



248th ACS National Meeting
Aug 2014

Revising the Topliss decision tree

...based on 30 years of medicinal chemistry
literature

Noel O'Boyle and Roger Sayle

NextMove Software

Jonas Boström

AstraZeneca





American Chemical Society

Division of Medicinal Chemistry

Hall of Fame



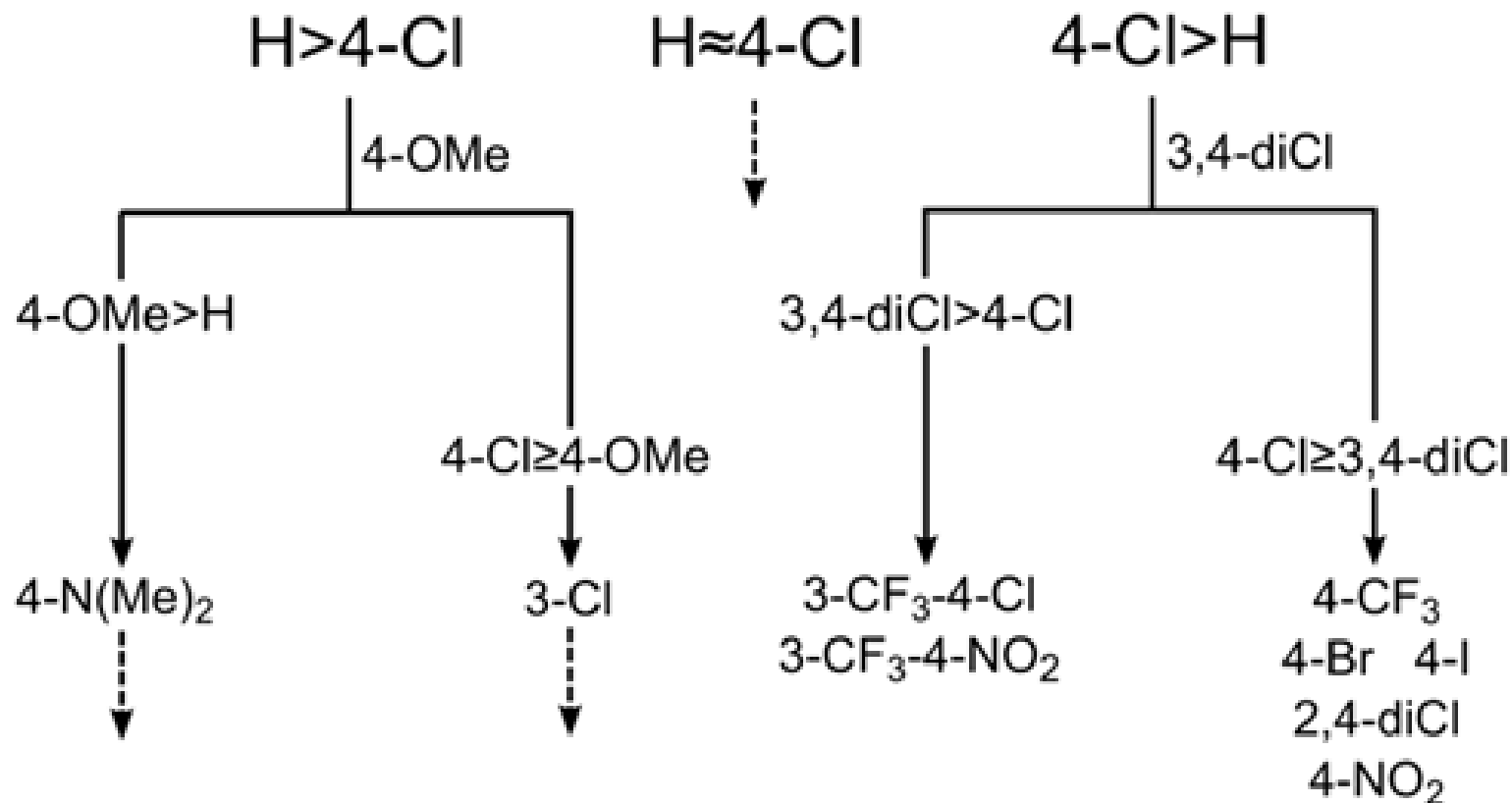
John G. Topliss, Ph.D.

Dr. John G. Topliss was born near Mansfield, England in 1930. He received a BSc degree with honors in chemistry, first class, in 1951 from The University of Nottingham and a Ph.D. degree in organic chemistry with Professor F. E. King on the total synthesis of tricyclic diterpenoids from the same institution in 1954. He then did postdoctoral research with Professor Holger Erdtman at The Royal Technical College in Stockholm, Sweden, on the isolation and structure determination of heartwood constituents, and with Professor Gilbert Stork at Columbia University on natural product total synthesis.

Dr. Topliss joined the Schering Corporation (now Schering-Plough) as a synthetic medicinal chemist in 1957, and in the following years worked primarily in the diuretic, antihypertensive, CNS, and antiandrogen areas. In a roughly 10 year period he and his research group synthesized and patented 5 drugs (trichlormethiazide, diazoxide, halazepam, quazepam, and flutamide) which were subsequently marketed.

Seeking a more rational, theoretically based approach to analog synthesis than was generally employed at that time, Dr. Topliss was one of the early medicinal chemists in the pharmaceutical industry to use quantitative structure-activity relationships (QSAR) methodology. This led him to formulate Operational Schemes for Analog Synthesis in Drug Design (later known as the Topliss Tree) published in 1972, and also a related Manual Hansch Approach published in 1977, which are simplified non-mathematical approaches for rapidly optimizing benzene ring

TOPLISS TREE FOR SUBSTITUTED PHENYL



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



FEATURES OF THE TOPLISS TREE

- Maximize the chances of synthesizing the **most potent compound** in the series as soon as possible
- Based on inferring Hansch structure-activity relationship from relative potencies of R groups
 - Electronic (σ), hydrophobic (π), steric (E_s)
- General scheme
 - for **any target**
 - for **any scaffold**



CHEMBL BIOACTIVITY DATABASE

- July 2008 - **ChEMBL established** with Wellcome Trust grant
 - John Overington, EMBL-EBI
- **Open access** source of bioactivity data abstracted from the literature
 - Chemical structures, activity values, activity type, assay description, journal article name, target
 - www.ebi.ac.uk/chembl/



CHEMBL BIOACTIVITY DATABASE

- ChEMBL 19 – July 2014



- 57k papers

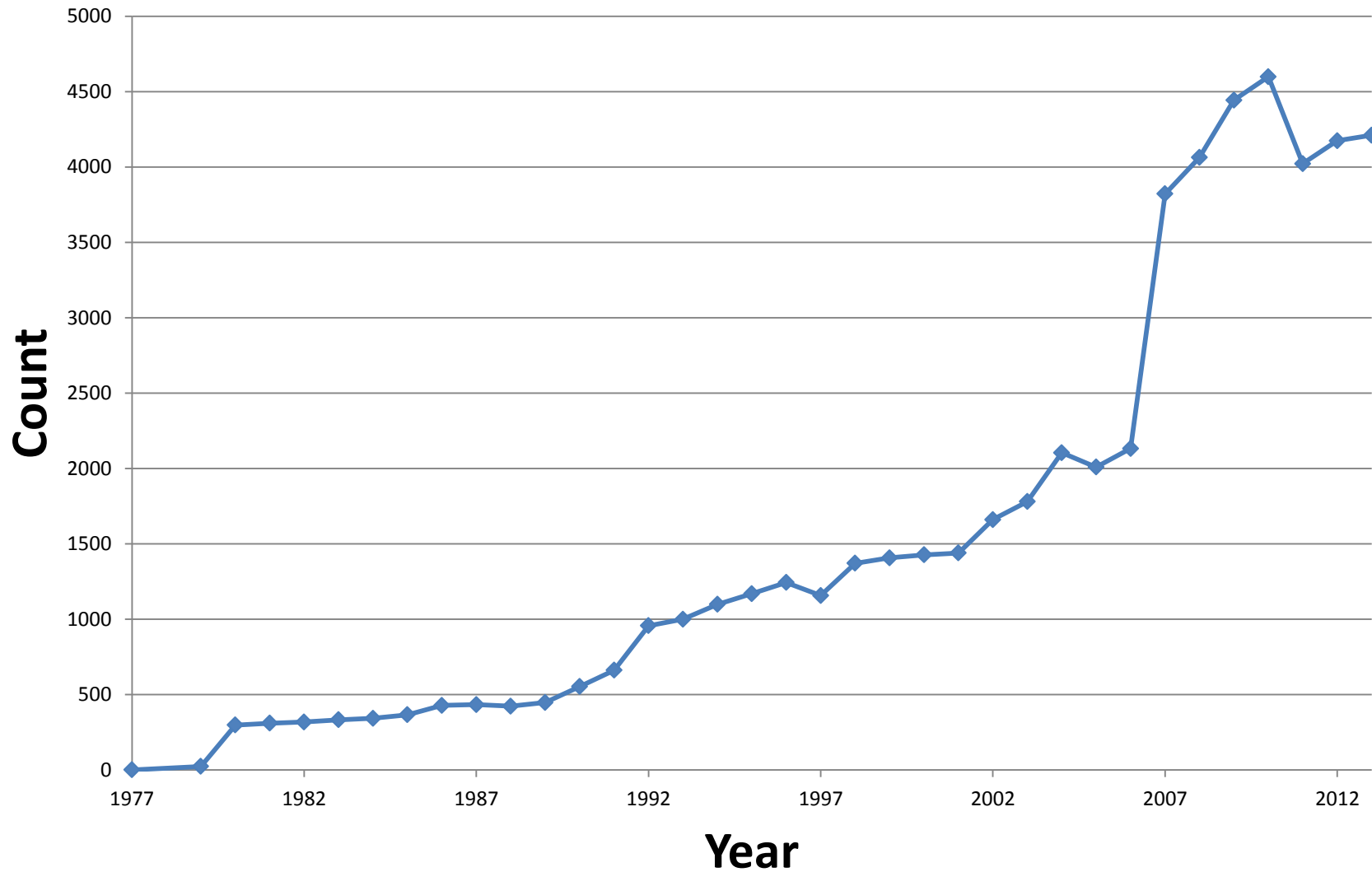
- 94% from *Bioorg. Med. Chem. Lett.*, *J. Med. Chem.*, *J. Nat. Prod.*, *Bioorg. Med. Chem.*, *Eur. J. Med. Chem.*, *Antimicrob. Agents Chemother.*, *Med. Chem. Res.*

