

BioDig ADME:

Using matched molecular pairs to find structural changes to improve your compounds properties

Jameed Hussain

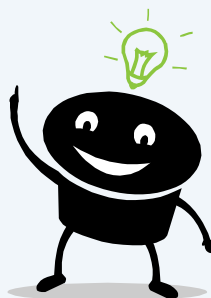
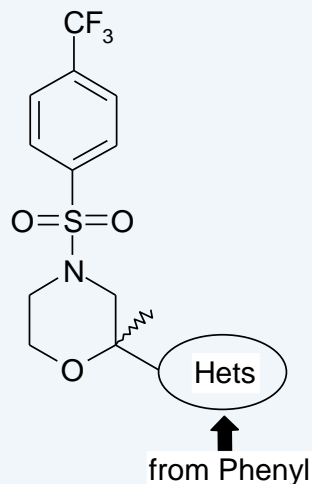
Outline

- Why matched molecular pair analysis is useful
- How the analysis is carried out
- How matched molecular pairs are identified
 - Advantages of the approach
- How BioDig has been used at GSK

BioDig – Why ?

- Common problem in medicinal chemistry..

Key Issue: High intrinsic clearance



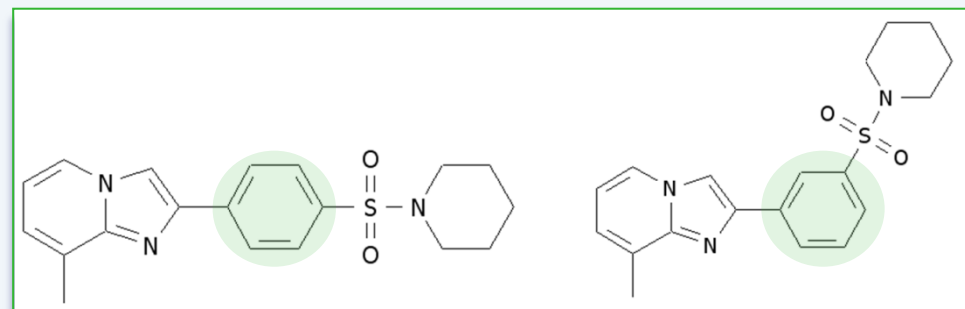
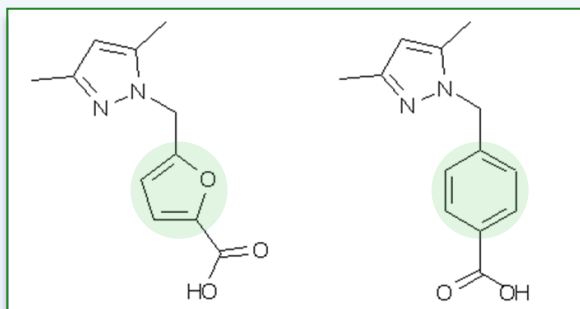
‘Why don’t you change it for an oxadiazole – it worked for us in Prog X’.

Can we systematically mine existing data to find structural changes that lead to better clearance ?

“Standing on the shoulders of giants” Isaac Newton

Mining our data

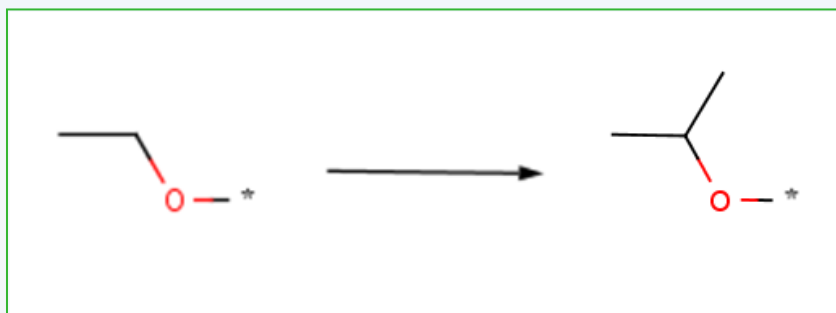
- How ?
- How do we analyse our SAR ?
 - One way is to look for pairs of compounds which only differ by a single change



The change in activity can then be attributed to the structural change

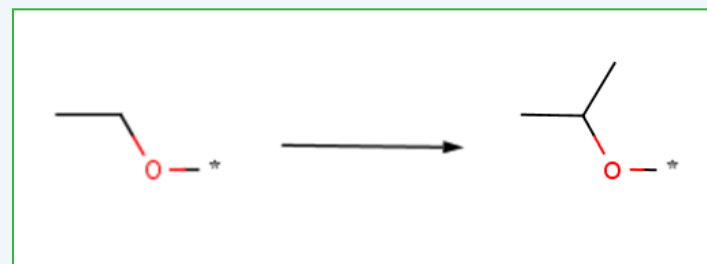
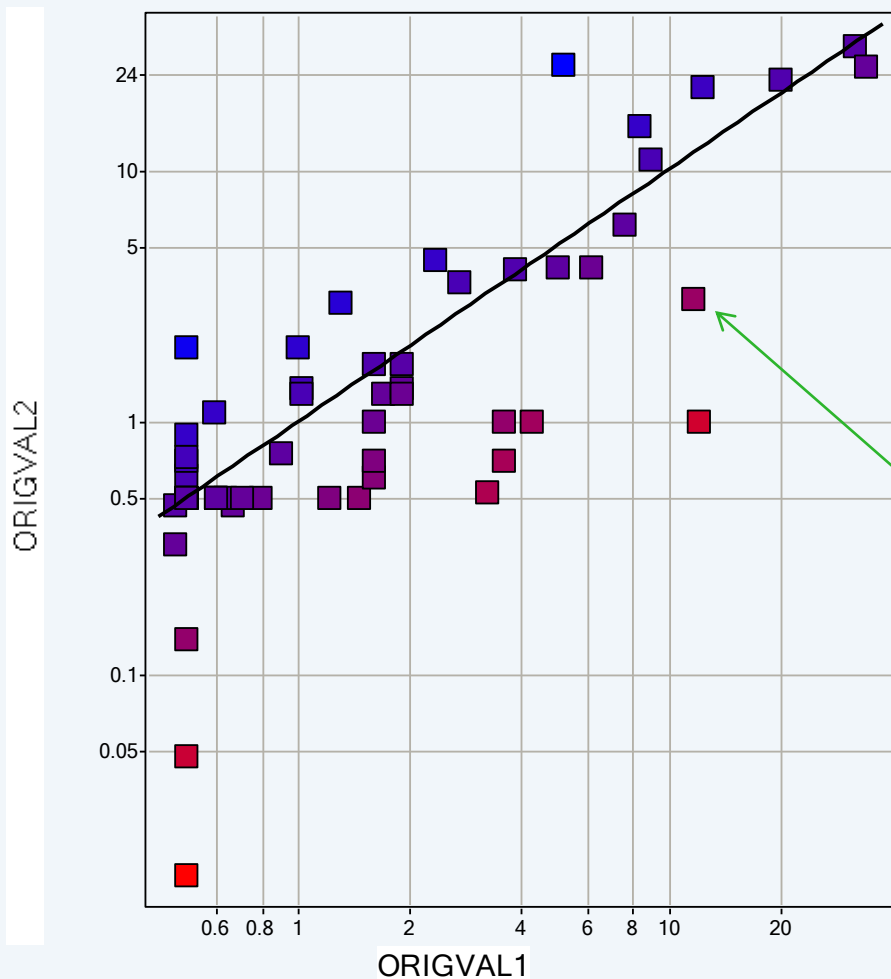
Real example..

