# Applying Matsy to predict new optimisation strategies

**14th April 2015**

# Todays speakers

- Noel O'Boyle – NextMove Software Limited
  - Senior Software Engineer
  - PhD in Computation Chemistry
  - Postdoctoral positions at Cambridge University, CCDC and University College Cork
- Ed Champness – Optibrium Limited
  - Chief Scientific Officer
  - Formerly GlaxoWellcome, Camitro and BioFocus
  - Co-founded Optibrium


Optibrium Webinar
Cambridge, Apr 2015

# Beyond matched pairs

## Applying Matsy to predict new optimisation strategies…

**Noel O'Boyle**

NextMove Software

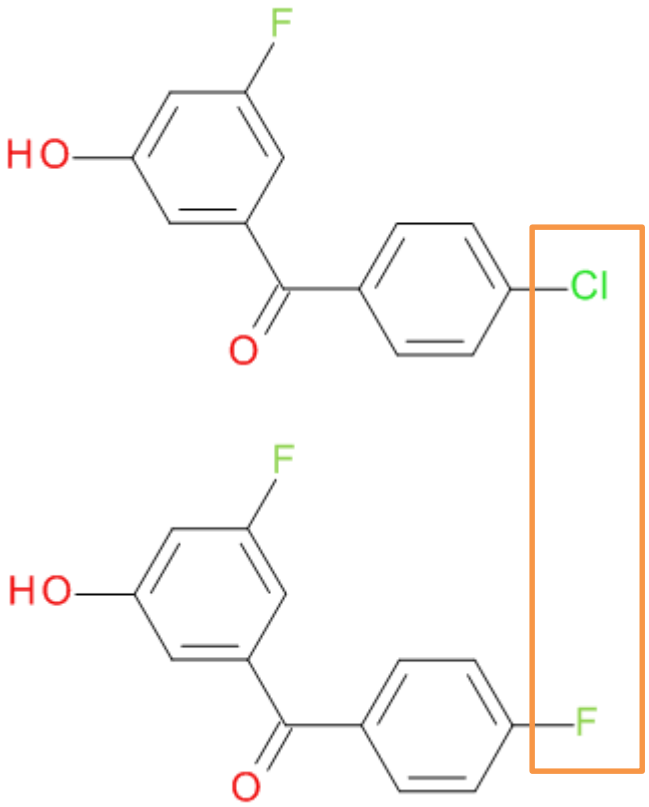
**Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity** *J. Med. Chem.* **2014**, *57*, 2704.

# 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 suggestion:
  - Take advantage of the wealth of *experience* and *trends* contained in 60K 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

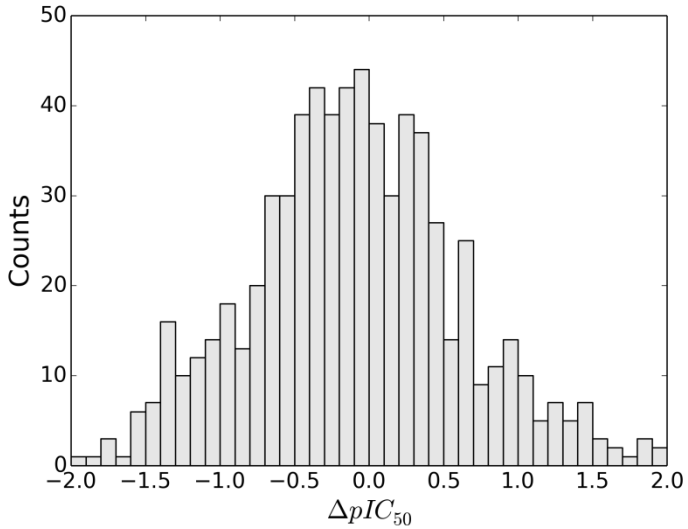
* Chemoinformatics in drug discovery, Wiley, 271–285.

# 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

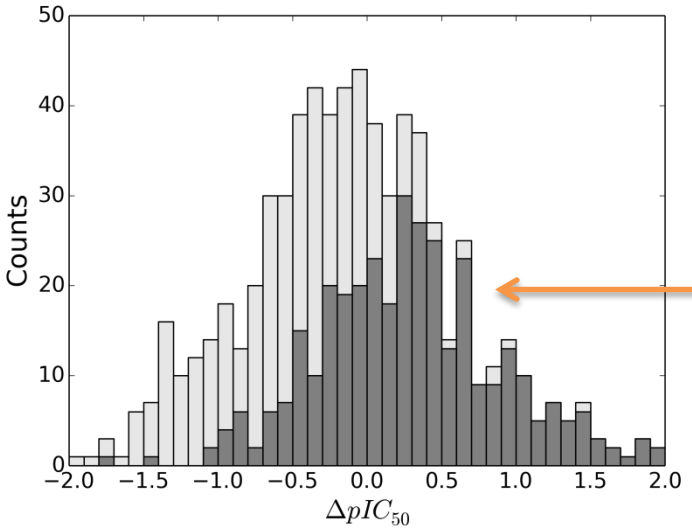
# MATCHED PAIRS AND ACTIVITY

$pIC_{50}(Et)-pIC_{50}(Bu)$




# MATCHED PAIRS AND ACTIVITY

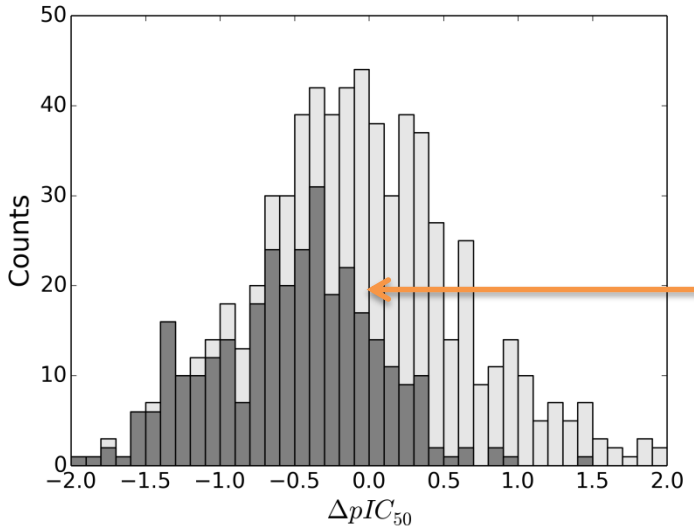
$pIC_{50}(Et)-pIC_{50}(Bu)$




For those cases where:
[Pro > Bu]

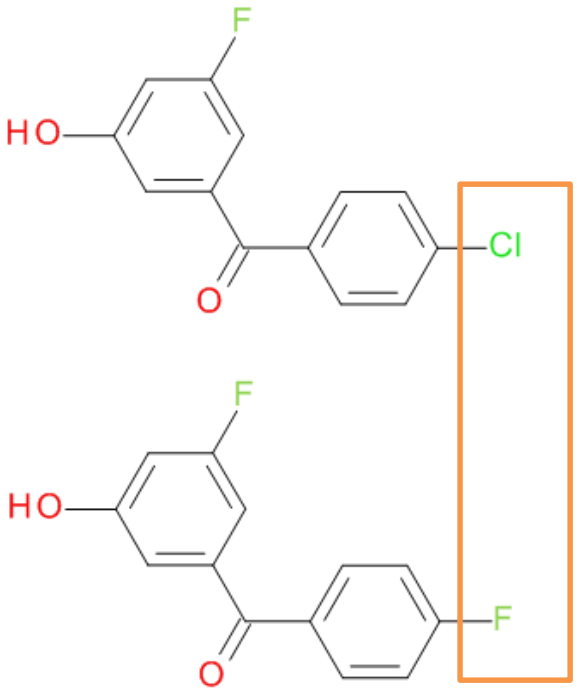
# MATCHED PAIRS AND ACTIVITY

$pIC_{50}(Et)-pIC_{50}(Bu)$




For those cases where:
[Pro < Bu]

