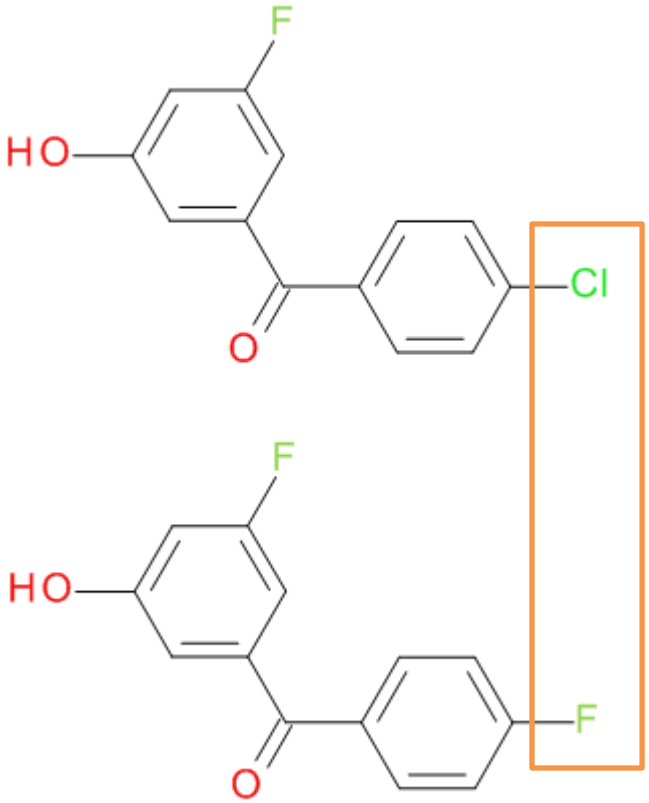
250th ACS National Meeting, Boston
16th Aug 2015

Visualization and manipulation of Matched Molecular Series for decision support

Noel O'Boyle and Roger Sayle

NextMove Software

MATCHED (MOLECULAR) PAIRS



1.6

3.5

[Cl, F]

Coined by Kenny and Sadowski in 2005*

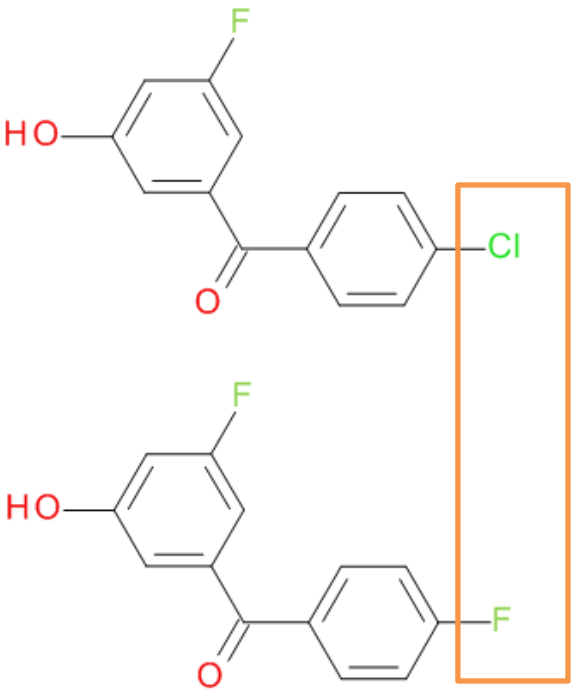
Easier to predict **differences** in the values of a property than it is to predict the value itself

* 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
- Need to look beyond matched pairs

