3rd RDKit Meeting
Merck, Darmstadt, Oct 2014

Using RDKit for Matched Molecular Series Analysis

When two are not enough

Noel O'Boyle

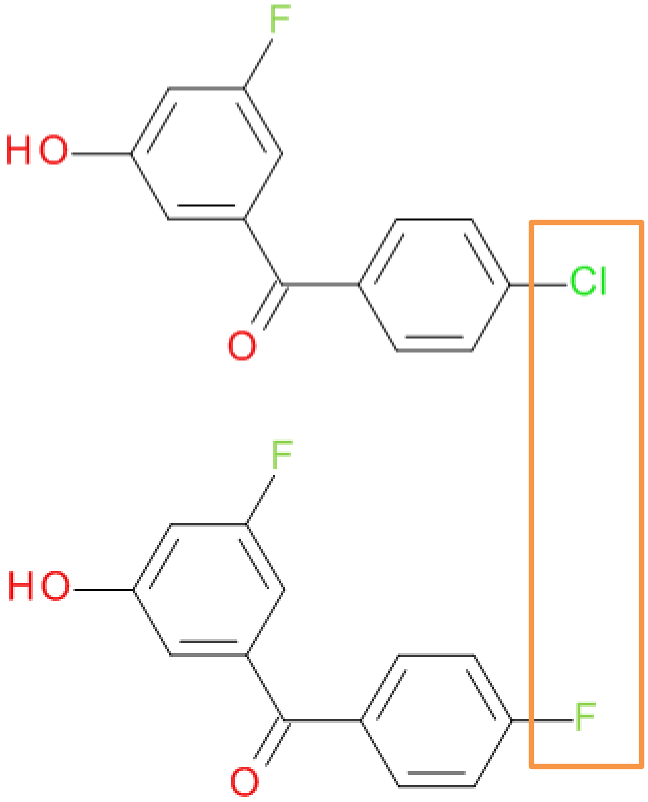
NextMove Software

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



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

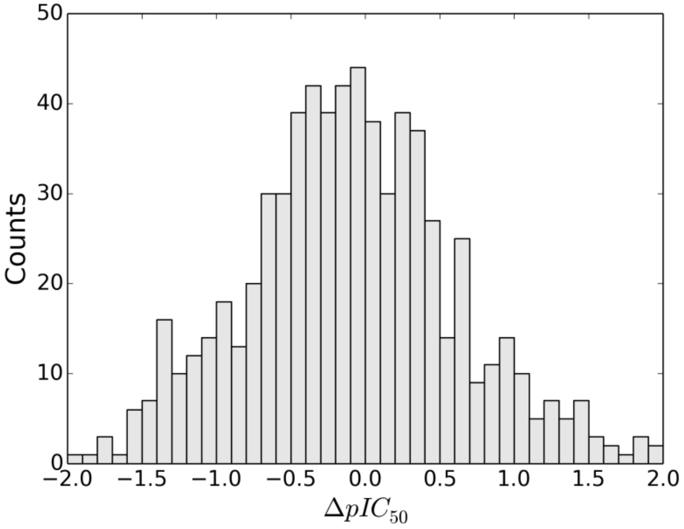
* Chemoinformatics in drug discovery, Wiley, 271–285.

MATCHED PAIR USAGE

- Successfully used for:
  - 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

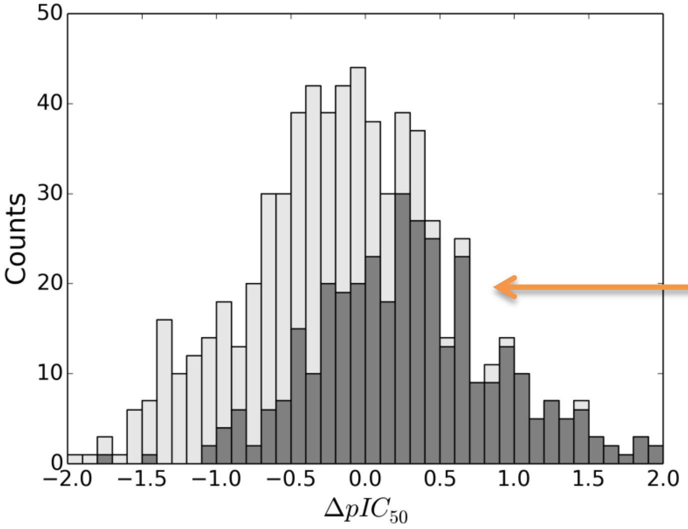
MATCHED PAIRS AND ACTIVITY

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



MATCHED PAIRS AND ACTIVITY

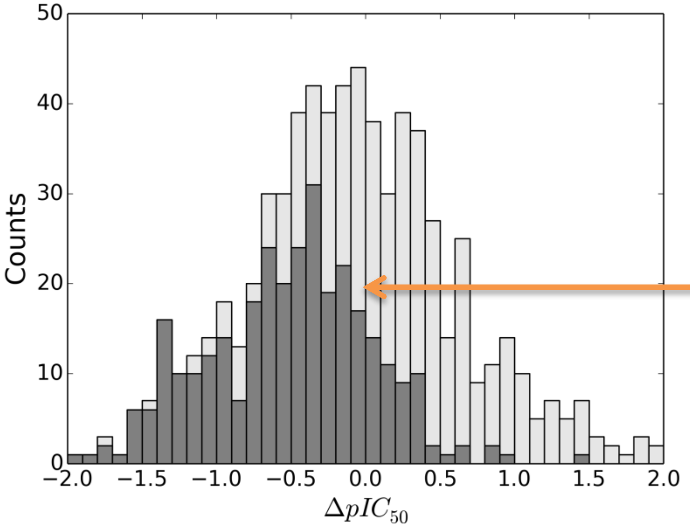
$pIC_{50}(CC)-pIC_{50}(CCCC)$



For those cases where:
[CCC > CCCC]

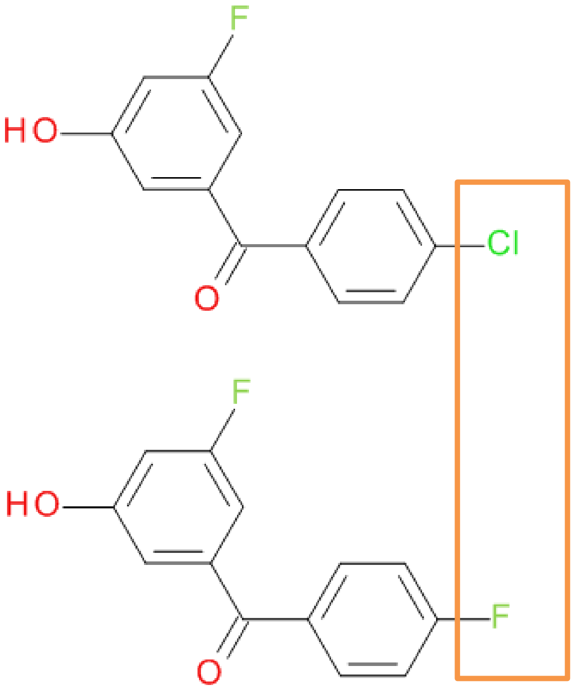
MATCHED PAIRS AND ACTIVITY

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



For those cases where:
[CCC < CCCC]

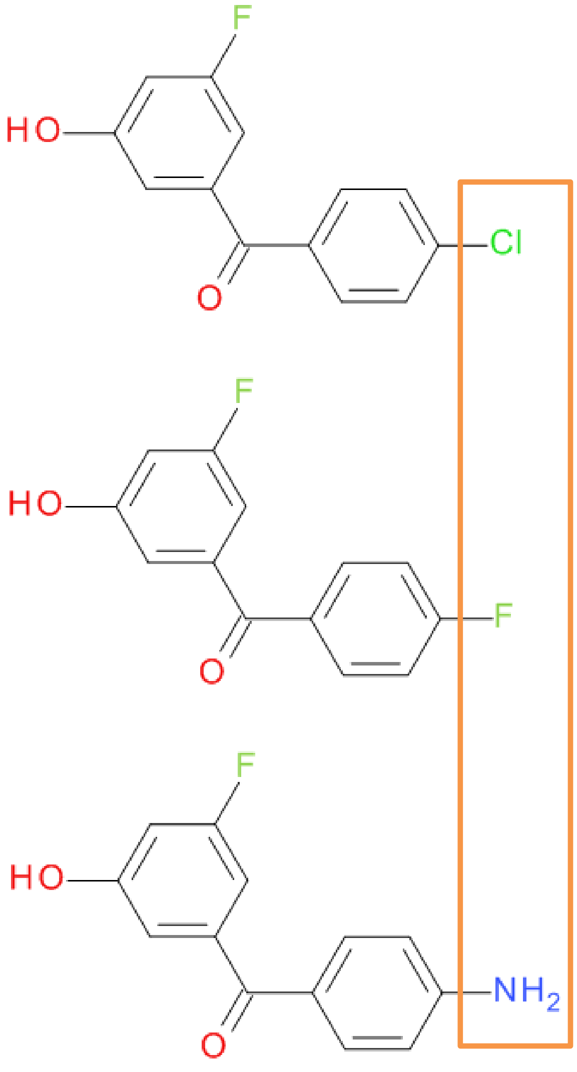
Matched series of length 2 = matched pair



[Cl, F]

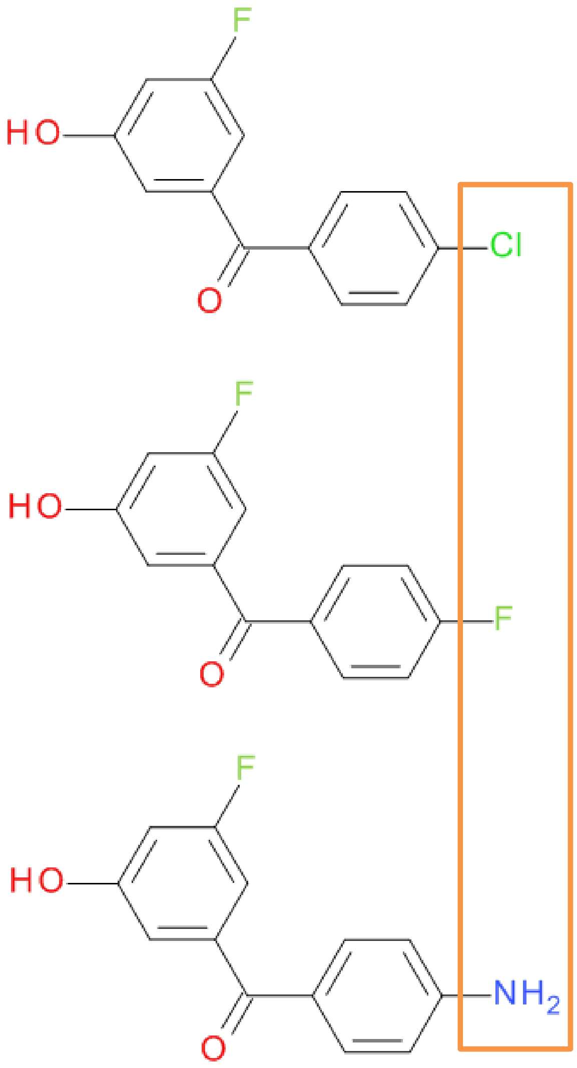
“Matching molecular series” introduced by Wawer and Bajorath, *J. Med. Chem.* **2011**, *54*, 2944

