



# *De novo* design of synthetically feasible compounds using reaction vectors and evolutionary multiobjective optimization.

Ben Allen



# Computational *de novo* Design

- Introduction to *de novo* design
- Methods
  - Reaction Vectors
  - Genetic Algorithms
  - Multiobjective Optimization
  - Computational Development
- Objective Functions and Test Sets
  - Objective Function Criteria
- Results
  - Sulmazole Lead Optimisation
  - Thrombin Fragment to Drug
- Conclusions



# What is *de novo* design?

- Generation of new chemical structures that satisfy design goals:
  - Activity and selectivity
  - Physiochemical parameters
  - Synthetic feasibility



# Previous Approaches

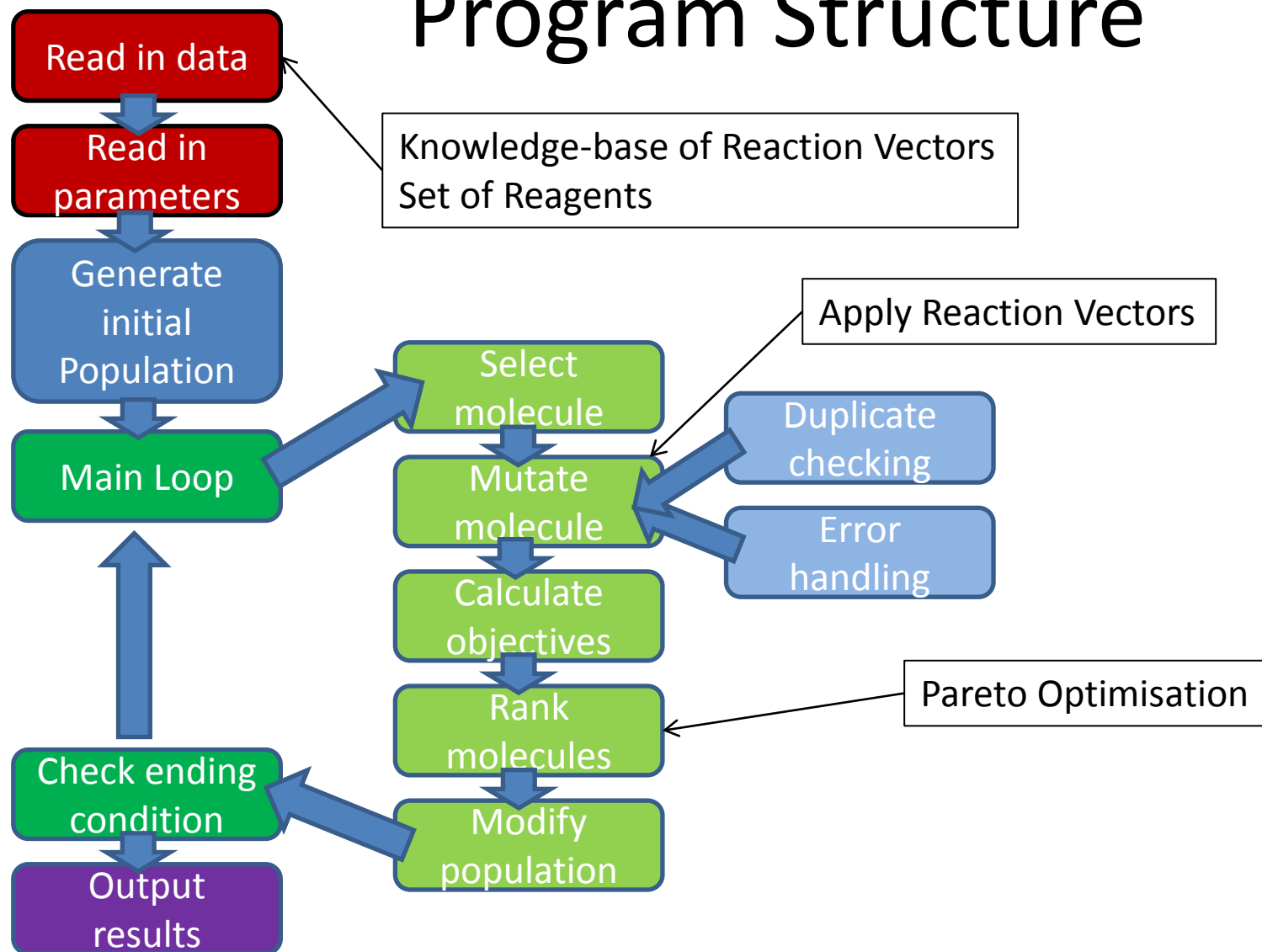
- Over 2 decades old
- More than 20 published computational tools
  - HSITE/2D Skeletons, 3D, LEGEND, LUDI, NEWLEAD, CONCEPTS, SPROUT, MCSS&HOOK, SMOG, CONCERTS ,LEA, LigBuilder, TOPAS, F-DycoBlock, ADAPT, SYNOPSIS, CoG, BREED
- Methods include:
  - Receptor vs Ligand
  - Atom vs Fragment
  - Growing vs Linking



# Methods

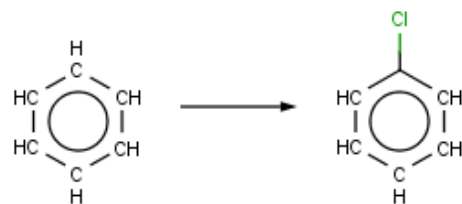
- Synthetic feasibility
  - Transformations based on user-definable knowledge-base of known reactions
- Effective searching in large chemical spaces
  - Genetic algorithm
- Generate active, drug-like molecules
  - Multiobjective optimisation

# Program Structure



# Reaction Vectors

Reagent Atom Pair	Count
C(2,2,1)-2(4)-C(2,2,1)	6
C(2,2,1)-3-C(2,2,1)	6



Product Atom Pair	Count
C(2,2,1)-2(4)-C(2,2,1)	4
C(3,2,1)-2(1)-Cl(1,0,0)	1
C(3,2,1)-2(4)-C(2,2,1)	2
C(2,2,1)-3-C(2,2,1)	4
C(3,2,1)-3-C(2,2,1)	2
C(2,2,1)-3-Cl(1,0,0)	2

