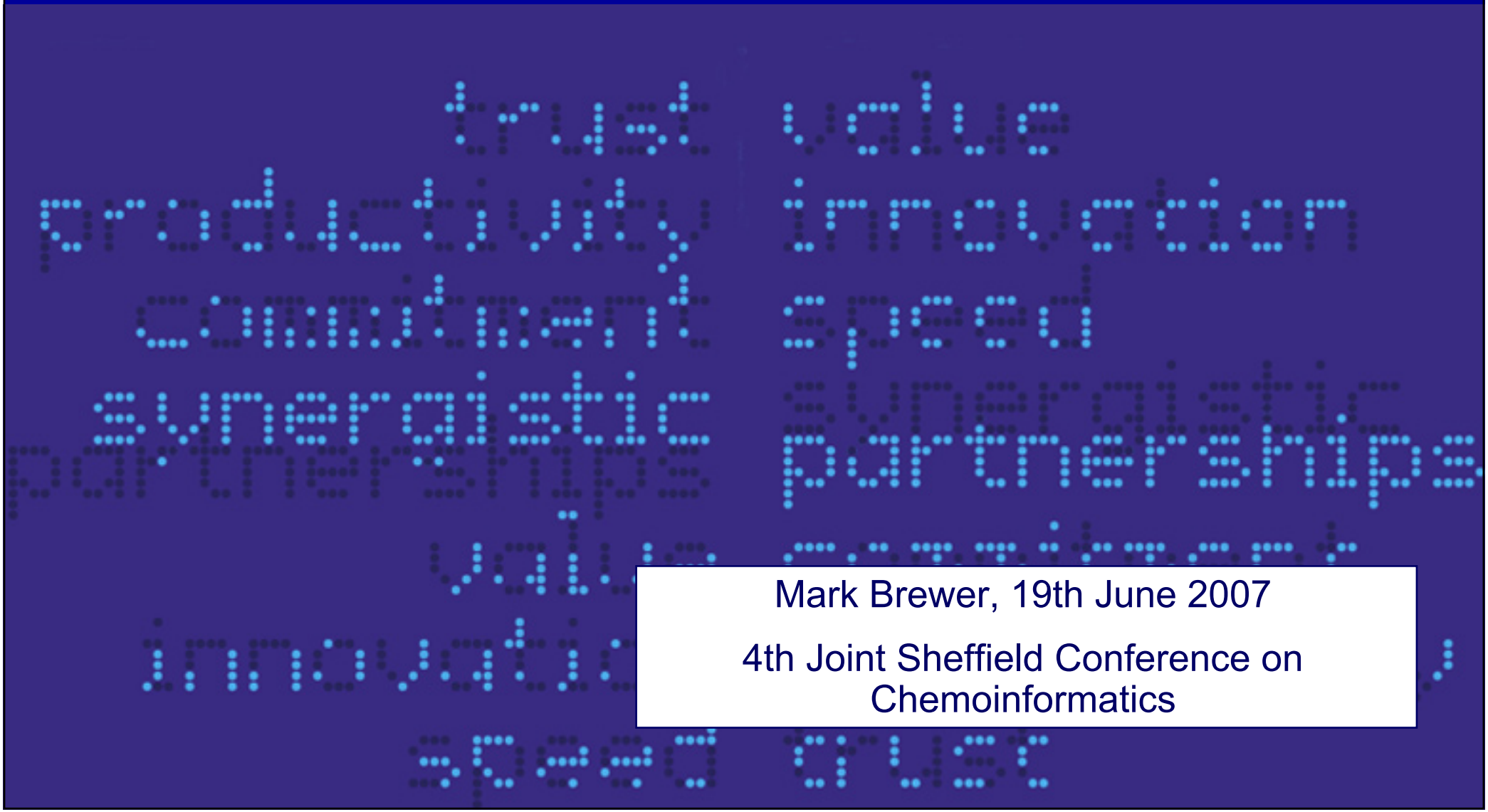
Tomorrow's Drugs Today™



Mark Brewer, 19th June 2007

4th Joint Sheffield Conference on Chemoinformatics

A spectral clustering approach for the analysis of screening data

- Techniques for generating screening data
- Analysing screening data
- A spectral clustering approach
- Example application to COX-2 inhibitors
- Provision of results and applications
- Some possibilities for future work and summary

evotec

High throughput screening

- High throughput screening factory that can generate 100,000 data points a day per machine
- Developed with partners:

NOVARTIS

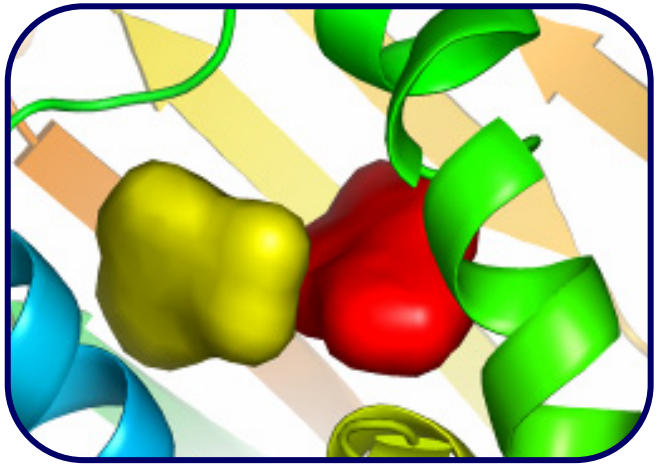
Pfizer

gsk GlaxoSmithKline

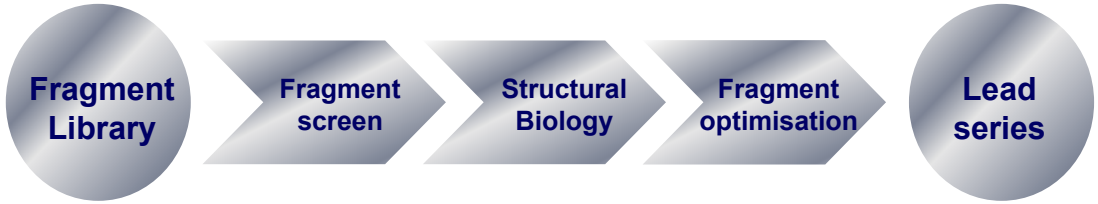
evotec

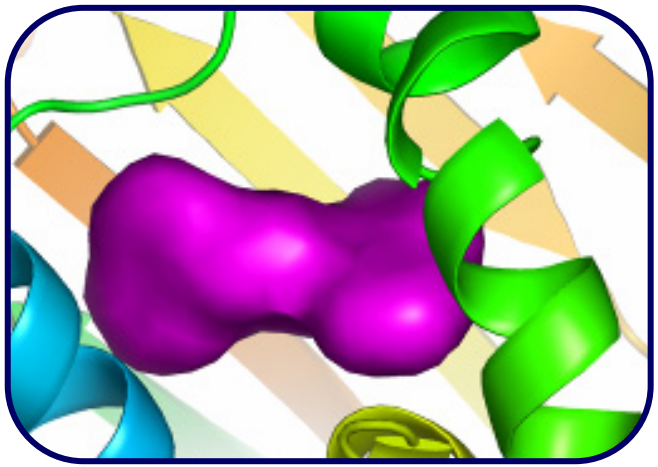
Fragment screening

- Confocal Fluorescence Spectroscopy
- NMR Spectroscopy
- 30000 fragment library



Fragment based drug discovery





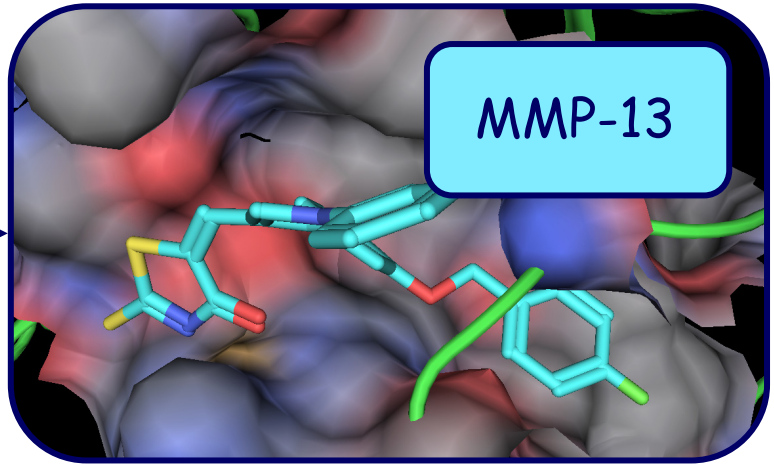
evotec

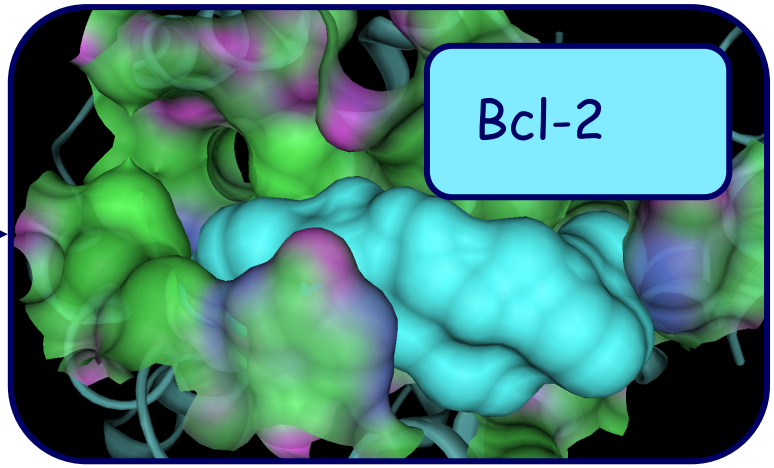
vita nuova

Virtual screening approaches

- Substructure searches
- Similarity searches
- Pharmacophore-based approaches
- Molecular docking
- Physiochemical properties
- *In silico* ADMET
- .....

Screening database → High throughput docking employs a computing grid distributed across 400PCs





evotec

Clustering for the analysis of screening data

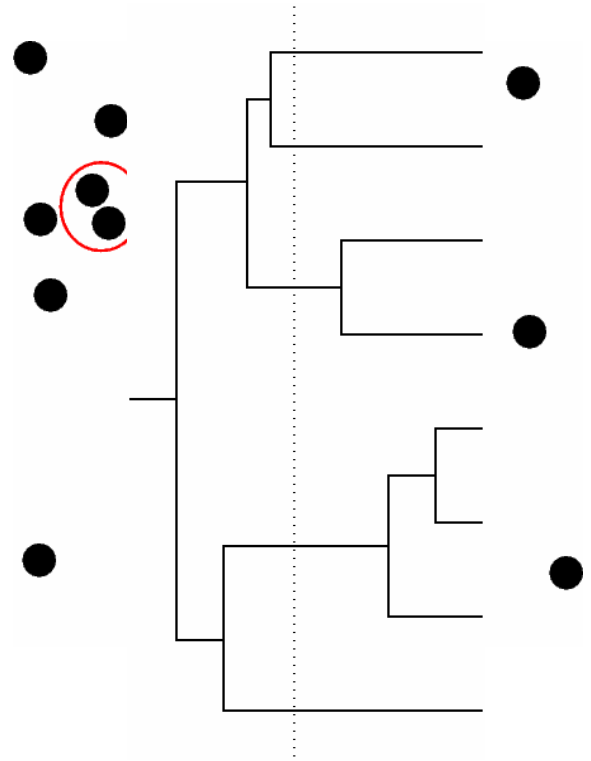
- Both *in vitro* and *in silico* screening yield lists of compounds
- Medicinal chemists will inspect *in vitro* hits to identify compounds to take into hit-to-lead programs
  - Assess number and representation of different scaffolds
  - Molecules with similar structure constitute hit series
  - Realised by software packages which cluster molecules
- Virtual screening hit lists contain list of compounds for purchase and/or synthesis
  - Diversity can be an important caveat so clustering can help here also
- Lots of other applications of clustering e.g. profiling of compound collections

evotec

Clustering molecular datasets

- Many types of clustering method available: hierarchical, non-hierarchical, nearest neighbour, relocation, etc.
- There is a common general procedure
  1. Assign descriptors
  2. Use descriptors to calculate similarity
  3. Process similarities with a clustering algorithm

Hierarchical clustering: e.g. Wards Method



P. Willett, *Similarity and Clustering in Chemical Information Systems,* Research Studies Press, Letchworth (1987)
G.M. Downs and J.M.Barnard, Rev. Comput. Chem., 18, 1 (2002)

evotec

Clustering molecular datasets

- How many clusters?
- How to interpret and distribute clustering output?
- Ties in proximity?
- How to determine the validity of clusters?

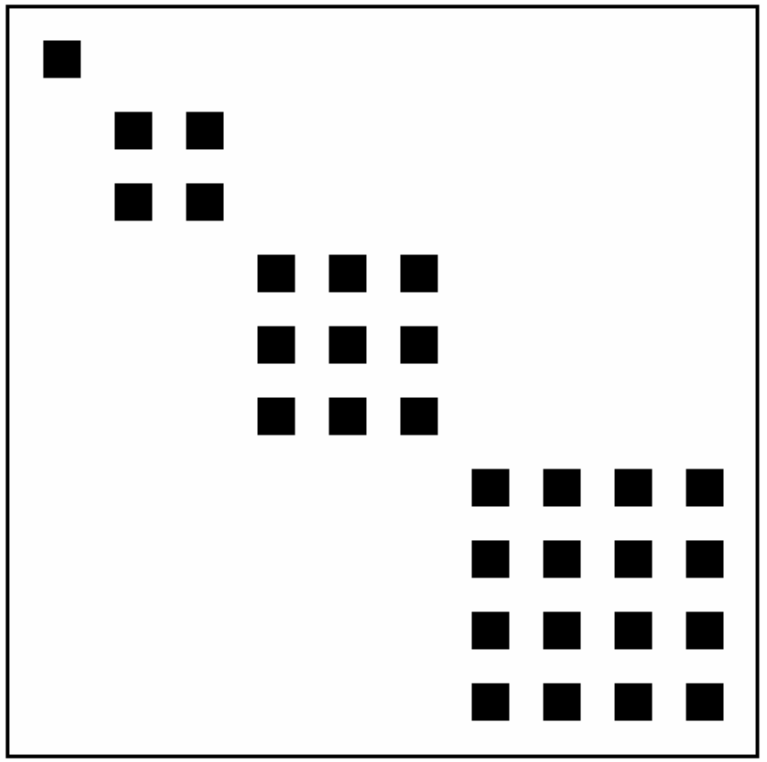
> .... the best measure of the validity (and hence choice of clustering algorithm) has been the usefulness of the resulting classification ...†

D.J. Wild and C.J. Blankley, JCICS, 40, 155 (2000), J. M. Barnard and G. M. Downs, JCICS, 32, 644 (1992)†, J. MacCuish *et al.*, JCICS, 41, 134 (2001)