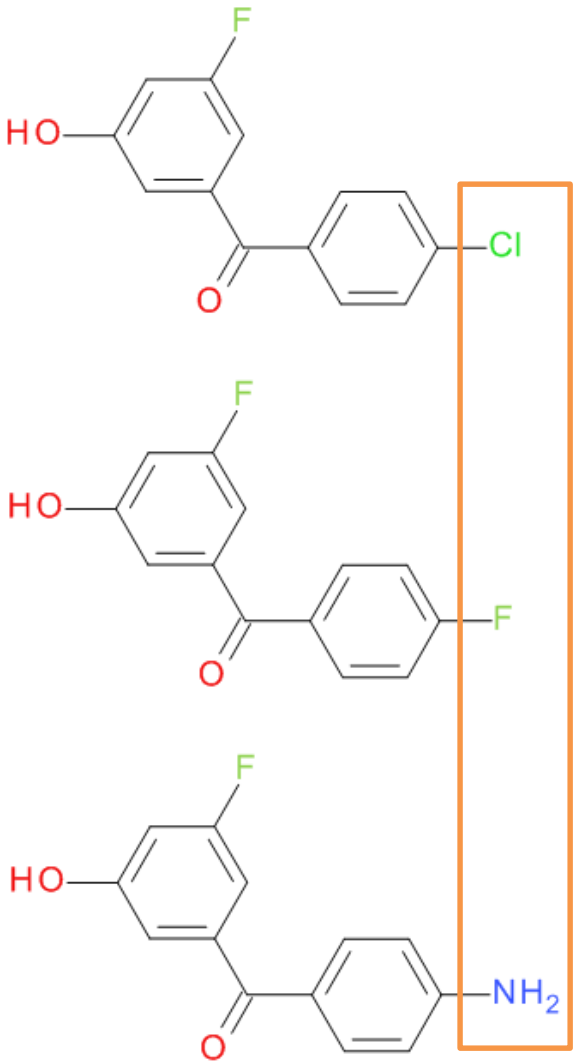
# MATCHED SERIES OF LENGTH 2 = MATCHED PAIR




**[Cl, F]**

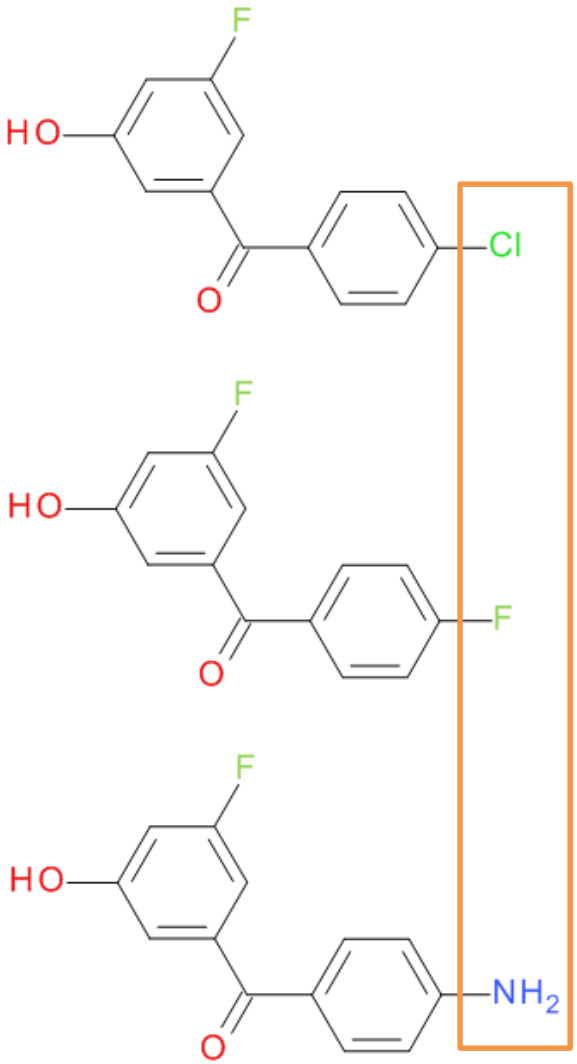
"Matching molecular series" introduced by Wawer and Bajorath, *J. Med. Chem.* **2011**, *54*, 2944

