

---

# Selecting Combinatorial Libraries to Optimise Diversity and Physical Properties

Val Gillet  
University of Sheffield

# Overview

---

- ◆ Scoring schemes for HTS
- ◆ Selecting structurally diverse combinatorial libraries
  - reactant vs product based selection
  - selecting diverse combinatorial libraries from product space
  - selection of libraries with optimised diversity and physical property profiles

# Scoring Schemes for HTS

---

- ◆ Scoring schemes (bioactivity profiles) have been developed to discriminate between two classes of compounds
  - eg. actives and inactives
  - based on ideas developed in SAR methods
  - compounds are ranked according to their likelihood of being active
  - uses generalised descriptors/features
- ◆ Scoring schemes can be used to:
  - determine the order in which compounds should be screened

# Data

---

- ◆ WDI: 30 K compounds
  - removed molecules labelled as trial-prep; pesticides; plant hormones; zootoxins; toxins; surfactants; diagnostics; chelators; absorbents
- ◆ SPRESI: 1.7 million compounds
  - removed molecules that occur in WDI
- ◆ Clean-up
  - charges neutralised; parent structures only; duplicates removed
- ◆ Filtering
  - molecules with uncommon elements removed
    - » C, N, O, F, P, S, Cl, Br and I only
  - restricted to:  $100 < MW < 1000$
- ◆ 14K WDI; 170K subset of SPRESI

# Generalised Features

---

- ◆ Features must be
  - easy to calculate for large numbers of compounds, e.g., 2D substructural features; physicochemical properties
  - relevant to biological activity

Feature	Description	Source
HBD	H Bond Donors	SMARTS
HBA	H Bond Acceptors	SMARTS
RB	Rotatable Bonds	SMARTS
AR	Aromatic Rings	Daylight Tools
MW	Molecular Weight	Daylight Tools
$^2K\alpha$	Kier Shape Index	Own Program
ClogP	Calculated LogP	CLogP

# Representing the Features

---

- ◆ Each feature is divided into a series of bins
- ◆ Each bin represents a given value (or range of values) of the feature
  - counts, eg. HBDs
  - measured property, eg, MW
- ◆ Weights are assigned to the feature bins

HBD weights					
0	1	2	3	.....	>19
$w_0$	$w_1$	$w_2$	$w_3$	.....	$w_{19}$

- ◆ A molecule is scored by:
  - calculating the value of each feature
  - summing the appropriate weights over all features

# Calculating the Weights

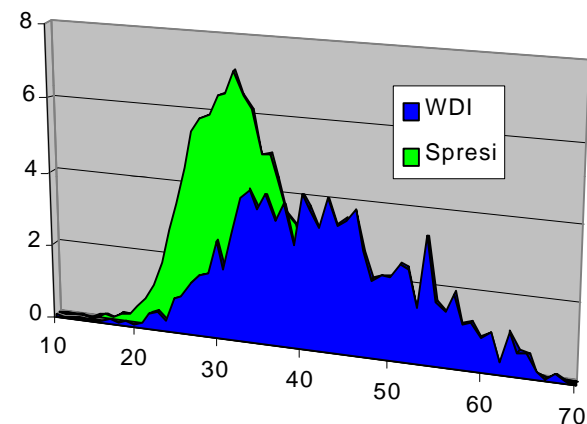
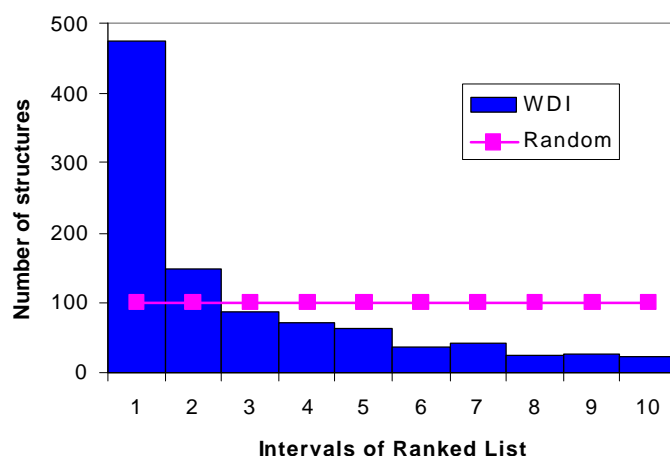
---

- ◆ Genetic Algorithm is used to derive optimum weights
- ◆ Population of chromosomes:
  - each chromosome represents a set of weights
  - weights are initially assigned random values
- ◆ Training set made up of two classes of molecules
- ◆ Fitness function of the GA
  - score all molecules in the training set according to weights in chromosome
  - rank molecules according to score
  - calculate the average rank for the molecules in one class
- ◆ GA optimises the discrimination between two classes of compounds by minimising the average rank