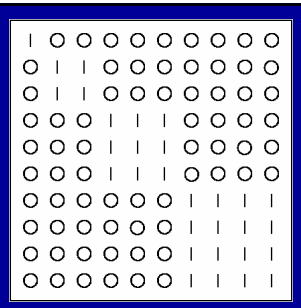
evotec

A thought experiment

A hypothetical similarity matrix, where ■ denotes a pair of 'similar' molecules

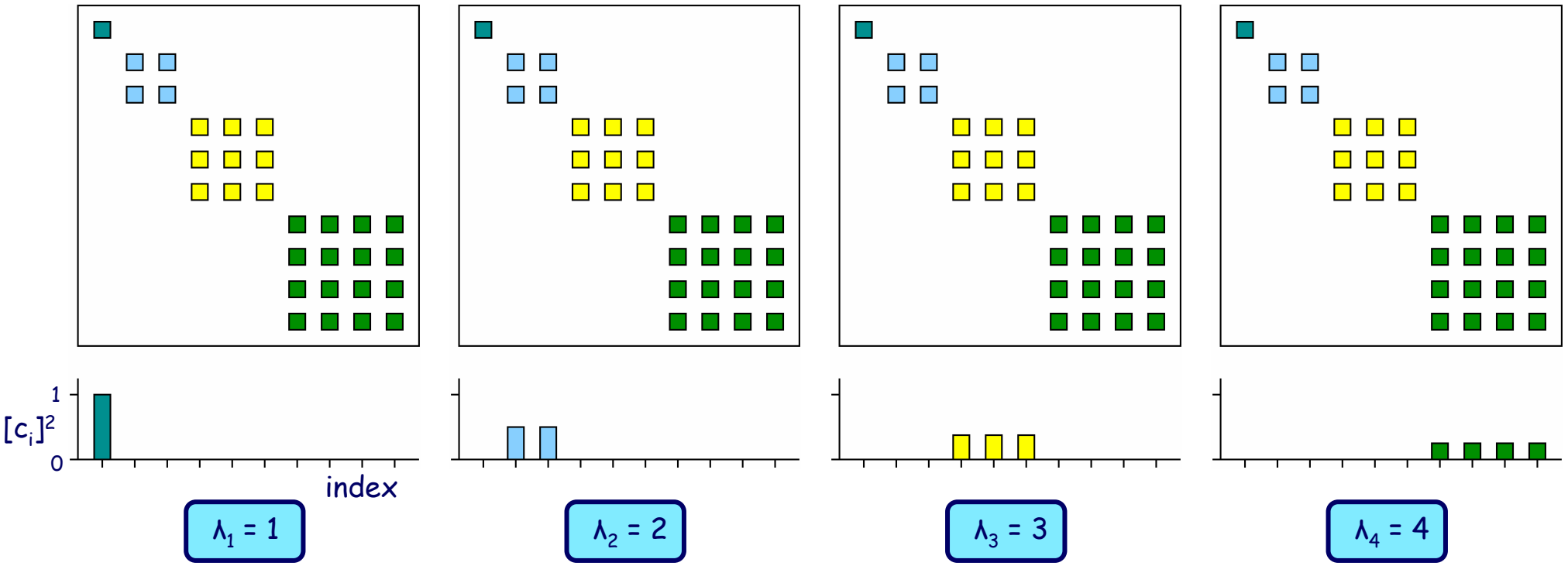


Ordered ideal matrix



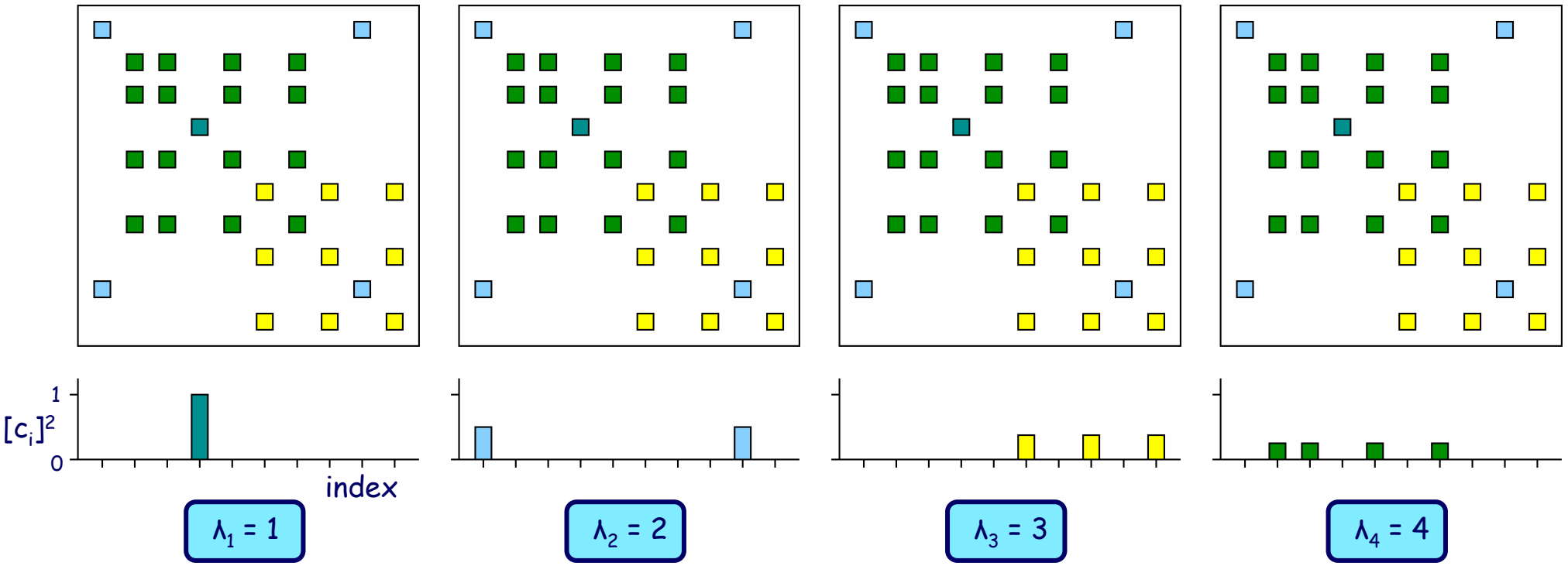
$= S = C \Lambda C^T$

evotec



evotec

Permuted ideal matrix

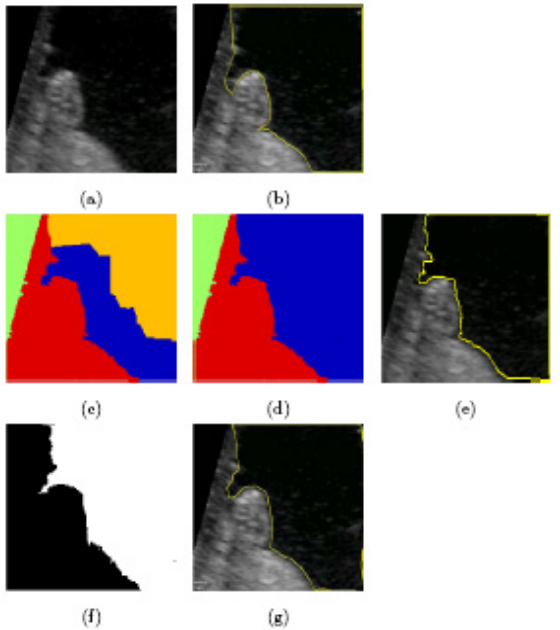


evotec

Existing applications of spectral clustering

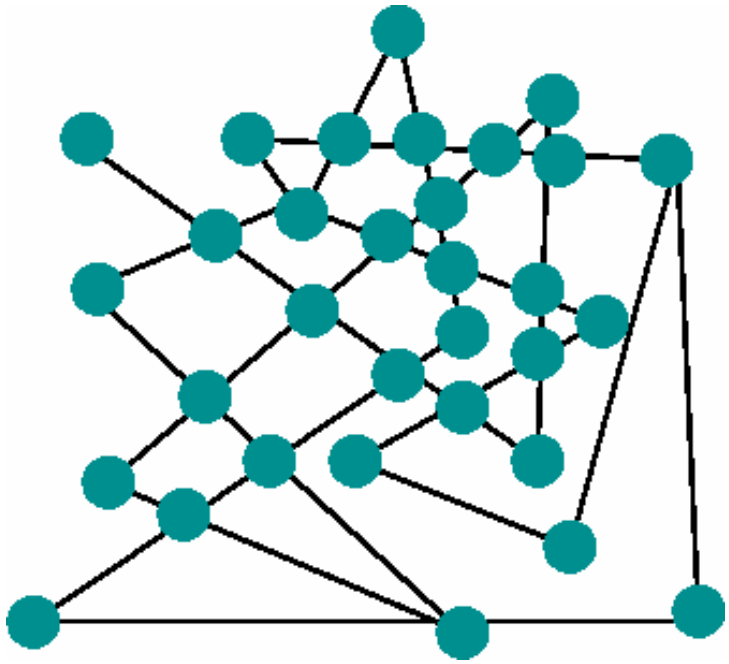
- Eigenvalues and eigenvectors form the basis of spectral clustering algorithms
- Spectral clustering algorithms are used routinely to process digital images
  - Segmentation: To obtain a compact representation of what is useful in the image
- Spectral clustering has also been used in other cheminformatics applications
  - Guha and Wild have used singular value decomposition to cluster molecular datasets

Foetal images processed using spectral clustering



evotec

Further background



If we recognise the similarity matrix (S) as a weighted adjacency matrix and diagonalise:

$S = C\Lambda C^{T}$.

Results from theory of graph spectra show that

1) Eigenvalues (λ) relate to cluster cohesiveness

2) Eigenvectors (C) relate to cluster contribution and are normalised $CC^{T} = 1$