- 1.4 million compounds with 12 million activities

- 1.1 million assays against 10k targets



Number of articles extracted from a particular year



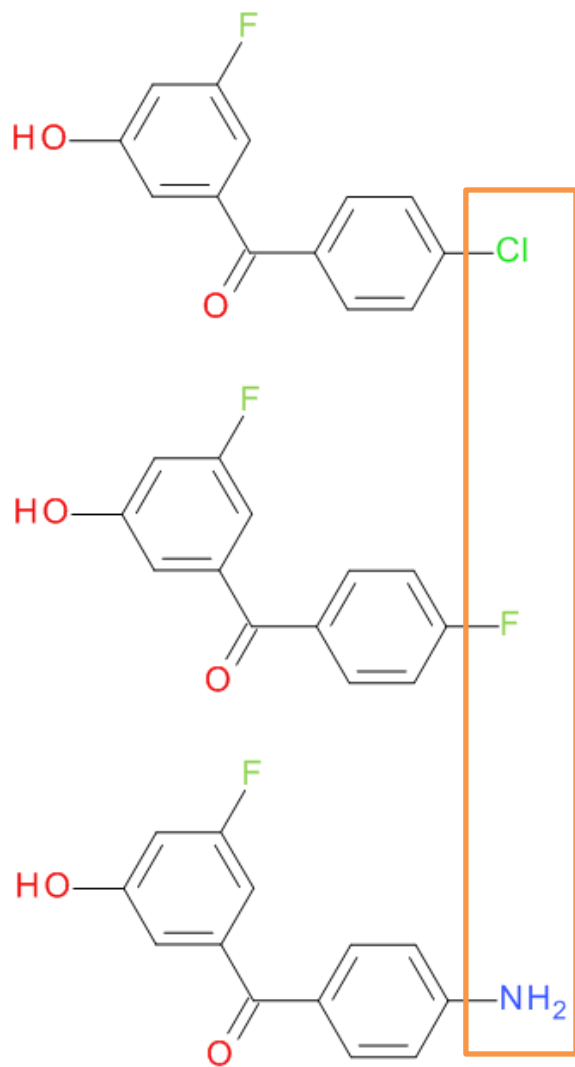
MATCHED (MOLECULAR) SERIES

- Recent concept in cheminformatics (*)
 - ... not so recent in medicinal chemistry
- Series of structural analogs
 - same scaffold
 - different R groups at a single position

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



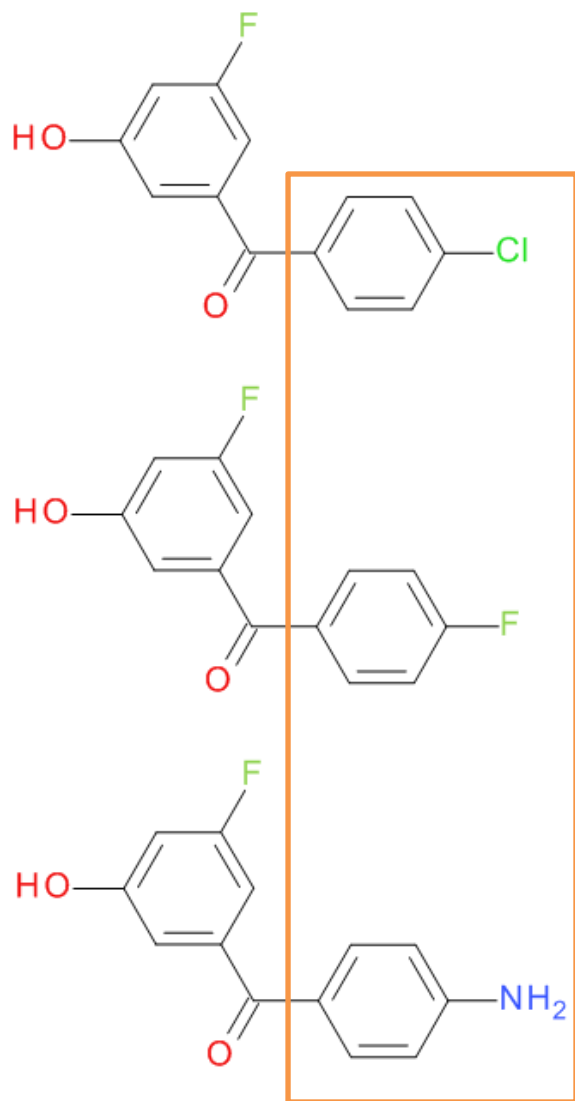
MATCHED SERIES OF LENGTH 3



[Cl, F, NH₂]



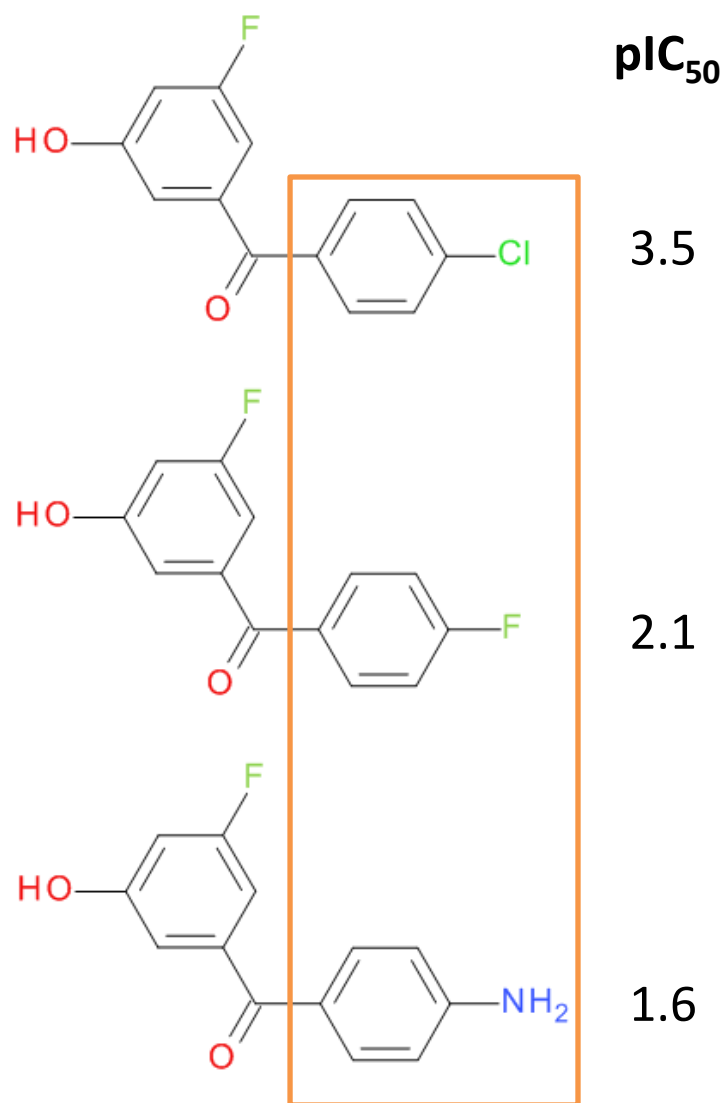
MATCHED SERIES OF LENGTH 3



[4-Cl-Ph, 4-F-Ph, 4-NH₂-Ph]



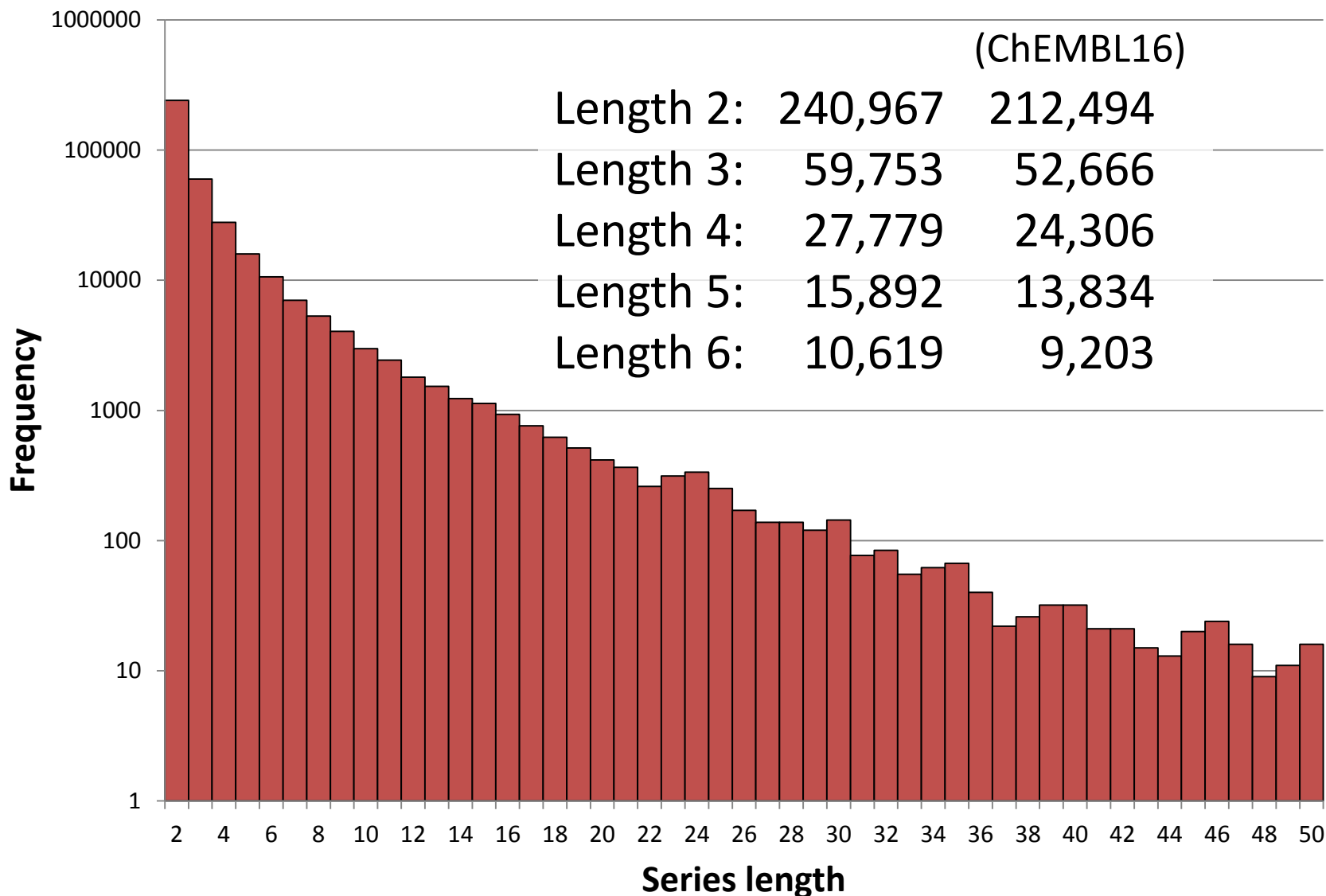
ORDERED MATCHED SERIES



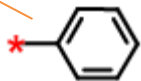
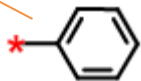
[4-Cl-Ph > 4-F-Ph > 4-NH₂-Ph]