# MATCHED SERIES OF LENGTH 3




**[Cl, F, $NH_2$]**

# ORDERED MATCHED SERIES OF LENGTH 3



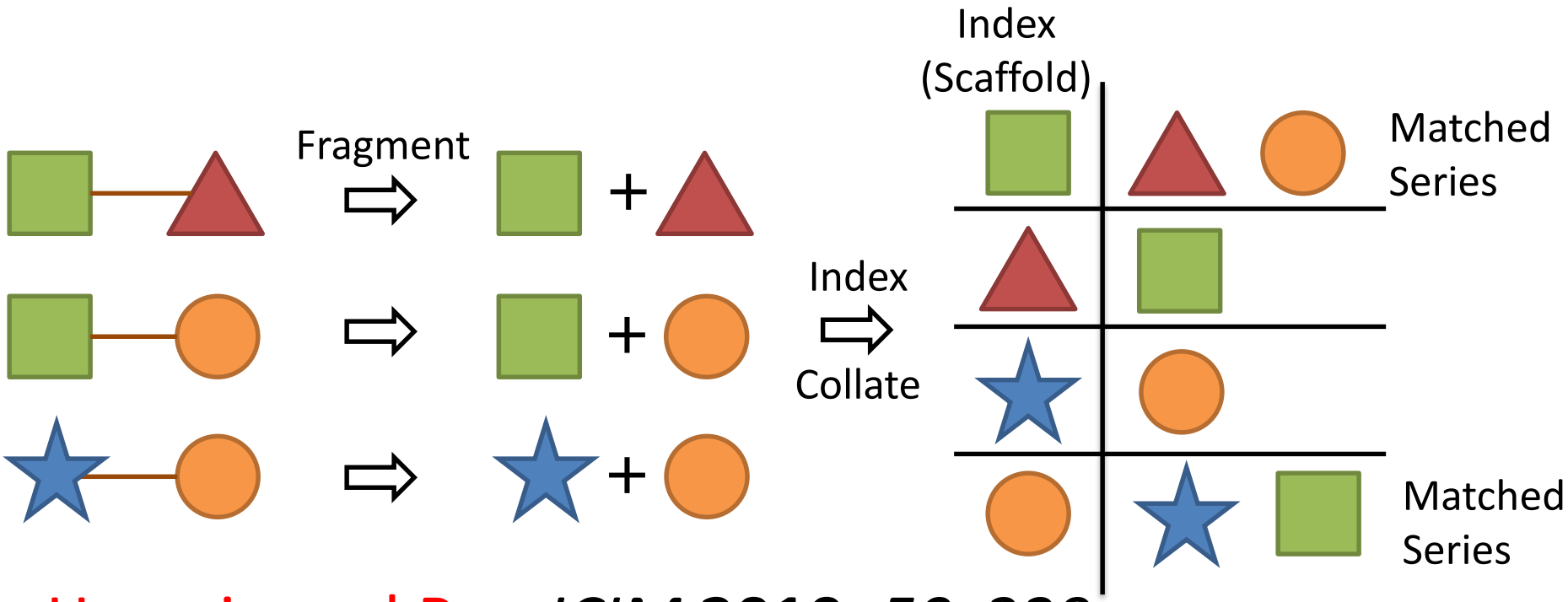

| | $pIC_{50}$ |
|---|---|
| Cl | 3.5 |
| F | 2.1 |
| $NH_2$ | 1.6 |

**[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, 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

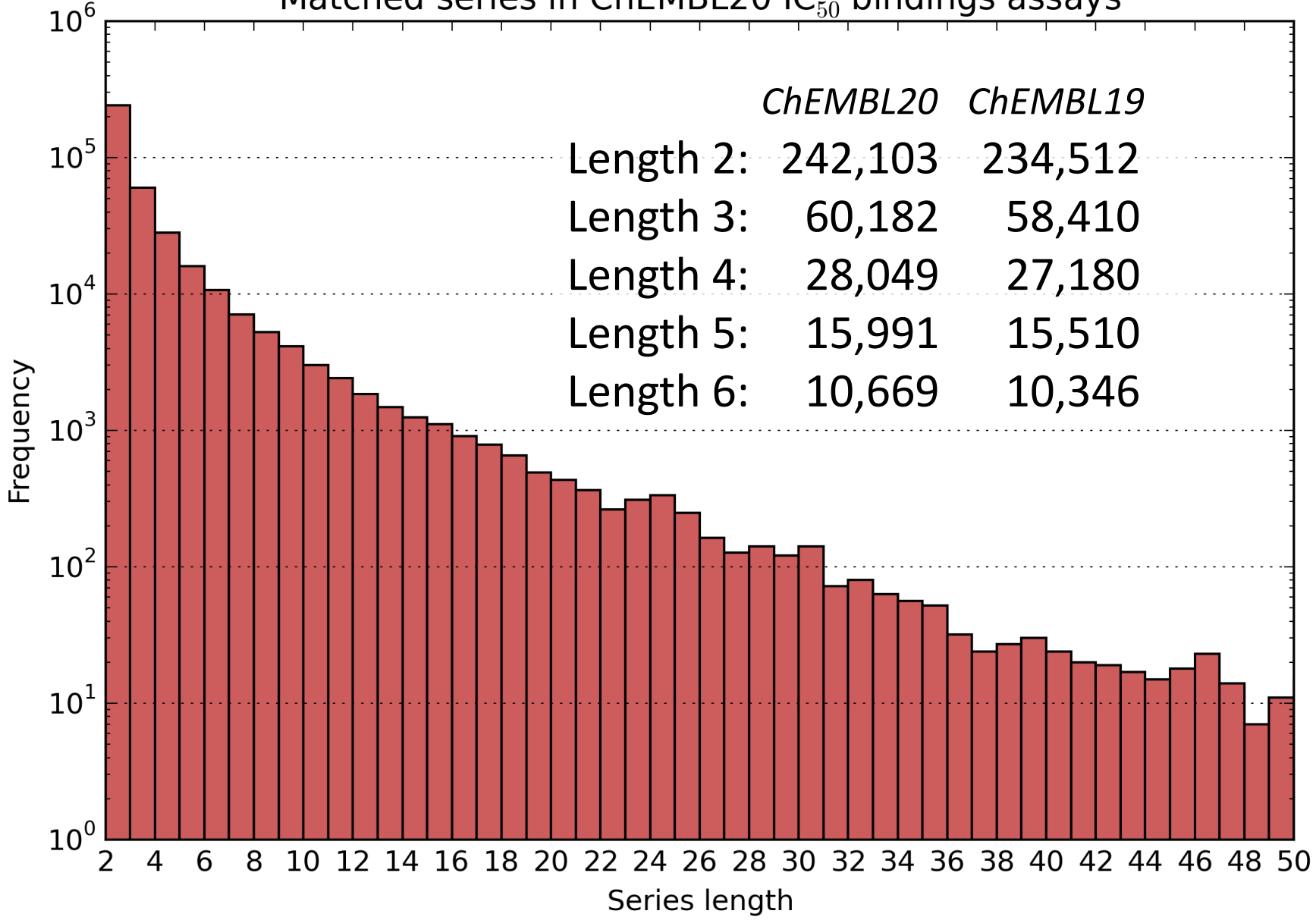



- Hussain and Rea *JCIM* **2010**, *50*, 339
  - Fragment molecules at acyclic single bonds
    - Single-cut only, scaffold >= 5, R group <= 12, preserve stereochemistry at break point
  - Index each fragment based on the other
    - A matched series will be indexed together

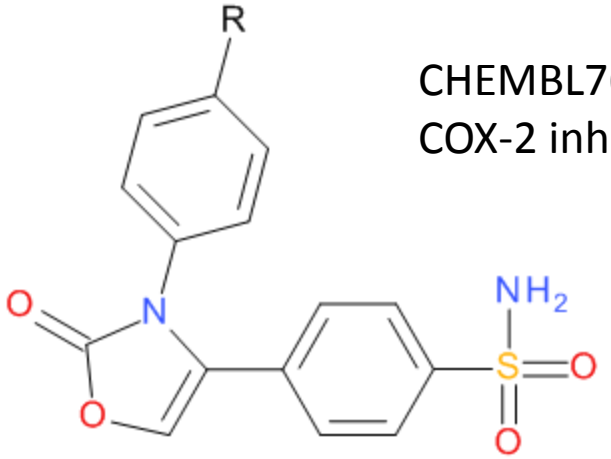
# CHEMBL BIOACTIVITY DATABASE

- ChEMBL 20 – Feb 2014
  - 60k 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.*
  - PK data from AstraZeneca, NTD screening data from Novartis, GSK, and others
  - 1.7 million compounds with 13.5 million activities
  - 1.1 million assays against 11k targets

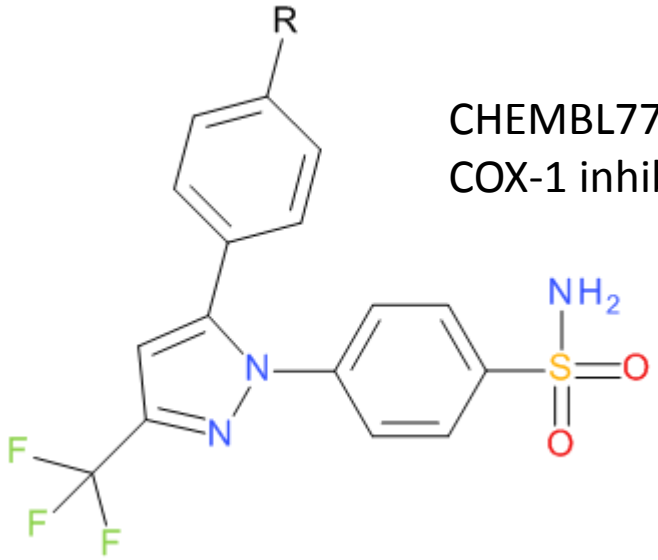
Gaulton et al. *Nucleic Acids Res.* **2012**, *40*, D1100


Matched series in ChEMBL20 IC50 bindings assays
ChEMBL20
ChEMBL19
Length 2: 242,103 234,512
Length 3: 60,182 58,410
Length 4: 28,049 27,180
Length 5: 15,991 15,510
Length 6: 10,669 10,346
Frequency
Series length

# 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 | Rank order | 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 |

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 | 9639 |
| H > F | 8558 |

Would expect 9098 for each assuming the order is random

- We can calculate enrichment

*Dataset is ChEMBL20 $IC_{50}$ 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* | 9639 |
| H > F | 0.94* | 8558 |

Would expect 9098 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* | 1455 |
| H > F > Cl | 1.06 | 811 |
| F > Cl > H | 0.89* | 685 |
| Cl > H > F | 0.77* | 587 |
| F > H > Cl | 0.76* | 582 |
| H > Cl > F | 0.63* | 480 |

# PREFERRED ORDERS: HALIDES (N=4)

| 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** |

N=2: Max = 1.06, Min = 0.94
N=3: Max = 1.90, Min = 0.63
N=4: Max = 5.36, Min = 0.17

Longer series exhibit greater preferences

If [H>F>Cl] is observed, will Br increase activity further?
141 observations of [H>F>Cl] but only 8 where [Br>H>F>Cl]

