The University Of Sheffield.

Diversity Analysis and Library Design

Val Gillet
Department of Information Studies,
University of Sheffield, UK

Outline

- Diversity Analysis
  - Measuring diversity
  - Selecting diverse subsets
  - Computational filtering
- Combinatorial Library Design
  - Designing libraries optimised on multiple properties
- Reduced Graphs as Molecular Descriptors

High-Throughput Technologies

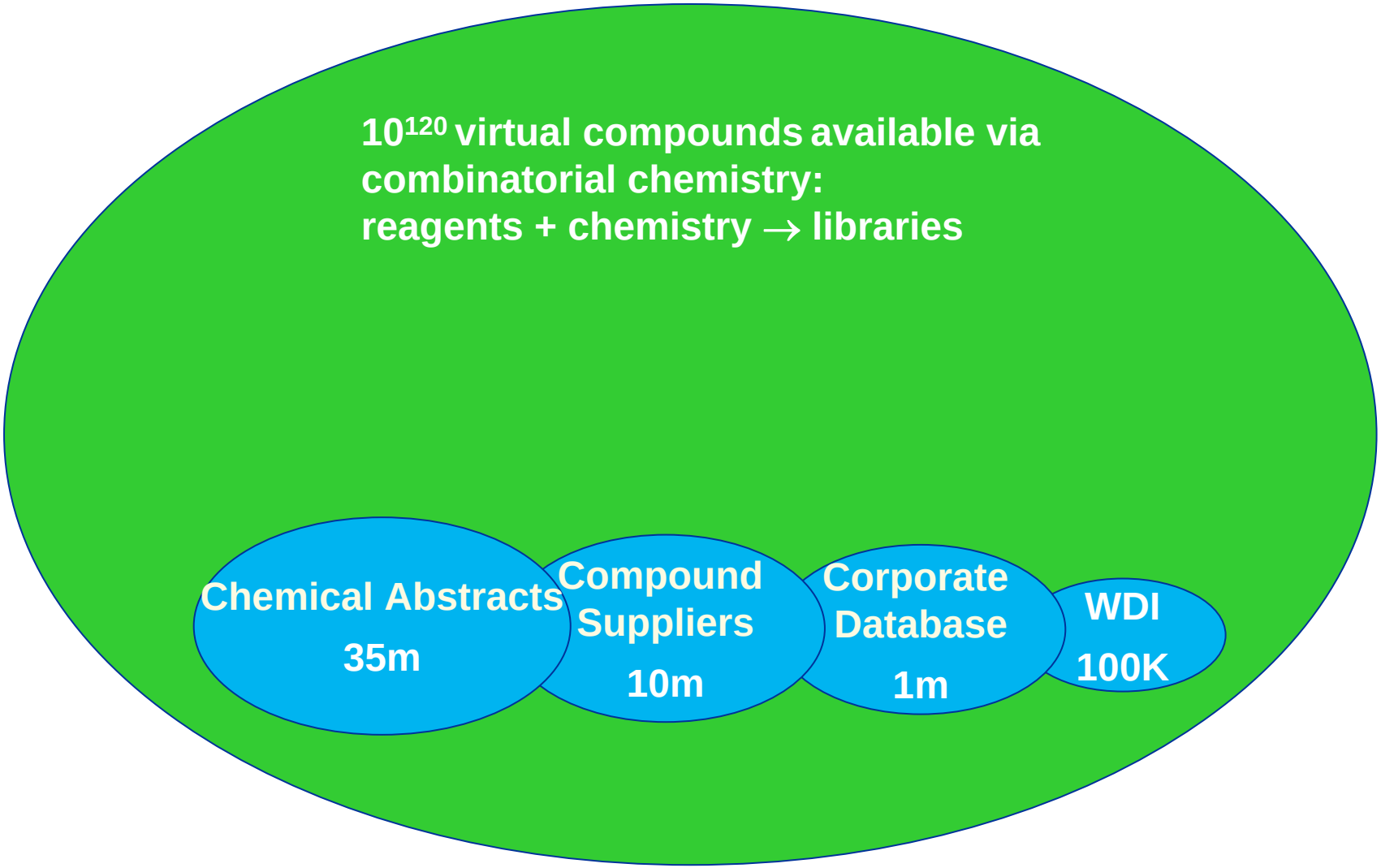
- High-throughput screening has massively increased the rate at which compounds can be tested for activity
- Combinatorial synthesis allows parallel synthesis of large numbers of compounds
- Have these increased numbers resulted in more hits?

The Combinatorial Explosion

- The Available Chemicals Directory (www.mdli.com) contains the following:
  - 85000 carboxylic acids
  - 44000 primary alkyl amines
  - 12000 aldehydes
  - 2000 fmoc amino acids

- So we could make…
  - 3.7×10^9 monoamides (acid + amine)
  - 5.3×10^8 secondary amines (reductive amination)
  - 7.5×10^{12} diamides (fmoc amino acid + amine + acid)

The Chemical Universe



Number of seconds since big bang = ca. 10^{17}

Increasing Throughput is not Enough!

- Chemical space is huge!
- Success rates of large libraries have been low
  - Insoluble; high molecular weight; too flexible; etc; etc
  - Costs are high: $1 per compound = $1 million for library of 10^6 compounds
- Assay does not always permit HTS
- Despite initial enthusiasm for large numbers the current trend is towards smaller carefully designed libraries

Virtual Screening

- Virtual screening
  - In-silico prioritisation of compounds
- Virtual screening can be used to
  - Select compounds for screening from in-house databases
  - Choose compounds to purchase from external suppliers
  - Design combinatorial libraries

- The technique applied depends on the aim and on the knowledge available, for example, about the particular disease target
  - Usually there are multiple criteria to consider

Virtual Screening: Focused Libraries

- Targeted/focused libraries
  - Selection of compounds that are similar to a lead compound or that fit a QSAR
  - Selection of compounds focused on a single therapeutic target using structure-based drug design
  - Selection of compounds focused on a family of related targets

- Even with focused libraries still need to have some diversity

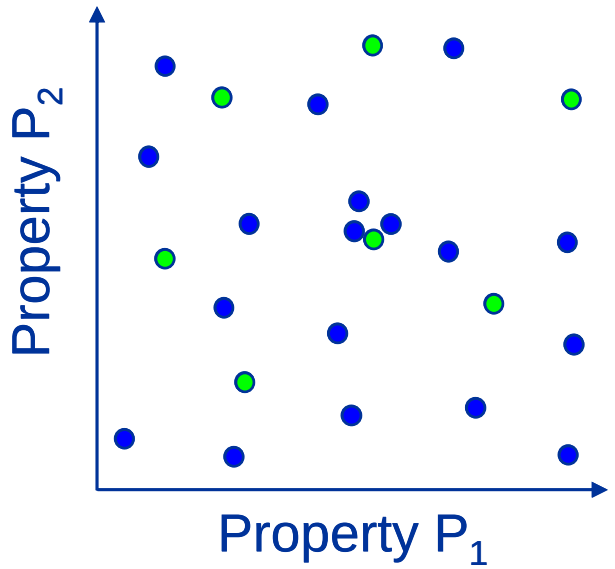
Virtual Screening: Diversity

- Lead generation
  - Selection of compounds when little is known about a particular target
  - Selection of compounds for screening against several targets
- Compound acquisition
  - Selection of compounds to purchase to augment an existing collection

Similarity and Diversity

- Similar Property Principle
  - Structurally similar compounds tend to exhibit similar properties

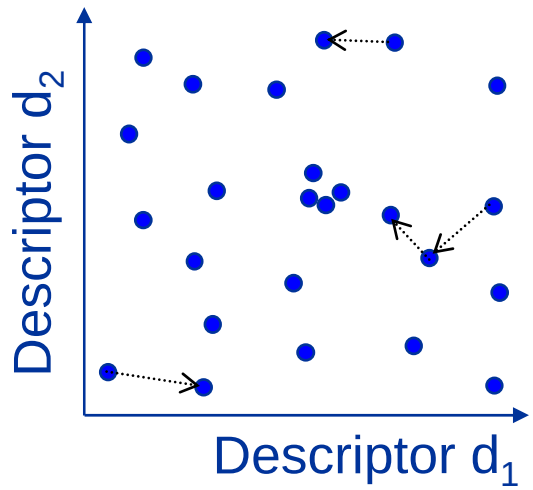
- Similarity
  - If we have a known active (literature, competitor compound etc) then compounds that are similar to it are likely to show similar activity
- Diversity
  - A diverse subset of compounds should maximise the coverage of biological activity and minimise redundancy



Similarity is the property of a **pair** of compounds
Diversity is the property of a **library** of compounds

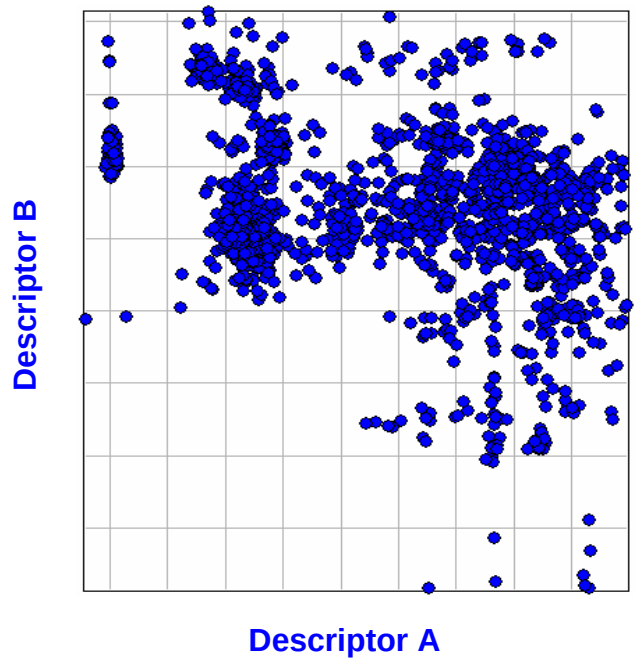
Measuring the Diversity of a Compound Library: 1

- Calculating (dis)similarities
  - Dissimilarity = 1 – Similarity
  - Euclidean distance
- Requires molecular descriptors and similarity coefficient
  - Whole molecule properties; 2D fingerprints; 3D pharmacophores
  - Tanimoto coefficient
- Example diversity measures
  - Sum of pairwise (dis)similarities/distances
  - Average NN dissimilarity/distance



Measuring the Diversity of a Compound Library: 2

- Coverage of a pre-defined chemistry-space
- Requires definition of a low (1D to 6D) dimensional chemistry-space
  - Physicochemical properties
  - BCUTs
  - pharmacophore coverage
  - scaffold coverage (number of unique scaffolds)



Selecting a Diverse Subset of Compounds

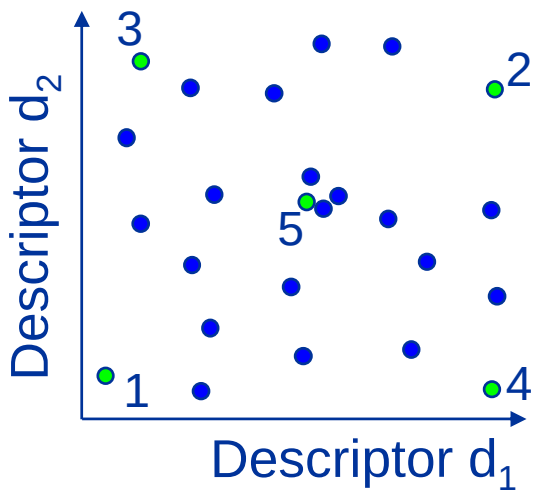
- Selecting a subset of size *n* from dataset of size *N* requires evaluation of $\frac{N!}{n!(N-n)!}$ subsets

  There are ~2x10^{13} ways of selecting 10 cmpds from 100!