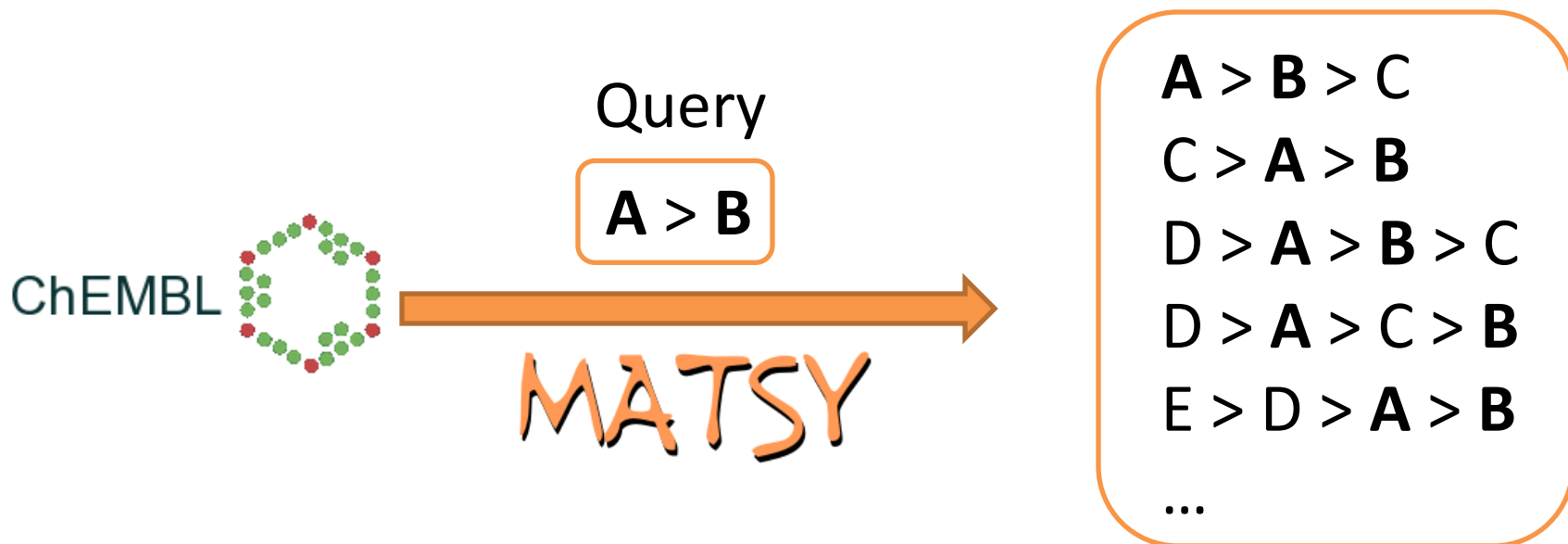
Matched series in ChEMBL19 IC50 binding assays



Method described in O'Boyle, Boström, Sayle, Gill. Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *J. Med. Chem.* **2014**, 57, 2704.



FIND R GROUPS THAT INCREASE ACTIVITY

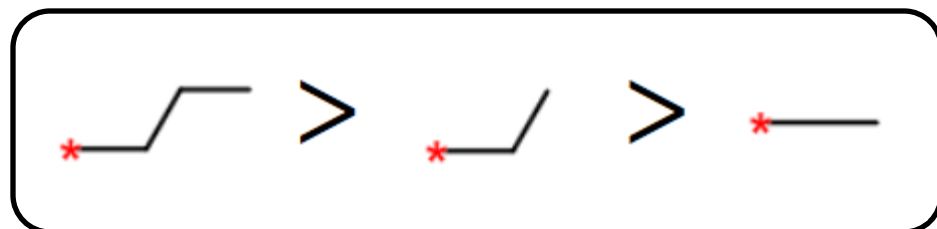
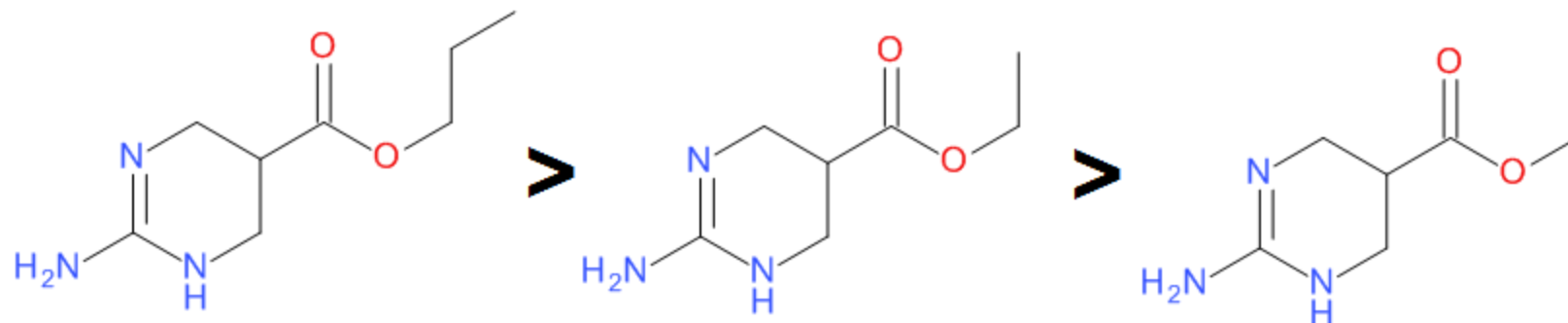


R Group	Observations	Obs that increase activity	% that increase activity
D	3	3	100
E	1	1	100
C	4	1	25
...

O'Boyle, Boström, Sayle, Gill. Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *J. Med. Chem.* **2014**, 57, 2704.

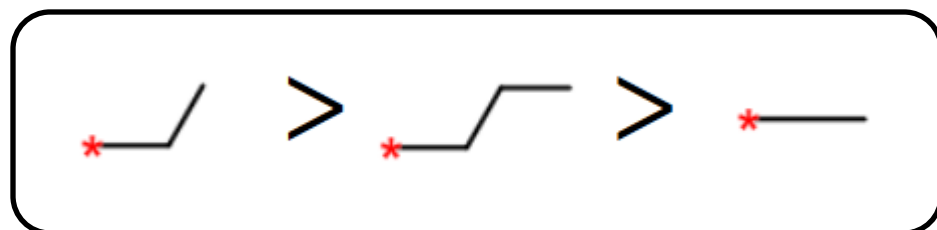
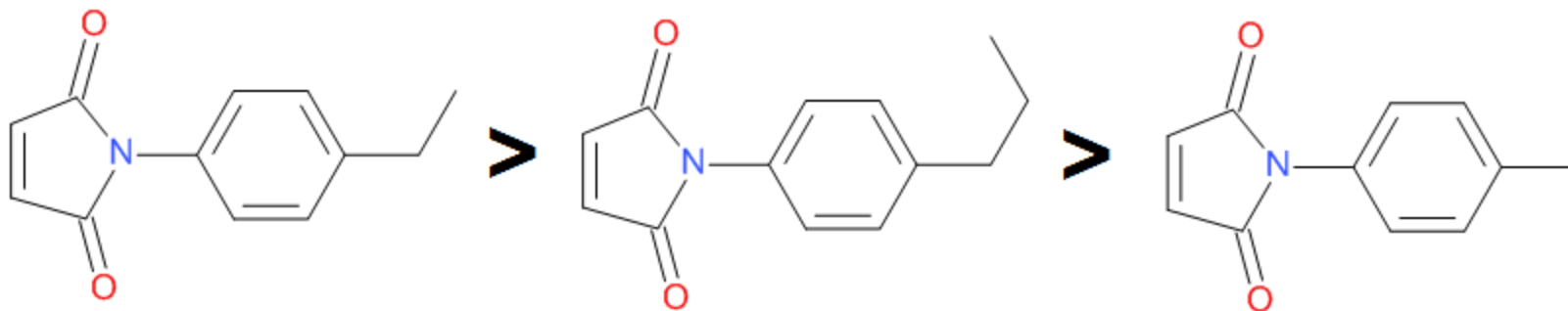


EXAMPLE



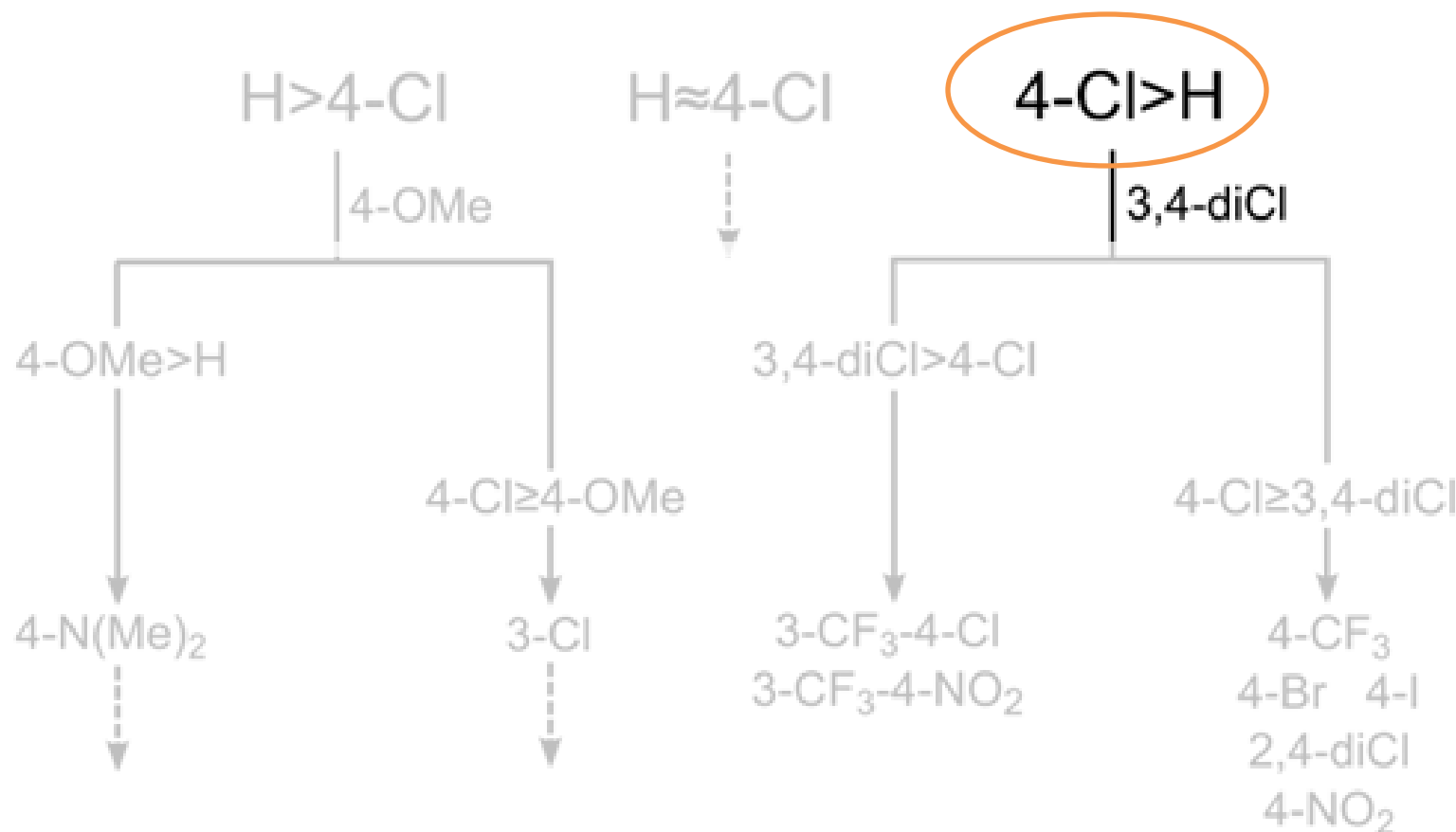
	% >	Counts	ΔLogP
	90	21	+3.3
	72	60	+1.7
	69	32	+2.8
	63	27	+1.6
	60	40	-0.1