# MATSY:
# PREDICTION USING MATCHED SERIES

# FIND R GROUPS THAT INCREASE ACTIVITY

ChEMBL

In-house

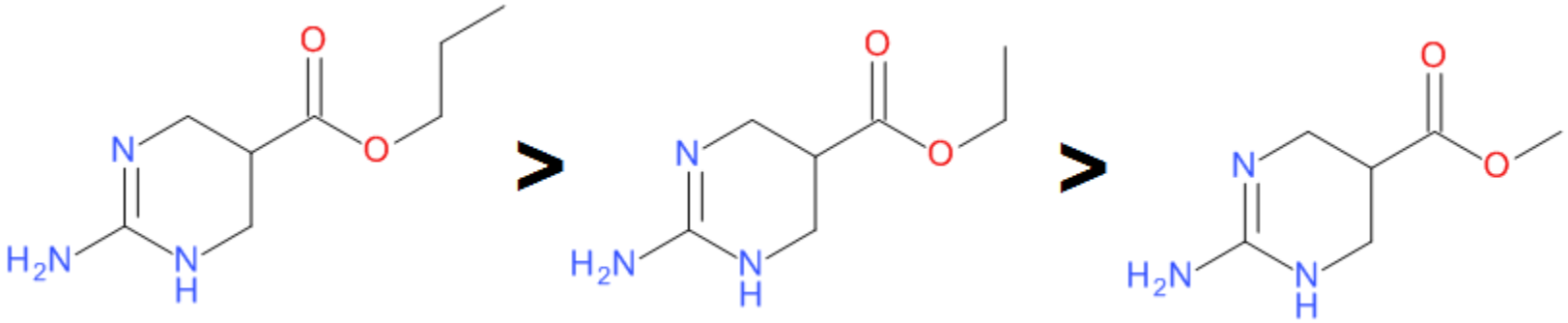
Query

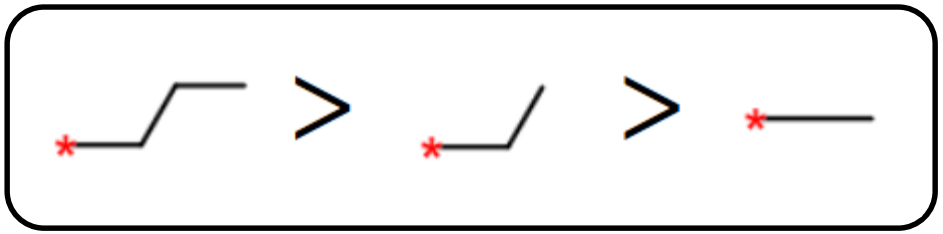
**A > B**

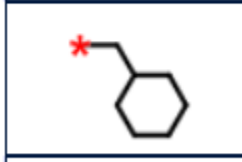

**A > B > C**
C > **A > B**
D > **A > B** > C
D > **A** > C > **B**
E > D > **A > B**
…

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

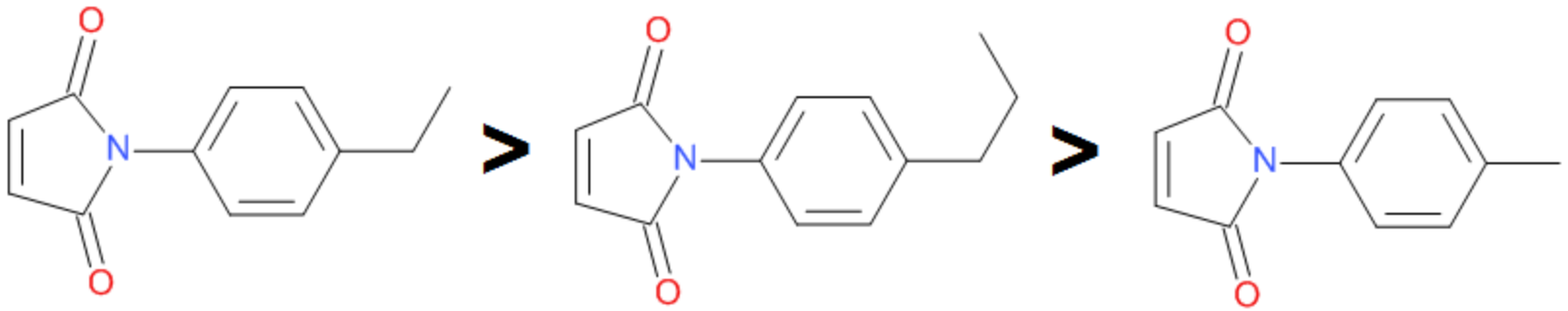
# EXAMPLE

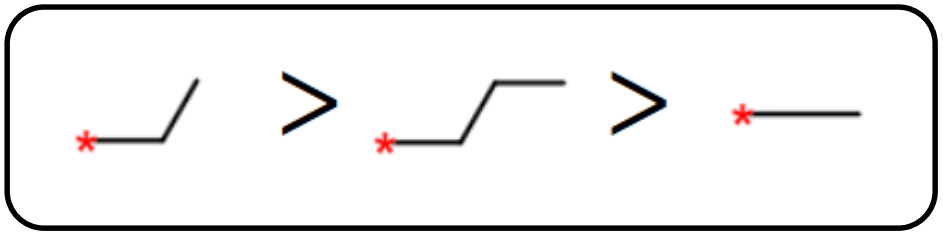


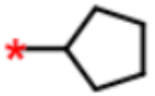
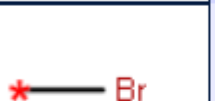
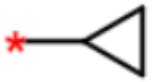
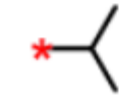


| | % > | Counts | ΔLogP |
|---|---|---|---|
|  | 90 | 21 | +3.3 |
|  | 73 | 59 | +1.7 |
|  | 69 | 32 | +2.8 |
|  | 62 | 26 | +1.6 |
|  | 59 | 44 | +2.2 |