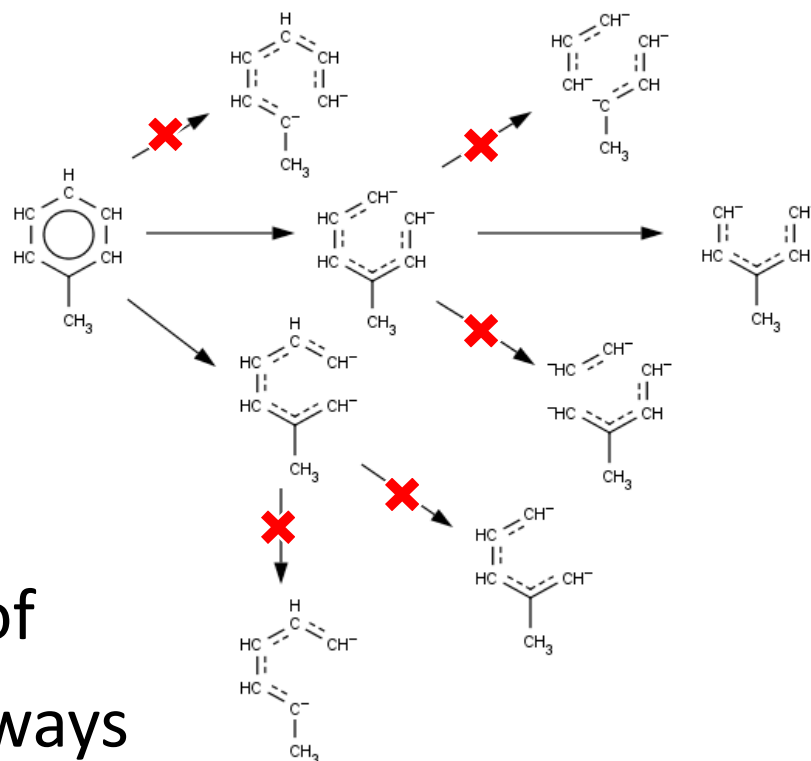
Reaction Vector			
Positive Atom Pairs		Negative Atom Pairs	
C(3,2,1)-2(4)-C(2,2,1)	+2	C(2,2,1)-2(4)-C(2,2,1)	-2
C(3,2,1)-2(1)-Cl(1,0,0)	+1	C(2,2,1)-3-C(2,2,1)	-2
C(3,2,1)-3-C(2,2,1)	+2		
C(2,2,1)-3-Cl(1,0,0)	+2		

**Atom Pairs:** X1(h,p,r)-S(o)-X2(h,p,r) where:

- X1 and X2 are the atomic symbols.
- S is the path separation between the atoms.
- h is the number of non-hydrogen connections.
- p is the number of  $\pi$  electrons.
- r is the no. of rings
- o is the bond order (only relevant for S=2).

# Reaction Vectors

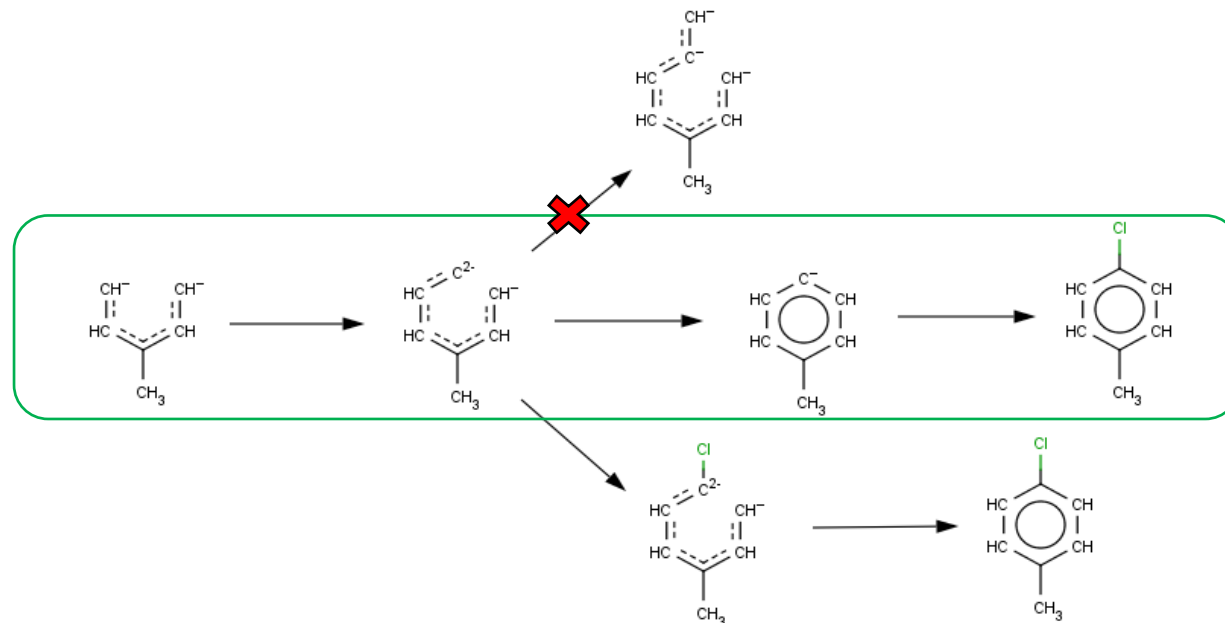
- Generation of Reagent Fragments
  - Breadth first search



- Elimination of incorrect pathways

# Reaction Vectors

- Growth of Product Fragments





# Reaction Vector Knowledge-base

- Stored as SQL database
- Includes:
  - the reaction vector
  - the environment around each broken bond
  - the fragmentation path
  - the reconstruction path
  - the original reagent and reactants
- Benefits:
  - Very rapid database searching
  - Fast de novo generation