# Discriminating WDI from SPRESI

---

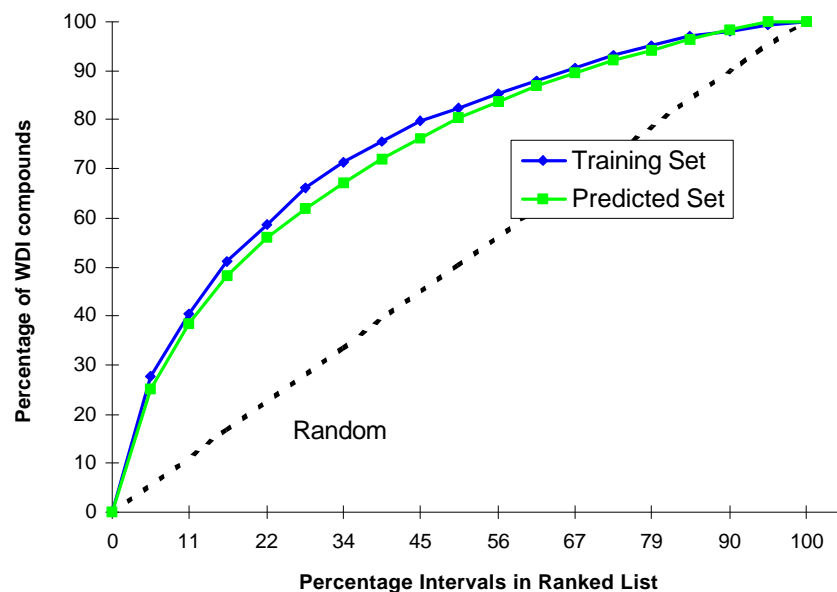
- ◆ 1 000 WDI; 16 661 SPRESI
- ◆ By chance, expect 57 WDI compounds to rank in top 1000
- ◆ Actually find 331 WDI in top 1000: 5.8 fold increase on chance
- ◆ 50% of WDI occur in the top 11% of the list



# Predictive Ability

---

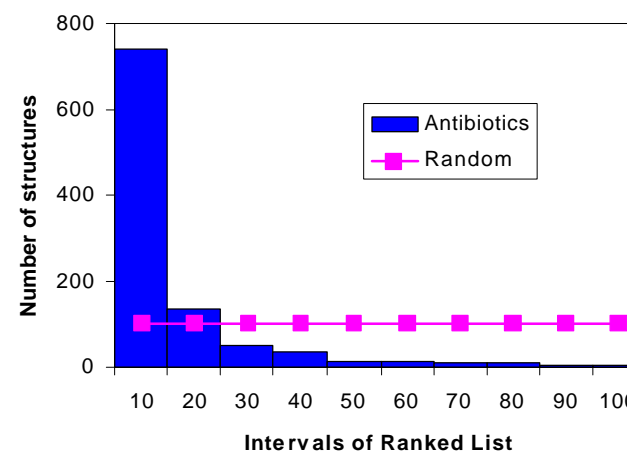
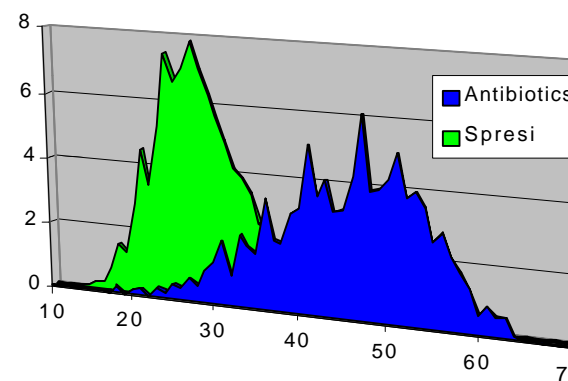
- ◆ Training set
  - 1000 WDI; 16661 SPRESI
- ◆ Predicted set
  - 10000 WDI; 166610 SPRESI