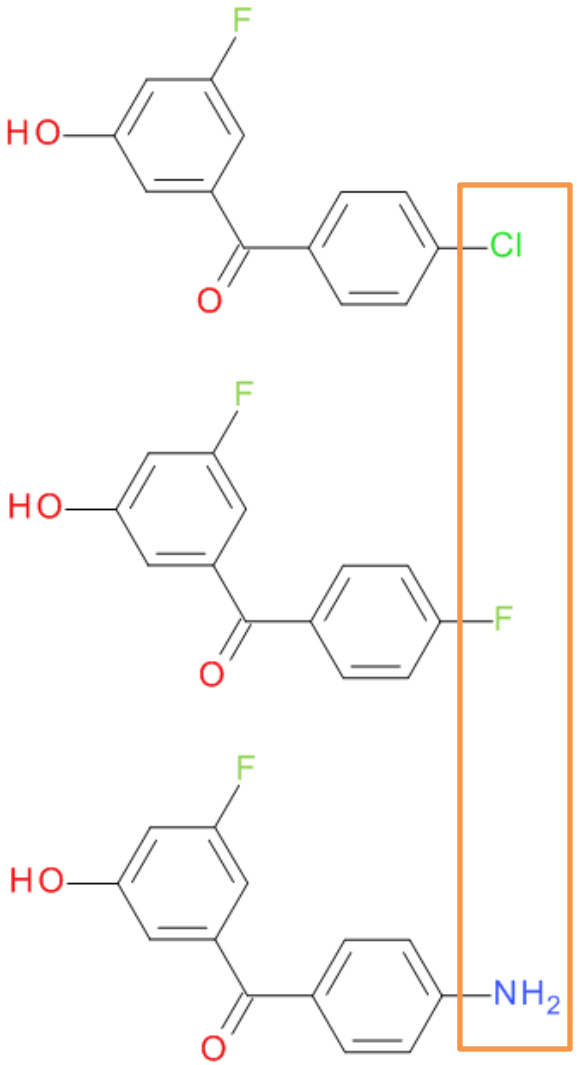
MATCHED SERIES OF LENGTH 2 = MATCHED PAIR



[Cl, F]

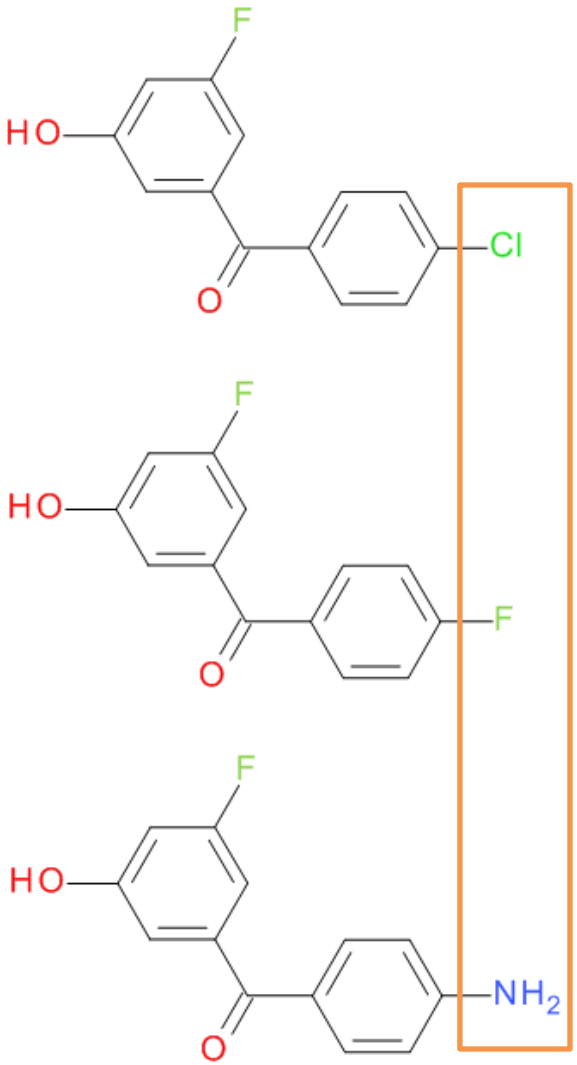
"Matching molecular series" introduced by Wawer and Bajorath, *J. Med. Chem.* **2011**, *54*, 2944

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]

THE MATCHED PAIR MENTALITY

- There can only be two
  - Like inhabitants of Flatland ignorant of a third dimension
- What is the equivalent of pair for three?
  - Triad, trio, triple?
- A matched pair represents a transformation from A->B
  - How would that work if it there were three?