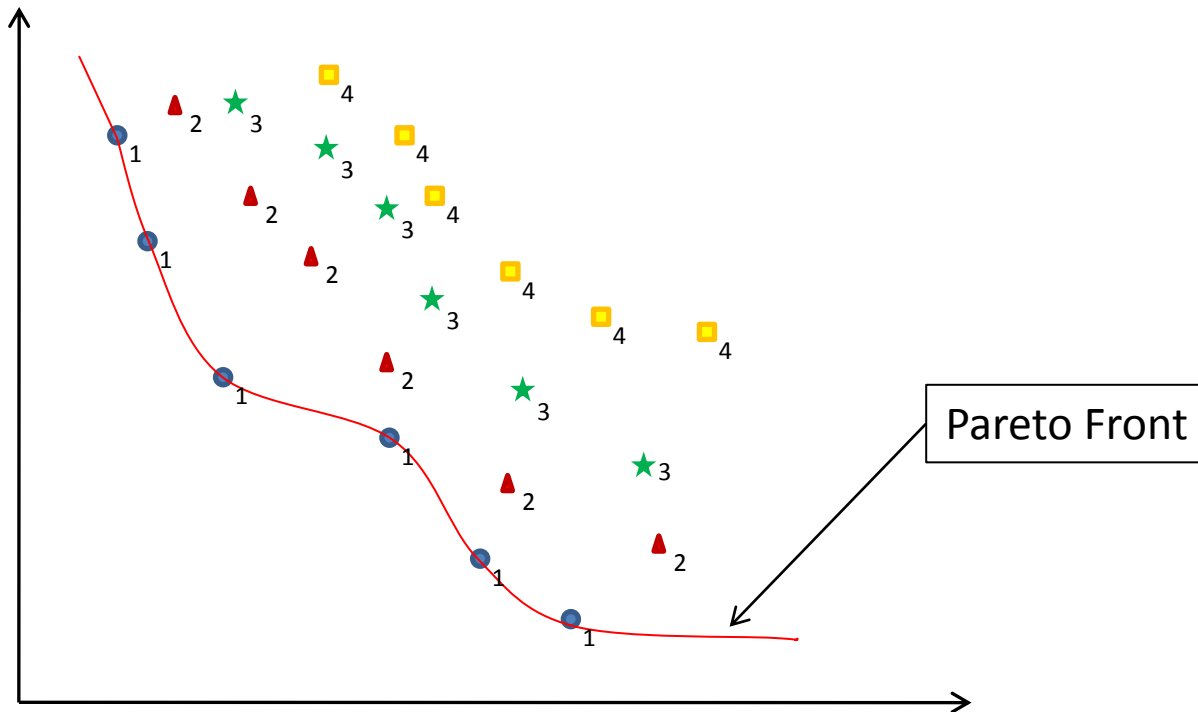
# Reaction Diversity

Alcohol dehydration	Claisen rearrangement	Beckmann rearrangement
Aldol condensation	Fischer indole	Diels-Alder cycloaddition
Dieckmann condensation	Ether halogenation	Olefination
Friedel-Crafts acylation	Hetero Diels-Alder	Claisen condensation




# Multiobjective Optimization

- Pareto front for two objective example
  - Modified NSGA II algorithm



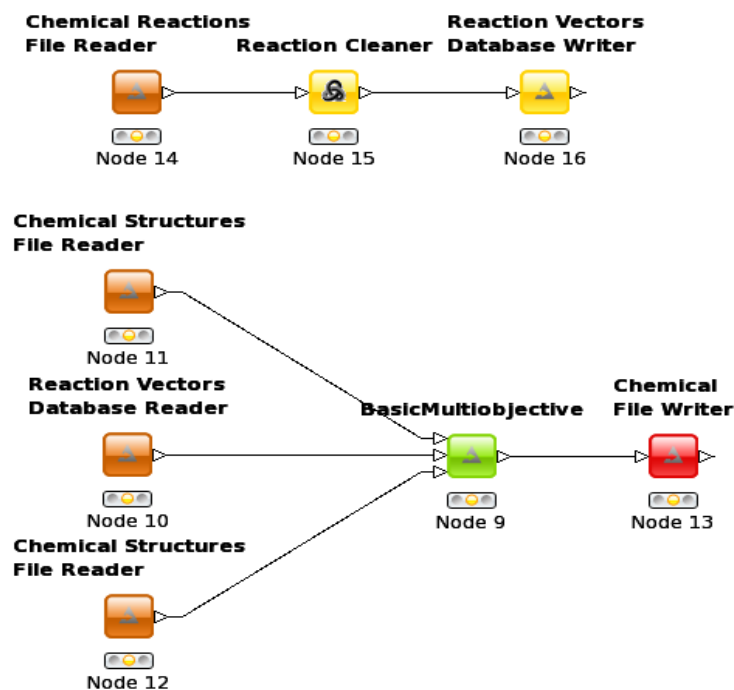
# Genetic Algorithm

- Genetic operators
  - Apply reaction vector
  - Replace last reaction vector
- Roulette wheel selection 
- Ratio controlled by selection pressure

