

Multi-objective approaches to challenges in drug discovery

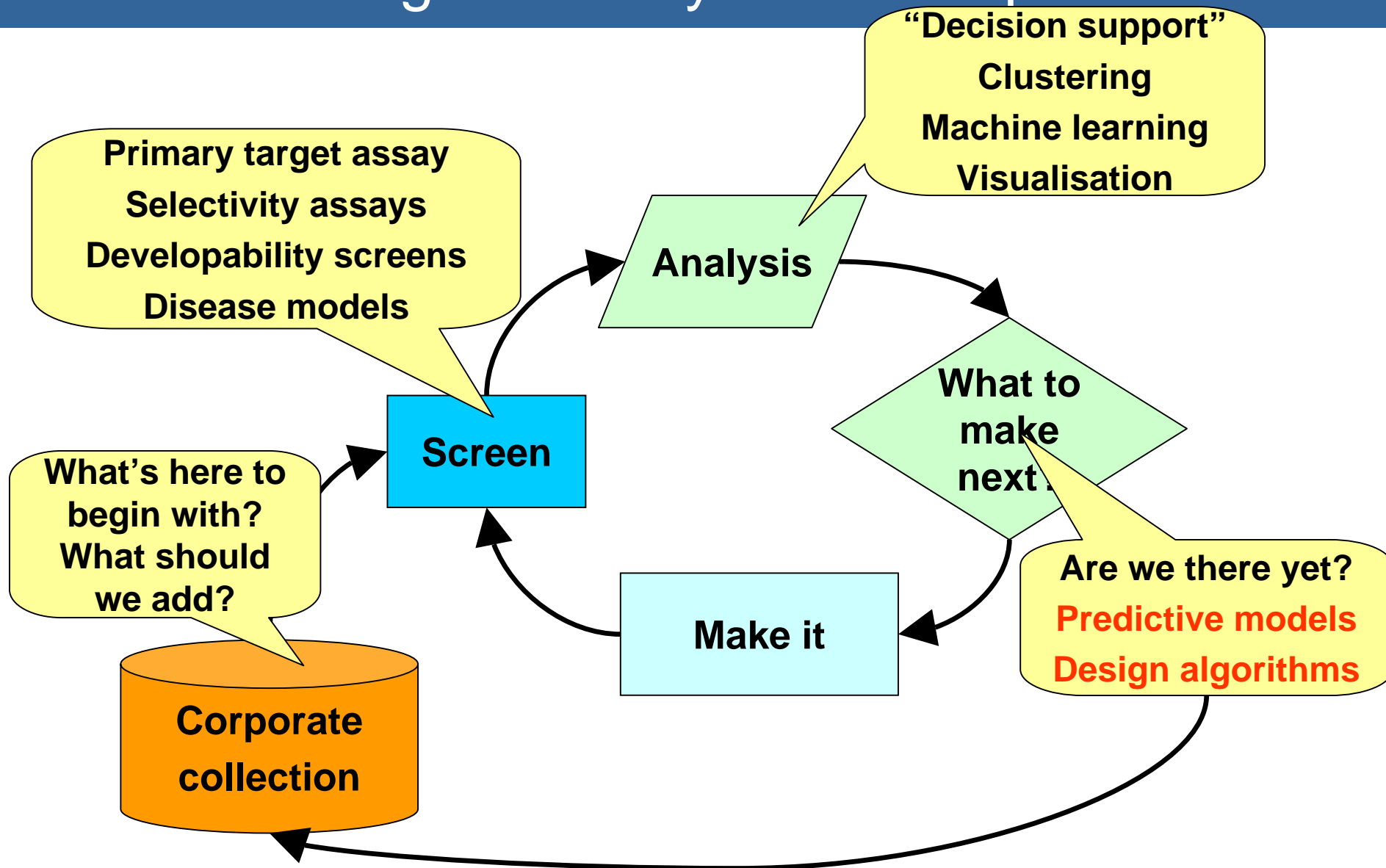
Stephen Pickett
Cheminformatics
GlaxoSmithKline

The World we live in ...

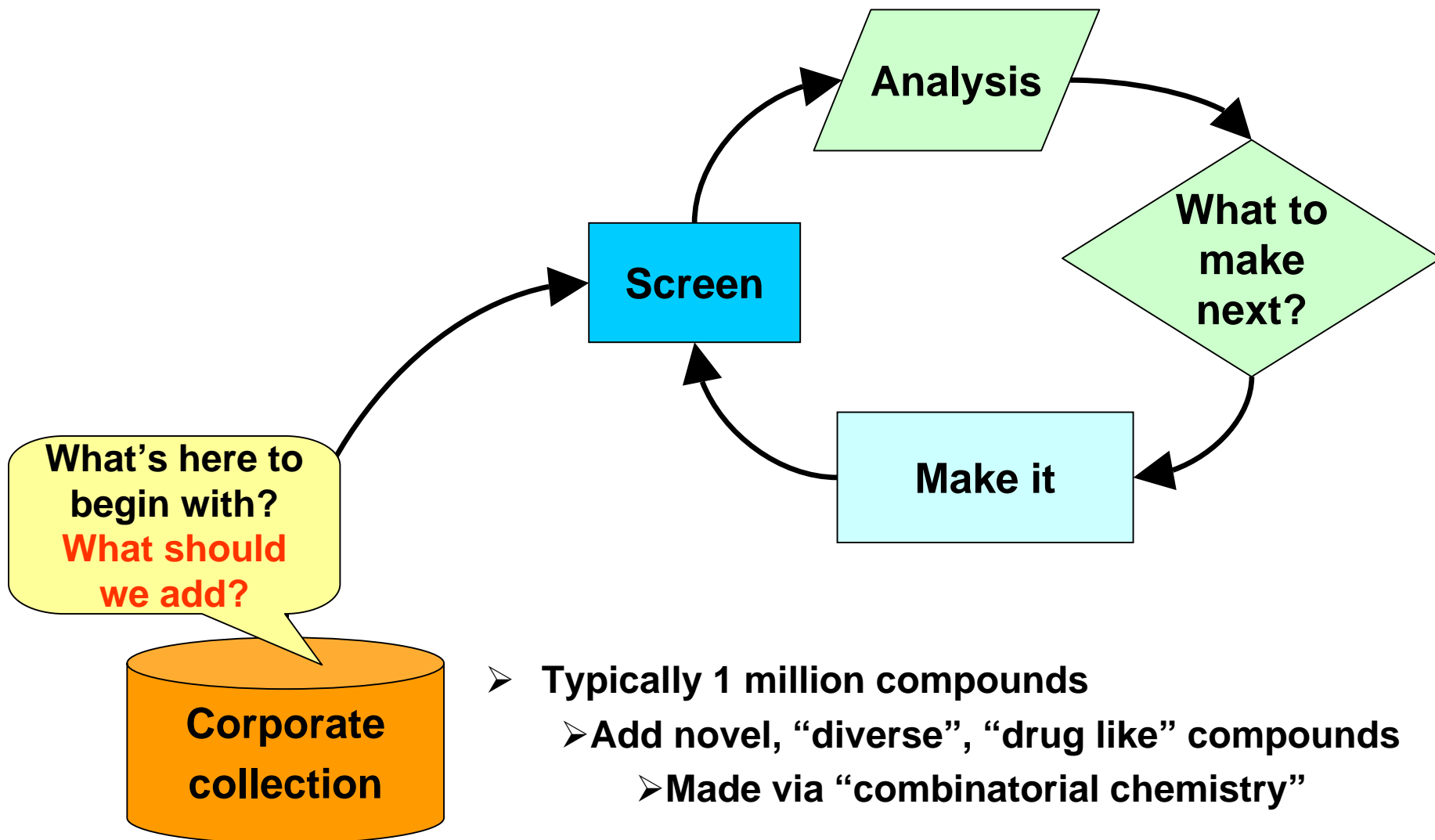
- **is not 1-dimensional**
- **Most decisions we make involve more than one parameter**
 - Time and place
 - Size, pattern, colour, style
- **Relative importance varies with season, current trends, likely use, personal preference (bias)**

- **The drug discovery process is no different!**

Drug discovery made simple



Start at the beginning



What is Combinatorial Chemistry?

