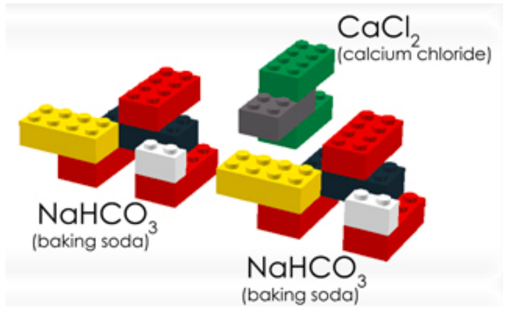
The Institute of Cancer Research

# Lessons learnt using RDKit for multiobjective optimisation

EBI Hinxton 03/10/2013

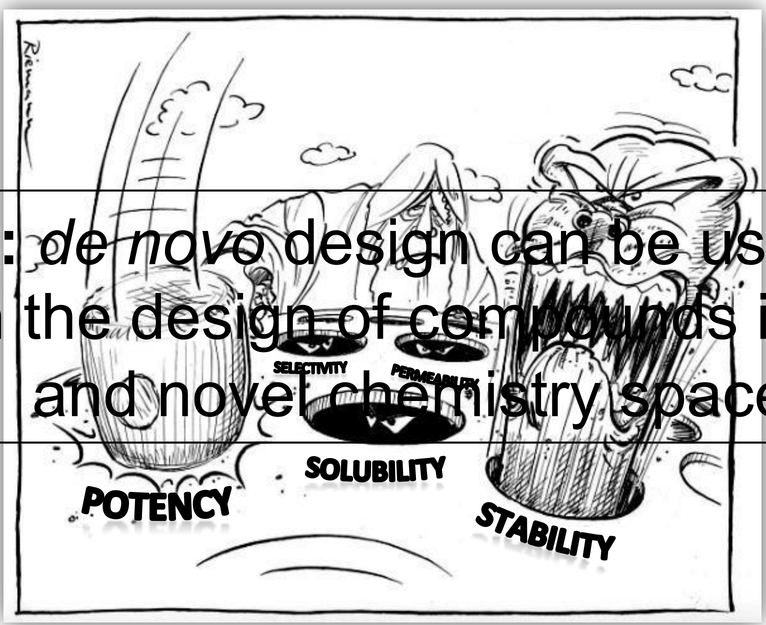
## Outline

- Introduction
    - Multiobjective optimisation
    - My previous work
- Methods
    - Workflow
    - Individual components
- Case study
- Ongoing work/Conclusion



## Multiobjective Optimisation

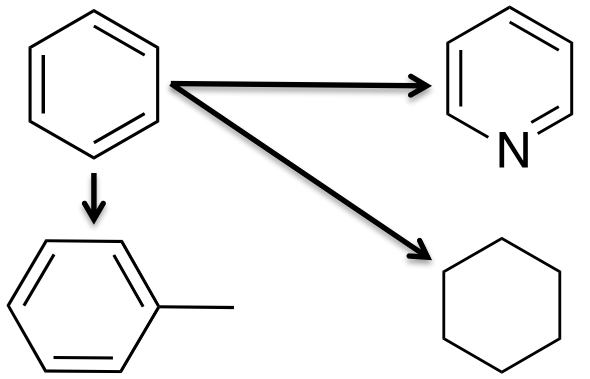
- One of the main pitfalls in drug discovery is that preclinical development candidates often maintain features of the hits from which they are derived
- Optimising in multiple objectives simultaneously often leads into synthetically challenging chemical space that is more time consuming to explore

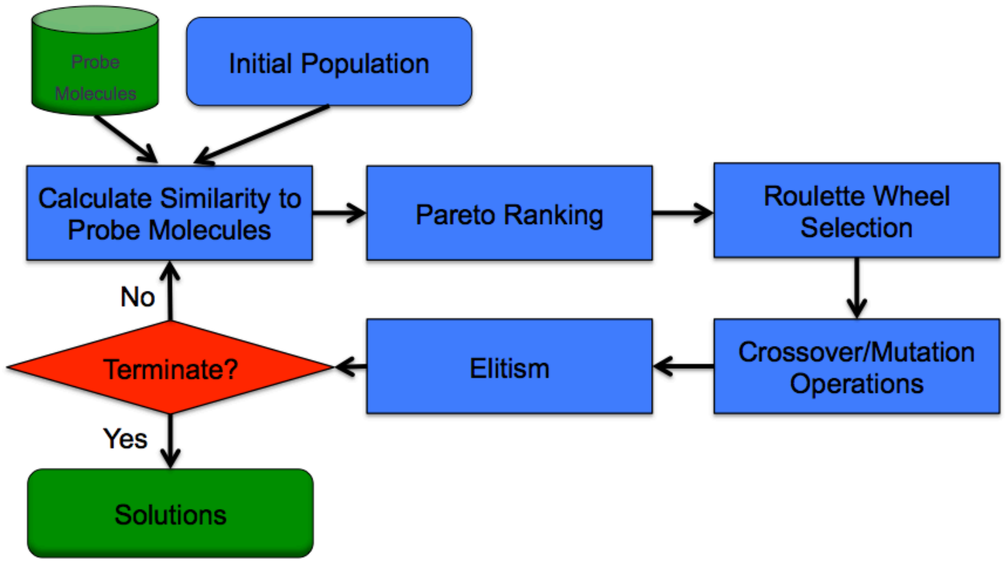


**Hypothesis:** *de novo* design can be used to increase confidence in the design of compounds in more diverse and novel chemistry space

## Previous Work

- I previously used atom-based *de novo* design



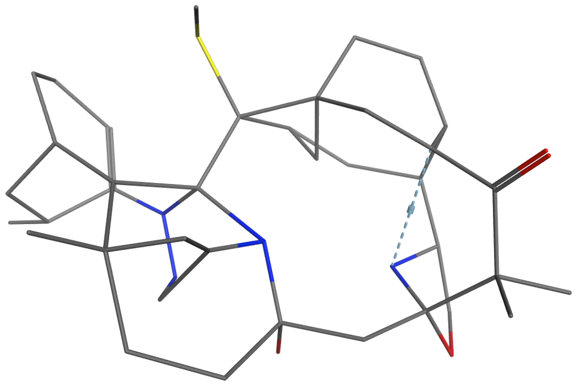


- Started using RDKit C++ API
  - Intermediate molecules cause errors

- Moved away from RDKit and represented molecules using a simple C++ class
  - SMILES parser (RDKit's)
  - Valency model
  - SDF writer