$$P_n = \frac{S(N + 1 - R_n) + R_n - 2}{N(N - 1)}$$

# Computational Development

- Knime environment
  - Pre-processing of inputs
  - Analysis of output

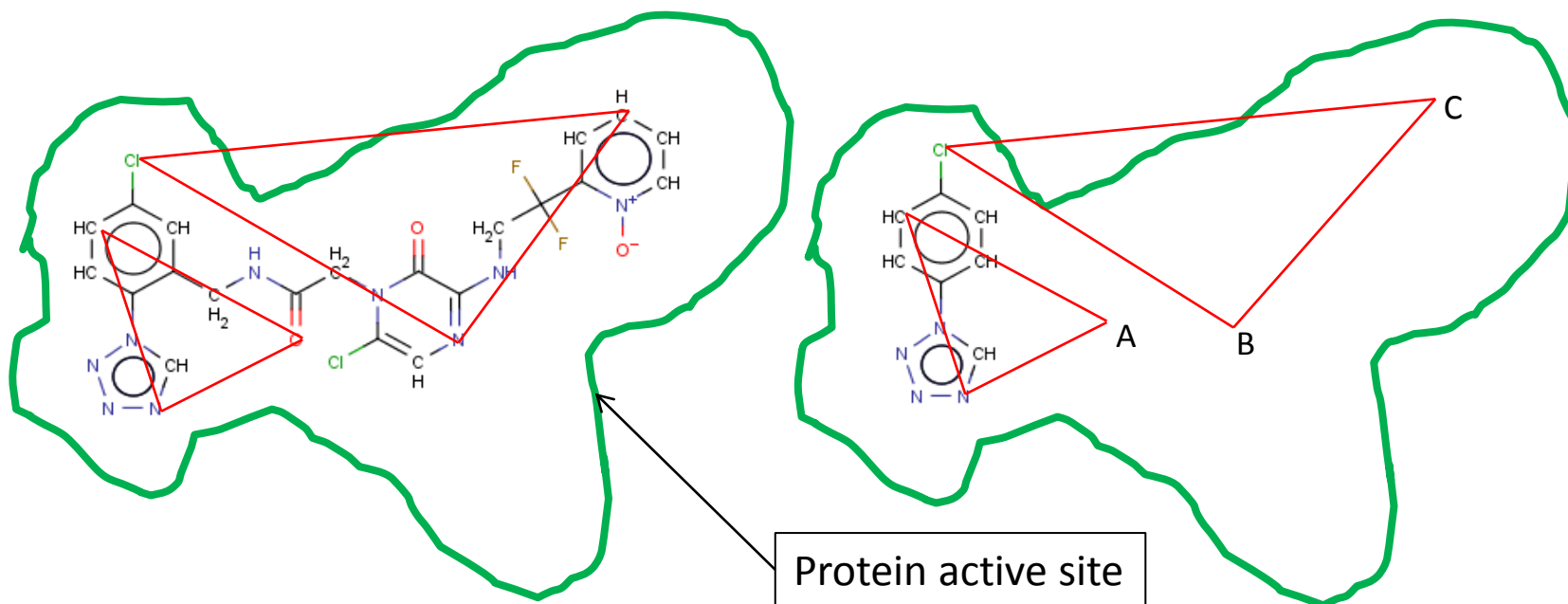




# Objective Functions

- Flexible approach
  - Currently generated from 2D structures
    - 3D methods would require conformer generation
- For example:
  - Physiochemical parameters
    - logP, mass, atom counts etc
  - Similarity to known target(s)
  - Fit to QSAR

# 2D Pharmacophore



- Pharmacophore triplets using:
  - Donor, Acceptor, Polar, Anion, Cation or Hydrophobe
- Provide a growth route for *de novo* design:  $A \rightarrow B \rightarrow C$



# Test Sets

- Initial test set: Lead Optimization
  - Simple problem, with clear objectives
  - Sulmazole
- More complex problem: Fragment to Drug
  - Objective to generate drug like molecules
  - Fragment based design:
    - Input fragments as starting molecules
  - Thrombin



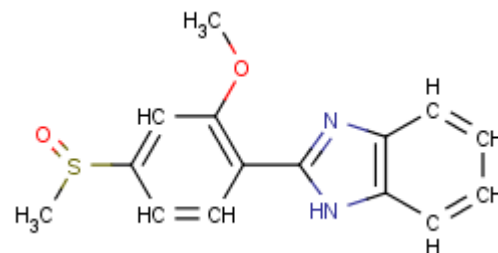
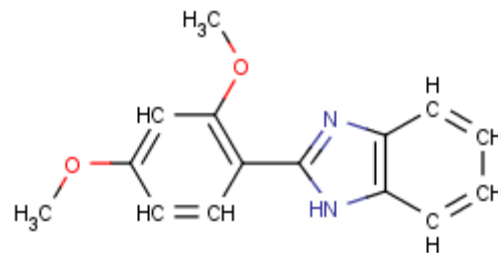
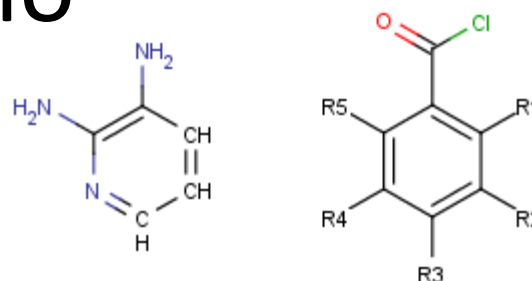
# Parameter Setup

- Reaction knowledge-base
  - 25K reactions extracted from J Med Chem
  - 5K reagents (mw < 150; contains Carbon)
- GA parameters
  - Population 20-200
  - Iterations 5-20
  - Selection Pressure 1.1-1.25



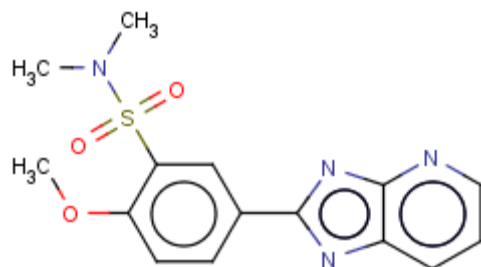
# ARL57 to Sulmazole: typical L.O. scenario

- Initial population generated from:
  - Dataset of targeted acid chlorides
- Optimise based on:
  - Maximise Atom Pair Tanimoto similarity to ARL 57
  - Minimize LogP
- Goals:
  - Reproduce sulmazole
  - Generate ARL 57 analogues with improved LogP

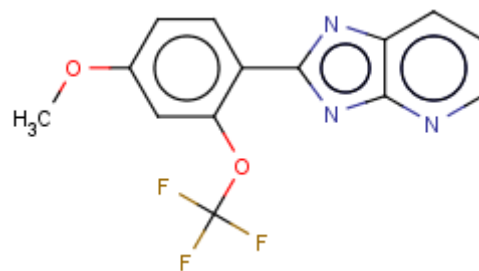


# Sulmazole Results

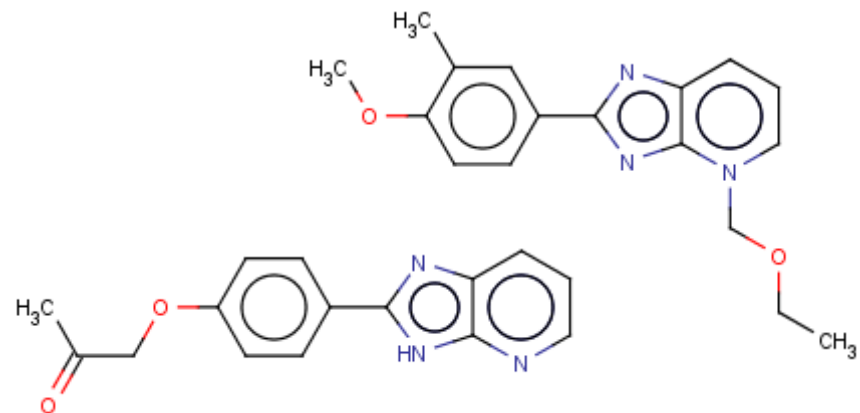
- Reliably generated molecules with better objective values than Sulmazole
- Produced novel, synthetically realistic structures
- Capable of regenerating both sulmazole and ARL 57



