



The  
University  
Of  
Sheffield.

# How similar is similar? A study of the Similarity Principle in the context of lead optimisation using molecular fingerprints

George Papadatos  
g.papadatos@shef.ac.uk

◆  
University of Sheffield

◆  
GlaxoSmithKline



# Overview

- Background
  - Objectives
  - Lead Optimisation
  - Array design
  - Similarity principle
  - Fingerprints
  - Neighbourhood behaviour
- Experiments
  - Descriptors
  - Datasets
  - Methods
  - Results
  - Conclusions



# Overview

- **Background**
  - Objectives
  - Lead Optimisation
  - Array design
  - Similarity principle
  - Fingerprints
  - Neighbourhood behaviour

- Experiments
  - Descriptors
  - Datasets
  - Methods
  - Results
  - Conclusions

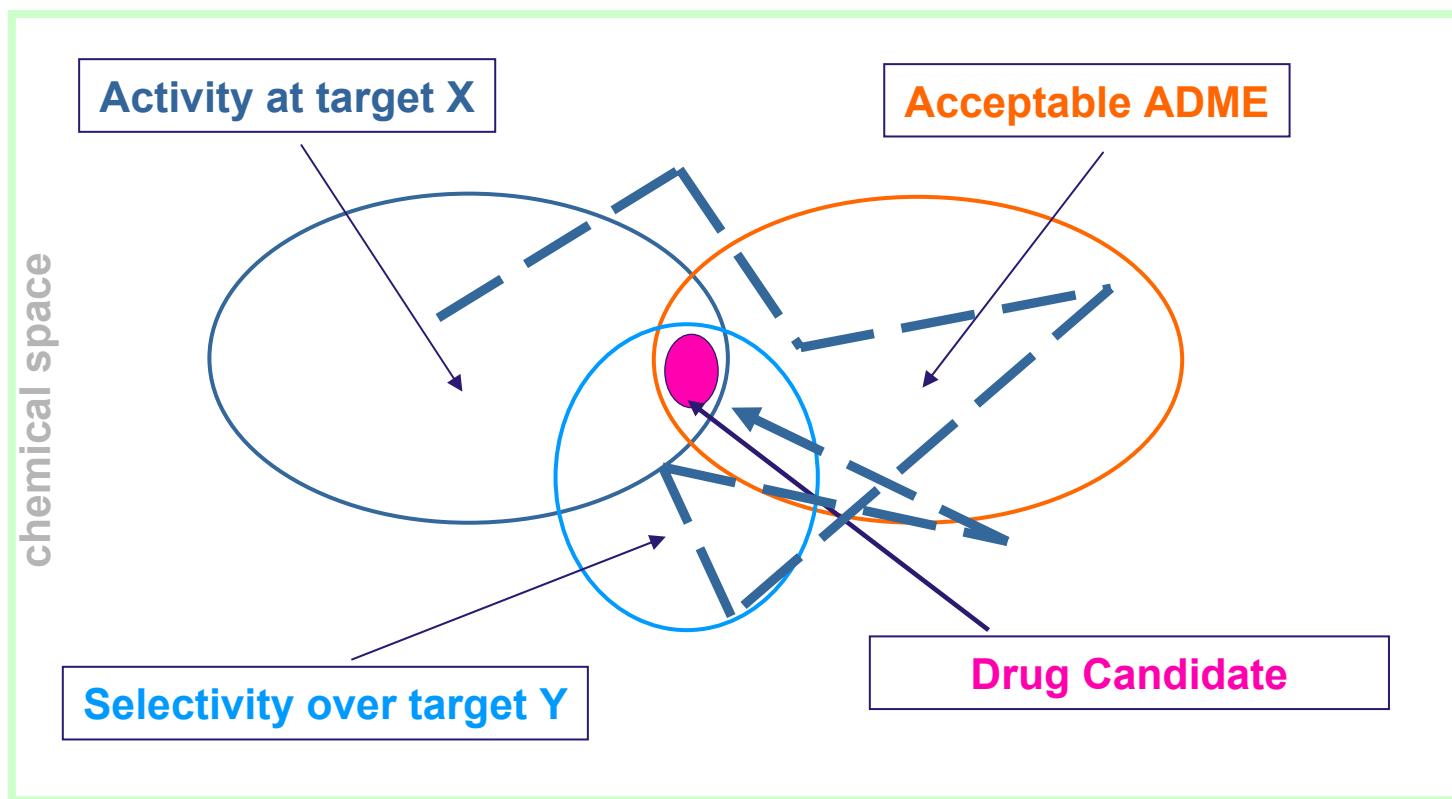


# Objectives

- Understand how different descriptors perceive molecular similarity
- Compare the existing methodologies and come up with a robust one to evaluate descriptors in this matter
- Find which descriptor correlates best with biological activity or any other important property
- Apply the findings to the improvement of chemical array design during lead optimisation



