



The  
University  
Of  
Sheffield.

# Designing Libraries Optimised on Multiple Properties

Val Gillet

UKQSAR Oct 2007



# Outline

- Using Pareto Ranking
  - Optimise libraries on multiple properties
    - Predicted activity, diversity, ADME properties,....
    - Size, configuration, number of combinatorial subsets
  - Explore trade-offs in lead optimisation
- Using FW analysis to assess additivity in library design and its implications for lead optimisation



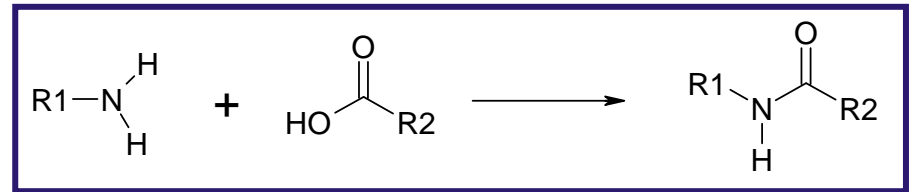
# Criteria for Library Design

- Library design is a multiobjective optimisation problem
  - Diverse or focused (similarity to a target; fit to a QSAR etc) or both, Cheap, Drug-like, good ADME properties, small, .....
- 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?

$$f = w_1 f_1 + w_2 f_2 + w_3 f_3 + \dots$$

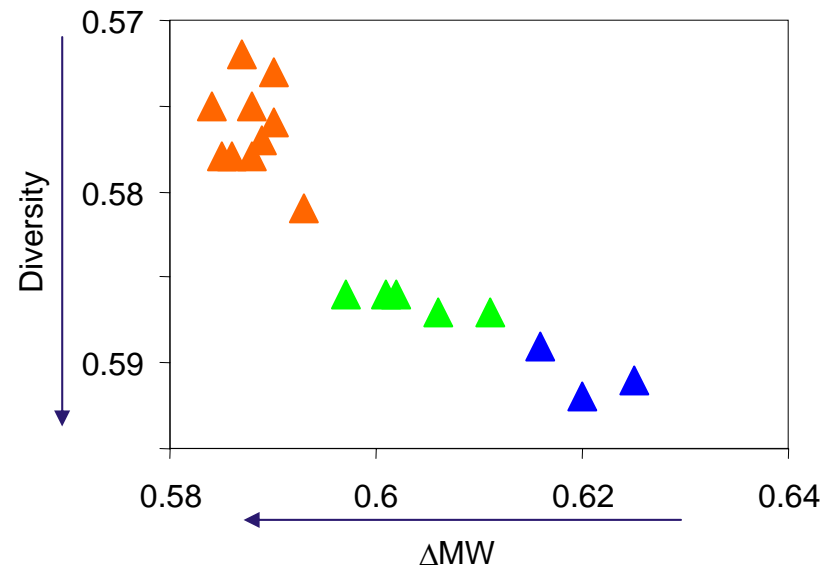
# Limitations of Weighted-Sum Approach

- 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)



$$f(n) = w_1 \cdot \text{diversity} + w_2 \cdot \Delta MW$$

▲  $w_1=1.0; w_2=1.0$  ▲  $w_1=1.0; w_2=0.5$  ▲  $w_1=10; w_2=1.0$



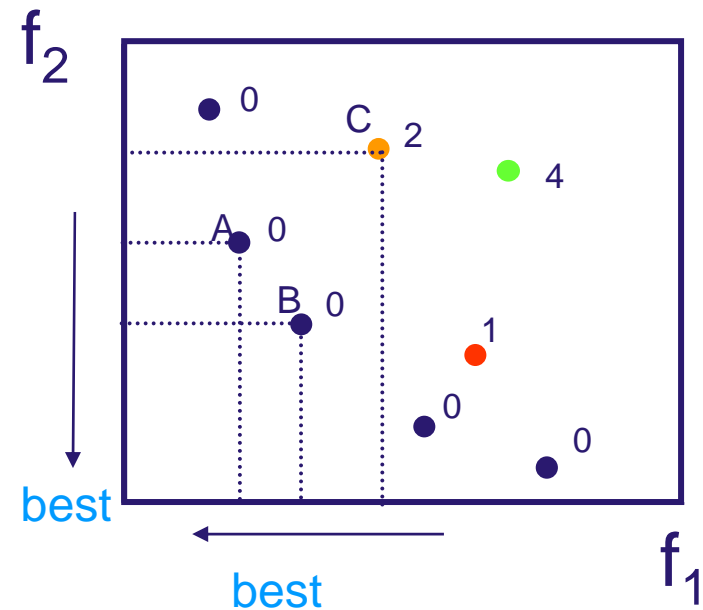


# Multiobjective Optimisation using a MOEA

- 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)
- MOGA for combinatorial library design
  - 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

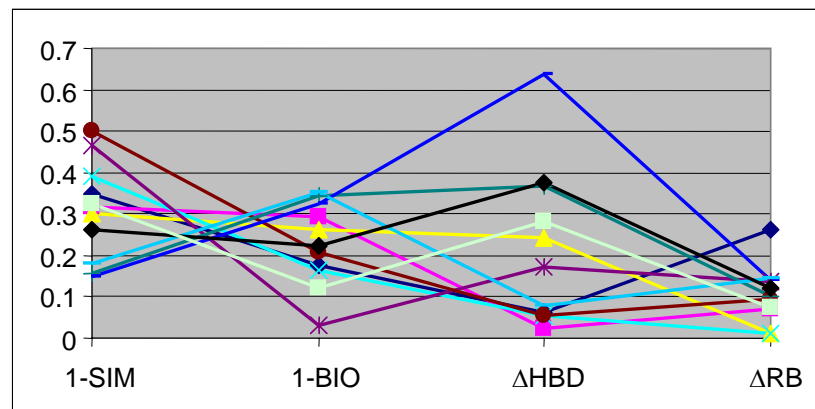
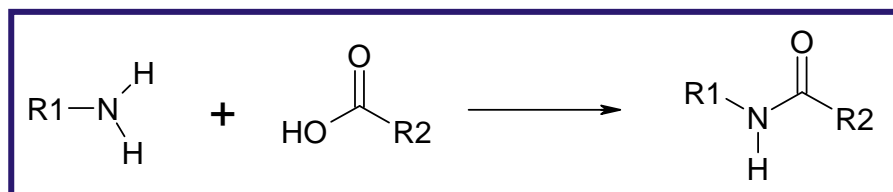
- Each objective is optimised independently
- One solution dominates another if it is better in both objectives
- Solutions are ranked according to dominance value
- Solutions where no other solutions are greater in all objectives are **non-dominated** and form the **Pareto frontier**