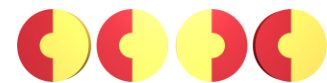
6/28/2011



20







# Thrombin

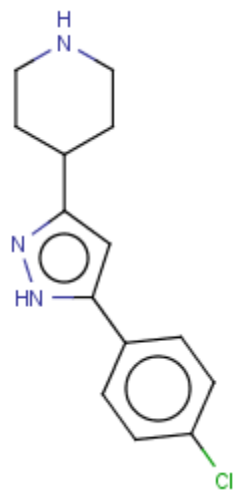
- Benefits:
  - Widely studied real world problem
  - Large selection of fragments and inhibitors in the literature
  - Well described protein target allows for comparison with docking methods

# What is FBDD?

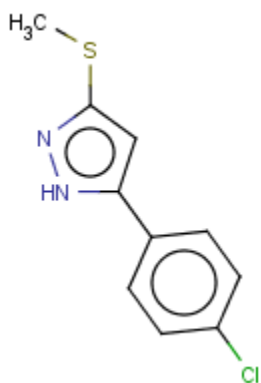
- Use fragments (<15 H.A., <160Da) for initial screening
- Generate drugs by building on or linking fragments
- Benefits:
  - Smaller search space:  $10^7$  fragments v's  $10^{60}$  drugs
  - Chemical simplicity
  - Good ligand efficiency
- For *de novo* design:
  - Fragments as starting materials
  - Known compounds as targets

# Thrombin Set 1

- Thrombin fragments



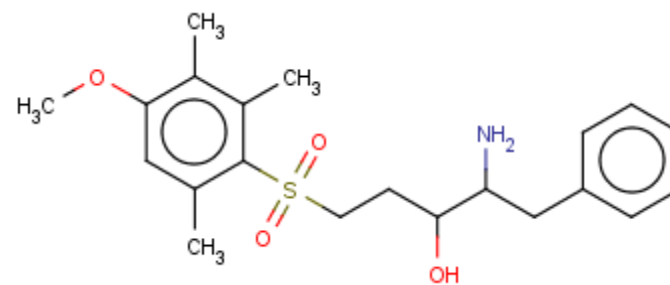
$IC_{50} = 400\mu M$



$IC_{50} = 1000\mu M$



$IC_{50} = 330\mu M$



$IC_{50} = 12\mu M$

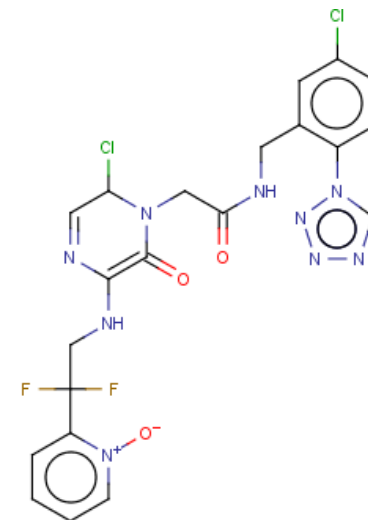
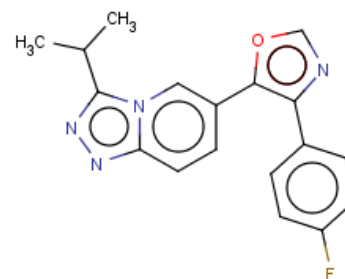
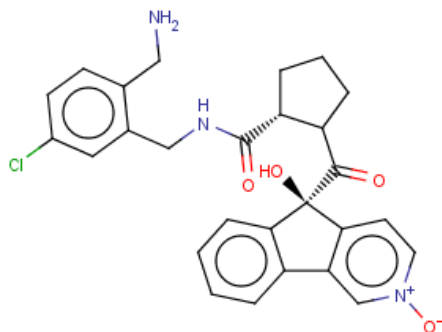
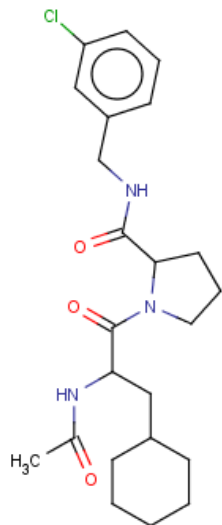
– Diverse set

- Different binding modes?