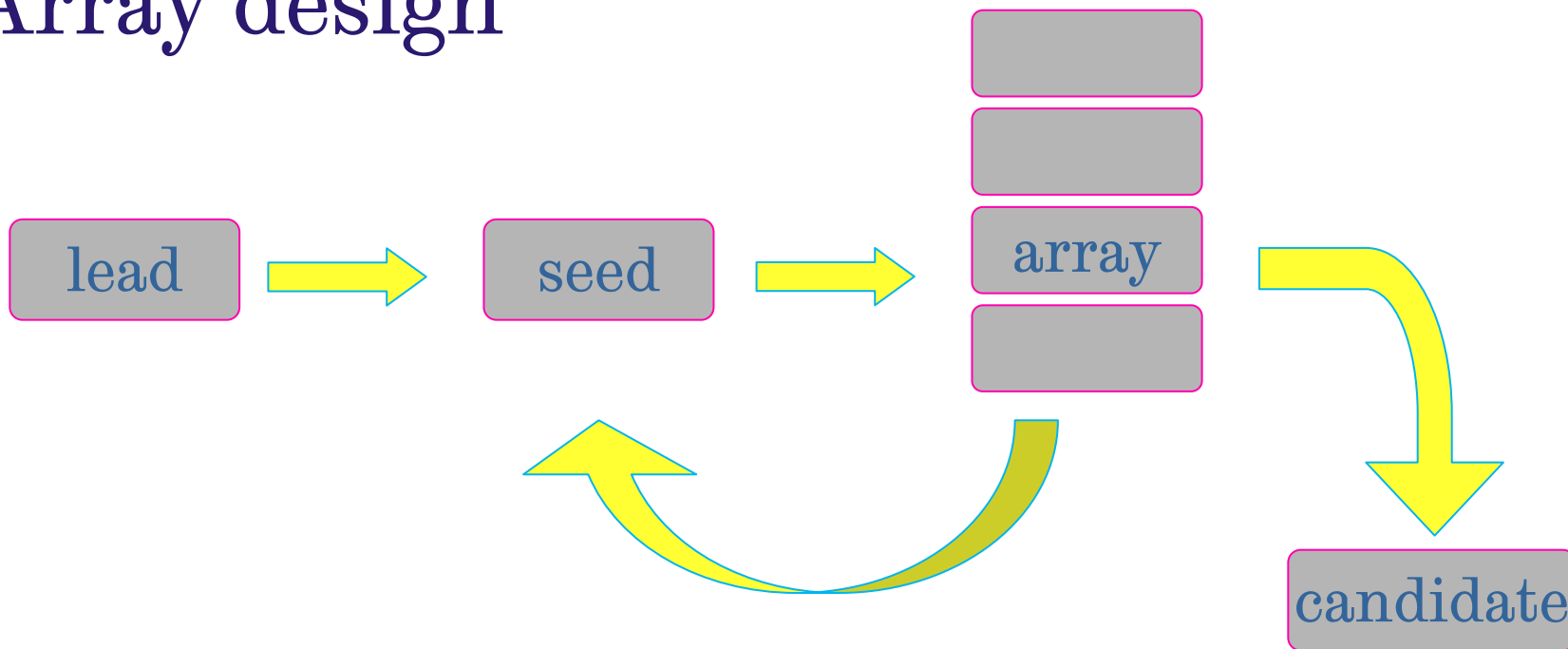
# Lead optimisation



Delaney (2009). Drug Discovery Today, 14, 3/4



# Array design



- Array-able chemistry
- Iterative procedure
- Screening cascades
- Multi-objective optimisation



# Similar Property Principle (SPR)

- Structurally similar compounds tend to exhibit similar properties → similar biological activity
  - Despite the exceptions, it's a powerful rule of thumb
    - Similarity searching
    - Property prediction
  - Without it, lead optimisation projects would be futile
- Is the inverse valid?
  - Dissimilar compounds → dissimilar activity?
  - Similar activity → similar structure?

Sheridan (2007). *Expert Opinion in Drug Discovery*, 2 (4), 423-430.

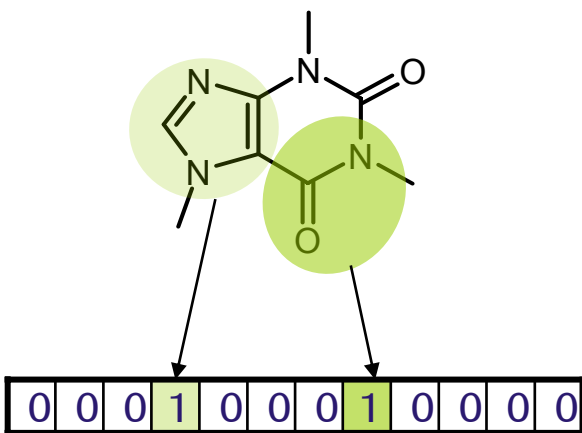
Fliri et al. (2005). *Proceedings of the National Academy of Sciences*, 102 (2), 261-266.



# Chemical similarity

## Representation

- Way to characterise a molecule in a computer-friendly format

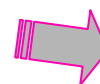


## Similarity coefficient

- Quantifies the degree of resemblance



$$T_c(A,B) = \frac{A \cap B}{A \cup B} = \frac{5}{9} = 0.56, 0 \leq T_c \leq 1$$



similarity metric

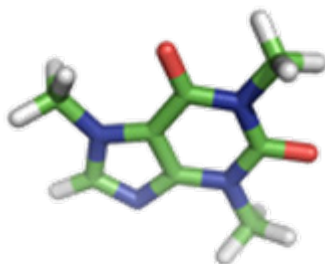


# Molecular descriptors

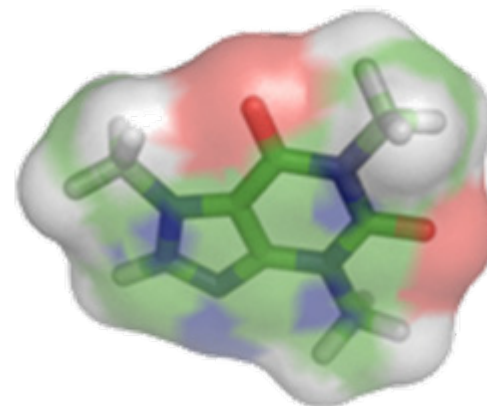
Connectivity



Pharmacophores



Shape

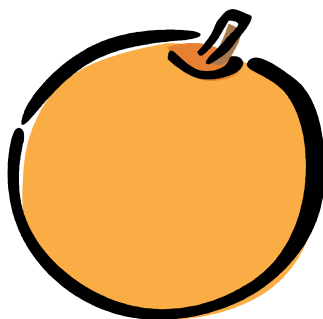


- The good news: numerous descriptors are available
- The bad news: numerous descriptors are available

Adapted from: Brown (2009). *ACM Computing Surveys*, **41** (2).



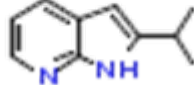
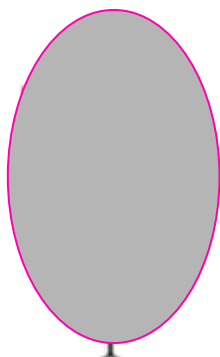
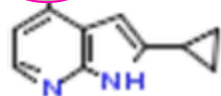
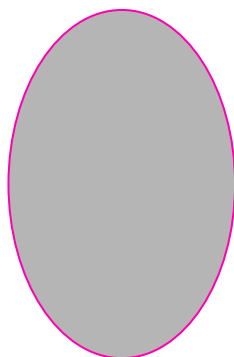
# Similarity is subjective...



Similar?

Shape/Colour/Texture: Yes

Function/Taste: NO!