- Computationally efficient methods are required
  - Distance-based methods
    - Dissimilarity-based compound selection; sphere exclusion; clustering
  - Coverage-methods
    - Partitioning-schemes; optimisation-methods (maximise or minimise diversity measure)

DBCS: General Algorithm

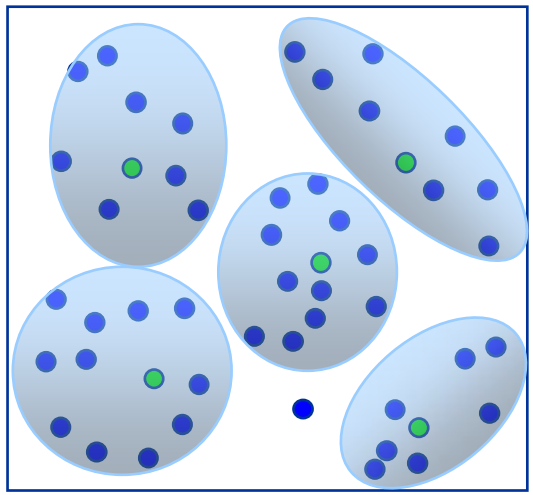
1. Select a compound and place in subset (random, centroid, most diverse)
2. Calculate dissimilarity between each remaining compound and compounds in subset
3. Choose next compound that is most dissimilar to compounds in subset (MaxMin, MaxSum)
4. If less than n compounds in subset, return to 2

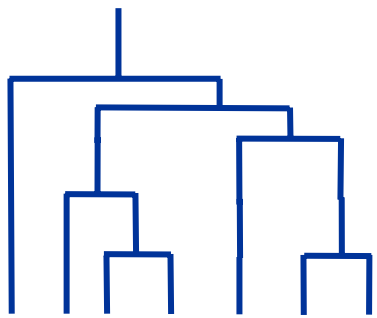


- Characteristics
  - fast enough to be applied to large dataset
  - based on (dis)similarities therefore can be used with high dimensionality data *e.g.* fingerprints
  - have a tendency to select outliers

Clustering

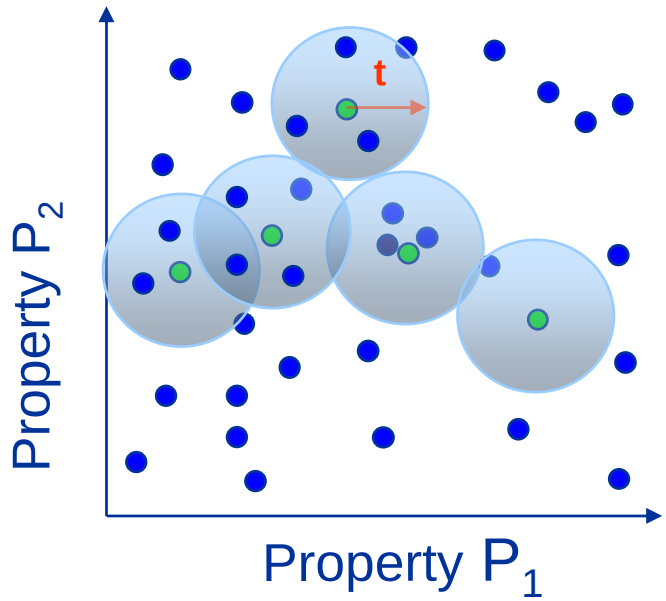
- Group molecules
  - molecules within a cluster are similar
  - molecules from different clusters are dissimilar
- Choose one or more from each cluster
- Characteristics
  - good for high dimensionality data (fingerprints)
  - reveals natural clustering in dataset
  - limited to small(ish) datasets
  - difficult to add new compounds





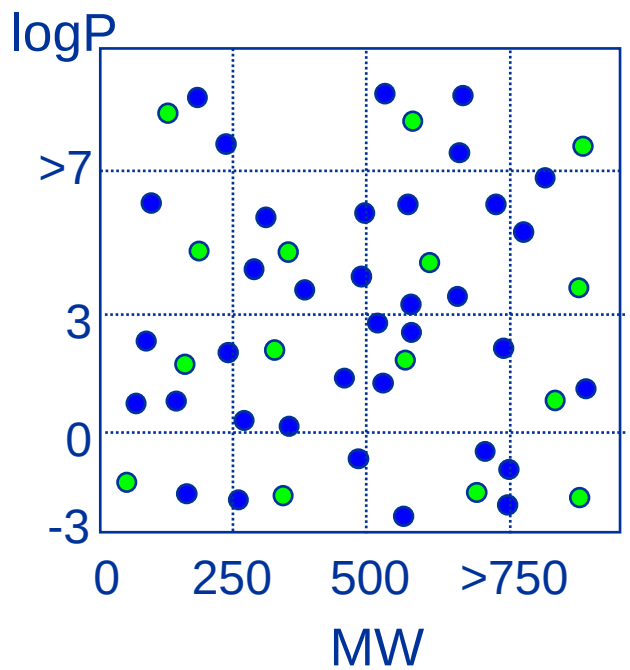
Sphere Exclusion

1. Define a threshold similarity t
2. Select a compound and place in subset
3. Remove all compounds with dissimilarity < t
4. If compounds left in data set, return to 2

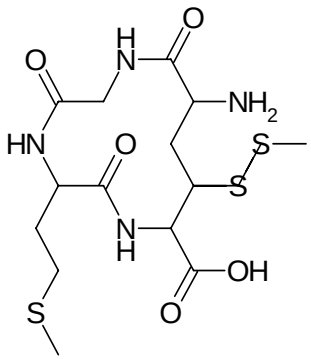


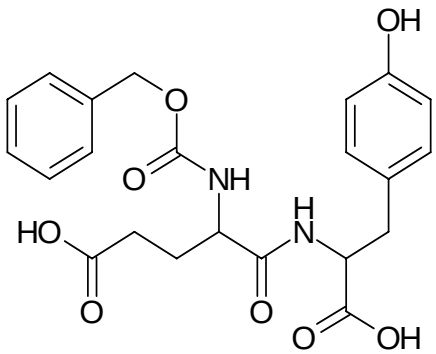
Partitioning/Cell-Based

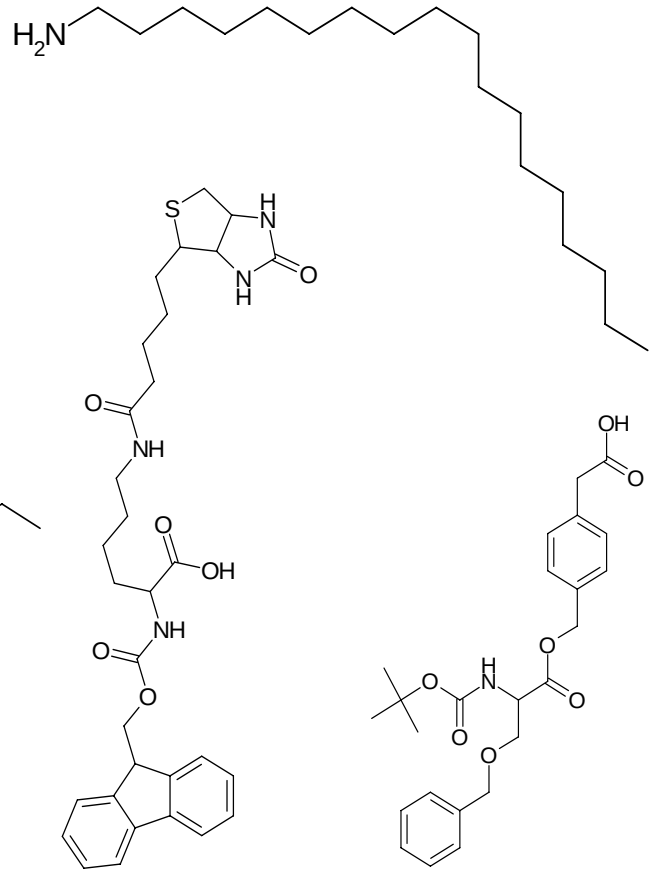
- Define a low dimensional space
  - e.g. physicochemical properties (logP, MW,…); BCUT descriptors
  - assign each compound to a cell
  - choose one or more from each cell
- Characteristics
  - fast but restricted to low dimensional descriptors
  - diversity voids are easily identified
  - easy to add new compounds
  - cell boundaries are arbitrary

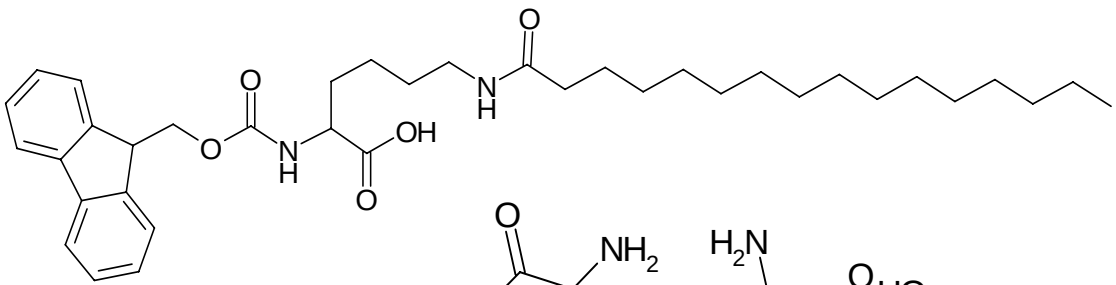


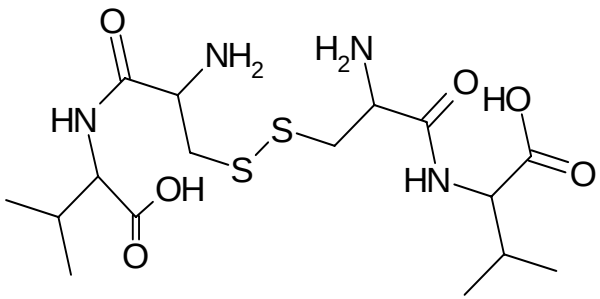
A Diverse Set of Compounds!