# Thrombin Set 1

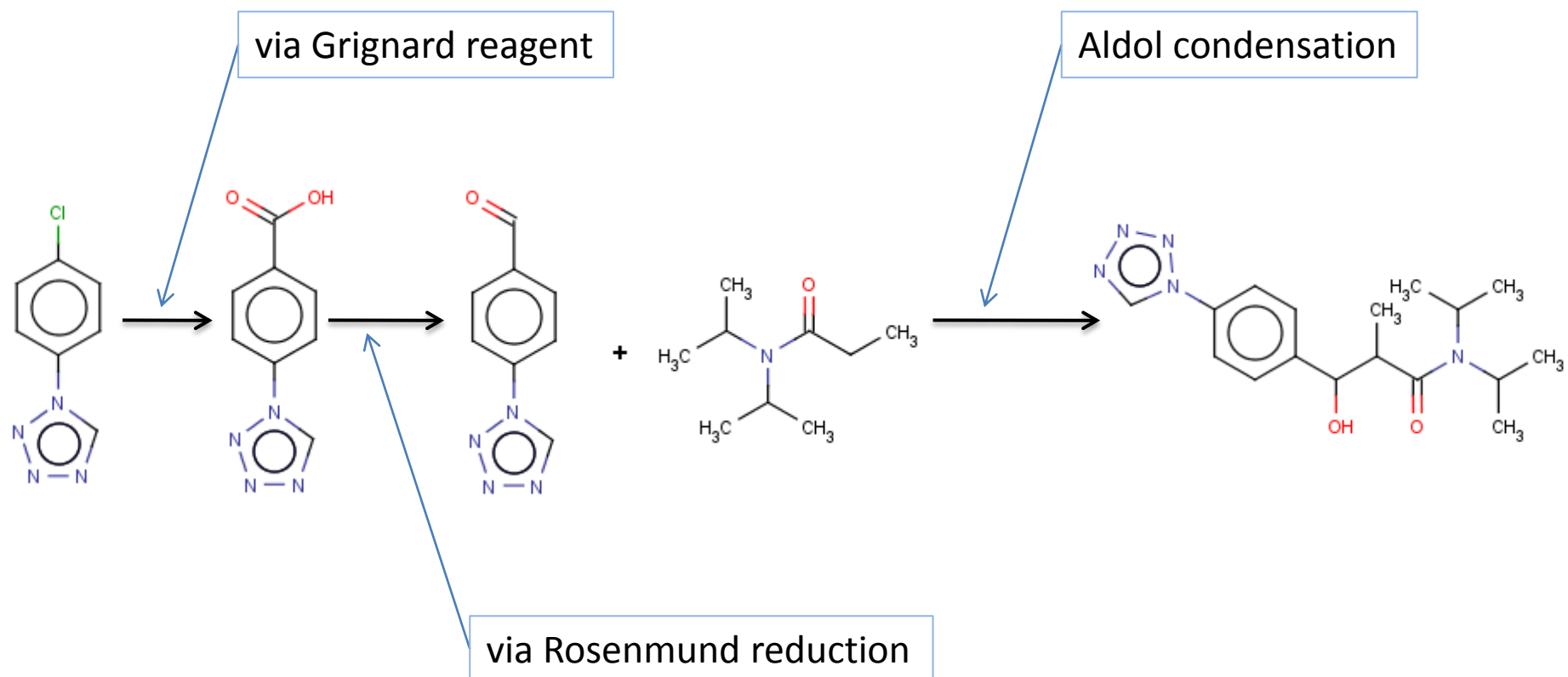
- Objectives:
  - Pharmacophore triplet similarity to a known thrombin inhibitor
  - Plausible synthetic pathways
  - Drug-like physicochemical properties





Target Inhibitor	<i>De novo</i> Molecule	Similar Known Inhibitor

# Synthetic Route Generation



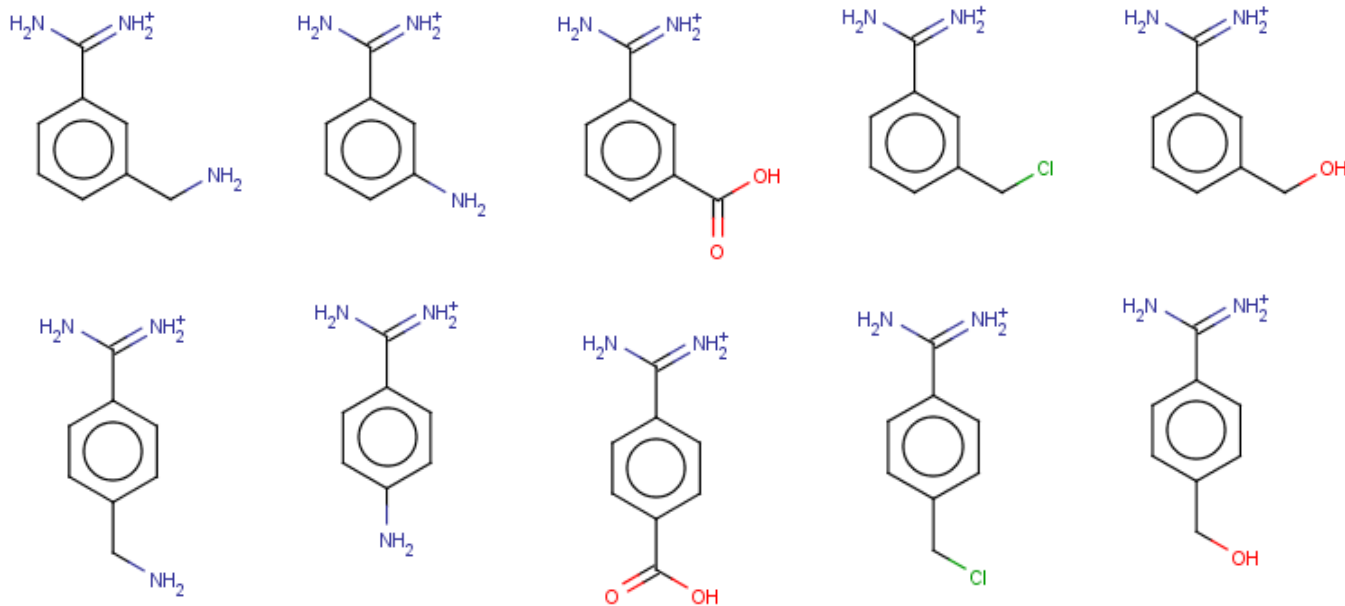


# Thrombin Set 1 Results

- Results
  - Novel
  - Similar to known inhibitors
  - Plausible synthetic pathways
- Problems
  - Elitism
  - Not sufficiently thrombin-inhibitor like (expert opinion)