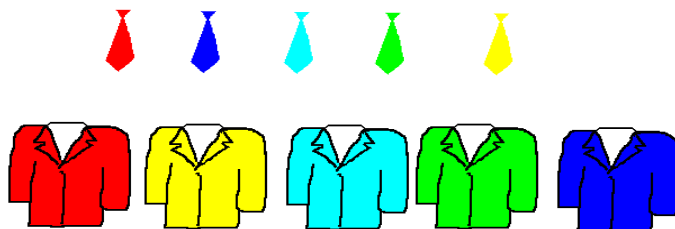
- A Generic label for a diverse set of chemistry technologies
- Automation of solution and solid-phase chemistry
- Enhancement of traditional synthesis (choose a suit and tie, then make it)



- Parallel synthesis



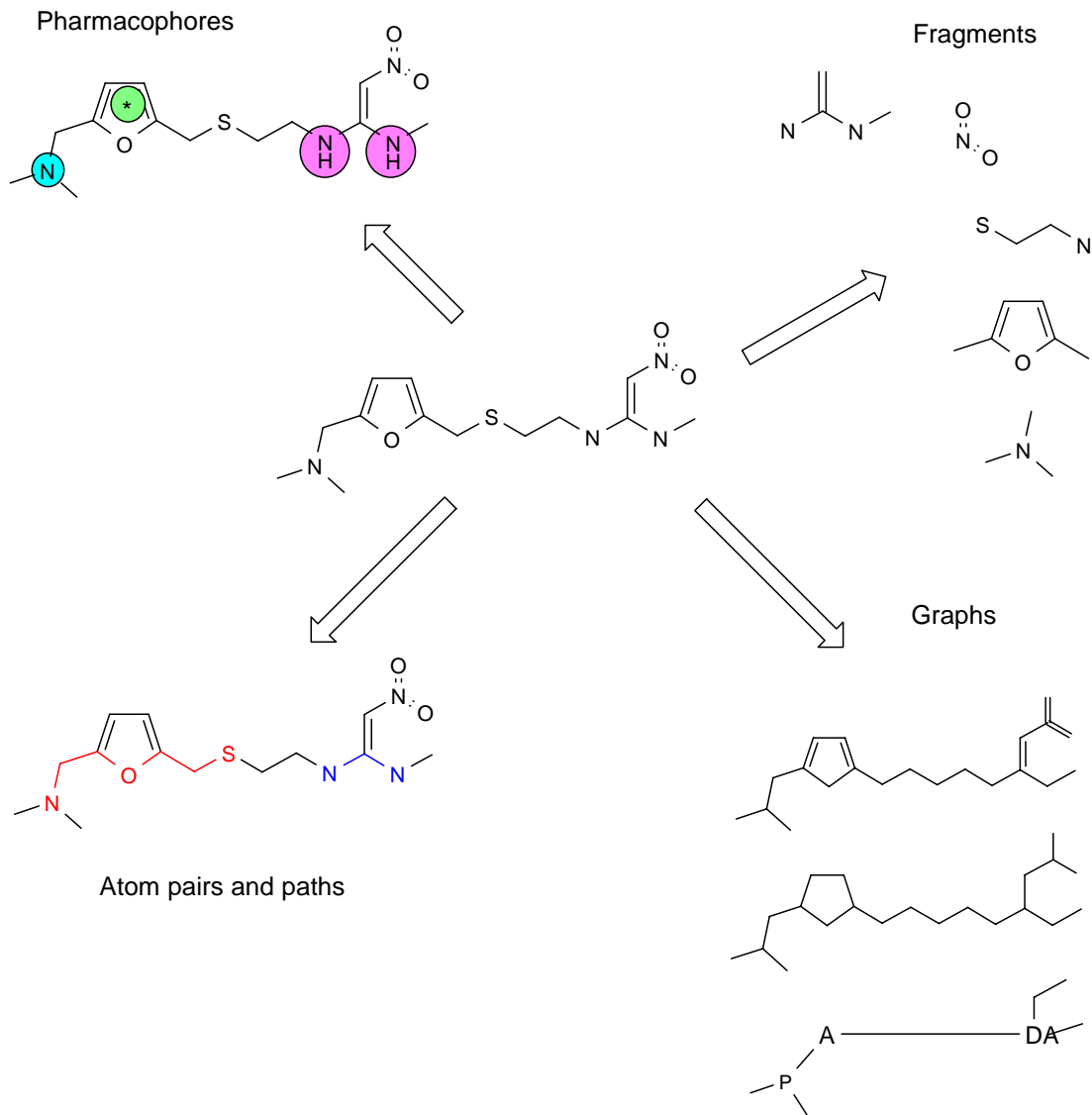
- *Combinatorial* Synthesis



Screening Collection Design Objectives

- **Novelty descriptor**
 - Minimise chemical structure similarity to current GSK screening set
- **Diversity descriptors**
 - Maximise
 - Number of new clusters (chemical structure descriptors)
 - Number of new “pharmacophores”
- **“Drug like” descriptors**
 - “Lipinski rules” and such like
 - Size (Molecular weight), lipophilicity (logP), number of polar groups (HBA, HBD)
 - Often used as property distributions
 - Minimise number of compounds outside of these guidelines
- ***Complex optimisation process***

Examples of Chemical structure descriptors

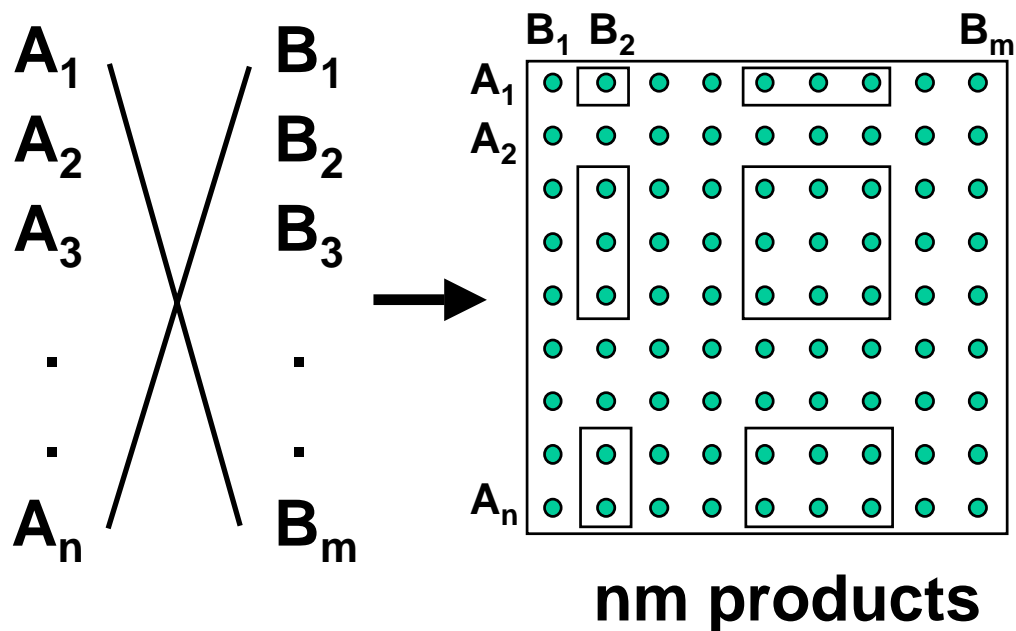


Combinatorial Synthesis

Traditional Synthesis



Combinatorial Synthesis



6 × 4 combinatorial subset

1st Generation Library Design - SELECT

Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci.* (1999), 39(1), 169-177

$$\text{score} = \sum (w1.\text{diversity} + w2.\text{property1} + w3.\text{property2}...)$$

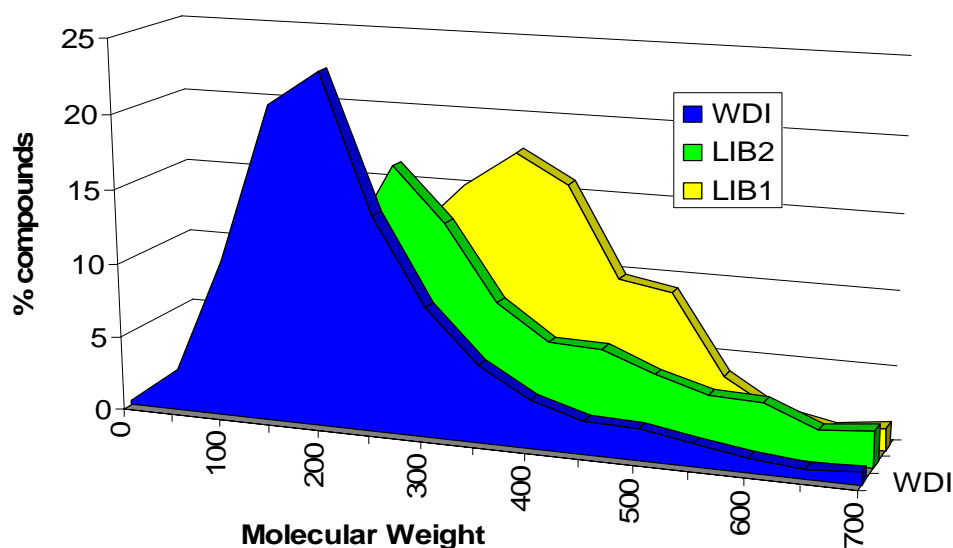


Figure 1. The distribution of molecular weights is shown in the World Drugs Index (blue); in a library optimised on diversity alone (yellow); and in a library optimised on diversity and molecular weight profile simultaneously, via the weight-sum fitness function in SELECT (green). Thus a more “drug-like” library has been selected (at the expense of some loss in diversity).