ECFP\_6: 0.65

Daylight: 0.96

$\Delta pIC_{50}=0.2$

**How similar is similar?**



# Neighbourhood behaviour (NB): why?

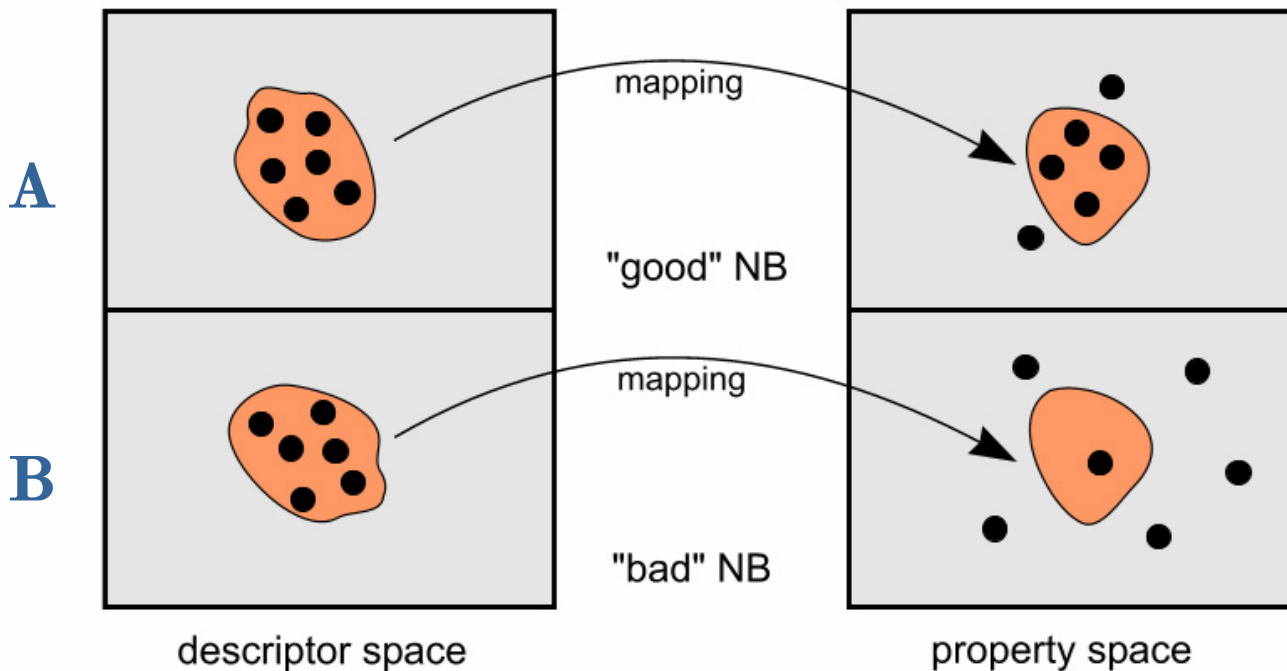
- Neighbourhood principle
  - Molecules in the same neighbourhood of chemical space tend to have similar values of a property
  - cf. *smoothness assumption*: if two points  $x_1, x_2$  are close, then so should be the corresponding outputs  $y_1, y_2$
- NB: the extent to which a given descriptor satisfies the neighbourhood principle → the SPR
  - Without this, any attempt to model computationally the array design would be in vain

Patterson et al. (1996). *Journal of Medicinal Chemistry*, **39** (16), 3049-59.

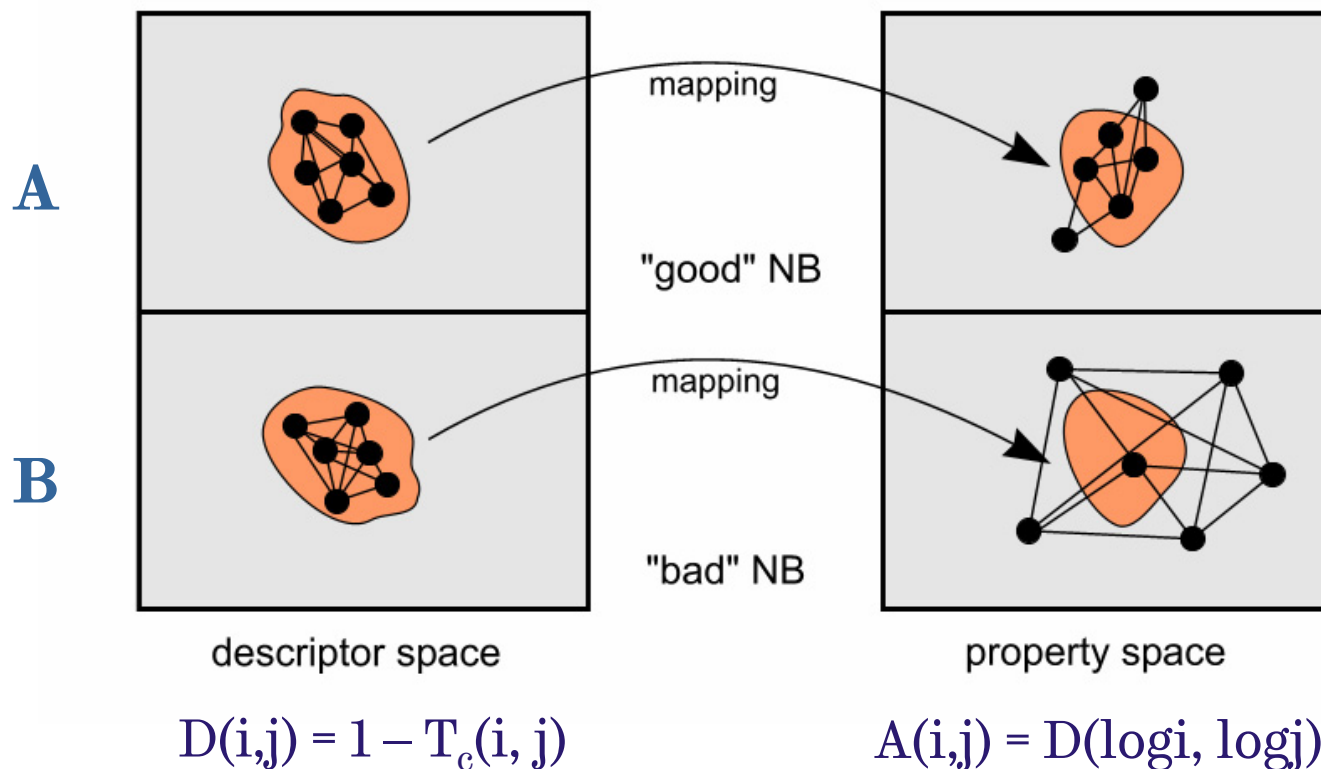
Barbosa and Horvath (2004). *Current Topics in Medicinal Chemistry*, **4**, 589-600.



# NB examples



# NB examples



Horvath and Jeandenans (2003). *Journal of Chemical Information and Computer Sciences*, **43** (2), 680-690.