- What happens to intrinsic clearance when you make the following structural change ?



- Pull out the clearance data for compound pairs with this structural change..

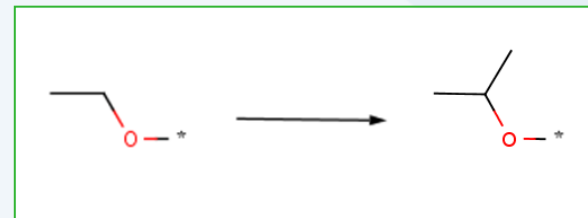
Intrinsic Clearance Data..



Cmpd 1	Property	Cmpd 2
11.7	clearance	3.1
372.9	MW	386.9
4.9	clog P	5.2

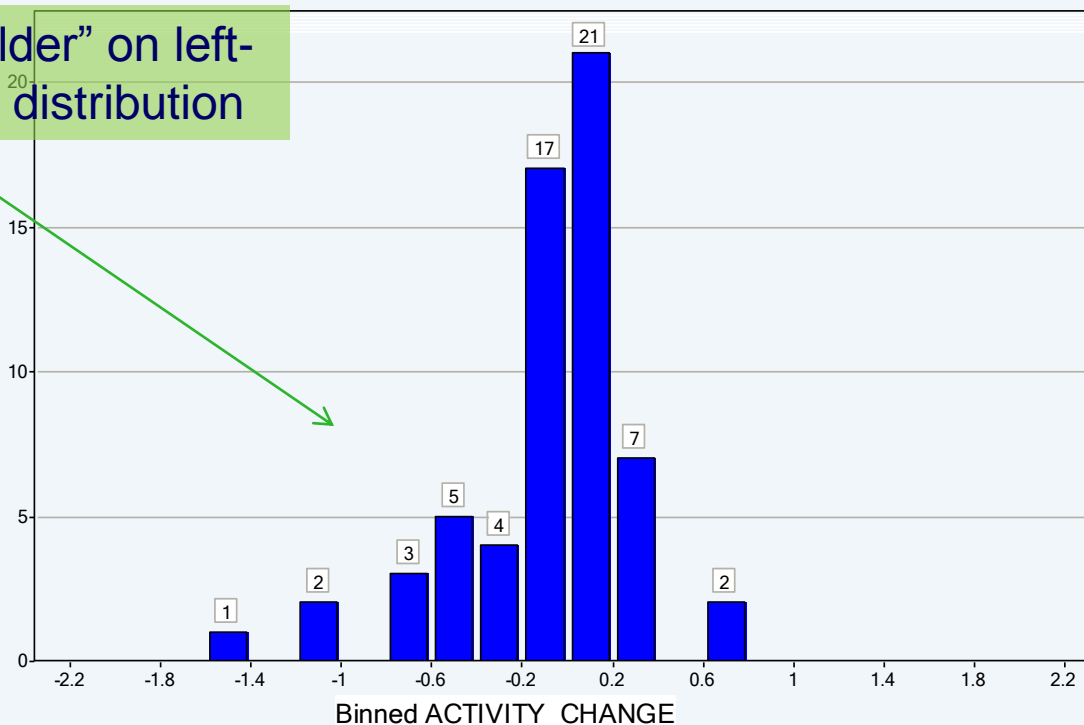
Intrinsic Clearance Data..

- Change in intrinsic clearance (log)
 - Appears to *reduce* intrinsic clearance



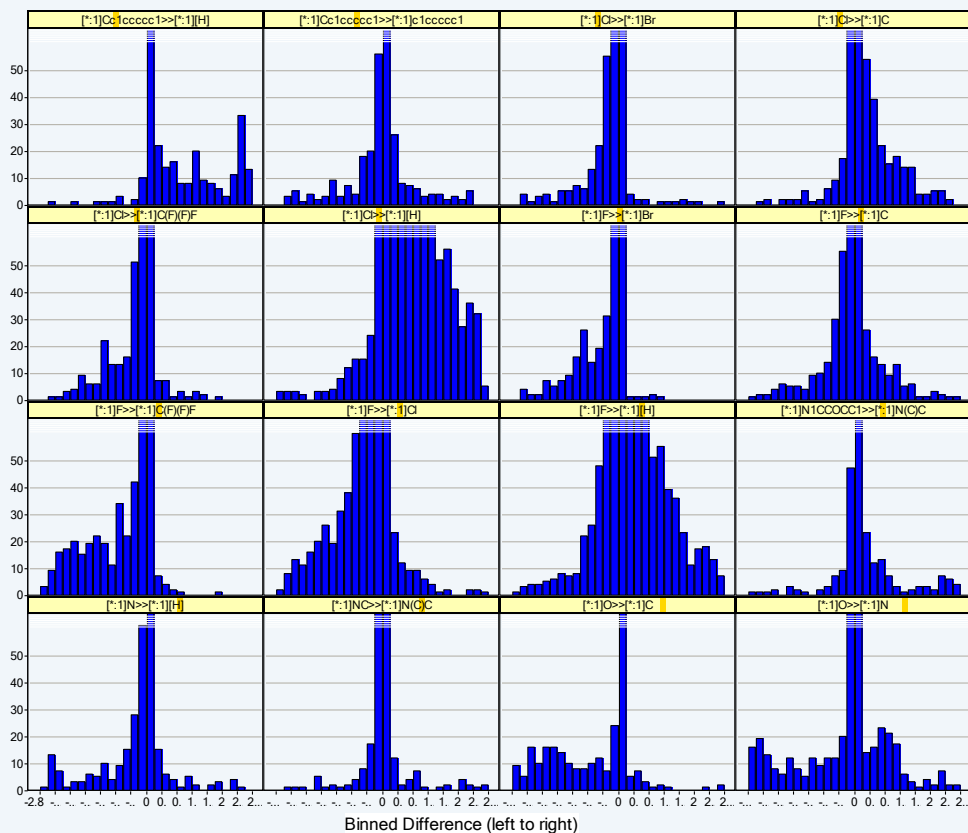
Notice the “shoulder” on left-hand side of the distribution