EXAMPLE II

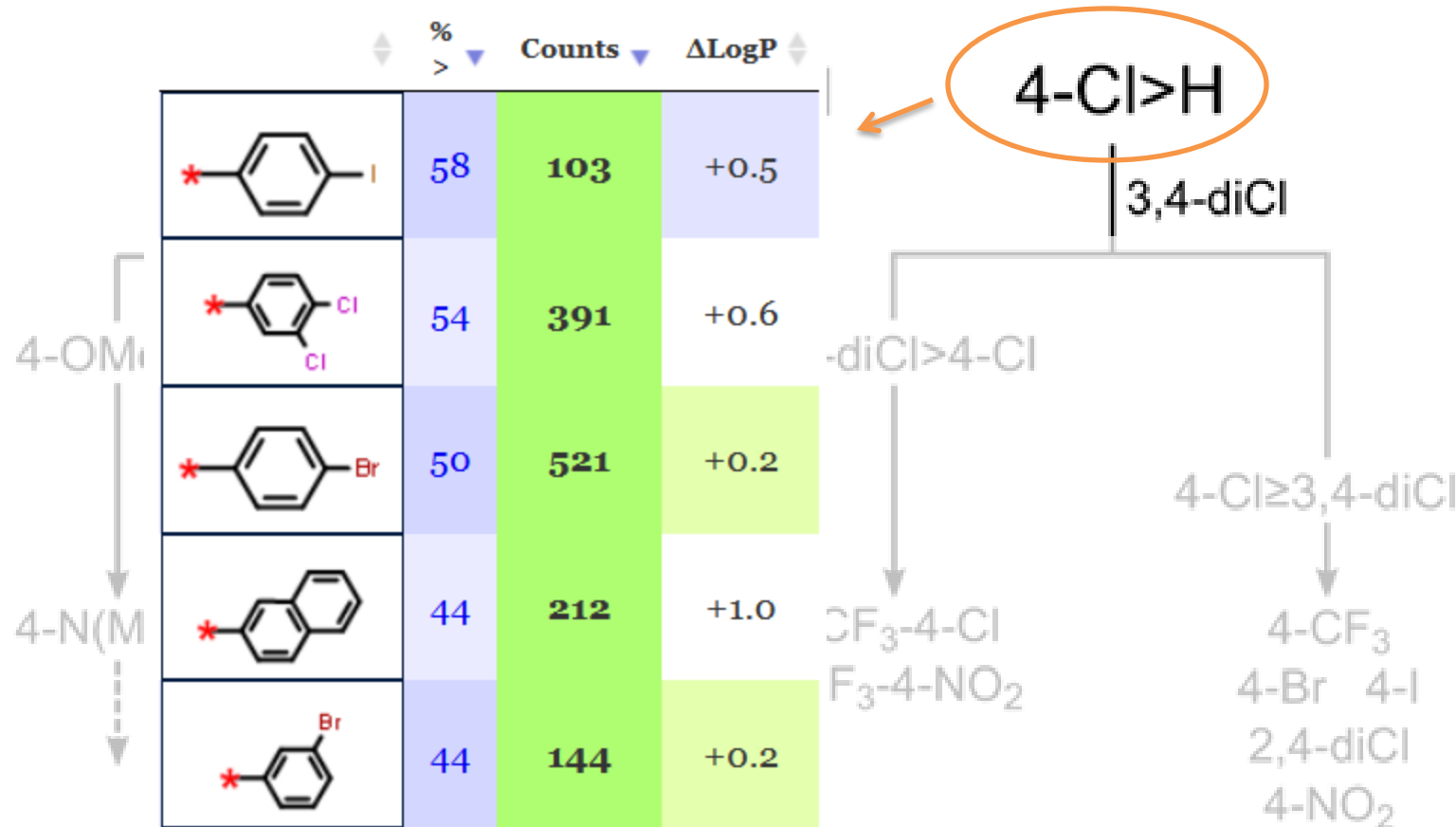


	%	Counts	ΔLogP
Br	38	21	-0.8
	37	27	+0.9
	33	111	+0.3
	33	27	+1.0
	33	21	-1.6

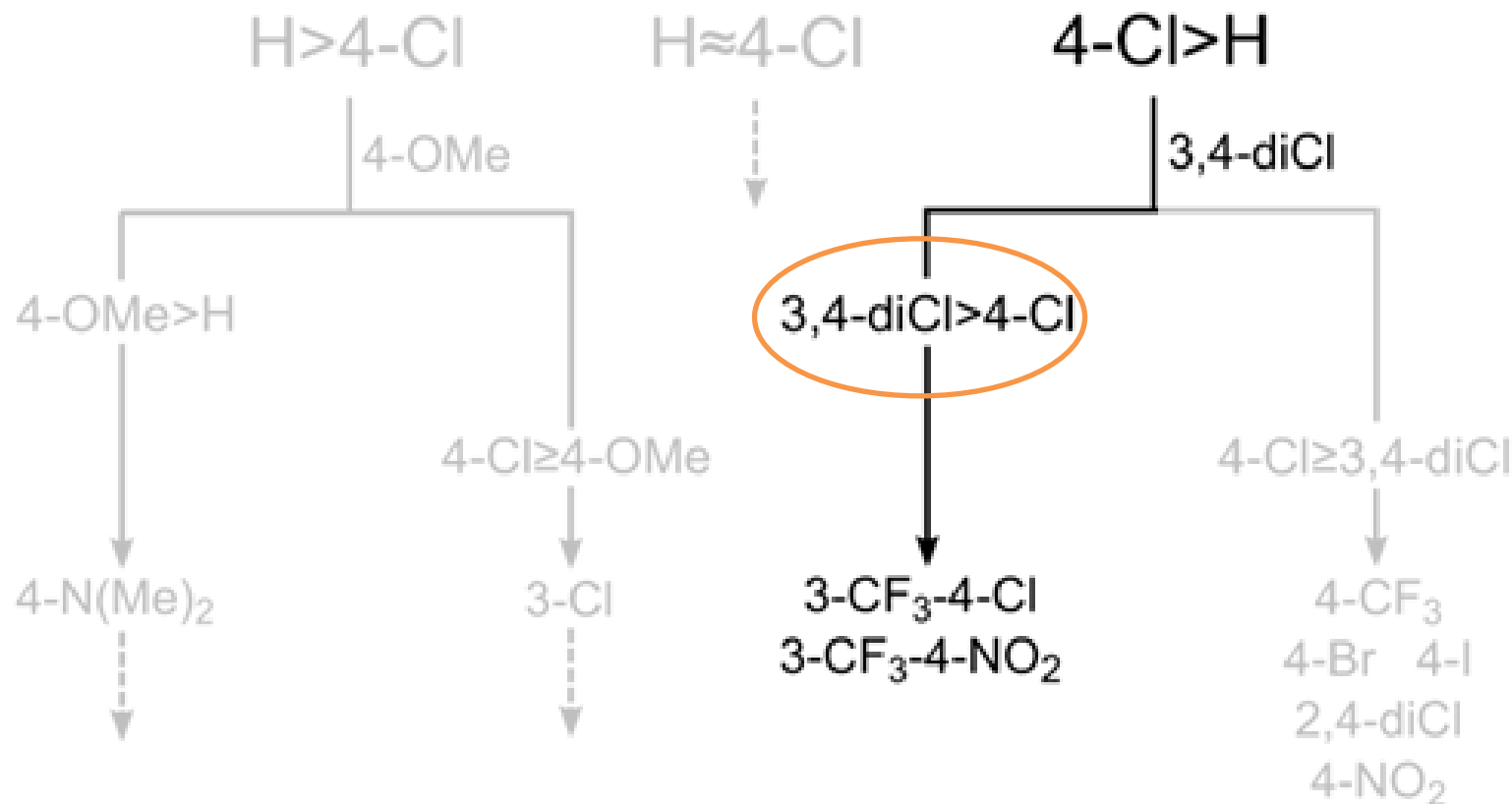
TOPLISS DECISION TREE



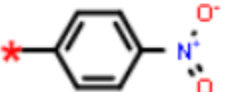
TOPLISS DECISION TREE



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

H ≈ 4-Cl

4-Cl > H

3,4-diCl


3,4-diCl > 4-Cl

3-CF₃-4-Cl
3-CF₃-4-NO₂

4-Cl ≥ 3,4-diCl

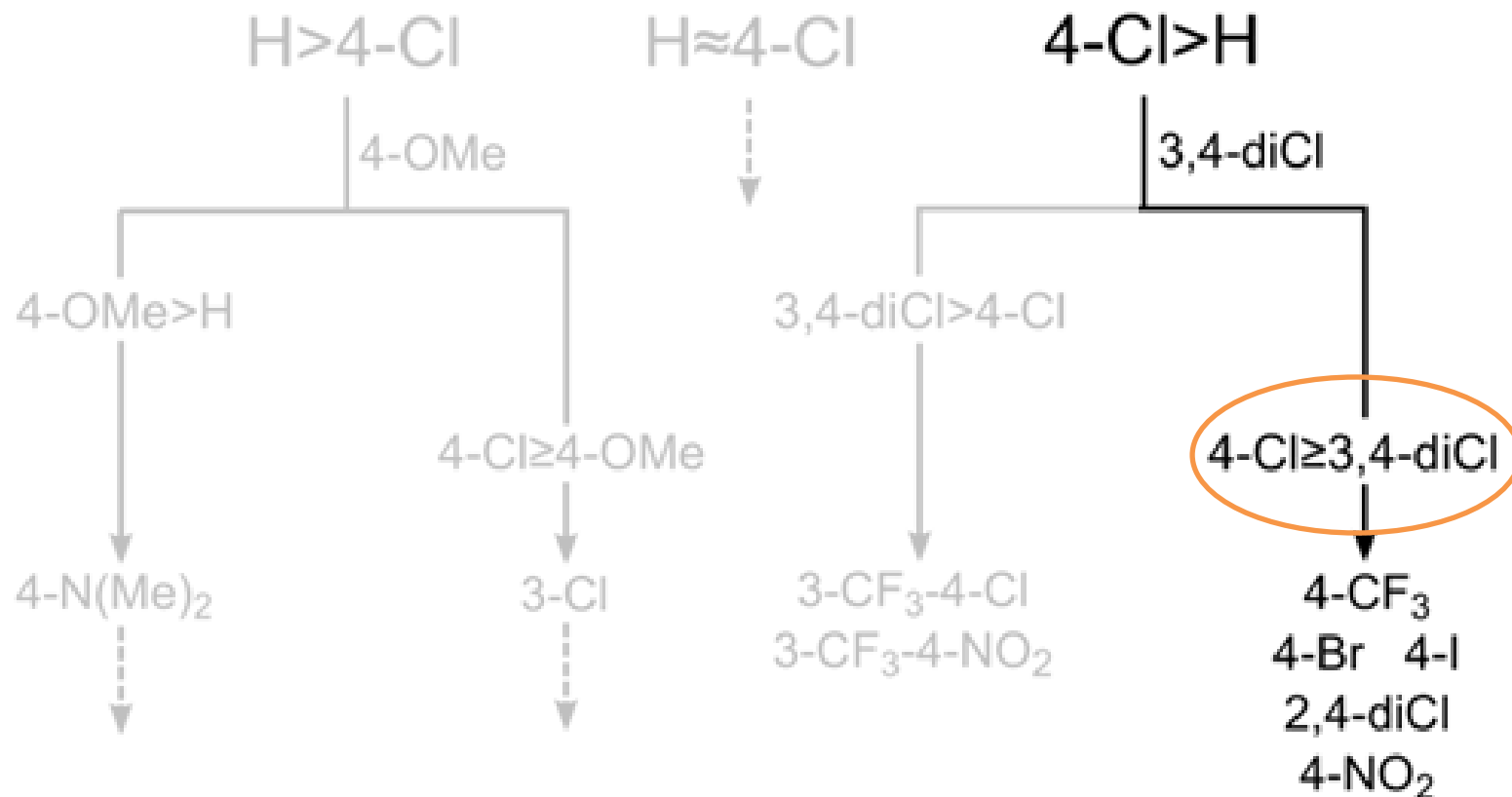
4-CF₃
4-Br 4-I
2,4-diCl
4-NO₂

(1st if lower cutoff)