# Overview

- Background
  - Rationale and aim
  - Lead Optimisation
  - Array design
  - Similarity principle
  - Fingerprints
  - Neighbourhood behaviour
- Experiments
  - Datasets
  - Fingerprints
  - Methods
  - Results
  - Conclusions



# Details of the study

- 3 GSK LO projects
- 9 diverse chemotypes
  - 35-901 cmpds each
  - 5-90 arrays each
- 12 structural descriptors
  - 2- and 3D
- Bioactivity (6 targets)
- Permeability
- Metabolic stability
- Lipophilicity
- Evaluate two existing methodologies in the context of LO and discuss the respective advantages and drawbacks
- Assess the performance of several descriptors using both methodologies



# Datasets

Chemotype	Project I	
	Target 1	Target 2
1	355	488
2	183	149
3	162	159
4	35	124
Total	<b>735</b>	<b>920</b>

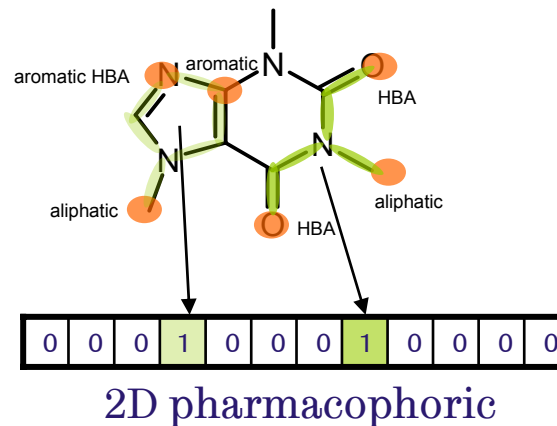
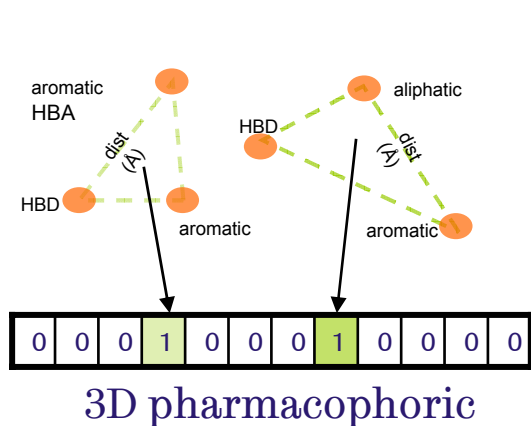
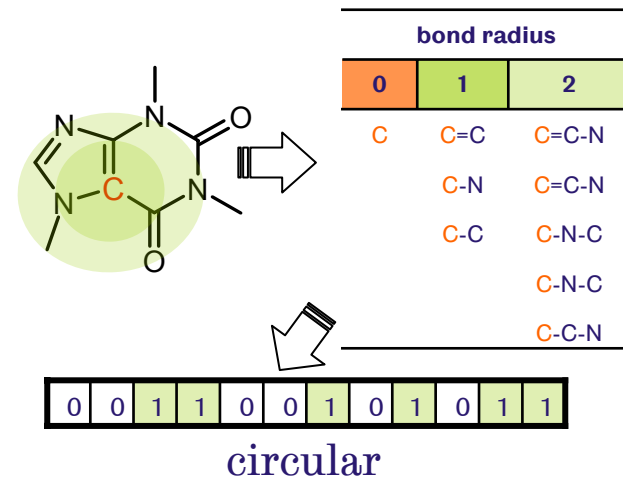
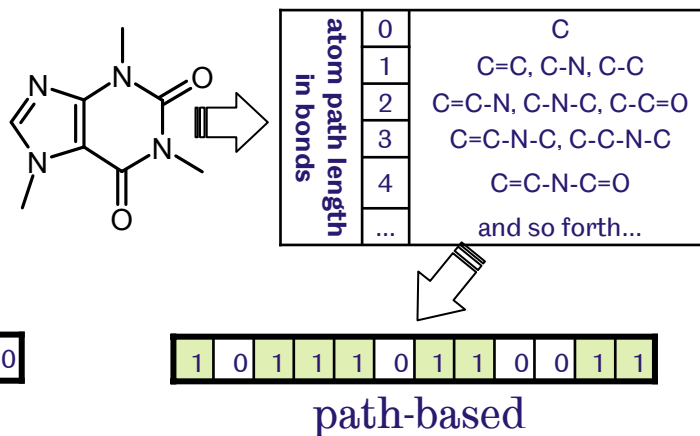
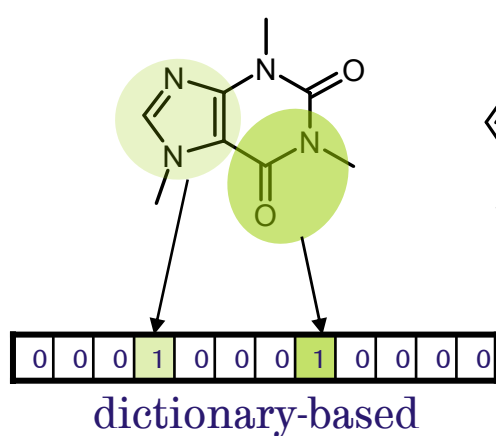
Chemotype	Project III
	Target 6
8	222
9	146
Total	<b>368</b>

Chemotype	Project II		
	Target 3	Target 4	Target 5
5	175	179	146
6	322	320	288
7	892	901	655
Total	<b>1389</b>	<b>1400</b>	<b>1089</b>

Chemotype	Project II		
	Permeability	Metabolic Stability	Lipophilicity
5	32	87	43
6	53	233	147
7	177	482	183
Total	<b>262</b>	<b>802</b>	<b>373</b>