MATCHED SERIES OF LENGTH 3



[Cl, F, NH_2]

ORDERED MATCHED SERIES OF LENGTH 3



pIC_{50}

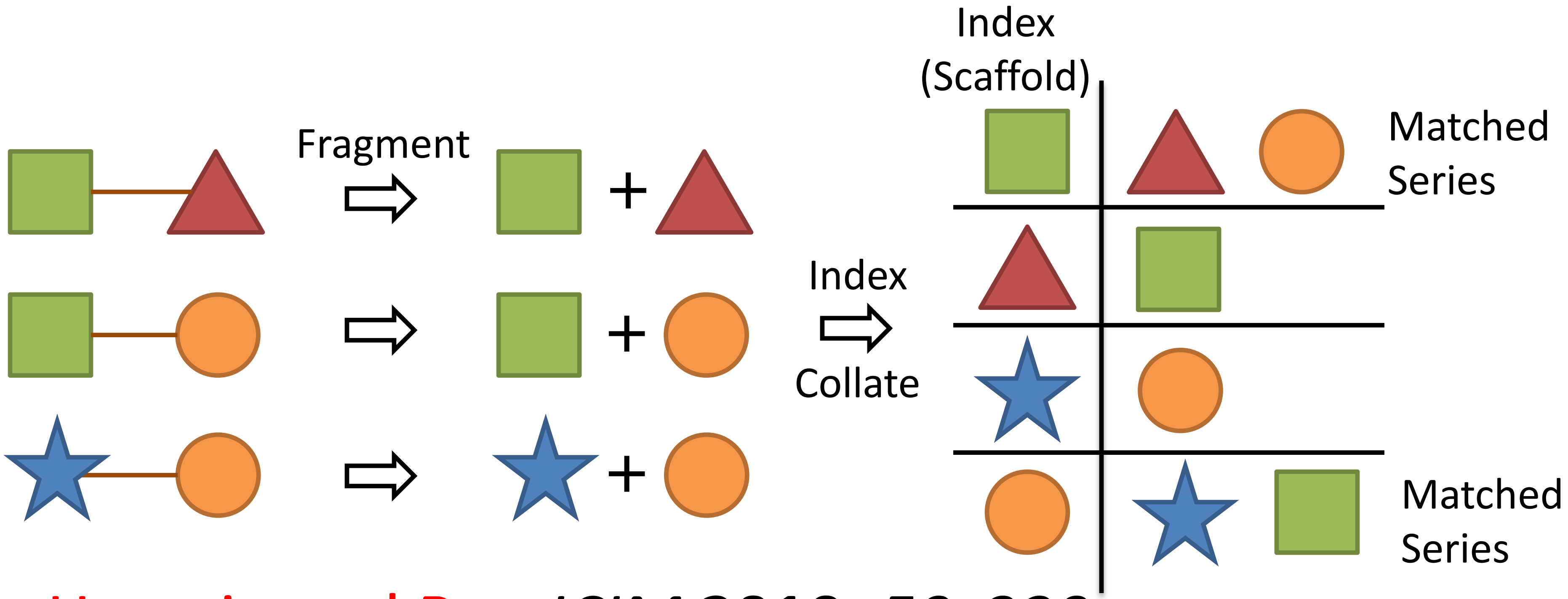
3.5

2.1

1.6

[Cl > F > NH_2]

ALGORITHM TO FIND MATCHED SERIES



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

FIND MATCHED SERIES IN JAMEED'S FRAGMENTS

```
import rdk
import sys
from collections import namedtuple
Frag = namedtuple('Frag', ['id', 'scaffold', 'rgroup'])

class Series():
    def __init__(self):
        self.rgroups = []
        self.scaffold = ""

def getFrags(filename):
    frags = []
    for line in open(filename):
        broken = line.rstrip().split(",")
        if broken[2]: # should be blank for single-cut
            continue
        smiles = broken[-1].split(".")
        mols = [rdk.readstring("smi", smi) for smi in smiles]
        numAtoms = [mol.Mol.GetNumAtoms() for mol in mols]

        if numAtoms[0] > 5 and numAtoms[1] < 12:
            frags.append(Frag(broken[1], smiles[0], smiles[1]))
        if numAtoms[1] > 5 and numAtoms[0] < 12:
            frags.append(Frag(broken[1], smiles[1], smiles[0]))
    frags.sort(key=lambda x:(x.scaffold, x.rgroup))
    return frags
```

```
def getSeries(frags):
    oldfrag = Frag(None, None, None)
    series = Series()
    for frag in frags:
        if frag.scaffold != oldfrag.scaffold:
            if len(series.rgroups)>=2:
                series.scaffold = oldfrag.scaffold
                yield series
            series = Series()
        series.rgroups.append( (frag.rgroup, frag.id) )

        oldfrag = frag
    if len(series.rgroups)>=2:
        series.scaffold = oldfrag.scaffold
        yield series

if __name__ == "__main__":
    filename = sys.argv[1]

    frags = getFrags(filename)
    it = getSeries(frags)

    for series in it:
        print "# %s" % series.scaffold
        for rgroup in sorted(series.rgroups):
            print "%s %s" % (rgroup[0], rgroup[1])
```

FIND MATCHED SERIES IN JAMEED'S FRAGMENTS

sample_fragmented.txt

```
# [*:1]CNc1ncnc2sccc21
[*:1]c1ccccc1 2139597
[*:1]c1cccnc1 2531831
# [*:1]Cn1nc(C)cc1C
[*:1]c1ccc(C(=O)O)cc1 615212
[*:1]c1ccc(C(=O)O)o1 658387
# [*:1]NC(=O)C1COc2ccccc2O1
[*:1]c1ccc(C(=O)O)cc1 2881039
[*:1]c1ccc(C(N)=O)cc1 2787356
....
```

pickett_fragmented.txt

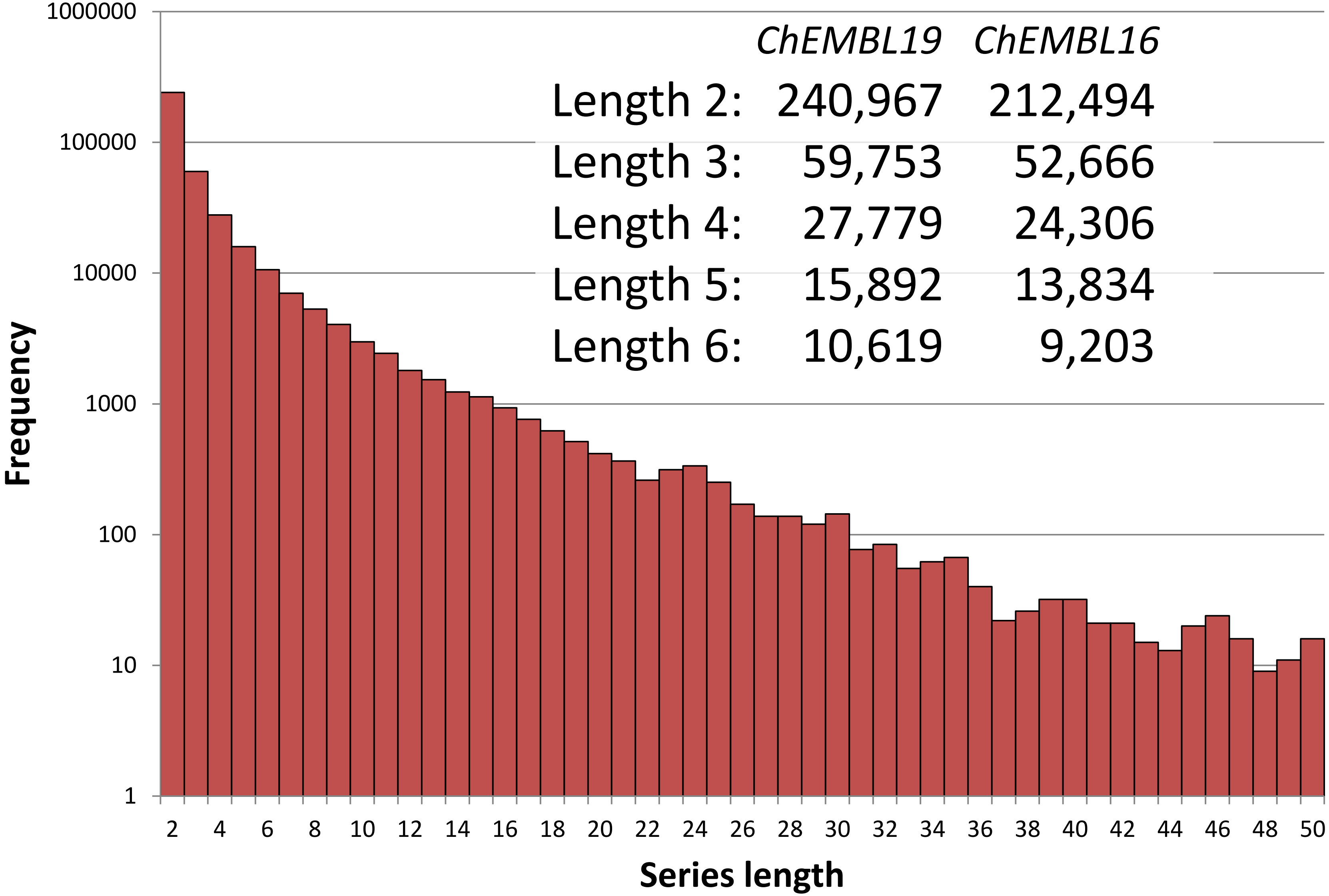
```
# [*:1]C[C@H](NS(=O)(=O)c1ccc(-c2c(C)cccc2C)cc1)C(=O)O
[*:1]C(=O)OC(C)(C)C A18B22
[*:1]C(C)C A04B22
[*:1]C(N)=O A29B22
[*:1]C1CC1 A42B22
[*:1]CC(=O)NC(C)C A41B22
[*:1]CC(=O)OC(C)(C)C A19B22
[*:1]CC(=O)OCC A14B22
[*:1]CCC A31B22
[*:1]CCCNC(C)=O A06B22
[*:1]CS(C)=O A24B22
[*:1]NC(N)=O A09B22
[*:1]OC(C)(C)C A33B22
[*:1]OCc1ccccc1 A32B22
[*:1]SC A02B22
[*:1]SCCc1ccncc1 A17B22
[*:1]SCc1ccccc1 A03B22
[*:1]Sc1ccccc1 A43B22
[*:1]Sc1cccs1 A50B22
[*:1]Sc1nccs1 A49B22
[*:1]c1c[nH]c2ccccc12 A28B22
[*:1]c1ccc(O)cc1 A01B22
[*:1]c1cccc(C#N)c1 A44B22
[*:1]c1ccccn1 A39B22
[*:1]c1cccs1 A27B22
[*:1]n1cccn1 A46B22
[*:1]n1cncn1 A47B22
```