Limitations of a Weighted-Sum

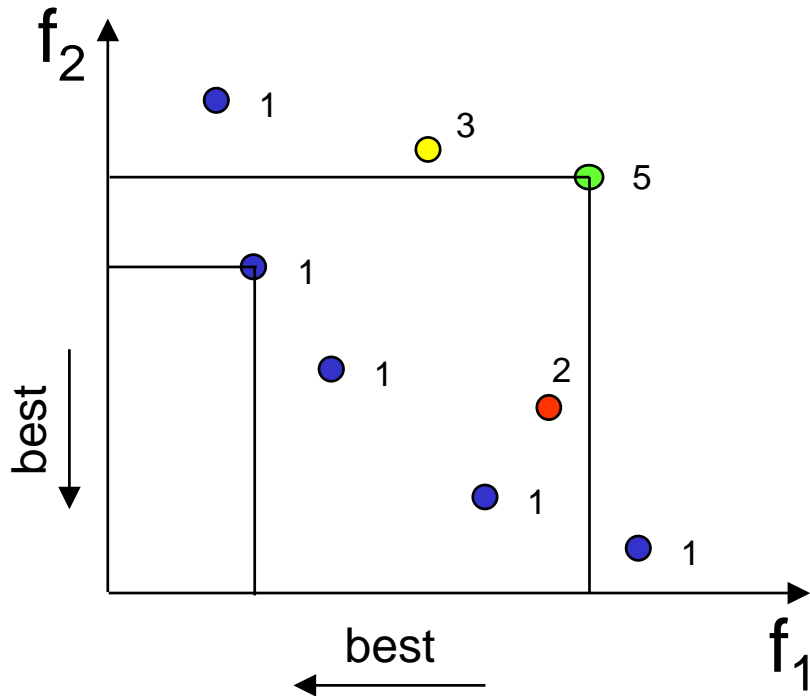
- **Definition of fitness function difficult especially for non-commensurate objectives**
 - e.g. molecular weight profile and cost
- **Setting of weights is non-intuitive**
- **Can result in regions of search space being obscured especially when objectives are in competition**
- **Difficult to monitor progress since >1 objective to follow simultaneously**
- **A single “best” solution is found per run**

Multiobjective Optimisation

- **Evolutionary algorithms, e.g. GAs**
 - operate with a population of individuals
 - suited to search for multiple solutions in parallel
 - readily adapted to deal with multiobjective optimisation
- **MOGA: MultiObjective Genetic Algorithm**
 - Fonseca & Fleming. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 28(1), 1998, 26-37.

Dominance & Pareto Ranking

- A **non-dominated** solution is one where an improvement in one objective results in a deterioration in one or more of the other objectives when compared with the other solutions in the population



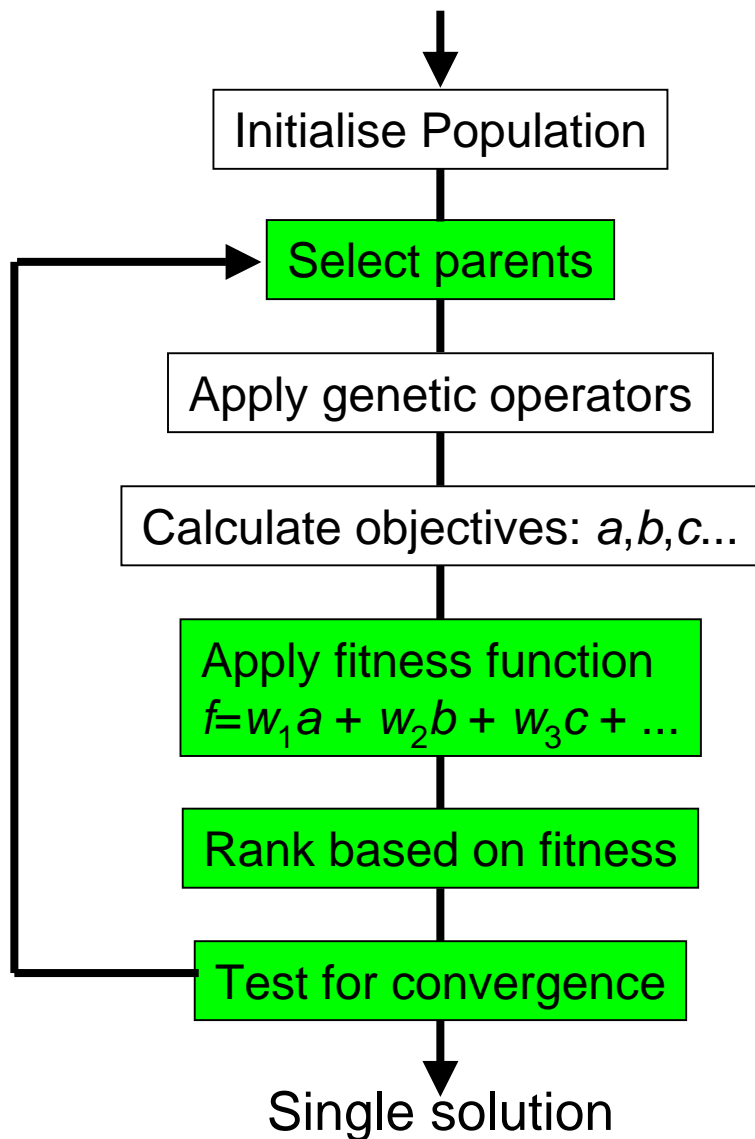
Pareto ranking: an individual's rank corresponds to the number of individuals in the current population by which it is dominated

Multiple objectives are handled **independently** **without** summation and **without** weights

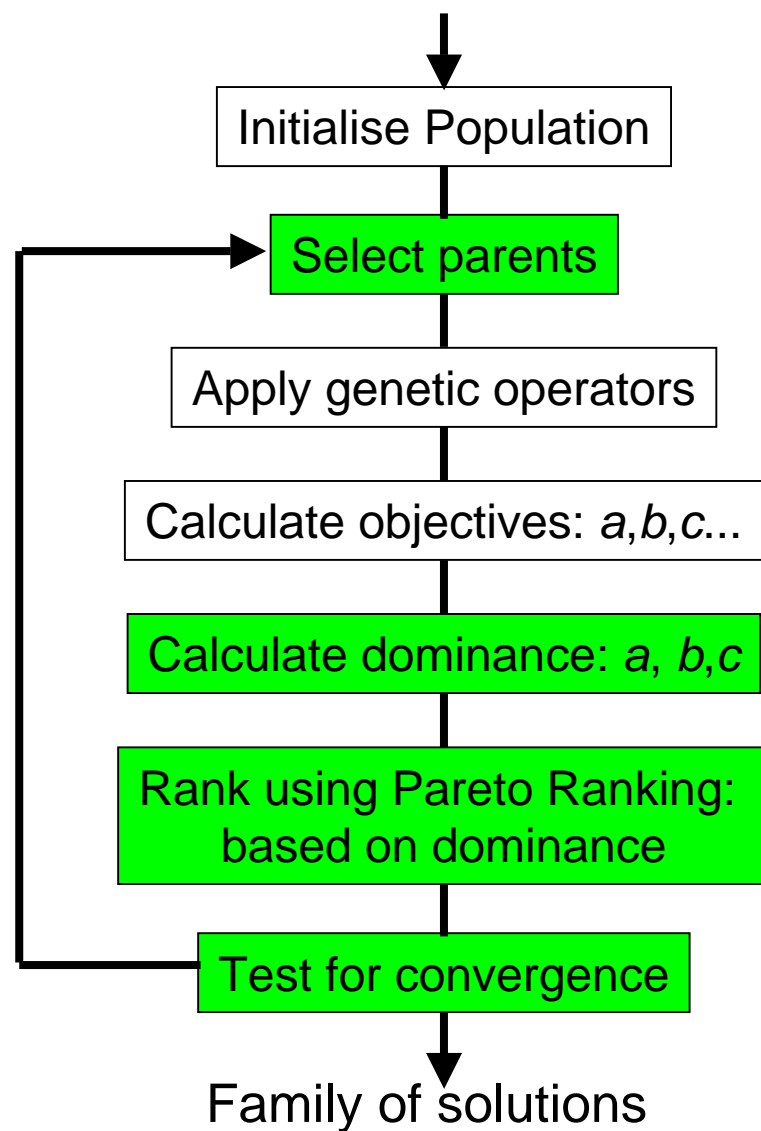
MOGA

- Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. S. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. JCICS (2002), 42(2), 375-385
- **Multiple objectives are handled independently without summation and without weights**
- **The GA means that you get many solutions**
 - represents a continuum of solutions where all solutions are seen as equivalent
 - represents compromises or trade-offs between the various objectives
- **A family of non-dominated solutions is found rather than a single solution**
- **Visualisation of the search progress allows trade-offs between objectives to be observed**
 - the user can make an informed choice on which solution(s) to explore

SELECT



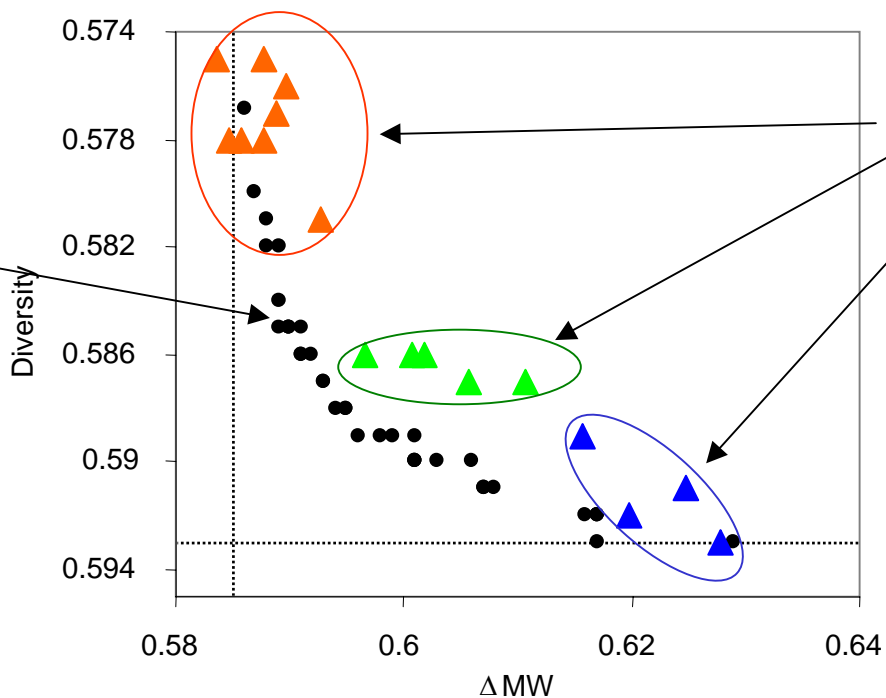
MoSELECT



Why MoGA?