# Types of fingerprints





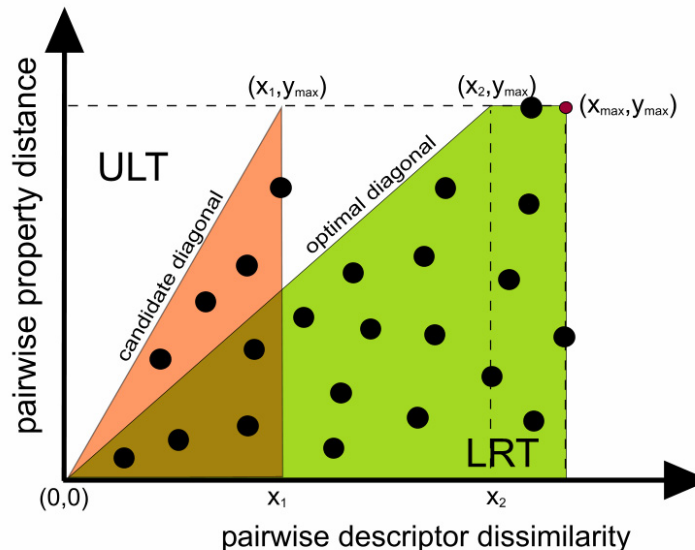
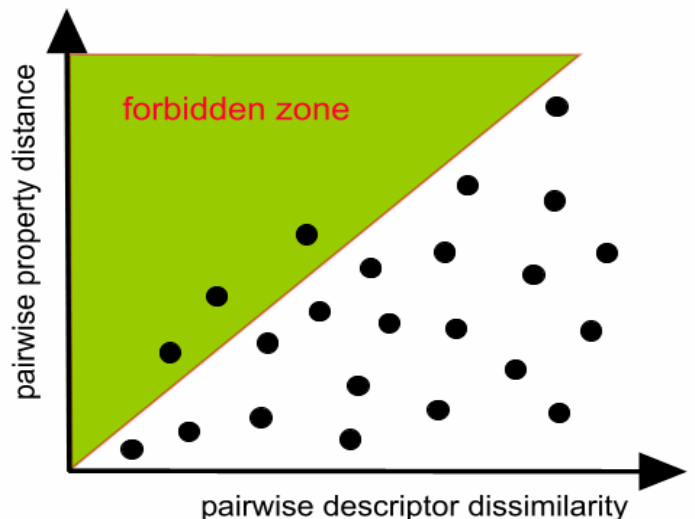
# Patterson plots

- Patterson plots
  - Pairwise descriptor dissimilarity vs. absolute differences in values
  - $n$  compounds  $\rightarrow n(n-1)/2$  unique pairs – data points
- Optimal diagonal algorithm
- NB score and  $\chi^2$  test

$$\text{Density} = \frac{\text{population}}{\text{area}}$$

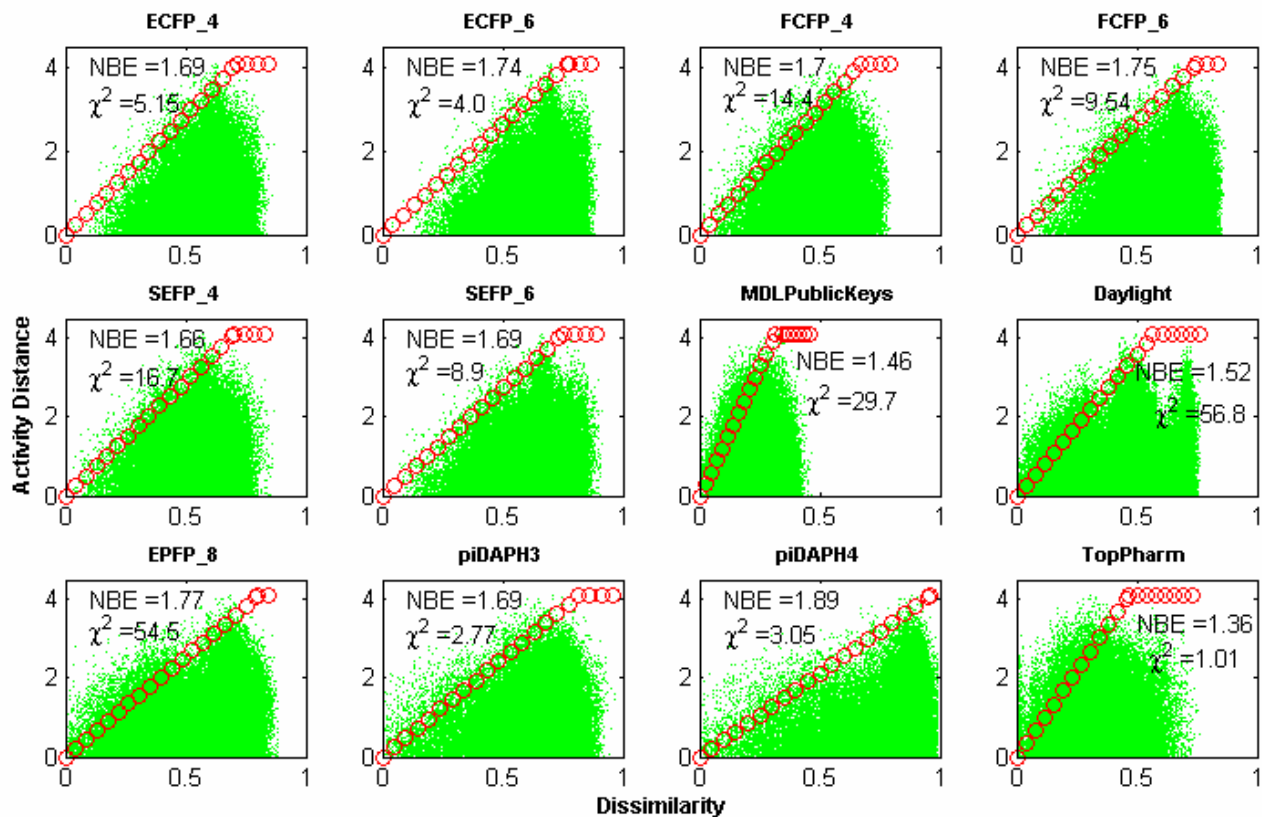
$$\text{NBE} = \frac{\text{Density}_{\text{LRT}}}{\text{Density}_{\text{LRT} \cup \text{ULT}}}$$

$$\chi^2 = \frac{(N_{\text{LRT}} - n_{\text{LRT}})^2}{n_{\text{LRT}}}$$





# Patterson plots - example



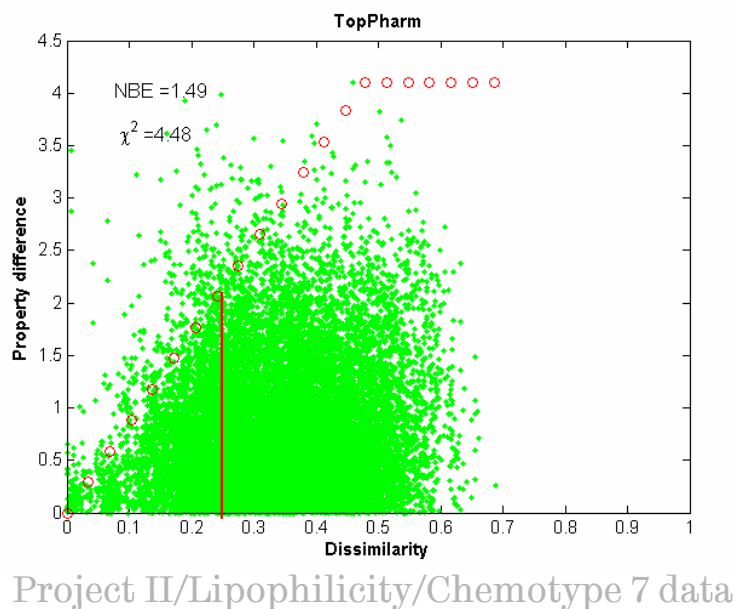
Optimal diagonal plots for the Project II/Target 3/Chemotype 7 data