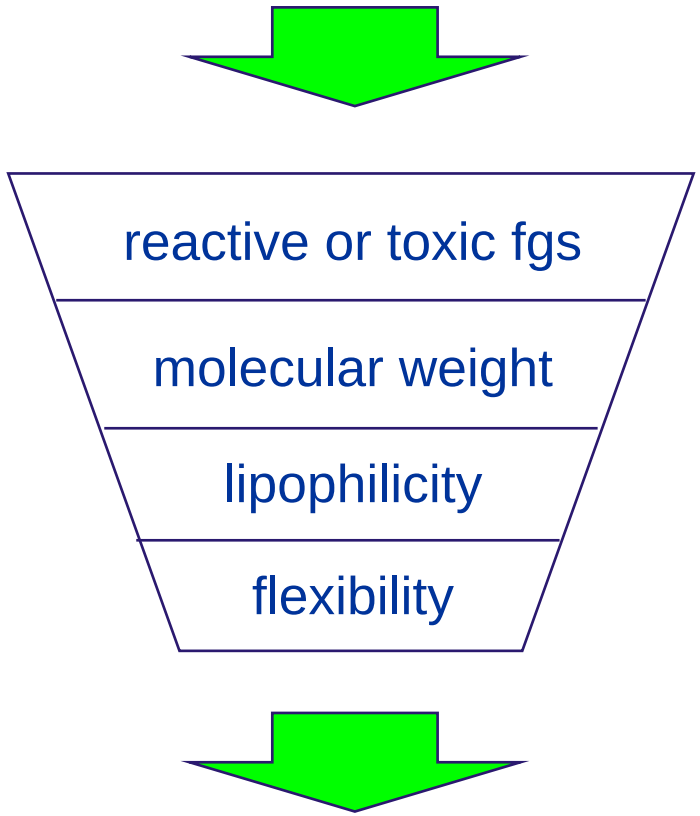
Drug-likeness

- The early approaches to designing HTS experiments based on large diverse sets of compounds gave disappointing hit rates
  - low numbers of hits
  - hits unattractive for lead optimisation:
    - Poor ADME properties - insoluble; lipophilic; too flexible; high molecular weight

- Whether designing diverse or focused libraries molecules should also be constrained to have "drug-like" physicochemical properties

Computational Filters

- Set of computational techniques to eliminate molecules that have inappropriate characteristics
- Reduce the number of compounds that need to perform calculations on
- “Badlist” of reactive or toxic substructures (cf “goodlist” of “privileged substructures”)



Drug-like and Lead-like

- Drug-likeness: eliminate compounds with non-drug-like physicochemical properties
  - Lipinski “Rule Of Five”, in which a molecule is assumed unlikely to be orally absorbed if at least two of the following conditions are met
    - MW>500, ClogP>5, HBD>5, HBA>10

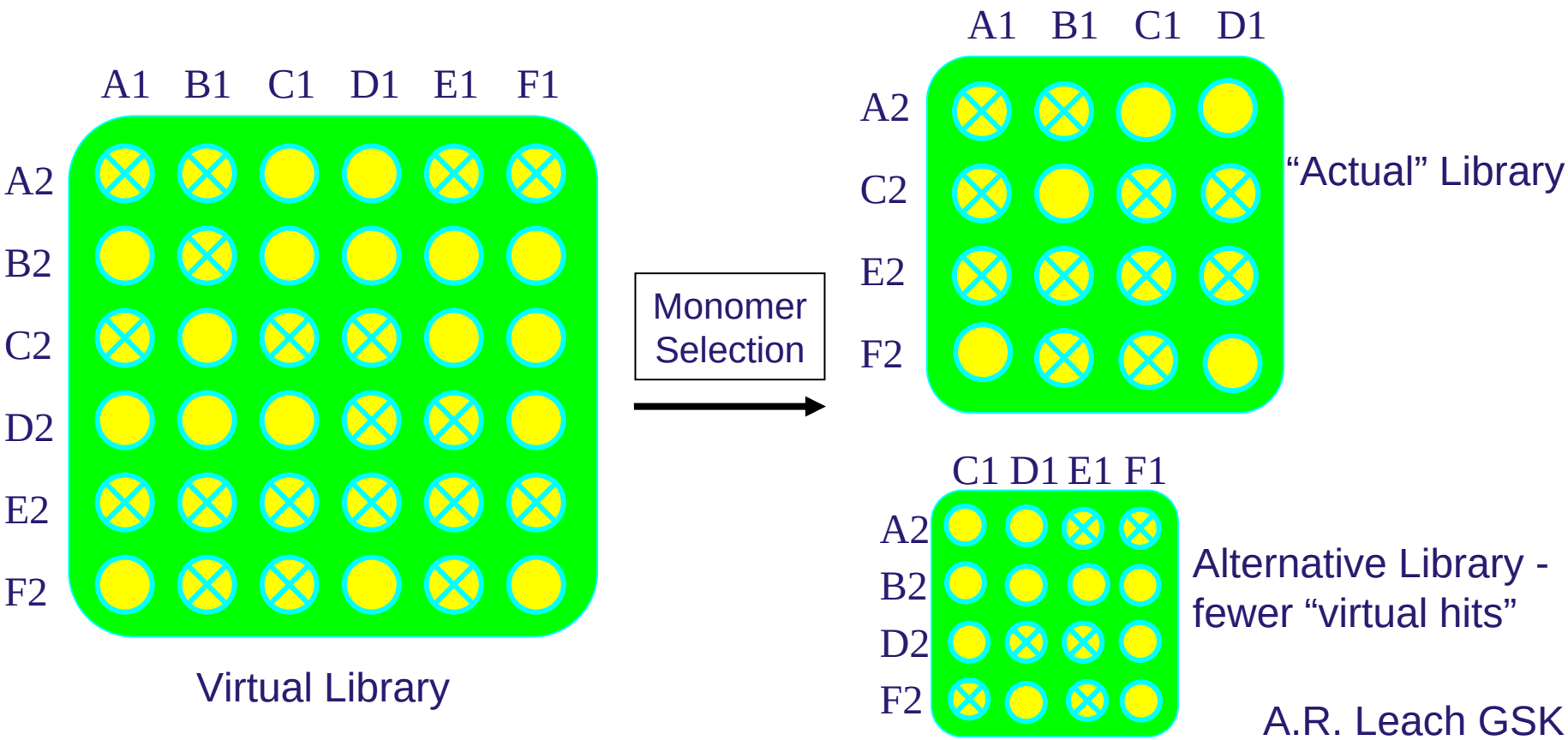
- Lead-likeness
  - “leads” tend to be smaller and less complex than “drugs”

The University Of Sheffield.

Combinatorial Library Design

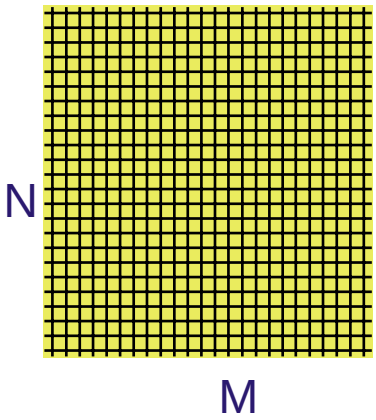
Building Block Selection

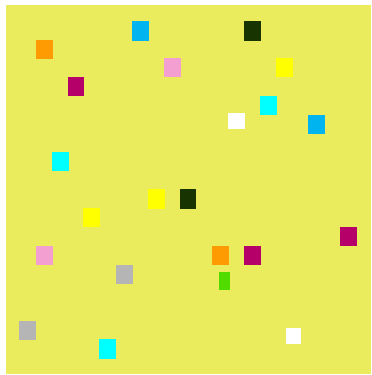
- Compromise: Select reagents for the “real” library that satisfy the layout but whose products perform well in the virtual screen



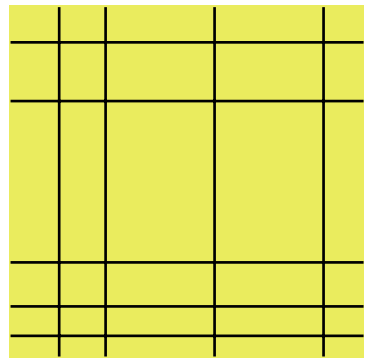
A.R. Leach GSK

Combinatorial chemistry imposes an important constraint on building block selection





Cherry pick for best subset



Select m × n for best combinatorial subset

There are 4×10^{54} ways to select a 36 × 36 library from 100 × 100 possible building blocks ($^{N}C_{n}$. $^{M}C_{m}$)

The optimal library cannot usually be derived by considering the reagents alone*: product-based library design*

Gillet et al J. Chem. Inf. Comput. Sci. (1997), 37(4), 731-740

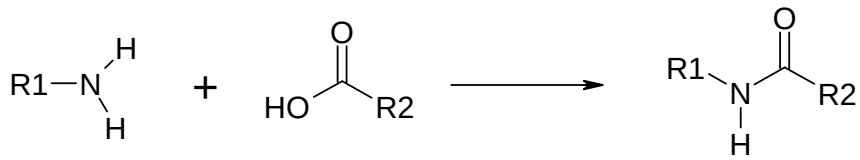
Combinatorial Library Design

- Library design is a multi-objective optimisation problem
  - Diverse or focused or both
  - Cheap, small, combinatorially efficient
  - Drug-like, good ADME properties,
- Many in-silico methods exist for calculating the various properties
- Applying computational filters sequentially can lead to sub-optimal designs
- How do we find a good balance in the objectives?

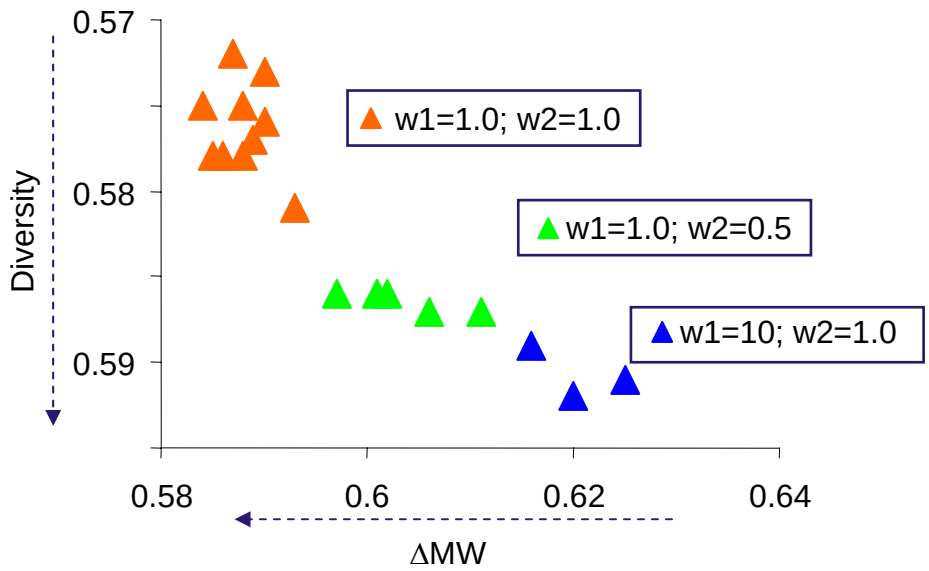
Weighted-Sum Approach

$$f(n) = w_1.diversity + w_2.\cost + w_3.property1 + w_4.property2 + ...$$

- Limitations
  - Setting of weights is difficult especially for different types of objectives
  - The objectives are often in competition
  - A single compromise solution is found when usually a family of alternative solutions exist that are all equivalent (trade-offs)



