



The
University
Of
Sheffield.



Extracting SARs using a Multi-Objective SMARTS Evolutionary System

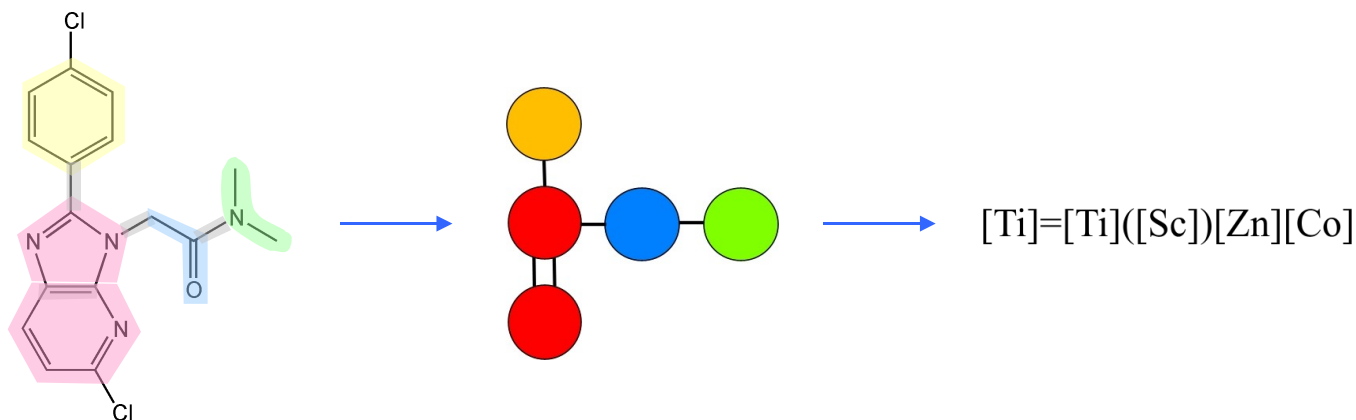
Kristian Birchall

lip03kb@sheffield.ac.uk

The Bigger Picture

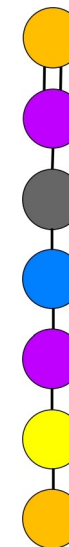
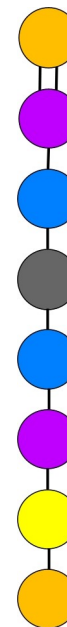
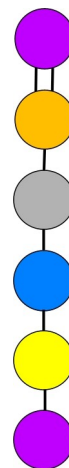
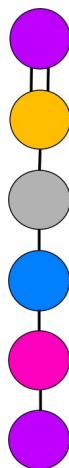
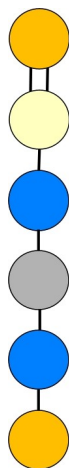
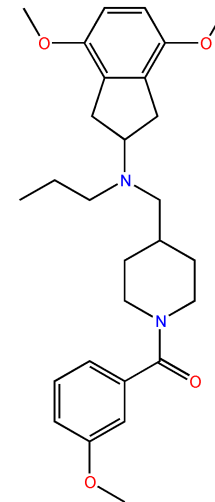
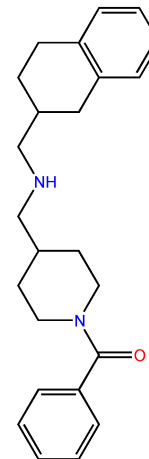
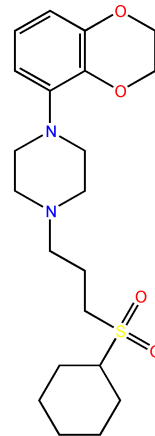
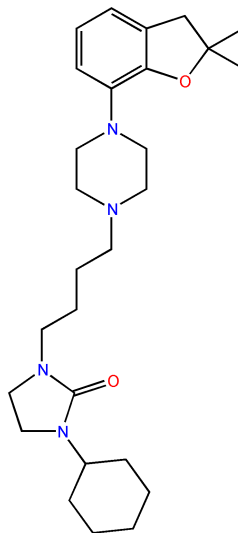
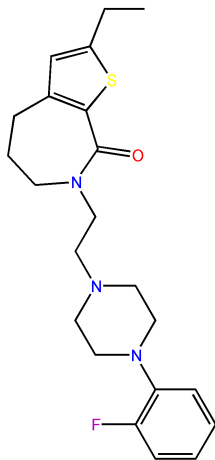
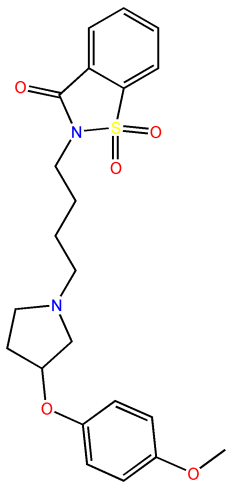
- Increase chances of finding bioactives
 - Based on primary screening data
 - Ideally want some understanding of why active
- Activity determined by structure
 - Which parts of the structure are important?
- Machine learning approach to answer this
 - Which structural features can accurately distinguish actives from inactives