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

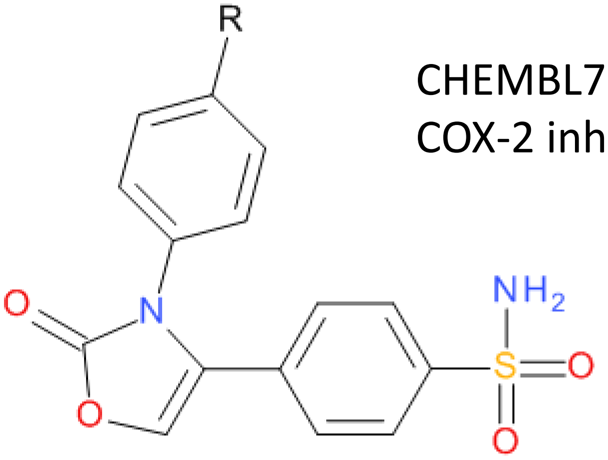
Gaulton et al. *Nucleic Acids Res.* **2012**, *40*, D1100

Matched series in ChEMBL19 IC50 binding assays

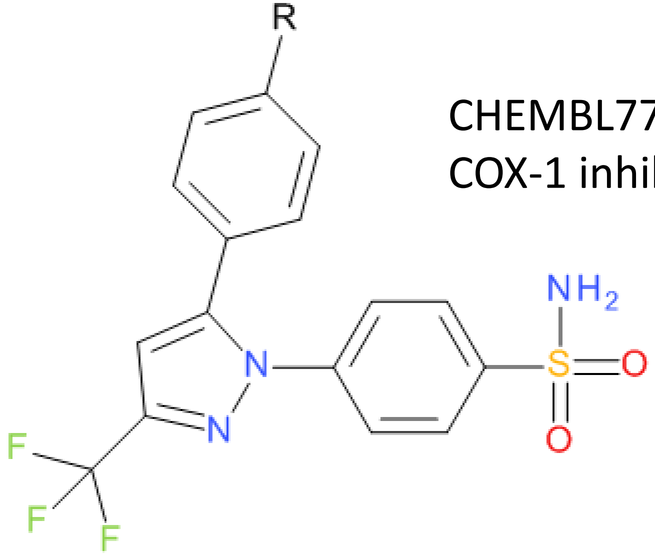


	ChEMBL19	*ChEMBL16*
Length 2:	240,967	212,494
Length 3:	59,753	52,666
Length 4:	27,779	24,306
Length 5:	15,892	13,834
Length 6:	10,619	9,203

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

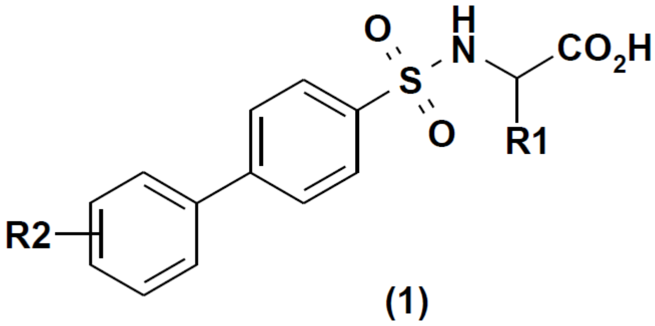
Rank order

Potential SAR transfer

0.93 rank order correlation

50x50 Matrix from Pickett et al.

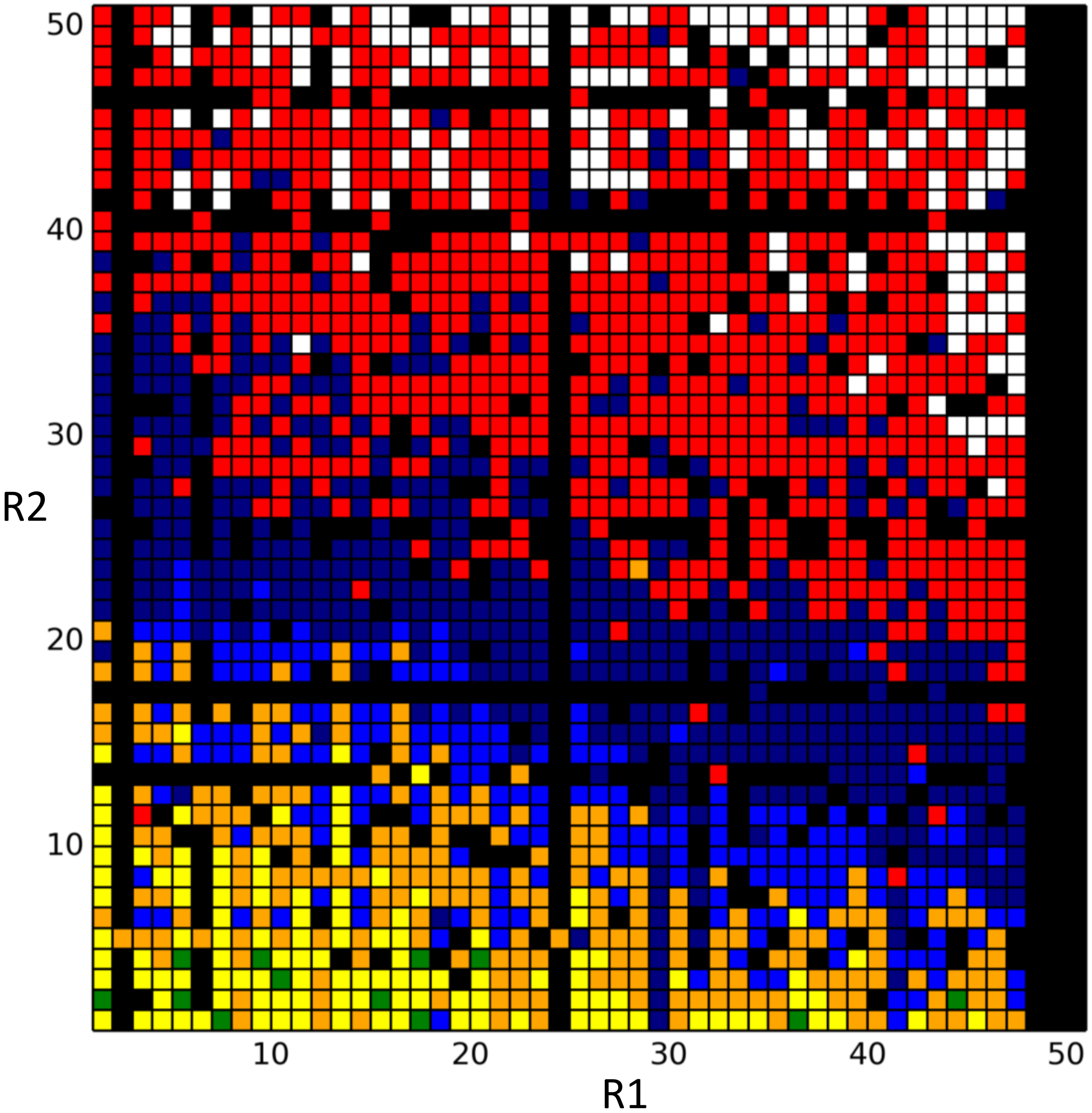
Pickett, Green, Hunt, Pardoe, Hughes. *ACS Med. Chem. Lett.* **2011**, *2*, 28.



(1)