# Discriminating Antibiotics from SPRESI

---

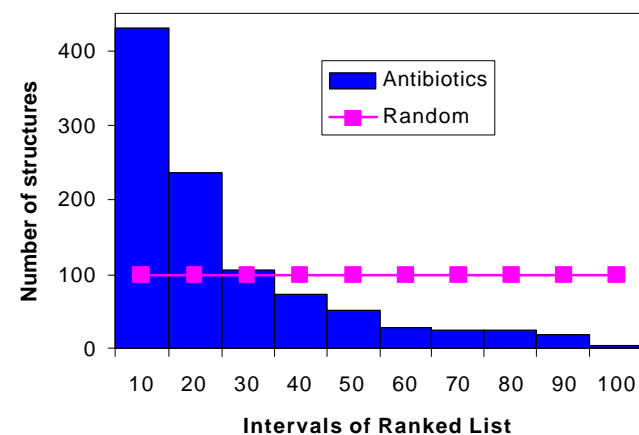
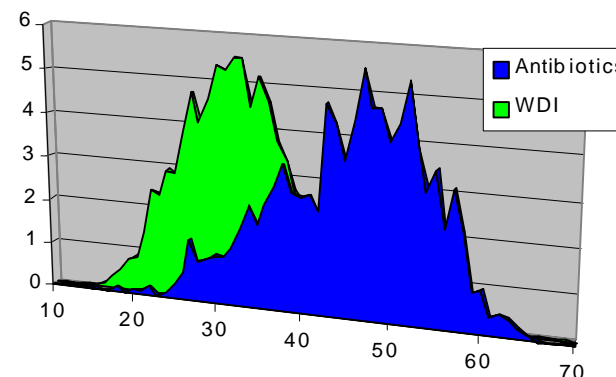
- ◆ 1 000 Antibiotics; 16661 SPRESI
- ◆ 50% of the antibiotics occur in the top 4% of the list



# Discriminating Antibiotics from rest of WDI

---

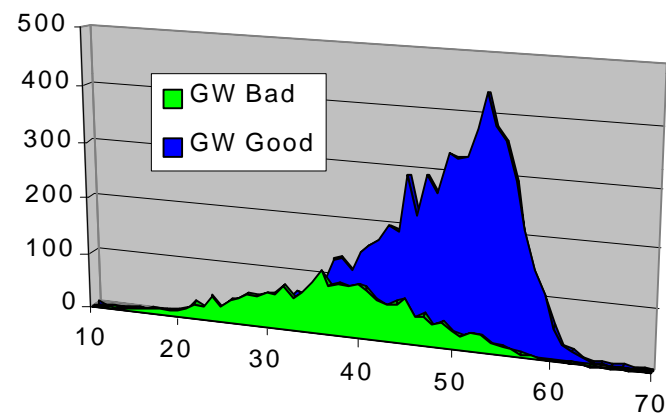
- ◆ GA trained to discriminate antibiotics from other classes of bioactive molecules
- ◆ 1 000 Antibiotics; 11910 WDI
- ◆ 50% of the antibiotics occur in the top 12% of the list



# GlaxoWellcome Data

---

- ◆ 8216 diverse compounds
- ◆ Manually labelled by GW chemists as:
  - 6 195 Good; 2 021 Bad
- ◆ Ranking reflects “drug-like” character in agreement with chemists’ intuition
- ◆ Some molecules in overlapping region identified as possible misclassifications