Number of compounds



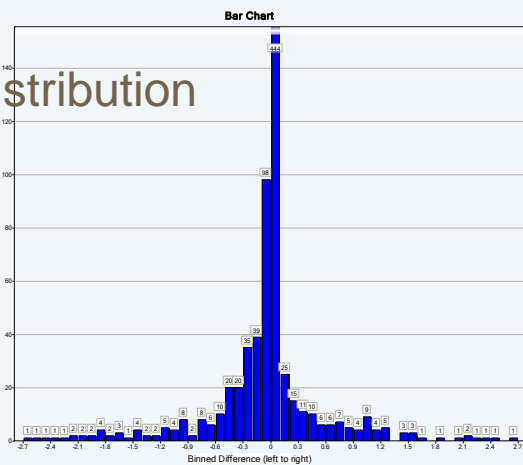
Analysis

- How do we quantify (the extent) of this effect?
- Lets look at some of the other distributions observed..

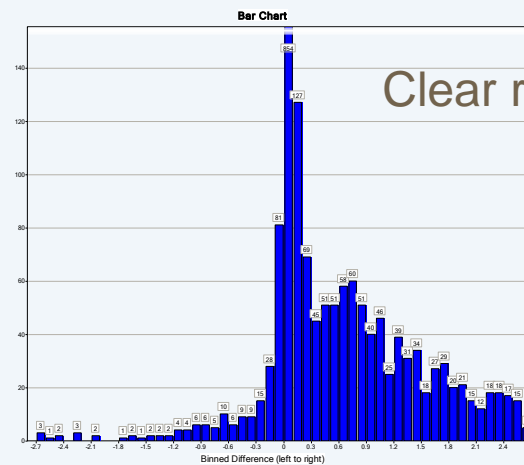


Four clear scenarios

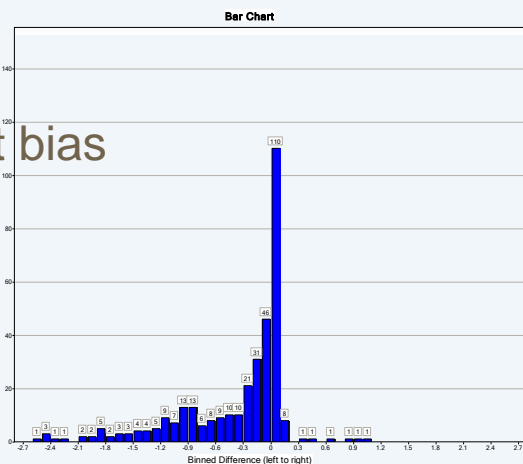
Normal distribution



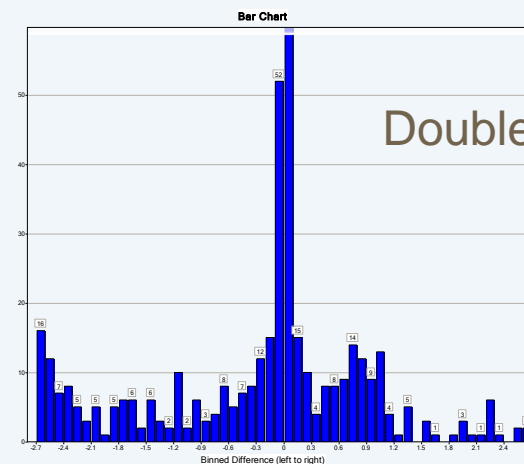
Clear right bias



Clear Left bias

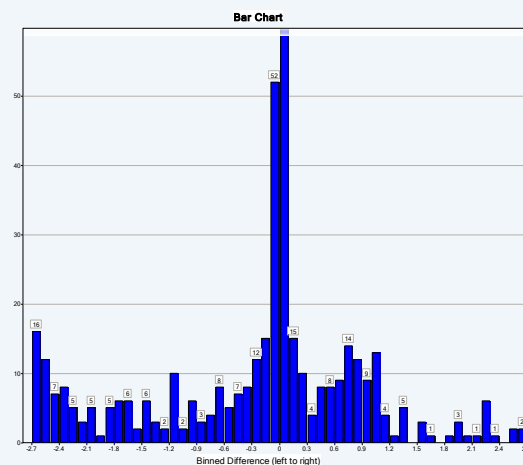


Double-sided bias



Literature approaches

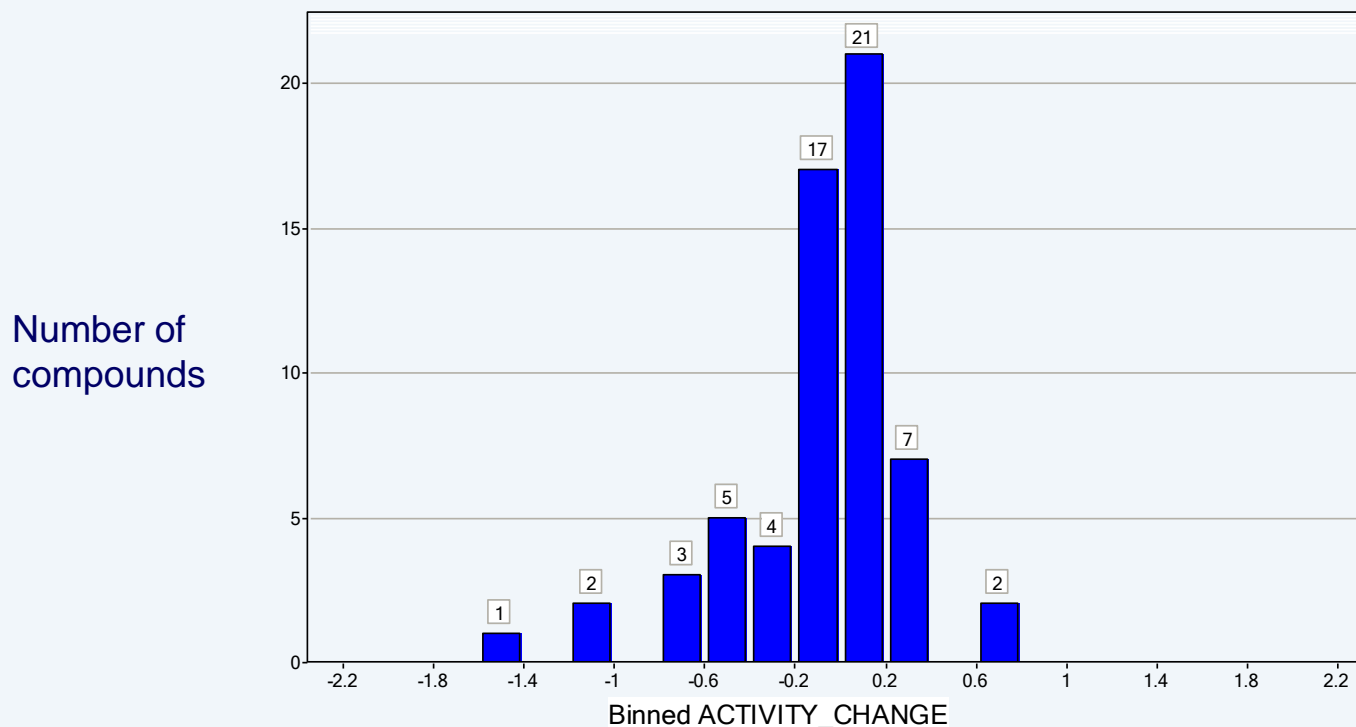
- Calculate Mean difference and SD
 - Leach *et al* (AZ) & Gleeson *et al* (GSK)
- Compare the distribution against a reference distribution – is this transform more likely to change the property than the reference transformation
 - Hajduk *et al* (Abbott)
- Numerical analysis
 - Lewis *et al* (Pfizer)