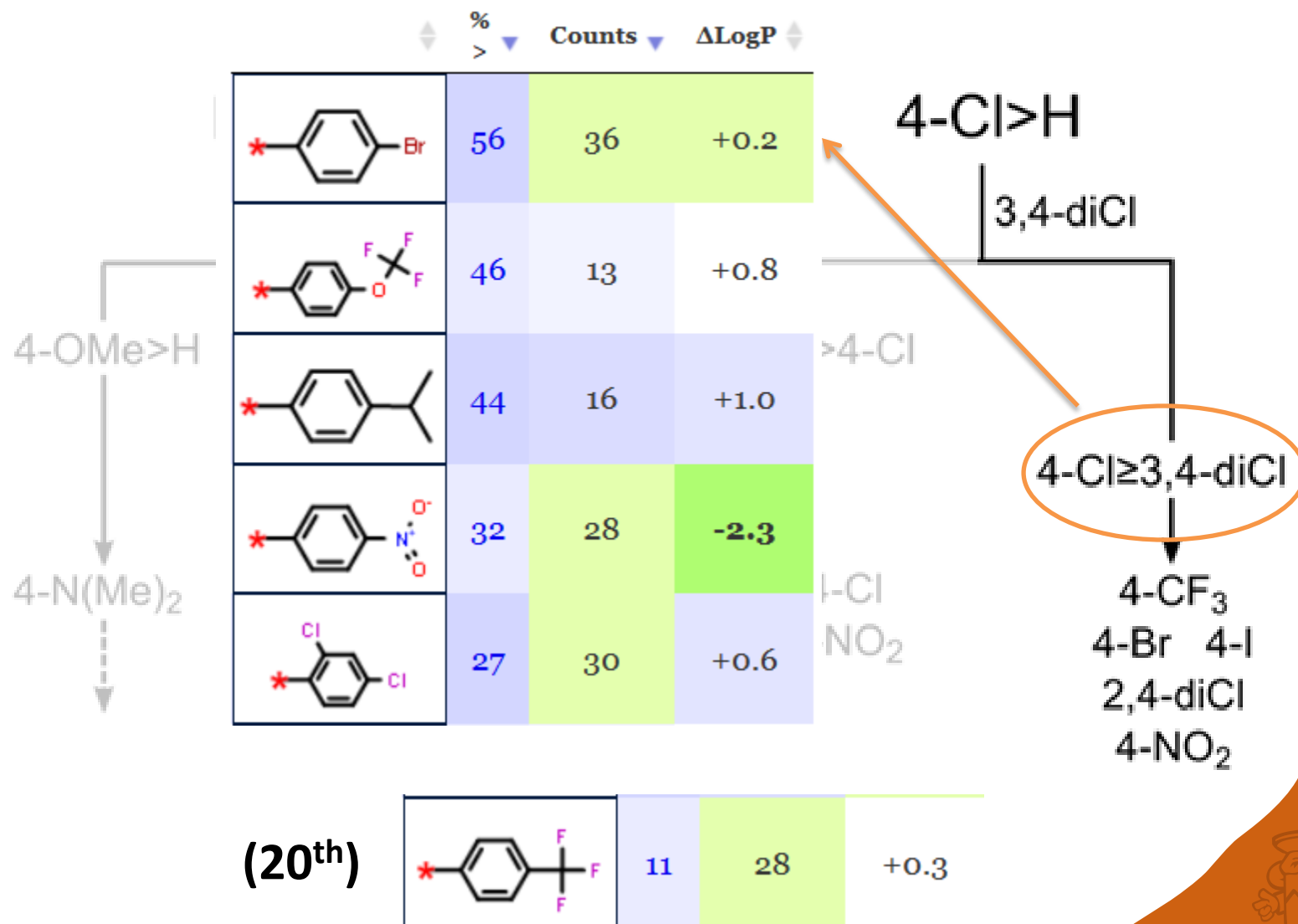
	33	5.00	15	+0.3
---	----	------	----	------