MATSY:
PREDICTION USING MATCHED SERIES

MATCHED SERIES HAVE PREFERRED ORDERS

Series	Enrichment	Observations
Br > Cl > F > H	5.36*	256
Cl > Br > F > H	3.14*	150
H > F > Cl > Br	**1.53***	**73**
Br > Cl > H > F	1.40	67
F > Cl > Br > H	1.36	65
Cl > F > Br > H	0.96	46
…	…	…
H > F > Br > Cl	**0.77**	**37**
…	…	…
H > Br > F > Cl	**0.48***	**23**
Cl > H > F > Br	0.48*	23
Cl > F > H > Br	0.48*	23
H > Cl > F > Br	0.42*	20
Br > F > H > Cl	0.40*	19
F > H > Br > Cl	0.40*	19
H > Cl > Br > F	0.38*	18
F > Br > H > Cl	0.36*	17
Br > H > F > Cl	**0.17***	**8**

The fact that certain orders are preferred may be used as the basis of a predictive method

FIND R GROUPS THAT INCREASE ACTIVITY

ChEMBL

In-house

Query matched series

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**
…	…		**…**

O'Boyle, Boström, Sayle, Gill. *J. Med. Chem.* **2014**, *57*, 2704.

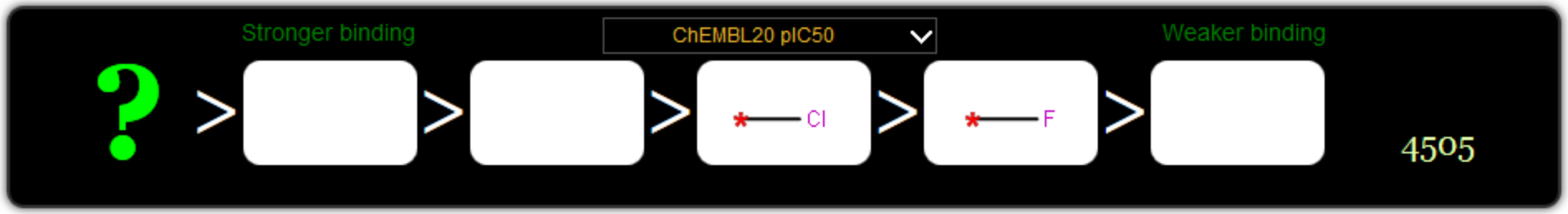
THE DATASET-CENTRIC APPROACH

- "Here is my dataset of molecules with activities – now tell me what to make next"
- Pro:
  - Easy for users to get up-and-running
  - Fits with their existing way of thinking
    - Don't need to think too much about matched series
- Con:
  - User is one step removed from the matched series data on which the predictions are actually based
  - Dataset is fixed: cannot play with around with the prediction input

GOALS FOR THE INTERFACE

- Visual interface based around R-Groups as first-class objects arranged in ordered series
  - Promote new paradigm
  - Make it clear that the scaffold is not involved
- Should help break the “matched pair” mentality
  - Just a particular case of matched series
- Should be easy to play with
  - Easy to manipulate and quick to respond

- Drag-and-drop R Groups into slots to represent observed activity order
  - The query matched series



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

ChEMBL20 pIC50

Weaker binding

? > > > > >

4505

PROS

- Easy to play around with
  - Swap around order of R groups
  - See what happens if you follow the predictions
- May suggest hypotheses
- Useful for searching (not just for predictions)
- Tablet-friendly

CONS

- The user needs to be able to provide an ordered matched series as a query
  - You can't just provide a dataset of molecules

NO CHEMISTRY REQUIRED

- Predictions are solely based on the order of R groups in a matched series
  - Not using any calculated properties
- Images of all R groups in ChEMBL can be generated in advance (~65K)
- ⇒ A cheminformatics toolkit is not required for the interface or even for making predictions
- In practice, we do use a toolkit to allow the user to enter R groups as SMILES

USE CASE #1
ARE MATCHED SERIES PREDICTIONS SYMMETRIC?

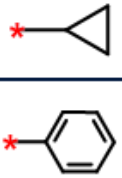
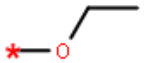
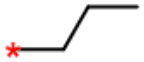
ARE MATCHED SERIES PREDICTIONS SYMMETRIC?

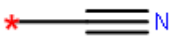
- If A>B>C>D is a highly preferred order
  - Then D>C>B>A also tends to be preferred
- Hypothesis:
  - if A reduces the activity given D>C>B

  ⇒ it will also improve the activity given B>C>D

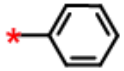
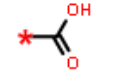
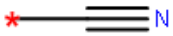
- If true, then we have twice as much data to use for predictions
  - Let's find out….