Numerical Analysis

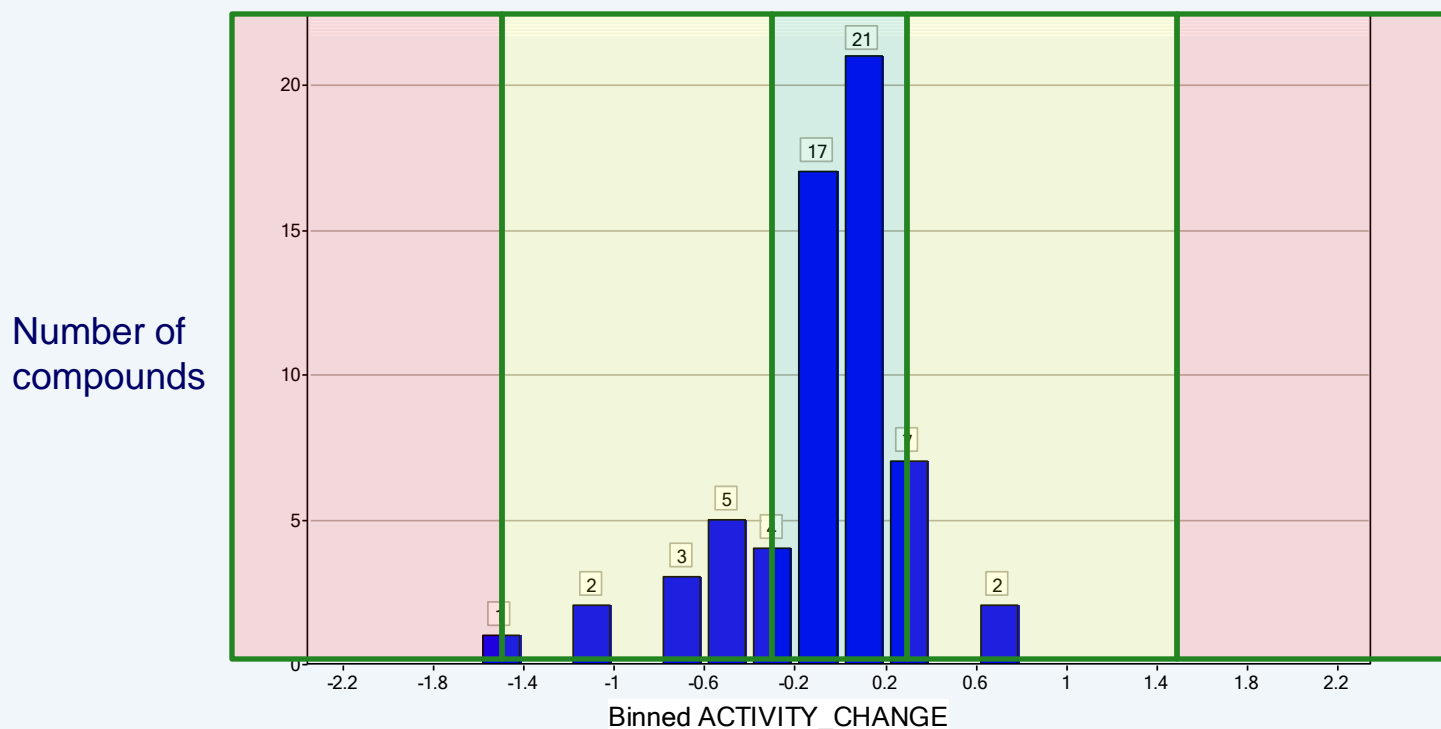
- Each compound pair has an associated activity change
 - This is binned into 5 categories:
 - Large negative change
 - Small negative change
 - No change
 - Small positive change
 - Large positive change
- % of compound pairs in each category is reported
 - This gives the probabilistic assessment