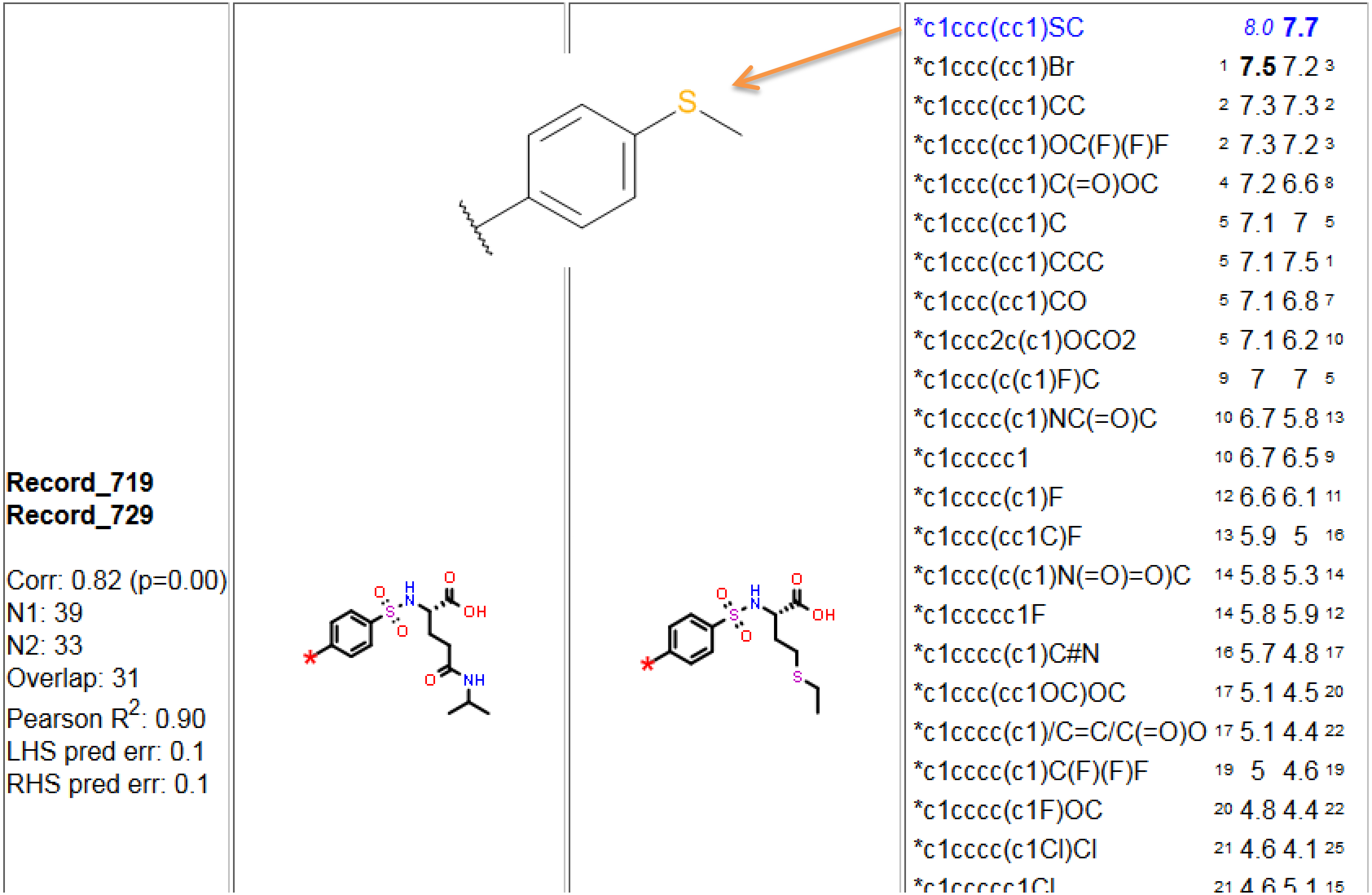
MMP-12 pIC50

- ⬛ Null
- ⬜ Inact
- 🟥 <5
- 🟦 5-6
- 🟦 6-6.5
- 🟧 6.5-7
- 🟨 7-7.5
- 🟩 >7.5



INTERNAL SAR TRANSFER

Do an all-against-all comparison of the series



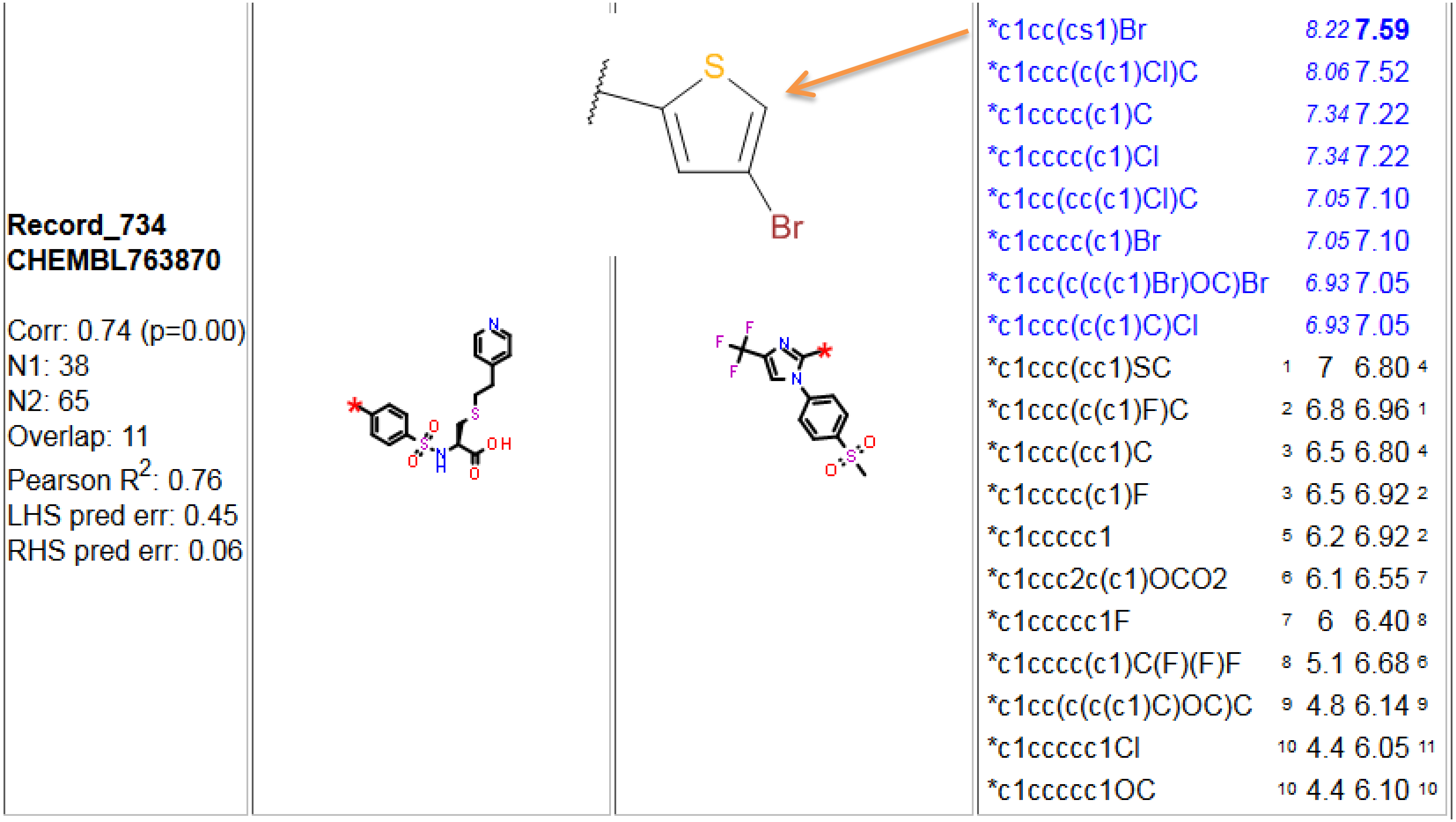
Record_719
Record_729

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

*c1ccc(cc1)SC		*8.0*	**7.7**	
*c1ccc(cc1)Br	1	**7.5**	7.2	3
*c1ccc(cc1)CC	2	7.3	7.3	2
*c1ccc(cc1)OC(F)(F)F	2	7.3	7.2	3
*c1ccc(cc1)C(=O)OC	4	7.2	6.6	8
*c1ccc(cc1)C	5	7.1	7	5
*c1ccc(cc1)CCC	5	7.1	7.5	1
*c1ccc(cc1)CO	5	7.1	6.8	7
*c1ccc2c(c1)OCO2	5	7.1	6.2	10
*c1ccc(c(c1)F)C	9	7	7	5
*c1cccc(c1)NC(=O)C	10	6.7	5.8	13
*c1ccccc1	10	6.7	6.5	9
*c1cccc(c1)F	12	6.6	6.1	11
*c1ccc(cc1C)F	13	5.9	5	16
*c1ccc(c(c1)N(=O)=O)C	14	5.8	5.3	14
*c1ccccc1F	14	5.8	5.9	12
*c1cccc(c1)C#N	16	5.7	4.8	17
*c1ccc(cc1OC)OC	17	5.1	4.5	20
*c1cccc(c1)/C=C/C(=O)O	17	5.1	4.4	22
*c1cccc(c1)C(F)(F)F	19	5	4.6	19
*c1cccc(c1F)OC	20	4.8	4.4	22
*c1cccc(c1Cl)Cl	21	4.6	4.1	25
*c1ccccc1Cl	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^2: 0.76
LHS pred err: 0.45
RHS pred err: 0.06

*c1cc(cs1)Br		8.22	**7.59**	
*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
*c1ccccc1	5	6.2	6.92	2
*c1ccc2c(c1)OCO2	6	6.1	6.55	7
*c1ccccc1F	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
*c1ccccc1Cl	10	4.4	6.05	11
*c1ccccc1OC	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>…), 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_{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*	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
H > F > Cl > Br	**1.59***	**77**
Br > Cl > H > F	1.43	69
F > Cl > Br > H	1.40	68
Cl > Br > H > F	0.85	41
…	…	…
H > F > Br > Cl	**0.76**	**37**
…	…	…
H > Br > F > Cl	**0.50***	**24**
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
Br > H > F > Cl	**0.23***	**11**

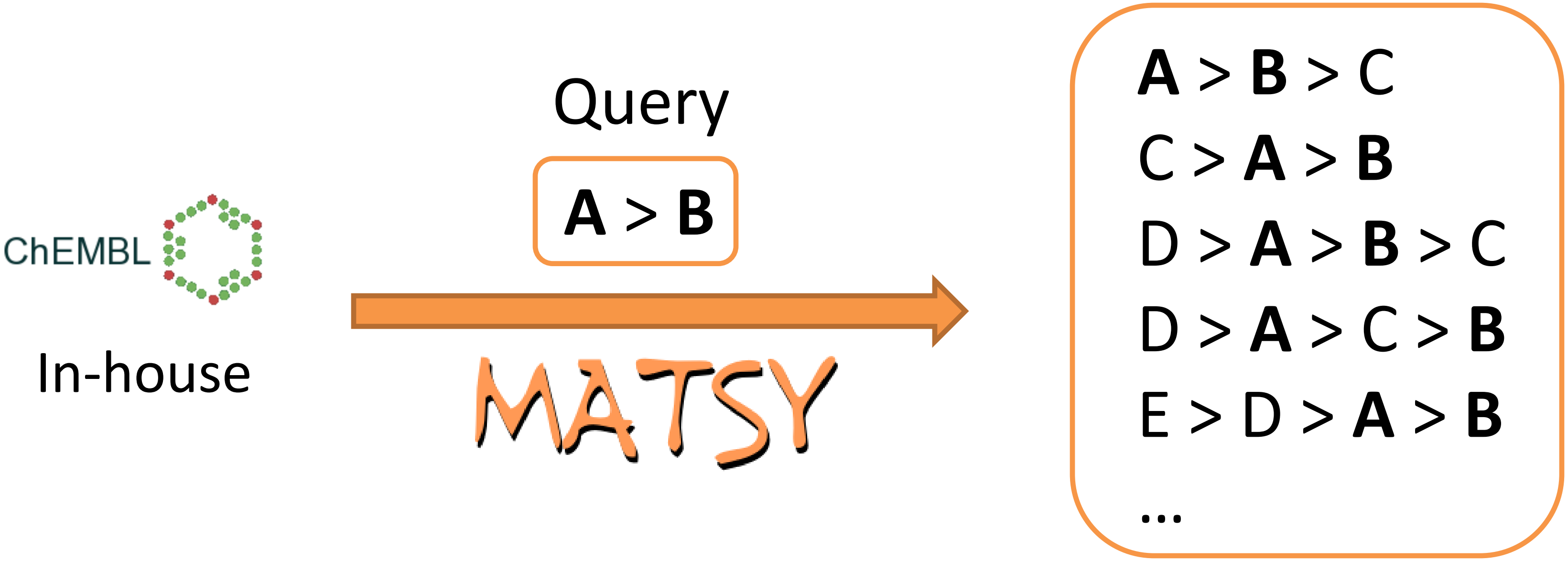
N=2: Max = 1.06, Min = 0.94
N=3: Max = 1.90, Min = 0.63
N=4: Max = 5.43, Min = 0.232

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



ChEMBL

In-house

Query

A > **B**

MATSY

A > **B** > C

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

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

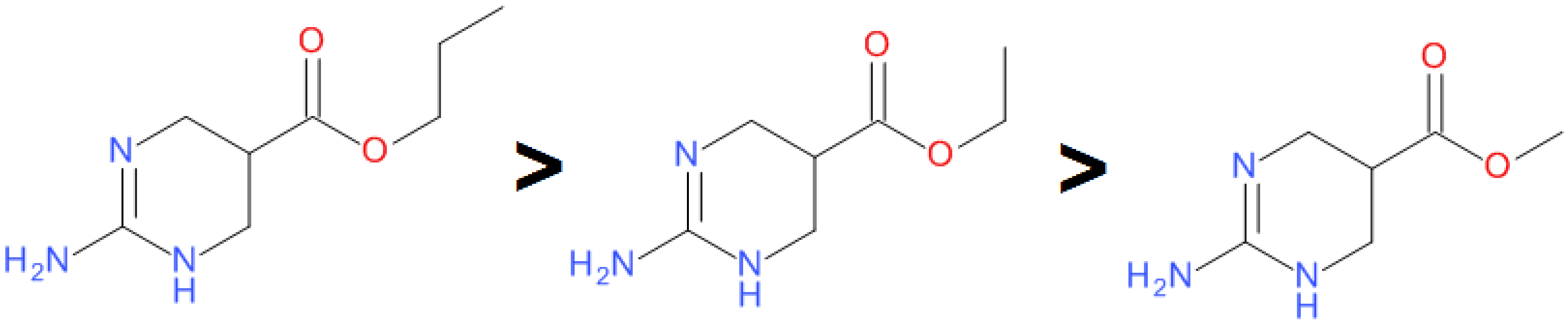
D > **A** > C > **B**

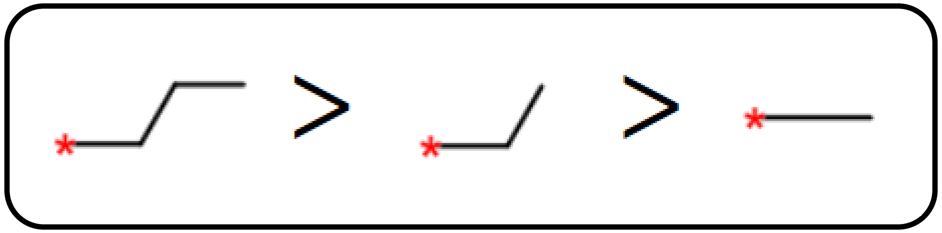
E > D > **A** > **B**

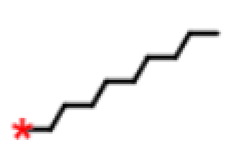
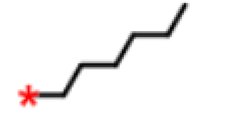
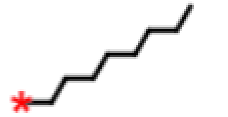
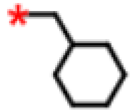
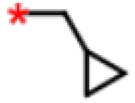
…