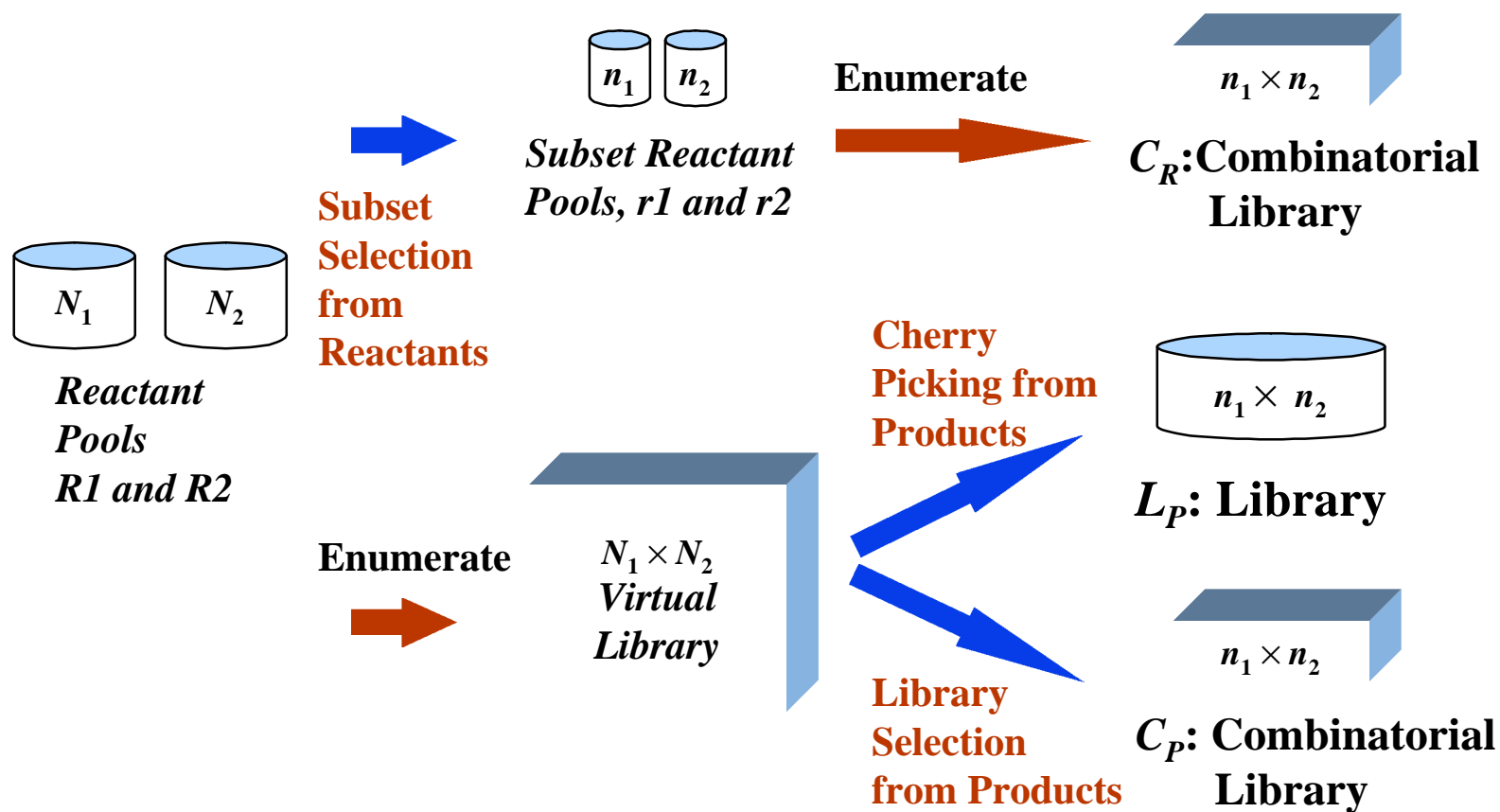
# Conclusions for HTS

---

- ◆ Simple structural features provide surprisingly good discrimination between drugs and non-drugs
- ◆ The GA provides a flexible approach
- ◆ Scoring schemes can be applied efficiently to very large collections of molecules
- ◆ Can be used to screen molecules in rank order to increase the likelihood of finding actives

# Combinatorial Library Design

## ◆ Product vs Reactant space selection



# Compound Selection

---

- ◆ Reactant space
- ◆ Cherry picking in product space

	y1	y2	y3	y4	y5
x1	x1 y1	x1 y2	x1 y3	x1 y4	x1 y5
x2	x2 y1	x2 y2	x2 y3	x2 y4	x2 y5
x3	x3 y1	x3 y2	x3 y3	x3 y4	x3 y5
x4	x4 y1	x4 y2	x4 y3	x4 y4	x4 y5
x5	x5 y1	x5 y2	x5 y3	x5 y4	x5 y5

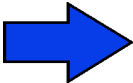
- DBCS
- diversity is measured as the normalised sum of pairwise dissimilarities using the cosine coefficient
- 1024 Daylight fingerprints as structural descriptors

# Selection in Product Space

---

- ◆ Cherry-picking in product space
  - synthetically inefficient
- ◆ Selecting a combinatorial library in product space
  - a library can be selected by intersecting rows and columns
  - exploring all combinatorial libraries is equivalent to permuting the rows and columns of the matrix

	y1	y2	y3	y4	y5					
x1	x1 y1	x1 y2	x1 y3	x1 y4	x1 y5					
x2	x2 y1	x2 y2	x2 y3	x2 y4	x2 y5					
x3	x3 y1	x3 y2	x3 y3	x3 y4	x3 y5					
x4	x4 y1	x4 y2	x4 y3	x4 y4	x4 y5					
x5	x5 y1	x5 y2	x5 y3	x5 y4	x5 y5					



	y2	y5	y1	y3	y4					
x1	x1 y2	x1 y5	x1 y1	x1 y3	x1 y4					
x4	x4 y2	x4 y5	x4 y1	x4 y3	x4 y4					
x2	x2 y2	x2 y5	x2 y1	x2 y3	x2 y4					
x3	x3 y2	x3 y5	x3 y1	x3 y3	x3 y4					
x5	x5 y2	x5 y5	x5 y1	x5 y3	x5 y4					