## Previous Work

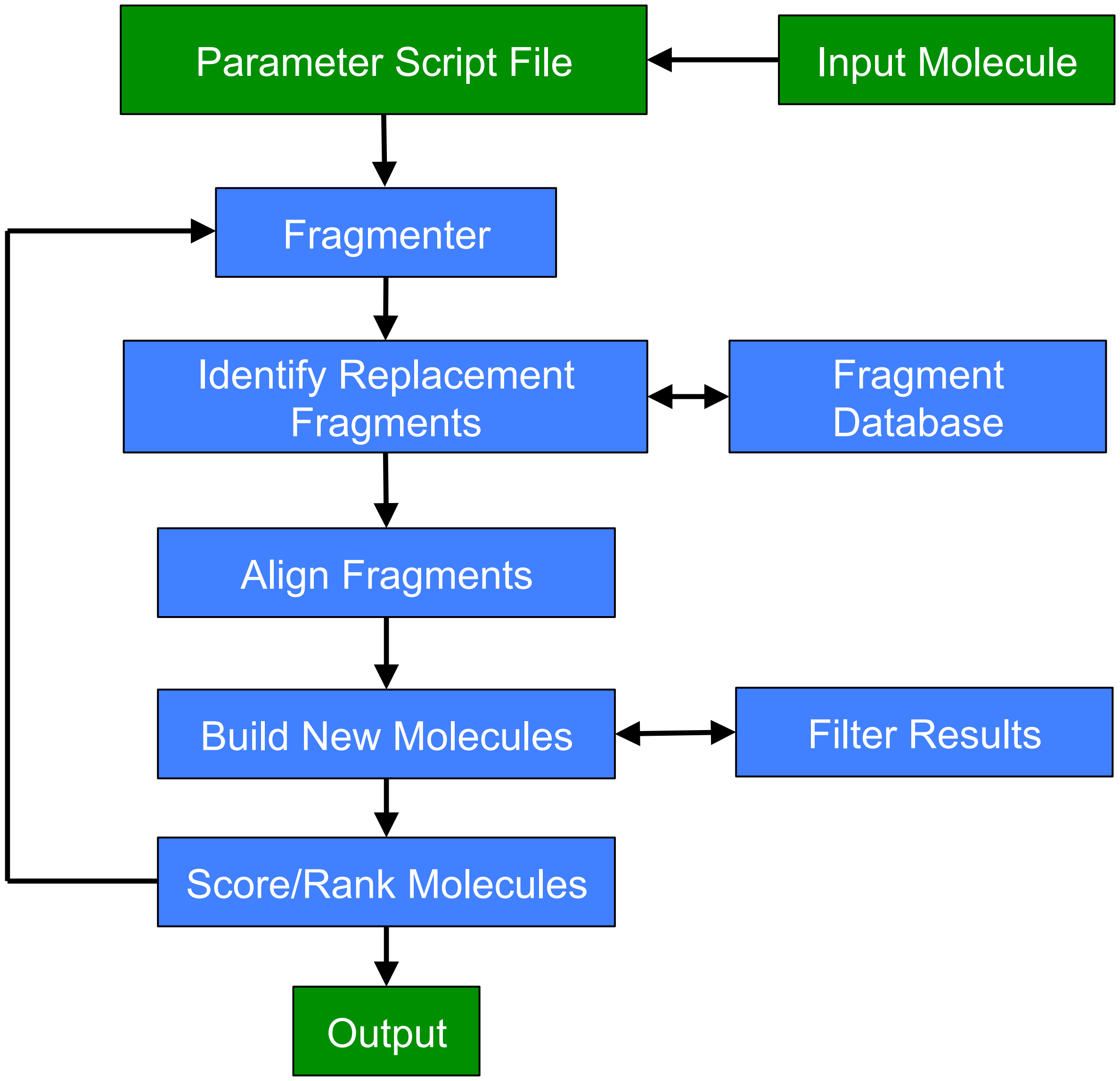
- Unfortunately the problem is not as simple as this
    - Ring perception
    - Synthetically inaccessible molecules



We can employ learning from IADE to develop a workflow for multiobjective optimisation

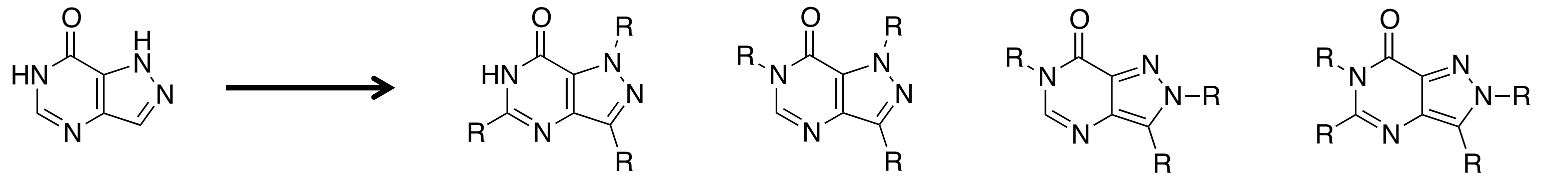
Ertl, Peter, and Richard Lewis. "IADE: a system for intelligent automatic design of bioisosteric analogs." *J. Comp. Aid. Mol. Des.* 26.11 (2012): 1207-1215.