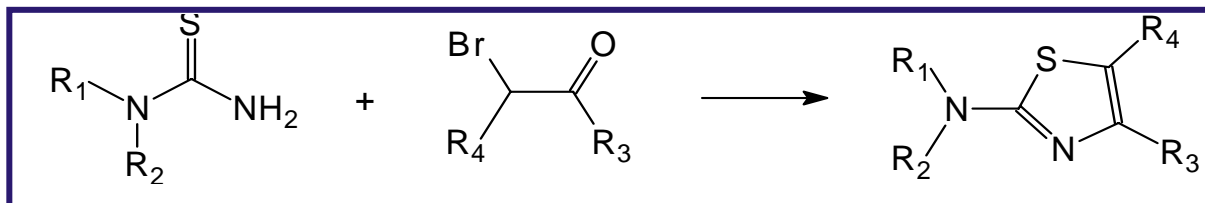
# Focused Libraries: Amides

- 100 × 100 virtual library
- MOGA 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





# Varying Library Size

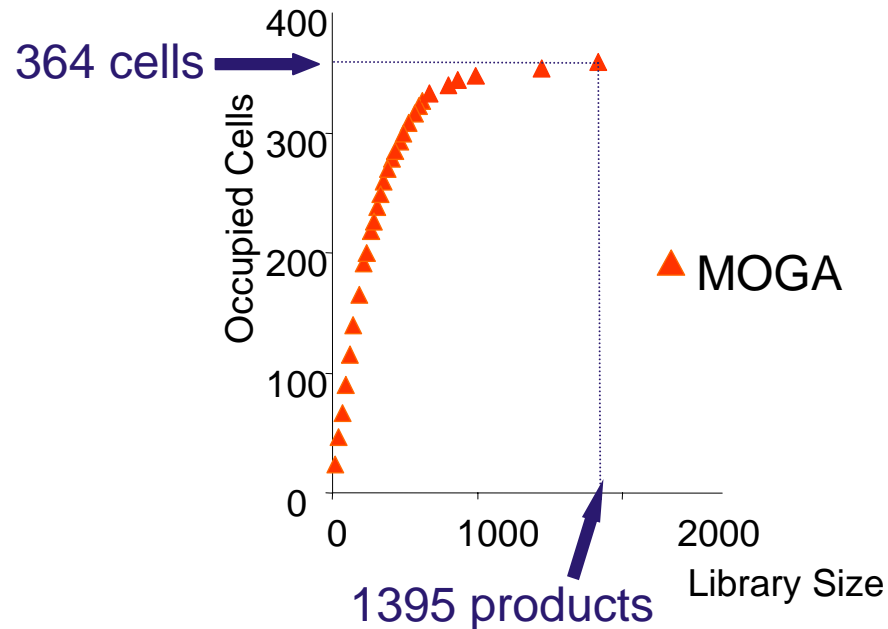


## Aminothiazole Library:

Virtual library of 12850 products

170 thioureas X 74 α-bromoketones

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



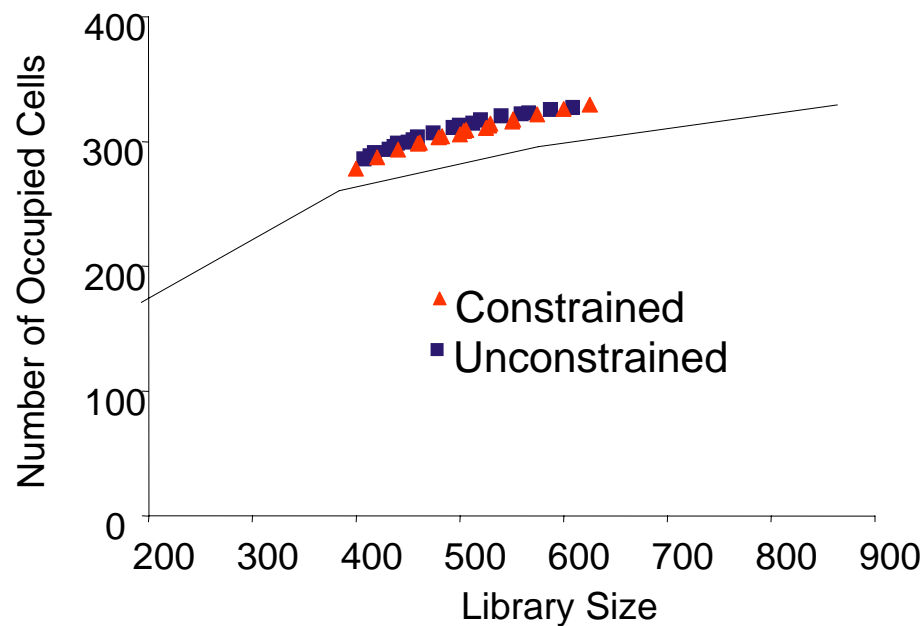
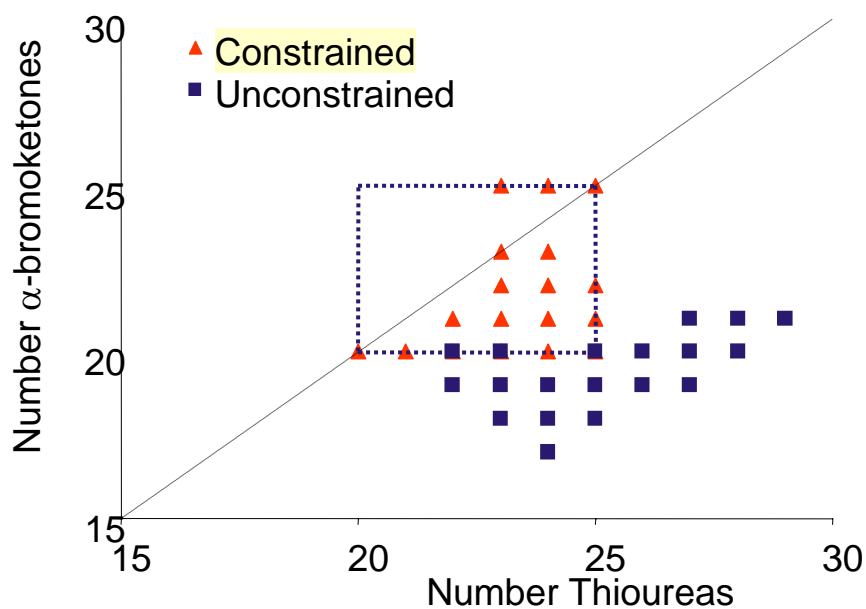