- Finding the same coverage of solutions using **SELECT** method would require multiple runs using various combinations of weights

MOGA spreads solutions out over whole surface in one run

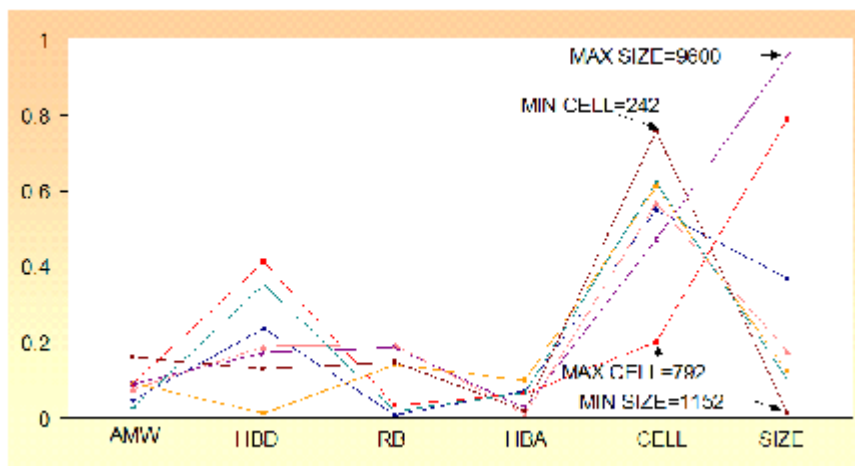


Note how using different weights clusters the results

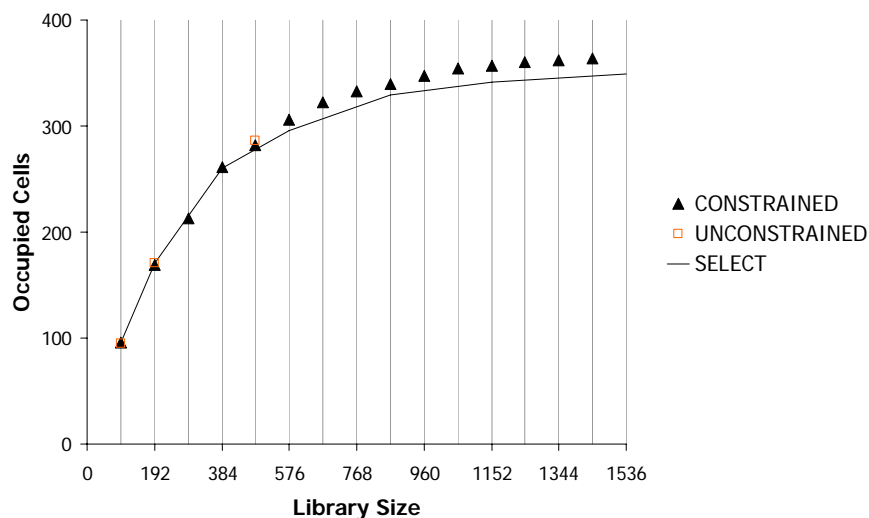
MoGA

- **MoGA allows us to ask 'smarter' questions**

- Optimize library configuration ($m \times n \times p$) and size
 - often decided before hand so can bias results!
- Multiple small arrays vs. one large array for a chemistry
- Wright, T.; Gillet, V. J.; Green, D. V. S.; Pickett, S. D. Optimizing the Size and Configuration of Combinatorial Libraries. *JCICS* (2003), 43(2), 381-390.



Identifying optimal library size
Combine size with other objectives



Satisfying experimental constraints
Constraining plate coverage

adept/adept - Microsoft Internet Explorer

Tools Help

Search Favorites Media

adept/adept

[X syp36393] [Expert Users](#) [New Job](#) [Jobs](#) [Moga](#) [Solutions](#) [Help](#)

MOGA R Group Range Choices for Job : jayshree

*Choose the dimensions of the library to be selected from the virtual library.
Individual R Group numbers must be within the ranges indicated.*

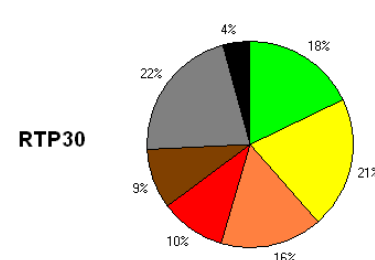
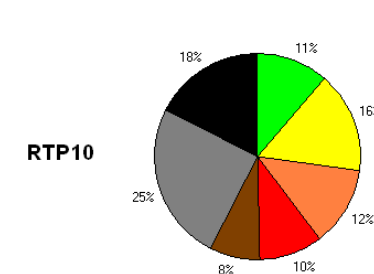
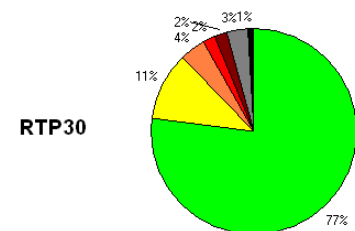
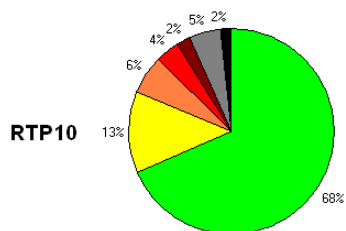
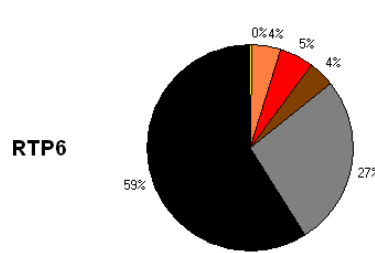
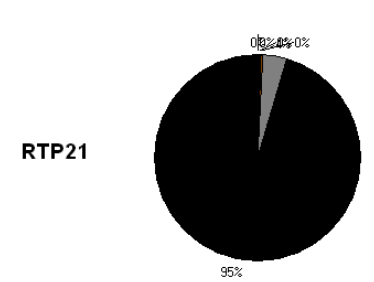
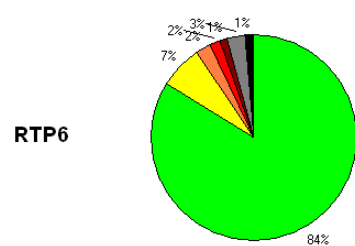
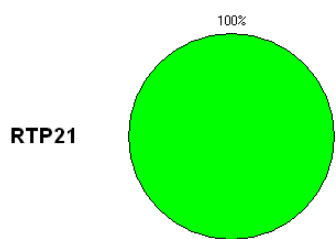
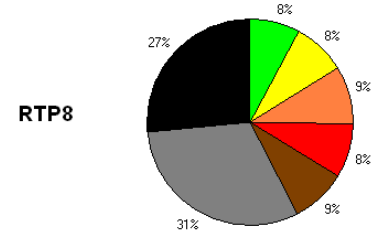
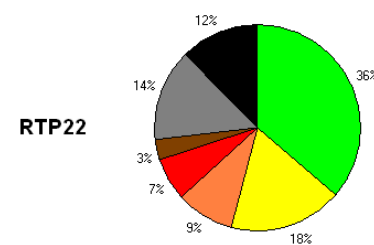
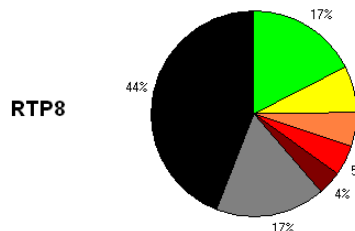
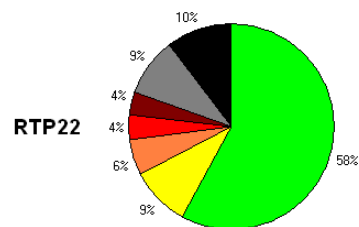
R Group 1 contains 1 Building Blocks. Choose between 1 and 1.	<input type="text" value="1"/>	<input type="button" value="View Structures"/>
R Group 2 contains 150 Building Blocks (14 Forced). Choose between 14 and 150.	<input type="text" value="24"/>	<input type="button" value="View Structures"/>
R Group 3 contains 24 Building Blocks. Choose between 1 and 24.	<input type="text" value="24"/>	<input type="button" value="View Structures"/>