(analogous to fuzzy clustering membership functions)

Sorting by λ and C provide a means to inspect clusters

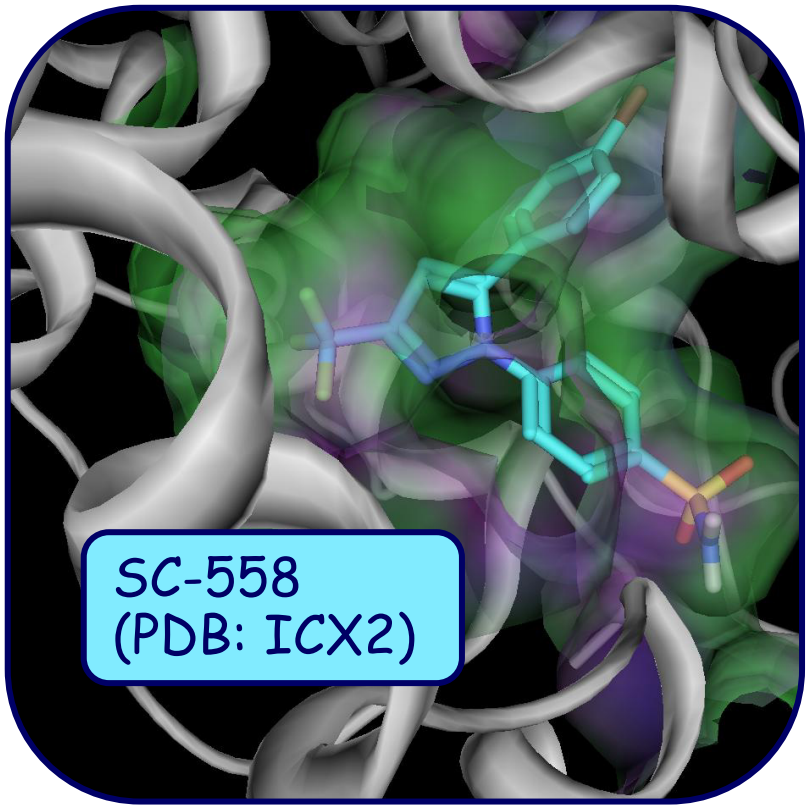
N.B. λ does not depend on the order of molecules

S. Sarkar and K. L. Boyer, Computer Vision and Image Understanding, 71, 110 (1998)
D. Forsyth and J. Ponce, *Computer Vision a Modern Approach,* Prentice-Hall, India (2003)

evotec

Example application: COX-2 inhibitors

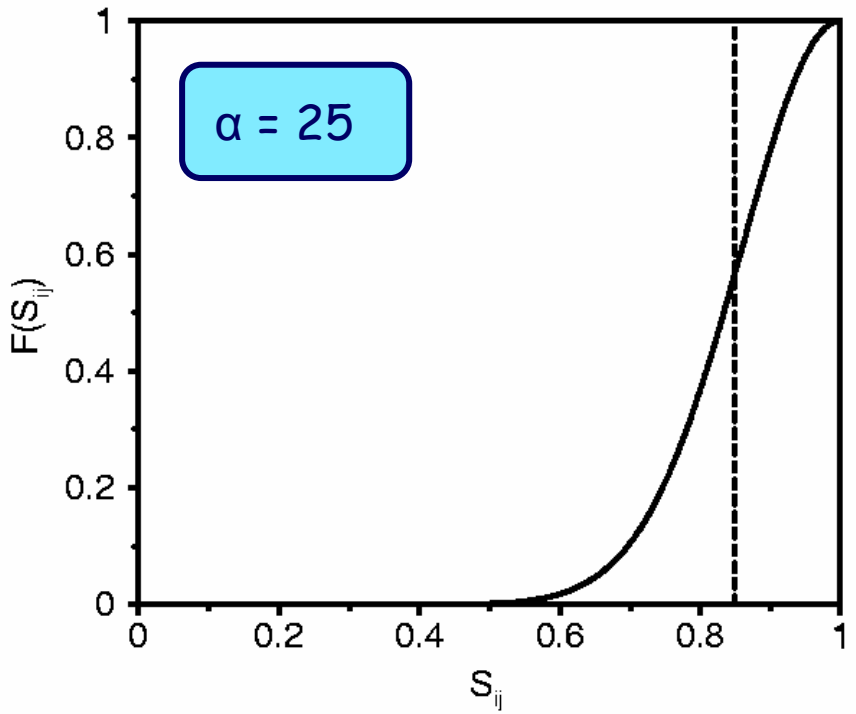
- COX-2 is an important target for inflammation
- COX-2 dataset downloaded from the cheminformatics website
- Standard Tripos UNITY 2D molecular fingerprints for 125 unique chemical structures
- Programs written to compute Tanimoto similarities and diagonalise filtered similarity matrices
- Example clusters presented using the five largest eigenvalues and $\alpha = 25$



evotec

Filtering function

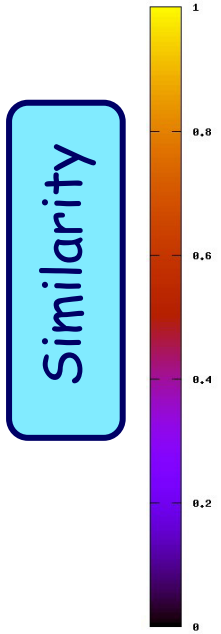
- Filtering function was found to be necessary to obtain useful clusters
- Removes background similarity but maintains chemically relevant levels of similarity
- Several filtering strategies investigated
- Gaussian filtering function selected
  - Single parameter with correct limiting behaviour