- 2D structural fragments reduced to graph nodes based on predefined functional classifications
 - Aromatic/Aliphatic rings, H-bond Donors/Acceptors etc.
 - Ar/F(4) definitions in Gillet *et al* 2003 **JCICS** 48 338-345
 - Topology retained – including ring fusion

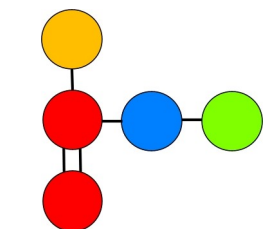




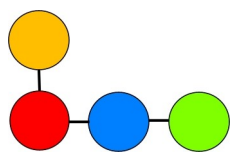
Graph Reduction – 5HT1A Agonists



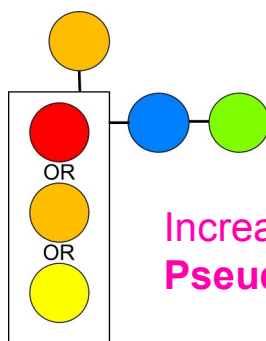
Structural Queries



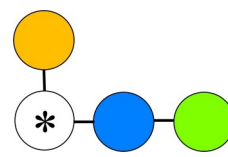
Active RG



Fixed sub-structural RG query

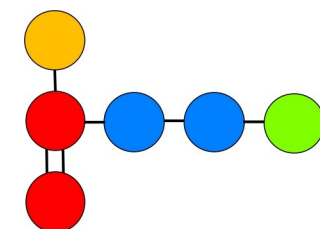
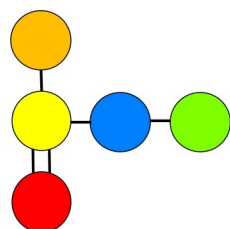
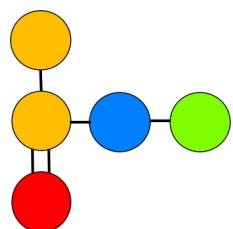


Increased matching flexibility using Pseudo RG SMARTS

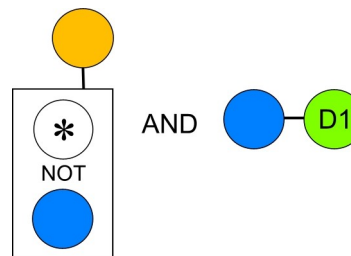


Actives not found by fixed sub-structural query

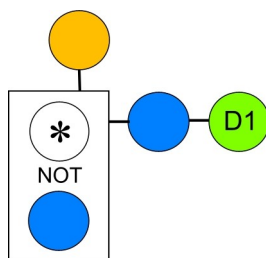
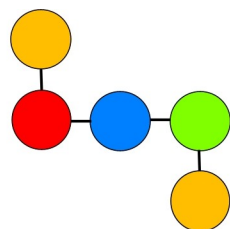
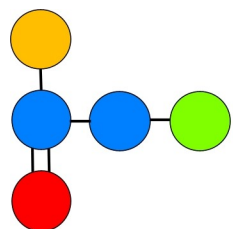
- Alternative (OR) node types possible at a particular position
- Any node type matched by wildcard (*)



- Disjoint structures can be matched (AND)



Inactives to be excluded



- Ability to exclude matching (NOT)
- Specification of connectivity (D)

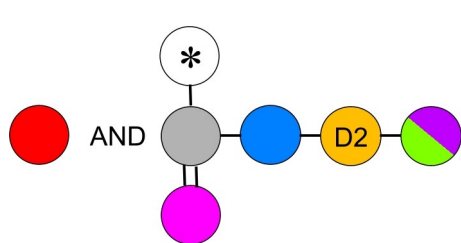


Aim - Clarified

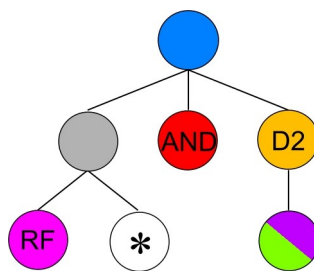
Given a mixture of known active and inactive RGs, train a genetic program to evolve flexible sub-structural RG queries, which are best able to separate the actives from the inactives

➤ Genetic Programming (GP) <http://www.genetic-programming.org/>

- Pseudo RG SMARTS queries directly represented and manipulated as trees
- Flexible searching operators simply “tagged” to nodes