## Methods Workflow

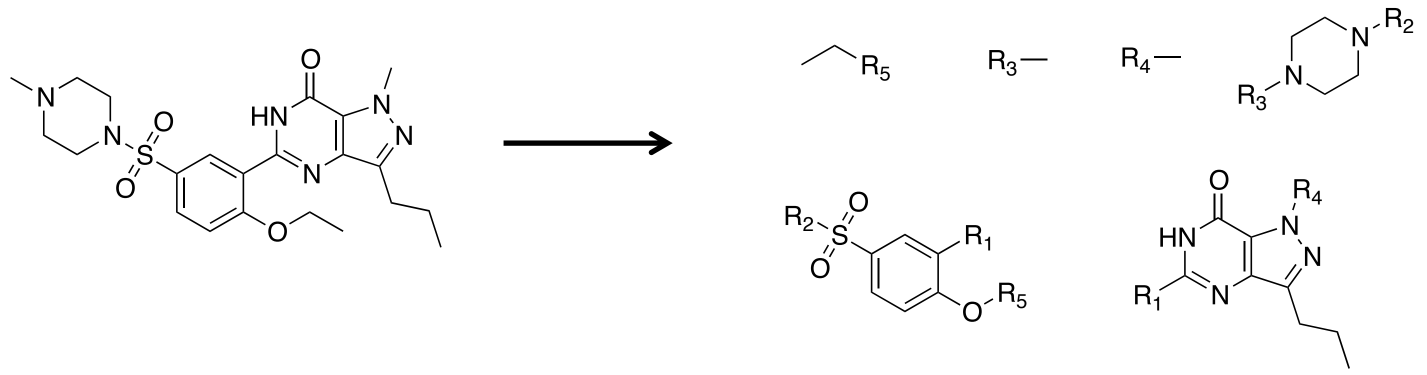


# Fragmentation

- We wanted a fragmentation scheme that produced fragments suitable for the *de novo* design of synthetically accessible molecules



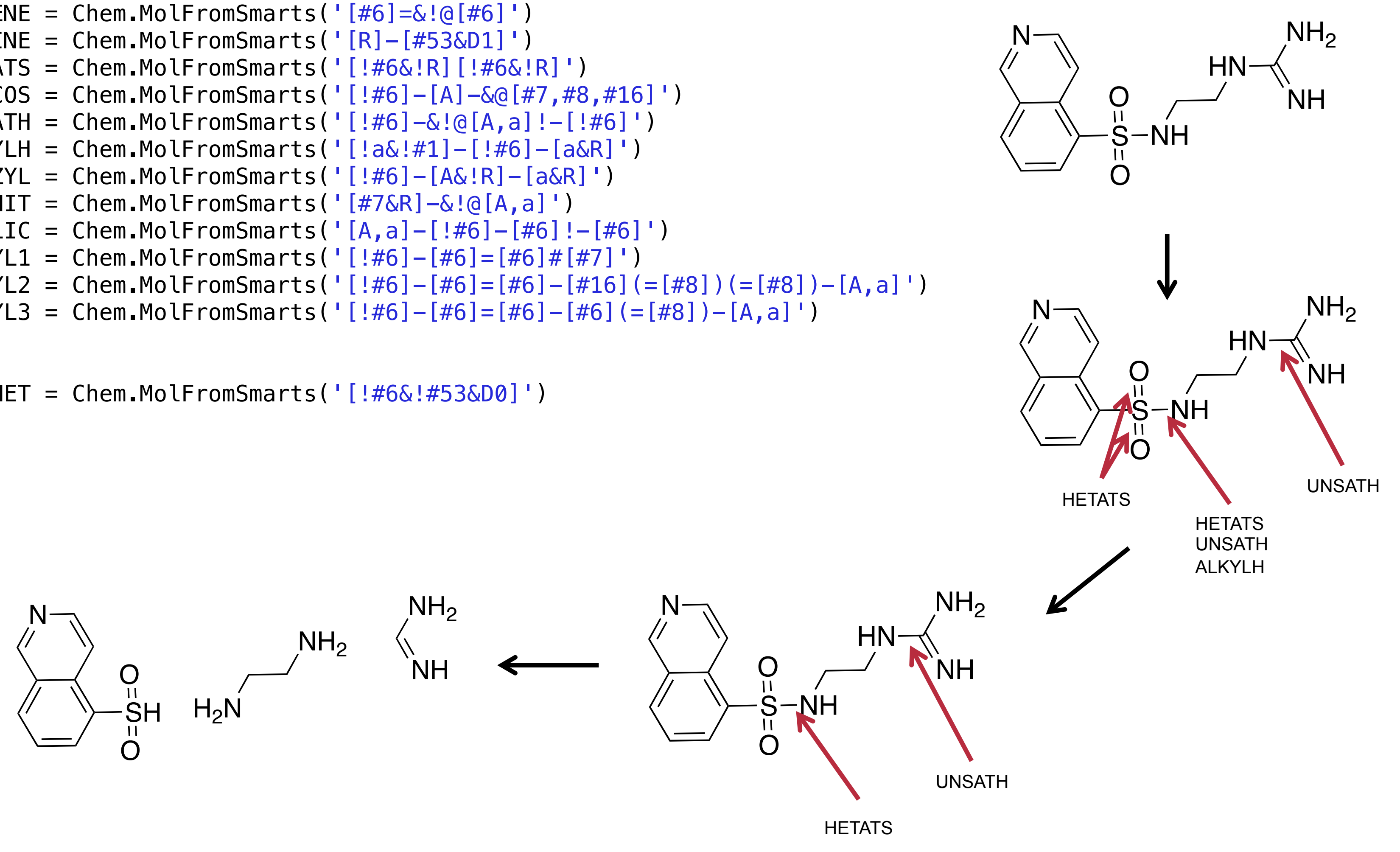
- Also wanted to break down input molecules to make them suitable for this cut and paste style *de novo* design



# Fragmentation

```
BIARYL = Chem.MolFromSmarts('[a]-&!@[a]')
ALKENE = Chem.MolFromSmarts('[#6]=&!@[#6]')
IODINE = Chem.MolFromSmarts('[R]-[#53&D1]')
HETATS = Chem.MolFromSmarts('[!#6&!R][!#6&!R]')
GLYCOS = Chem.MolFromSmarts('[!#6]-[A]-&@[#7,#8,#16]')
UNSATH = Chem.MolFromSmarts('[!#6]-&!@[A,a]!-[!#6]')
ALKYLH = Chem.MolFromSmarts('[!a&!#1]-[!#6]-[a&R]')
BENZYL = Chem.MolFromSmarts('[!#6]-[A&!R]-[a&R]')
EXONIT = Chem.MolFromSmarts('[#7&R]-&!@[A,a]')
ENOLIC = Chem.MolFromSmarts('[A,a]-[!#6]-[#6]!-[#6]')
VINYL1 = Chem.MolFromSmarts('[!#6]-[#6]=[#6]#[#7]')
VINYL2 = Chem.MolFromSmarts('[!#6]-[#6]=[#6]-[#16](=[#8])(=[#8])-[A,a]')
VINYL3 = Chem.MolFromSmarts('[!#6]-[#6]=[#6]-[#6](=[#8])-[A,a]')


ISOHET = Chem.MolFromSmarts('[!#6&!#53&D0]')
```