# Incorporating Constraints

Size constraint: 400 to 600 products

Combinatorial efficiency constraint:  $20 \leq \alpha\text{-bromoketones} \leq 25$

$20 \leq \text{thioureas} \leq 25$





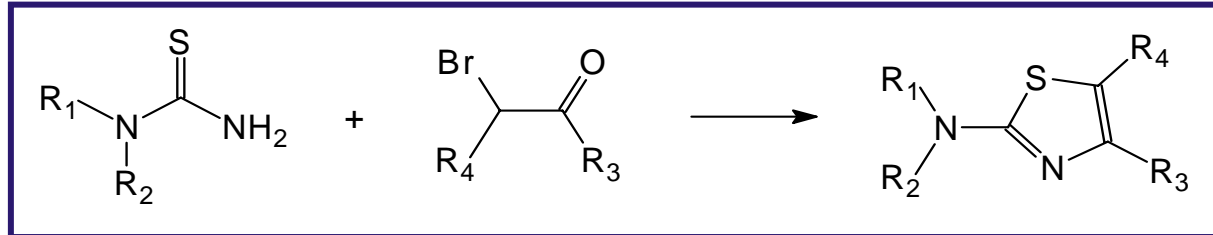
# Selecting Multiple Combinatorial Subsets

	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	A <sub>6</sub>
B <sub>1</sub>	■	■			■	■
B <sub>2</sub>						
B <sub>3</sub>	■	■			■	■
B <sub>4</sub>	■	■			■	■
B <sub>5</sub>						
B <sub>6</sub>	■	■			■	■
B <sub>7</sub>	■	■			■	■
B <sub>8</sub>	■	■			■	■
B <sub>9</sub>						
B <sub>10</sub>						

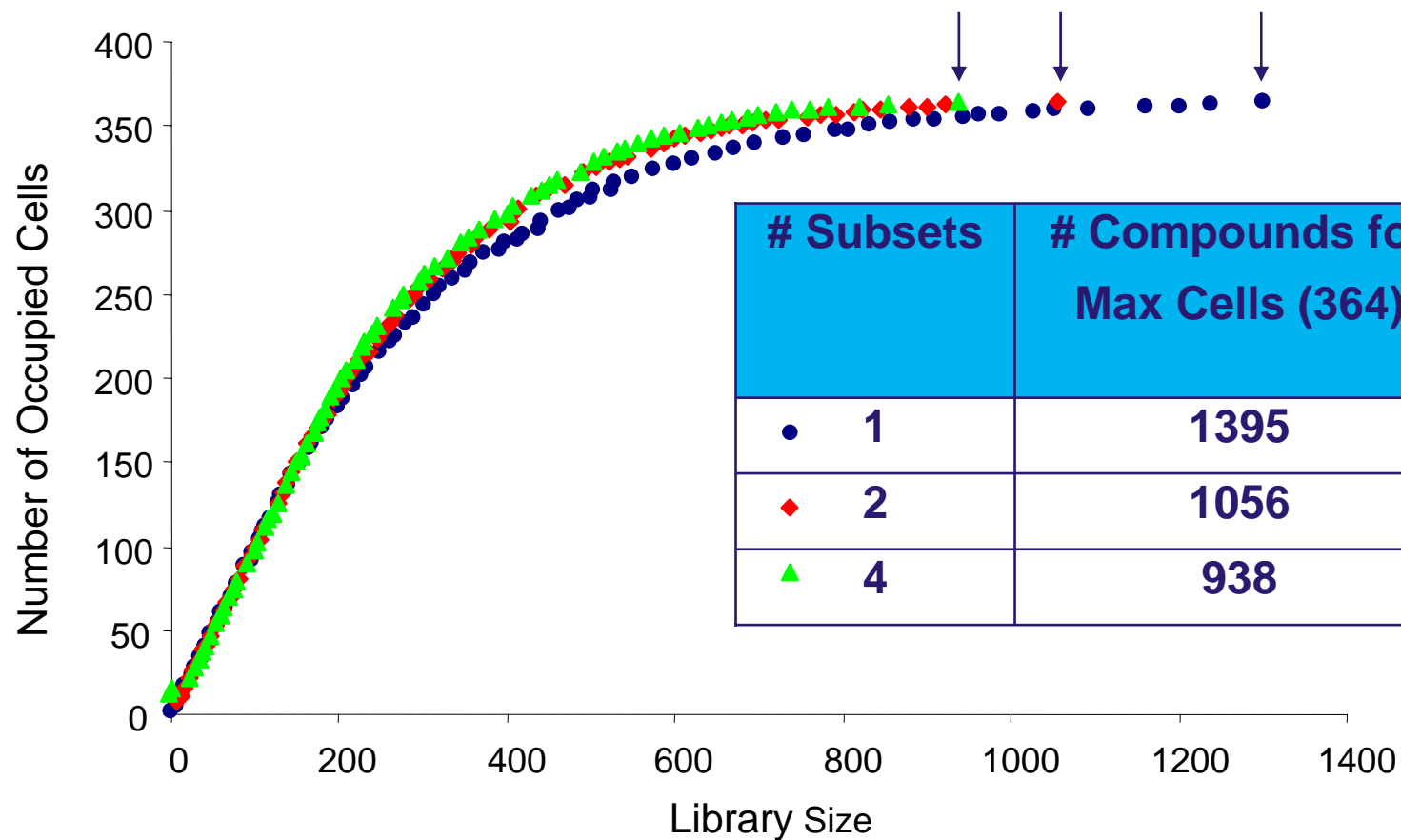
24 products constructed from one 4 × 6 subset

	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	A <sub>6</sub>
B <sub>1</sub>	■	■				
B <sub>2</sub>						
B <sub>3</sub>	■	■				
B <sub>4</sub>	■	■				
B <sub>5</sub>			■	■	■	■
B <sub>6</sub>			■	■	■	■
B <sub>7</sub>			■	■	■	■
B <sub>8</sub>	■	■				
B <sub>9</sub>	■	■				
B <sub>10</sub>	■	■				

