Joint CICAG and Cambridge Cheminformatics Network Meeting
19th Feb 2014

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

Noel O'Boyle and Roger Sayle

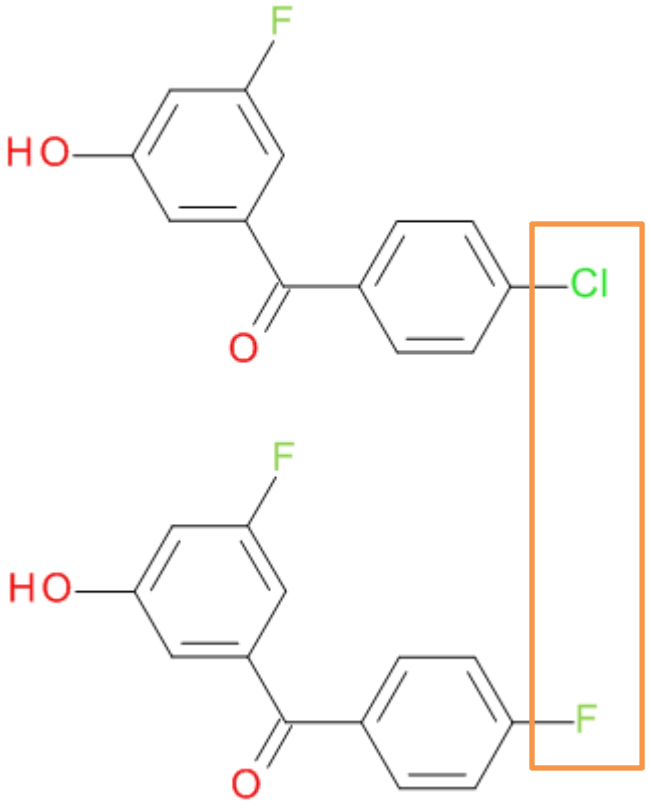
NextMove Software

Jonas Boström and Adrian Gill

AstraZeneca

MATCHED PAIRS & SERIES

MATCHED (MOLECULAR) PAIRS



1.6

3.5

[Cl, F]

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

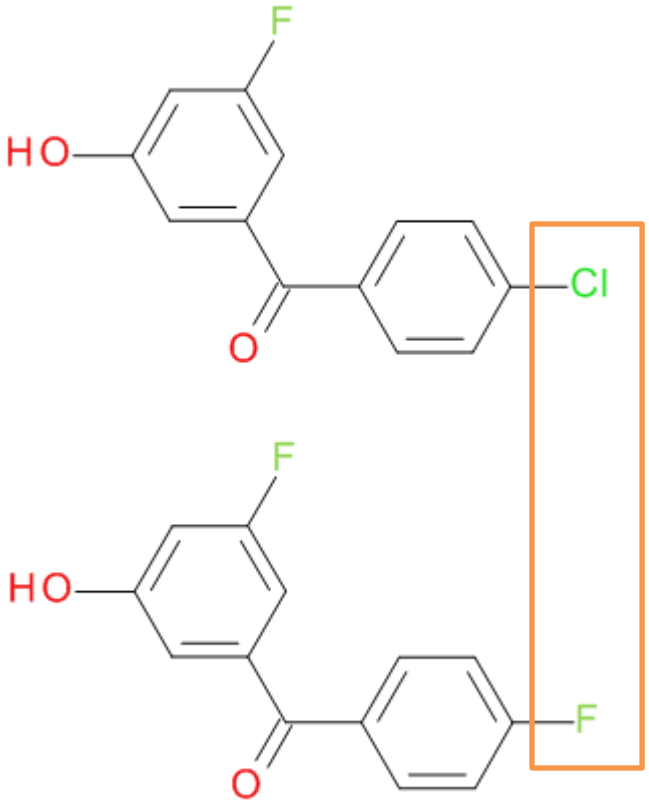
- Successfully used for:
  - Rationalising and predicting physicochemical property changes
  - Finding bioisosteres
- Not very successful in improving activity
  - Activity changes dependent on binding environment
- Various approaches to address this
  - Incorporate atom environment (WizePairZ and Papadatos et al JCIM, 2010, 50, 1872)
  - Incorporate protein environment (VAMMPIRE and 3D Matched Pairs)

LOOKING BEYOND MATCHED PAIRS

- Consider the following ‘trivial’ inference
  - If we know that [Cl>F] in a particular case, it would increase the likelihood that [Br>F]

- Using known orderings of matched pairs, we can make improved inferences about other matched pairs
  - Not captured by matched pair analysis
- Matched (Molecular) Series