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

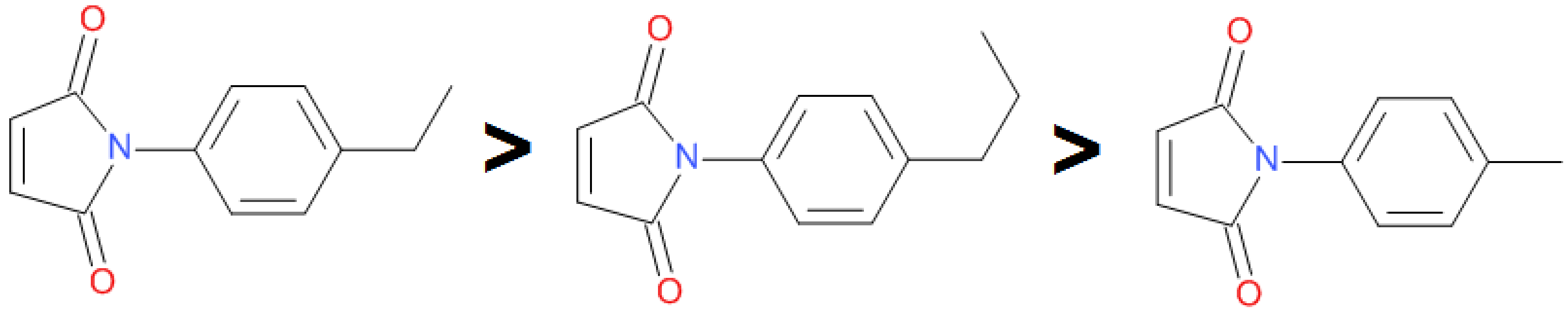
EXAMPLE

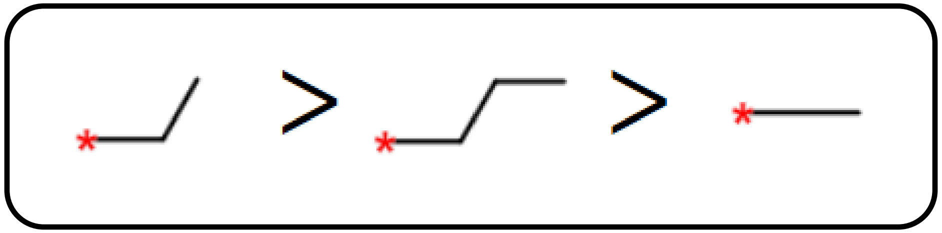


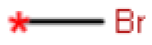
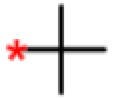
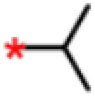
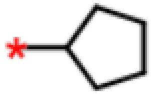
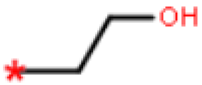


	% >	Counts	ΔLogP
	90	21	+3.3
	72	60	+1.7
	69	32	+2.8
	63	27	+1.6
	60	40	-0.1

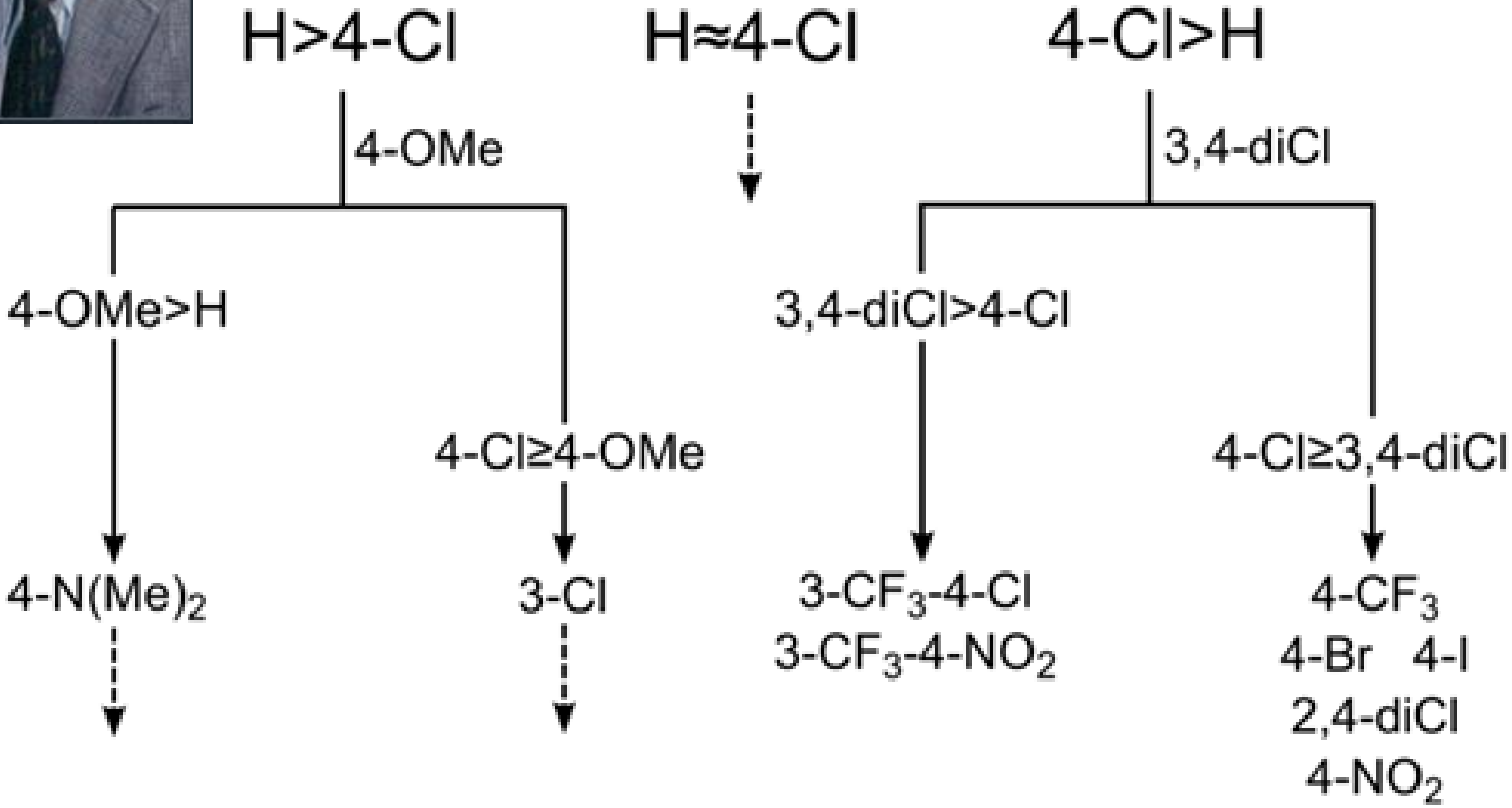
EXAMPLE II





	% >	Counts	ΔLogP
	38	21	**-0.8**
	37	27	+0.9
	33	**111**	+0.3
	33	27	+1.0
	33	21	**-1.6**