Reduced Graph Query



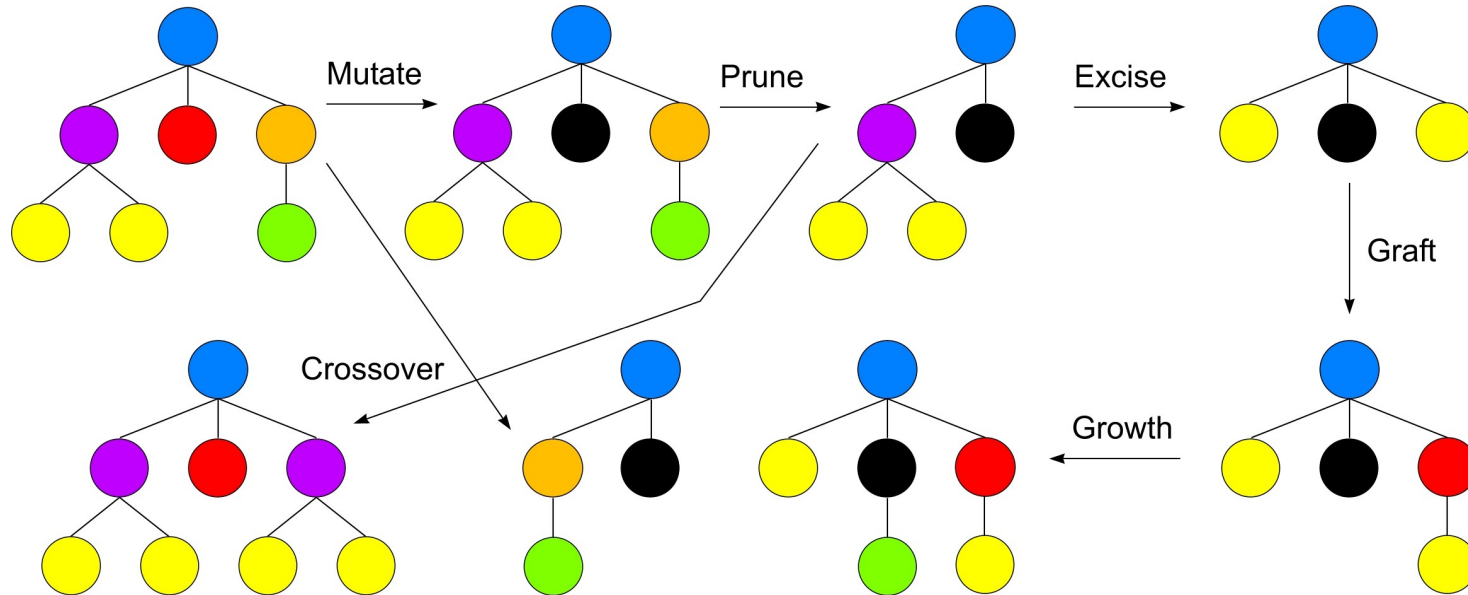
Tree-based chromosome

[Zn]([W](=[V])[*])(.[Ti])[Sc;D2]([\$([Co,Y])])

Pseudo SMARTS query string

- Iterative modification and evaluation of queries to evolve ones better able to separate active and inactive RGs

Evolutionary Operators



➤ Intra-node operations

- Expand/Contract OR list, Add/Remove node tag

➤ Query validity checks

➤ Generate new queries - favour better performance

- For each query search through a dataset containing a mix of active and inactive RGs:

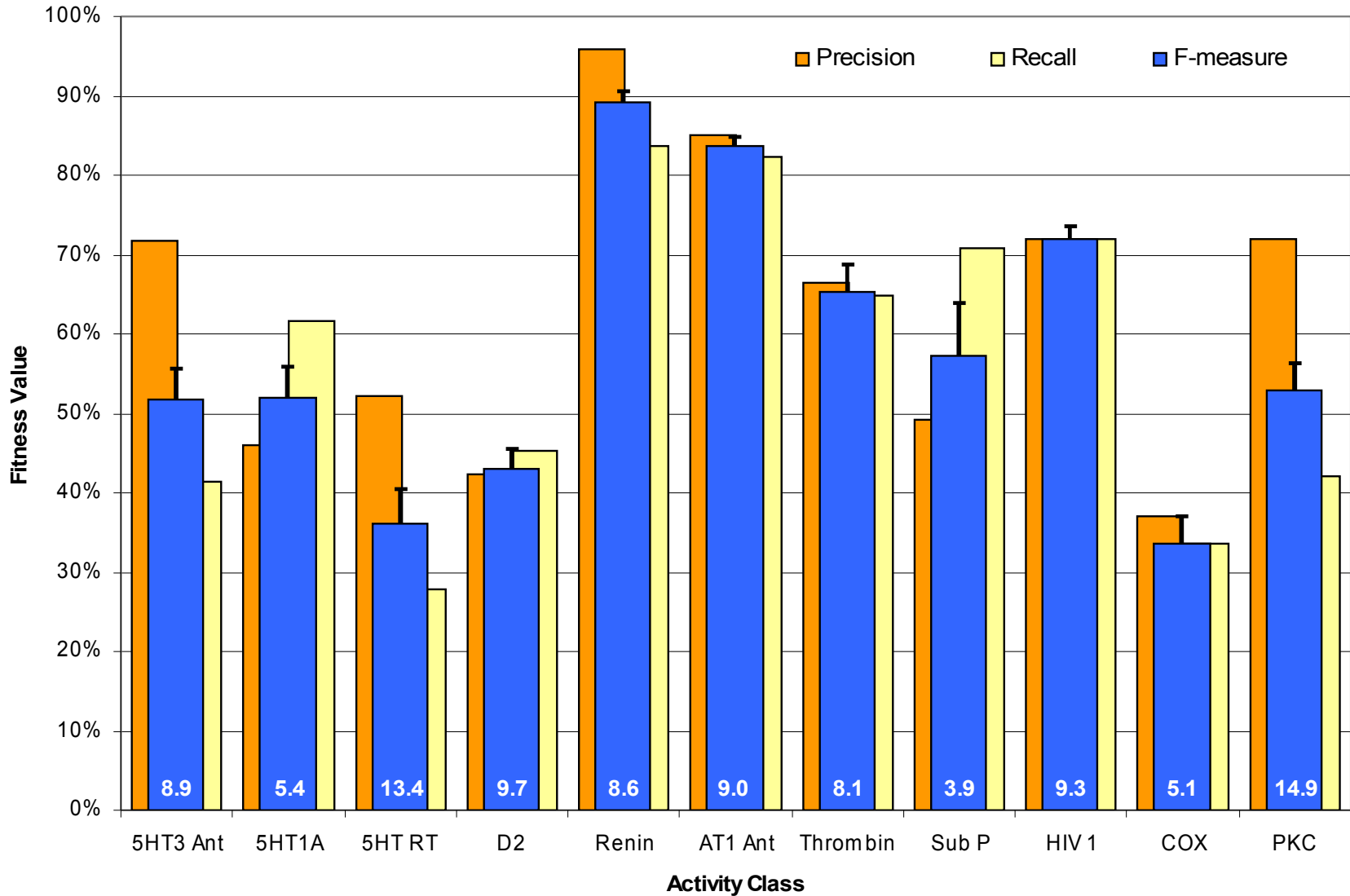
A = Actives **I** = Inactives
F = False **T** = True