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



# Aminothiazole Library





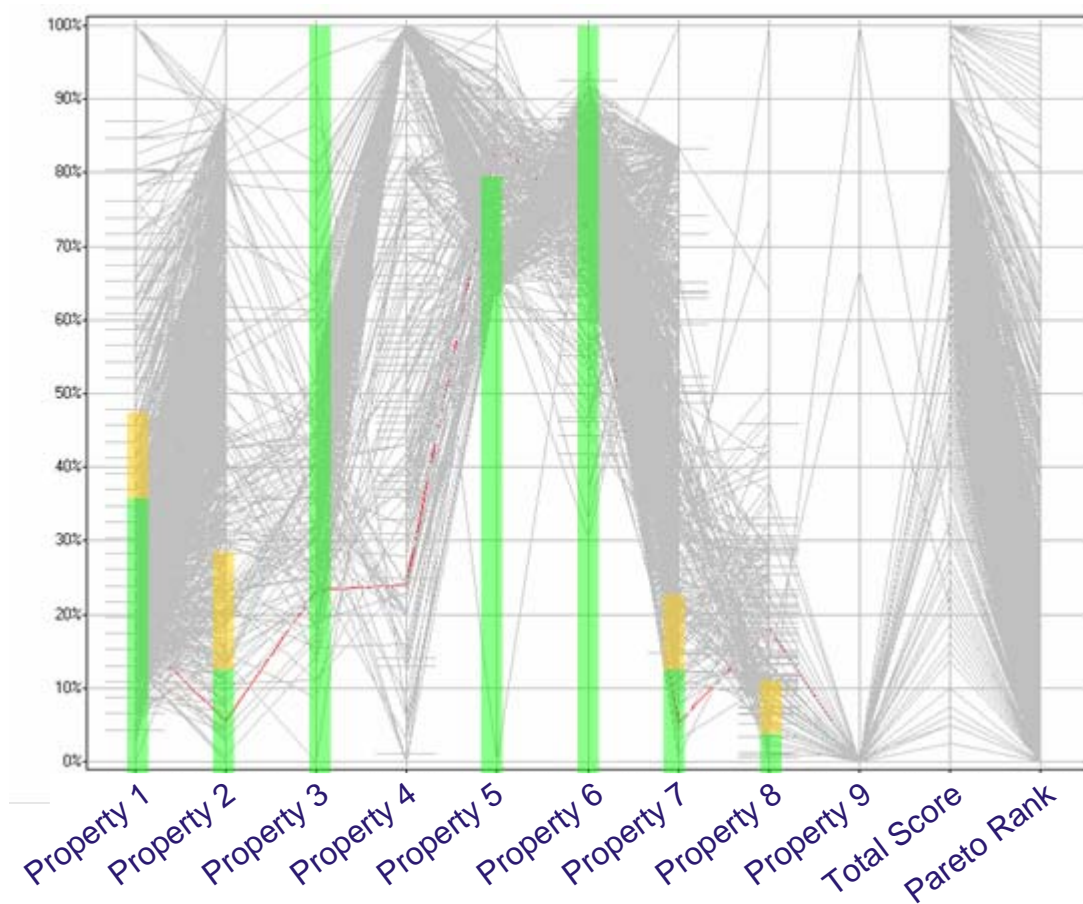
The  
University  
Of  
Sheffield.