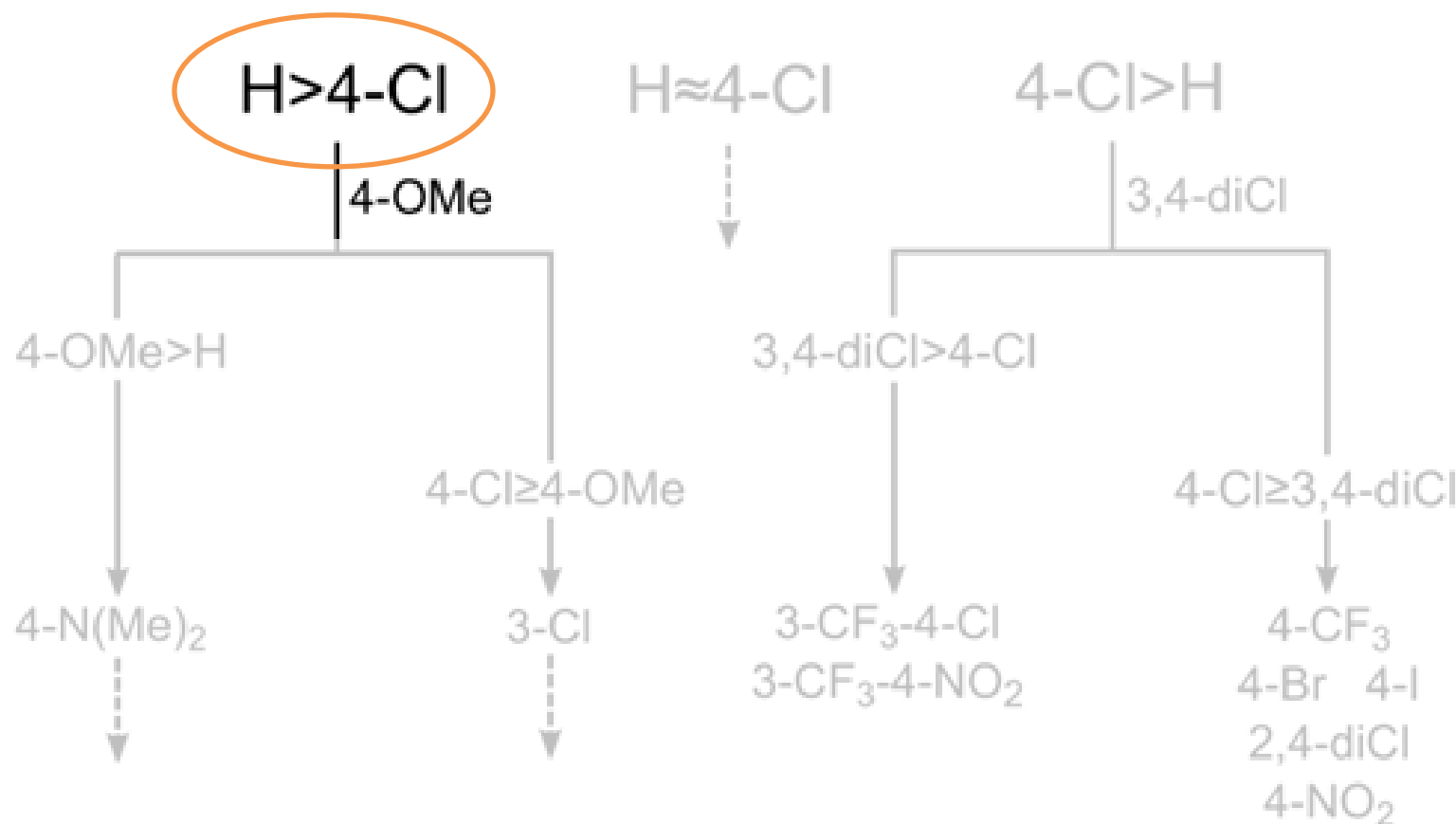
TOPLISS DECISION TREE



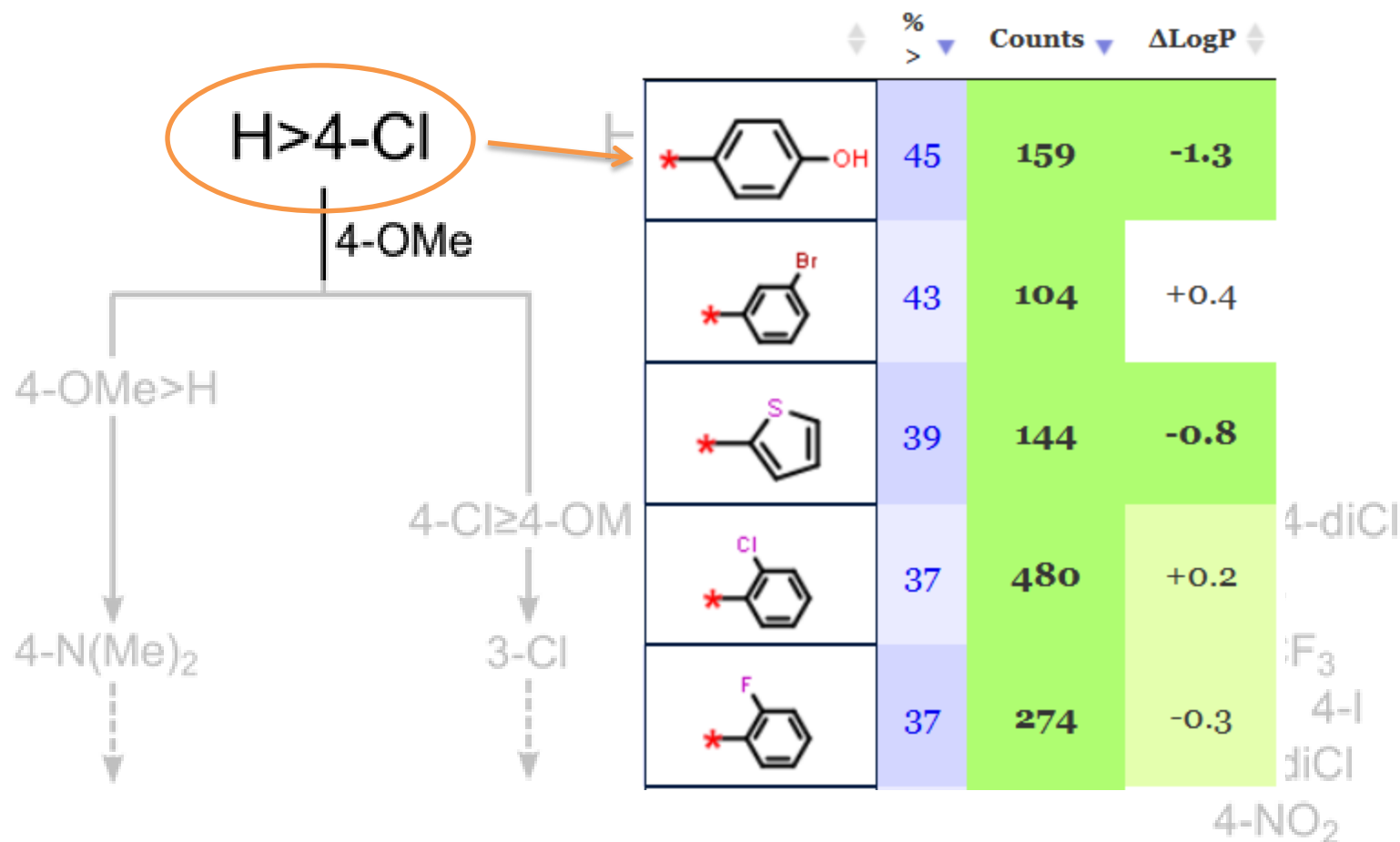
TOPLISS DECISION TREE



TOPLISS DECISION TREE



TOPLISS DECISION TREE



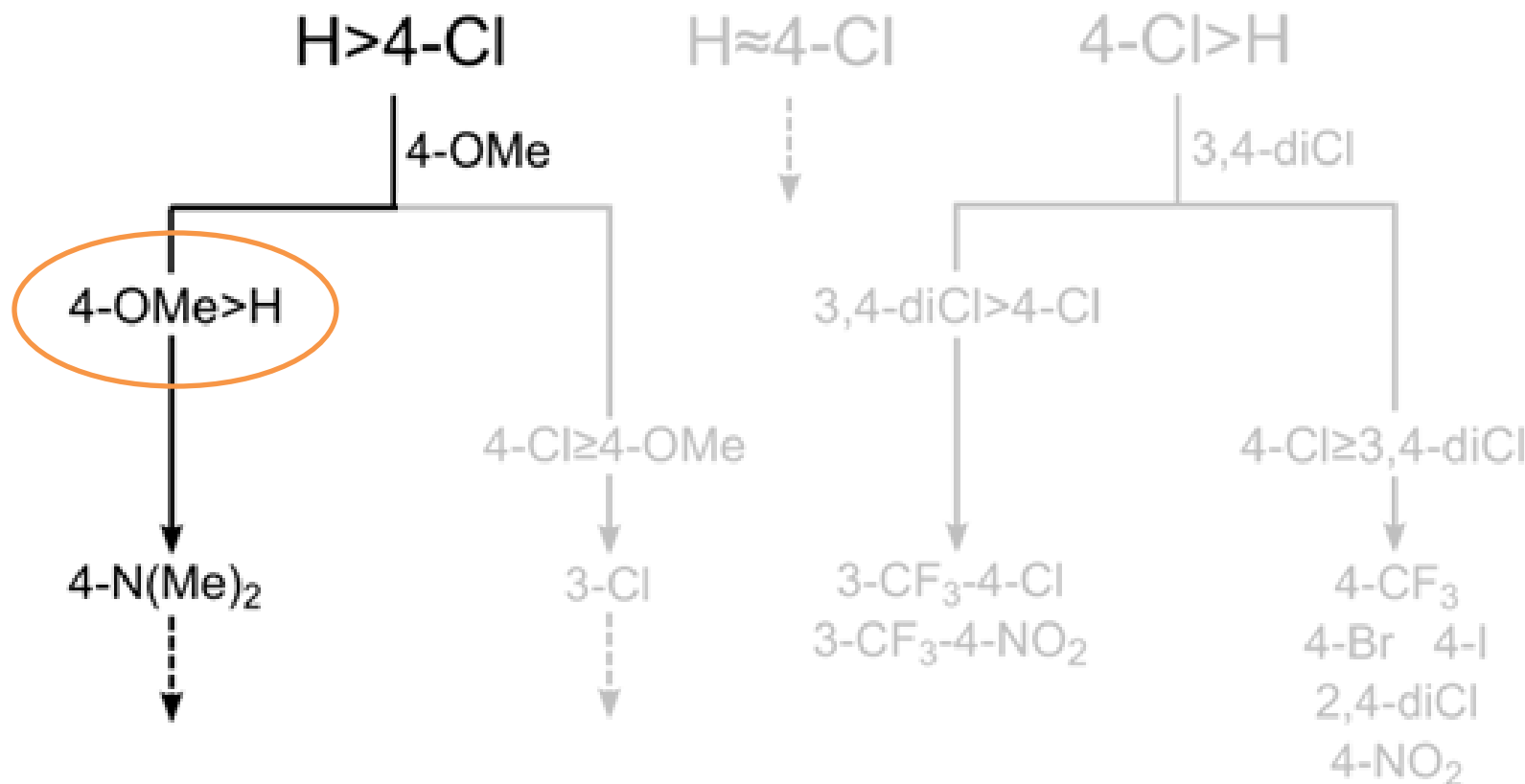
(21st)

	27	803	-0.5
--	----	-----	------

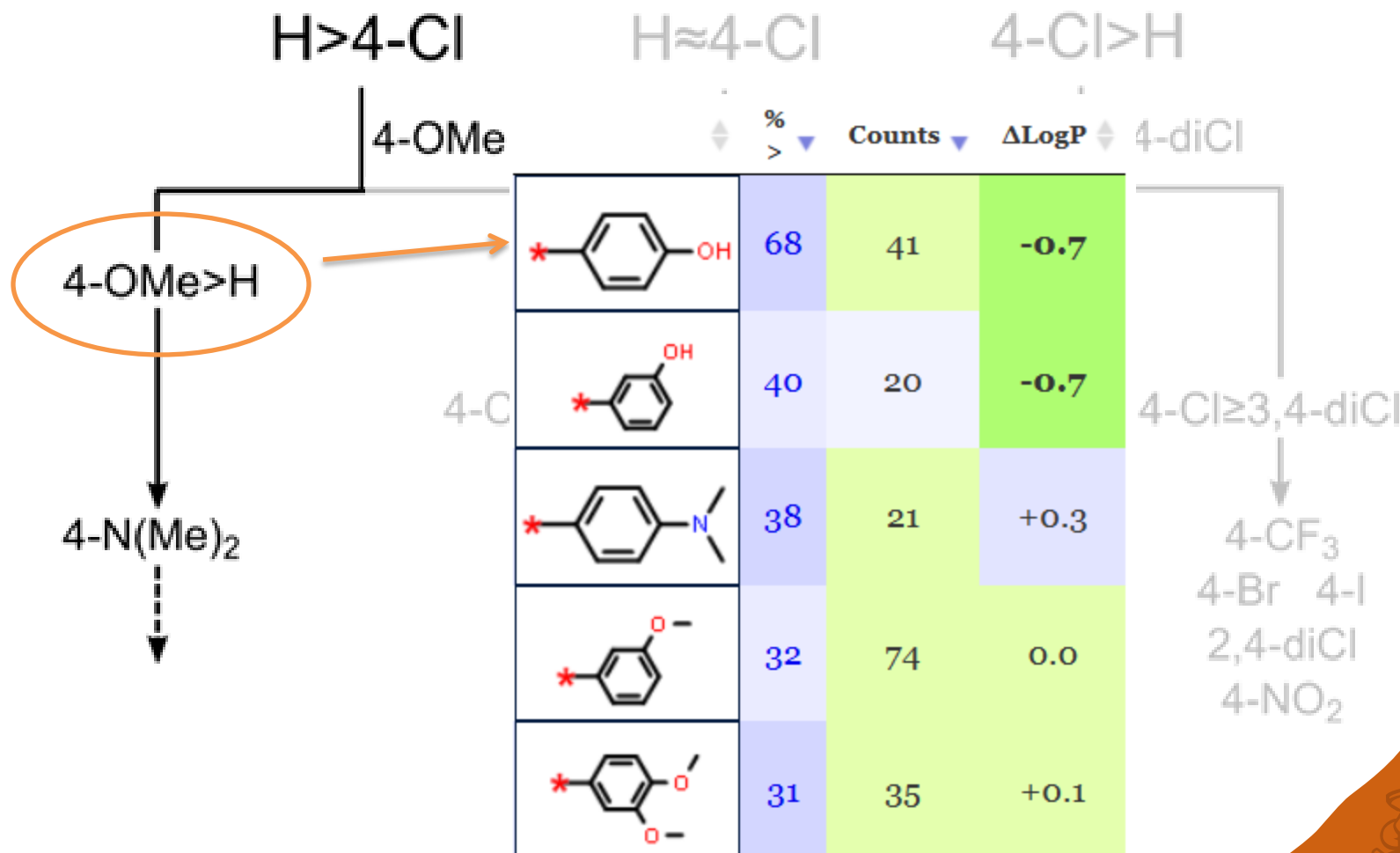
"Assuming that the -σ effect is the most probable explanation..."