		Predicted Class	
		Active	Inactive
Actual Class	Active	TA	FI
	Inactive	FA	TI

- $Recall(R) = TA / (TA + FI)$
 - Proportion of the total number of dataset actives retrieved
- $Precision(P) = TA / (TA + FA)$
 - Proportion of retrieved molecules that are actually active
- $F\text{-measure} = 2PR / (P + R)$
 - Equally weighted “average” of precision and recall

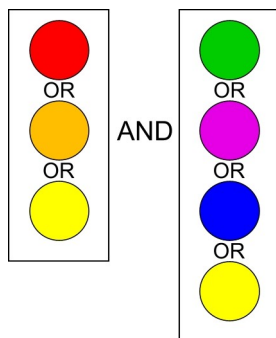
- 11 MDDR activity classes*
 - Divide each class equally into training, test and validation sets and mix each subset with 3000 “inactives”
- Objective: maximise f-measure fitness of queries
 - Train to convergence and select best query using test set
 - Prevents “overtraining” leading to non-generalising queries
 - Quoted fitness judged using validation set
 - Independent measure, i.e. not used in derivation or selection
 - Use 3 different partitions of the actives with 3 runs of the GP on each – 9 queries for each activity class

MDDR Results

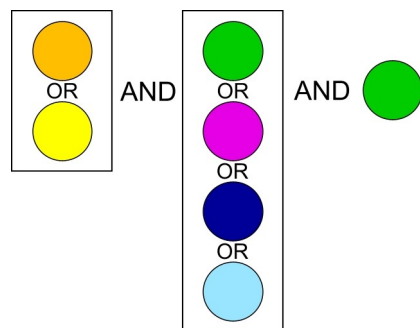


- IC50 data (5727 mols) for developability assay
- Continuous to binary data conversion
 - 3 activity cut-offs
 - low (>4.3) 1684 mols, med (>5) 937 mols, high (>6) 269 mols
 - “Inactives” all those below activity cut-off!
 - E.g. Medium “inactives” incorporate 747 low “actives”
- Train, test and validation sets, with 5 GP runs