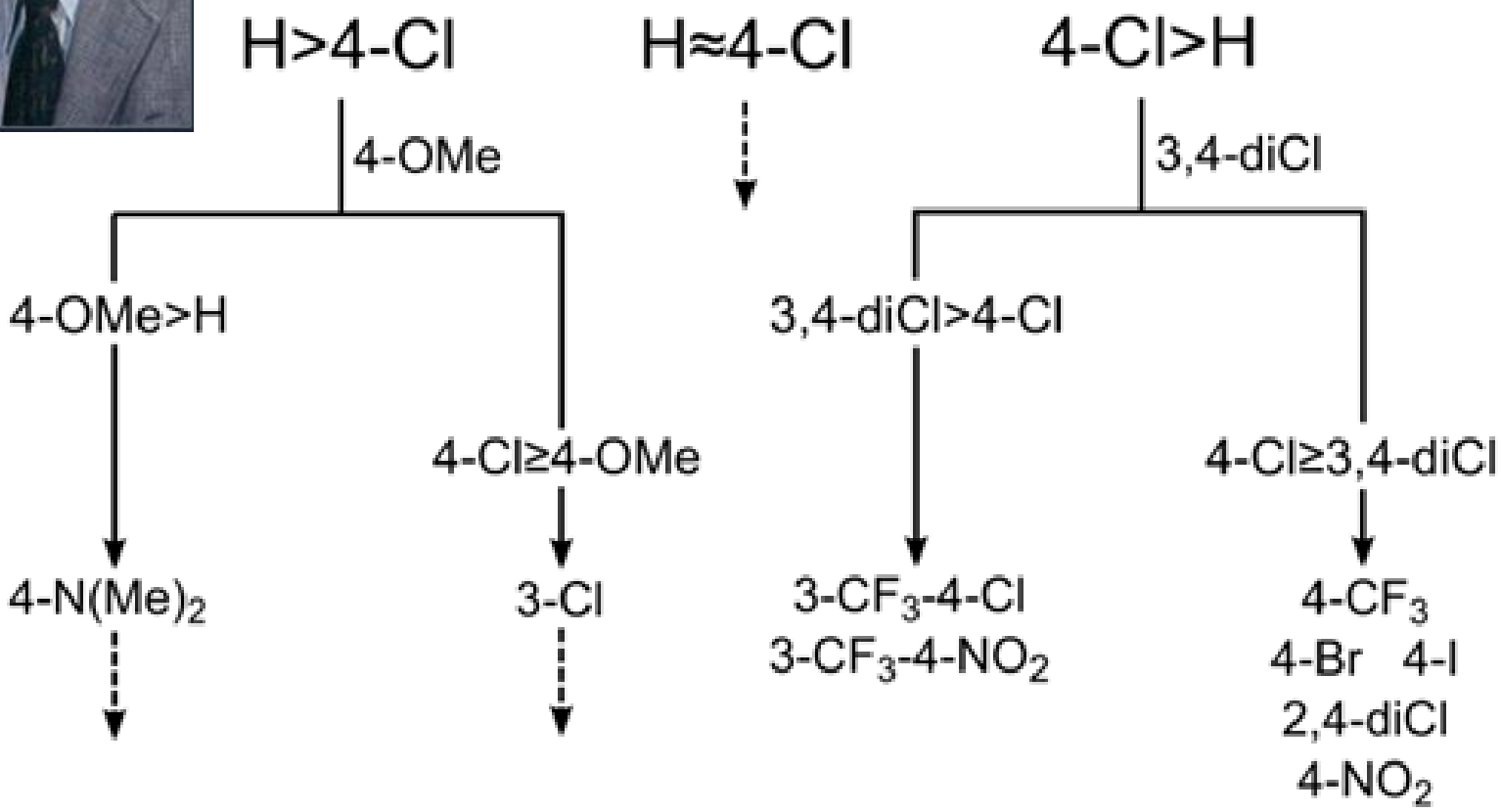
# EXAMPLE II





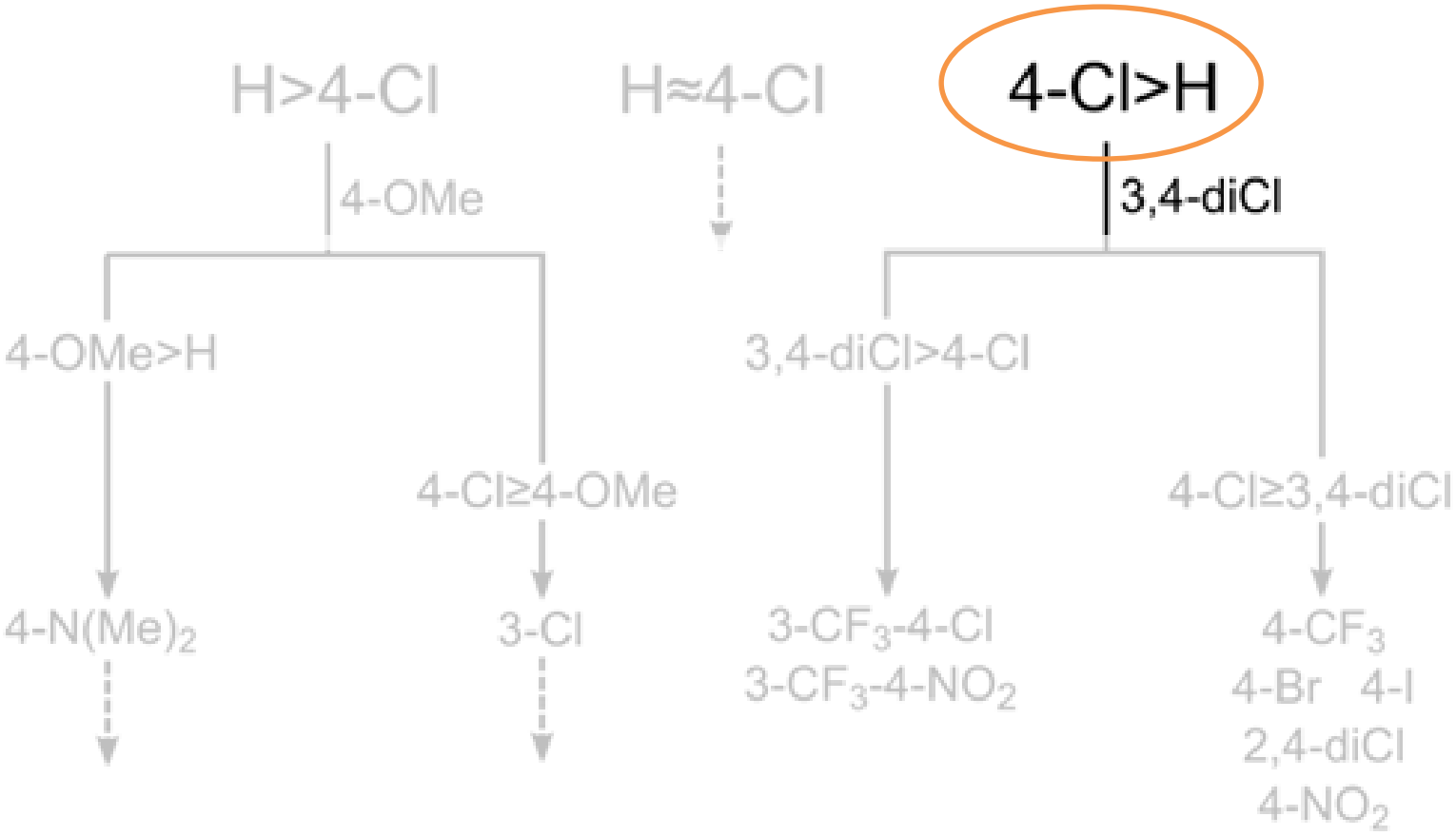
| | % > | Counts | ΔLogP |
|---|---|---|---|
|  | 39 | 23 | +1.0 |
|  | 38 | 21 | **-0.8** |
|  | 35 | 26 | +0.9 |
|  | 35 | 20 | -0.1 |
|  | 33 | **107** | +0.3 |

# TOPLISS DECISION TREE

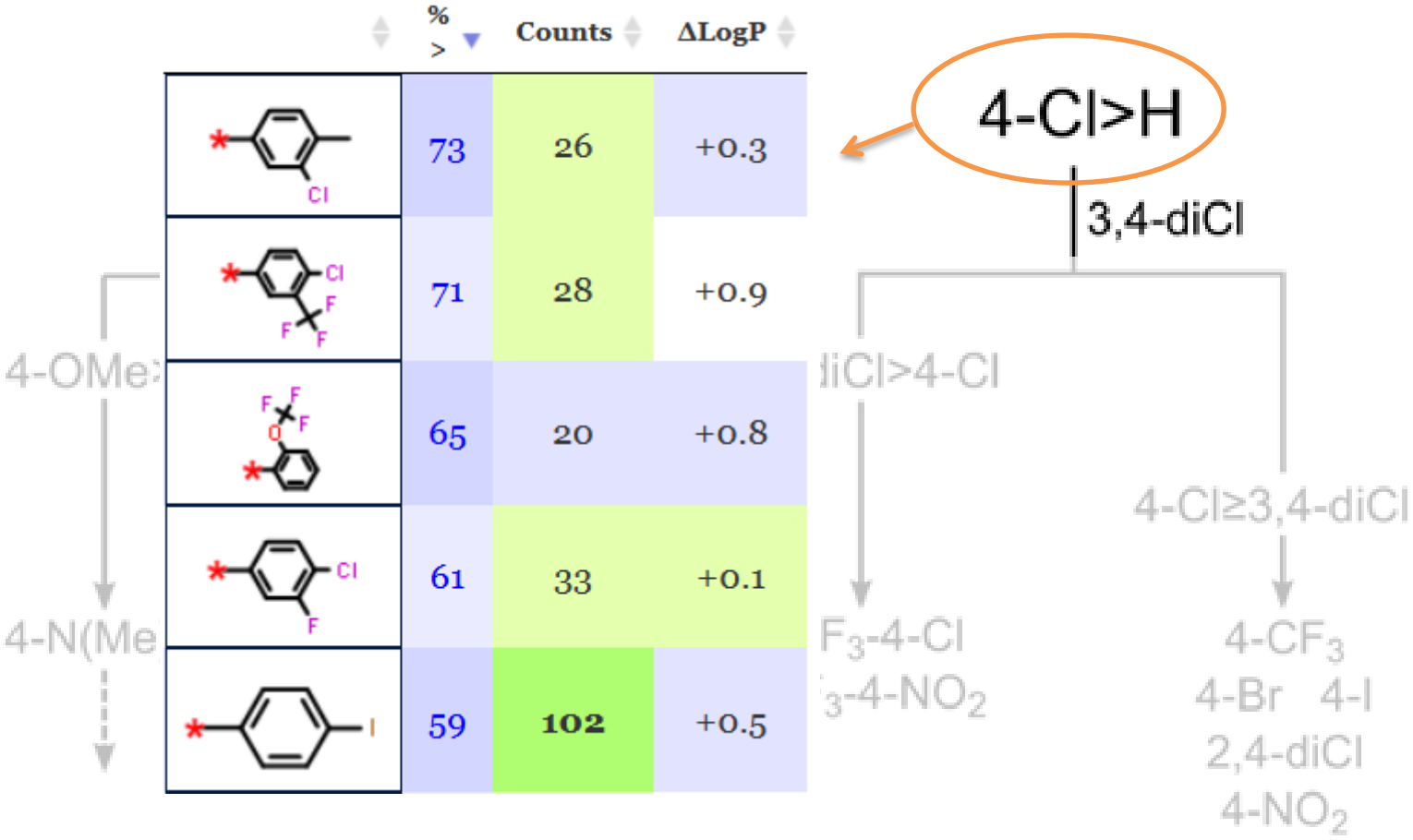


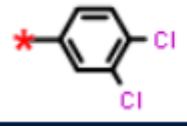

Topliss, J. G. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Chem.* **1972**, *15*, 1006–1011.