MATCHED SERIES OF LENGTH 2 = MP

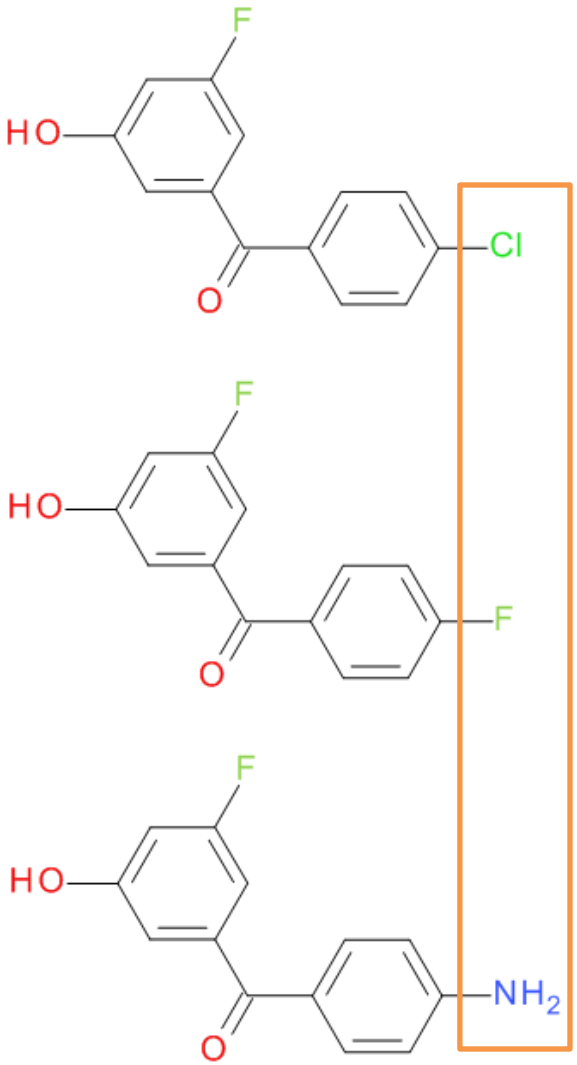


1.6

3.5

[Cl, F]

MATCHED SERIES OF LENGTH 3



1.6

3.5

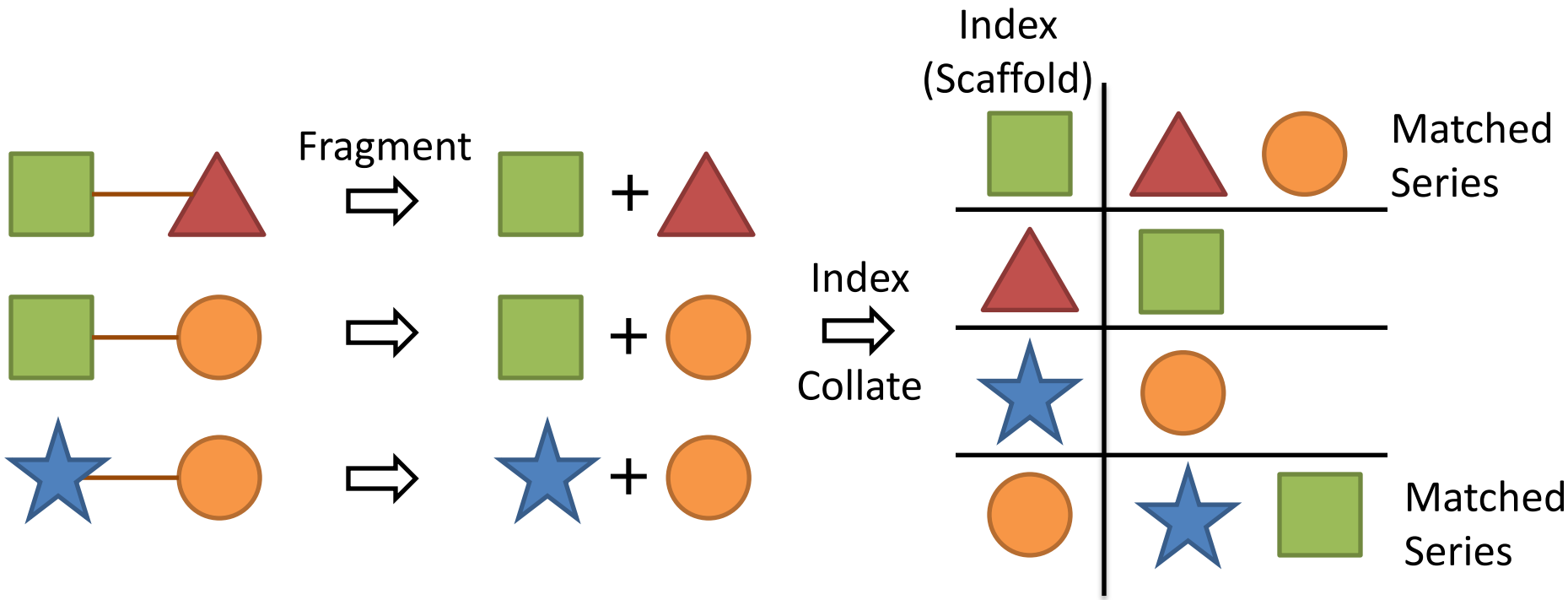
2.1

[Cl, F, NH_2]