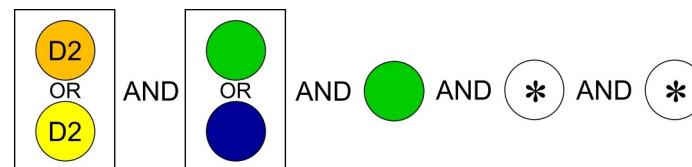
Low Activity Cut-off Query



Medium Activity Cut-off Query



High Activity Cut-off Query



● = Aromatic +ve ionisable
● = Aliphatic +ve ionisable
● = Acyclic +ve ionisable

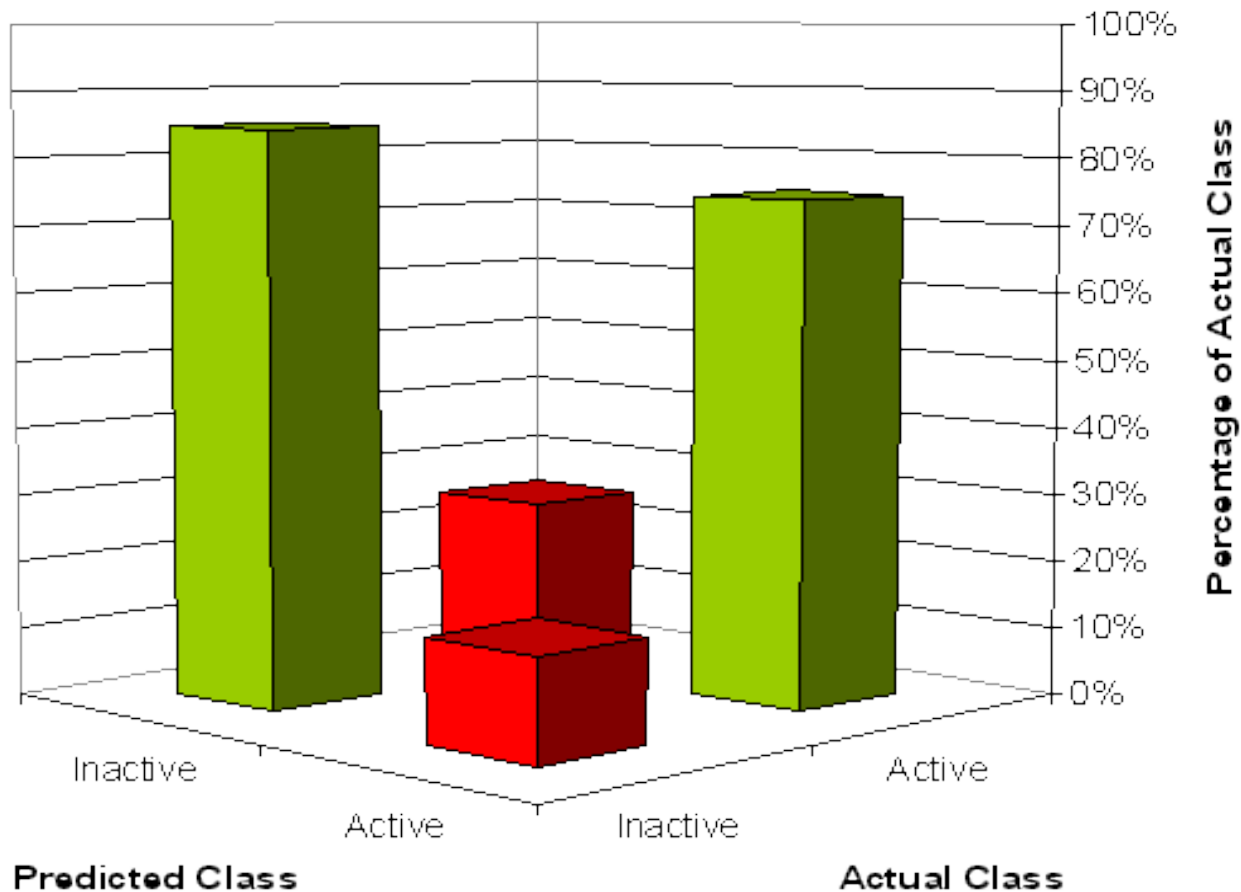
● = Aromatic donor
● = Aliphatic donor
● = Acyclic donor

● = Aromatic ring (no feature)
● = Aromatic donor/acceptor
● = Aromatic acceptor

- Some noise in OR lists— minimise
 - Computationally: Keep simpler but equally fit queries
 - Manually: Identify key features across queries from repeat runs
- SAR apparent even in raw queries & activity sensitive!