# TOPLISS DECISION TREE



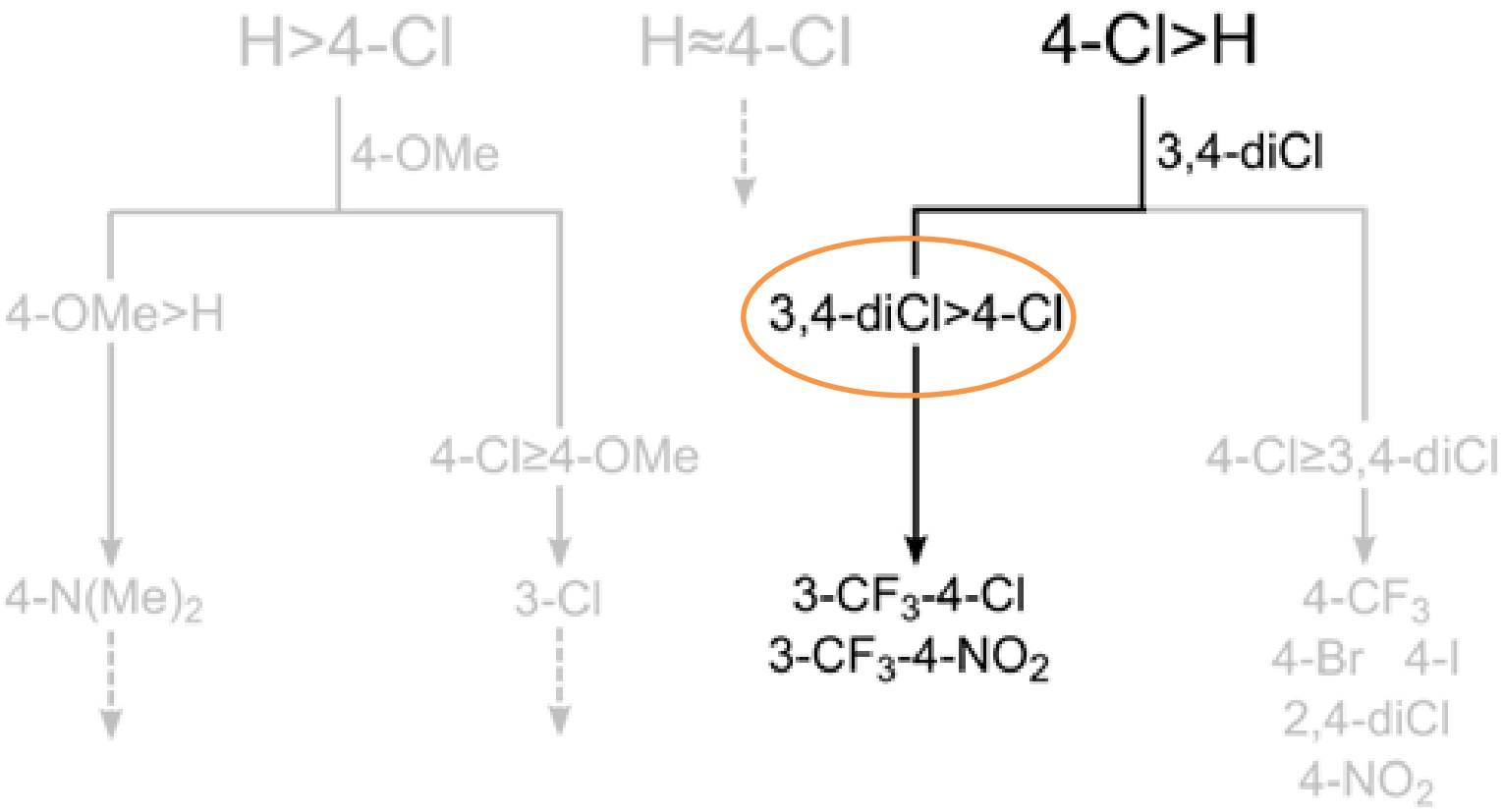
# TOPLISS DECISION TREE



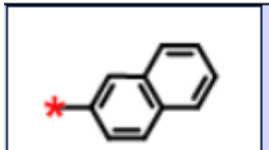
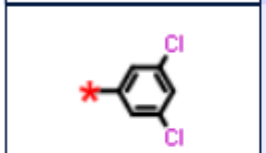
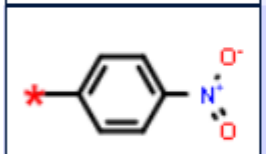
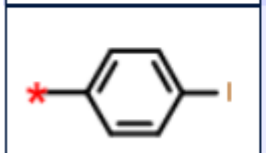
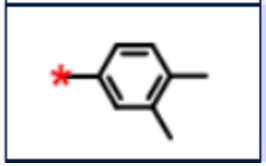

| | % > | Counts | ΔLogP |
|---|---|---|---|
|  | 73 | 26 | +0.3 |
|  | 71 | 28 | +0.9 |
|  | 65 | 20 | +0.8 |
|  | 61 | 33 | +0.1 |
|  | 59 | 102 | +0.5 |
|  | 56 | 380 | +0.6 |

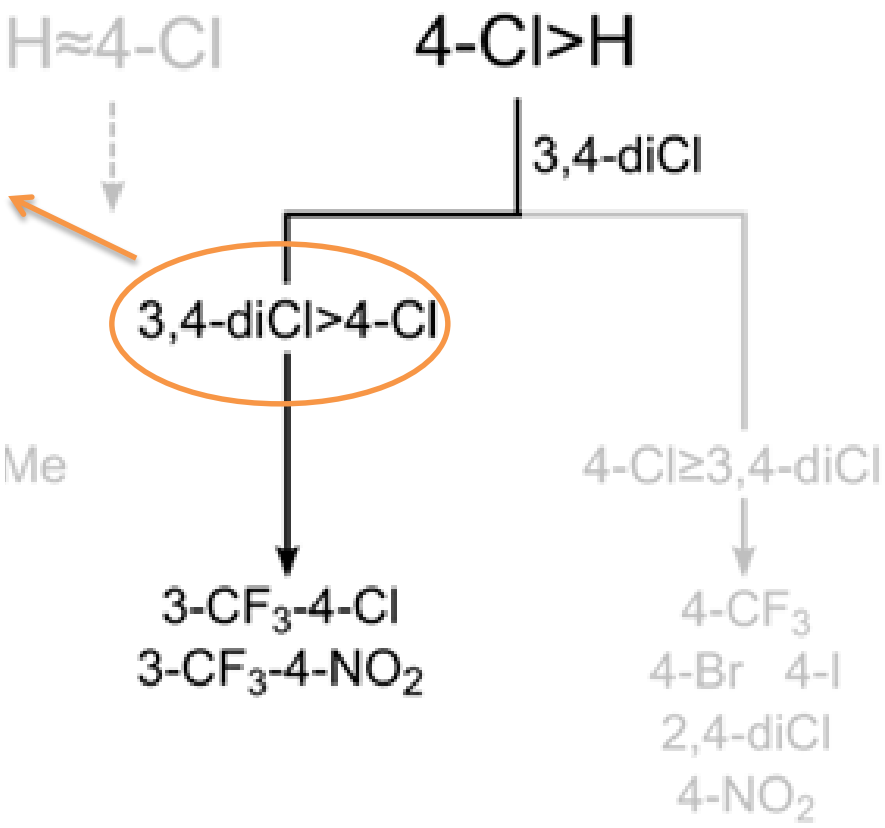
(9th)

# TOPLISS DECISION TREE

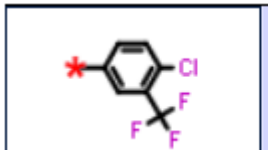


# TOPLISS DECISION TREE

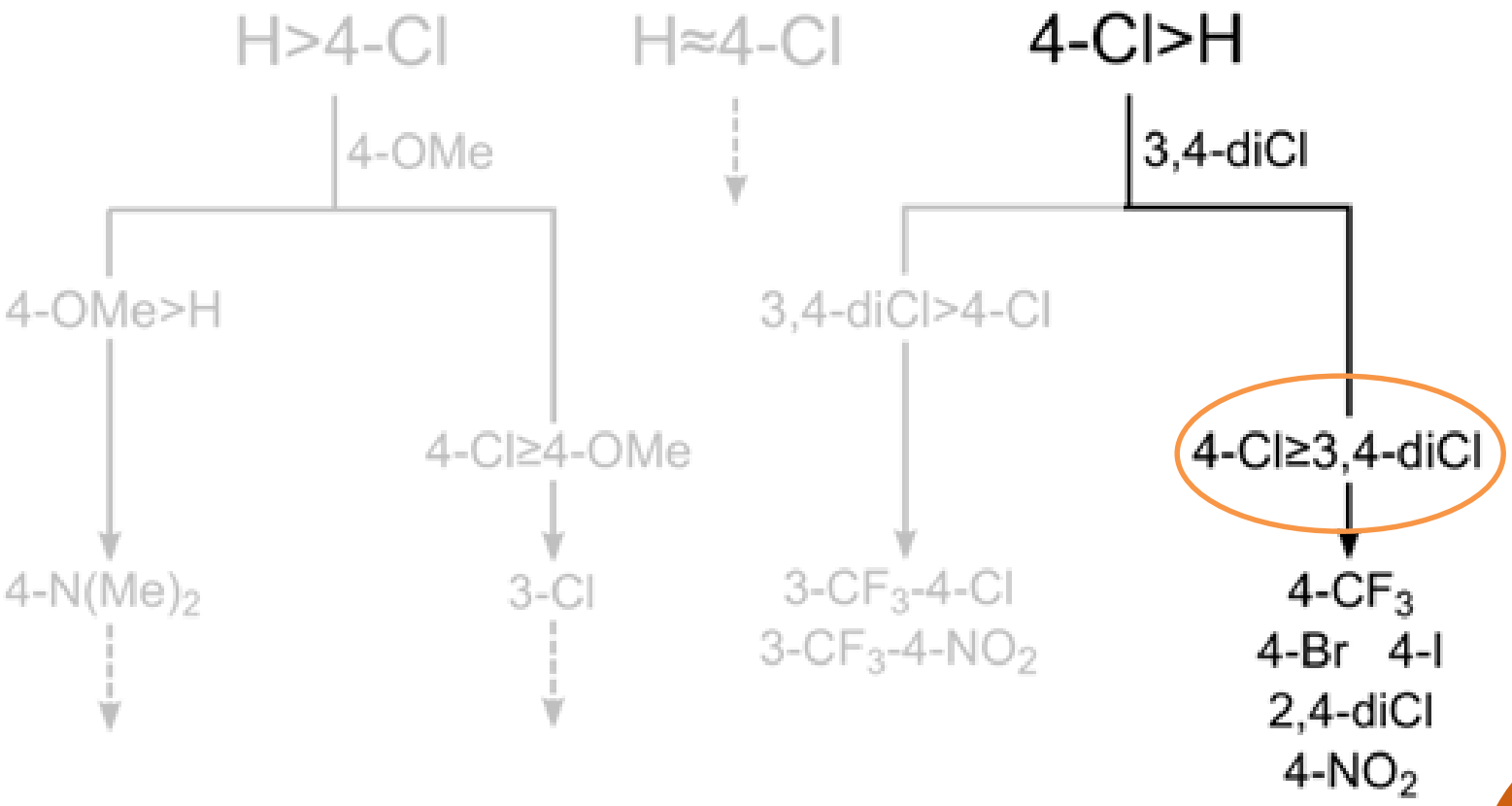
| | % > | Counts | ΔLogP |
|---|---|---|---|
|  | 32 | 56 | +0.4 |
|  | 24 | 38 | 0.0 |
|  | 23 | 47 | **-2.9** |
|  | 23 | 22 | -0.1 |
|  | 19 | 21 | -0.1 |