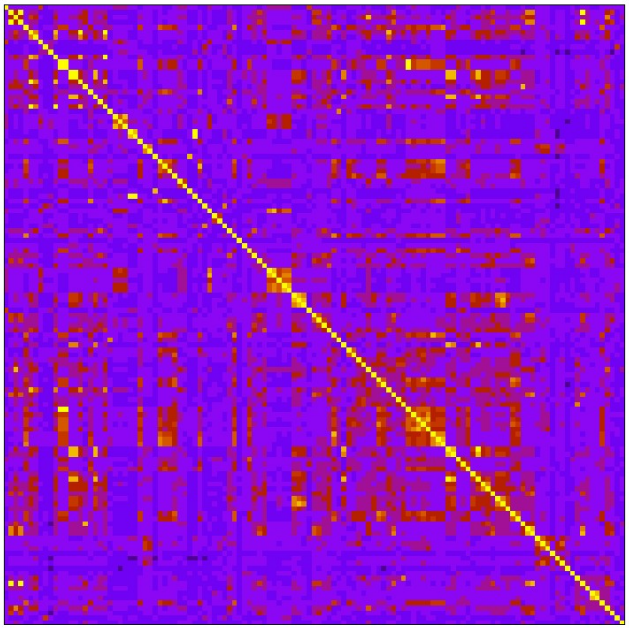


$$F(S_{ij}) = \exp[-\alpha(S_{ij}-1)^2]$$

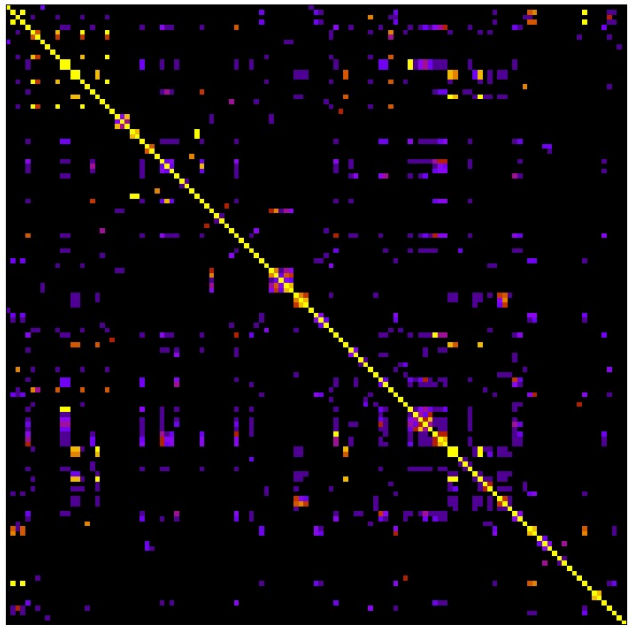
evotec

Effect of filtering function





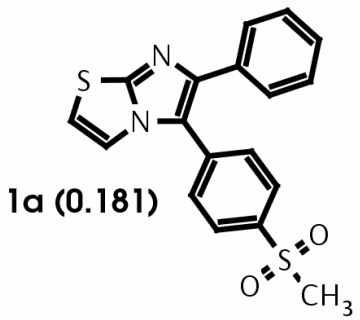
Input similarity matrix

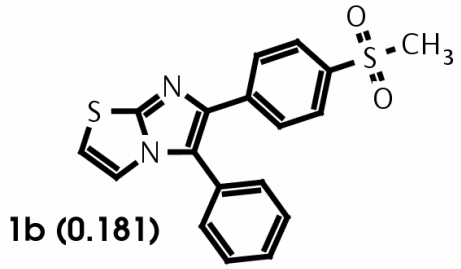


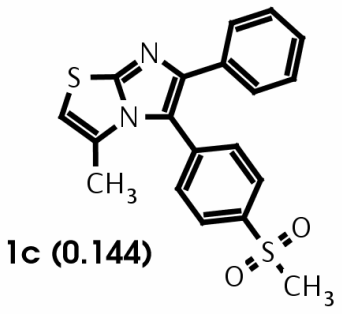
Gaussian filtering (α = 25)

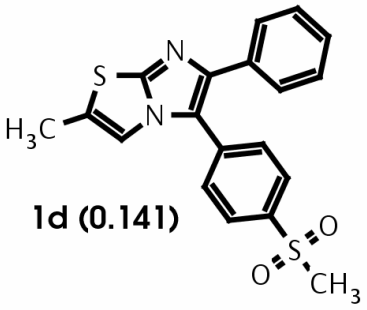
evotec

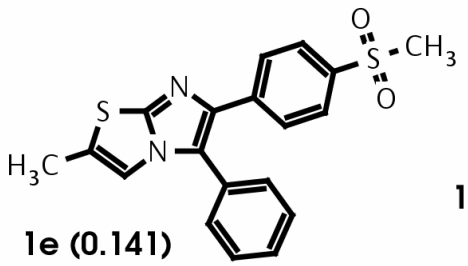
Cluster 1: Imidazothiazoles

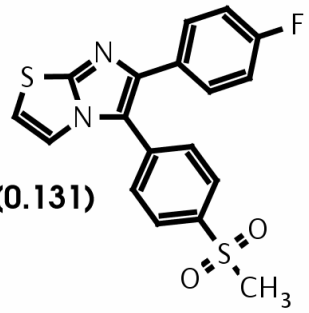


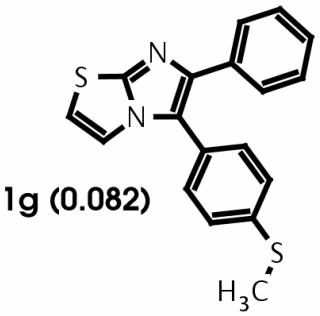


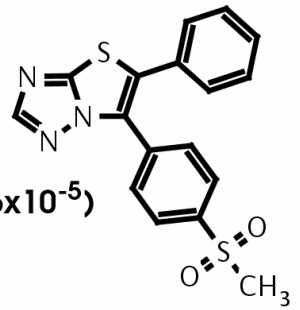






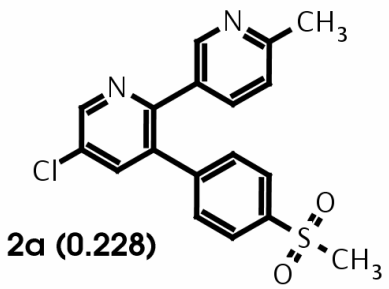




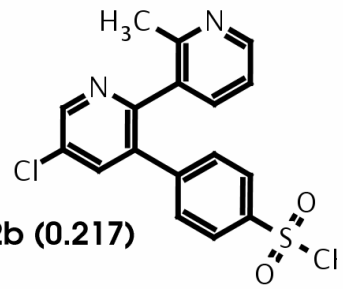


evotec

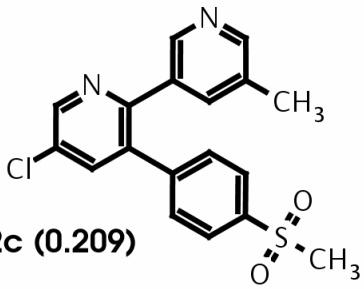
Cluster 2: Chloropyridines



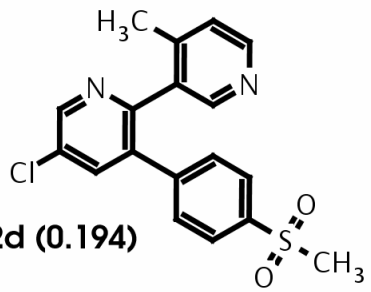
2a (0.228)



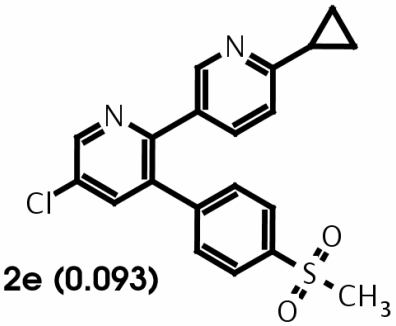
2b (0.217)