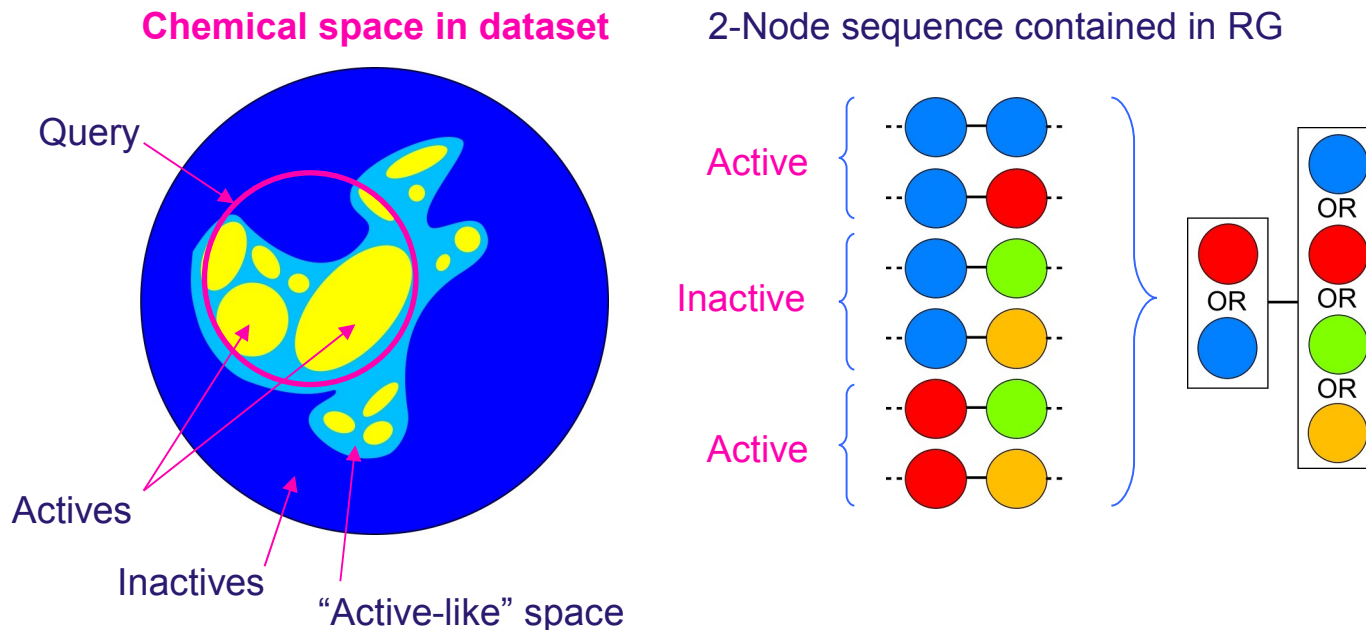
Medium Activity Query Performance



Precision = 52% **F-measure = 61%**
Recall = 75% **Enrichment = 2.9**

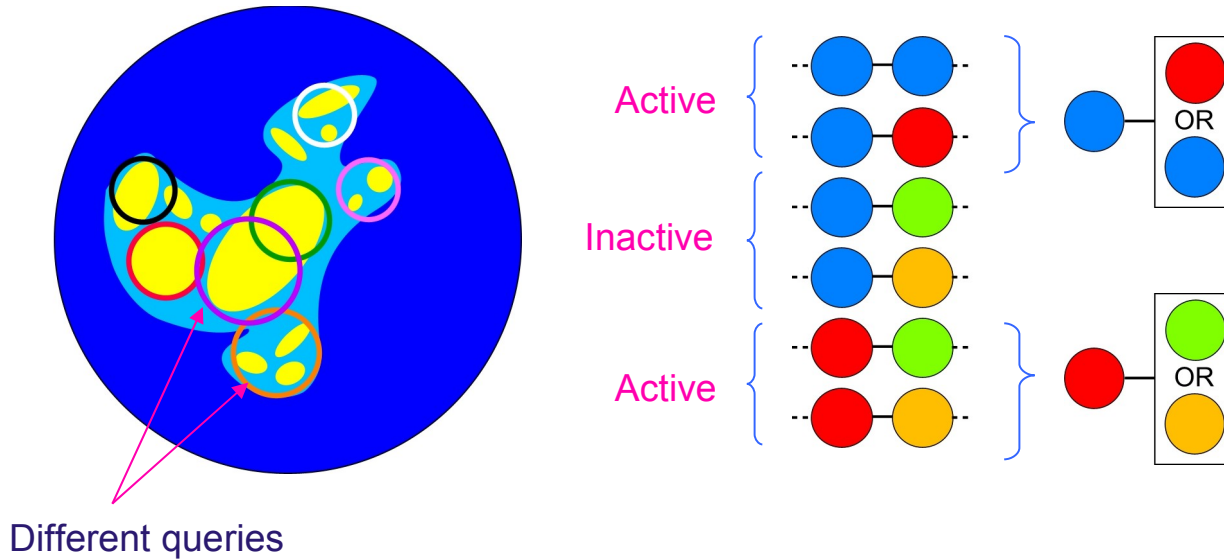
Multi-subclass complexity problem

- Several different types of RG within an activity class
 - High structural variability, different binding modes?
- Queries have limited flexibility



Multi-Query Approach

- Combine results from two or more queries
 - RG retrieved if found by query A **or** query B ...
 - Potential increases in recall, precision and diversity



➤ Queries must be complementary

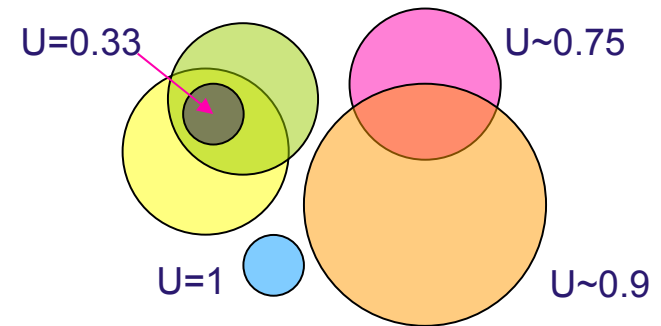
- Uniqueness score related to this

$$Uniqueness(Q) = \left(\sum_{i=0}^{i=a} \frac{1}{q} \right) / a$$

- Higher uniqueness for queries retrieving actives found by few (or no) other queries