Library Prioritisation

Number of compounds similar to library product

good **bad**

0 1 2 3 4 5-10 >10



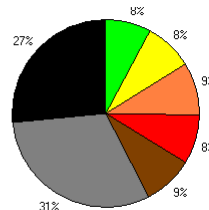
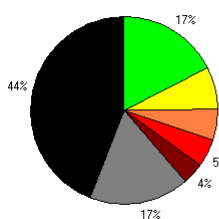
GSK

Internal

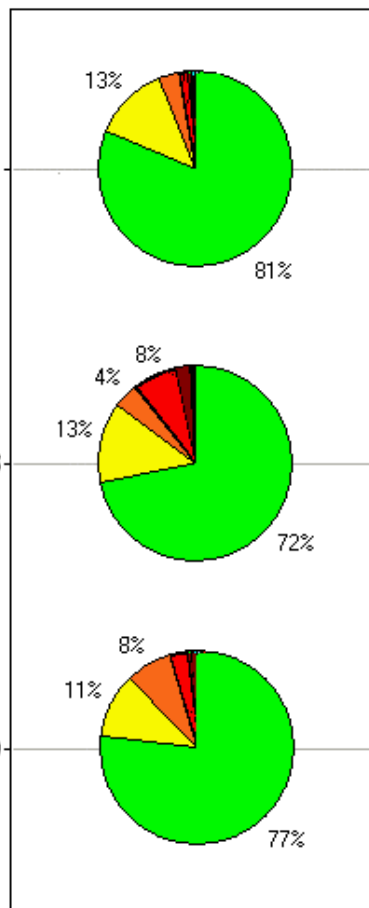
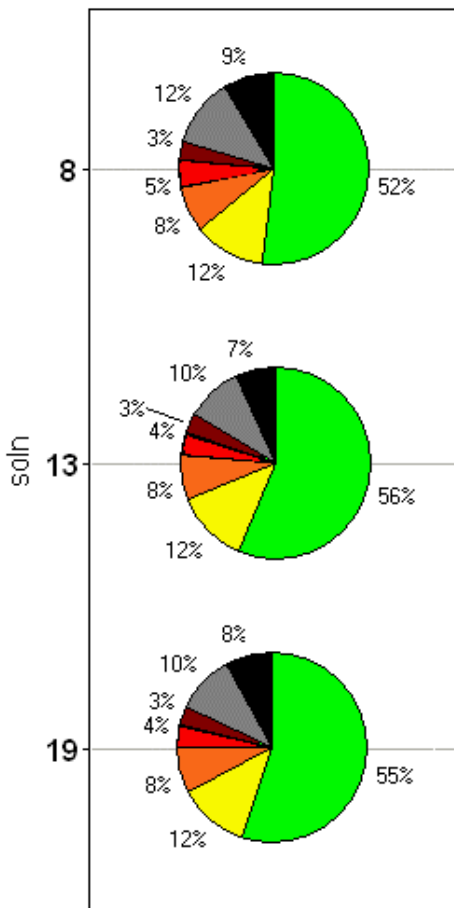
Number of compounds similar to
good library product **bad**