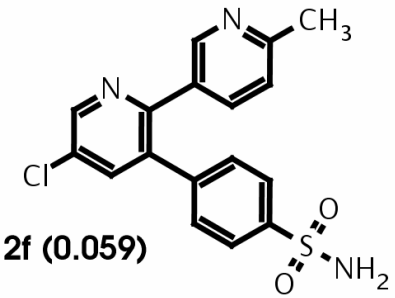
2c (0.209)



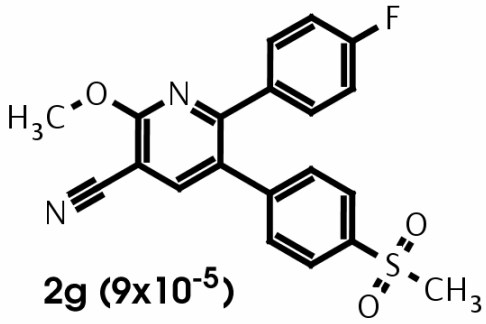
2d (0.194)



2e (0.093)



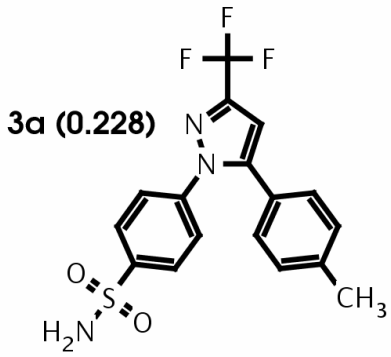
2f (0.059)

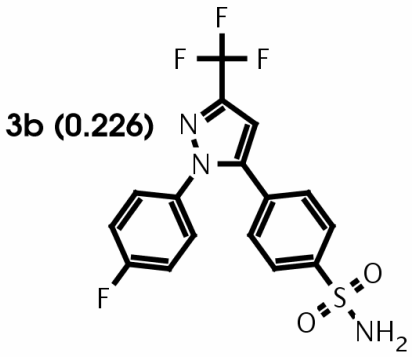


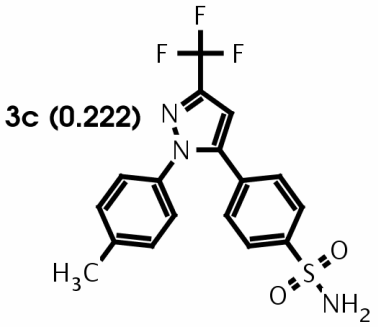
2g (9x10^{-5})