Topliss Decision Tree



H>4-Cl

H≈4-Cl

4-Cl>H

4-OMe

3,4-diCl

4-OMe>H

4-Cl≥4-OMe

3,4-diCl>4-Cl

4-Cl≥3,4-diCl

4-N(Me)$_2$

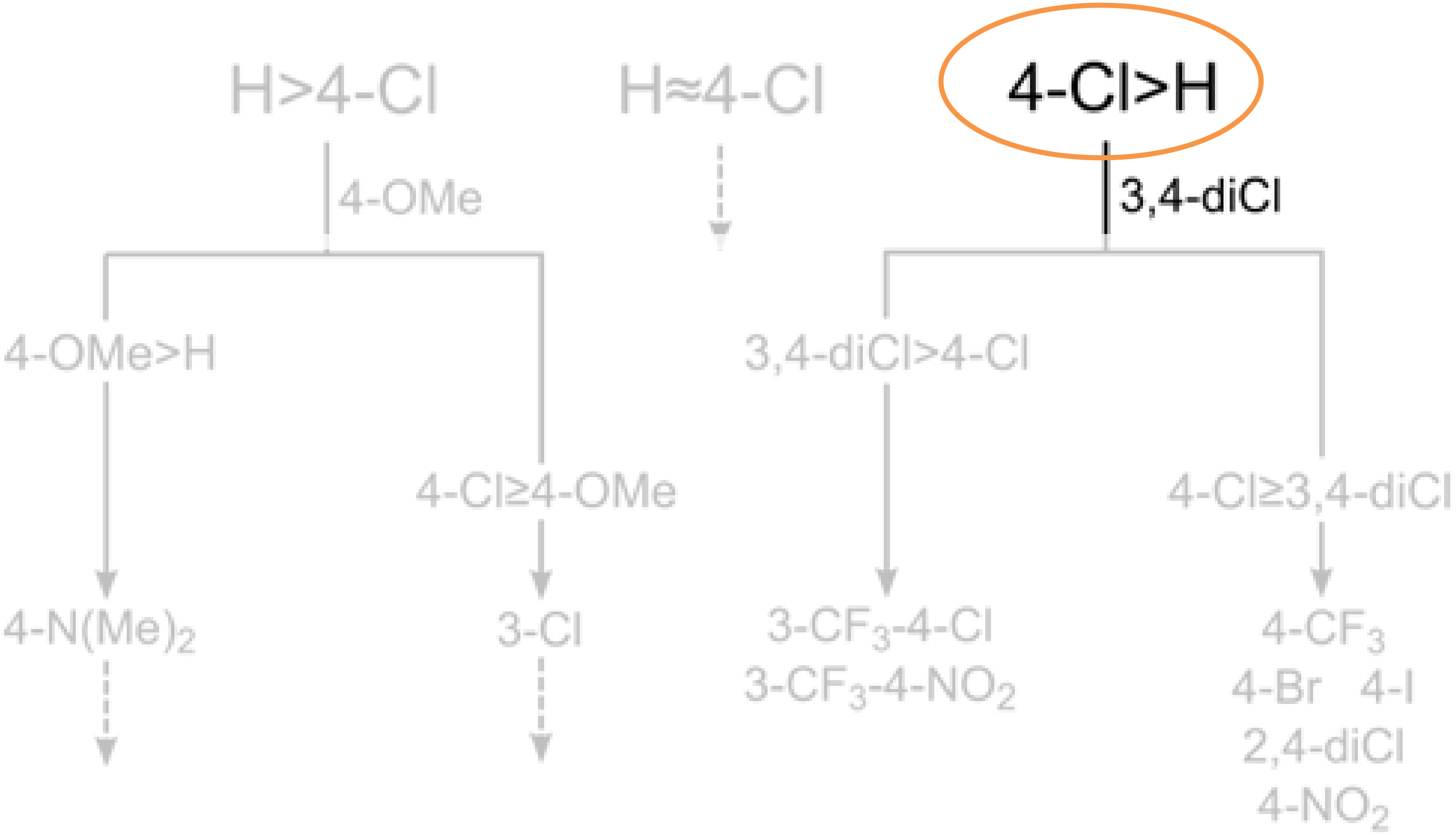
3-Cl

3-CF$_3$-4-Cl
3-CF$_3$-4-NO$_2$

4-CF$_3$
4-Br 4-I
2,4-diCl
4-NO$_2$

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

TOPLISS DECISION TREE



H>4-Cl

H≈4-Cl

4-Cl>H

4-OMe

3,4-diCl

4-OMe>H

4-Cl≥4-OMe

3,4-diCl>4-Cl

4-Cl≥3,4-diCl

4-N(Me)$_2$

3-Cl

3-CF$_3$-4-Cl

3-CF$_3$-4-NO$_2$

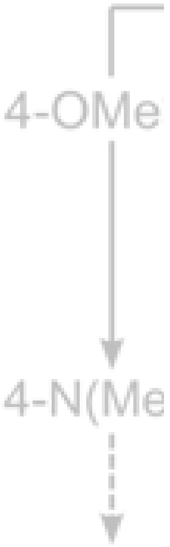
4-CF$_3$

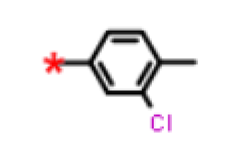
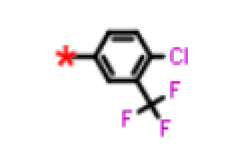
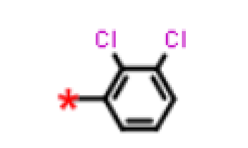
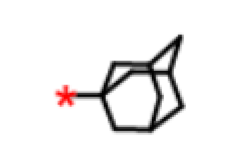
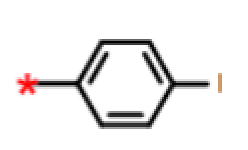
4-Br 4-I

2,4-diCl

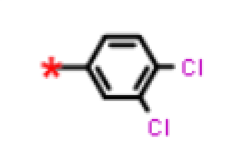
4-NO$_2$

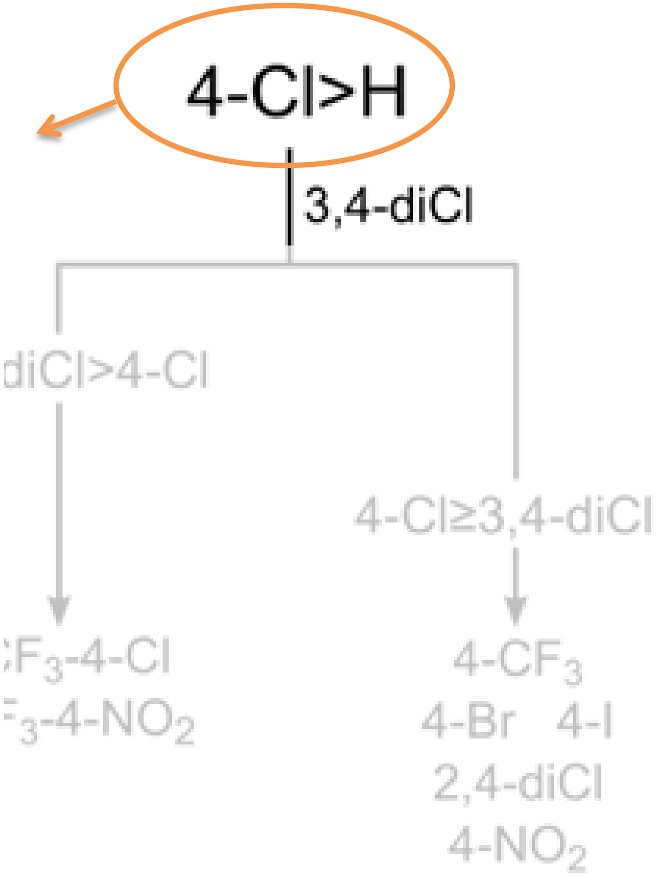
TOPLISS DECISION TREE



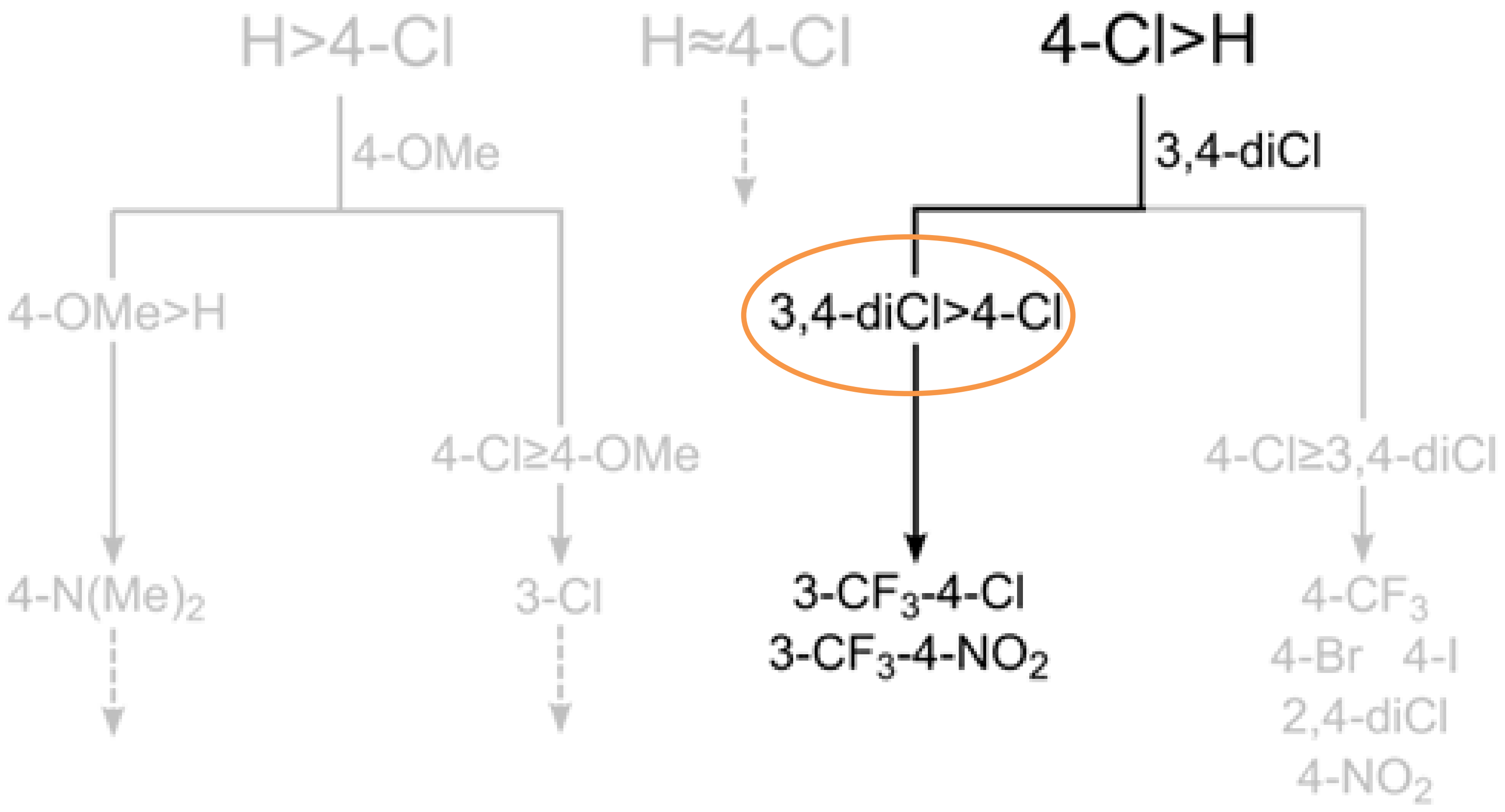
	% >	Counts	ΔLogP
	74	27	+0.3
	71	28	+0.9
	61	51	+0.6
	61	23	+1.4
	58	103	+0.5

(11th)

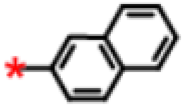
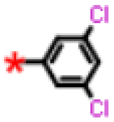
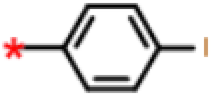
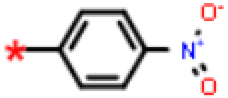
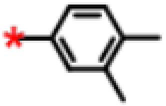
	54	391	+0.6