MATCHED SERIES LITERATURE

- "Matching molecular series" introduced by Wawer and Bajorath *JMC* **2011**, *54*, 2944
  - Subsequent papers use MMS to investigate SAR transfer, mechanism hopping, visualisation of SAR networks and SAR matrices
- Only a single other paper on MMS
  - Mills et al *Med Chem Commun* **2012**, *3*, 174

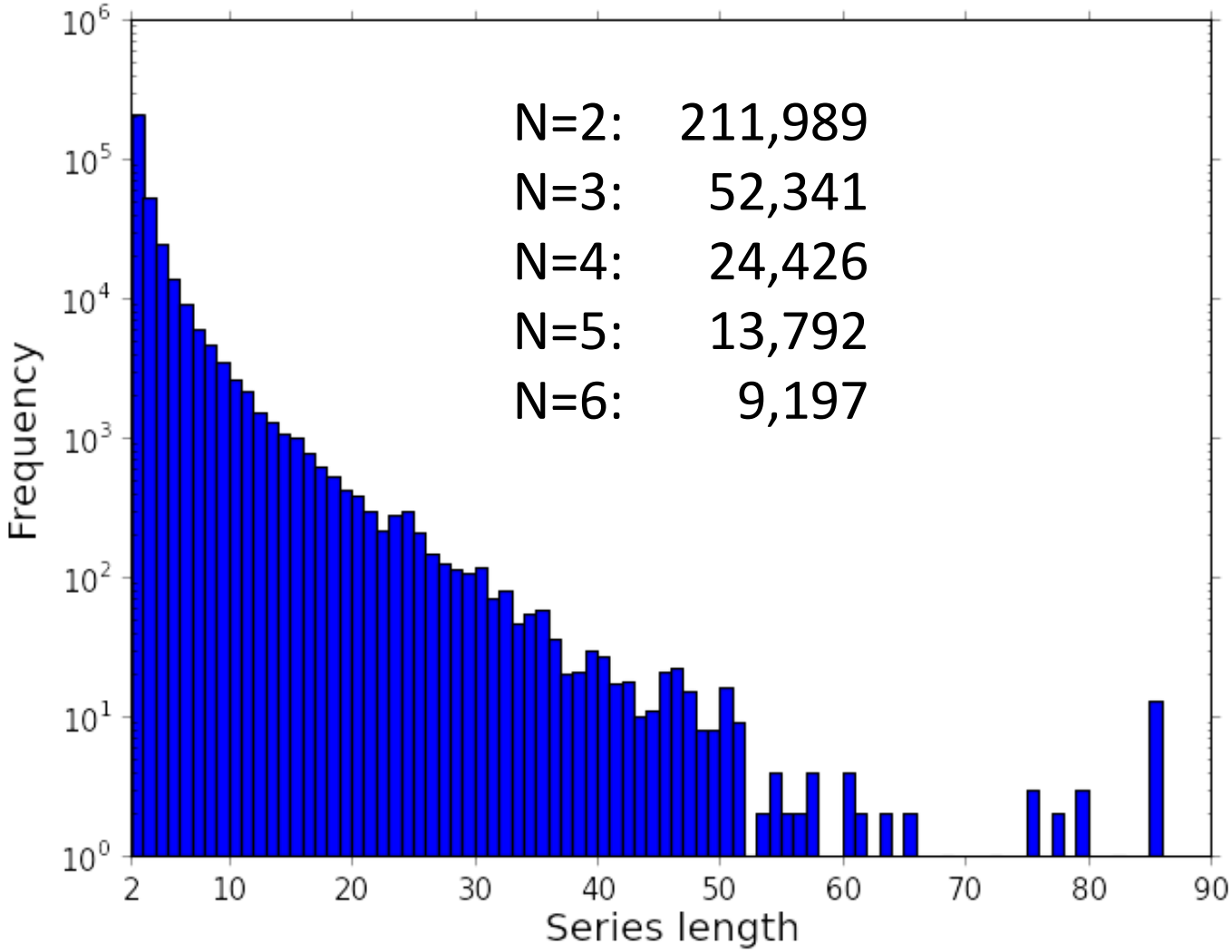
ALGORITHM TO FIND MATCHED SERIES



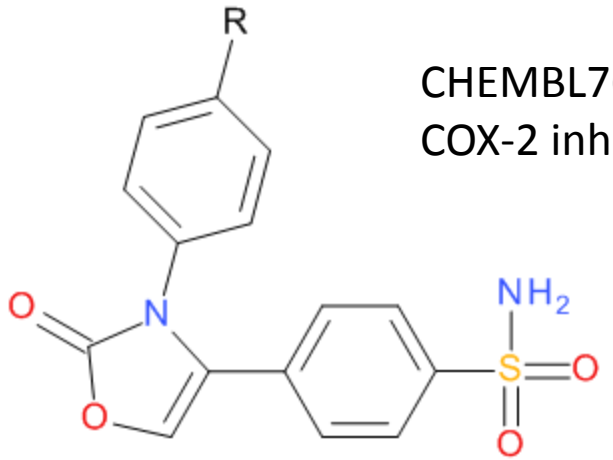
- Hussain and Rea *JCIM* **2010**, *50*, 339
  - Fragment molecules at acyclic single bonds
    - Single-cut only, scaffold >= 5, R group <= 12
  - Index each fragment based on the other
  - A matched series will be indexed together

DATASET

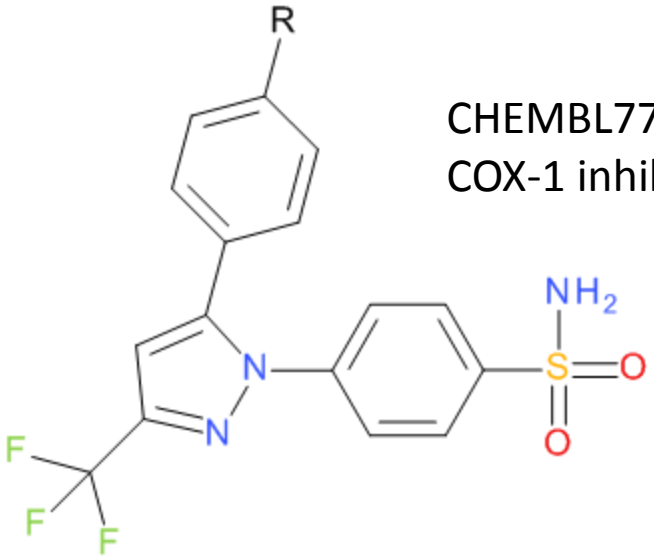
Matched series from ChEMBL16 IC_{50} binding assays



SAR TRANSFER



CHEMBL768956
COX-2 inhibition



CHEMBL772766
COX-1 inhibition

R Group	CHEMBL768956 (pIC_{50})	CHEMBL772766 (pIC_{50})
SMe	??	5.92
NH_2	??	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_3	5.70	<4.00
H	5.62	4.26
COOH	4.23	<3.60

Rank order

Potential SAR transfer

0.93 rank order correlation

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>…), there are N! ways of arranging the R Groups:

Series	Observations*
F > H	8250
H > F	7338