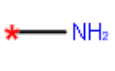
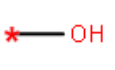
? > Cl > F > H

	% >
	54
	50
	46
	39
	37
	36
	35
	33
	33
	33

	% >
	30
	30
	26
	26
	26
	24
	23
	20
	15
	15

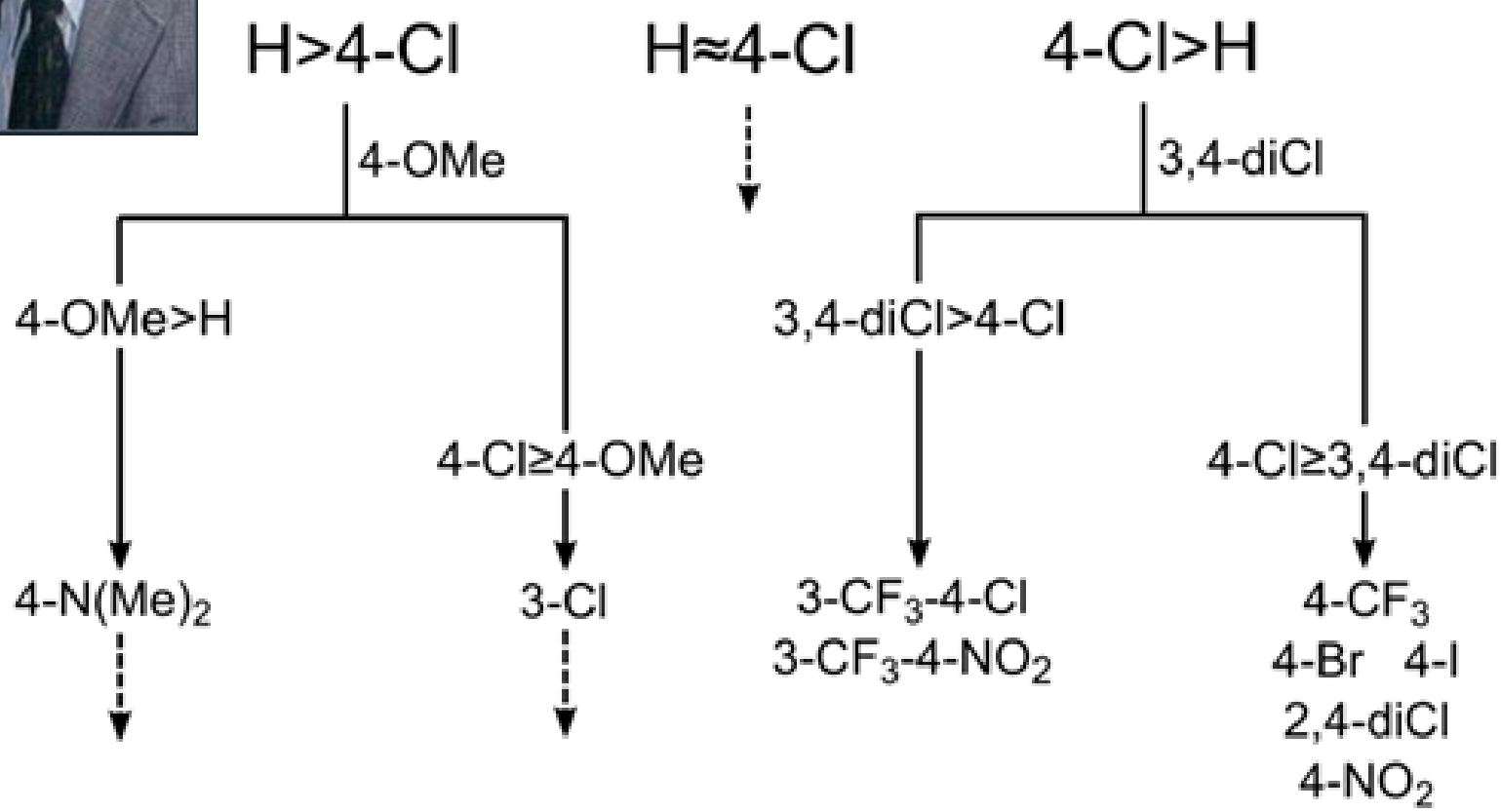
H > F > Cl > ?