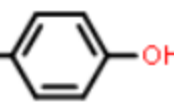
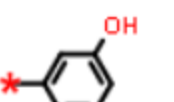
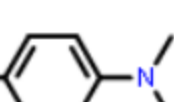
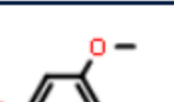
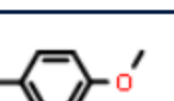


TOPLISS DECISION TREE



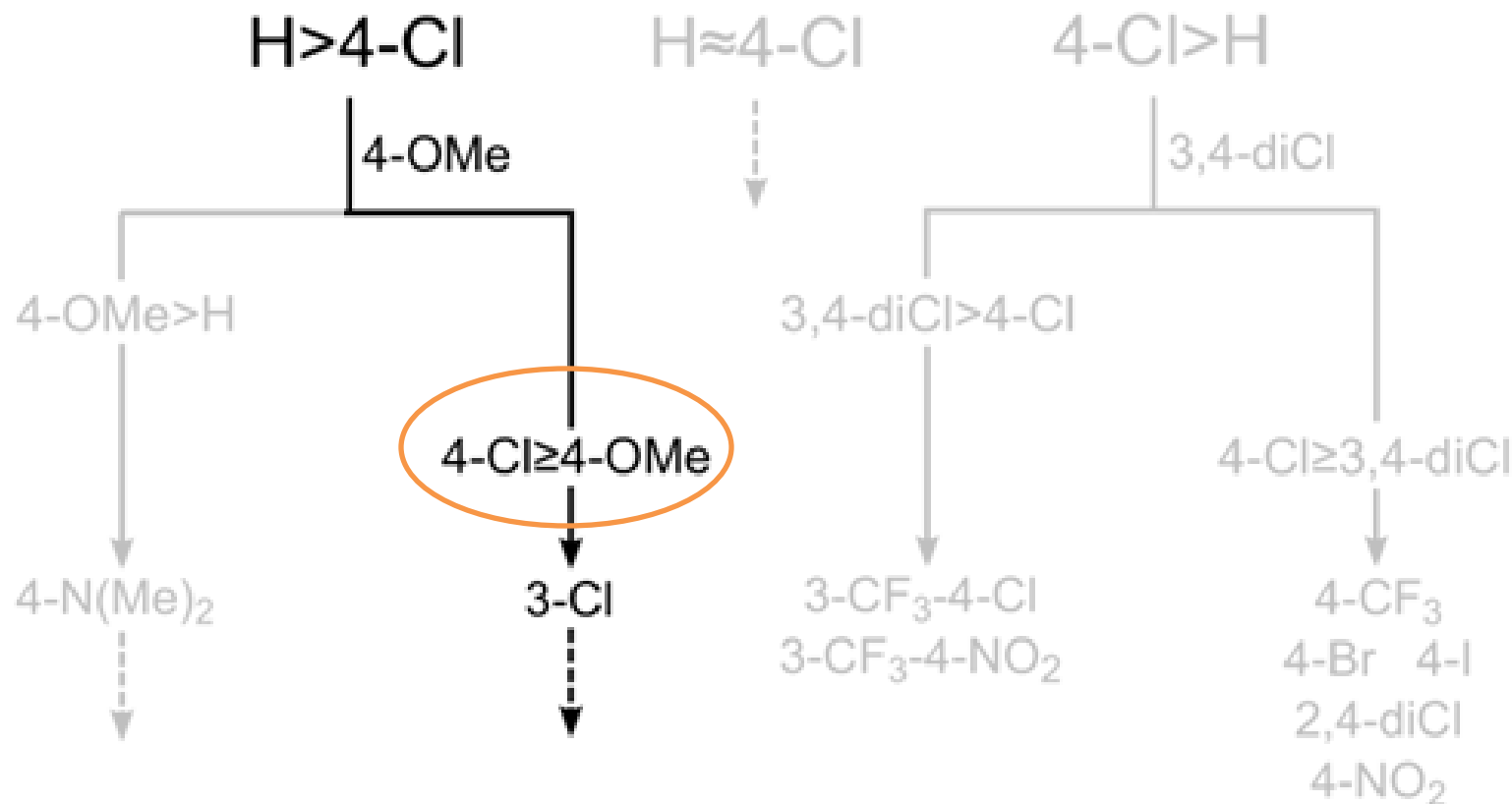
TOPLISS DECISION TREE



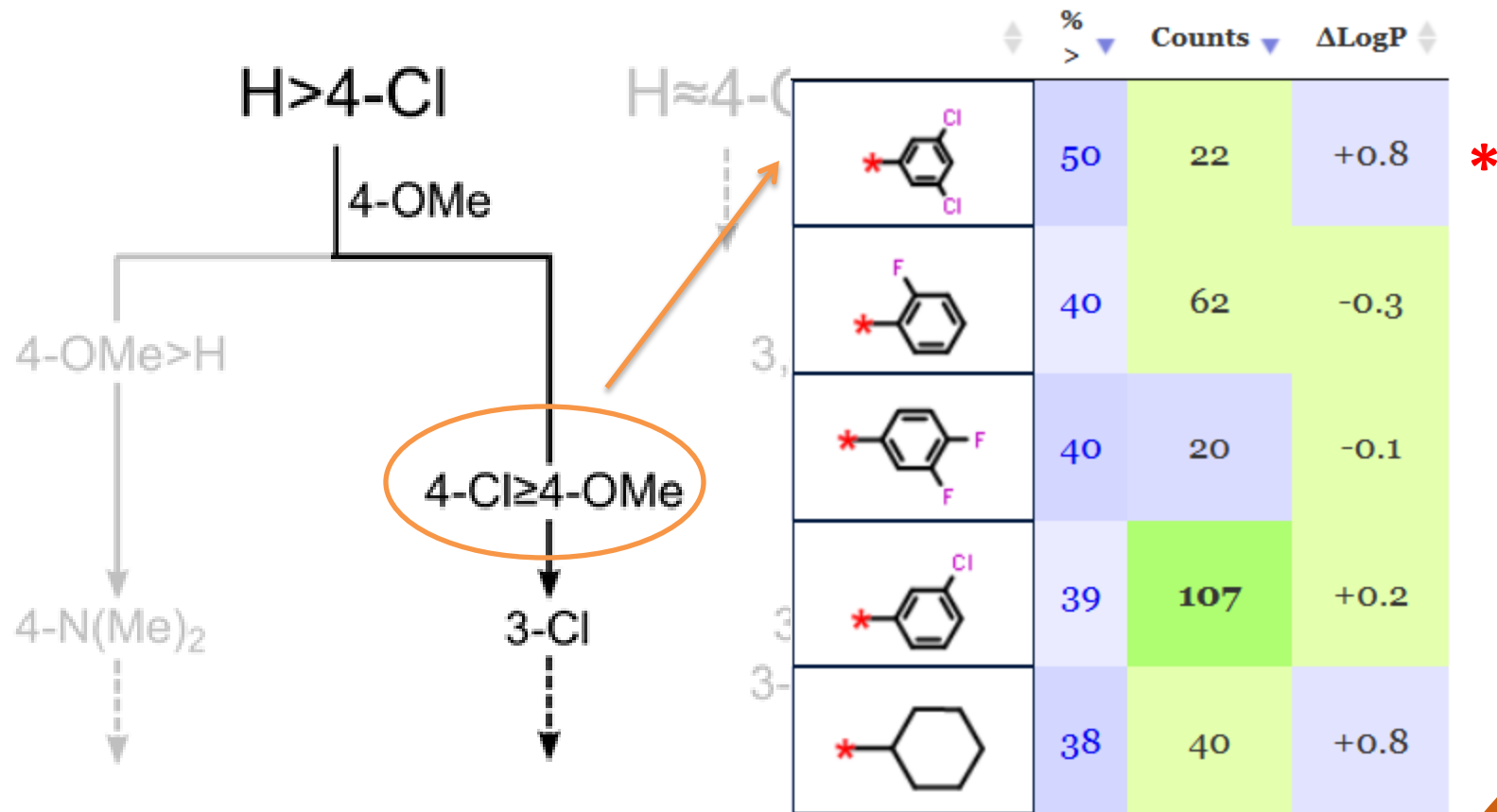
	%	Counts	ΔLogP
	68	41	-0.7
	40	20	-0.7
	38	21	+0.3
	32	74	0.0
	31	35	+0.1