# Using Pareto Ranking to Explore Trade-offs in Lead Optimisation

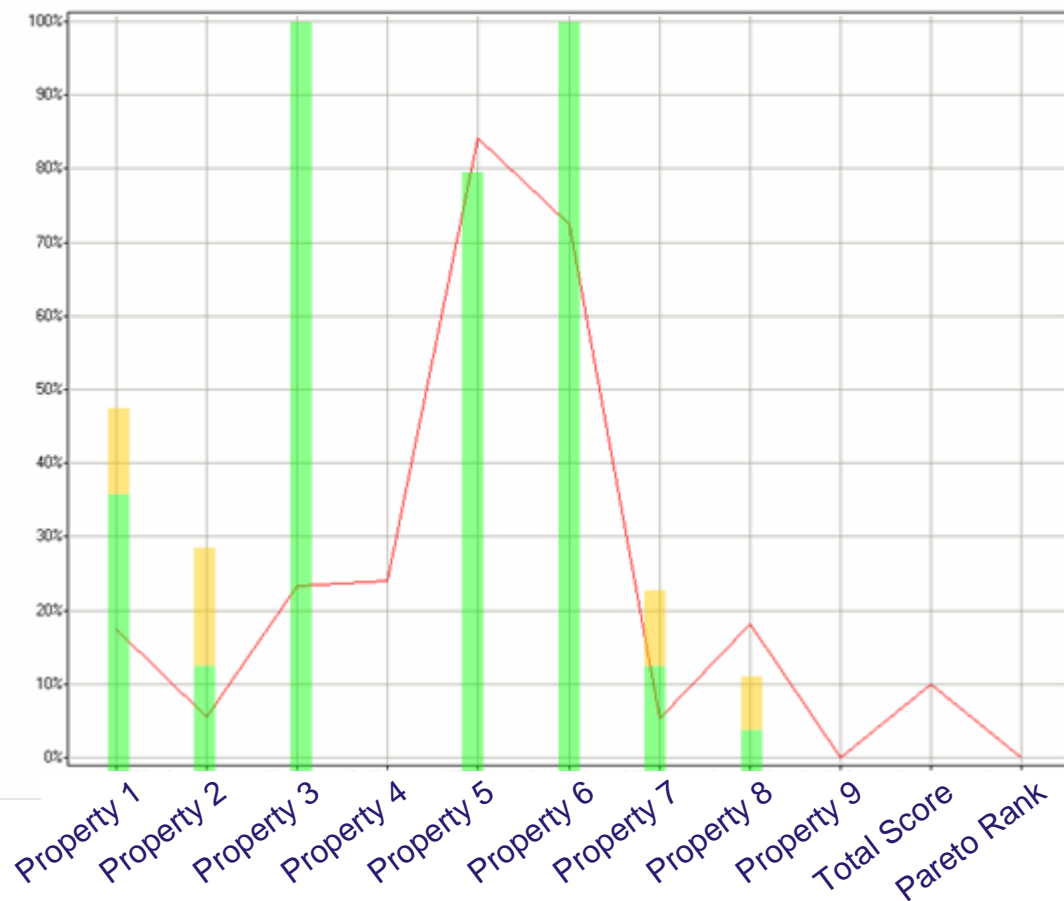


# Optimisation Profile



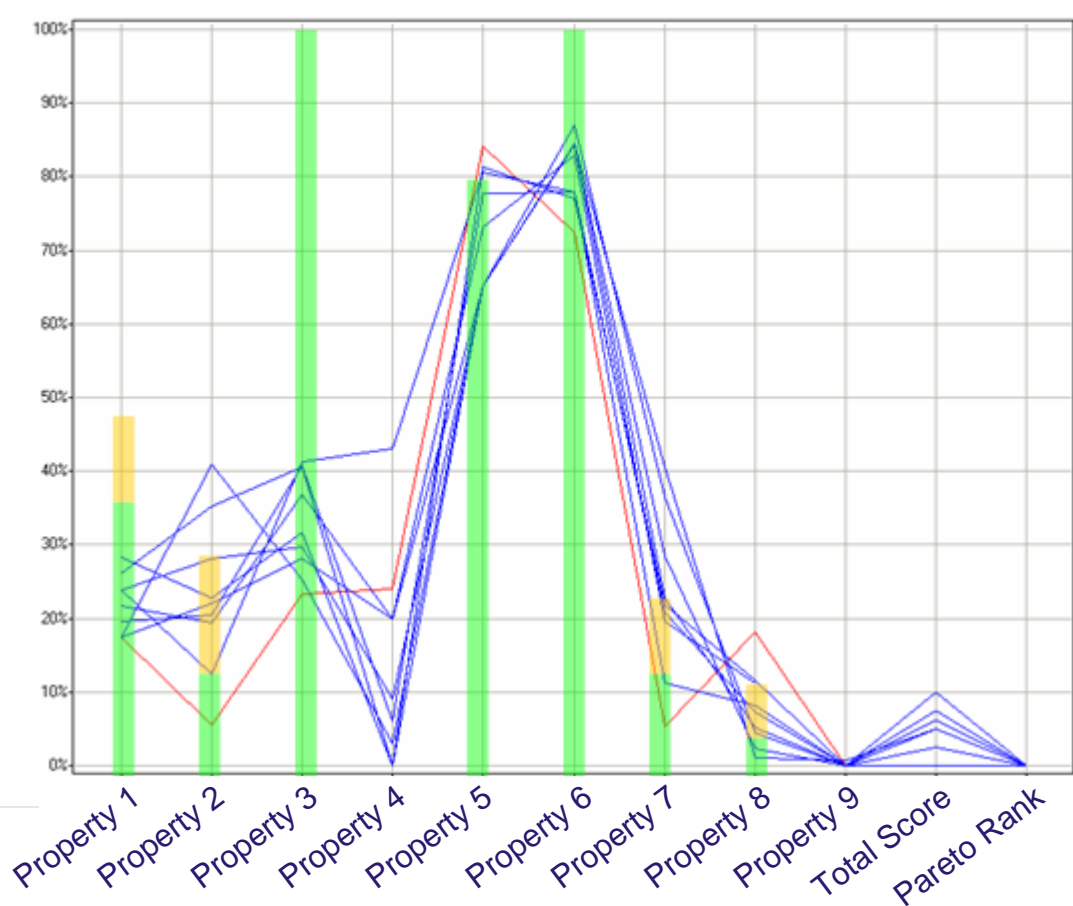


# Profile of Drug Candidate





# Other High Scoring Compounds





# Conclusions: Pareto Ranking

- The MOGA provides a very flexible tool for library design
- The arbitrary nature of weighted-sum approaches is avoided
- The user is presented with a number of libraries which represent different balances of the properties
- The user may then make an informed choice on an appropriate compromise
- Pareto ranking also provides a useful way of exploring trade-offs that may exist in LO projects.



# Assessing Additivity: FW Analysis

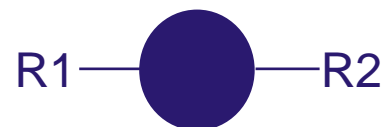
- Is it possible to identify additivity effects in (near) complete combinatorial libraries using FW analysis?
- Can we determine the minimum requirements (number of compounds, distribution of properties, etc) required to determine additivity based on retrospective experiments?
- Yogi Patel (Sheffield), Peter Willett
- Julen Oyarzabal\*, Trevor Howe (J&J)

(\*Spanish National Cancer Centre, [www.cnio.es](http://www.cnio.es))



# Free-Wilson Analysis

Molecule ID	x1	x2	.....	y1	y2	.....
1	1	0	.....	1	0	.....
2	0	1	.....	1	0	.....
3	1	0	.....	0	1	.....
4	0	1	.....	0	1	.....
.....	.....	.....	.....	.....	.....	.....



R1 = x1, x2,.....

R2 = y1, y2, ....

$$y = C + a_1x_1 + a_2x_2 + a_3x_3 + \dots + b_1y_1 + b_2y_2 + b_3y_3 + \dots$$

$$= C + \sum_{i=1}^n a_i x_i + \sum_{j=1}^m b_j y_j$$

$a_i$  and  $b_j$  are coefficients that represent the contribution made by each R-group to the activity of a compound;

$y$  is the property value under investigation (e.g. biological activity).



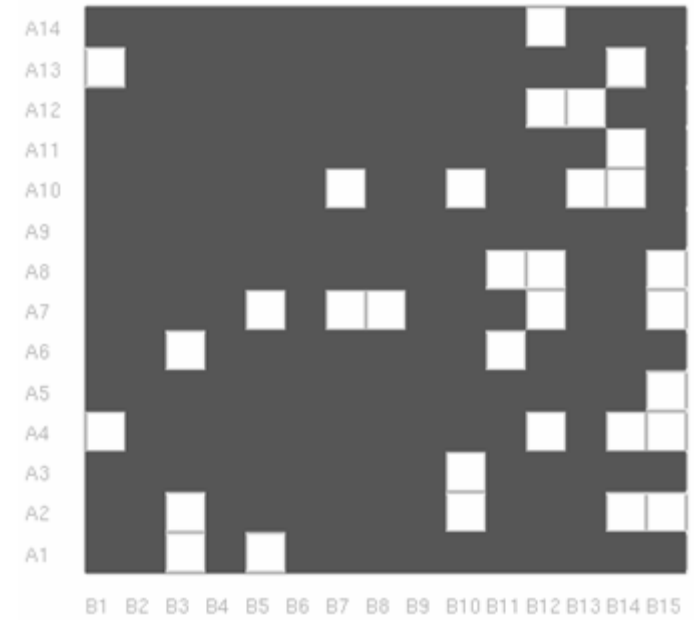
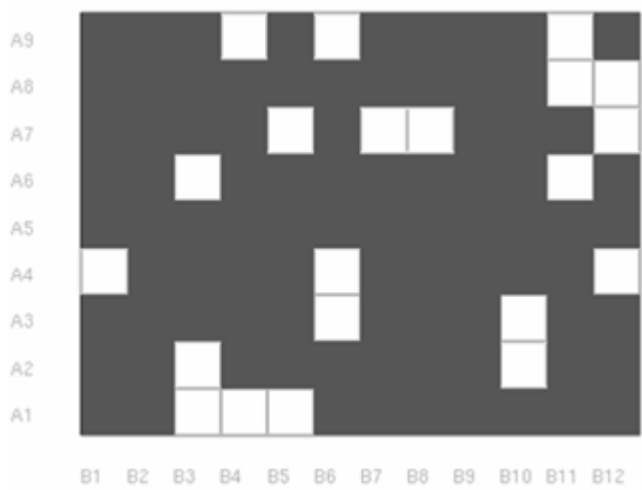
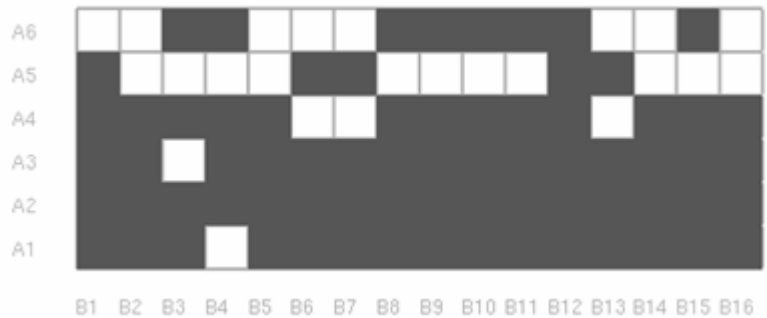
# Datasets