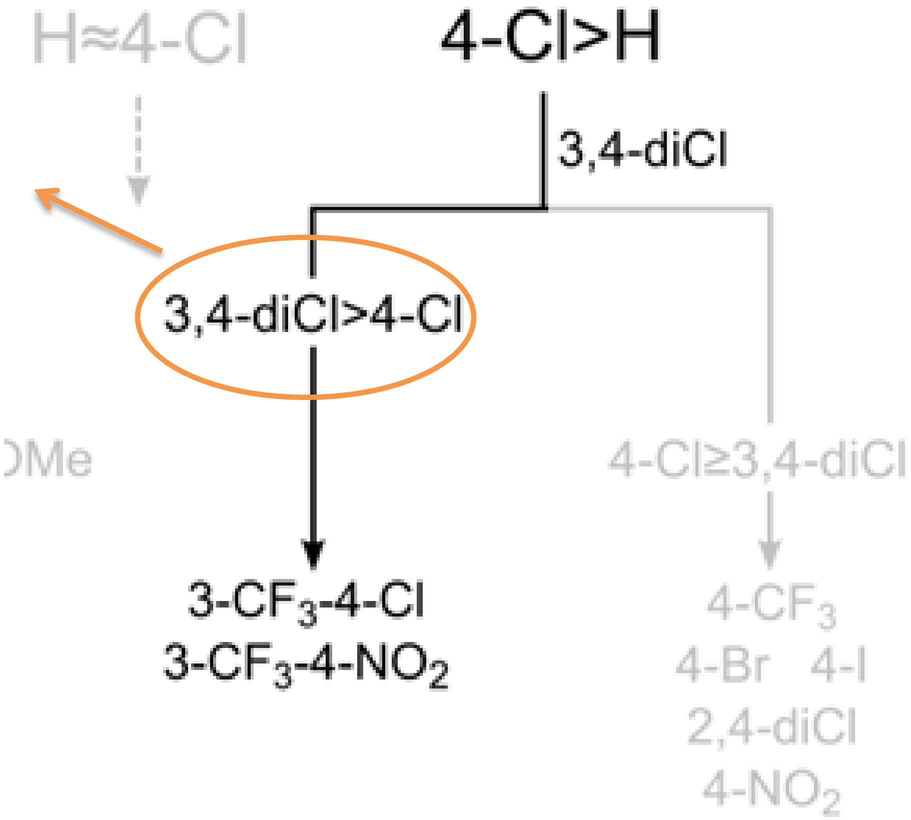


TOPLISS DECISION TREE

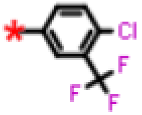


TOPLISS DECISION TREE

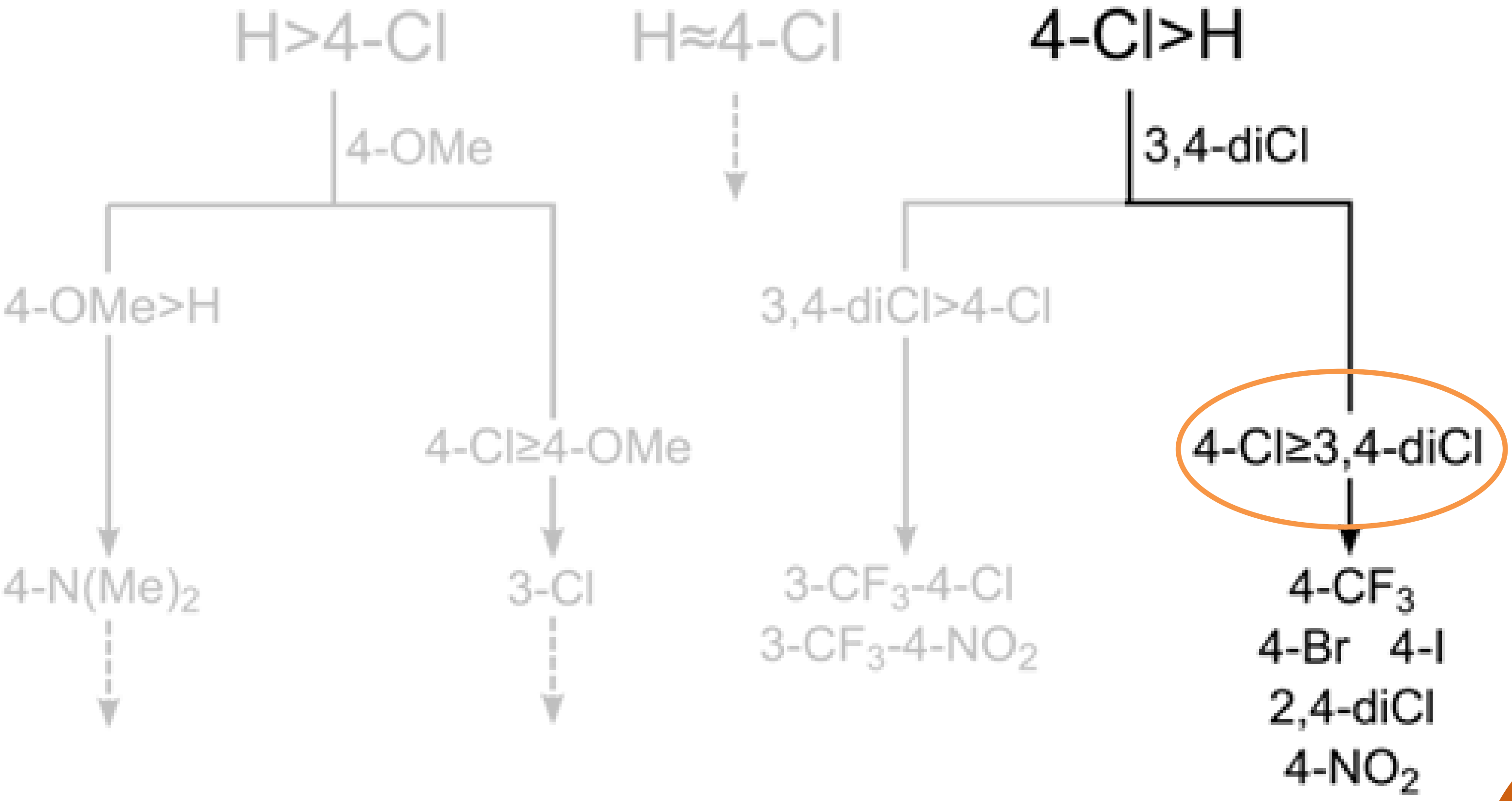
	% >	Counts	ΔLogP
	32	57	+0.4
	24	37	0.0
	22	23	-0.1
	21	47	-2.9
	20	20	-0.1



(1st if lower cutoff)

	33	5.00	15	+0.3

TOPLISS DECISION TREE



H>4-Cl

4-OMe

4-OMe>H

$4\text{-}N(Me)_2$

4-Cl≥4-OMe

3-Cl

H≈4-Cl

4-Cl>H

3,4-diCl

3,4-diCl>4-Cl

$3\text{-}CF_3\text{-}4\text{-}Cl$

$3\text{-}CF_3\text{-}4\text{-}NO_2$

4-Cl≥3,4-diCl

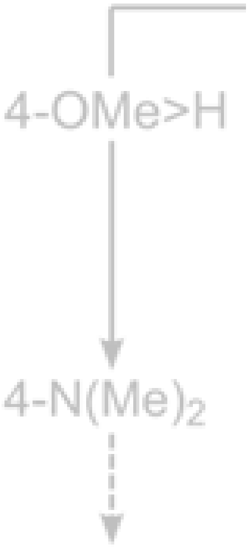
$4\text{-}CF_3$

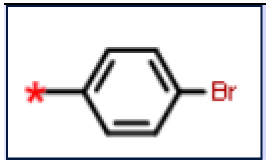
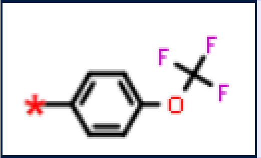
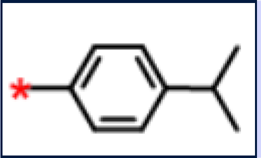
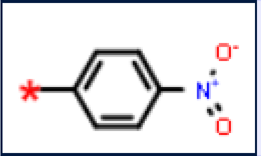
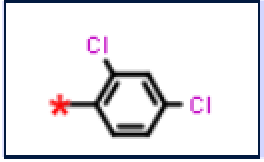
4-Br 4-I

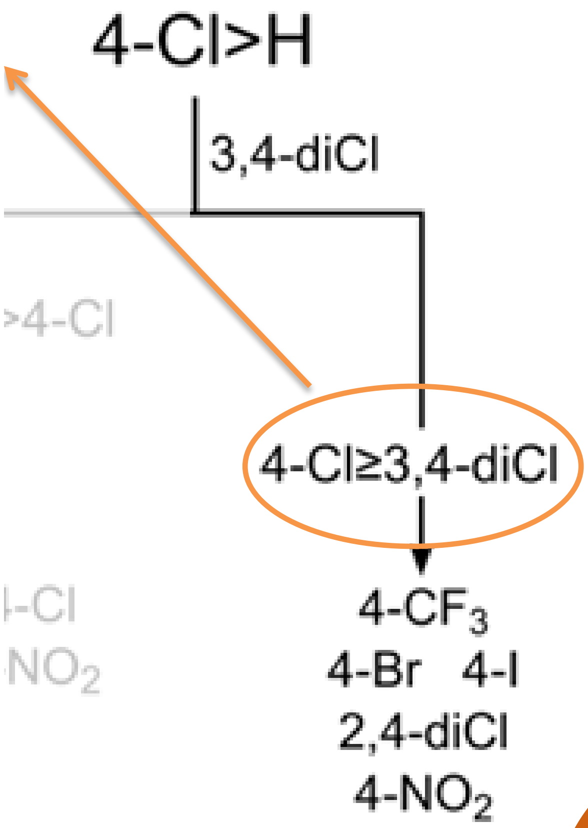
2,4-diCl

$4\text{-}NO_2$

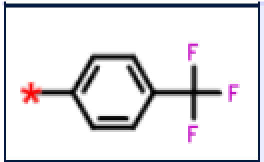
TOPLISS DECISION TREE



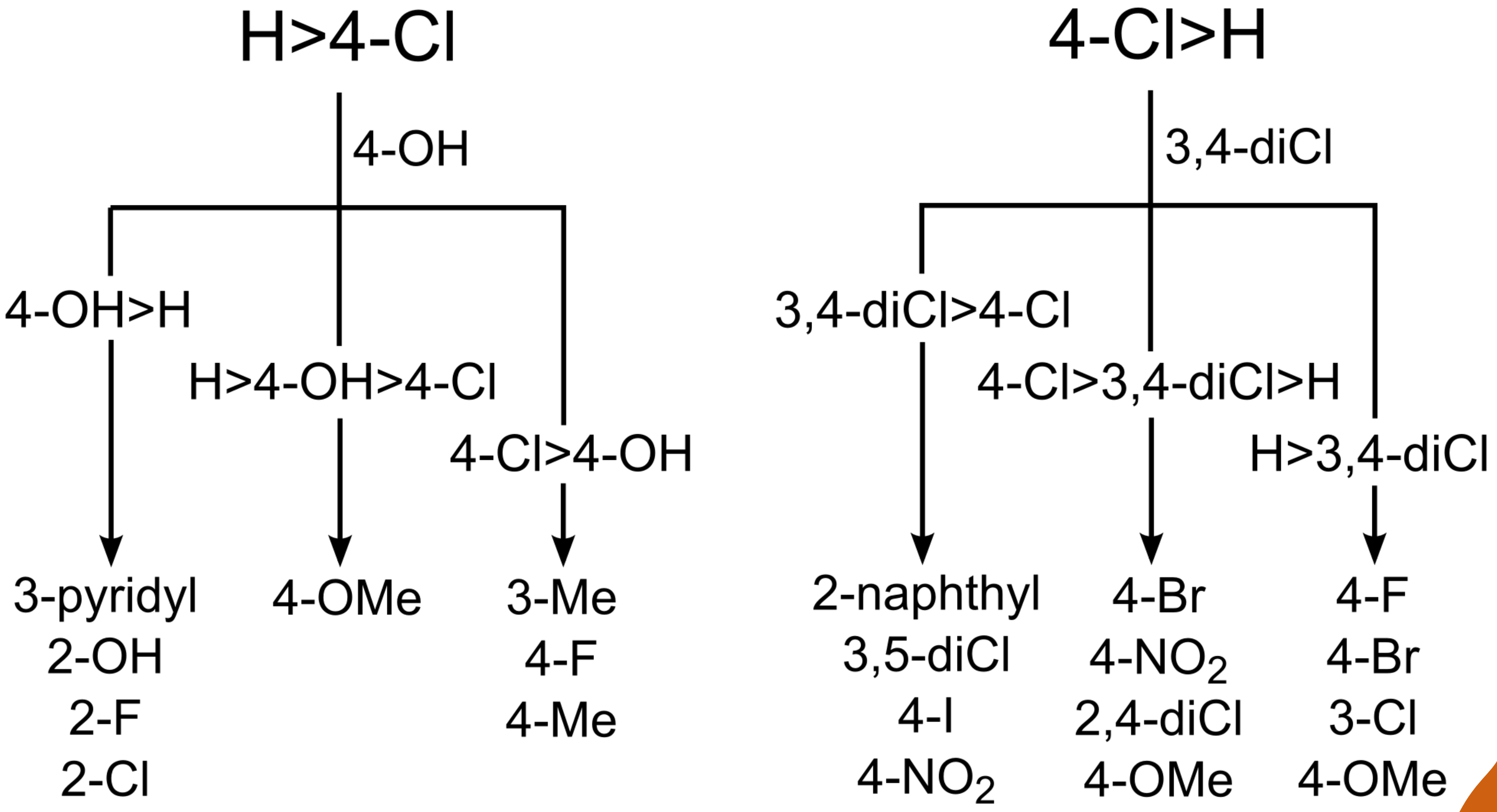
	% >	Counts	ΔLogP
	56	36	+0.2
	46	13	+0.8
	44	16	+1.0
	32	28	-2.3
	27	30	+0.6