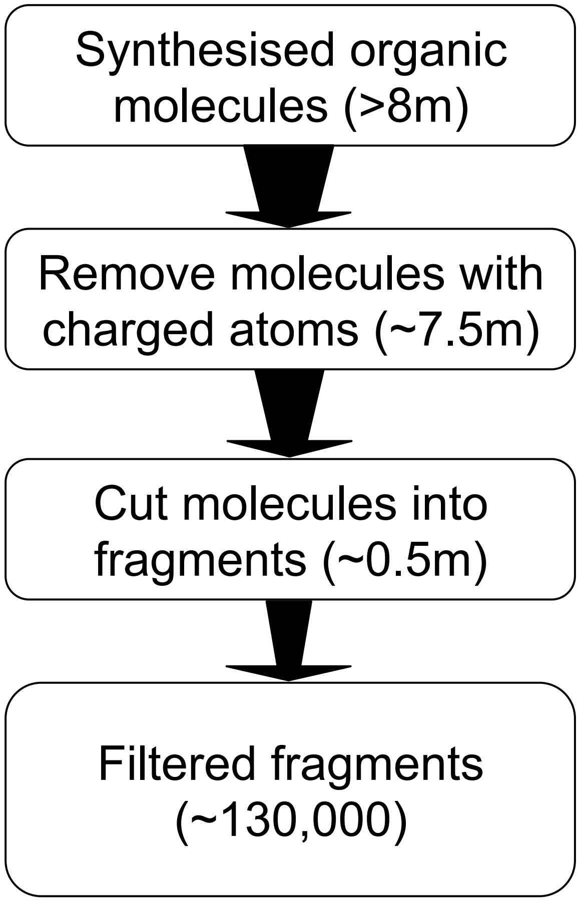
## Fragmenter and Database

- Similar to bioisosteric replacement it is important to search a relevant collection of compounds for *de novo* design

- To increase the number of fragments generated, we chose to use as many starting molecules as possible

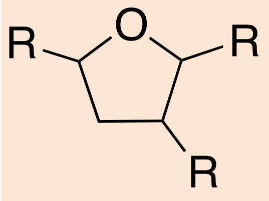
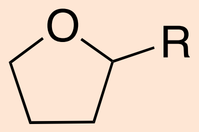
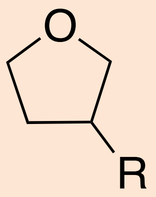
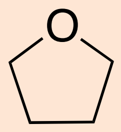
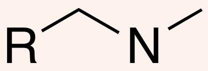
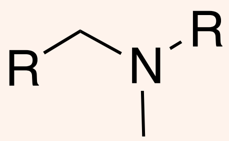
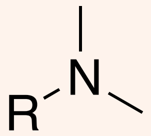
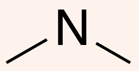
- Charged atoms were removed due to the early valence model. This has now changed

- All cut point locations are retained and these fragments stored in a PostgreSQL database


Synthesised organic molecules (>8m)
Remove molecules with charged atoms (~7.5m)
Cut molecules into fragments (~0.5m)
Filtered fragments (~130,000)

# Database View

Parent Fragment

| RG1 | R1_Fre | RG2 | R2_Fre | RG3 | R3_Fre | Par_Smi | P_Fre |
|---|---|---|---|---|---|---|---|
|  | 5 |  | 175 |  | 386 |  | 581 |
|  | 31 |  | 2,465 |  | 72,630 |  | 75,216 |

Repeated for all R-group patterns

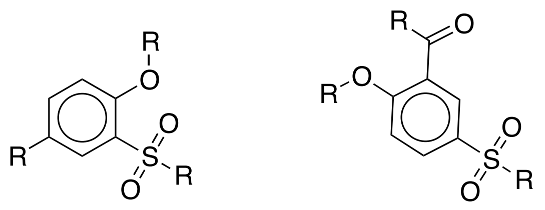
SMILES with cut point location and corresponding frequency

Overall frequency of fragment

- Cut locations and frequency can be used in choosing replacements
- Also contains some physicochemical properties and the 3D structure

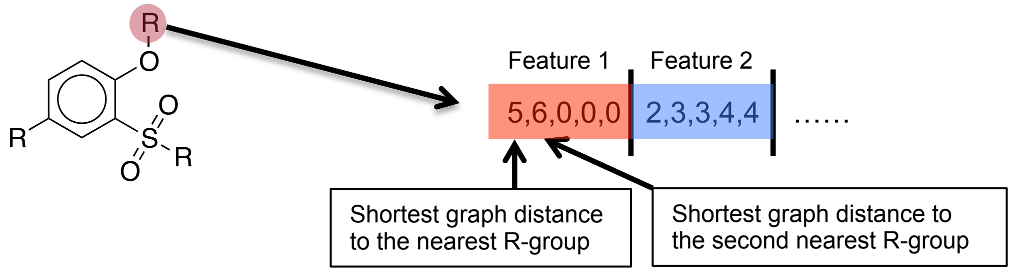
# Fragment Alignments

- Automated alignment of R-groups between a fragment and its replacement is a non-trivial problem