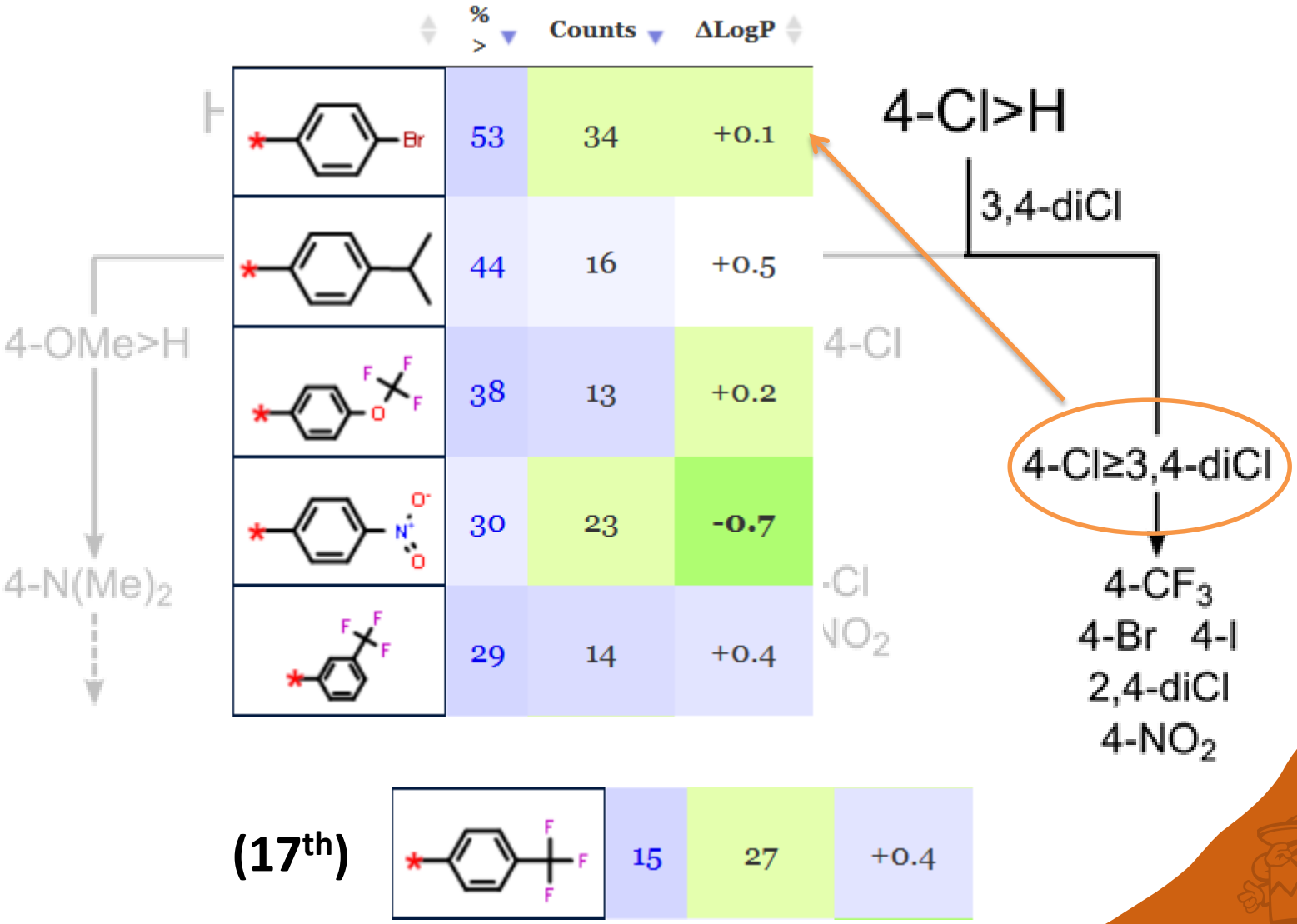

**(1st if lower cutoff)**

| | | | |
|---|---|---|---|
|  | 33 | 15 | +0.4 |

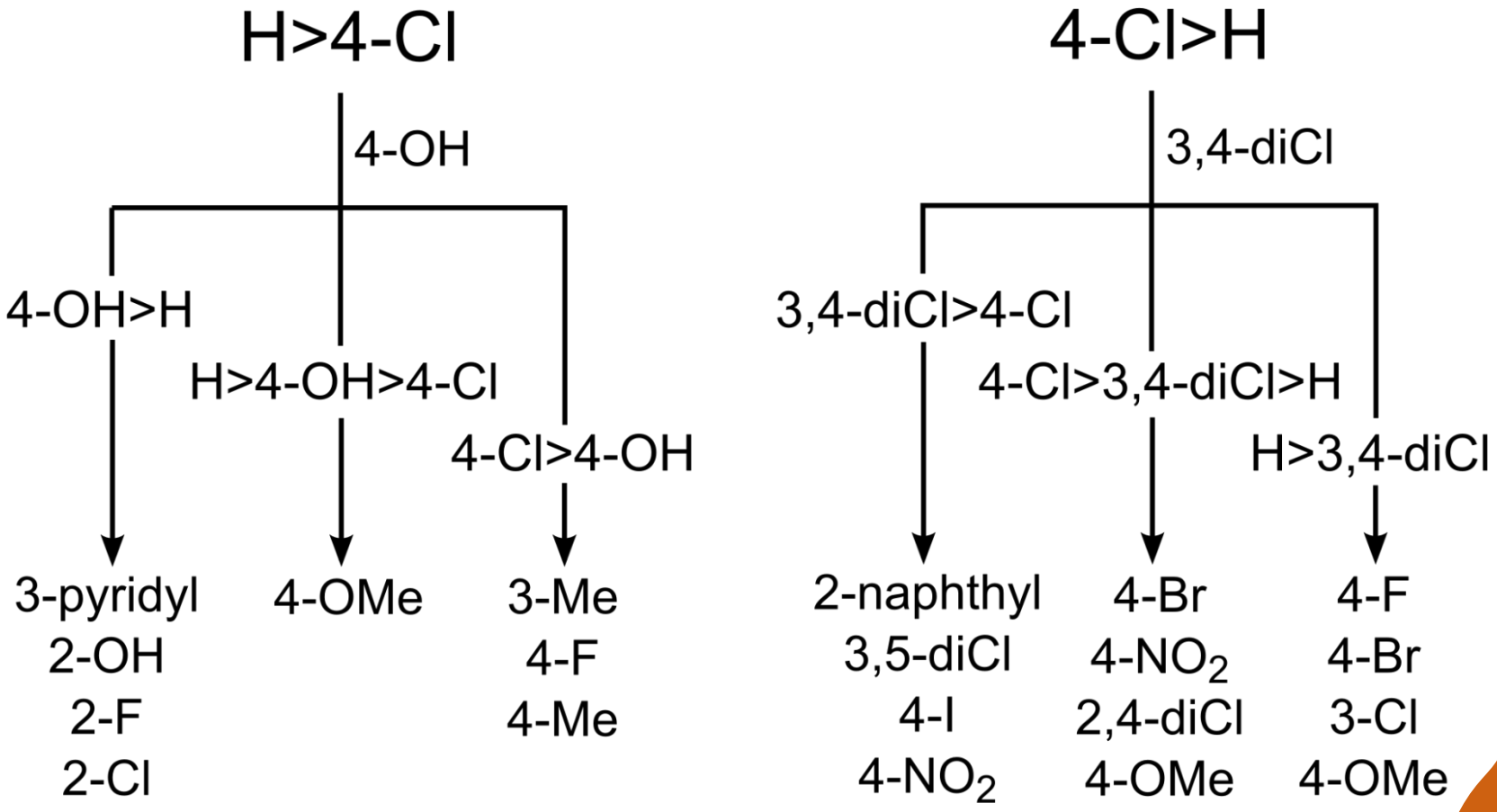
# TOPLISS DECISION TREE



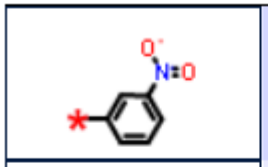
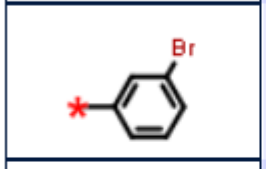
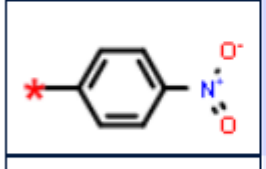
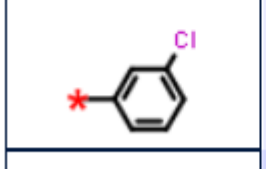
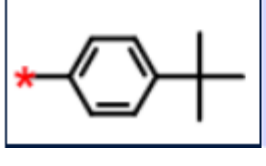
# TOPLISS DECISION TREE



# MATSY DECISION TREE (ONE OF MANY)

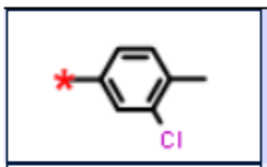
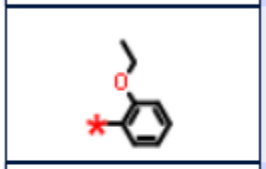
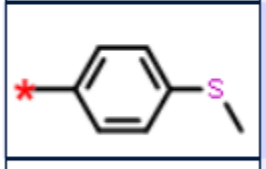
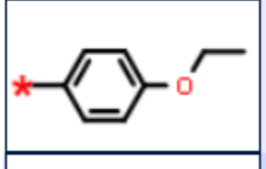
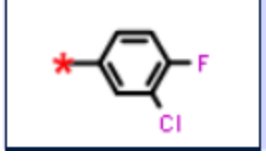


# MODIFYING THE PREDICTIONS FOR **4-Cl > H**

| | % > | Counts |
|---|---|---|
|  | 67 | 30 |
|  | 50 | 24 |
|  | 47 | 30 |
|  | 45 | 78 |
|  | 45 | 22 |

**Kinases**

Target-specific

| | % > | Counts | ΔLogP |
|---|---|---|---|
|  | 63 | 27 | +0.3 |
|  | 55 | 20 | -0.4 |
|  | 49 | 63 | 0.0 |
|  | 48 | 46 | -0.4 |
|  | 48 | 46 | +0.1 |

**ΔLiPE > 0**

Incorporate metrics

# 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

# 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 suggestion:
  - Take advantage of the wealth of *experience* and *trends* contained in 60K med chem papers
  - ‘evidence-based medicinal chemistry’