Numerical Analysis



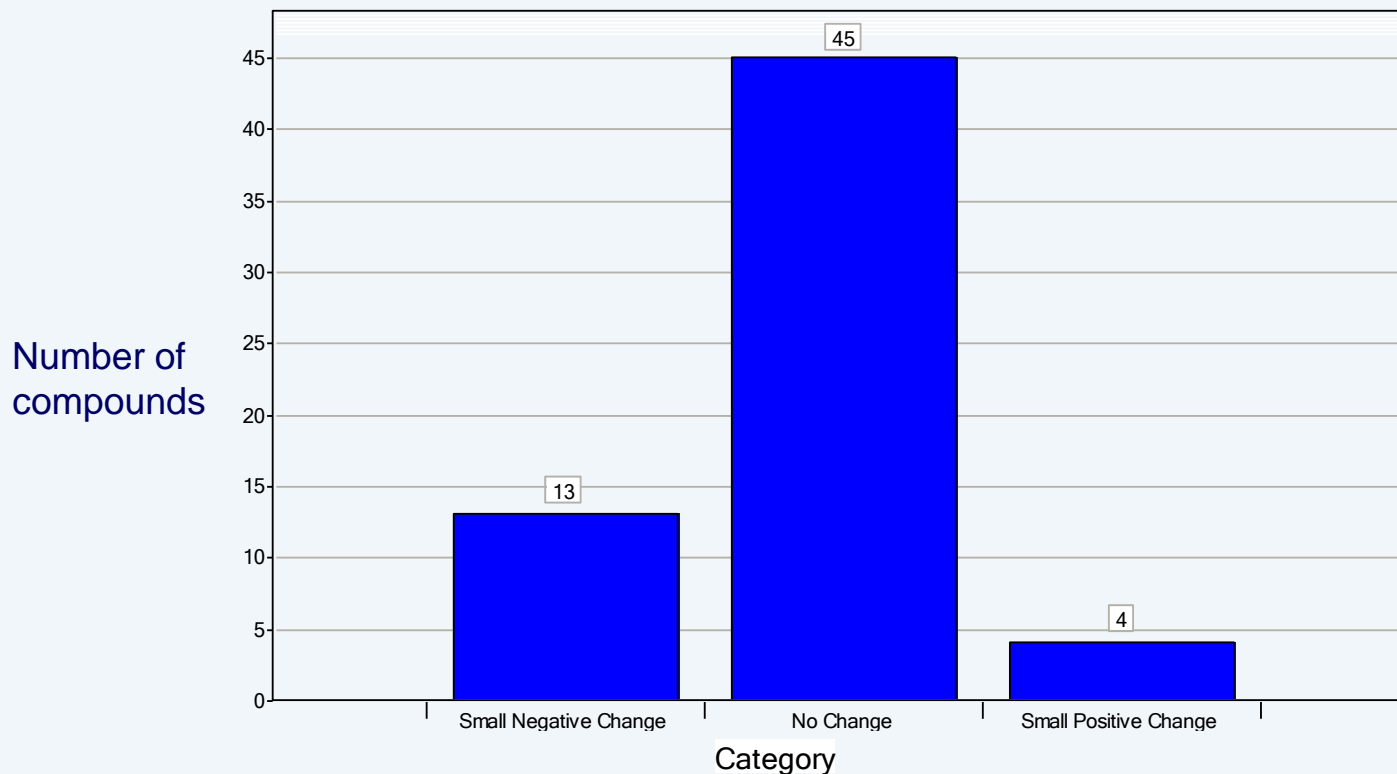
- Appropriate thresholds have been chosen for the different categories
 - Based on assay (experimental error) and pragmatic relevance
 - Activity change thresholds (Clearance) : -1.5, -0.3, 0.3, 1.5

Numerical Analysis



- Appropriate thresholds have been chosen for the different categories
 - Based on assay (experimental error) and pragmatic relevance
 - Activity change thresholds (Clearance) : -1.5, -0.3, 0.3, 1.5

Numerical Analysis



- Appropriate thresholds have been chosen for the different categories
 - Based on assay (experimental error) and pragmatic relevance
 - Activity change thresholds (Clearance) : -1.5, -0.3, 0.3, 1.5

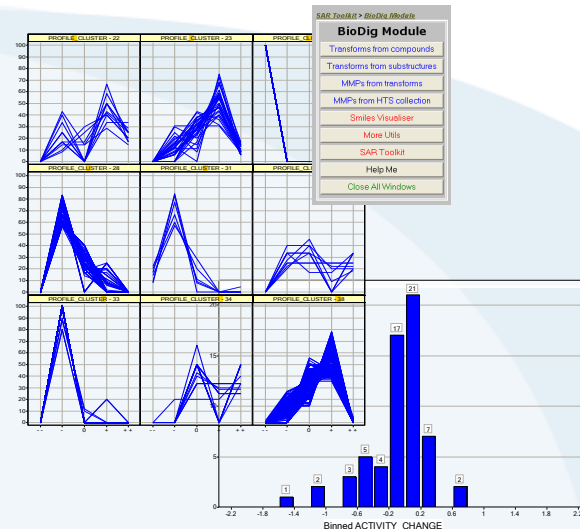
Numerical Analysis



- Appropriate thresholds have been chosen for the different categories
 - Based on assay (experimental error) and pragmatic relevance
 - Activity change thresholds (Clearance) : -1.5, -0.3, 0.3, 1.5

Further Analysis

- Next, some reality checks..
 - How general is effect likely to be ?
 - How many compound pairs in the set ?
 - How many chemotypes in the set ?
 - How many of the pairs contain data with modifiers ?
- Tools in place to see distributions and compound pairs
 - Use Spotfire (data visualiser) to really interrogate the data



MMP identification

BioDig Principle: Get the biological data to tell you what is important

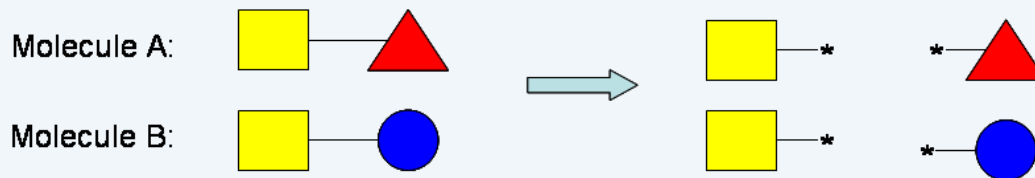
- Ideal MMP identification algorithm
 - Fast enough to cope with large datasets
 - Find all MMPs in the dataset
 - Do not require the structural change to be defined upfront
- MMP identification methods available (in literature) but not suitable
 - Some restricted to side chains only
 - Require the structural changes to be defined
 - Slow as require a N^2 search
 - Calculation of MCS between every pair of cmpds

MMP Identification






- Algorithm written in-house
 - Hussain & Rea, *J. Chem. Inf. Model.* **2010**, 50, 339-348
 - Wagener & Lommerse, *J. Chem. Inf. Model.* **2006**, 46, 677-685
- Already been adopted by others
- Capable of finding (almost) all MMPs in a compound set
 - No pre-definition of changes required
 - Fast
- Output of the program is a set of MMPs with a SMIRKS transform of the change

How it works

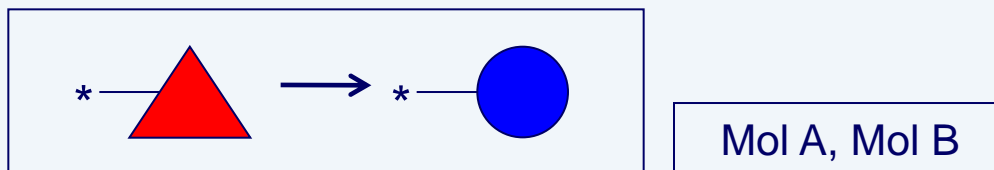
1) Take input molecule and enumerate all acyclic single cuts



2) Index

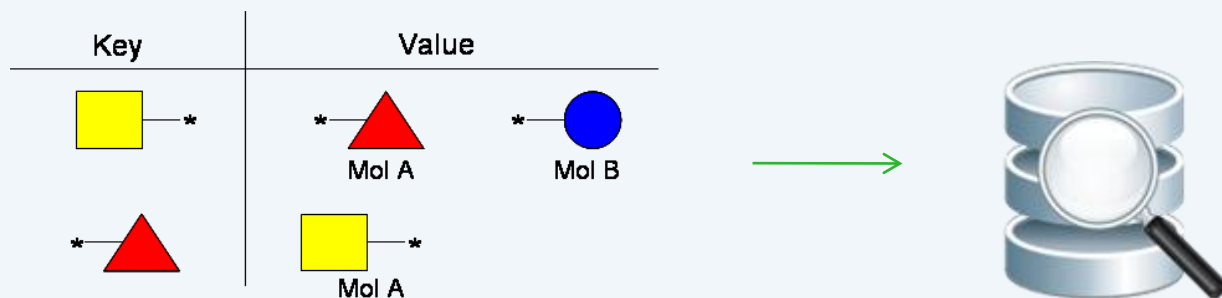
Key	Value
	*  Mol A *  Mol B
* 	 Mol A
...	...