8 sets of compounds with 2-6 properties  
19 datasets in total

Dataset	Property	# R <sub>x</sub> Groups	# R <sub>y</sub> Groups	% complete dataset	Min prop value	Max prop Value
+4R 1_136	Potassium Channel	7	11	70.1	5.05	7.17
+4R 1_56	Sodium Channel	6	16	75.0	5.06	7.07
+4R 1_73	Class A GPCR	7	17	68.1	6.04	8.98
+D1 1_62	Class A GPCR	14	15	84.8	5.01	7.86
+D1 1_64	Class A GPCR	14	15	84.8	5.16	8.99
+D2 1_62	Class A GPCR	9	12	80.6	5.30	7.86
+D2 1_64	Class A GPCR	9	12	80.6	5.52	8.99
+D2 58_7	Class A GPCR	9	12	80.6	5.01	7.91
+D3 1_73	Class A GPCR	22	5	89.1	5.46	8.98
*D3 221_5	Cellular Metabolism	22	5	74.5	12.00	100.00
-D4 100_152	Ion Channel	13	7	75.8	5.03	7.98
-D5 100_193	Class A GPCR	7	11	70.1	5.07	8.79
#D6 76_97	Ser/thr Kinase	19	6	84.2	6.79	8.90
*D7 221_5	Class A GPCR	12	5	70.0	0.00	91.5
+D7 359_3	P450 3A4	12	5	66.7	12.01	85.26
+D7 359_5	P450 2C9	12	5	66.7	16.81	110.79
+D7 359_6	P450 2D6	12	5	66.7	11.35	54.23
+D7 359_7	P450 1A2	12	5	66.7	3.93	102.75
+D7 76_97	Ser/Thr Kinase	12	5	70.0	7.18	8.87



# Examples





# Design of Test and Training Sets

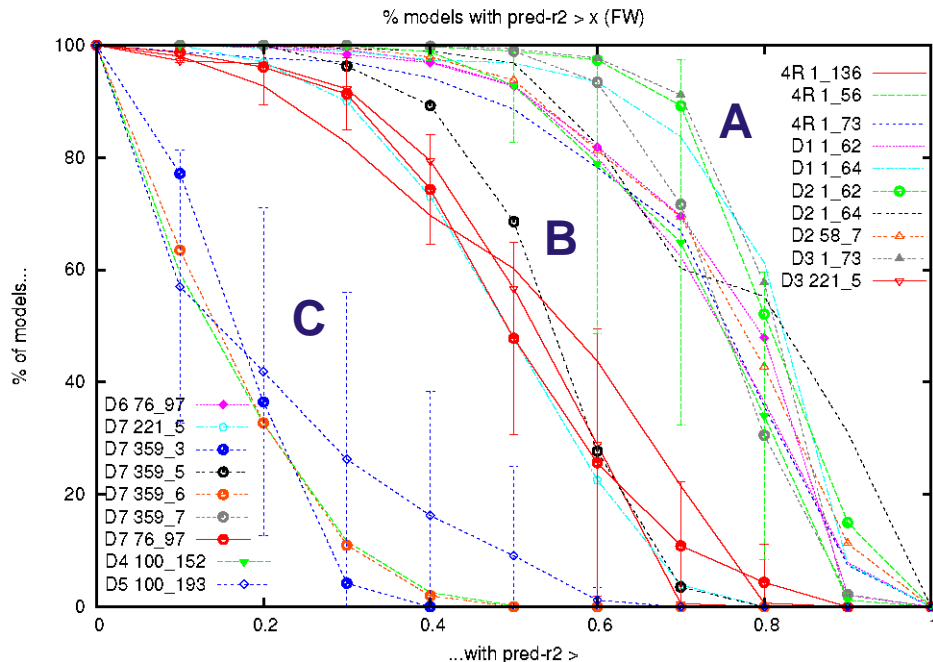
- Five test sets selected for each dataset
  - At least one occurrence of each Rx, Ry group
- For each test set
  - Training sets were selected with least one occurrence of each Rx, Ry
  - The following properties were then varied from min to max
    - Distribution of R groups (measured using Scaled Shannon Entropy)
    - Dataset size
    - Property range



# Design of Test and Training Sets

Dataset	# training sets per test set	Max Prop Range	Min Prop Range	Max $R_x$ SSE	Min $R_x$ SSE	Max $R_y$ SSE	Min $R_y$ SSE	Min % cmpds	Max % cmpds
4R_1_56	102	1.93	1.45	0.991	0.712	0.996	0.929	22	58
D1_1_62	77	2.85	1.74	0.998	0.870	0.997	0.929	14	71
D2_1_62	77	2.56	1.58	0.998	0.898	0.996	0.925	26	72
D3_1_73	62	3.52	2.59	0.999	0.921	0.999	0.924	39	68
D4_100_152	202	2.95	1.92	1.000	0.858	0.991	0.874	28	62
D5_100_193	87	3.72	2.26	0.999	0.725	0.993	0.921	27	56
D6_76_97	97	2.11	1.69	1.000	0.909	1.000	0.933	31	68
D7_359_3	92	70.79	67.11	0.986	0.928	0.998	0.871	38	47