Would expect 7794 for each assuming the order is random

- We can calculate enrichment

*Dataset is ChEMBL16 IC_{50} data for binding assays (transformed to pIC_{50} values)

PREFERRED ORDERS: HALIDES (N=2)

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

Series	Enrichment	Observations
F > H	1.06*	8250
H > F	0.94*	7338

Would expect 7794 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	1.85*	1185
H > F > Cl	1.08	690
F > Cl > H	0.88*	566
Cl > H > F	0.79*	504
F > H > Cl	0.78*	503
H > Cl > F	0.63*	401

PREFERRED ORDERS: HALIDES (N=4)

Series	Enrichment	Observations
Br > Cl > F > H	5.62*	230
Cl > Br > F > H	2.79*	114
H > F > Cl > Br	**1.69***	**69**
F > Cl > Br > H	1.47	60
Br > Cl > H > F	1.39	57
Cl > Br > H > F	0.88	36
…	…	…
H > F > Br > Cl	**0.73**	**30**
…	…	…
Cl > H > F > Br	0.49*	20
H > Br > F > Cl	**0.49***	**20**
Cl > H > Br > F	0.46*	19
Br > F > H > Cl	0.44*	18
H > Cl > Br > F	0.44*	18
F > H > Br > Cl	0.42*	17
H > Cl > F > Br	0.37*	15
F > Br > H > Cl	0.34*	14
Br > H > F > Cl	**0.22***	**9**

N=2: Max = 1.06, Min = 0.94
N=3: Max = 1.85, Min = 0.63
N=4: Max = 5.62, Min = 0.22

Longer series exhibit greater preferences

If [H>F>Cl] is observed, will Br increase activity further?
128 observations of [H>F>Cl] but only 9 where [Br>H>F>Cl]