(20th)

	11	28	+0.3

MATSY DECISION TREE (ONE OF MANY)



H>4-Cl

4-OH

4-OH>H

3-pyridyl
2-OH
2-F
2-Cl

H>4-OH>4-Cl

4-OMe

4-Cl>4-OH

3-Me
4-F
4-Me

4-Cl>H

3,4-diCl

3,4-diCl>4-Cl

2-naphthyl
3,5-diCl
4-I
4-NO_2

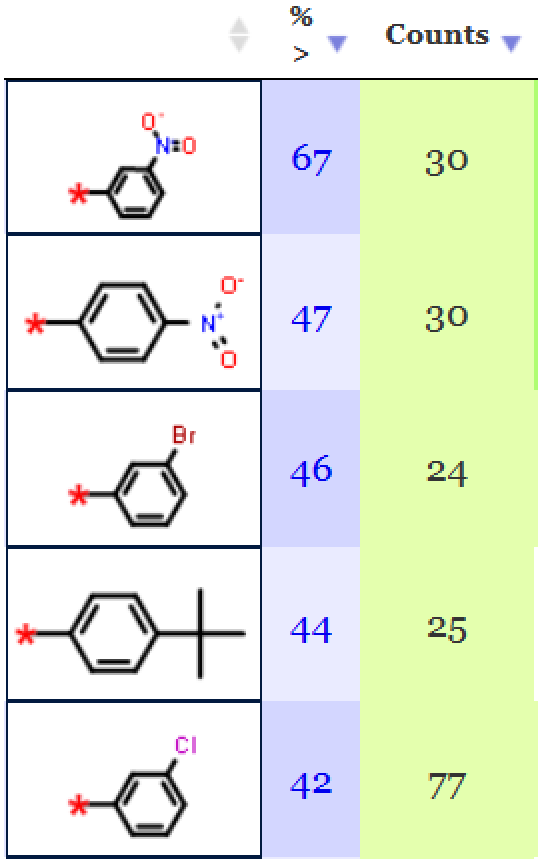
4-Cl>3,4-diCl>H

4-Br
4-NO_2
2,4-diCl
4-OMe

H>3,4-diCl

4-F
4-Br
3-Cl
4-OMe

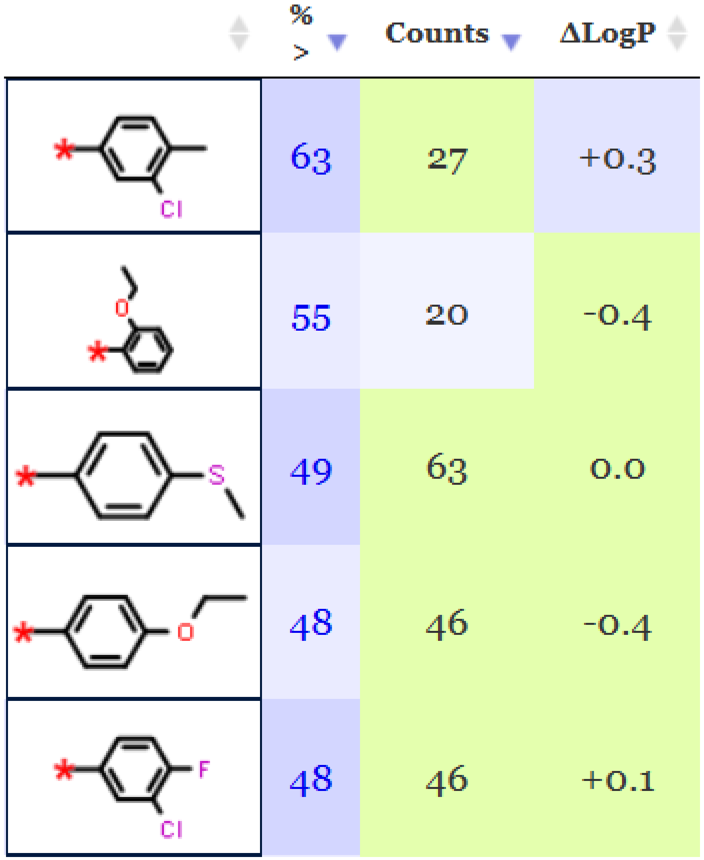
MODIFYING THE PREDICTIONS FOR 4-Cl > H



	% >	Counts
O^- N^+=O	67	30
O^- N^+ O	47	30
Br	46	24
	44	25
Cl	42	77

Kinases

Target-specific

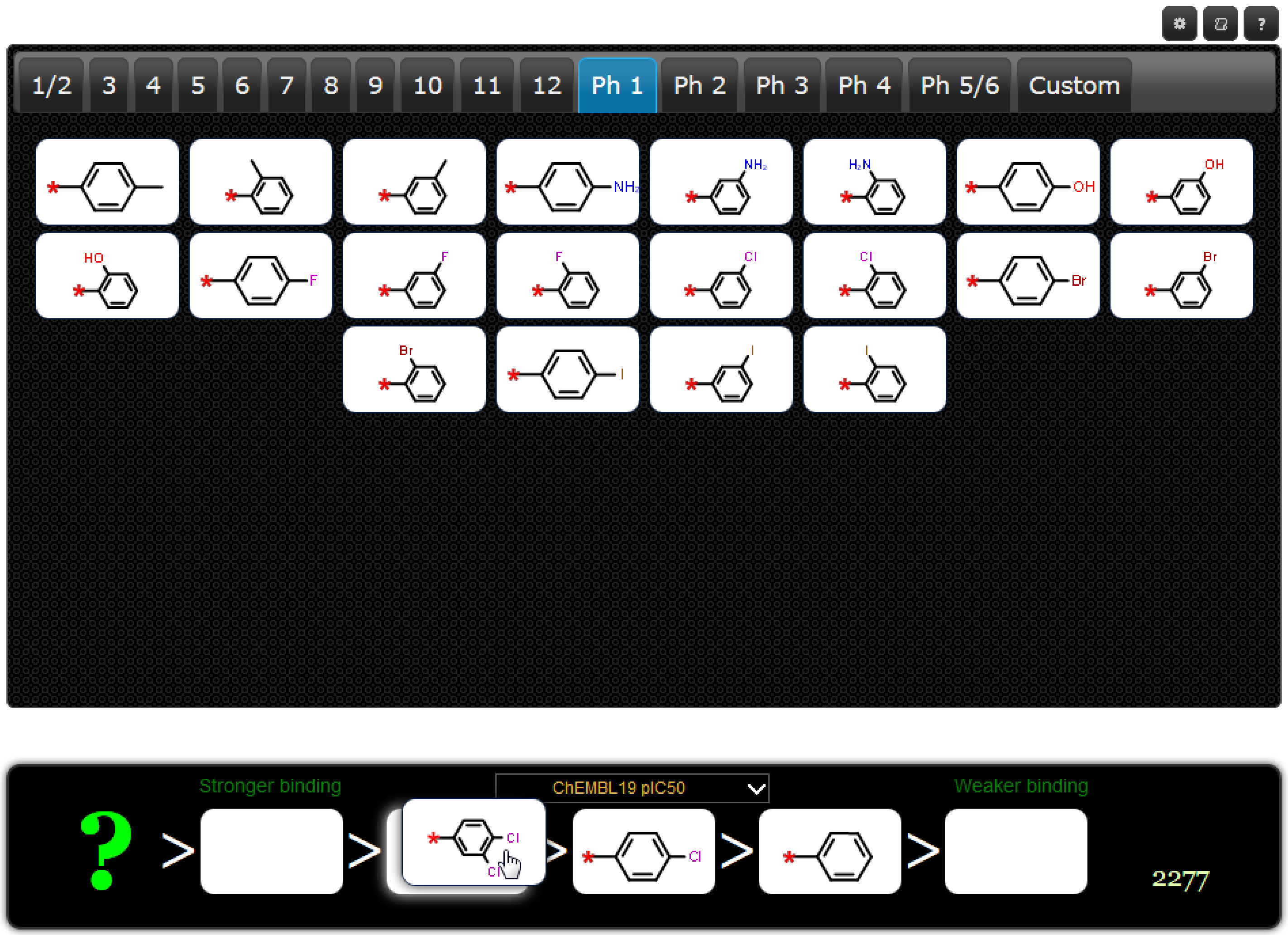


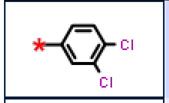
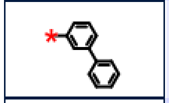
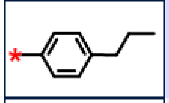
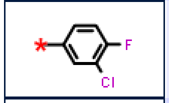
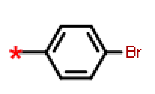
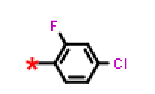
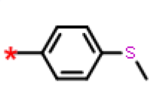
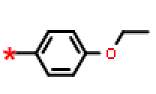
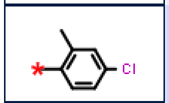
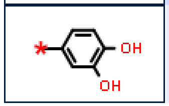
	% >	Counts	ΔLogP
Cl	63	27	+0.3
O	55	20	-0.4
S	49	63	0.0
O	48	46	-0.4
F Cl	48	46	+0.1

ΔLiPE > 0

Incorporate metrics

DRAG-AND-DROP INTERFACE TO MATSY



	% >	Counts	ΔLogP
	54	**391**	+0.6
	54	37	+2.0
	53	30	+1.1
	52	46	+0.1
	50	**521**	+0.2
	50	32	+0.1
	49	63	0.0
	48	46	-0.4
	48	25	+0.3
	48	21	**-2.2**

Showing 11 to 20 of 111 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

Using RDKit for Matched Molecular Series Analysis

When two are not enough

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.