$$f(n) = w_1.diversity + w_2.\Delta MW$$

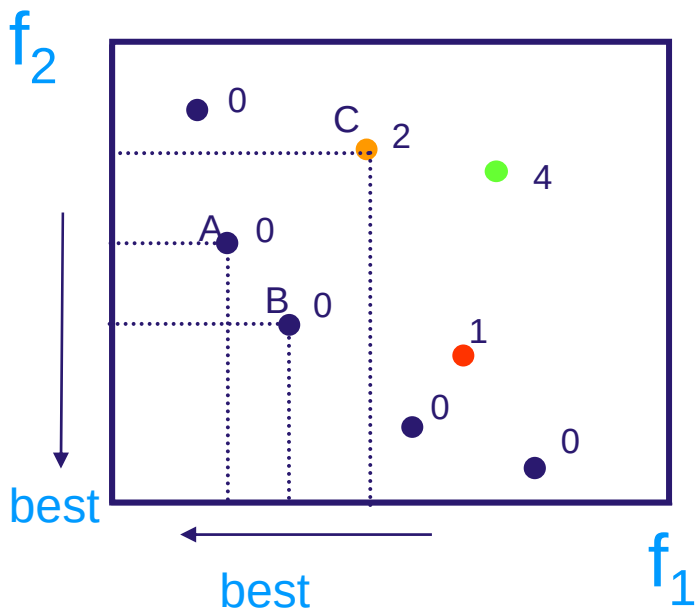


Multiobjective Optimisation using a MOEA

- Multiple objectives and handled independently
- Pareto optimality is used to explore the search space
- Multiple equivalent solutions are explored in parallel (exploiting the population nature of an EA)
- MOEA for combinatorial library design
  - Combinatorial subsets selected from virtual library of possible products
  - The objectives can include any property that can be calculated for a library of compounds,
    - Chemical properties: e.g. diversity; drug-like profiles; in-silico ADME properties
    - Physical properties: size, configuration, number of subsets, cost

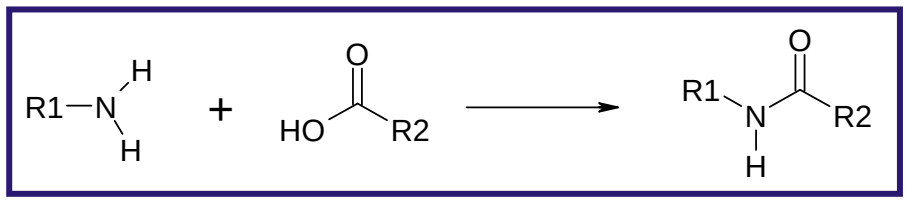
Pareto Ranking

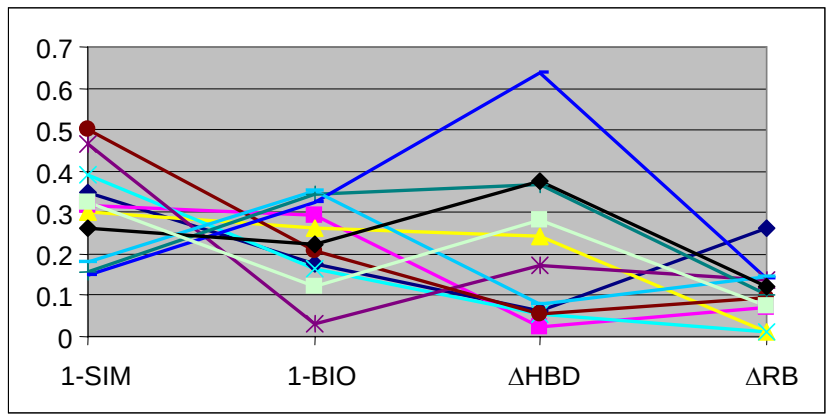
- Each objective is optimised independently
- One solution dominates another if it is better in both objectives
- Solutions are ranked according to dominance value
- Solutions where no other solutions are greater in all objectives are non-dominated and form the Pareto frontier



Combinatorial Library Design: Focused Libraries

- 100 × 100 virtual library
  - 10^4 potential products
- MOEA used to design 10 × 10 subsets
- Objectives
  - Similarity to a target
    - •Sum of similarities using Daylight fingerprints
  - Predicted bioavailability
    - •Each compound rated from 1 to 4
    - •Sum of ratings
  - Hydrogen bond donor profile
  - Rotatable bond profile

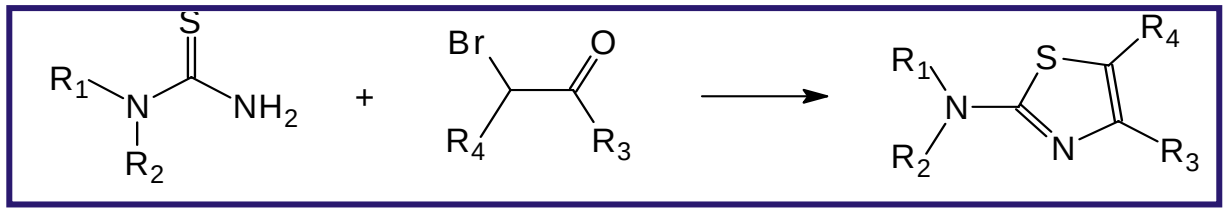




Combinatorial Library Design: Diverse Libraries

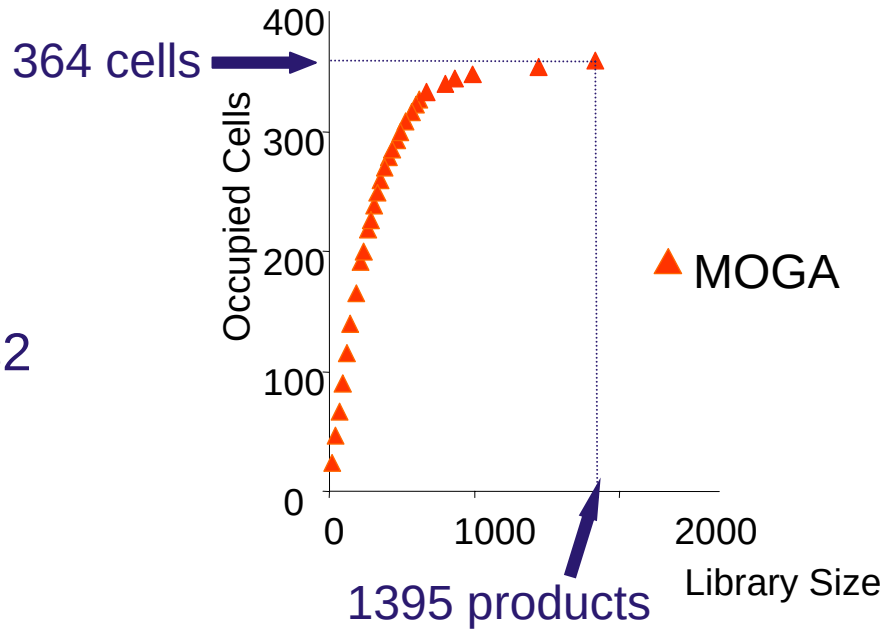
Exploring the trade-off in diversity and library size

Aminothiazole Library



Virtual library of 12850 products
170 thioureas × 74 α-bromoketones

Occupies 364 of 1134 cells (Cerius2 topological and physicochemical descriptors followed by PCA)



Incorporating Constraints

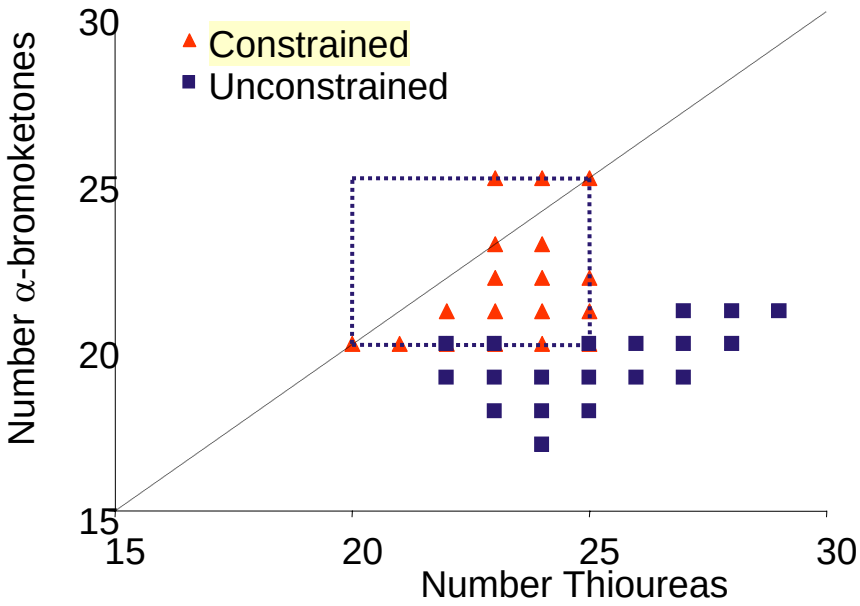
- Practical limits are often imposed on library design so that restricted regions of the search space are of interest
- Constraints can be applied to restrict the search to pertinent regions
  - Library size
  - Combinatorial efficiency
    - Number of reactants required to generate given number of products
    - 20×20; 40×10; 80×5; etc
  - Plate coverage