# Selecting CLs using a GA

---

- ◆ Permuting all possible libraries represents an enormous search space
- ◆ Chromosome encoding



- each chromosome represents a **combinatorial library**
  - one partition for each reactant pool
  - the number of genes in a partition equals the number of reactants required from the corresponding pool
- ◆ Genetic operators are used to evolve new potential solutions
  - ◆ Fitness function used to judge the value of a potential solution

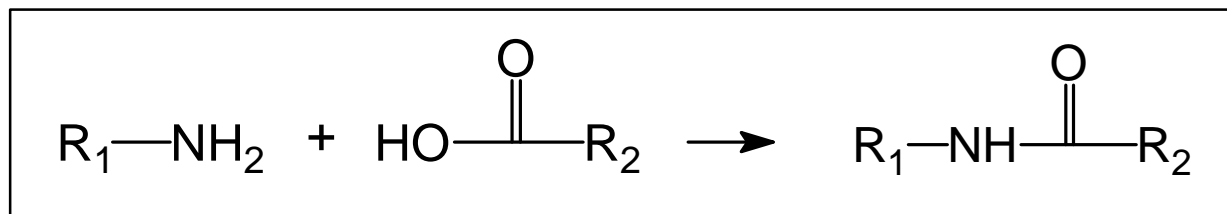
# Fitness Function

---

- ◆ Evaluating the “goodness” of a chromosome
- ◆ Chromosome decoding
  - enumerating the library
- ◆ Measuring the diversity of the library
  - sum of pairwise dissimilarities
    - » 1024 Daylight fingerprints as structural descriptors
    - » cosine coefficient
- ◆ Fitness function is applied frequently during the GA
  - efficiency is important

# Amide Library

---



- ◆  $C_P$ : select  $40 \times 40$  libraries from a  $400 \times 400$  amide library

Random	$C_R$	$C_P$	$L_p$
0.510 (0.004)	0.594	0.623	0.652

- ◆ Product based selection results in a more diverse library than reactant based selection

# Designing Diverse and “Drug-like” Libraries

---

- ◆ Structural diversity does not imply good coverage of biological activity
- ◆ A GA allows multiple objectives to be optimised simultaneously (SELECT)
  - diversity
  - physical properties for “drug-like” libraries
  - distance from existing libraries
- ◆ Niching can be used to find “good” solutions in different parts of the search space
- ◆ Memory requirements:
  - 152 + 4 bytes per physical property
  - $10^6$  virtual library requires ~160 Mbytes

# Multicomponent Fitness Function

---

$$f(n) = w_D(1 - D) + w_C(1 - C) + w_{p_1} \Delta p_1 + w_{p_2} \Delta p_2 + \dots$$

$$w_D(1 - D)$$

◆ Diversity term

$$w_C(1 - C)$$

◆ Combined diversity term

$$w_{p_1} \Delta p_1 + w_{p_2} \Delta p_2 + \dots$$

◆ Physical property terms

$$w_D, w_C, w_{f1}, \dots$$

◆ User defined weights

# Diversity Measure: $w_D(1 - D)$

---

- ◆ Diversity is measured as either:
  - the **sum of pairwise dissimilarities**,  $D_{SUM}$ , calculated from the centroid,  $A_c$ , of library  $A$ , consisting of  $N(A)$  compounds:
    - » the centroid of a library is a weighted vector

$$D_{SUM}(A) = 1 - \frac{DOTPROD(A_c, A_c)}{N(A)^2}$$

- » time complexity  $O(N)$
- the **average nearest neighbour distance**,  $D_{NN}$ 
  - » time complexity  $O(N^2)$

# Library Comparisons: $w_C(1 - C)$

---

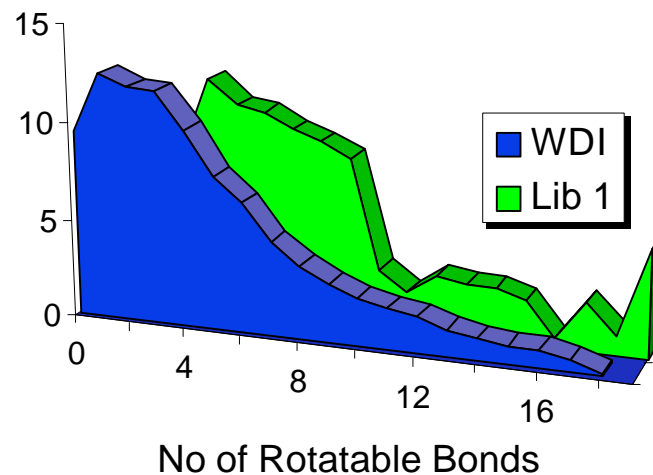
- ◆ A library can be optimised to complement an existing collection
- ◆ The diversity of the combined library,  $AX$ , can be measured efficiently using the centroids of  $A$  and  $X$ :

$$C = D(AX) = 1 - \left( \frac{DOTPROD(A_c, A_c) + DOTPROD(X_c, X_c) + 2 \times DOTPROD(A_c, X_c)}{(N(A) + N(X))^2} \right)$$