	% <
	73
	67
	65
	59
	57
	50
	50
	47
	46
	42

	% <
	36
	36
	34
	21
	19

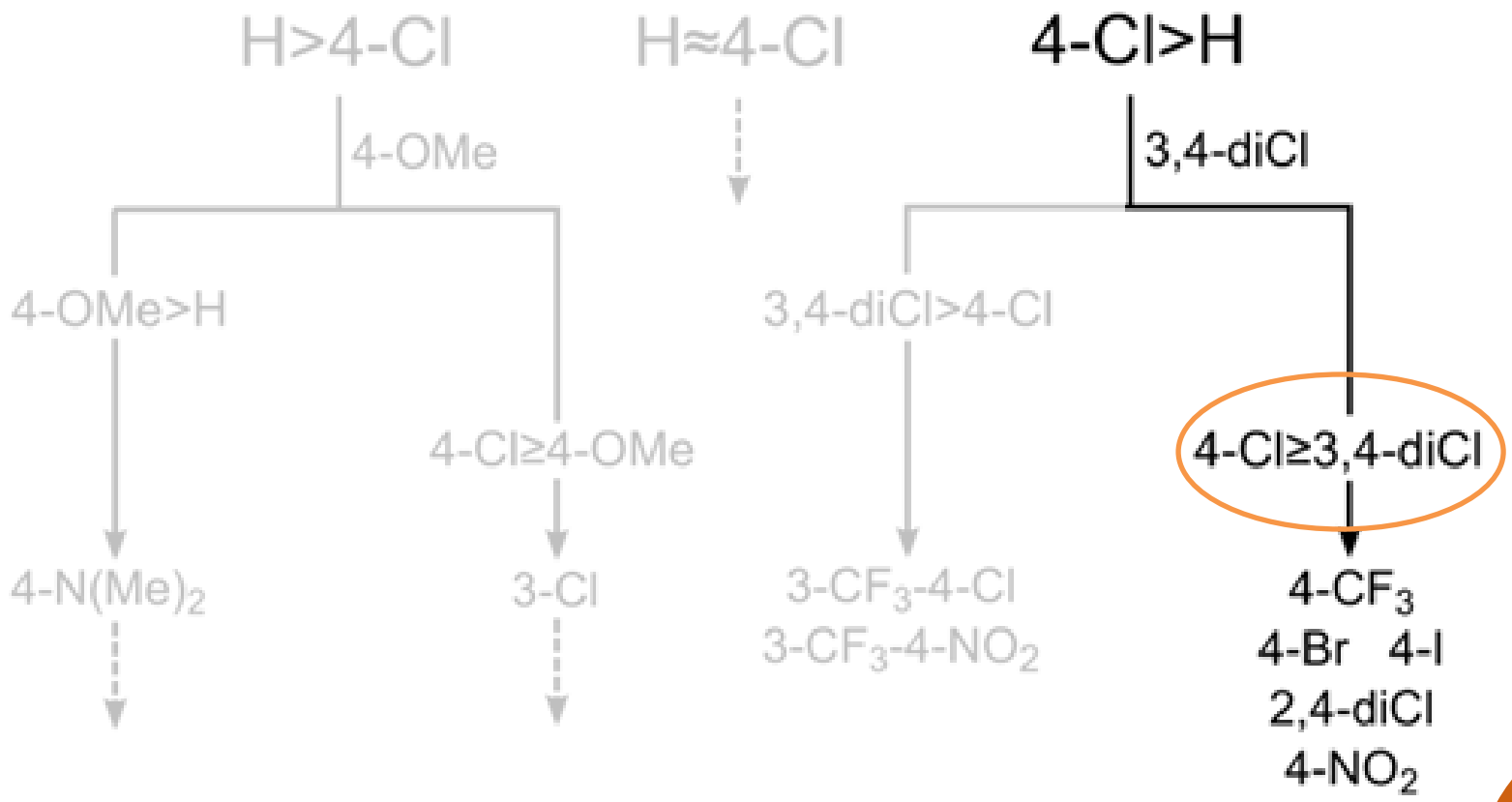
USE CASE #2
TOPLISS DECISION TREE