3) Structural change and pair is then:



MMP Identification

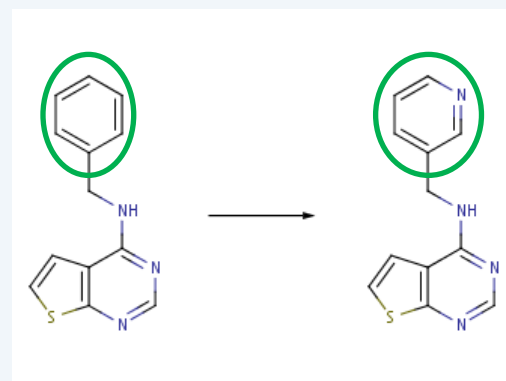
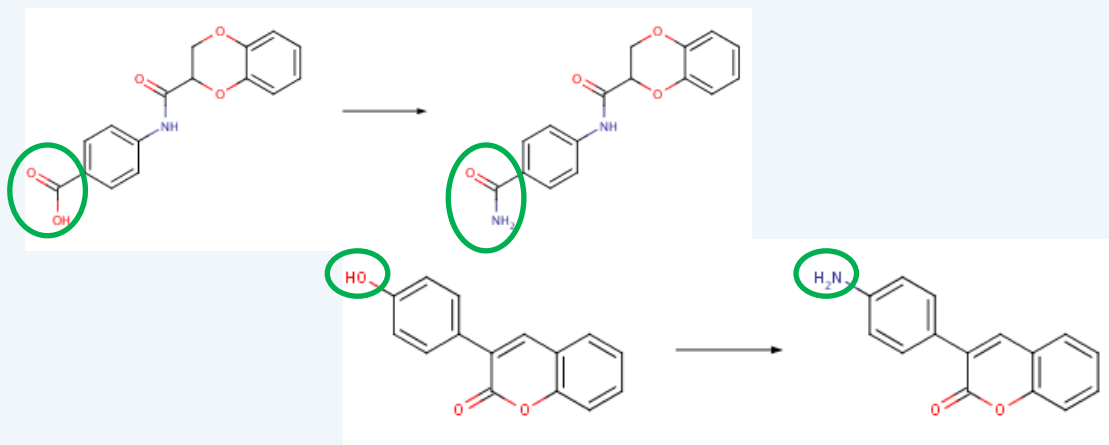
- Applicable to large datasets (thousands of compounds)



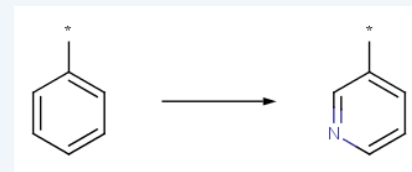
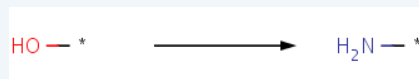
- Building the index as a SQL database makes the algorithm applicable to very large datasets (millions of compounds)
- Built for the GSK compound collection
 - Now easy to find MMPs for compounds within the GSK collection

Interpretability

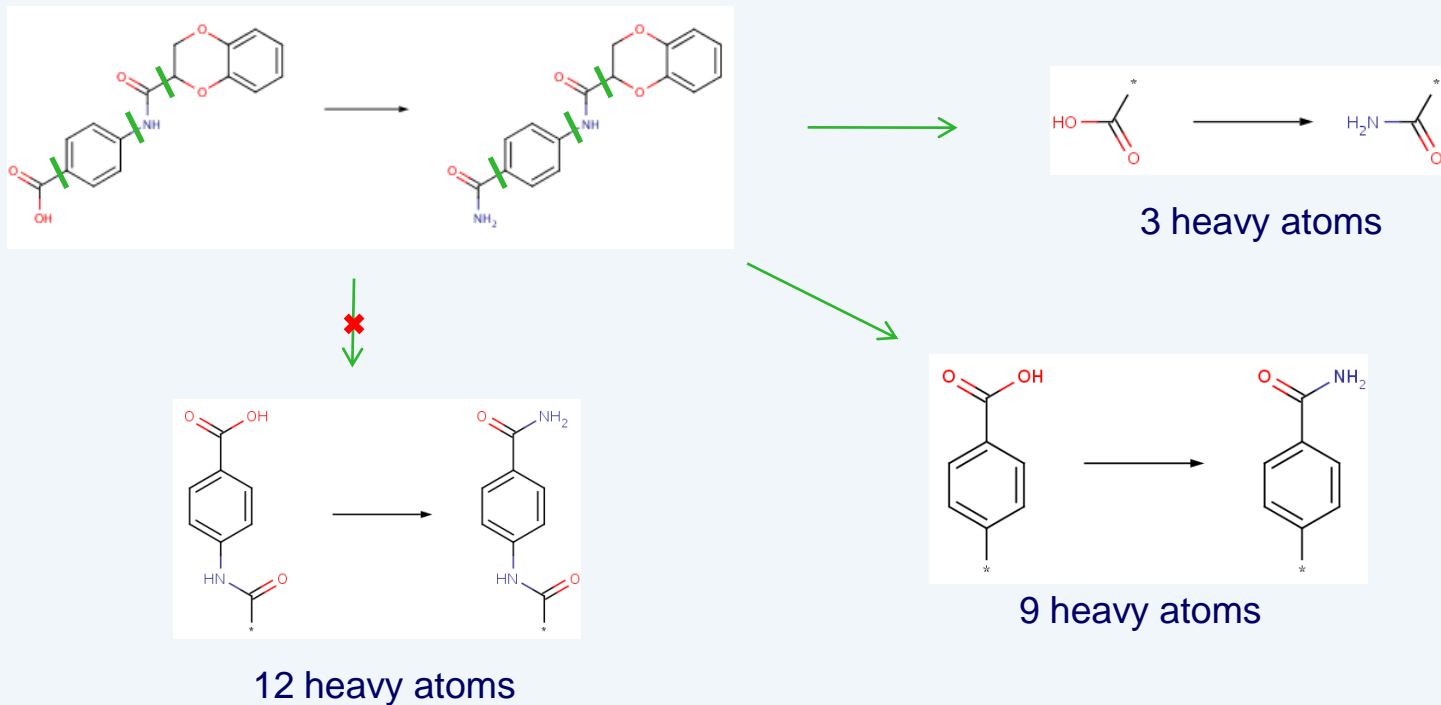
- The transforms formed by the algorithm are readily interpretable
 - Functional groups and rings are never fragmented