➤ Queries must be specific

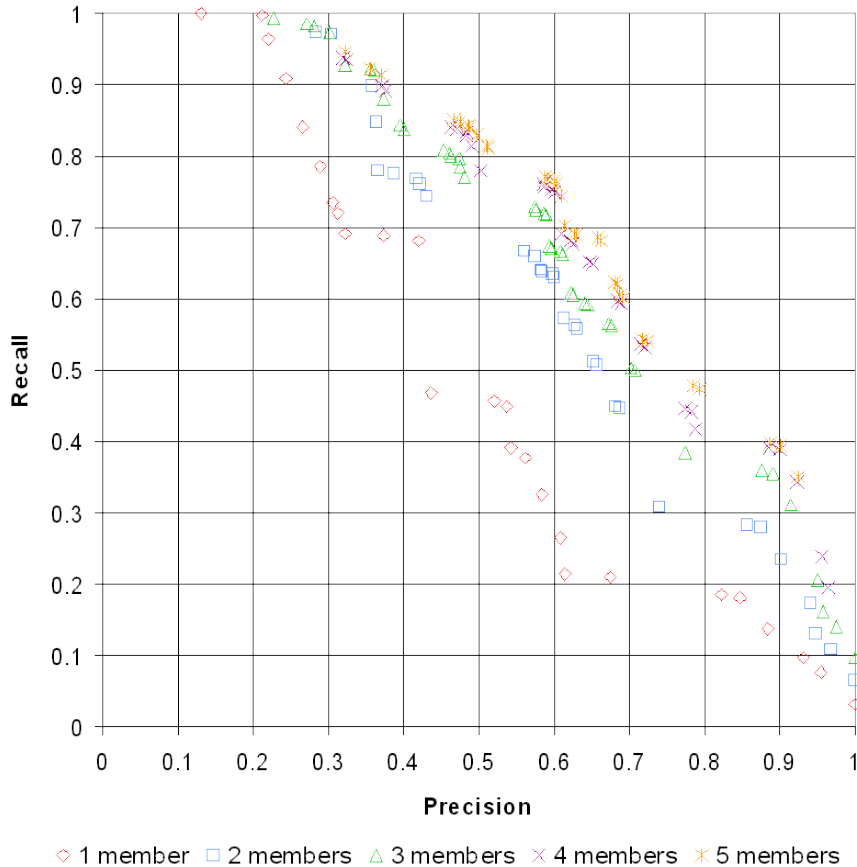
- To prevent large numbers of false actives resulting from the combination of several queries
- Typically increased specificity necessitates lower recall
 - Will tend not to evolve when using a single fitness objective!