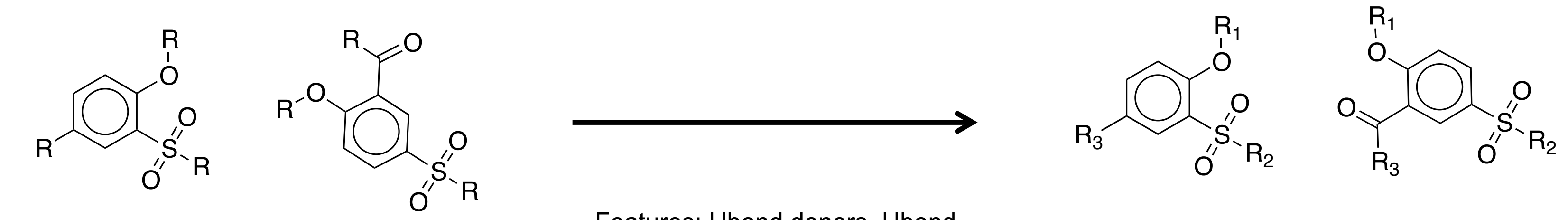
- To get around this problem I developed a new algorithm
  - Rapid 2D Alignment of Topological Scaffolds (RATS)

- Each R-group is abstracted as a fingerprint based on distances to other R-groups and pharmacophoric features

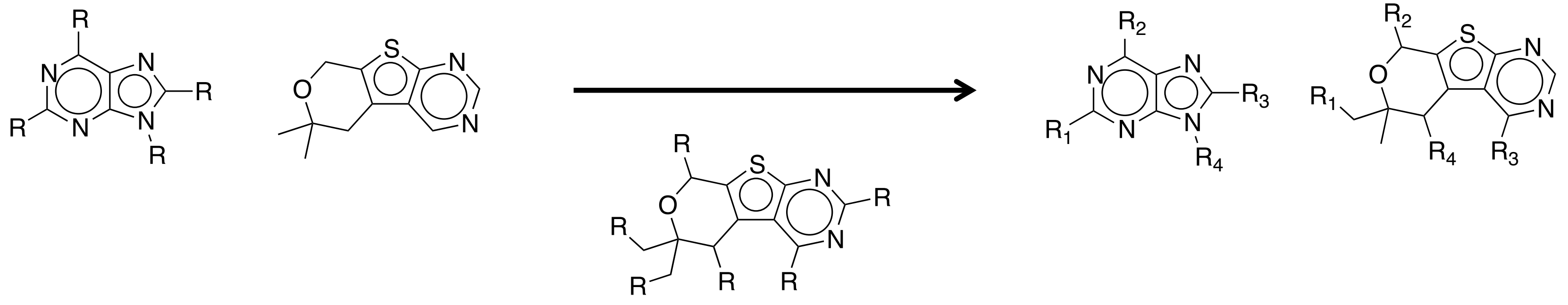

Feature 1
Feature 2
5,6,0,0,0
2,3,3,4,4
……
Shortest graph distance to the nearest R-group
Shortest graph distance to the second nearest R-group

- A similarity metric is then used to assess which 2D alignment is the best for the replacement

# Fragment Alignments - Examples



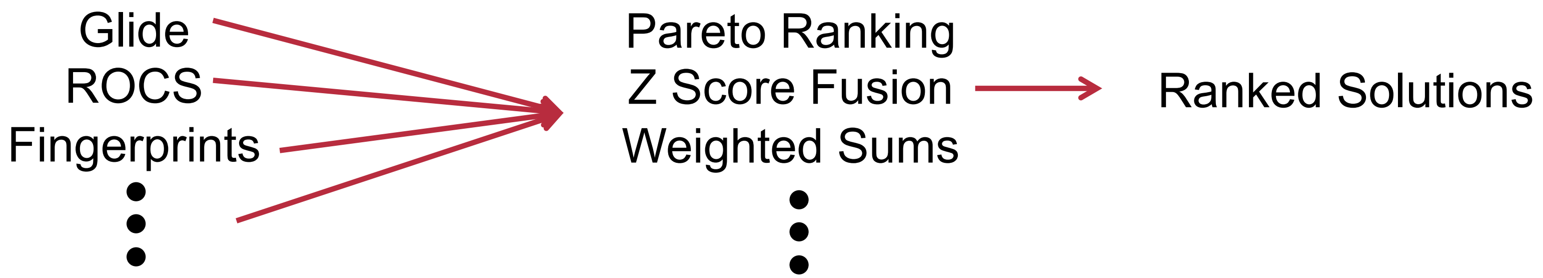
Features: Hbond donors, Hbond acceptors, aromatic atoms and RGroup distance



Features: Hbond donors, Hbond acceptors and RGroup distance

# Scoring and Filtering

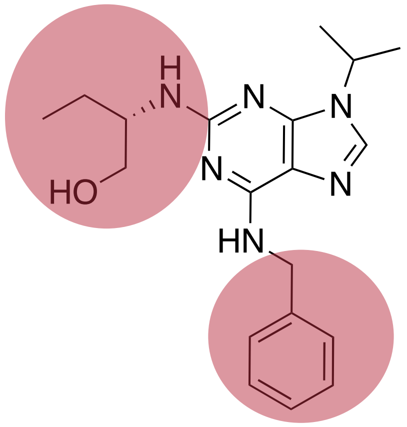
- Due to the modular nature of the program scoring functions can be changed as the user requires
- Currently available scoring methods are Glide (with LigPrep), ChemAxon cxcalc, ROCS (with OMEGA), scikit-learn statistical models, R functions and a variety of fingerprints (homebrewed CATS, ECFP and RDKit )
- A number of methods are available to perform data fusion on the array scoring methods



- Post processing includes substructure filtering *i.e.* PAINS

# Case Study: CDK2/Roscovitine

- A former drug discovery project at the ICR looked at optimising roscovitine to keep/improve activity against CDK2 and improve metabolic properties
- The project teams goal was to prevent oxidative metabolism of the alcohol-containing side and phenyl ring
    - Make these substituent more polar

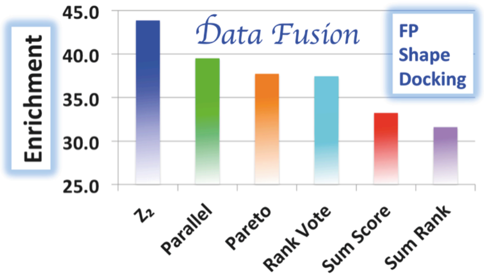


Roscovitine

- So we are mimicking this project by keeping the purine and isopropyl motif the same and optimising the substituents

## Case Study: Modelling Activity

- Given how well studied CDK2 is I wanted to use as much of the available information as possible
- A recent paper[1] published shows that a data fusion method using Z Scores of orthogonal methods gives a good enrichment over individual scores

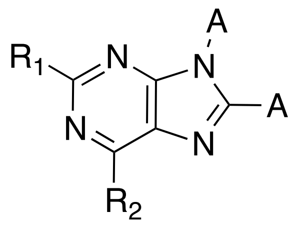


- I have combined Glide, ROCS and RDKit fingerprints to build an activity model

1. Sastry, G. Madhavi, VS Sandeep Inakollu, and Woody Sherman. "Boosting virtual screening enrichments with data fusion: Coalescing hits from 2D fingerprints, shape, and docking." *J. Chem. Inf. Model* (2013).