# Estimating Additive/Non-additive: pred-r<sup>2</sup>

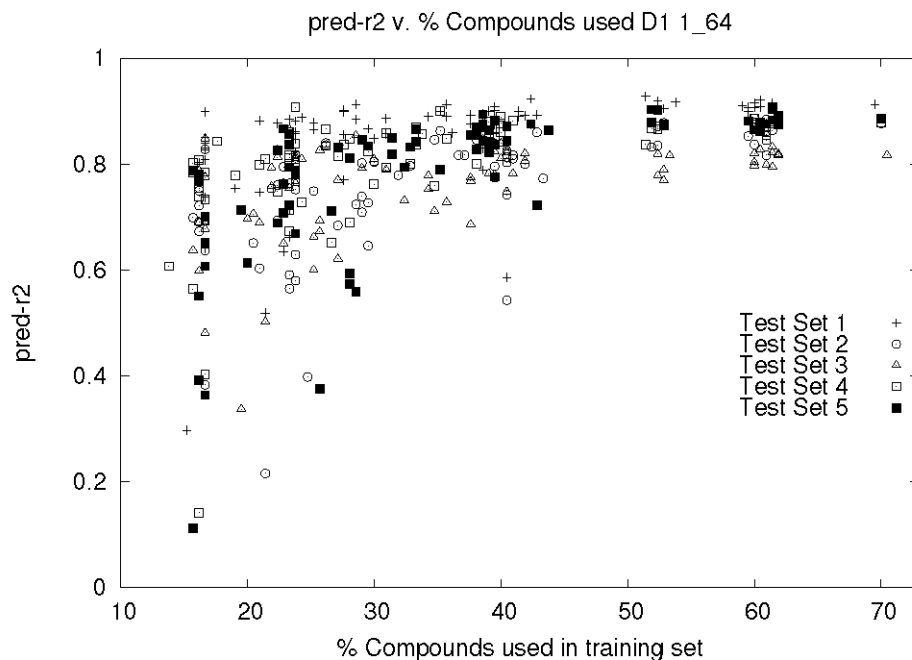


- Additive (10): > 50% models with pred-r<sup>2</sup> ≥ 0.7
- Non-additive (4): < 20% models with pred-r<sup>2</sup> ≥ 0.4
- Partially additive (5): < 50% models with pred-r<sup>2</sup> ≥ 0.6;  
and > 20% models with pred-r<sup>2</sup> ≥ 0.4