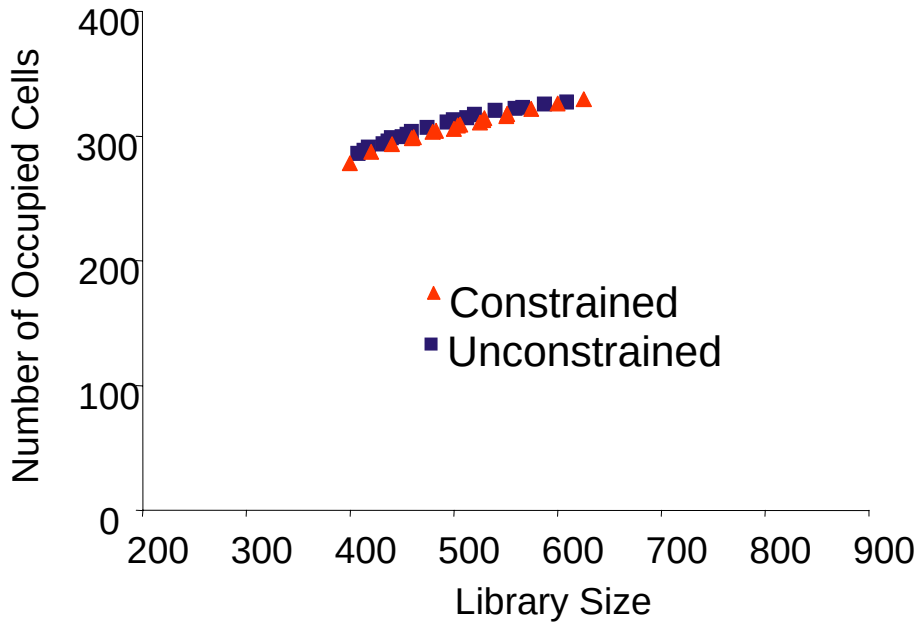
Combinatorial Efficiency

Size constraint: 400 to 600 products

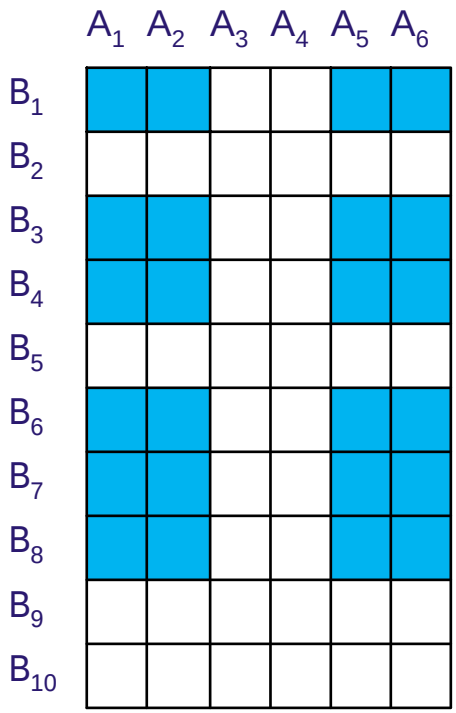
Combinatorial efficiency constraint: 20 <= α-bromoketones >= 25

20 <= thioureas >= 25

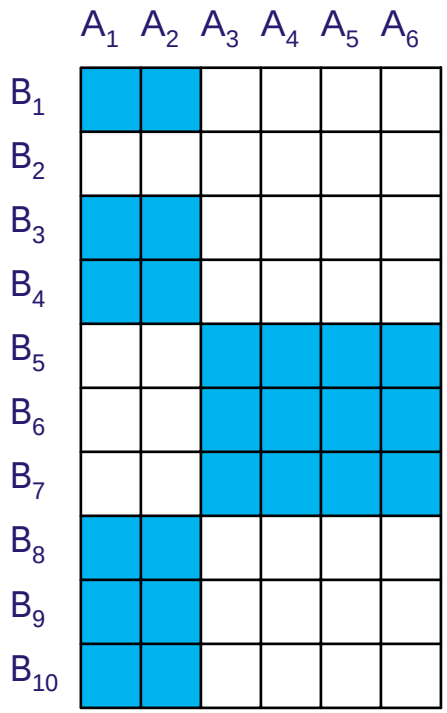




Selecting Multiple Combinatorial Subsets



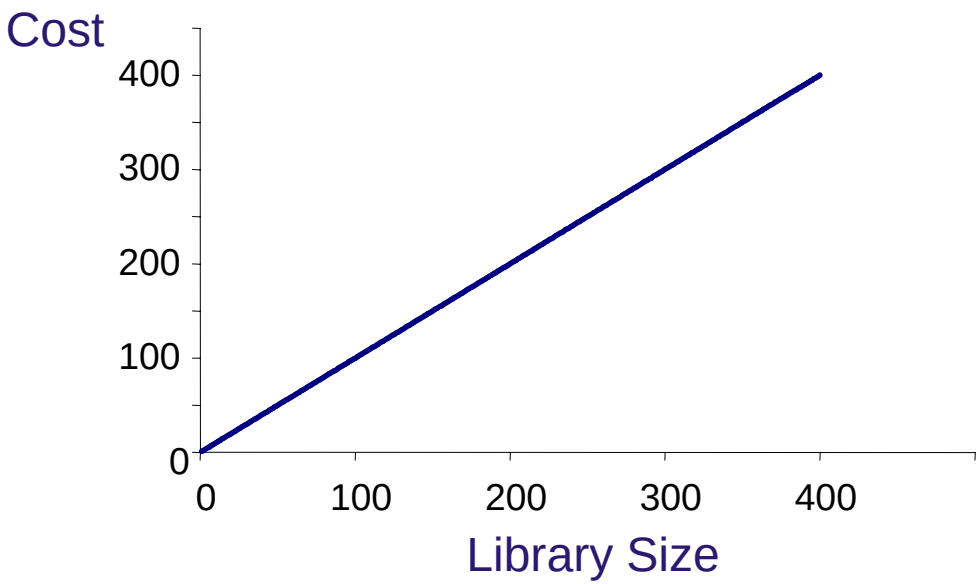
24 products constructed from one 4 × 6 subset



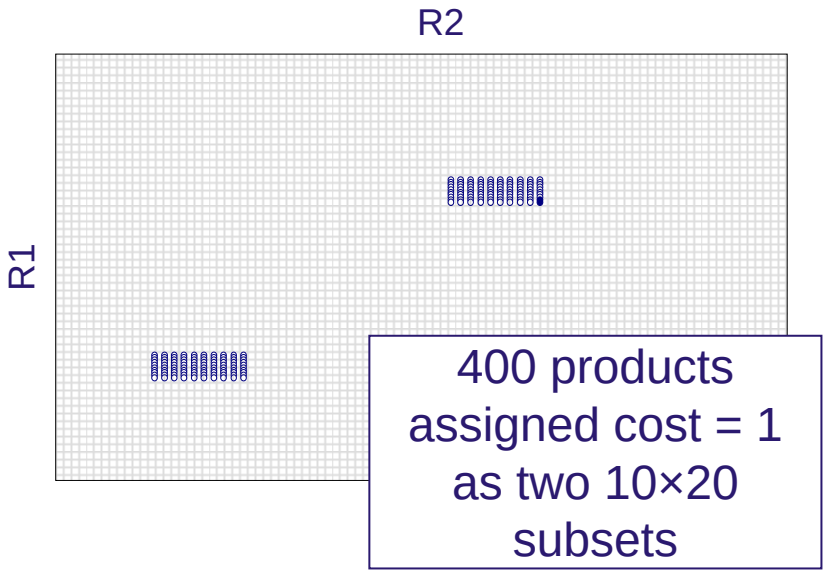
24 products constructed from two subsets: 2 × 6 and 4 × 3

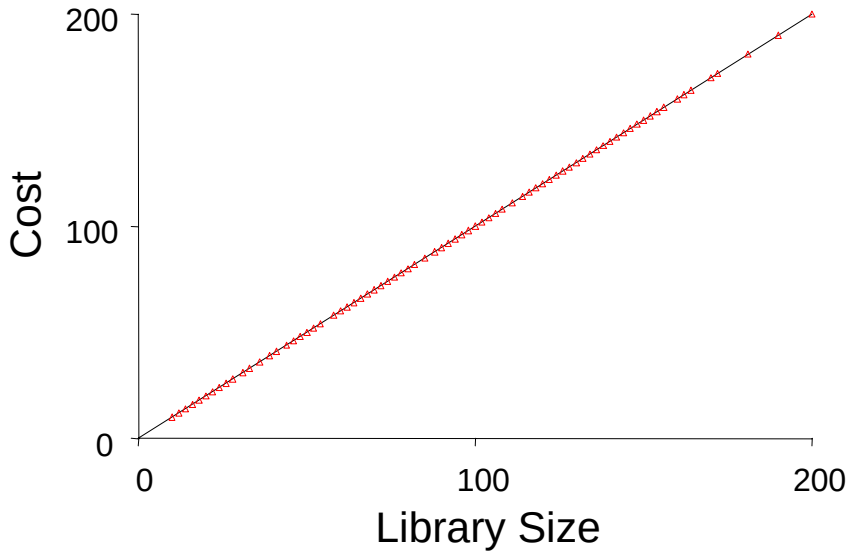
Design of Test Cases

- A cost value of 1 is assigned to various subsets of compounds in the virtual library, all other compounds assigned a cost of 0
- Design criteria
  - Identify libraries with maximum sum of cost and minimum size
  - Ideal solutions: sum of cost = library size