# Drawbacks

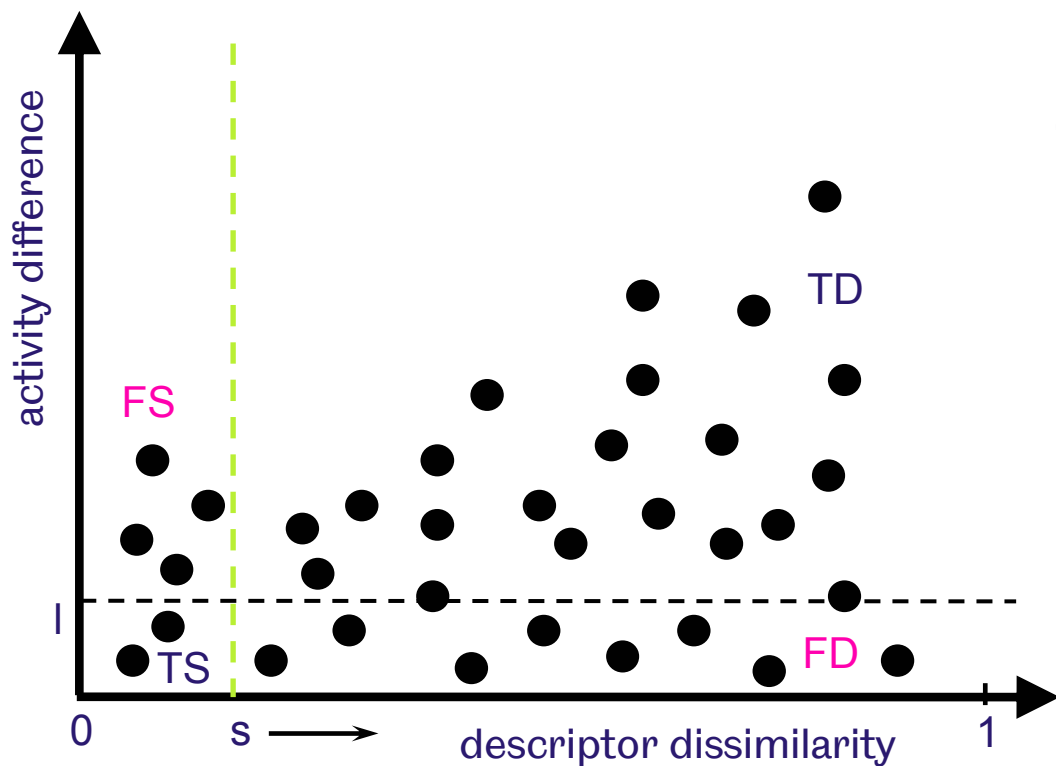
## Patterson method

- Tends sometimes to underestimate descriptors' performance
- Difficult to see trends
- Role of  $\chi^2$  is ambiguous
- Triangles and rectangles add complexity and subjectivity
- Can be very slow





# Optimality $\Omega$



Optimality  $\Omega$  = consistency + completeness

- For a specific  $s$ : minimal  $\Omega(s)$  better NB

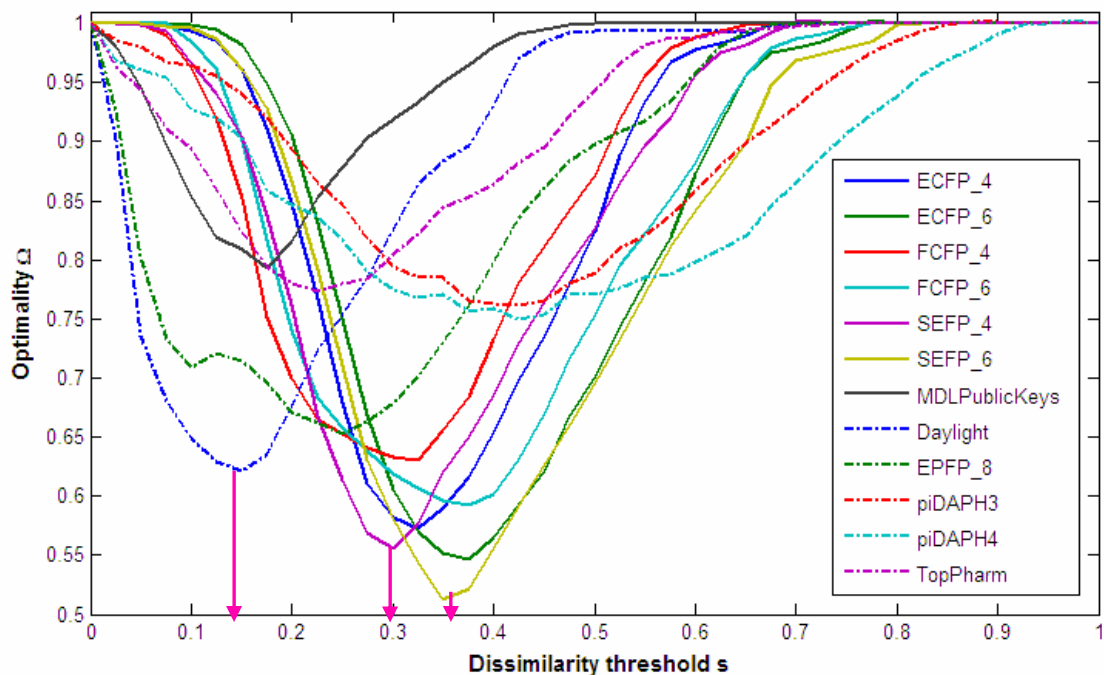
$$\Omega(s) = \frac{kN_{FS} + N_{FD}}{kN_{FS}^{\text{expected}} + N_{FD}^{\text{expected}}}$$