Don't forget sampling bias

MATSY:
PREDICTION USING MATCHED SERIES

FIND R GROUPS THAT INCREASE ACTIVITY

ChEMBL

Query

A > B

MATSY

A > **B** > C

C > **A** > **B**

D > **A** > **B** > C

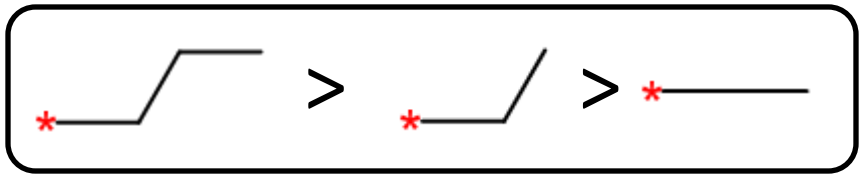
D > **A** > C > **B**

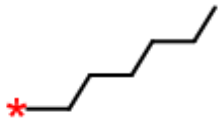
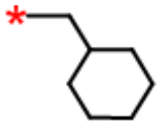
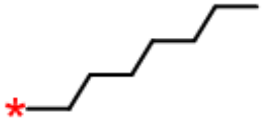
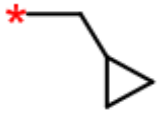
E > D > **A** > **B**

...

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

EXAMPLE

Query: 

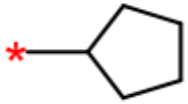
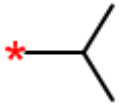
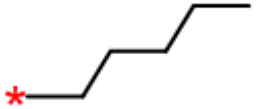
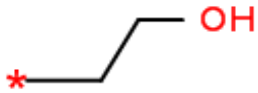
R Group	Observations	% that increase activity
	53	**75**
	28	**71**
	22	**63**
	41	**58**
	36	**58**

40 proteins including:

22 GPCRs (muscarinic acetylcholine, glucagon, endothelin, angiotensin)
5 oxidoreductases (cytochrome P450, cyclooxygenase)
3 acyltransferases
3 hydrolases

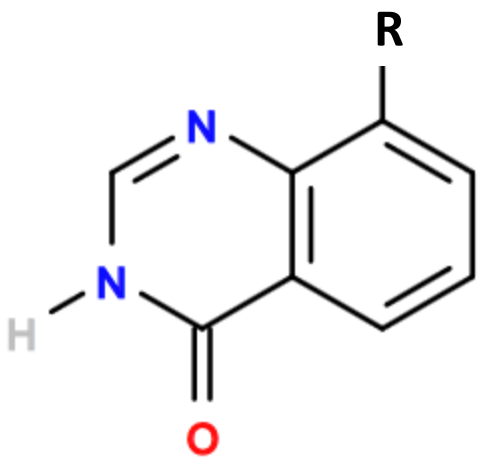
EXAMPLE

Query: 

R Group	Observations	% that increase activity
	23	**39**
	24	**37**
	97	**35**
	21	**33**
	21	**33**

9 proteins including:

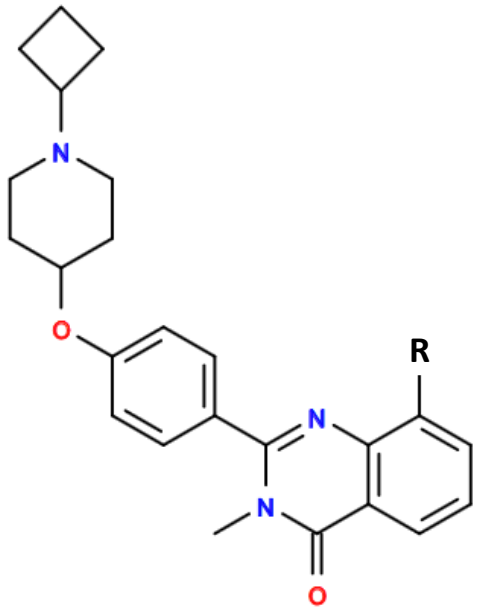
3 proteases (HIV-1, cathepsin K)
2 kinases (serine/threonine protein kinase ATR, CDK2)
1 GPCR



CHEMBL1953234
PARP-1 inhibition (Poly[ADP-Ribose] Polymerase 1)
[Me>Cl>H>F>CF_3]

Remove most active and predict:
[?>Cl>H>F>CF_3]