TOPLISS DECISION TREE

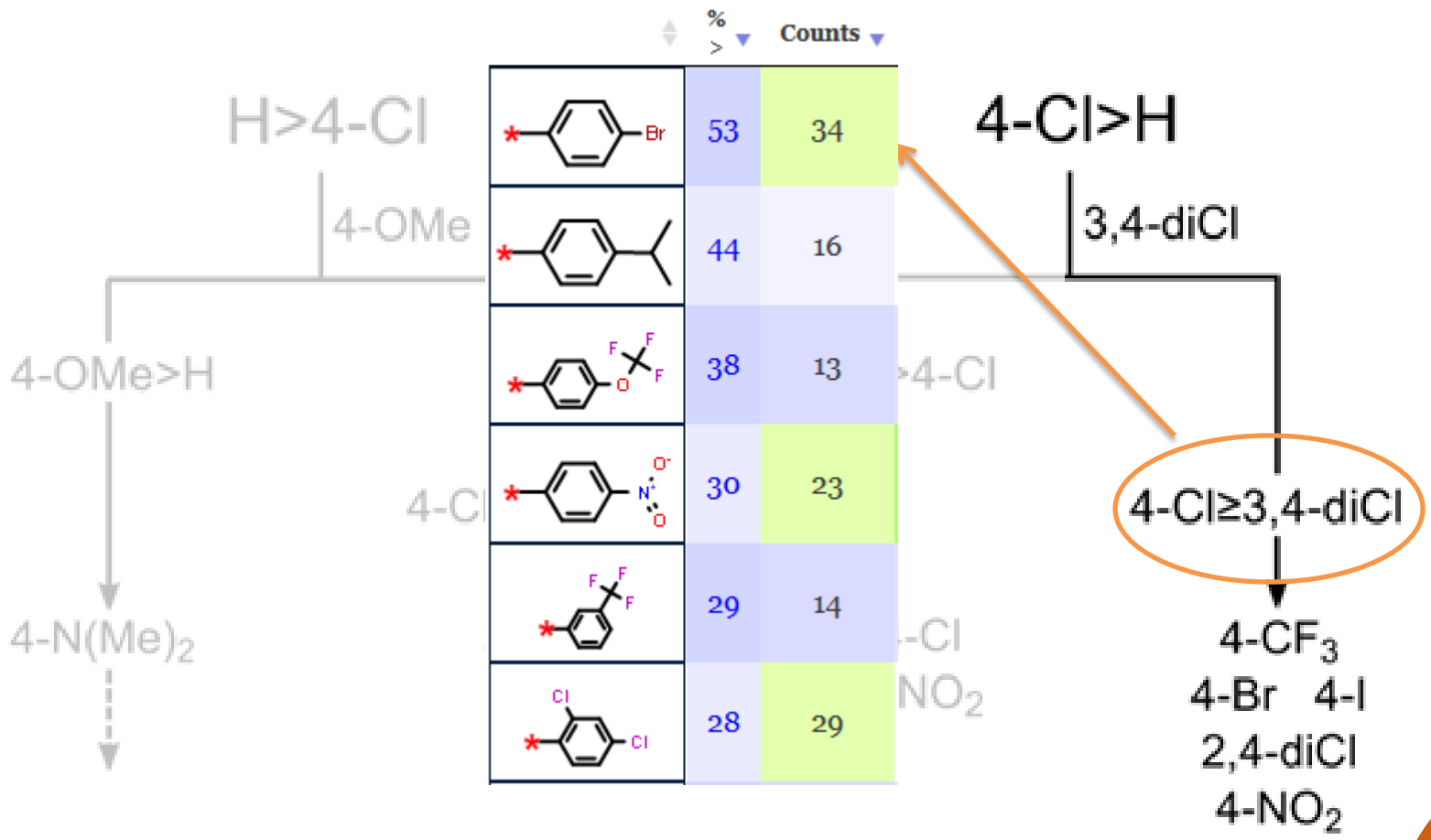


Topliss, J. G. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Chem.* **1972**, *15*, 1006–1011.