## Case Study: Modeling Activity

- To improve the activity model we will include a classifier
- The model will be local to the chemical space around the substructure:

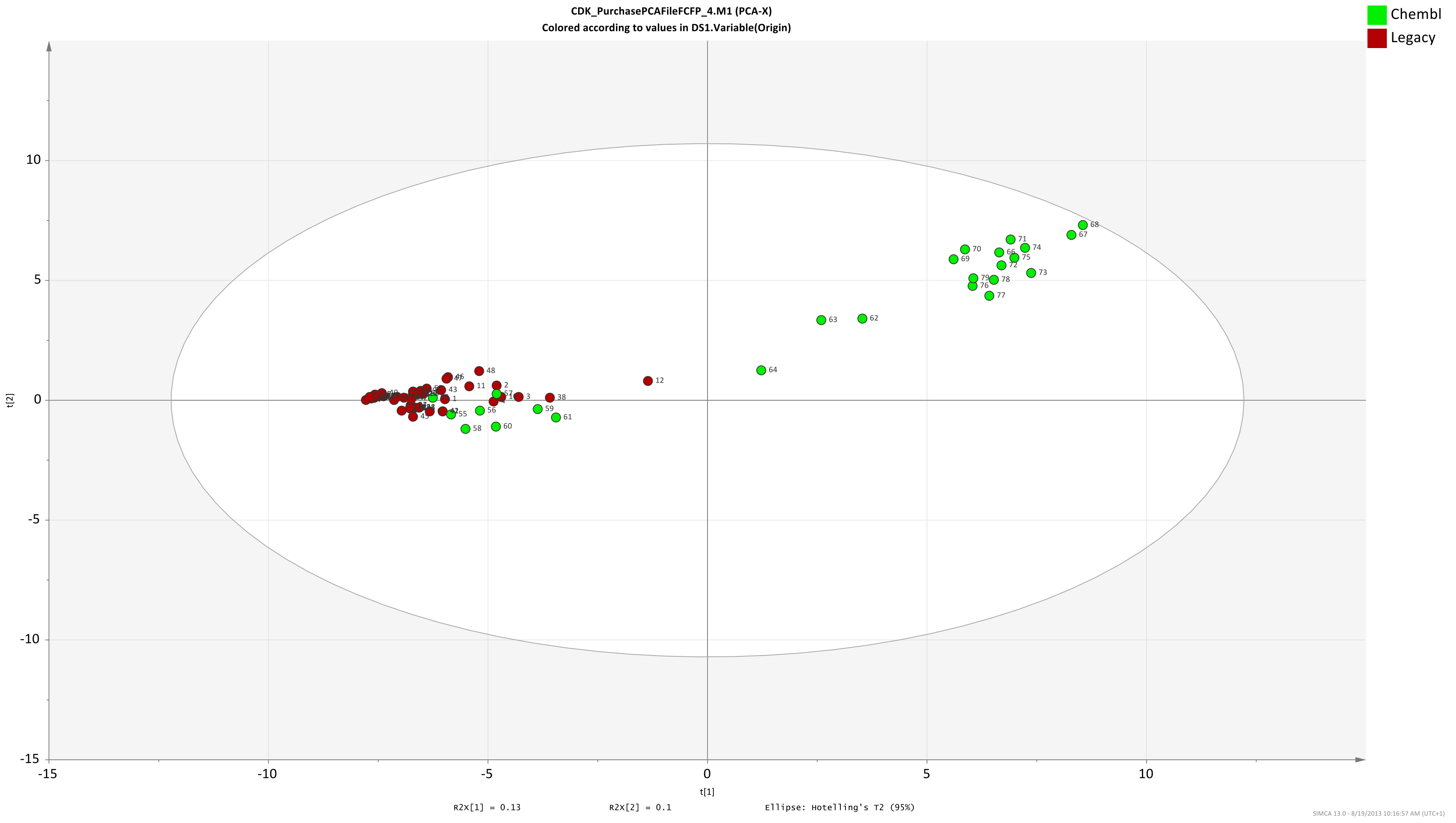


Where A is any atom and $R_1$ and $R_2$ are any group except H

- We have distinct 79 $IC_{50}$'s from in house data and mined from ChEMBL
- In order to make this a robust model I have purchased 119 compounds with the above substructure, I'm currently testing these compounds to get $IC_{50}$'s