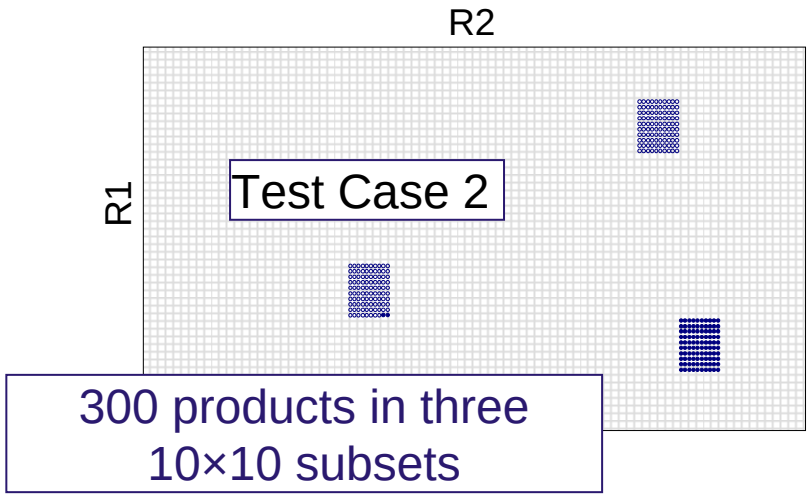
Test Case 1

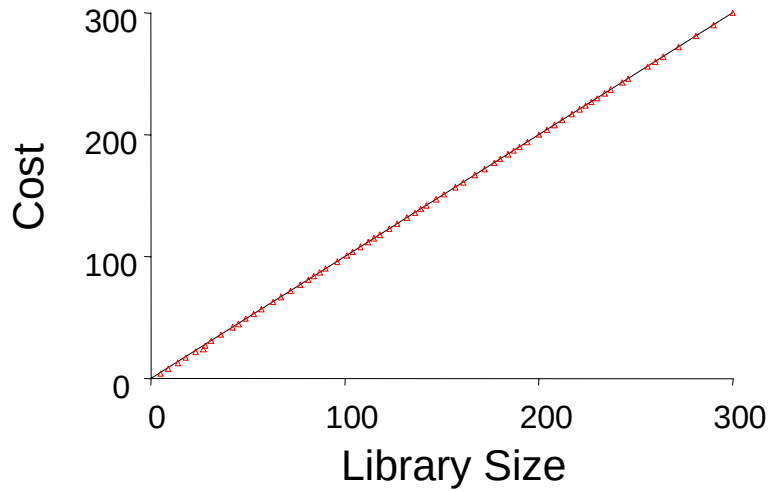


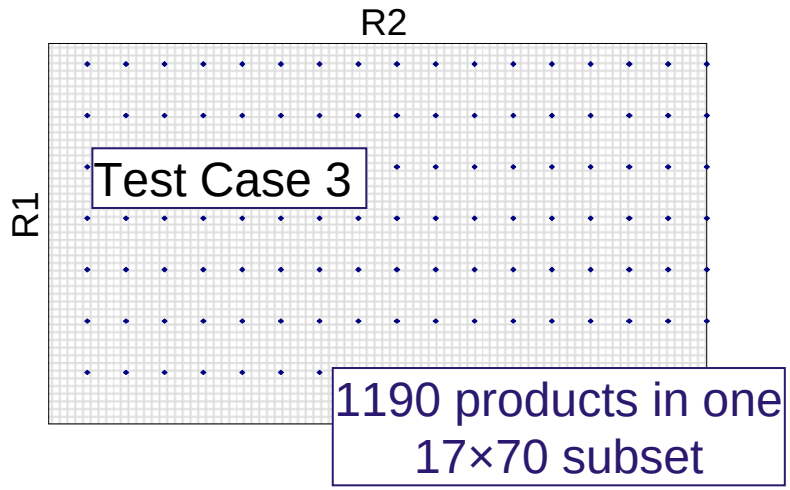


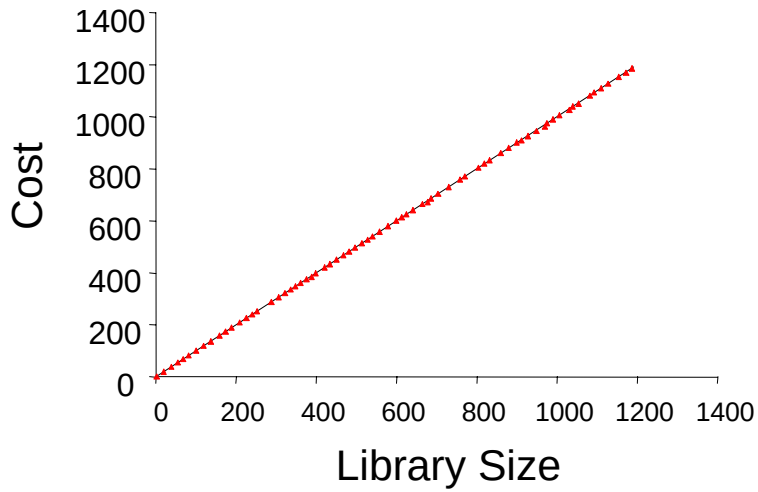
Subset 1 (R1 × R2)		Size	Subset 2 (R1 × R2)		Size	Total Size	Cost
2	6	12	10	4	40	52	52
8	5	40	10	6	60	100	100
10	10	100	10	10	100	200	200