# Property Profiles: $w_{p_1} \Delta p_1 + w_{p_2} \Delta p_2 + \dots$

---

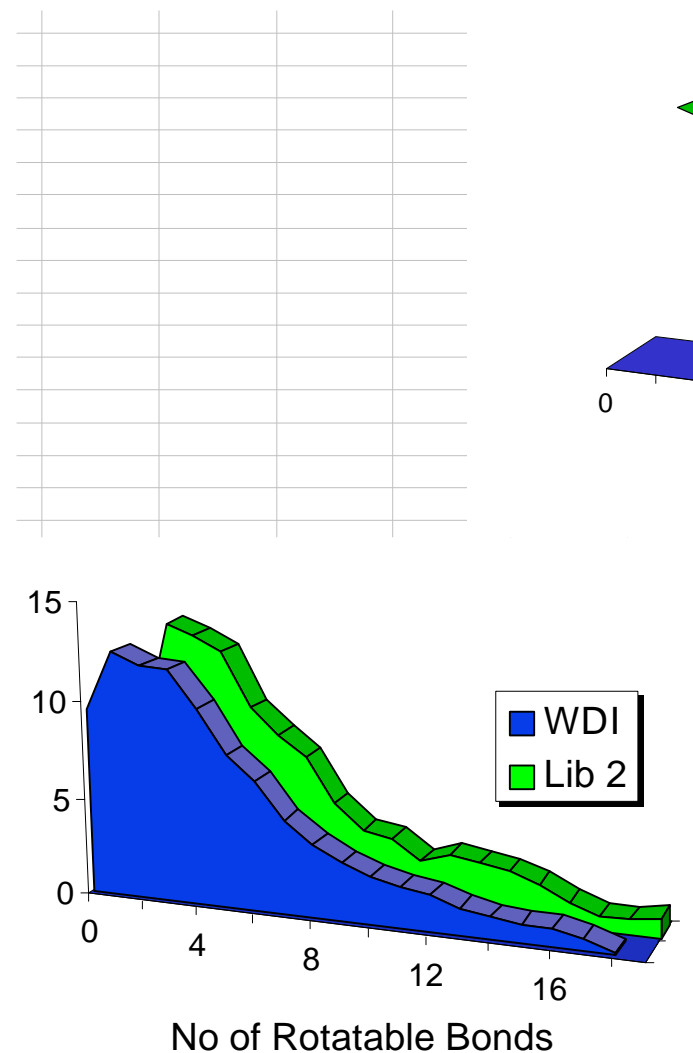
- ◆ The distribution of properties within the library are compared with a reference collection
  - e.g., “drug-like” distribution derived from WDI
- ◆ The RMS difference between the distribution in the library and the reference distribution is minimised



# Amide Library

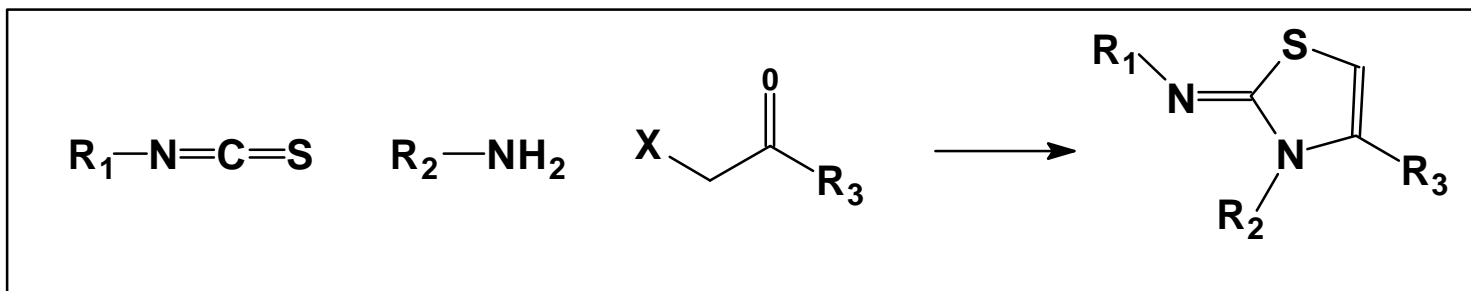
---

- ◆  $20 \times 20$  library selected from  $100 \times 100$  virtual library
- ◆ Lib 1 optimised on
  - diversity alone
  - $D = 0.595$ ;  $\Delta_{\text{RB}} = 0.381$
- ◆ Lib 2 optimised on
  - diversity and rotatable bond profile
  - $D = 0.574$ ;  $\Delta_{\text{RB}} = 0.170$
- ◆ 7.2 mins SG R10K 195MHz
- ◆ 1.56 Mbytes memory