# Pred-r<sup>2</sup> and Training Set Size

## Additive Dataset

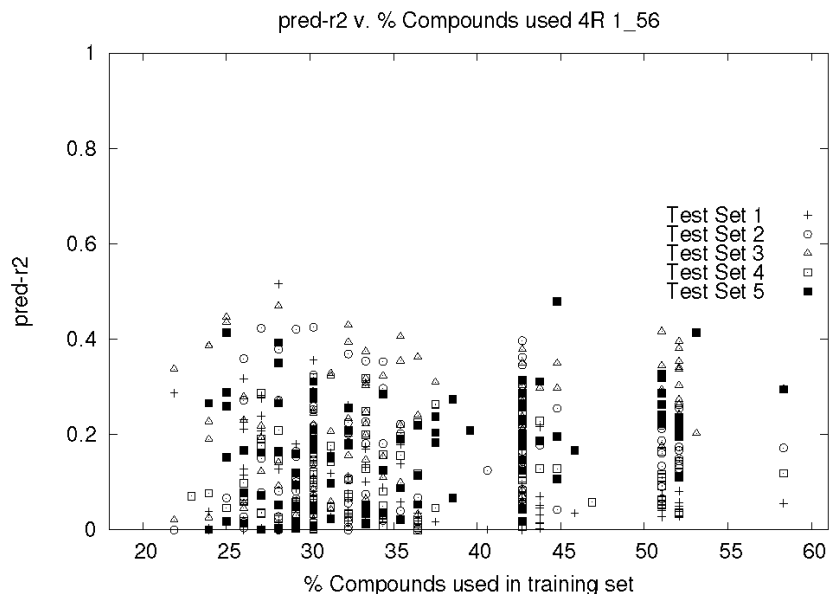


Models with high predictivity are found for 30% of dataset

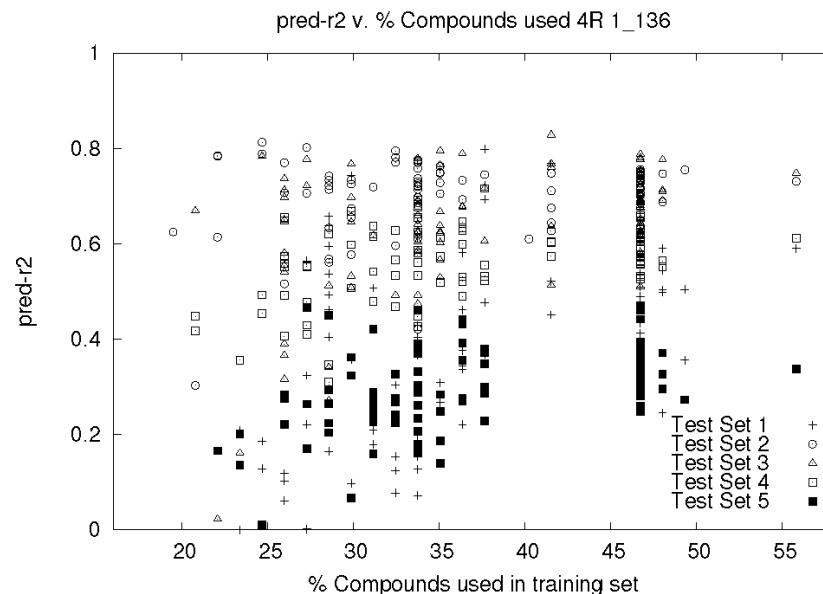


# Pred-r<sup>2</sup> and Training Set Size:

Non-Additive Dataset



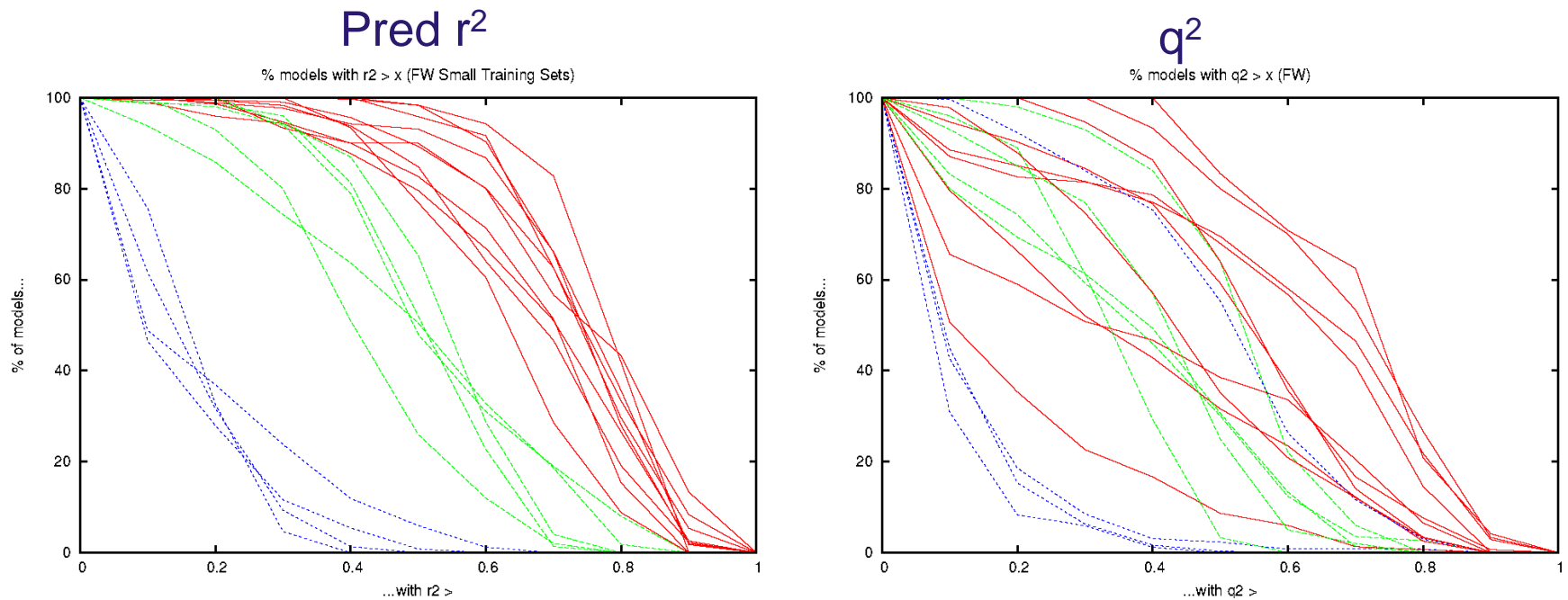
Partially Additive Dataset





# Small (~30%) Training Sets

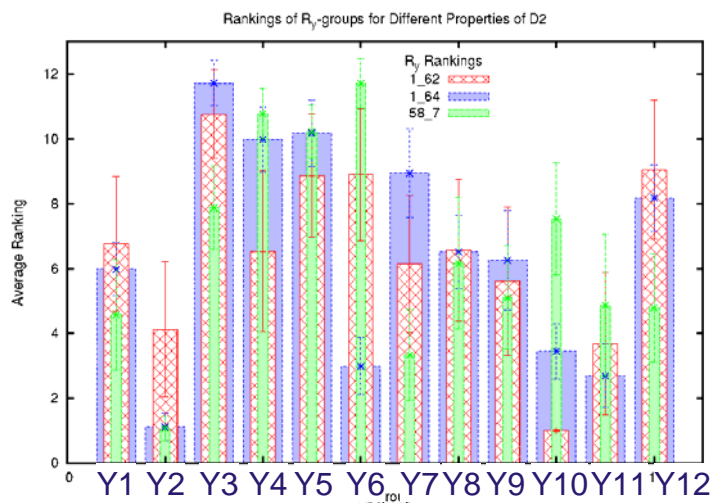
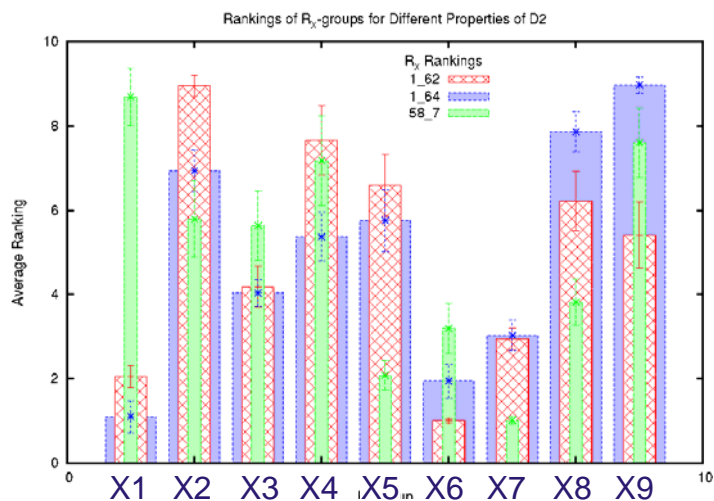
Additive: Red; Non-additive: Blue; Partially additive: Green



Clear separation of additive, non-additive and partially additive datasets



# Compound Profiling



5.3-7.9      5.5-9.0      5.0-7.9



R <sub>x</sub> Group	R <sub>y</sub> Group	1_62	1_64	58_7
X1	Y1	6.74	8.06	5.64
X1	Y2	6.68	8.94	5.98
X1	Y6	6.76	8.61	5.01
X1	Y7	6.96	7.69	6.06
X1	Y9	6.81	7.77	5.77
X1	Y8	6.79	8.06	5.59
X1	Y10	7.77	8.84	5.10
X1	Y11	6.85	8.55	5.17
X1	Y12	6.81	7.78	5.23



# Conclusions: Additivity Analysis

- $\text{Pred-r}^2$  can be used to determine additivity effects for training sets that cover 30% of the full array
- No correlations found between  $\text{pred-r}^2$  and distribution of R groups or property ranges in the training sets
- Further studies needed to
  - Determine robust attributes for  $q^2$
  - Establish good attributes for test sets
- An analysis of the coefficients in additive datasets can be a useful visualisation tool for multi-criteria library design



The  
University  
Of  
Sheffield.

# Acknowledgements

## **Sheffield**

Peter Fleming

Illy Khatib

Jeff Loo

Orazio Nicolotti

Yogi Patel

Peter Willett

Trudi Wright

## **GlaxoSmithKline**

Darren Green

Gavin Harper

Jameed Hussain

Iain Mclay

Stephen Pickett

## **Johnson & Johnson**

Trevor Howe

Julen Oyarzabal

## **Funding and software support**

BBSRC, Chemical Computing Group, Daylight, EPSRC, GSK,  
J&J, Royal Society, Tripos, Wolfson Foundation