- Closer to the way medicinal chemists view the changes

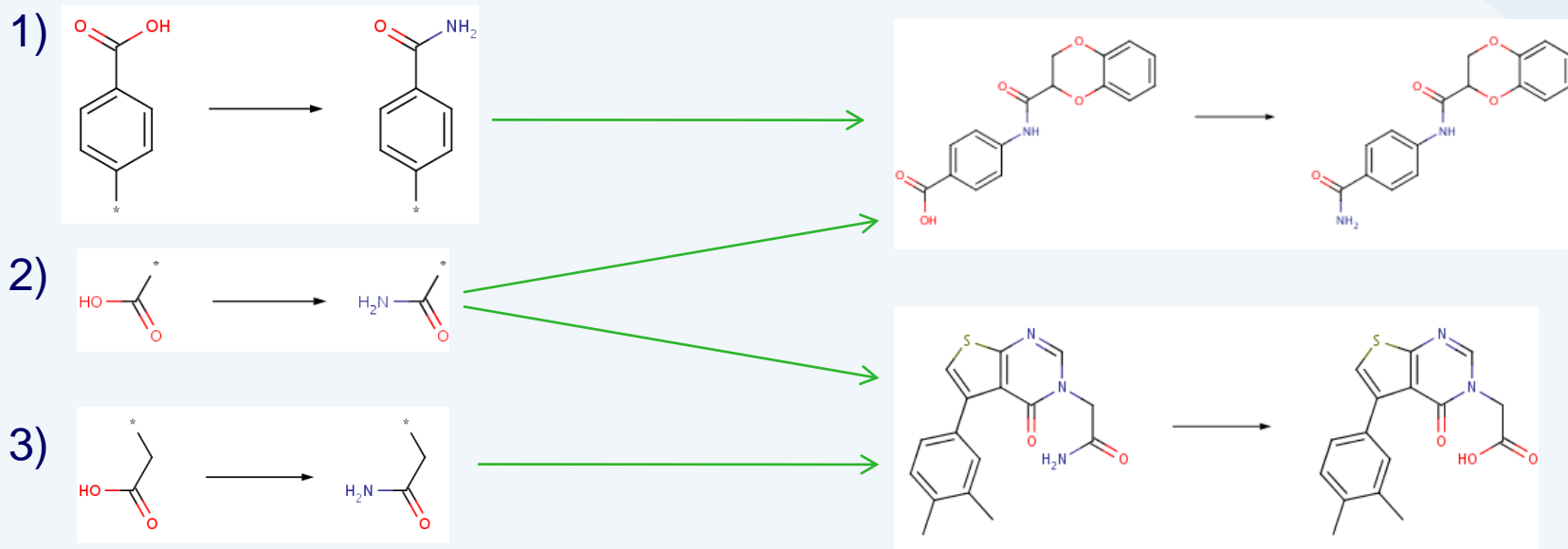


Transform Redundancy



- All cuts that result in substructures / transforms smaller than a heavy atom count specified are stored
 - 10 heavy atoms in cases above
- The activity change is stored for each transformation found

Transform Redundancy gives context



- If the aromatic ring is important for a property, that should be apparent in activity change distribution for 1) & 2)

BioDig Principle: Get the biological data to tell you what is important

BioDig ADME database



- A database has been created where we have collected together the compound pairs (and associated structural changes) in our ADME data

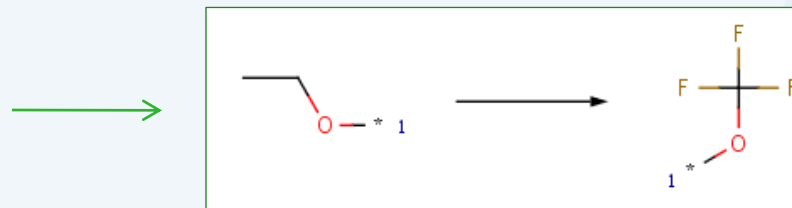
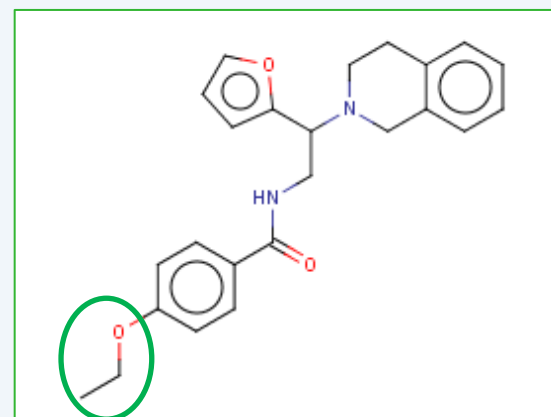
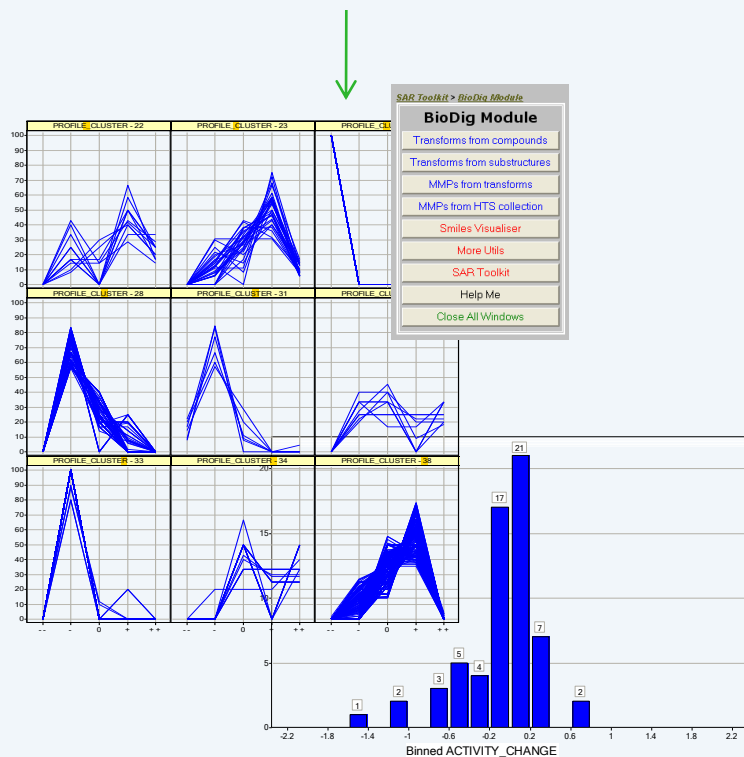
Property	Number of compounds	Number of matched molecular pairs	Number of transforms
Clearance	47k	8M	6M
Solubility	170k	96M	82M
Log D	158k	89M	68M
hERG	175k	98M	76M
PP Binding	188k	70M	55M
P450 3A4	240k	161M	136M
P450 2D6	231k	150M	128M