# Thiazoline-2-Imine Library I

---



- ◆  $8 \times 40 \times 20$  (6400) library selected from  $12 \times 99 \times 54$  (70092) virtual library
- ◆ SELECT was run to choose libraries optimised on
  - diversity alone
  - diversity and rotatable bond profile
  - diversity and Andrews' Binding Energy profile
- ◆ 1.4 hrs SG R10K 195Mhz      11.2 Mbytes memory

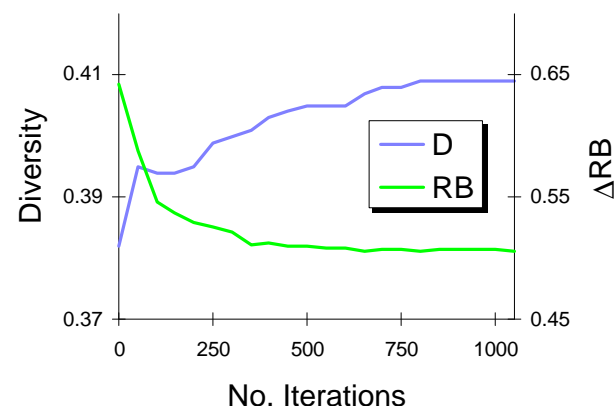
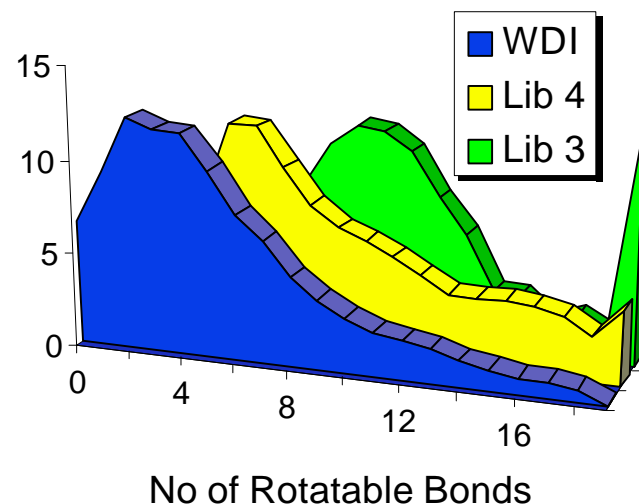
# Thiazoline-2-Imine Library II

- ◆ Lib 3 optimised on diversity alone

- diversity = 0.427
- $\Delta_{RB} = 0.701$

- ◆ Lib 4 optimised on diversity and rotatable bond profile

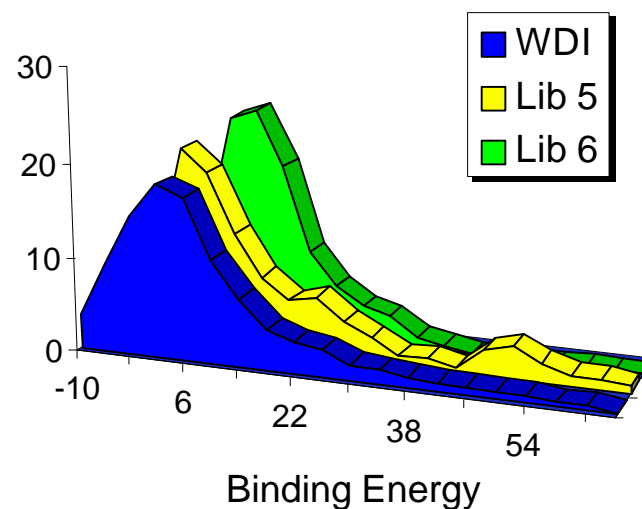
- diversity = 0.404
- $\Delta_{RB} = 0.507$



# Thiazoline-2-Imine Library III

---

- ◆ Lib 5 optimised on diversity alone
  - diversity = 0.427
  - $\Delta_{\text{BE}} = 0.603$
- ◆ Lib 6 optimised on diversity and Andrew's binding energies
  - diversity = 0.416
  - $\Delta_{\text{BE}} = 0.297$