0 1 2 3 4 5-10 >10

Virtual Library
 20K compounds



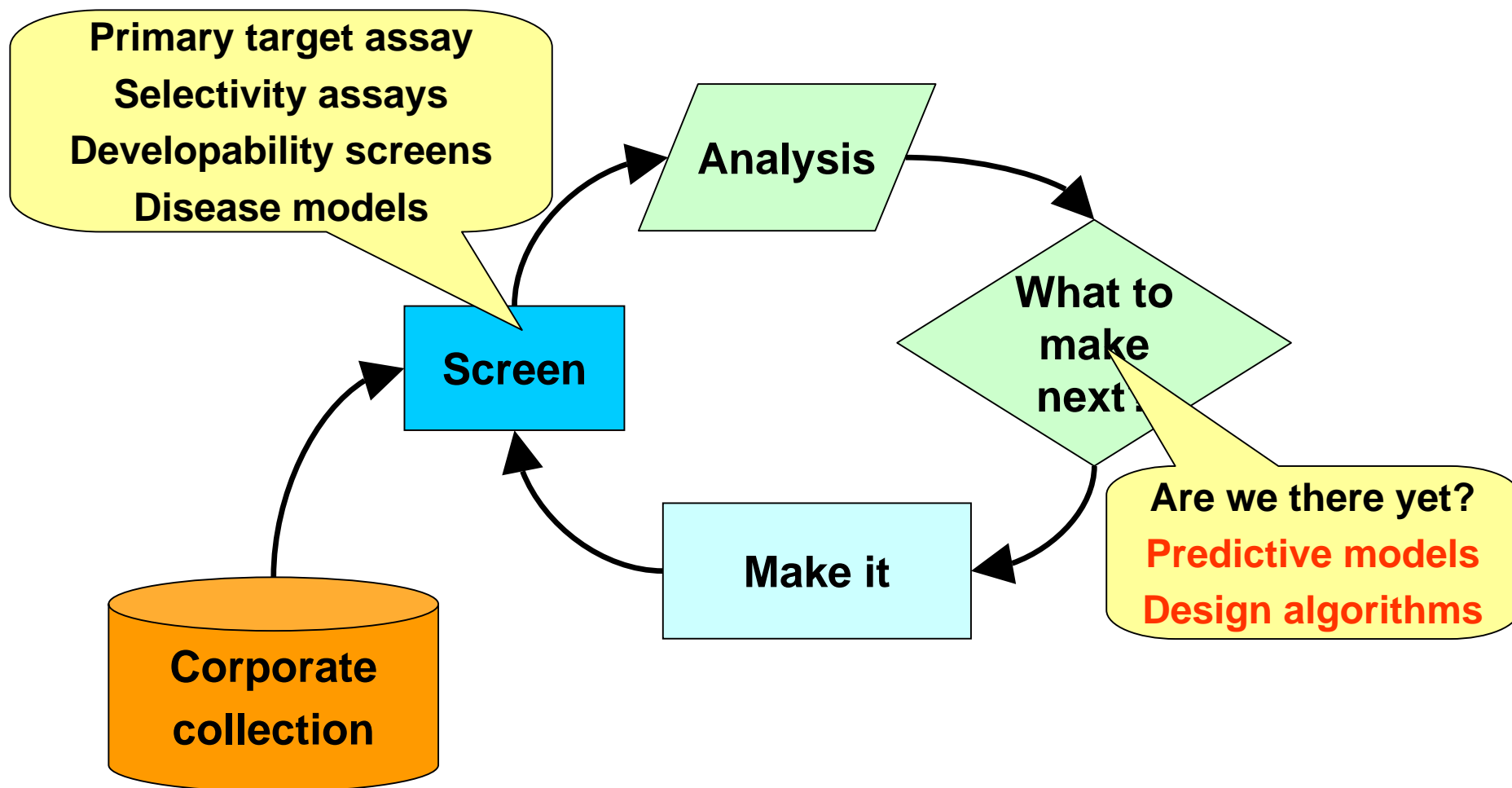
3 MOGA
 Solutions
 1920 compounds



GSK

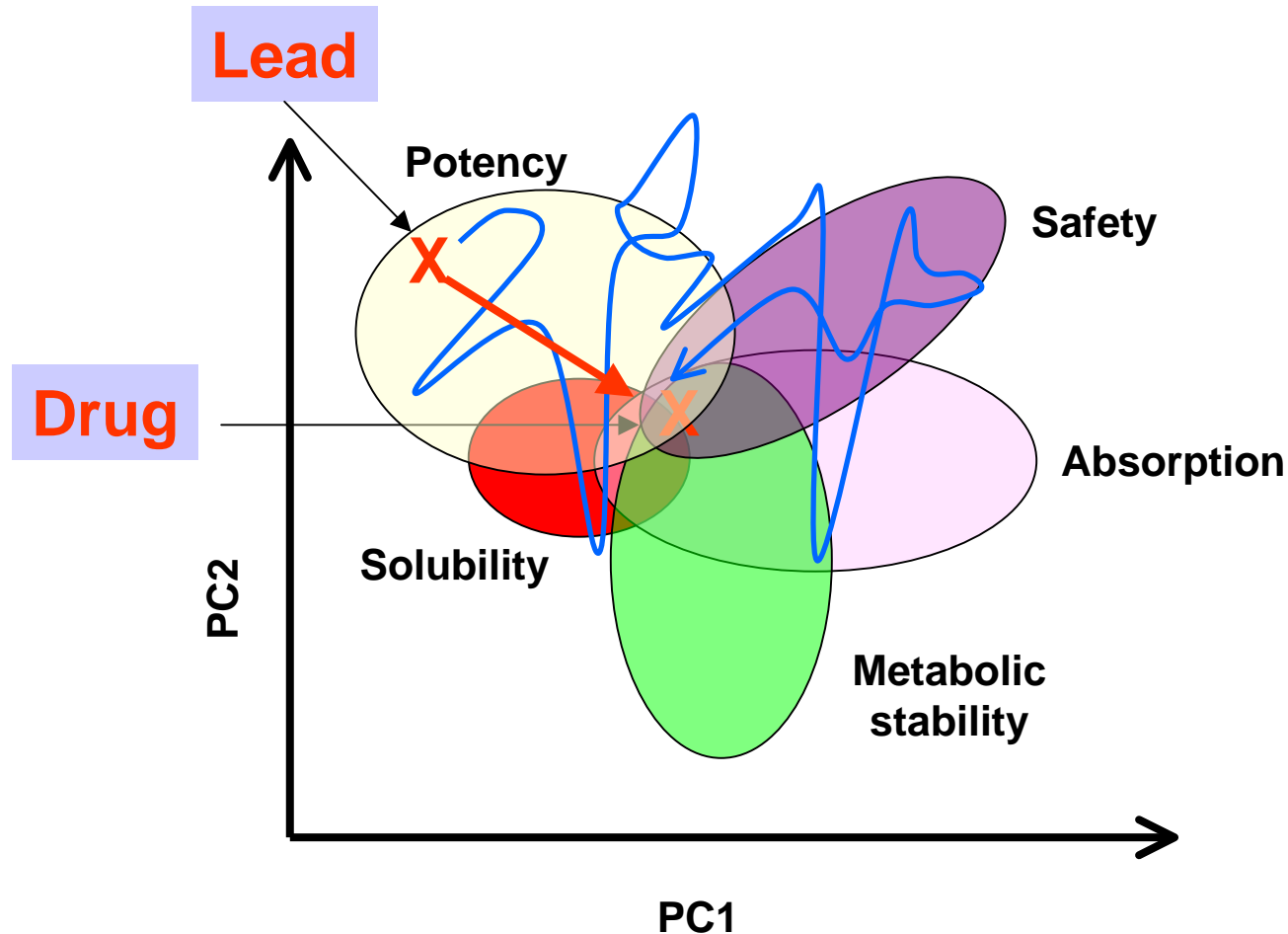
Internal

Drug discovery made simple



Why Predict?

A multi-objective optimisation process



in-silico: Faster way to navigate the route via prediction and knowledge

What are predictive models?

- **A predictive model quantitatively relates a number of *descriptors* (variable factors that are likely to influence future behaviour or results) to an *outcome*.**
 - In marketing, for example, a customer's gender, age, and purchase history (*descriptors*) might predict the likelihood of a future sale (*outcome*).
- **In drug discovery, descriptors tend to be derived from chemical structure, and outcomes are *in vitro* or *in vivo* phenomena**
 - the goal is to predict behaviour before synthesis
 - models can be built from experimental data too:
 - e.g. prediction of bioavailability from solubility, permeability and clearance data

Statistics

- **Various statistical methods are applied to find the mathematical relationship between the descriptors and the outcomes**
 - multiple linear regression, logistic regression K-nearest neighbours, PLS, linear discriminant analysis, decision trees, neural networks, Support Vector machines and many more
 - Choice depends on
 - data type/volume
 - the objectives for the model (see later)
 - personal preference
- **The resulting equations are generically termed QSAR (Quantitative Structure Activity Relationships) models**

Common uses of predictive models

- **Lead generation**

- Filtering of structures to remove poor start points from screening collection
 - “Lipinski’s rules”, sub-structure filters, hard to remove or critical properties like poor solubility, permeability
 - Even 70% predictive models are useful, as they can enrich the proportion of “good” compounds coming in

- **Hit to candidate**

- Used to guide medicinal optimisation
 - ***Predictive power and interpretability*** are key
 - Interpretability can often compensate for poor predictive power, as gives insights to the chemists as to what might solve the problem

Rethinking QSAR

- **Sheffield collaboration has helped us look at the problem from an alternative perspective**
 - a multi-objective optimisation process
- **Interpretability vs predictivity**
 - Which descriptors should we use?
 - Which modelling methods?
- **Approaches such as GFA**

Rogers, D.; Hopfinger, A. J. Application of Genetic Function *J. Chem. Inf. Comput. Sci.*, **1994**, 34, 854-856

$$\text{pIC50} = -2.215 + 0.677 \text{ LOGP} - 0.000083 \text{ MOFI_Z}$$

r2 = 0.610 q2 = 0.533

$$\text{pIC50} = 2.871 + 0.568 \text{ LOGP} - 0.013 \text{ SURF_A} + 0.810 \text{ ESDL3}$$

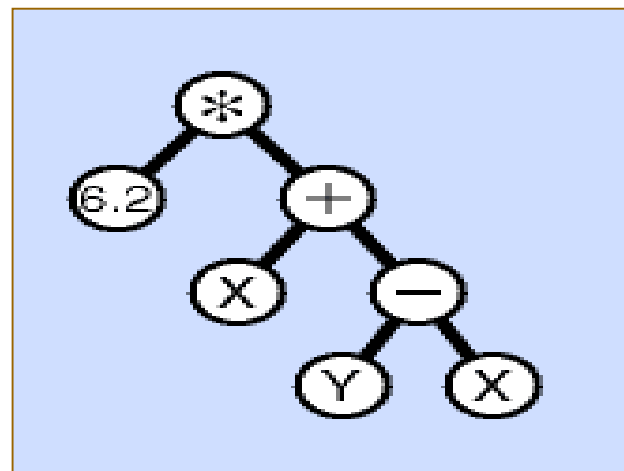
r2 = 0.719 q2 = 0.644

$$\text{pIC50} = -1.790 + 0.499 \text{ LOGP} + 2.807 \text{ ATCH4} - 0.693 \text{ ESDL3} - 0.199 \text{ PEAX_X}$$

r2 = 0.774 q2 = 0.636

Multiobjective QSAR

- **GP involves maximising or minimising a single fitness value to generate a single solution**



- **Multiobjective Genetic Programming (MoGP)**
 - Different objectives to be optimised are treated independently
 - Fitness, interpretability, complexity
 - Pareto ranking is used to evolve a family of solutions that represent the trade-offs that exist in the objectives

Nicolotti, O.; Gillet, V. J.; Fleming, P. J.; Green, D. V. S. Multiobjective Optimization in Quantitative Structure-Activity Relationships: Deriving Accurate and Interpretable QSARs. *J. Med. Chem.* (2002), 45(23), 5069-5080.

MoGP Results

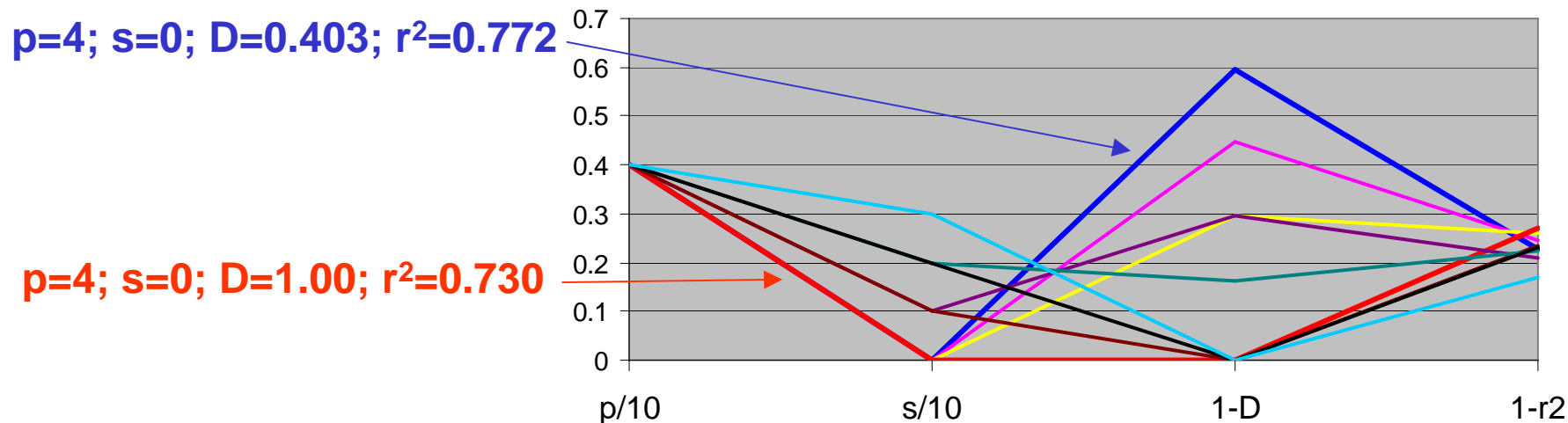
4 term models

Best linear model:

$$y_{calc} = 0.500 \text{ LOGP} - 2.808 \text{ ATCH4} + 0.842 \text{ ESDL3} + 0.199 \text{ PEAX_X} + 1.791$$

$r^2=0.774$ $q^2=0.636$

Optimize interpretability



Most Interpretable Linear Model:

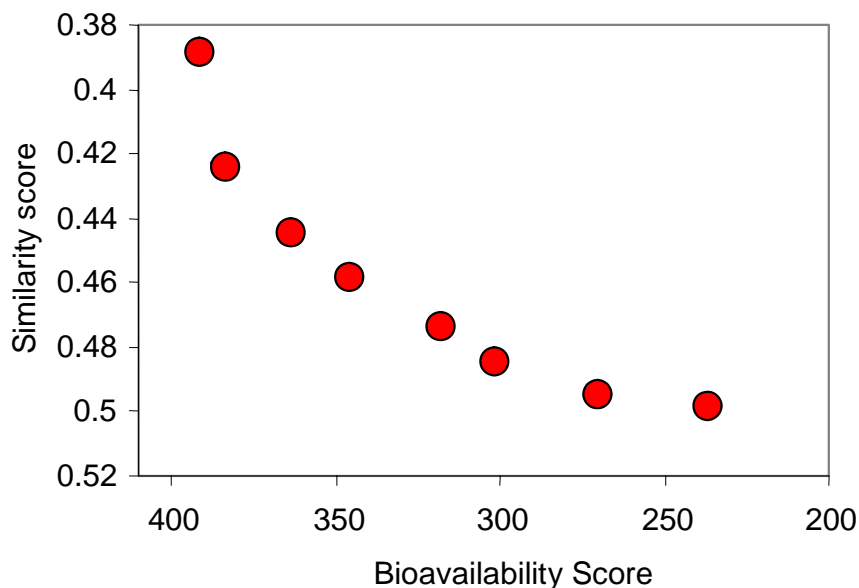
$$y_{calc} = 0.468 \text{ LOGP} - 1.904 \times 10^{-2} \text{ VDWWOL} + 7.007 \times 10^{-3} \text{ MOL_WT} + 1.364 \text{ SUM_F} + 0.141$$

$r^2=0.730$ $q^2=0.616$

excellent, fair, poor

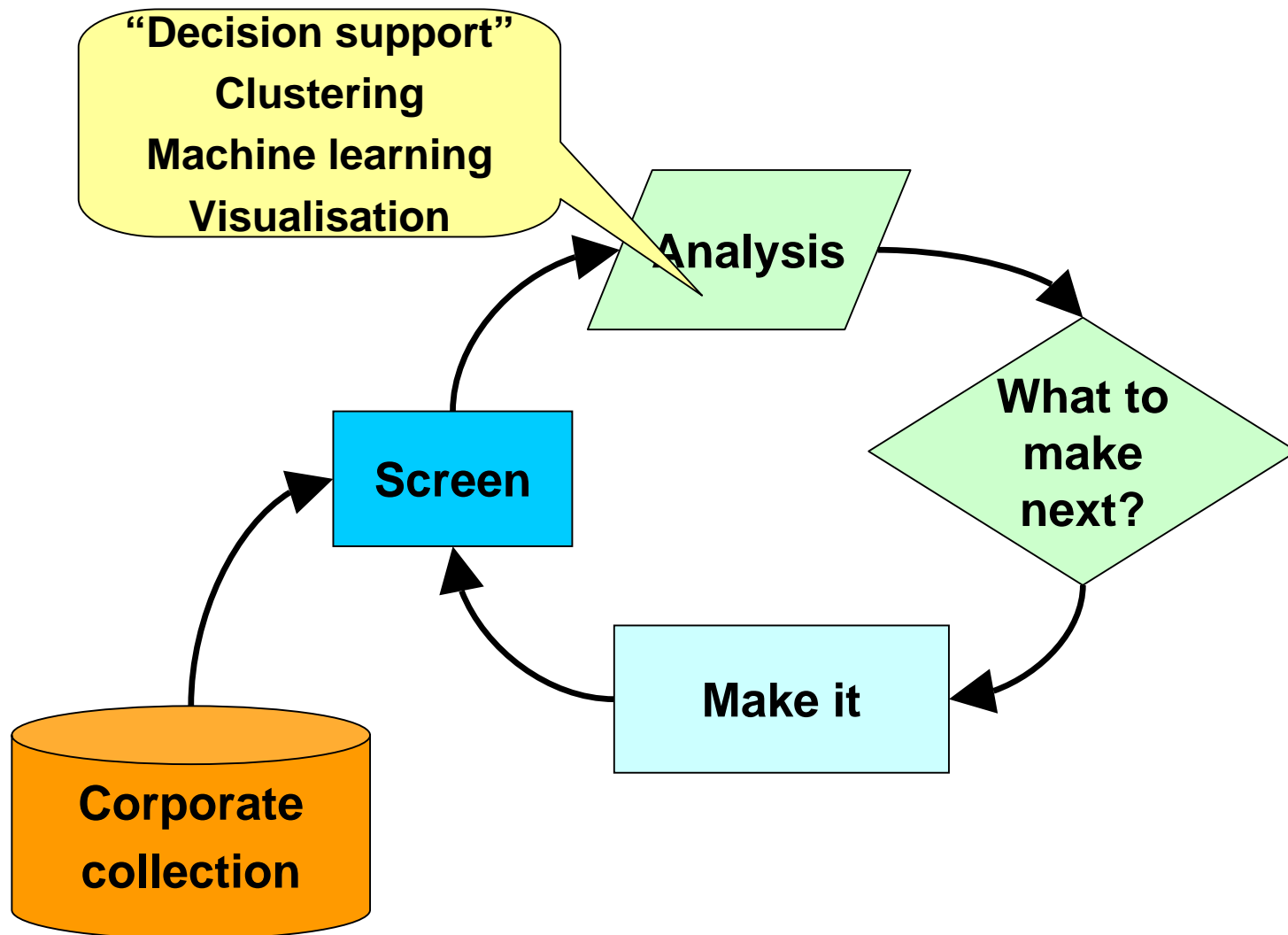
Application of MOGA and predictive QSAR

- **Algorithms applied to a variety of common examples, for both focussed and diversity libraries.**
- **MoGA copes quite happily with larger numbers of objectives.**
 - Combinations of continuous and categorical objectives
- **Output from a focussed library design with 6 objectives**
 - including similarity to a lead molecule and the bioavailability ranking for the library.

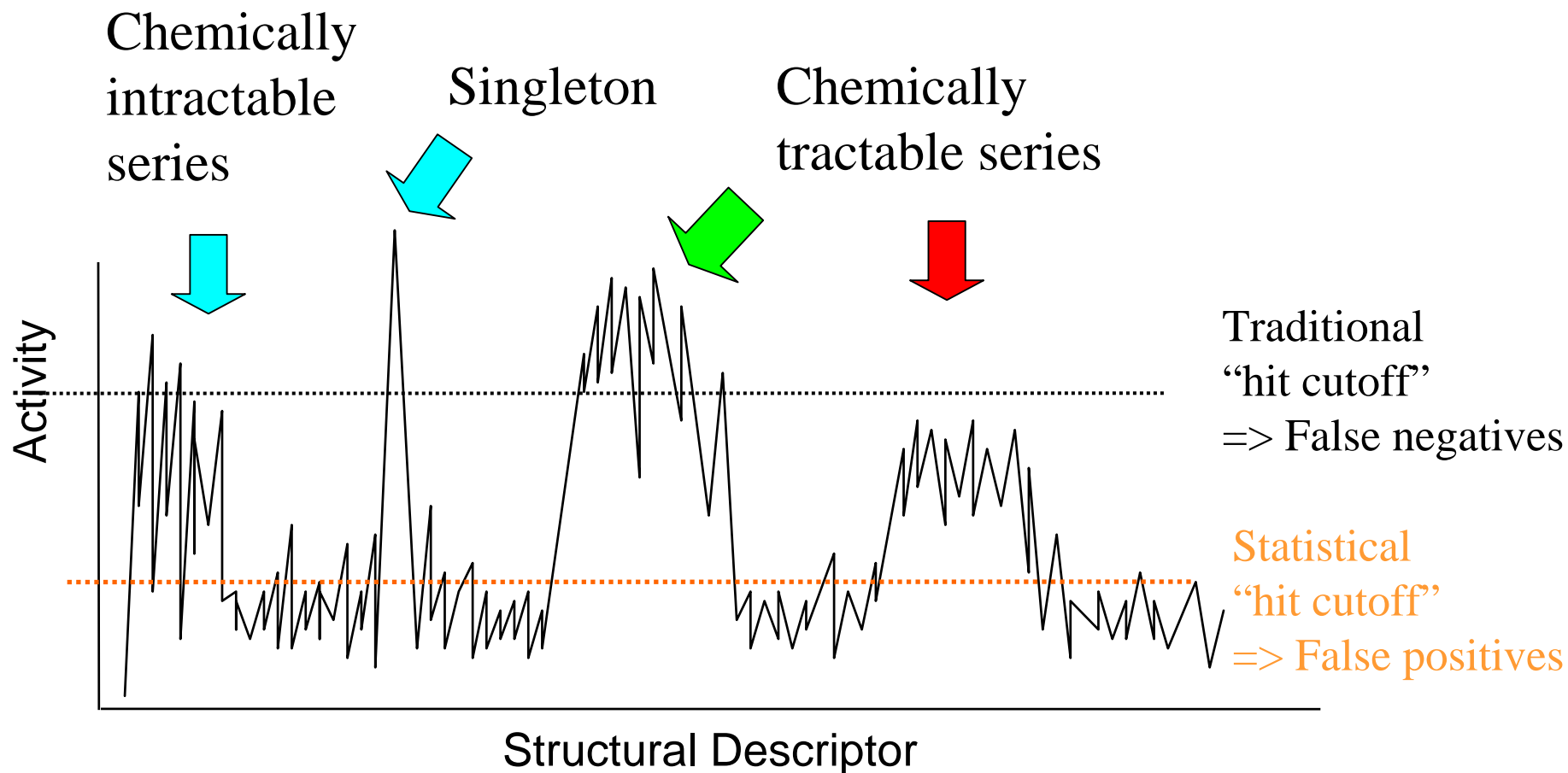


- The primary objectives of similarity and predicted bioavailability are in competition.
- May enable a user to drop one of the less important objectives in order to meet the primary goals of the library design.