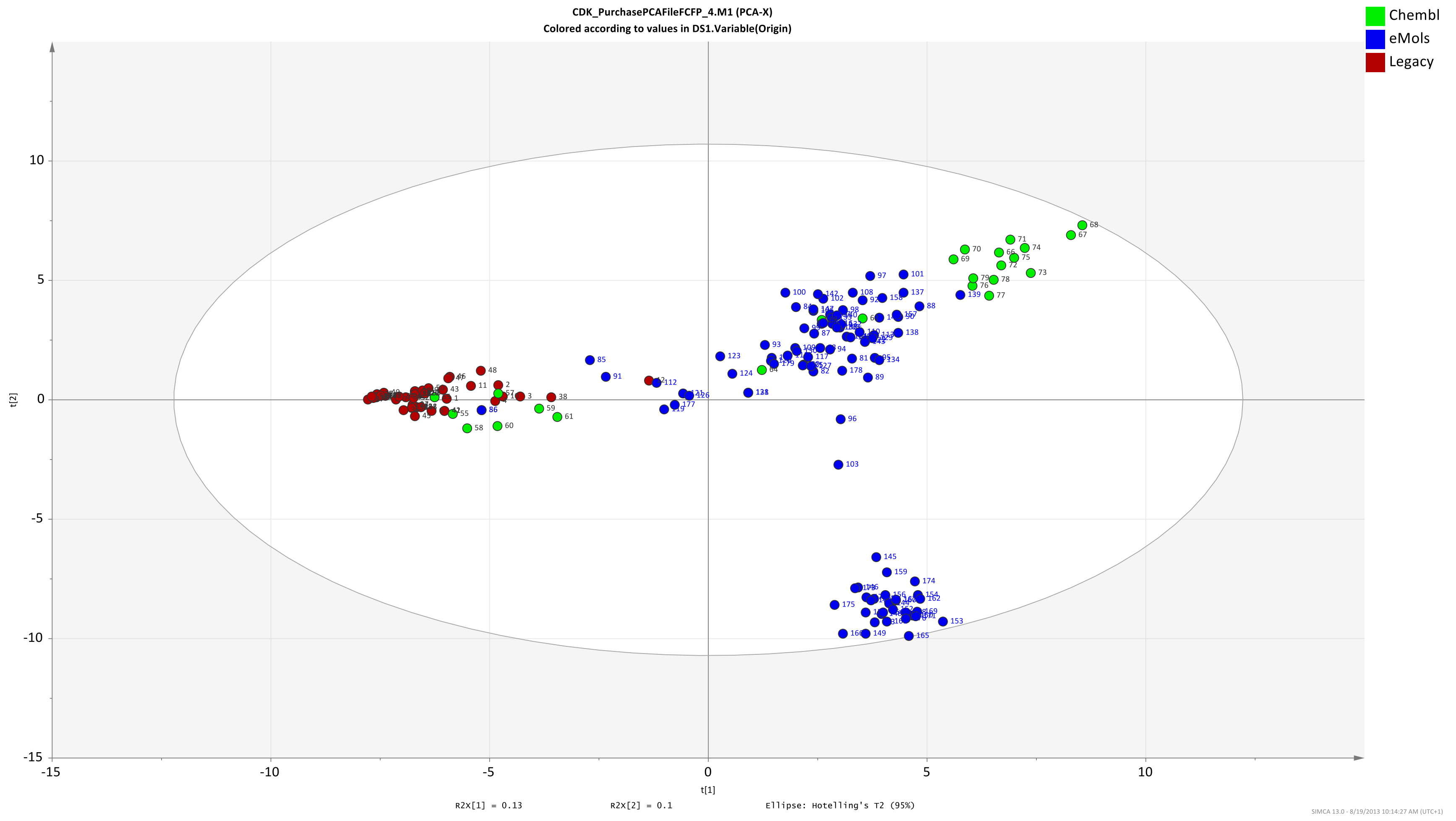
# Case Study: Modeling Activity

- Initially the chemical space that we have represented by our data

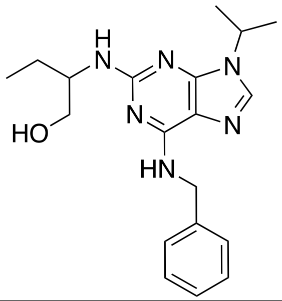
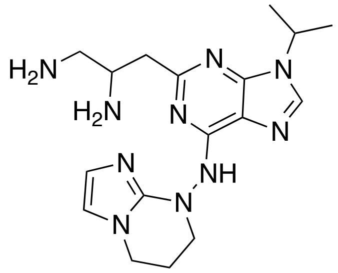
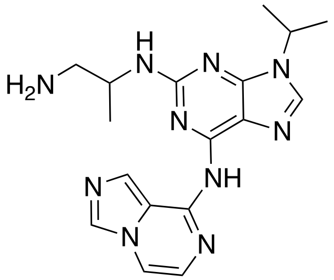
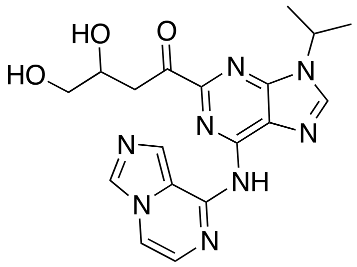
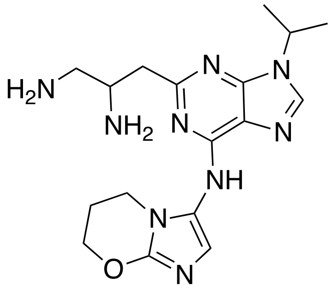
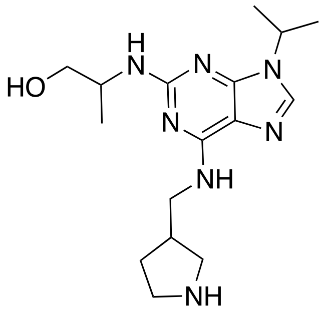


## Case Study: Modelling Activity

- When we add the compounds that we have purchased we need



# Initial Results

| Molecule | Glide | ROCS | FP Sim | Molecule | Glide | ROCS | FP Sim |
|---|---|---|---|---|---|---|---|
|  | -7.05 | 1.49 | 1 |  | -9.39 | 1.19 | 0.420 |
|  | -9.82 | 1.33 | 0.421 |  | -8.90 | 1.18 | 0.430 |
|  | -9.38 | 1.20 | 0.421 |  | -9.00 | 1.62 | 0.317 |

## Conclusions/Ongoing Work

- I have built a *de novo* design workflow which iteratively replaces fragments within a molecule
- I have begun prospective validation of this workflow by optimising a known CDK2 inhibitor
- I am continuing to improve the activity model used in this validation by the inclusion of more data to build a classifier
- I have started comparing initial results with this optimisation to those of a reaction based *de novo* design algorithm (DOGS)

# Acknowledgements

- Supervisors
  - Julian Blagg
  - Nathan Brown
  - Swen Hoelder

- *In Silico* Medicinal Chemistry
  - Sarah Langdon
  - Lewis Vidler
  - Ngai Yi Mok
  - Fabio Broccatelli