# Library Configurations

---

- ◆ SELECT can be used to determine an optimal configuration for a library
- ◆ Thiazoline-2-Imine libraries of 2880 compounds
  - pool 1 : isothiocyanates
  - pool 2: amines
  - pool 3: haloketones
- ◆ Maximum diversity is achieved with
  - at least 4 isothiocyanates

Pool 1	Pool 2	Pool 3	<i>D</i>
1	72	40	0.386
2	72	20	0.405
2	36	40	0.409
4	72	10	0.423
4	36	20	0.431
4	18	40	0.430
8	72	5	0.421
8	36	10	0.433
8	18	20	0.433
8	9	40	0.430

# Maximising the Distance Between Libraries

---

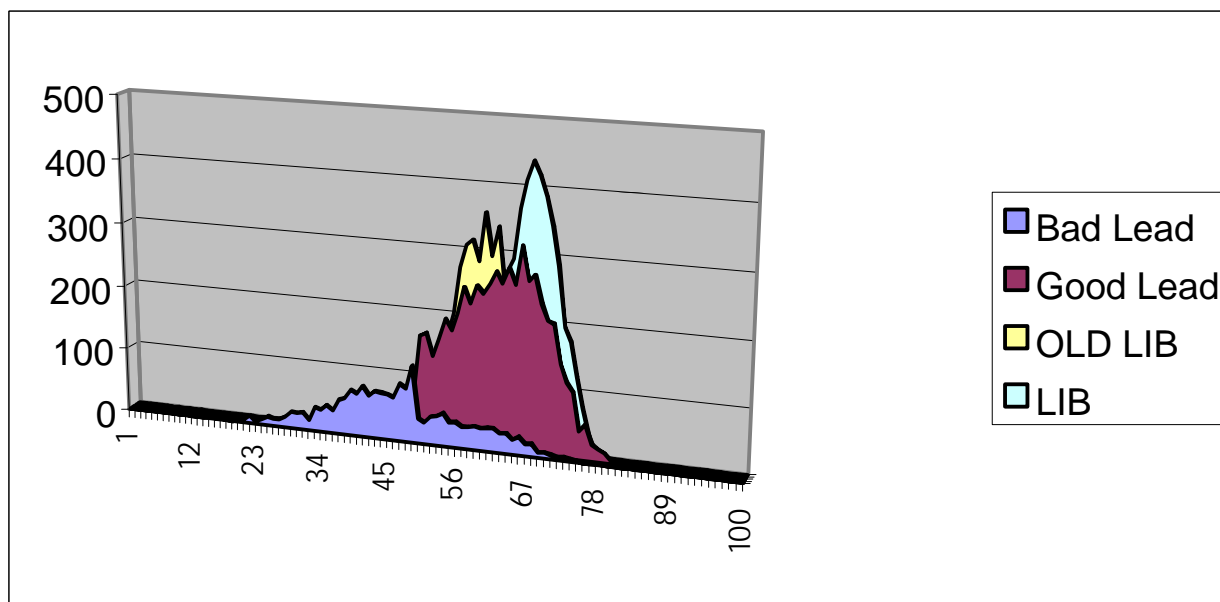
- $10 \times 10$  amide library **A** selected from a  $100 \times 100$  virtual library (10K)
- **B** selected to be internally diverse
- **C** selected to complement **A**

Library	$w_D$	$w_C$	$D_{SUM}$	overlap with <b>A</b> (%)
<b>B</b>	1.0	0.0	0.597	49
<b>C</b>	0.0	1.0	0.582	12

# GlaxoWellcome Data

---

- ◆ 800K virtual library
  - 3 components:  $100 \times 40 \times 200$
- ◆ SELECT
  - optimum configuration;
  - diverse and “drug-like” mw and rotatable bond profiles



# Conclusions

---

- ◆ Easy to calculate generalised molecular features can be used together with a GA to develop effective scoring schemes for HTS
- ◆ Product based selection results in better optimised libraries than reactant based selection
- ◆ SELECT provides an efficient and effective way of designing diverse and “drug-like” combinatorial libraries
- ◆ SELECT is currently being used to design libraries inhouse at GlaxoWellcome

# Acknowledgements

---

- ◆ Peter Willett
  - University of Sheffield, UK
- ◆ John Bradshaw and Darren Green
  - GlaxoWellcome, UK
- ◆ GlaxoWellcome for funding
- ◆ Daylight Chemical Information Systems Inc. for software