TOPLISS DECISION TREE

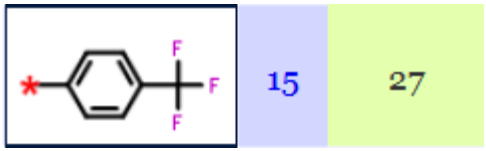


TOPLISS DECISION TREE

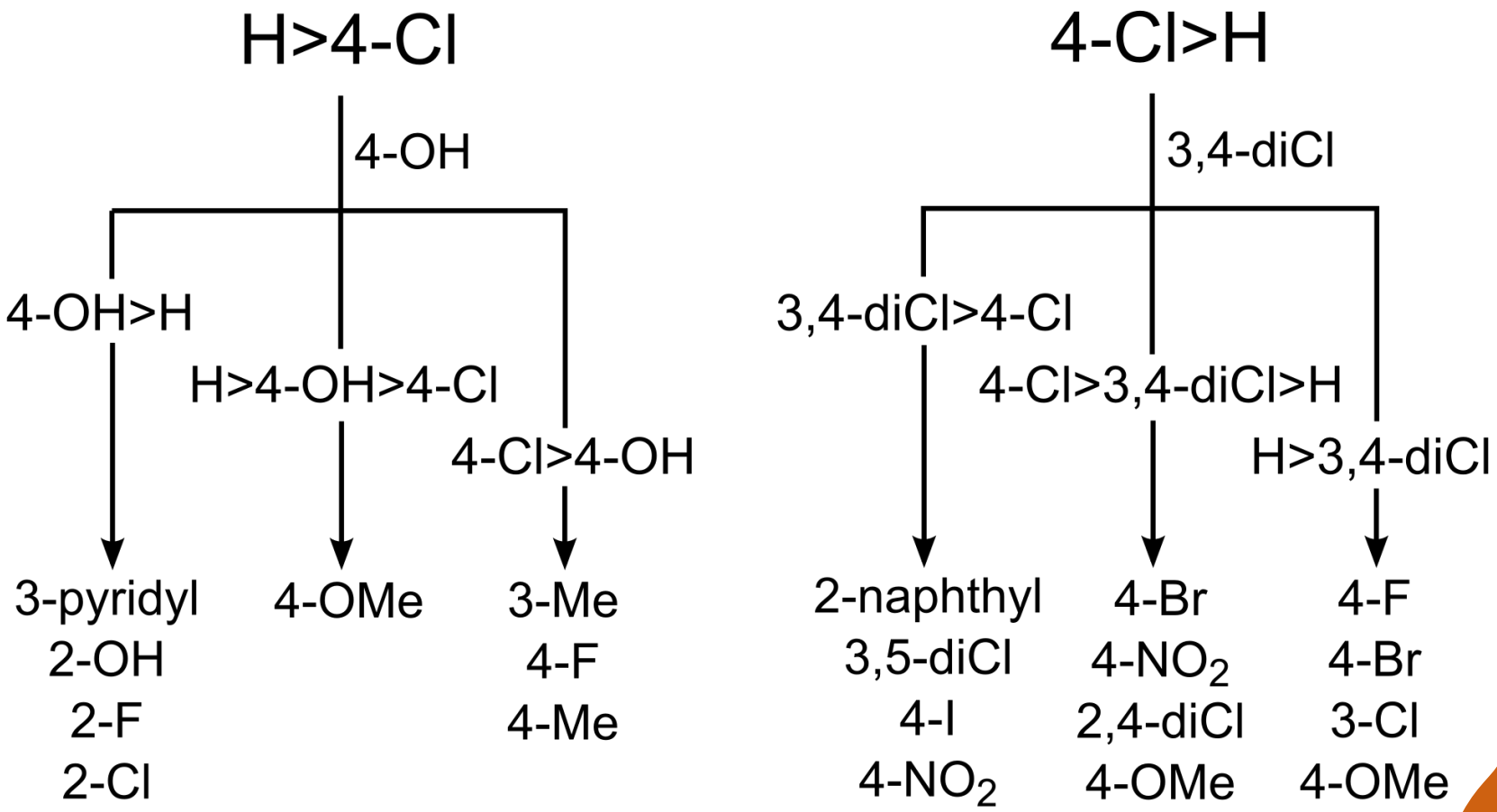


	% >	Counts
	53	34
	44	16
	38	13
	30	23
	29	14
	28	29

(17th)

	15	27

CHEMBL-BASED DECISION TREE (ONE OF MANY)



Visualization and manipulation of Matched Molecular Series for decision support

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