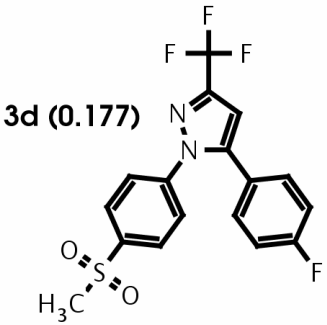
evotec

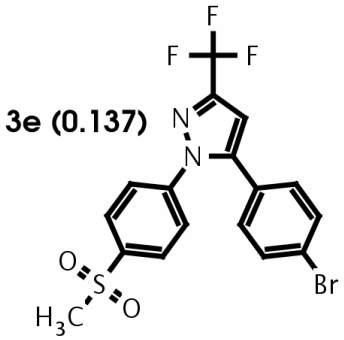
Cluster 3: Pyrazoles

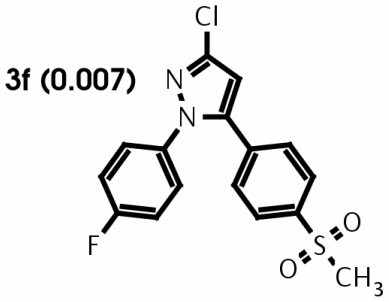


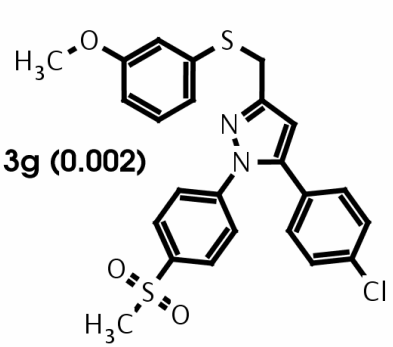


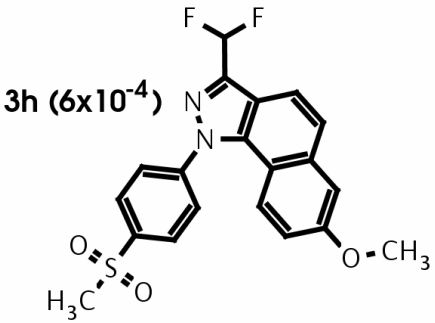






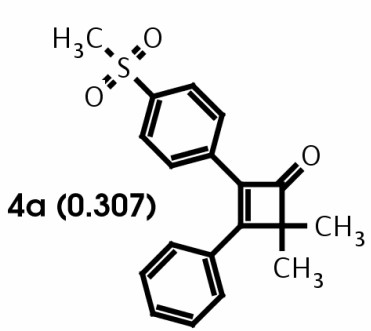


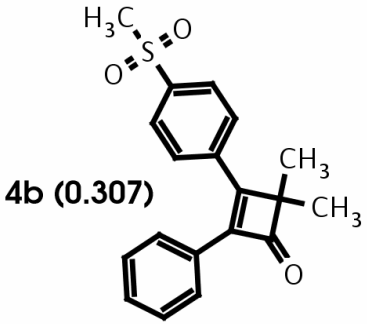


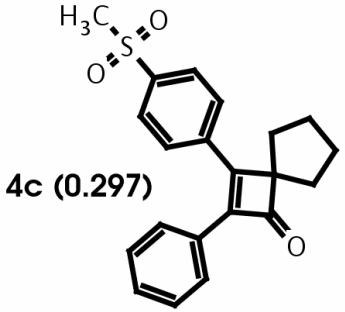


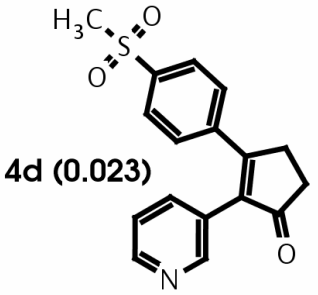
evotec

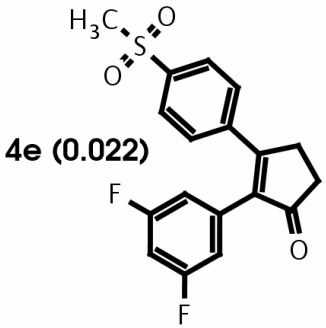
Cluster 4: Stilbenes

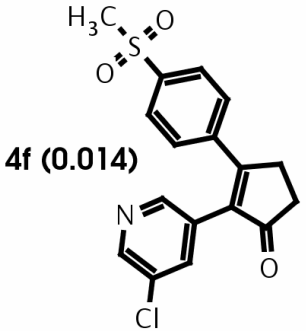


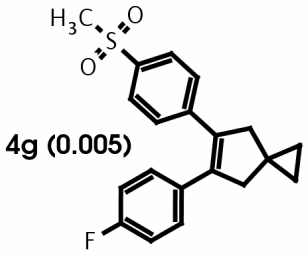


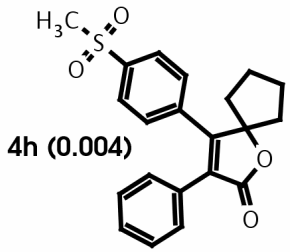






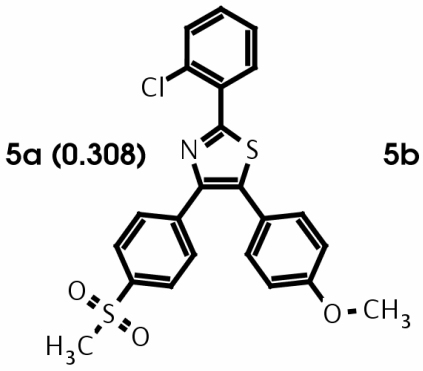


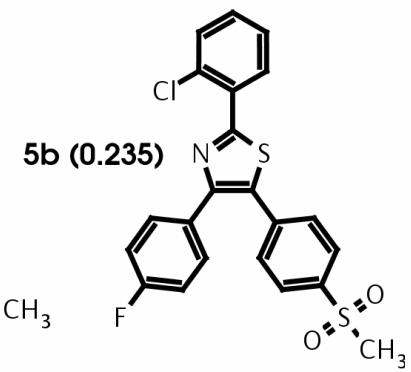


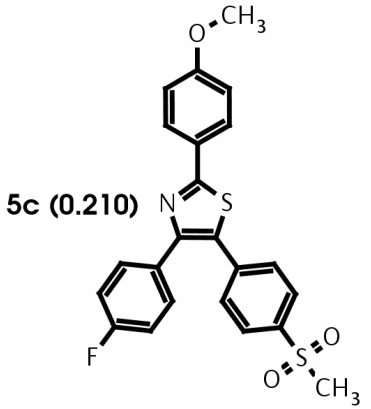


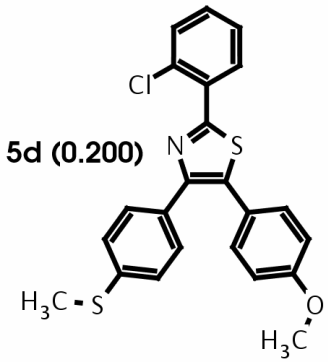
evotec

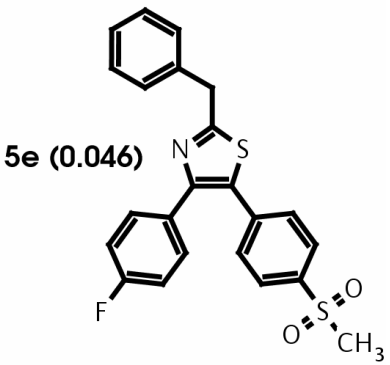
Cluster 5: Thiazoles



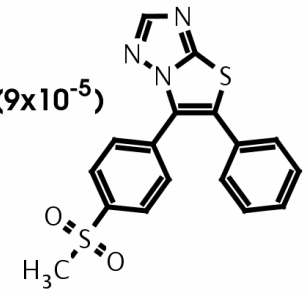








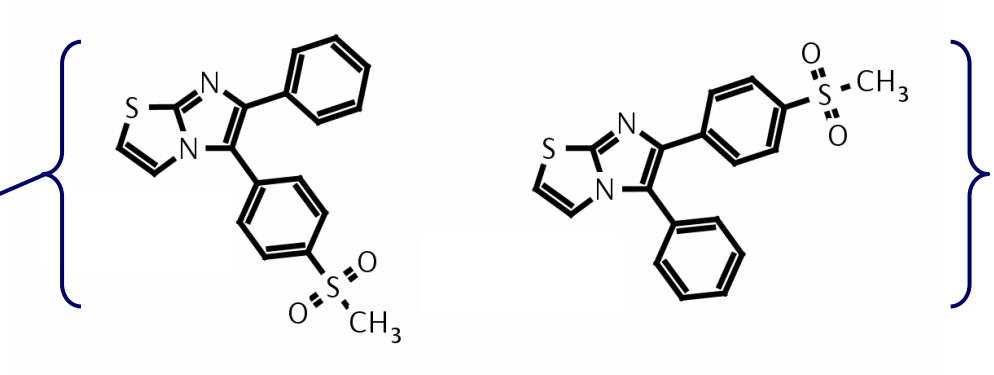
This molecule also made a minor contribution to cluster 1



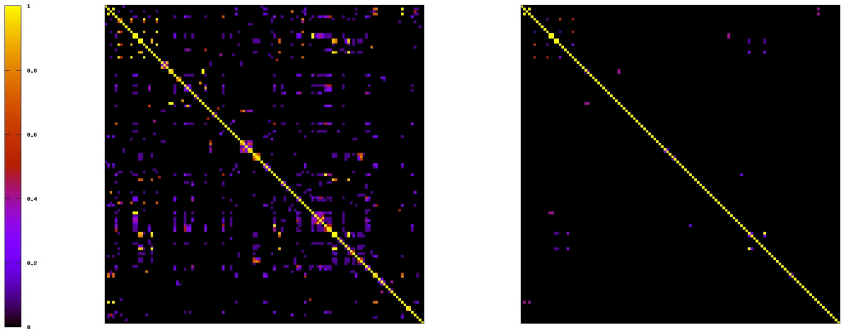
evotec

Effect of Gaussian width parameter (α)

	α					
	12.5	25	50	100	200	400
λ_1	6.367	5.586	4.752	3.806	2.917	2.272
1a	0.149	0.181	0.208	0.243	0.286	0.348
1b	0.149	0.181	0.208	0.243	0.286	0.348
1c	0.132	0.144	0.142	0.133	0.115	0.075
1d	0.131	0.141	0.139	0.134	0.119	0.097
1e	0.131	0.141	0.139	0.134	0.119	0.097
1f	0.127	0.131	0.118	0.098	0.073	0.034
1g	0.097	0.082	0.046	0.016	0.002	4×10^{-5}
1h	0.004	6×10^{-5}	3×10^{-8}	0	0	0