- ETH Zurich
  - Gisbert Schneider
  - Daniel Reker

- Others
  - Jonty Macdonald
  - Gary Nugent

- Funding

ICR The Institute of Cancer Research

ICR

# Cut Rules

| Cut Rule |
|---|
| 1 Biaryl bonds |
| 2 Alkene bonds |
| 3 Exocyclic iodine bonds |
| 4 Bonds between any two non cyclic heteroatoms |
| 5 Glycosidic linkages |
| 6 Bond between heteroatom and an unsaturated system with an alpha heteroatom |
| 7 Akyl heteroatoms |
| 8 Benzylic bonds |
| 9 Exocyclic bonds from a nitrogen in a cyclic system |
| 10 Enolic bonds |
| 11 Bonds between heteroatoms and vinyl/alkynyl when there is an electron withdrawing group in the $\beta$ position |

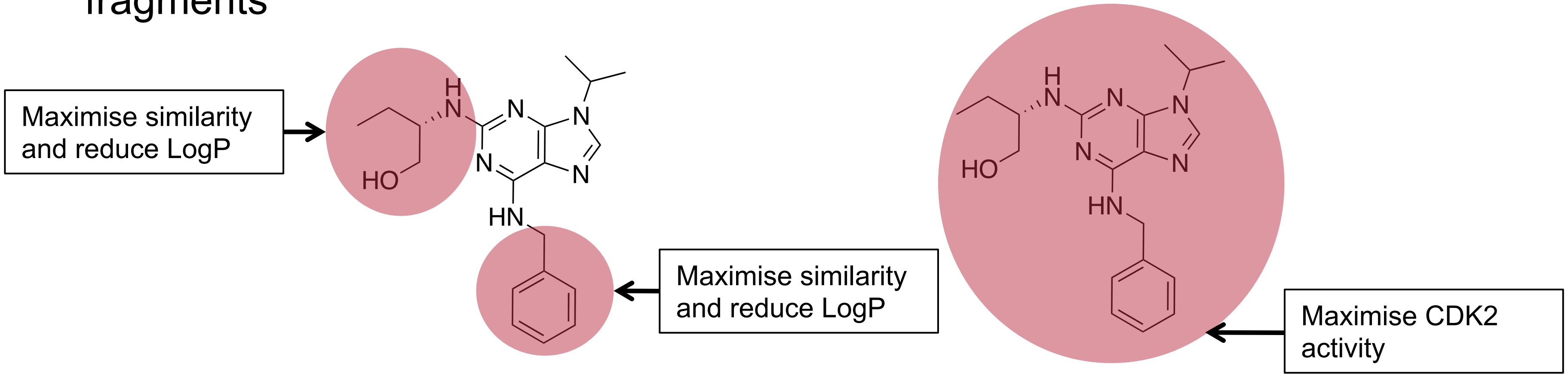
TABLE 6. Ordered set of cut rules

| Prevent Cut Rule |
|---|
| 1 Do not break any triple bonds |
| 2 Do not break any rings |
| 3 Do not break a bond which leaves a heteroatom as a fragment |

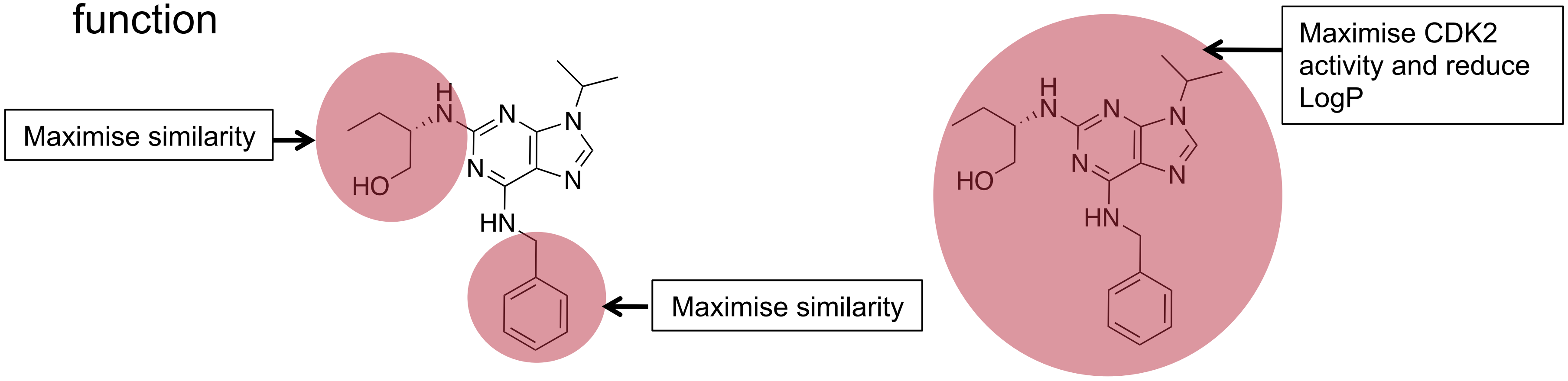
TABLE 7. Set of non cut rules.

# Case Study: Incorporating ClogP

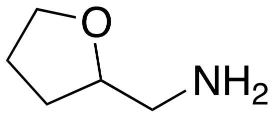
- Initially ClogP was considered during the choosing of potential replacement fragments


Maximise similarity and reduce LogP
Maximise similarity and reduce LogP
Maximise CDK2 activity
HO
H
N
N
N
N
N
HN

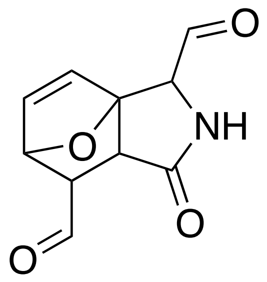
- I then compared results by using ClogP as an objective in the scoring function



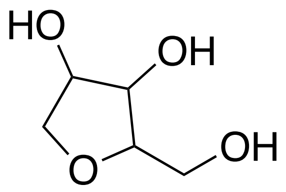
# Examples of THF



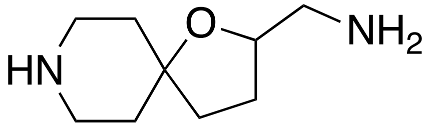
33,832



1,249



5,136



697