# Thrombin Set 2

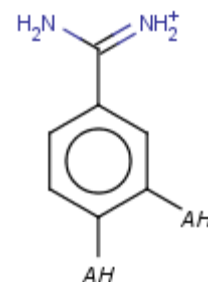
- Initial fragments



– Highly similar set

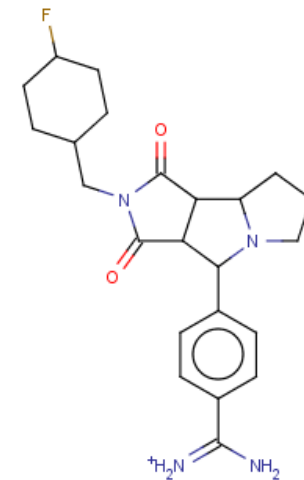
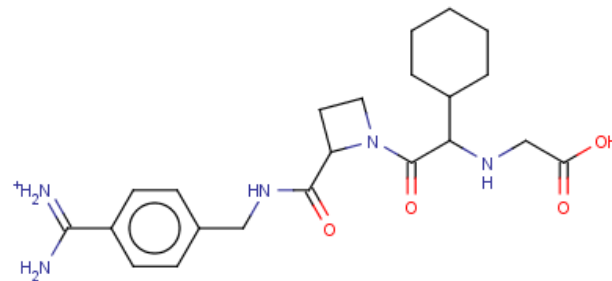
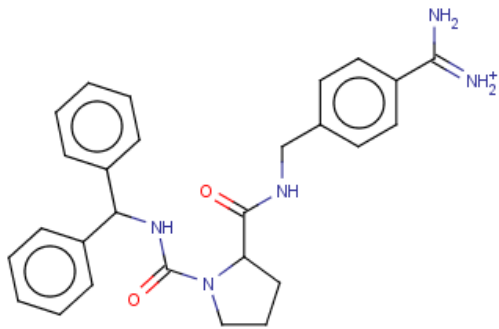
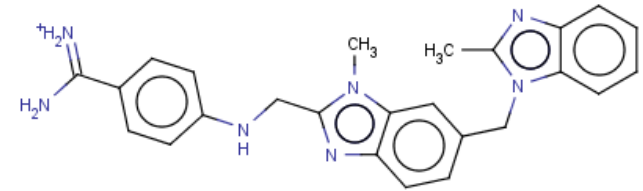
# Thrombin Set 2

- Substructure preservation
- Consensus pharmacophore
  - Generated by taking the set of fingerprints found in 3 out of 4 target inhibitors



