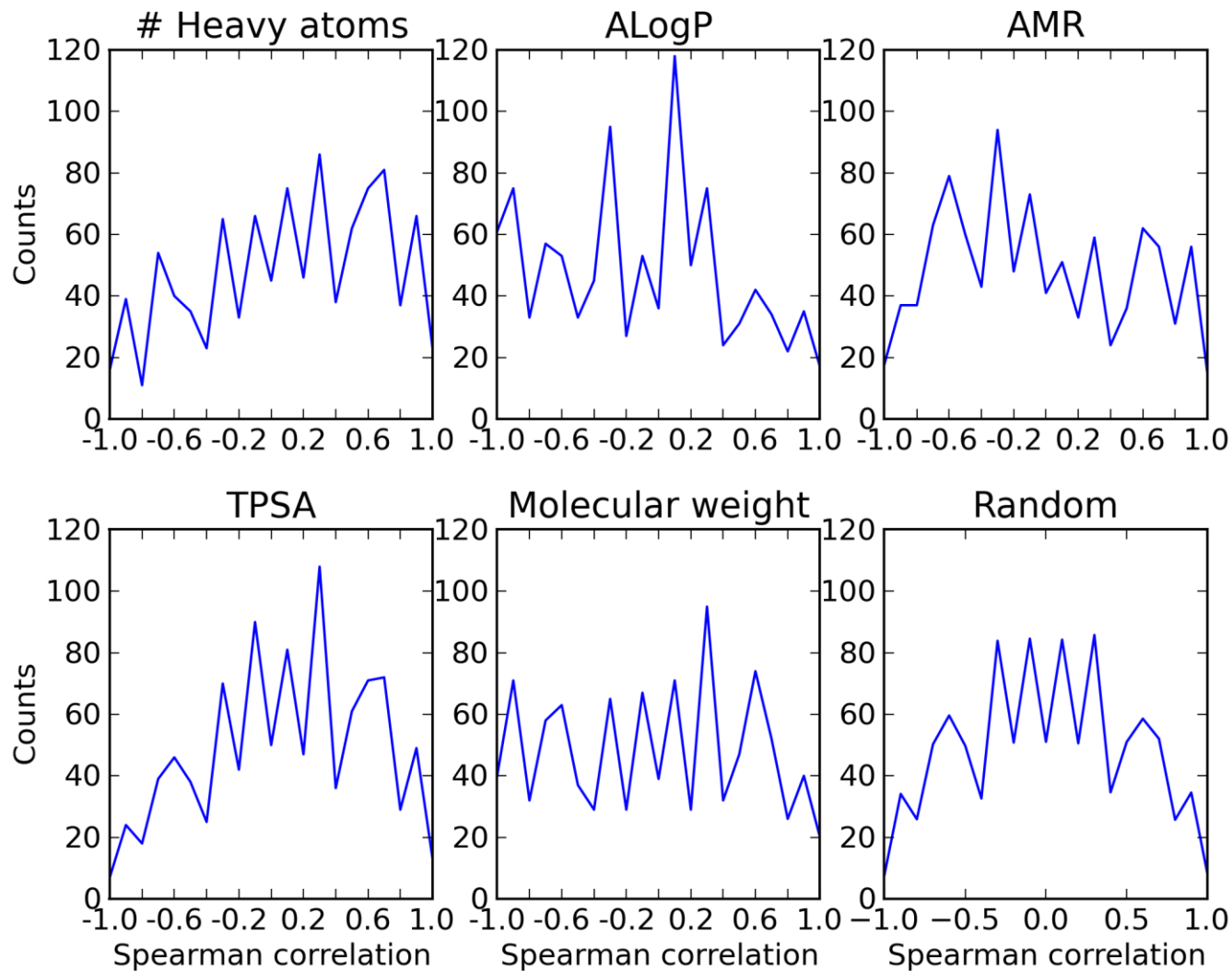


## Matched series predictions

Series length	Testset size	Predictions made	In top 5	% found predicted	% found overall
2	48699	39648 (81%)	2427	6	5
3	43450	21858 (50%)	4190	19	10
4	33705	8514 (25%)	3387	40	10
5	24273	1868 (8%)	1016	54	4
6	17379	76 (0%)	33	43	0

- Calculate Spearman correlation of the 1016 series against common descriptors
  - RDKit: ALogP, AMR, TPSA, MolWt, NumHvyAtoms





# IN SUMMARY

- Longer matched series ( $N > 2$ ) show an increased preference for particular activity orders
- This can be exploited to **predict R groups** that will increase activity
  - Predictions are typically based on data from a range of targets and structures
- Completely **knowledge-based**
  - Can link predictions to particular targets/structures
  - Predictions refined based on new results





# Beyond Matched Pairs

Using matched series for activity prediction

noel@nextmovesoftware.com

## Acknowledgements

Roger Sayle

Jonas Bostrom, AstraZeneca

Using Matched Molecular Series as a  
Predictive Tool To Optimize Biological  
Activity

*J. Med. Chem.* **2014**, *57*, 2704.