$$\Omega(0) = 1, \Omega(1) = 1, \Omega = 1 - \text{kappa}$$

$$l=0.5, k=5$$



# Optimality plots - example



Optimality plots for the Project II/Target 3/Chemotype 5 data

The lower the optimality, the better the NB



# Results I - Comparison of methods

## Patterson method

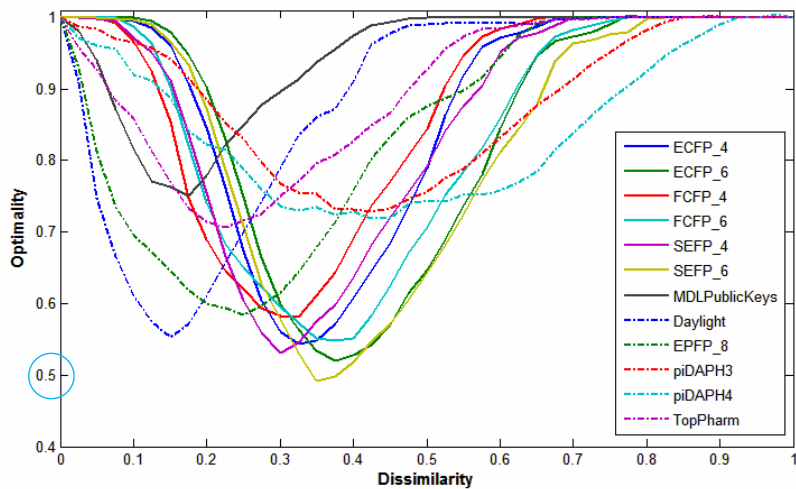
- Tends sometimes to underestimate descriptors' performance
- Difficult see trends
- Role of  $\chi^2$  is ambiguous
- Triangles and rectangles add complexity and subjectivity
- Can be very slow

## Optimality method

- Plots easier to compare
- Defines the optimal similarity cut-off
- Simpler, faster and based on an established statistic
- More robust framework for evaluating similarity metrics

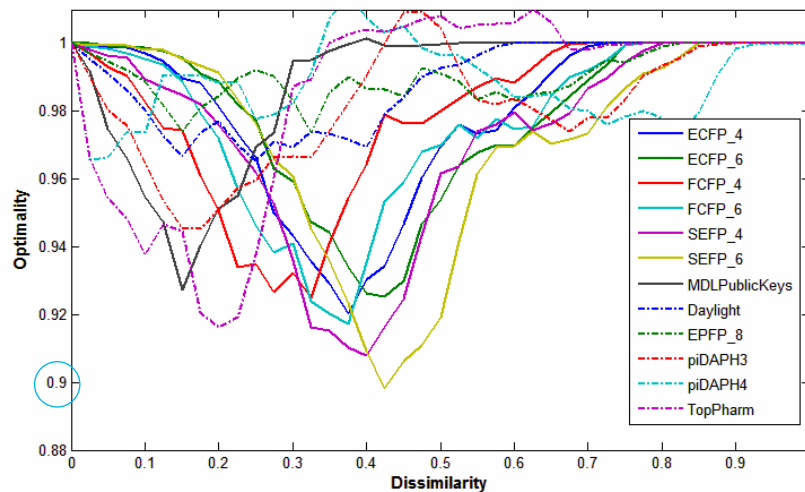


# Results II – Indications about SAR



Project II/Target 4/Chemotype 5 data

the “smooth hills of Kansas”?



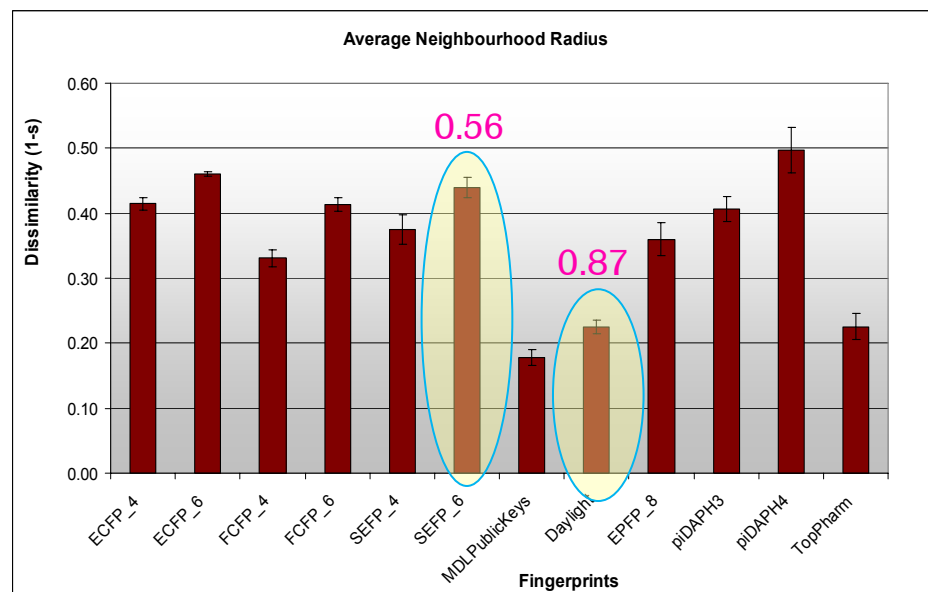
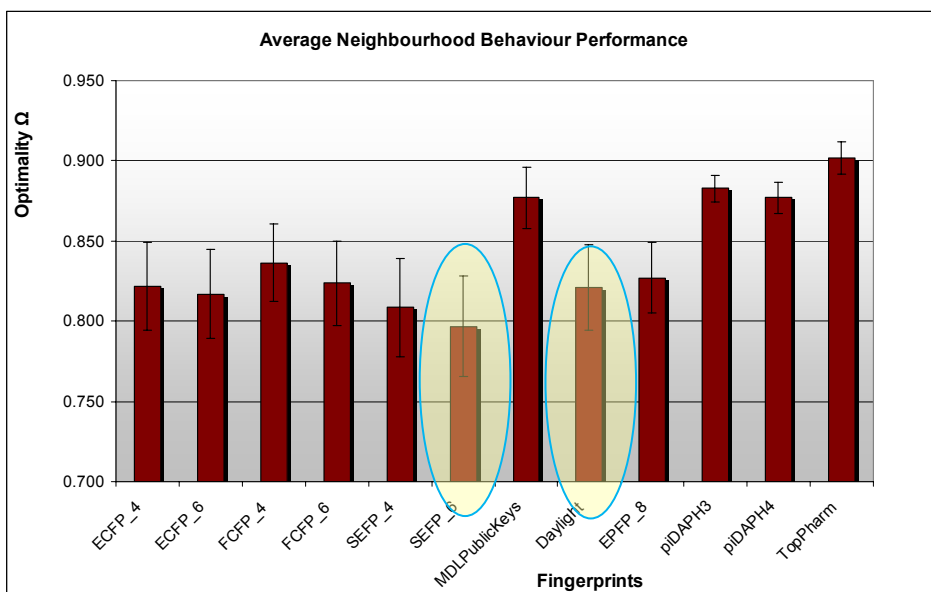
Project I/Target 1/Chemotype 2 data

the “sharp cliffs of Bryce Canyon”?