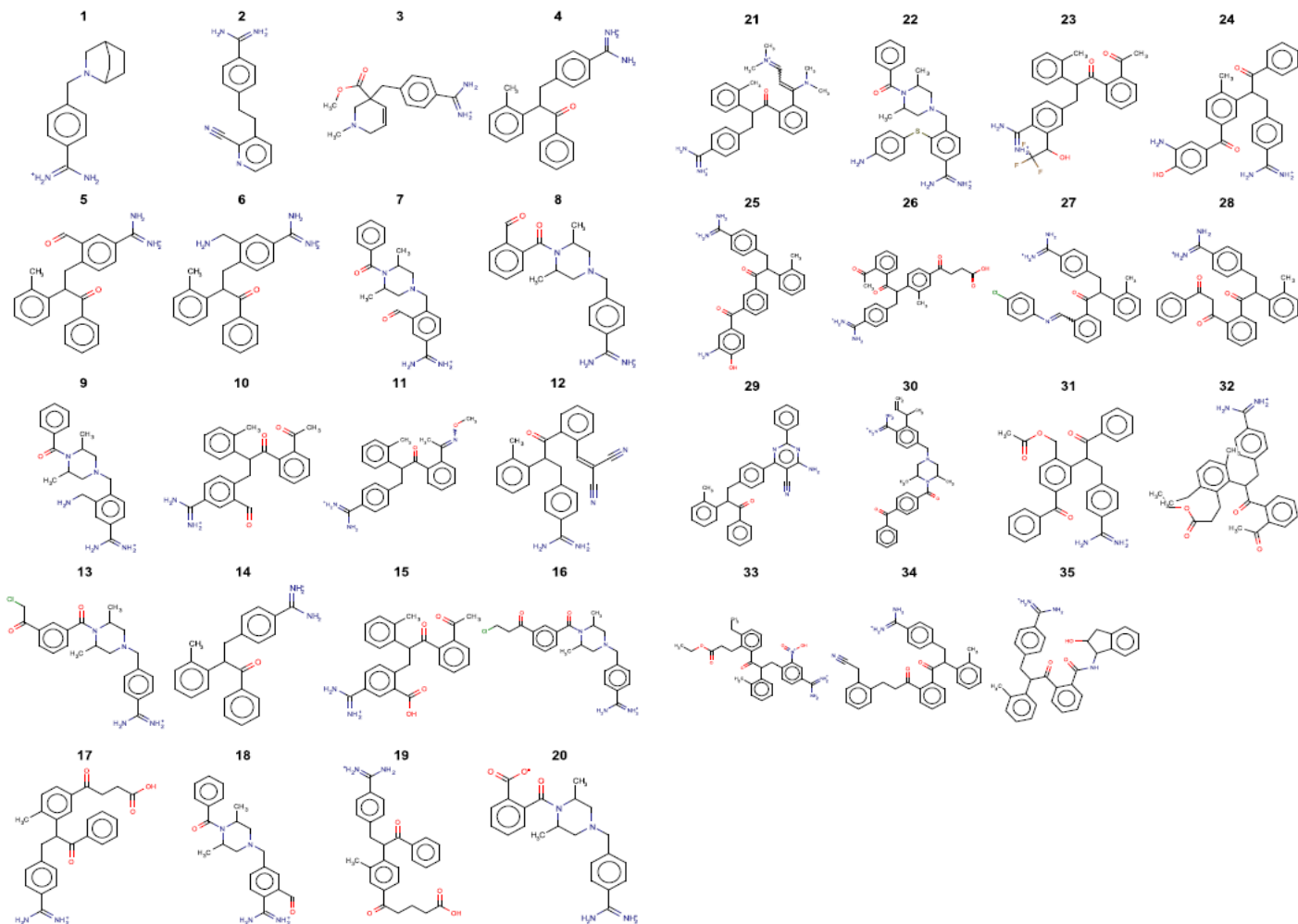
# Thrombin Set 2

- Known inhibitors
  - All contain conserved motif





# Thrombin Set 2 Results

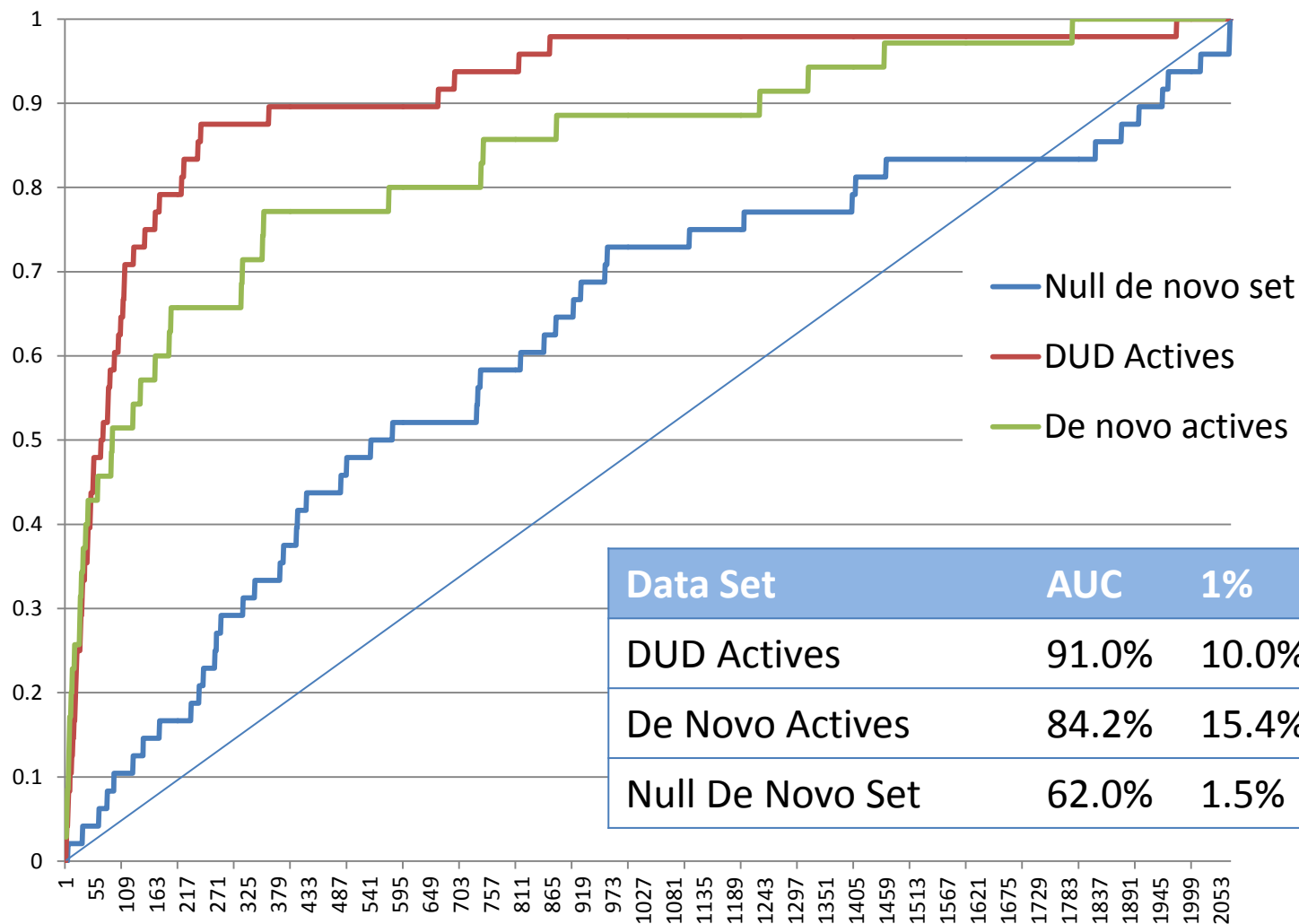


# Comparison with DUD dataset

- Dock thrombin actives and decoys to thrombin active site using GOLD
- Replace the active molecules with *de novo* molecules
- Dock *de novo* compounds and decoys to thrombin active site
- Null set – generated using *de novo* design with no pharmacophore objective
- Compare enrichments

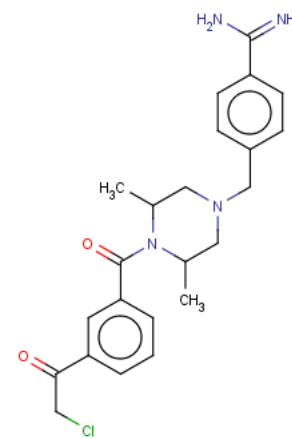
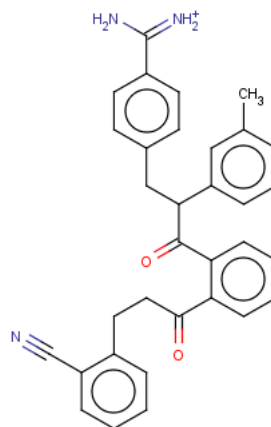
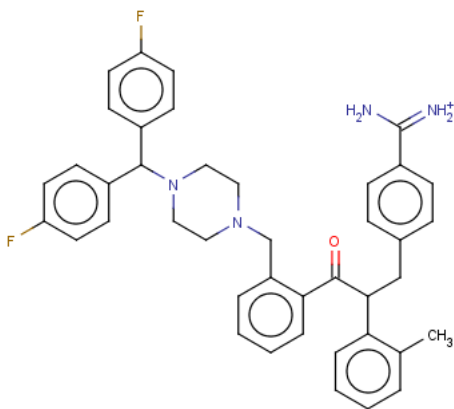
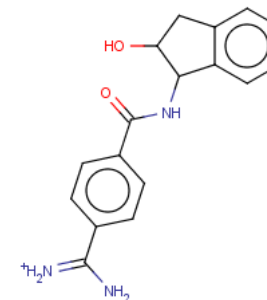


# Enrichment Plots



# Thrombin Set 2 Results

- Druglike
- Good physiochemical properties
- Active pharmacophore features
- **Novel** (1 of 37 found in ChEMBL Thrombin antagonists)





# Conclusions & Further Work

- Generates sets of molecules which satisfy the desired objectives
- Suggests plausible synthetic routes based on known reactions
- With suitable objectives/inputs, it can generate novel drug-like molecules
- Further work:
  - Structural niching to increase output diversity
  - Improved objective functions



# Acknowledgements

- University of Sheffield
  - Val Gillet
  - Bening Chen
- Eli Lilly
  - Mike Bodkin
  - Hina Patel (formerly University of Sheffield)
  - Dimitar Hristozov (now at FDA)
- Cambridge Crystallographic Data Centre
  - John Liebeschuetz
  - Jason Cole