Drug discovery made simple

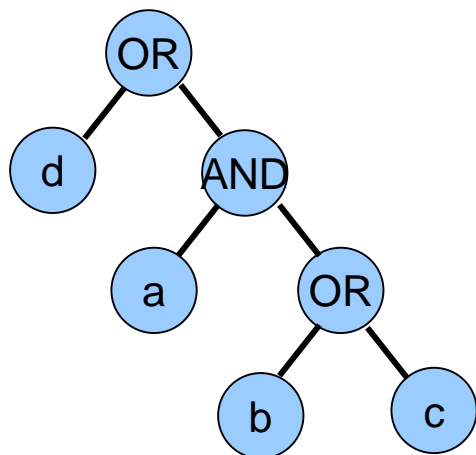


HTS data analysis: schematic



Using MoGP with Boolean Logic

- **External nodes (leafs) are molecular descriptors**
 - e.g. structural features
- **Internal nodes are boolean operators**
 - OR: allows alternative models to be built
 - AND: increases complexity of an individual model



= d OR (a AND (b OR c))

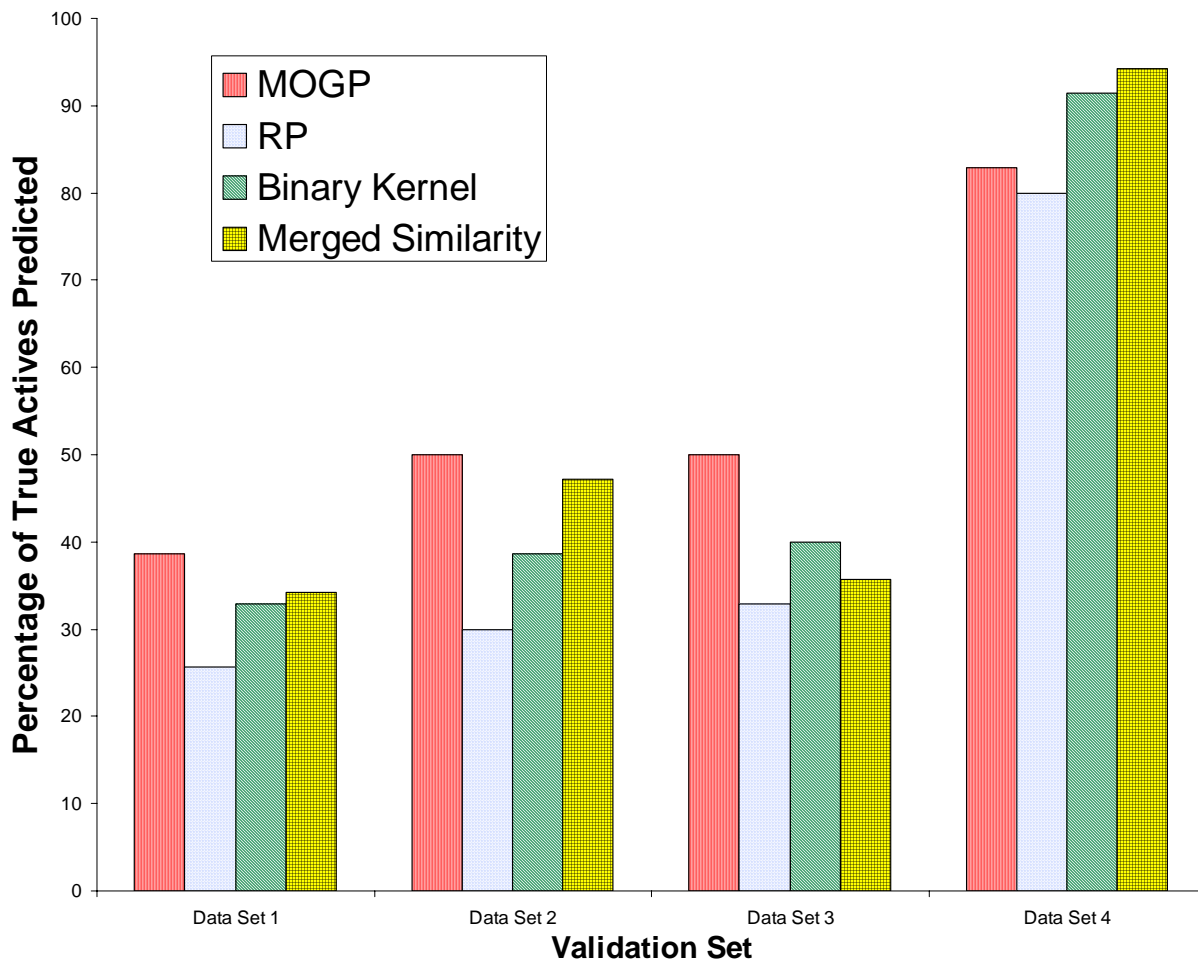
= d OR (a AND b) OR (a AND c))

- **Molecules satisfying features are predicted to be active**

MoGP for HTS

Antony Whitehead, Sheffield MSc

Comparison of Data Sets from GSK HTS Assays



Acknowledgements

- **Val Gillet, Peter Fleming & Peter Willett, University of Sheffield**
 - Wael “Illy” Khatib, Orazio Nicolloti, Trudi Wright, Kris Birchall
 - Antony Whitehead
- **Darren Green**
- **Stephen Pickett**
- **Gavin Harper, Jameed Hussain, Jimmy Chung, Nicola Richmond, Chris Luscombe, Richard Bolton, Andy Whittington, Andrew Leach, Giampa Bravi**
- **Colleagues in GSK DR**

QSAR Objectives

- **Accuracy (goodness of fit)**
 - r^2
- **Complexity**
 - number of terms (p)
- **Internal model complexity**
 - number of non-linear terms (s): +1 for pow(2); +2 for pow(3)
- **Interpretability**
 - user-defined desirability of descriptors
 - Individual descriptors are rated as excellent (3), fair (2) or poor (1)
 - Values are combined over all descriptors in the model to give a desirability score for the model (D)

Fitness Measure

- **Total number of molecules correctly classified**
 - $KA + I$
 - A number of actives correctly classified
 - I Number of inactives correctly classified
 - $K = N_I/N_A$
- **Number of models**
 - Minimise the number of different SARs
- **Model size**
 - Maximise the number of descriptors in the smallest model
- **Combined using MOGA (MOGP) approach**

Collaborations with Sheffield

GSK

- Wright, T.; Gillet, V. J.; Green, D. V. S.; Pickett, S. D. Optimizing the Size and Configuration of Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* (2003), 43(2), 381-390.
- Nicolotti, O.; Gillet, V. J.; Fleming, P. J.; Green, D. V. S. Multiobjective Optimization in Quantitative Structure-Activity Relationships: Deriving Accurate and Interpretable QSARs. *J. Med. Chem.* (2002), 45(23), 5069-5080.
- Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* (2003), 43(2), 338-345.
- Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des.* (2002), 16(8/9), 653-681.
- Gillet, V. J.; Willett, P.; Fleming, P. J.; Green, D. V. S. Designing focused libraries using MoSELECT. *J. Mol. Graph. Model.* (2002), 20(6), 491-498.
- Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. S. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* (2002), 42(2), 375-385.
- Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Persp. Drug Disc. Des.* (2000), 20 1-16.
- Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Further development of a genetic algorithm for ligand docking and its application to screening combinatorial libraries. *ACS Symp. Ser.* (1999), 719(Rational Drug Design), 271-291.

GW

- Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci.* (1999), 39(1), 169-177.
- Gillet, V. J.; Willett, P.; Bradshaw, J. Identification Of Biological Activity Profiles Using Substructural Analysis And Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* (1998), 38(2), 165-179
- Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* (1997), 37(4), 731-740.
- Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* (1997), 267(3), 727-748.
- Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* (1995), 9(6), 532-49.
- Jones, G.; Willett, P.; Glen, Robert C. Molecular recognition of a receptor sites using a genetic algorithm with a description of its application. *J. Mol. Biol.* (1995), 245(1), 43-53.

Wellcome

- R. D.; Jones, G.; Willett, P.; Glen, R. C. Matching two-dimensional chemical graphs using genetic algorithms. *J. Chem. Inf. Comput. Sci.* (1994), 34(1), 63-70
- Clark, D. E.; Jones, G.; Willett, P.; Kenny, P. W.; Glen, R. C. Pharmacophoric pattern matching in files of three-dimensional chemical structures: Comparison of conformational-searching algorithms for flexible searching. *J. Chem. Inf. Comput. Sci.* (1994), 34(1), 197-206.

Driving Organisational Change!