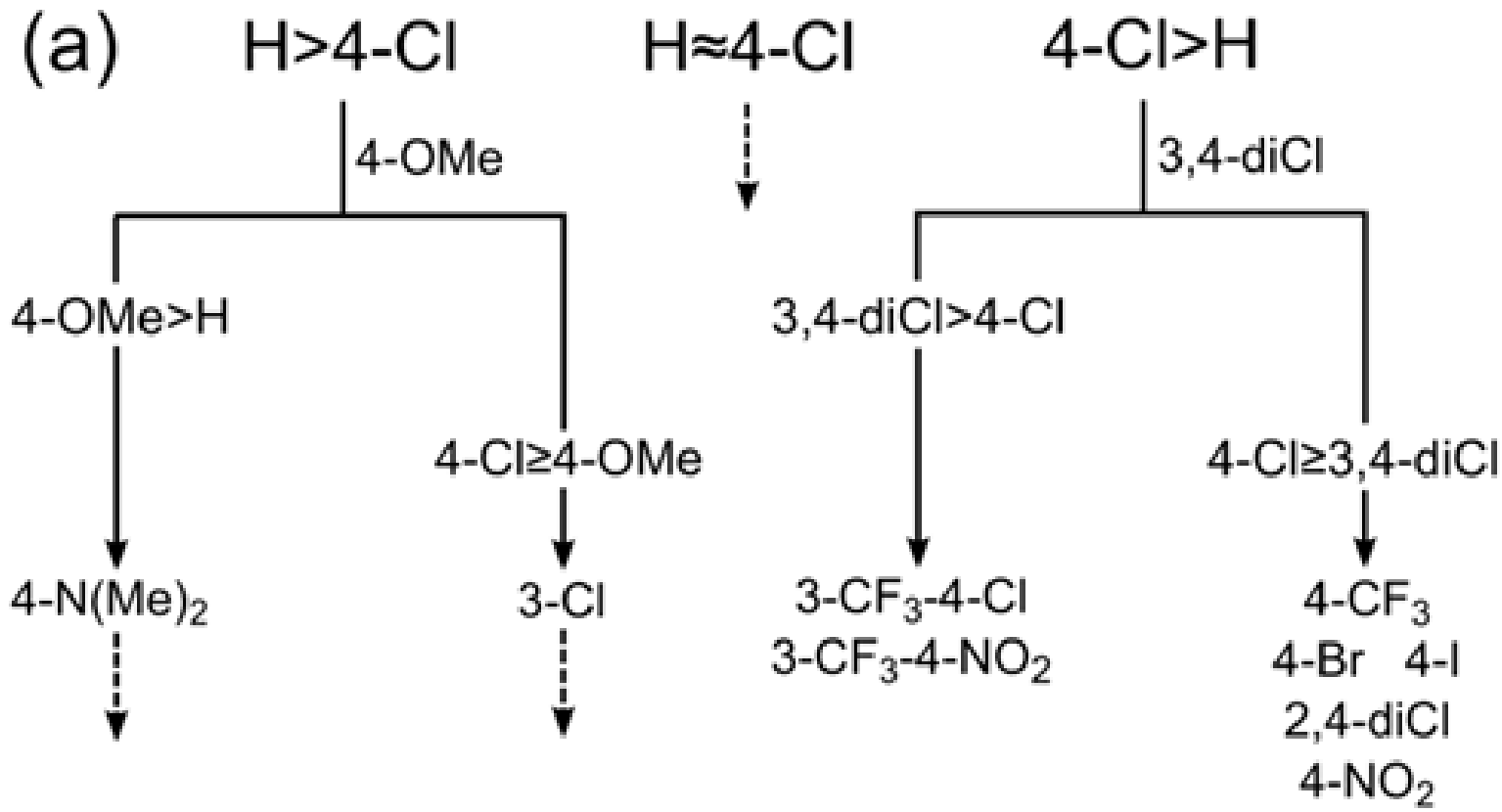
Prediction ranked Me as 2nd most likely, on the basis of 23 observations of which 7 (30%) showed improvement



CHEMBL956577
Inverse agonist at Histamine H3 receptor
[Me>Cl>H>F>CF_3]

TOPLISS DECISION TREE

RATIONAL STEPWISE SCHEME FOR SUBSTITUTED PHENYL



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

Bioorganic & Medicinal Chemistry

Volume 19, Issue 16, 15 August 2011, Pages 5031–5038

Novel benzofuroxan derivatives against multidrug-resistant *Staphylococcus aureus* strains: Design using Topliss' decision tree, synthesis and biological assay

Salomão Dória Jorge[a], Fanny Palace-Berl[a], Andrea Masunari[b], Cléber André Cechinel[c], Marina Ishii[a], Kerly Fernanda Mesquita Pasqualoto[a], Leoberto Costa Tavares[a]