Guha and van Drie (2008). *Journal of Chemical Information and Modeling*, **48**, 646-658



# Results III – Descriptors' overall performance



The lower the optimality, the better the NB

Dissimilarity=1-similarity

Papadatos et al., *Journal of Chemical Information and Modeling*, **49**, 195-208.

Muchmore et al. (2008). *Journal of Chemical Information and Modeling*, **48**, 941-948.



# Take-home message(s)

- Fingerprints have peculiarities
  - This affects how they perceive similarity
- Neighbourhood behaviour is a valid concept
  - Assesses the extent to which a descriptor follows the SPP
  - Applicable to any modelling effort
  - Highly suggested for array design analysis
  - Optimality seems to be the method of choice
  - Circular fingerprints tend to perform better, at a lower threshold



# Acknowledgements

- My supervisors:
  - Val Gillet, Peter Willett, Visakan Kadiramanathan
  - Chris Luscombe, Giampa Bravi, Iain McLay

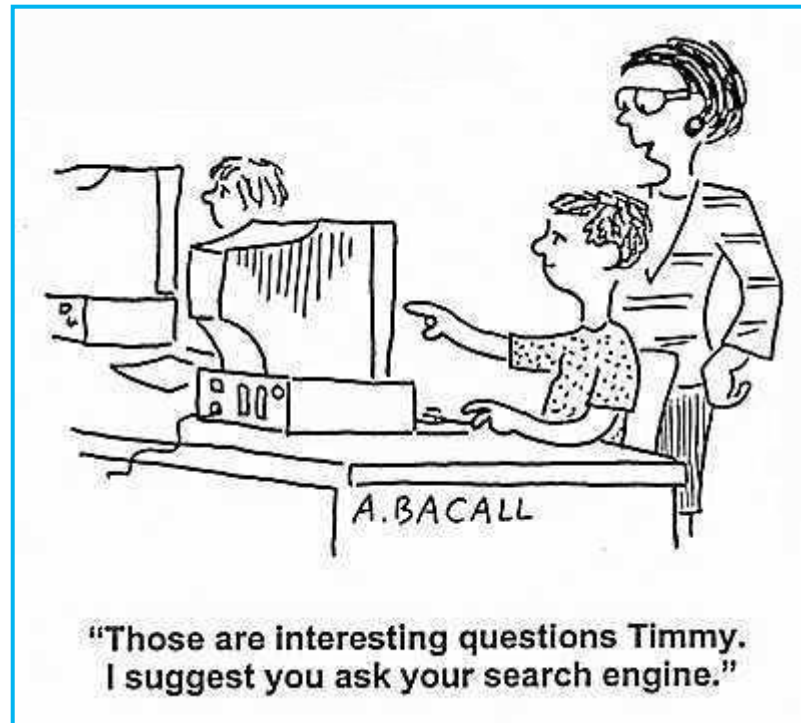
**EPSRC**

Engineering and Physical Sciences  
Research Council





# Any questions?





# Back-up slides

... as if it wasn't enough



# Datasets

Chemotype	Project I	
	Target 1	Target 2
1	355	488
2	183	149
3	162	159
4	35	124
Total	<b>735</b>	<b>920</b>

Chemotype	Project III
	Target 6
8	222
9	146
Total	<b>368</b>

Chemotype	Project II		
	Target 3	Target 4	Target 5
5	175	179	146
6	322	320	288
7	892	901	655
Total	<b>1389</b>	<b>1400</b>	<b>1089</b>

Chemotype	Project II		
	Permeability	Metabolic Stability	Lipophilicity
5	32	87	43
6	53	233	147
7	177	482	183
Total	<b>262</b>	<b>802</b>	<b>373</b>



# Endpoints

Property	Unit	Distance	Threshold I
Potency	pIC50	Euclidean $ \Delta pIC50 $	0.5
Lipophilicity	$\log D_{pH7.4}$	Euclidean $ \Delta \log D $	0.5
Permeability	nm/s	Euclidean $ \Delta \log Perm $	0.5
Metabolic Stability	% remaining	Euclidean $ \Delta \log Stab $	0.3



# Fingerprints

Type of fingerprint	Name	Abbreviation	Program used
2D structural keys	MDL Public Keys	MDLPublicKeys	Pipeline Pilot <sup>25</sup>
2D path substructures	Daylight	Daylight	Daylight <sup>26</sup>
	EPFP	EPFP_8	Pipeline Pilot
2D circular substructures	Extended Connectivity	ECFP_4 and _6	Pipeline Pilot
	Functional Class	FCFP_4 and _6	Pipeline Pilot
	Atom Environments	SEFP_4 and _6	Pipeline Pilot
2D pharmacophores	GSK Topological Pharmacophores	TopPharm	In-house
3D pharmacophores	3-point	piDAPH3	MOE <sup>27</sup>
	4-point	piDAPH4	MOE



# Pipeline Pilot implementation

- Database integration
- Data pre-processing
  - Remove duplicates
  - Remove missing/wrong values
  - SMARTS filtering
- SOAP integration
  - GSK web services
  - Daylight fps
- Fingerprint generation
- Property distance / fingerprint similarity matrices
- MOE integration
  - MOE fps
- MATLAB integration?
  - PP version 7.5!



# Optimality

		Structural similarity	
		Similar ( $\Delta S < s$ )	Dissimilar ( $\Delta S > s$ )
Activity similarity (assay)	Similar ( $\Delta \text{Act} < l$ )	$N_{TS}$	$N_{FD}$
	Dissimilar ( $\Delta \text{Act} > l$ )	$N_{FS}$	$N_{TD}$

$$N_{FS}^{\text{expected}} = \frac{(N_{TS} + N_{FS})(N_{TD} + N_{FS})}{N}$$

$$N_{FD}^{\text{expected}} = \frac{(N_{TS} + N_{FD})(N_{TD} + N_{FD})}{N}$$

$$\Omega(s) = \frac{kN_{FS} + N_{FD}}{kN_{FS}^{\text{expected}} + N_{FD}^{\text{expected}}}$$

$$l=0.5$$

$$k=5$$



# Optimality

- Consistency
  - Can a descriptor selectively pick activity-related pairs among the most structurally similar of them? (cf precision)
- Completeness
  - What is the proportion of all activity-related pairs among the most similar of them? (cf recall)
- Optimality  $\Omega$  = consistency + completeness
  - For a specific  $s$ : minimal  $\Omega(s)$   $\rightarrow$  better NB

$$\Omega(s) = \frac{kN_{FS} + N_{PFD}}{kN_{FS}^{expected} + N_{PFD}^{expected}}, \Omega(0) = 1, \Omega(1) = 1, \Omega = 1 - \kappa$$