Test Cases 2 and 3

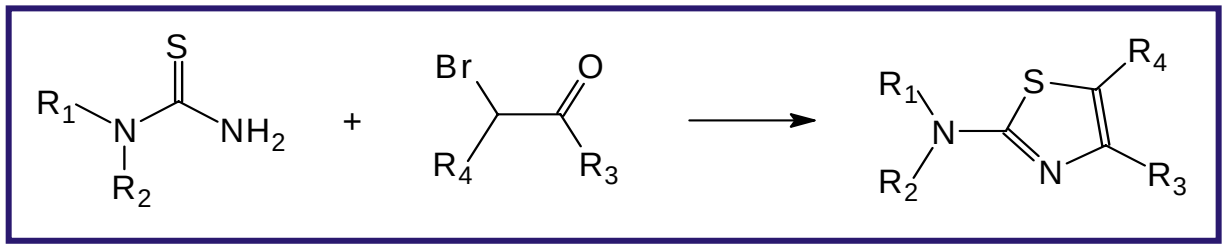


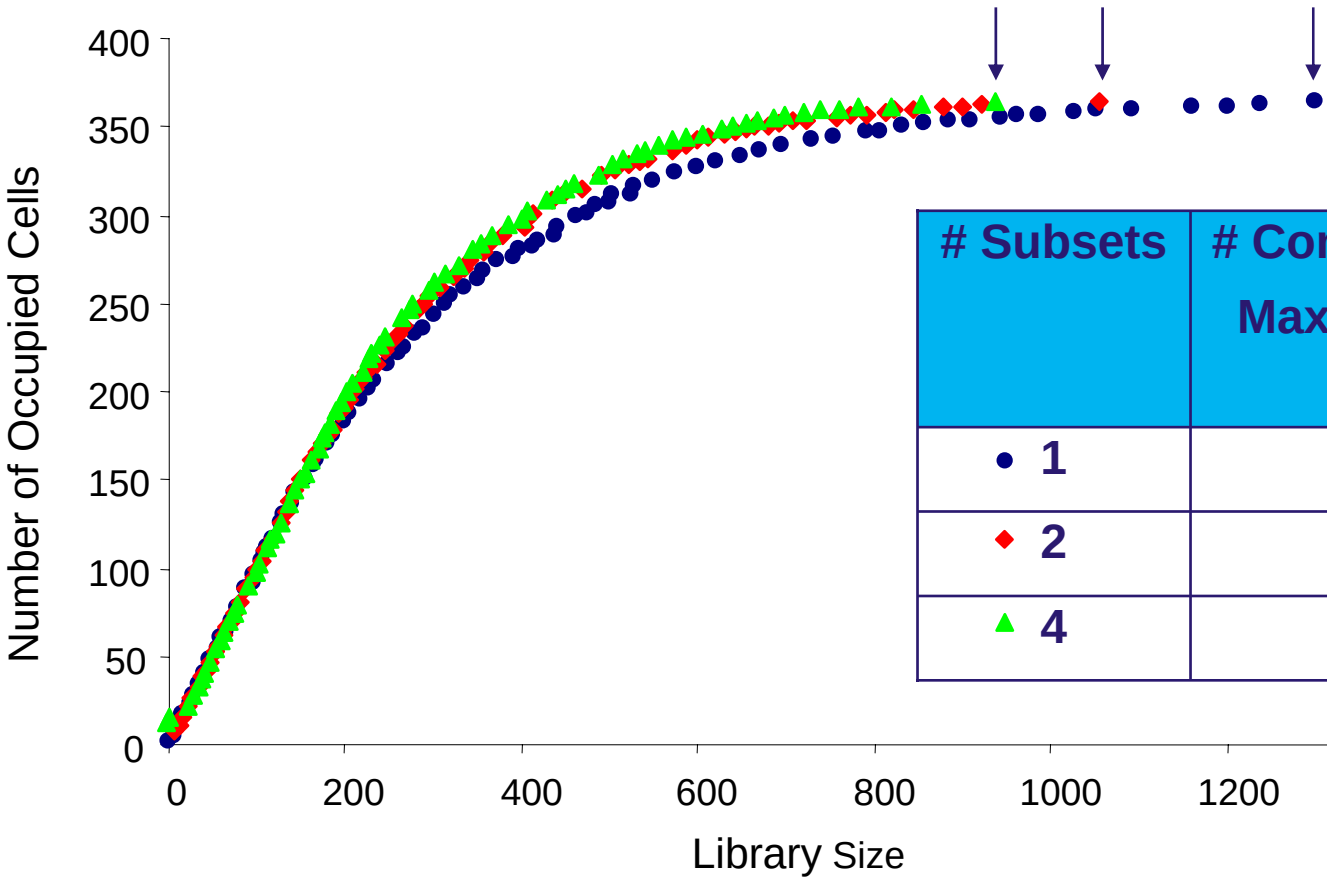






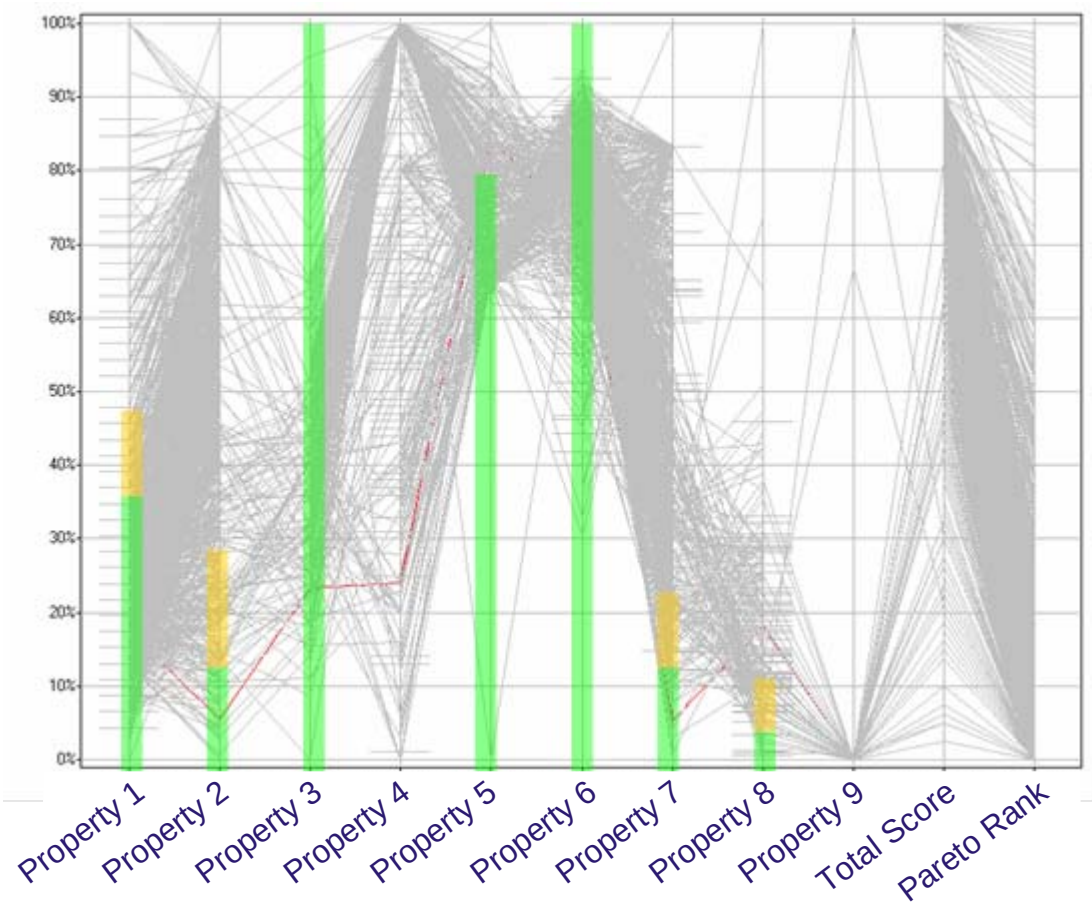
Aminothiazole Library





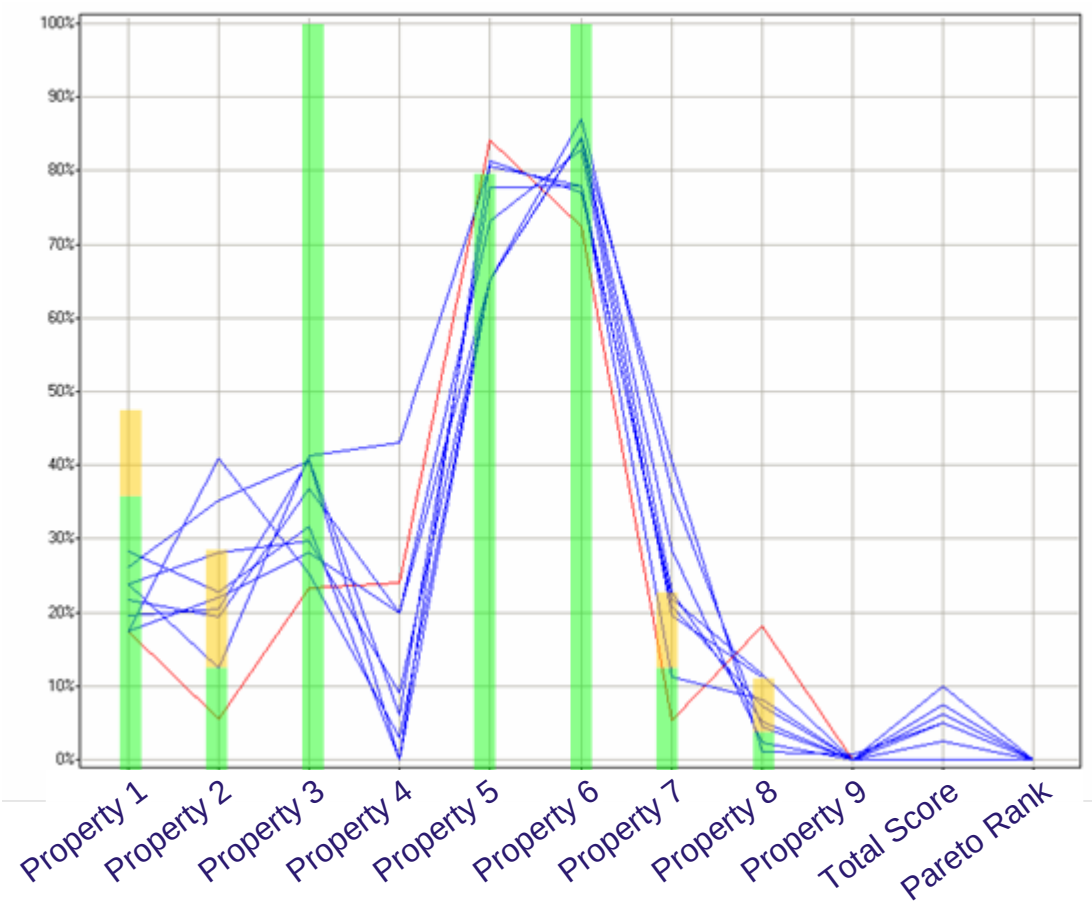
# Subsets	# Compounds for Max Cells (364)
● 1	1395
◆ 2	1056
▲ 4	938

Using Pareto Ranking to Profile Compounds Synthesised in Lead Optimisation



Profiles of High Scoring Compounds

Good solutions can be missed if property filters are applied sequentially

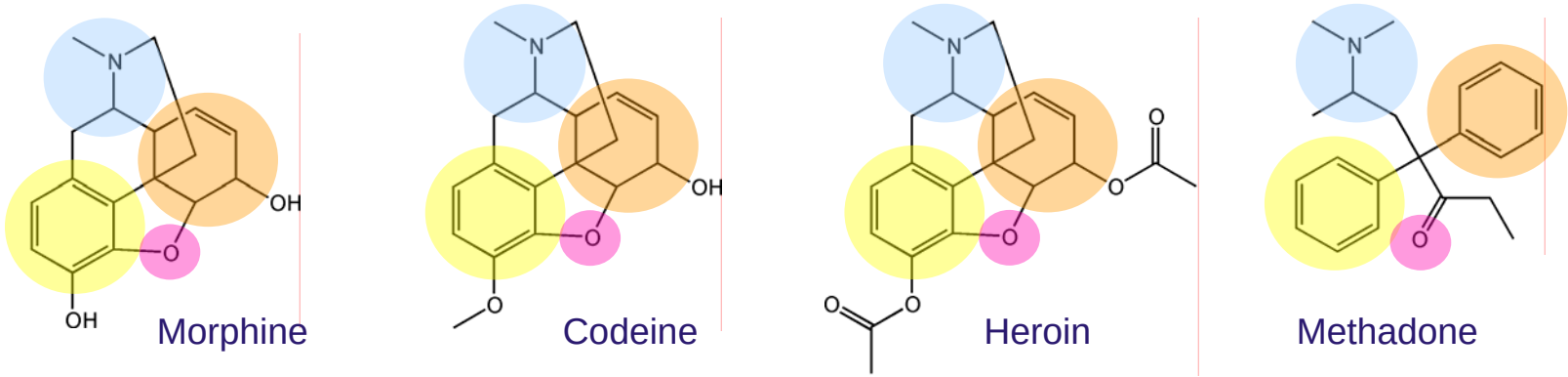


The University Of Sheffield.

Reduced Graphs as Molecular Descriptors

Reduced Graphs

- Emphasise functional over structural similarity



Morphine | Codeine | Heroin | Methadone

Daylight similarity: 0.99 | 0.95 | 0.20

- Applications
  - Identify compounds with similar activities but different skeletons
  - Cluster representation
  - Analysis of HTS to identify SAR

Fitness Evaluation

- For each query: search through a dataset containing active and inactive molecules represented as RGs:

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

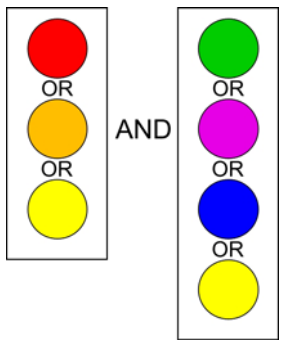
Actual Class \ Predicted Class	Active	Inactive
Active	TA	FI
Inactive	FA	TI

  - 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-measure=*2PR/(P+R)*
    - Equally weighted “average” of precision and recall

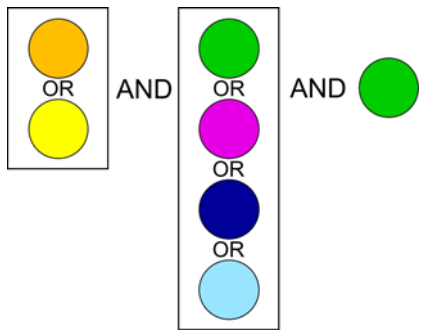
hERG Dataset

- 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 EA runs

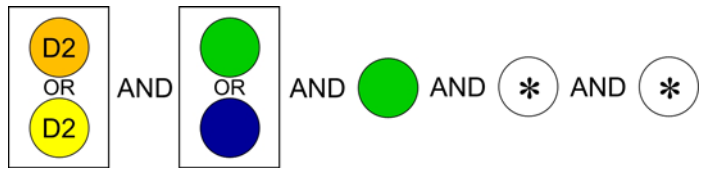
Low Activity Cut-off Query



Medium Activity Cut-off Query



High Activity Cut-off Query

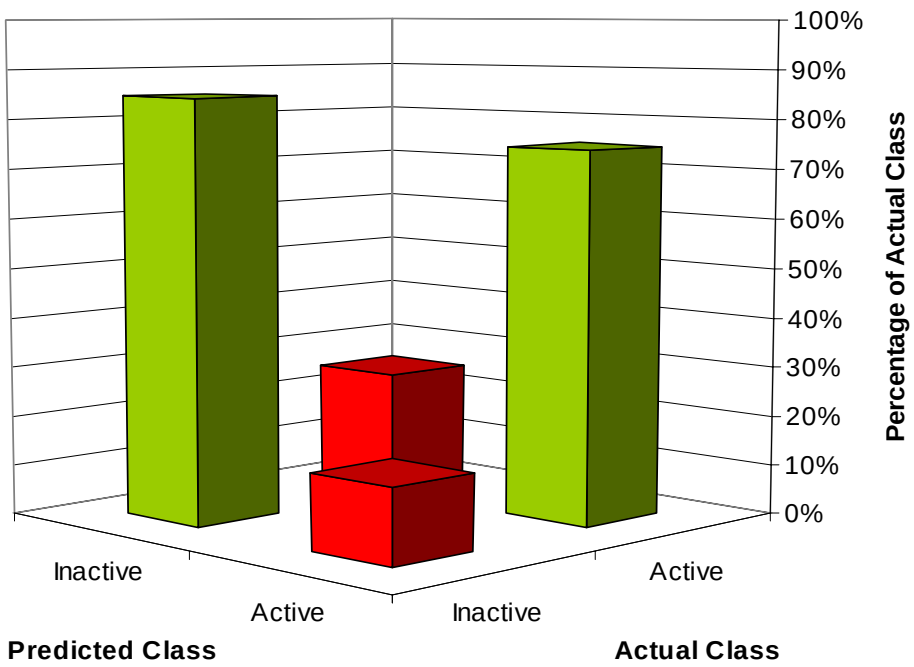


= Aromatic +ve ionisable
= Aliphatic +ve ionisable
= Acyclic +ve ionisable

= Aromatic donor
= Aliphatic donor
= Acyclic donor

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

Medium Activity Query Performance

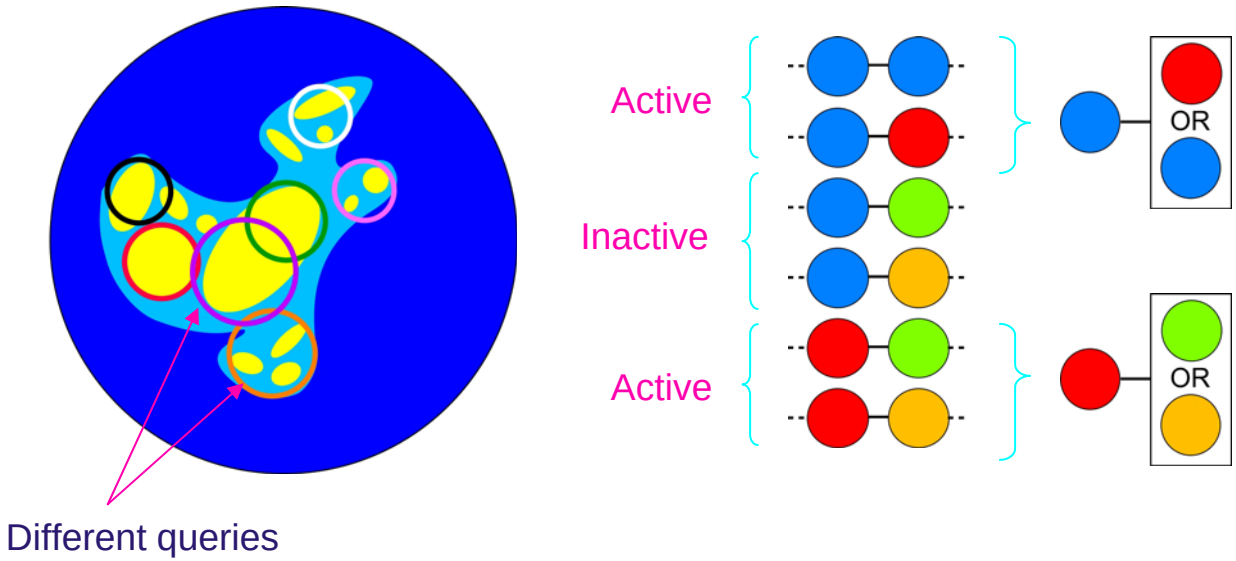


Precision = 52% **F-measure = 61%**

Recall = 75% **Enrichment = 2.9**

Evolving Multiple Queries

- Combine results from two or more queries
  - RG retrieved if found by query A or query B …
  - Potential increases in recall, precision and diversity
    - Several different types of RG can exist within an activity class due to high structural variability or different binding modes