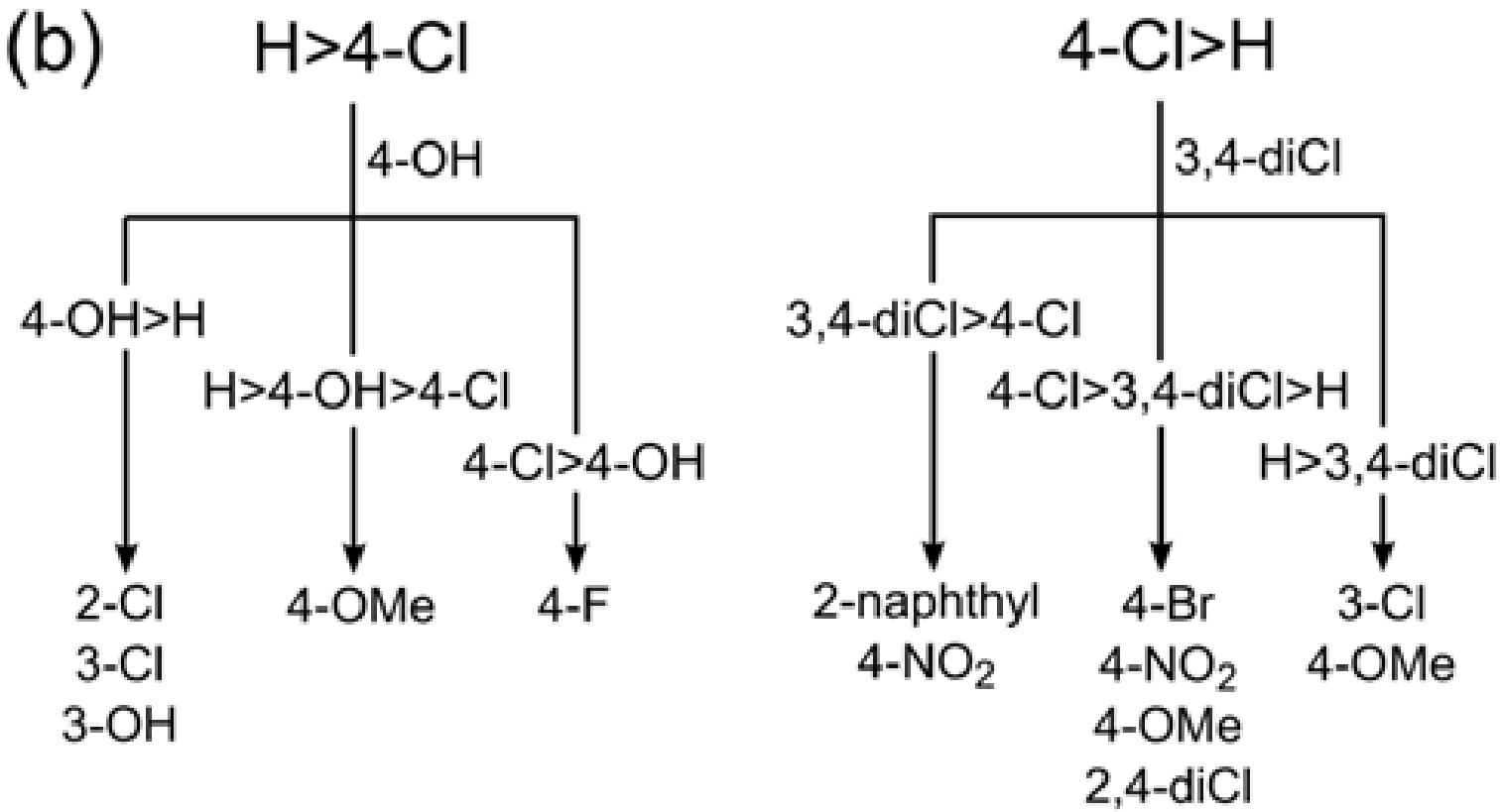
Bioorganic & Medicinal Chemistry Letters

Volume 21, Issue 21, 1 November 2011, Pages 6523–6526

Synthesis and preliminary biological evaluation of novel *N*-(3-aryl-1,2,4-triazol-5-yl) cinnamamide derivatives as potential antimycobacterial agents: An operational Topliss Tree approach