# Beyond Matched Pairs

## Applying Matsy to predict new optimisation strategies

noel@nextmovesoftware.com

**Acknowledgements**

Roger Sayle

Jonas Boström, AstraZeneca

**StarDrop Integration**

James Chisholm

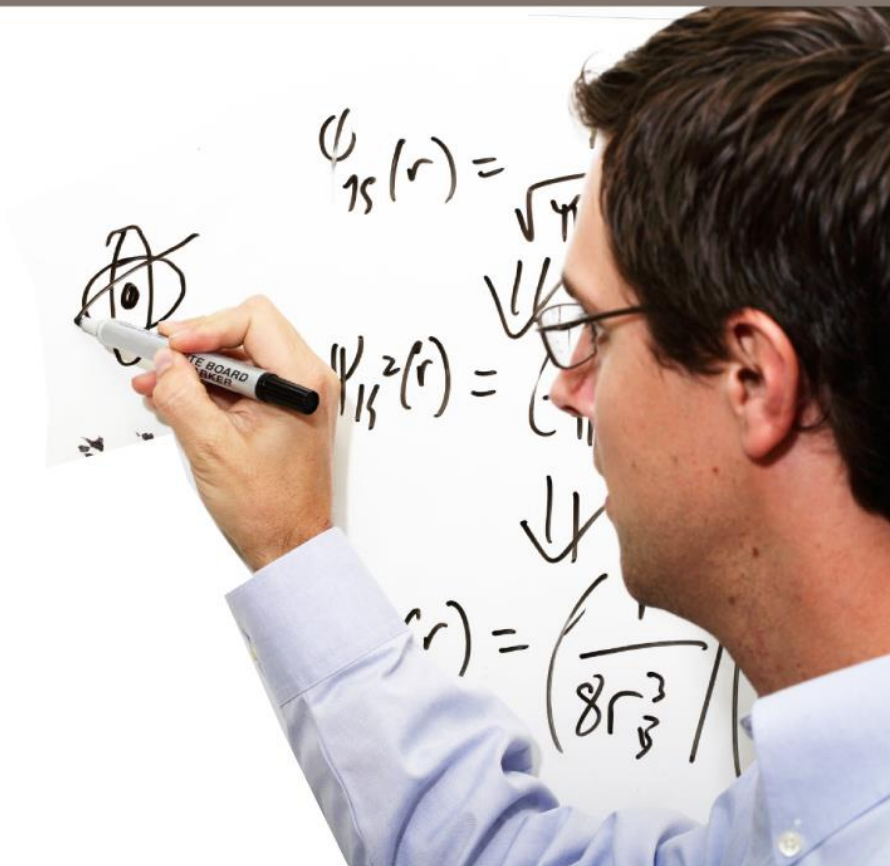
Sam Dowling

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

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

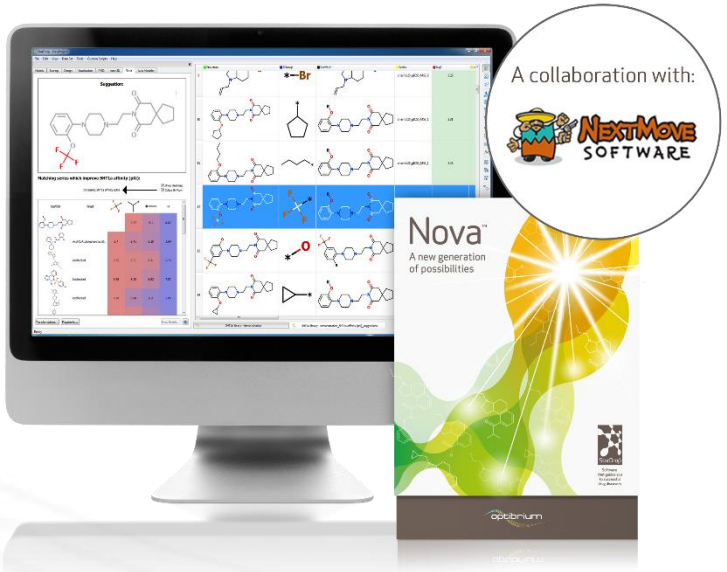
# Demonstration

optibrium



# StarDrop
## Matched Series Analysis

- Developed in collaboration with NextMove Software
- Identifies chemical substitutions that are most likely to improve target activity




- Goes 'beyond' matched pair analysis
  - Uses data from longer matched series to make more relevant predictions for your chemical series
- Two methods implemented within StarDrop's Nova Module
  - Matsy™
  - SAR transfer
- Based on ChEMBL activity database
  - Can be extended with matched series from your in-house database (provided as a service by NextMove)

# For more information…

- Noel O'Boyle
  - noel@nextmovesoftware.com
  - www.nextmovesoftware.com
- Nick Foster
  - nick@Optibrium.com
  - www.optibrium.com/stardrop



- A recording of the webinar will be made available as soon as possible on the Optibrium Community website at:
  - www.optibrium.com/community

Thank you for joining us today for our webinar!