Represents a huge amount of knowledge to tap into

BioDig ADME

The information in the database can be used to answer other questions

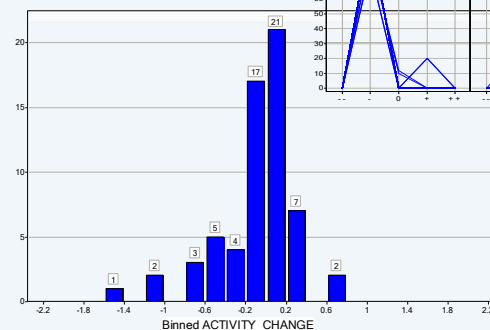
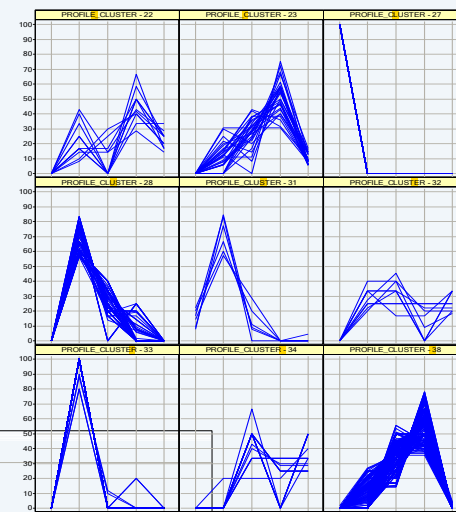
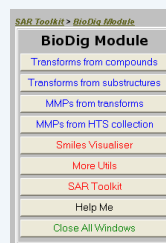
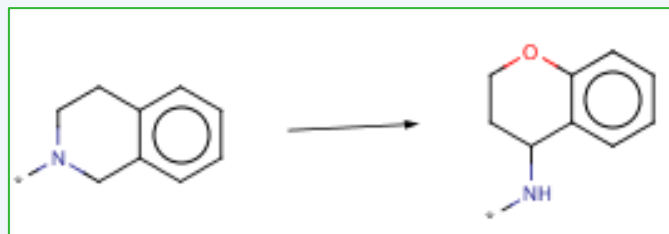
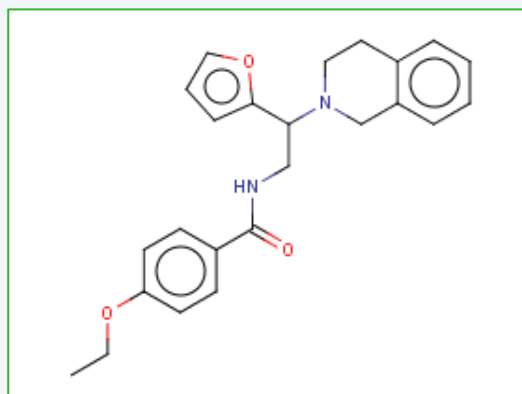
Substructure X: “What possible replacements have been tried and which improve clearance ?”



BioDig ADME

The information in the database can be used to answer other questions

“What possible changes could be made to the compound to improve its clearance ?”



How has it been used ?

- Version with clearance, solubility and log D has been available for ~1.5 years
 - Rolled out to a small number of expert users
 - Computational chemists and several medicinal chemists
- Roll out of a tool with more ADME properties is imminent
- Main use is to check an hypothesis about a particular structural change
 - Does the data back up the hypothesis ?
- Used to suggest changes to improve a particular property
 - Suggestions from BioDig have led to the synthesis of compounds with improved clearance and solubility

Summary

- Why matched molecular pair analysis is useful
- How the analysis of the activity change distributions is carried out
- How matched molecular pairs are identified
 - Advantages of the approach
- How BioDig has been used at GSK

Acknowledgements

- Ceara Rea
- Chris Luscombe
- Jordi Munoz-Muriedas

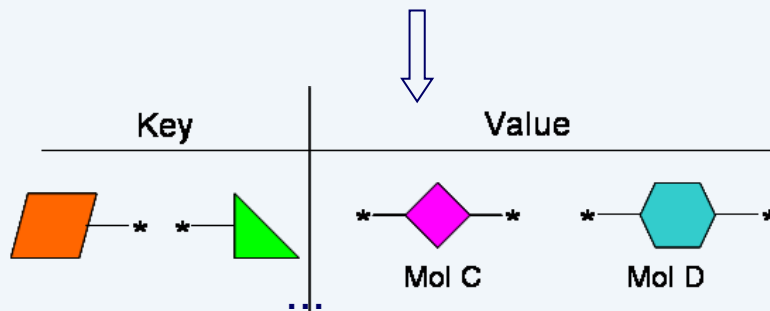
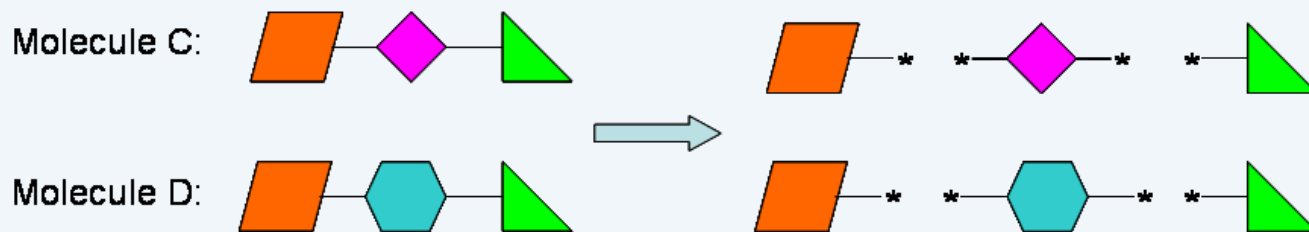
- Stefan Senger
- Gianpaolo Bravi
- Colin Edge

- The users who provided feedback for the tool

Backup Slides

Finding core changes:

- For the doublet cuts



Mol C, Mol D