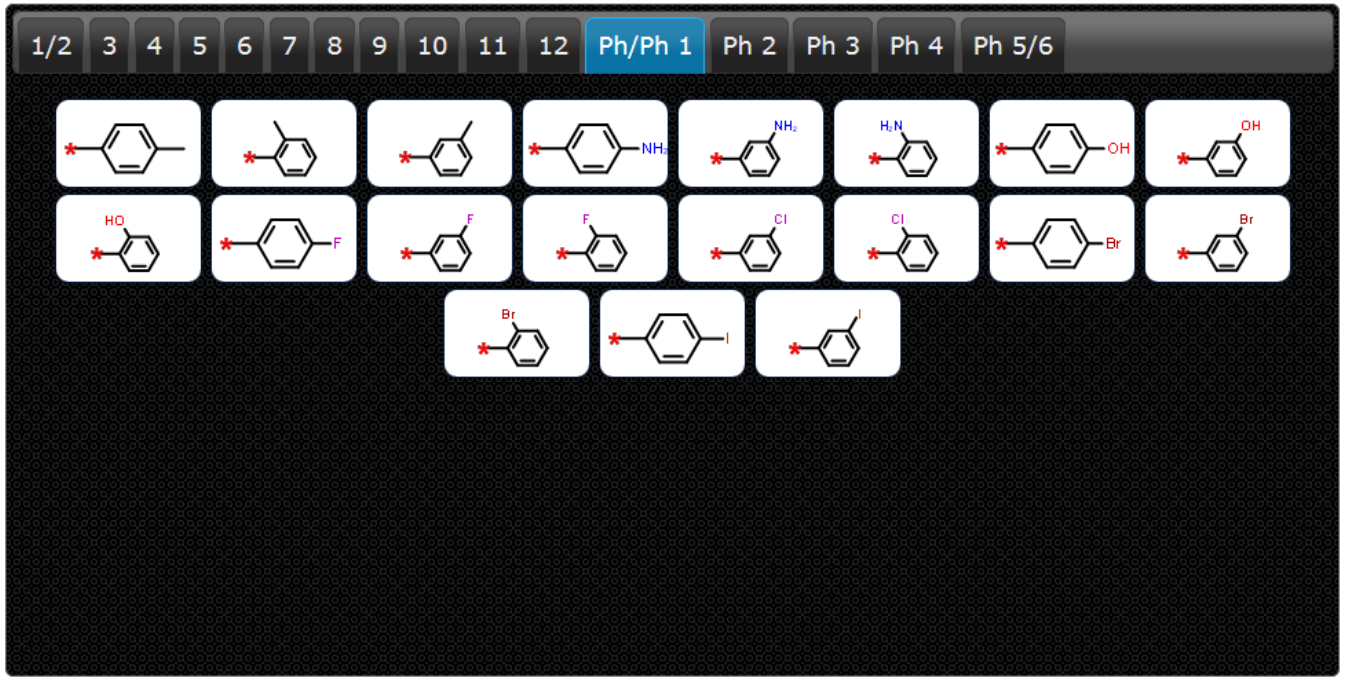
Manoj D. Kakwani[a], Nutan H. Palsule Desai[a], Arundhati C. Lele[a], Muktikant Ray[b], M.G.R. Rajan[b], Mariam S. Degani[a]

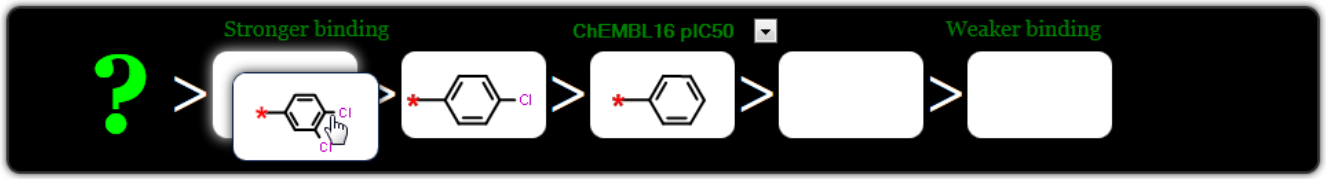
DATA-DRIVEN STEPWISE SCHEME FOR SUBSTITUTED PHENYL

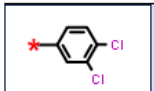
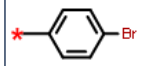
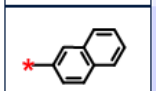
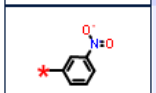
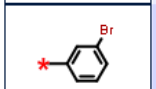
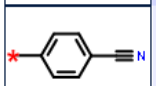
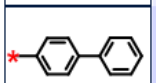
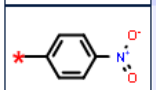
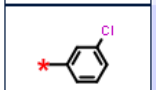
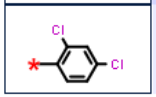


Using Matsy and ChEMBL 16 IC_{50} binding data

DEMO OF DRAG-AND-DROP INTERFACE





	% >	% >=	Counts
	54	56	326
	52	56	431
	48	51	186
	44	48	148
	44	48	124
	39	40	215
	39	40	141
	38	39	296
	37	40	556
	36	39	157

Showing 1 to 10 of 34 entries

Previous Next

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
  - Data-hungry

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

http://nextmovesoftware.com
noel@nextmovesoftware.com
@nmsoftware