Normalised contributions from molecules in cluster 1

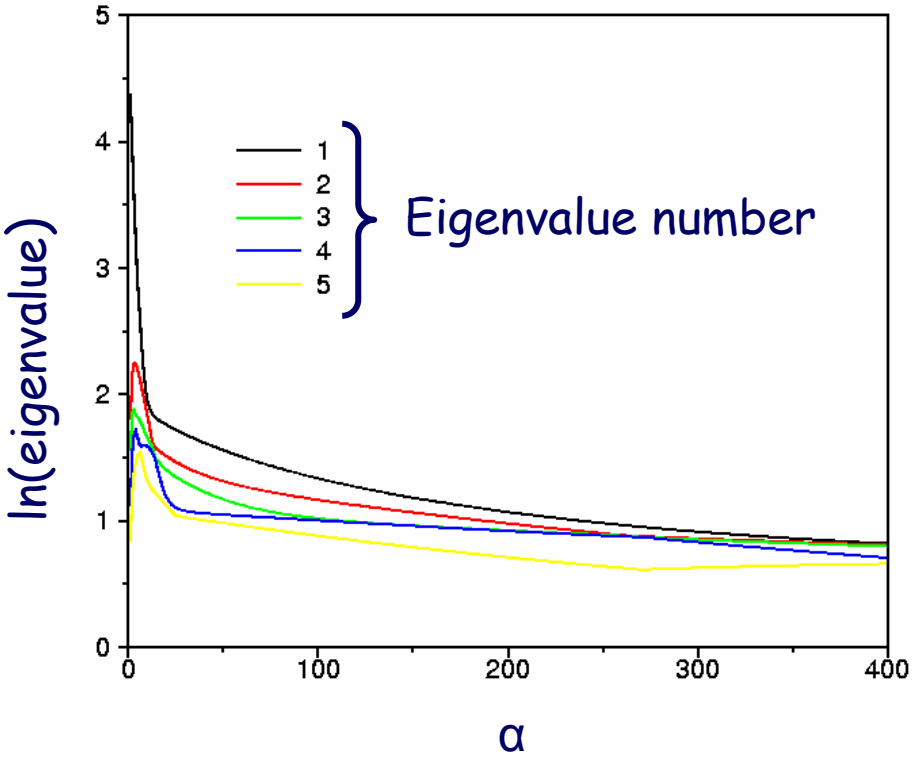


α = 12.5

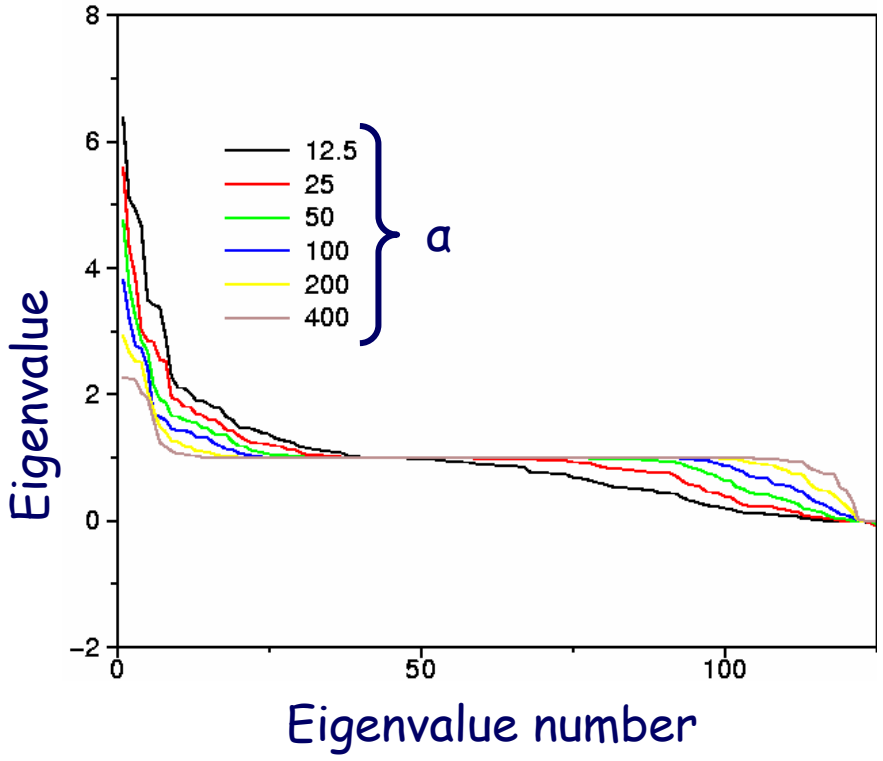
α = 400

evotec

Effect of Gaussian width parameter (α)



The five largest eigenvalues as a function of α

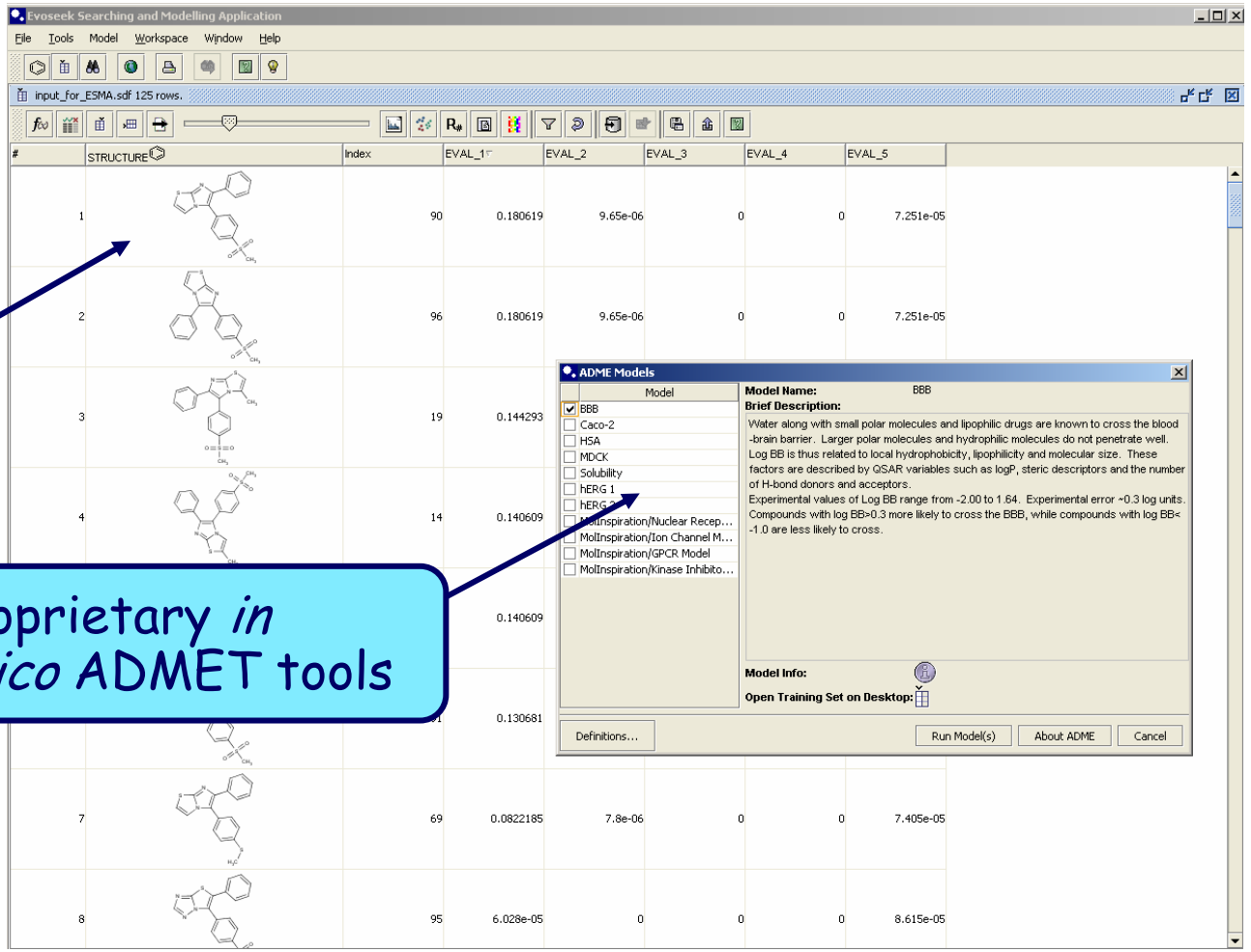


Magnitude of eigenvalues for different values of α

Provision of results

ESMA† Chemical Spreadsheet

Proprietary *in silico* ADMET tools



†EVOseek Searching and Modelling Application

evotec

Applications

- Analysis of screening data
  - Prioritisation of hit series
  - Visualising different scaffolds within a dataset
  - Identifying chemically related scaffolds within and between different clusters
- Diversity assessment
  - Virtual screening hit lists
  - Virtual libraries

evotec

Future and on-going work

- Investigate alternative molecular descriptors and/or similarity methods and/or datasets and/or filtering schemes
- Investigate alternative spectral clustering methods
- Implementations which scale more favourably with the number of molecules
  - Matrix diagonalisation scales as N^3 where N is the number of molecules
- Develop the intuitive understanding of spectral clustering methods
  - A global rather than local approach to clustering

evotec

Summary

- Demonstrated the benefit of spectral clustering
  - Eigenvalues and the normalised cluster contributions (eigenvectors) provide a natural way to analyse a dataset
  - Single tunable parameter
  - Overlapping and non-overlapping clusters possible
- The clusters of molecules produced have proven effective for medicinal chemists
- Further details of this work to appear soon:

  Mark Brewer, *"Development of a spectral clustering method for the analysis of molecular datasets"*, JCIM (accepted)

Acknowledgements

I would like to thank current and former colleagues at Evotec who have assisted with this work