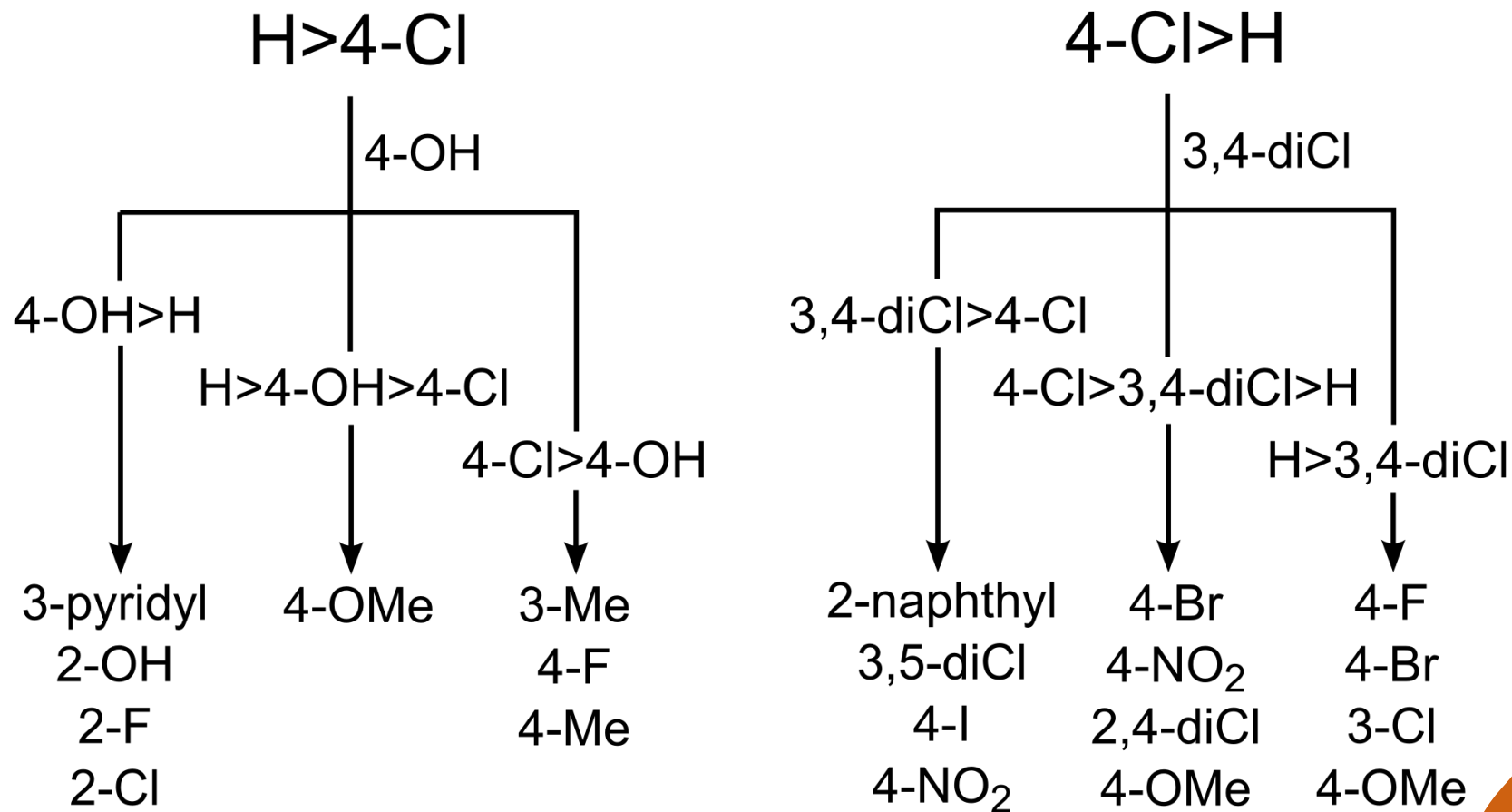
TOPLISS DECISION TREE



TOPLISS DECISION TREE



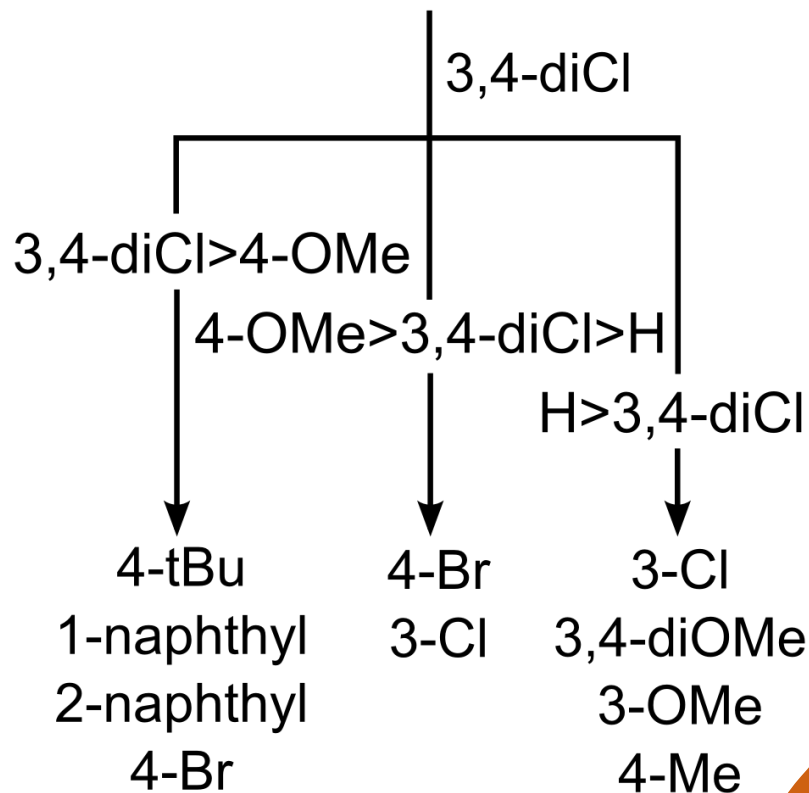
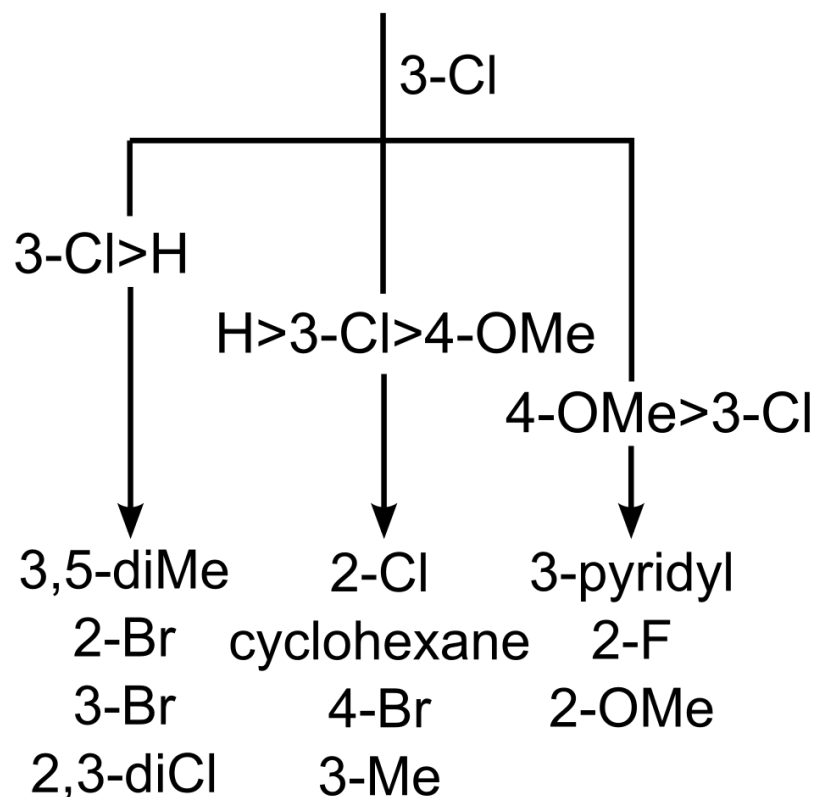
MATSY DECISION TREE



MATSY DECISION TREE (TAKE II)





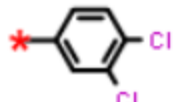

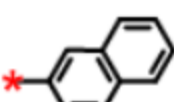
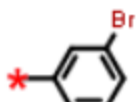
H>4-OMe

4-OMe>H




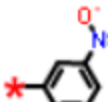
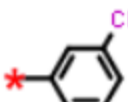


TARGET SPECIFIC SUBSETS




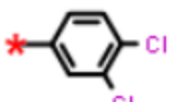
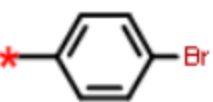

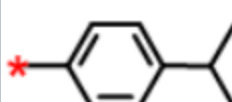
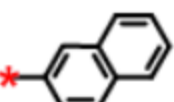
4-Cl > H

	 % > 	Counts 
	58	103
	54	391
	50	521
	44	212
	44	144

Everything

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

Kinases


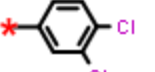

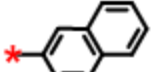
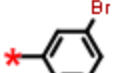
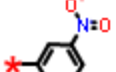
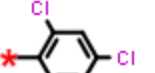
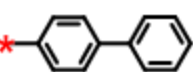
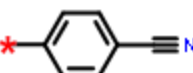
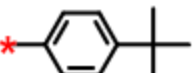
	 % > 	Counts 
	59	54
	48	40
	48	23
	48	21
	40	30

Class A GPCRs

ACCOUNT FOR LIPOPHILIC EFFICIENCY

- $\Delta\text{LiPE} = \Delta\text{pIC}_{50} - \Delta\text{LogP}$
- The “%>” value is based on the number of times a particular R group has greater pIC_{50}
 - i.e. $\Delta\text{pIC}_{50} > 0$
- Redefine it to only include cases where the increase in pIC_{50} was larger than any increase in LogP
 - i.e. $\Delta\text{pIC}_{50} > 0$ and $\Delta\text{LiPE} > 0$


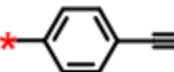

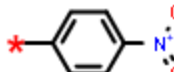
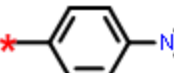

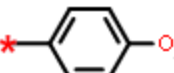
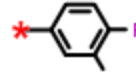
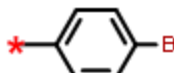
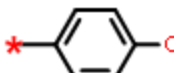


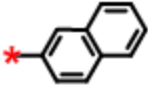

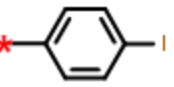
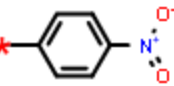
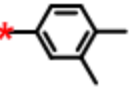
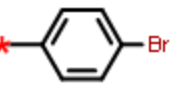
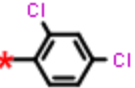

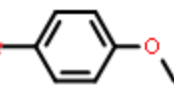

	% >	Counts	ΔLogP
	58	103	+0.5
	54	391	+0.6
	50	521	+0.2
	44	212	+1.0
	44	144	+0.2
	41	180	-2.3
	40	185	+0.6
	38	169	+2.0
	37	275	-0.8
	37	149	+1.6

4-Cl > H

$\Delta\text{LiPE} > 0$




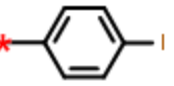
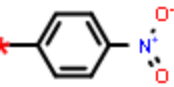
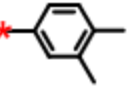
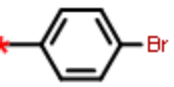


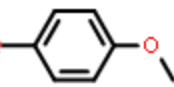

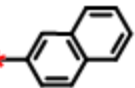
	% >	Counts	ΔLogP
	41	180	-2.3
	37	275	-0.8
	36	678	0.0
	35	347	-2.3
	33	110	-0.5
	32	104	-1.0
	30	958	-0.8
	27	124	-0.3
	27	521	+0.2
	26	172	-1.5

	% >	Counts	ΔLogP
	32	57	+0.4
	24	37	0.0
	22	23	-0.1
	21	47	-2.9
	20	20	-0.1
	18	83	-0.4
	17	42	0.0
	16	37	-1.4
	15	116	-1.4
	15	104	-0.6

3,4-diCl > 4-Cl > H

ΔLiPE > 0



	% >	Counts	ΔLogP
	24	37	0.0
	22	23	-0.1
	21	47	-2.9
	20	20	-0.1
	18	83	-0.4
	17	42	0.0
	16	37	-1.4
	15	116	-1.4
	15	104	-0.6
	14	57	+0.4

DATA-DRIVEN APPROACH

- **Not limited** to the two trees in the Topliss paper
- All predictions backed by **experimental data**
 - Can drill-down into the data, look at targets, scaffolds
 - Can restrict experimental data used to particular targets, use in-house data rather than ChEMBL
- Does not explain **why**, only that it happens



CONCLUSIONS

- In the main, the Topliss Tree is **supported** by published data
 - Largest difference is recommendation of 4-OMe rather than 4-OH
 - Suggestion of 4-CF₃ is also problematic
- We have generated the corresponding 'Matsy Tree' derived from experimental data



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

⚙ ?

⬆ ⬇ ⬆ Counts ΔLogP ⬆

*c1ccc(C)cc1

*c1cc(C)cc1

*c1cc(C)cc1

*c1ccc(N)cc1

*c1cc(N)cc1

*c1cc(N)cc1

*c1ccc(O)cc1

*c1cc(O)cc1

*c1cc(O)cc1

*c1ccc(F)cc1

*c1cc(F)cc1

*c1cc(F)cc1

*c1cc(Cl)cc1

*c1cc(Cl)cc1

*c1ccc(Br)cc1

*c1cc(Br)cc1

*c1cc(Br)cc1

*c1ccc(I)cc1

*c1cc(I)cc1

*c1cc(I)cc1

<chem>*c1ccc(I)cc1</chem>	58	103	+0.5
<chem>*c1cc(Cl)cc1</chem>	54	391	+0.6
<chem>*c1ccc(Br)cc1</chem>	50	521	+0.2
<chem>*c1ccc2ccccc2c1</chem>	44	212	+1.0
<chem>*c1cc(Br)cc1</chem>	44	144	+0.2
<chem>*c1cc([N+](=O)[O-])cc1</chem>	41	180	-2.3
<chem>*c1cc(Cl)cc1</chem>	40	185	+0.6
<chem>*c1ccc(cc1)-c2ccccc2</chem>	38	169	+2.0
<chem>*c1ccc(C#N)cc1</chem>	37	275	-0.8
<chem>*c1ccc(C(C)(C)C)cc1</chem>	37	149	+1.6

Stronger binding
ChEMBL19 pIC50
Weaker binding

?
>>

*c1cc(Cl)cc1

>

*c1ccc(Cl)cc1

>

*c1ccccc1

>

2277

Showing 1 to 10 of 40 entries
⬅ Previous Next ➡

Revising the Topliss decision tree

...based on 30 years of medicinal chemistry literature

noel@nextmovesoftware.com

Want to hear more?

Poster COMP 394

Tuesday 6:00-8:00pm Marriott Marquis

Interested in an evaluation copy of Matsy?

Come by our booth



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

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