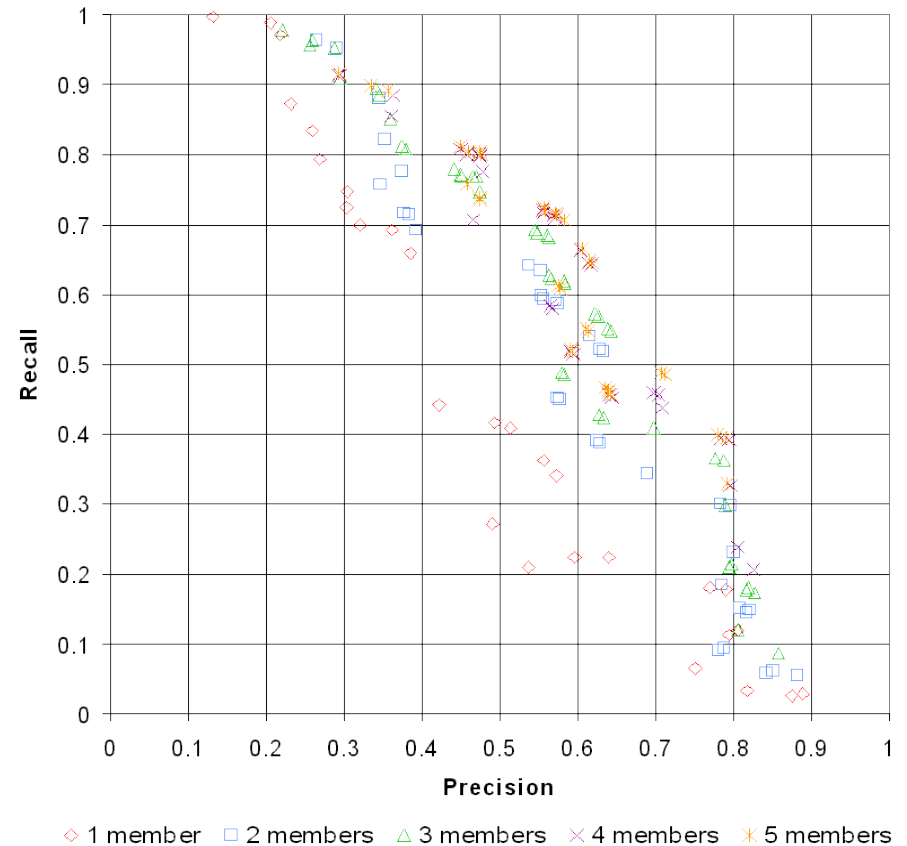
Multi-Objective GP

- Recall (R), Precision (P) and Uniqueness (U) objectives
 - Low R/high P queries equally valid as high R/low P queries
 - Uniqueness score encourages complementary queries
- Post-GP, non-dominated (best) queries combined iteratively
 1. Attempt all pairwise combinations & keep non-dominated
 2. Attempt to add third query to existing non-dominated pairs & keep non-dominated three-member solutions etc...
 - Stop when addition of queries no longer improves fitness, or at a defined number of members per solution

Iterative team formation on 5HT1A queries from 3 GP runs



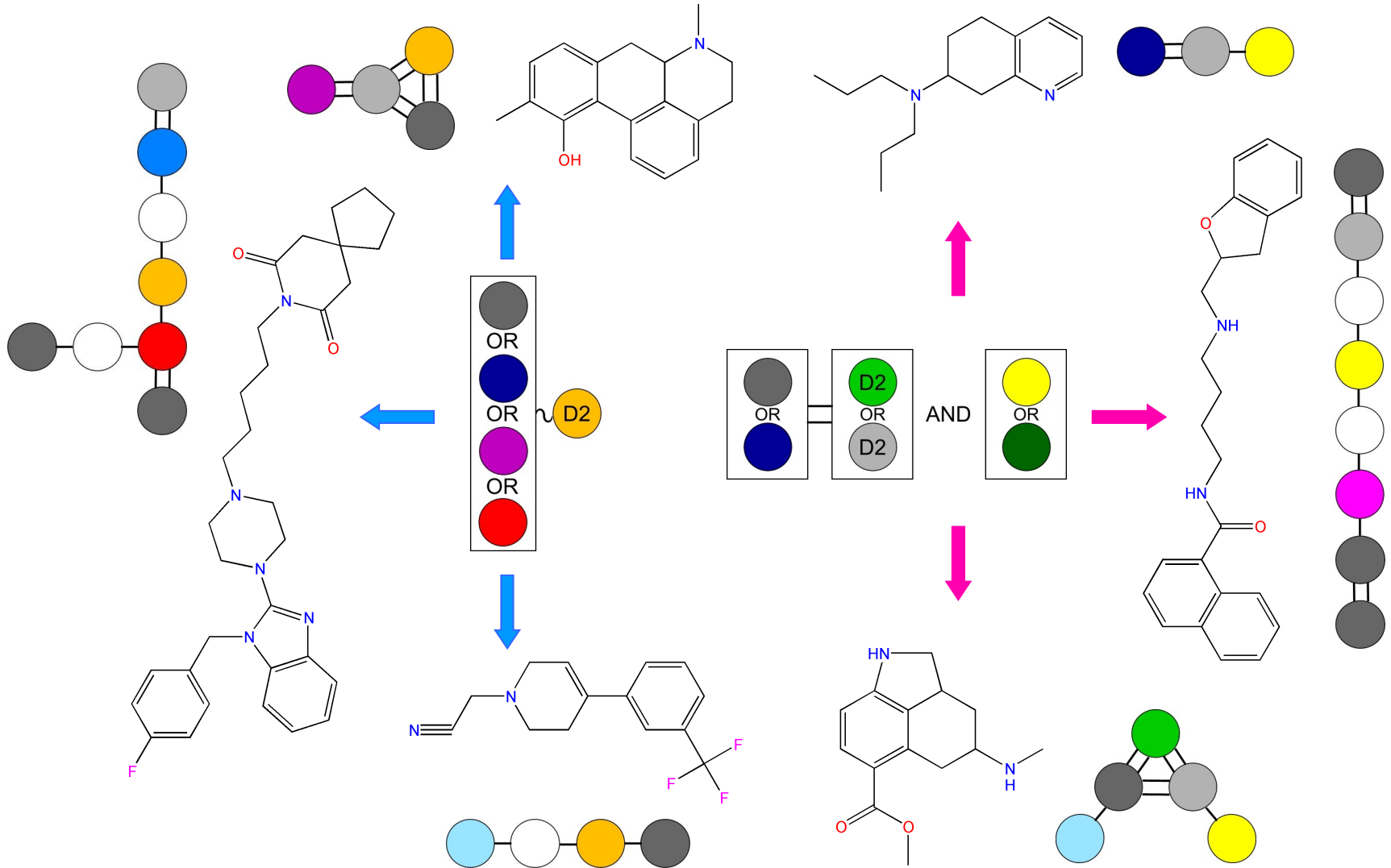
Test set results



Validation set results



2-member Solution Results





The University Of Sheffield.



**Stephen
Pickett**



**Gavin
Harper**

**Val
Gillet**