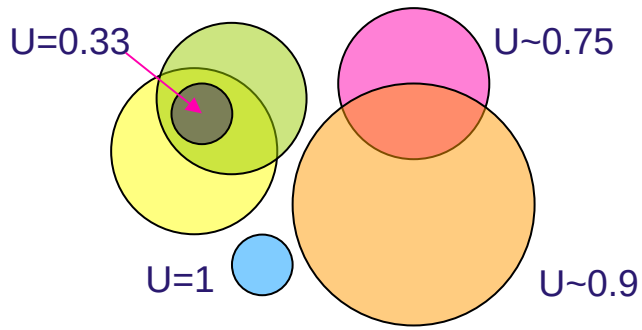
Evolving Multiple Queries

- Queries must be complementary
  - Uniqueness score

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

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

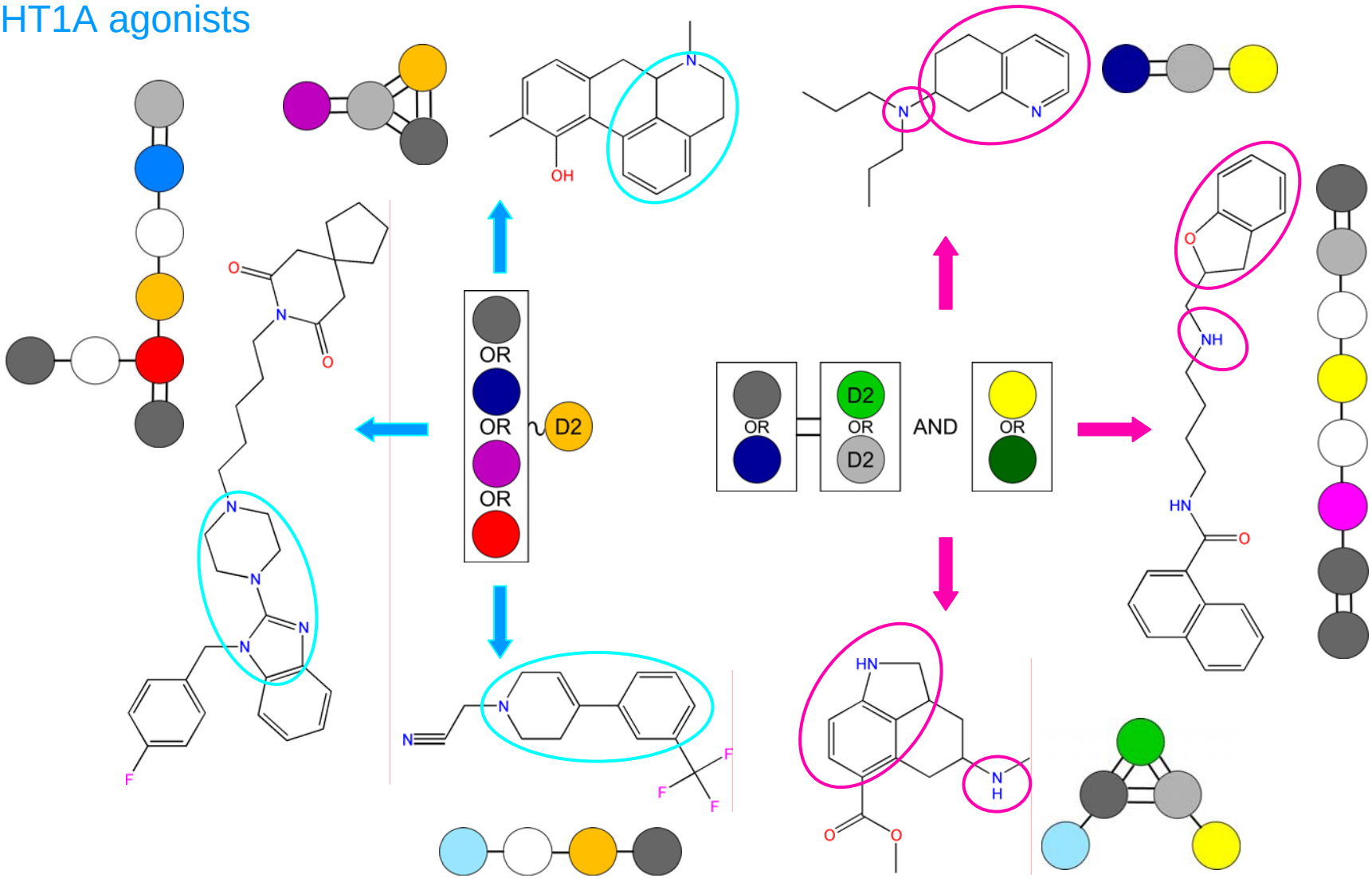
- Queries must be specific
  - To prevent large numbers of false actives resulting from the combination of several queries



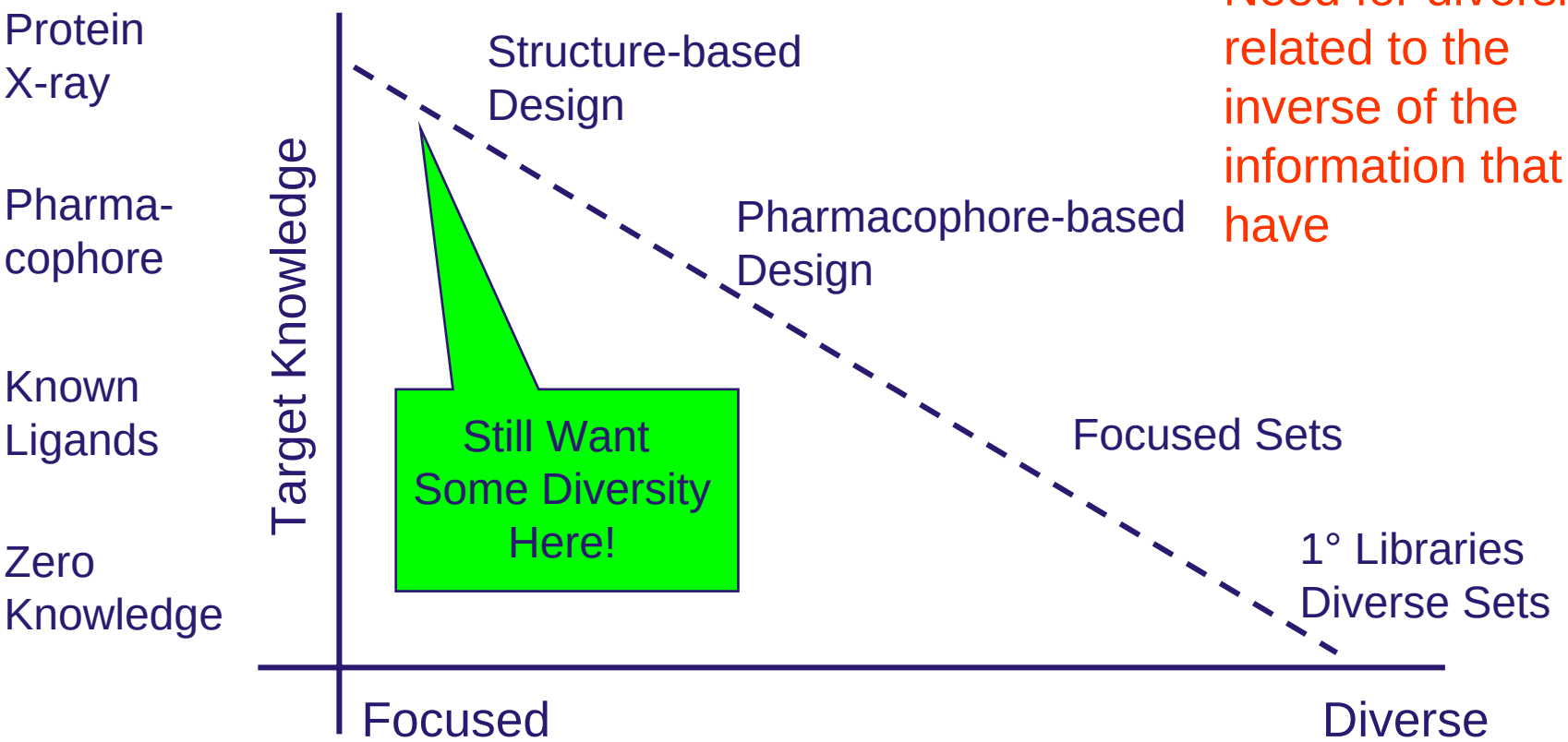
  - Typically increased specificity necessitates lower recall
- Pareto ranking used to evolve queries that maximise recall, precision and uniqueness

Extracting Multiple SARs

5HT1A agonists



Summary: Library Design Amount of Focus and/or Diversity Needed is Knowledge-Based



Need for diversity is related to the inverse of the information that you have

A.R. Leach GSK

Summary: II

- There are more compounds available for testing than ever before and therefore there is a great need to design screening sets and combinatorial libraries carefully
- HTS means that obtaining information about the activities of compounds is easier than it has ever been but it is still time consuming and expensive if done on a very large scale
- Virtual screening and computational filters provide a variety of ways of prioritising compounds
- Pareto ranking provides a useful way of exploring trade-offs in different properties to be optimised
- A wide range of molecular descriptors have been devised for similarity searching and diversity analysis
- Careful selection of descriptors and method is required

References

- Combinatorial Library Design
  - Gillet et al. *J. Chem. Inf. Comput. Sci.* 37, 1997, 731-740
  - Gillet et al. Designing Focused Libraries Using MoSELECT. *J. Mol. Graphics* Model. 20, 2002, 491-498.
  - Gillet et al. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Comp. Sci.* 42, 2002, 375-385.
  - Wright et al. *J. Chem. Inf. Comp. Sci.* 43, 2003, 381-390.
- Reduced Graphs
  - Gillet et al. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comp. Sci.* 43, 2003, 338-345.
  - Barker et al. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs *J. Chem. Inf. Model.* 46, 2006, 503-511.
  - Gardiner et al. Cluster Representation Using Reduced Graphs. *J. Chem. Inf. Model.* 47, 2007, 354-366.
  - Birchall et al. Evolving Interpretable Structure-Activity Relationship Models. Part 1: *R*educed Graph Queries *J. Chem. Inf. Model* In Press
  - Birchall et al. Evolving Interpretable Structure-Activity Relationship Models. Part 2: Using Multiobjective Optimisation to Derive Multiple Models *J. Chem. Inf. Model* In Press