

UK QSAR & Chemoinformatics

Spring Meeting

14th April 2005

Domain of Applicability

Quantitative Measure of Distance from the Domain

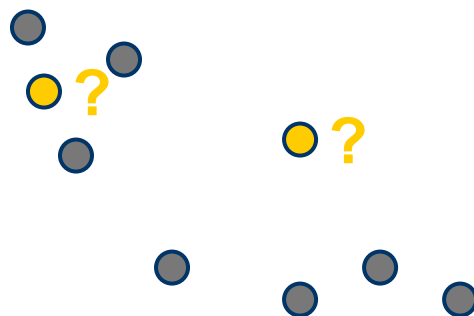
Robert Stanforth

Research Student

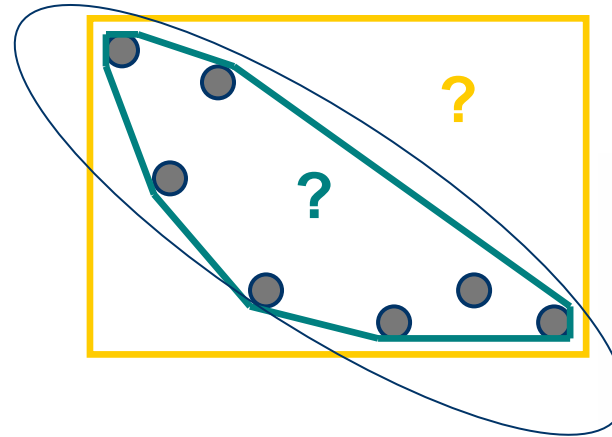
$$D_C = D_D + D_M + D_A - C$$

- Motivation
- Modelling the Dataset
- Measure of Distance from Domain
- Applications

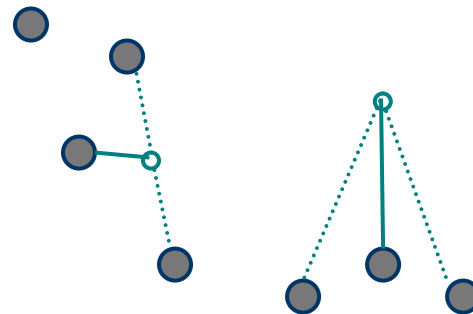
- QSAR model only valid in domain of its training set
- Measure membership of this 'domain of applicability'
- Provides assurance of:
 - External test set
 - Many-out validation
 - Dataset splitting into training and test sets
 - Prediction



- Bounding Box
- Convex Hull
- Distance to Centroid



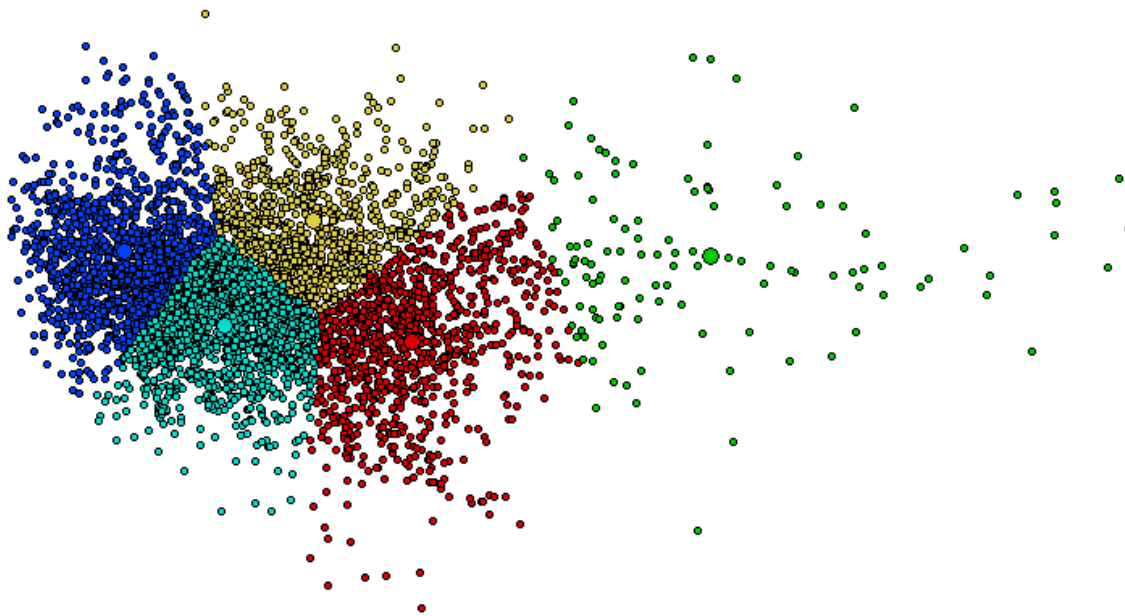
- Nearest Neighbour and k -NN Methods

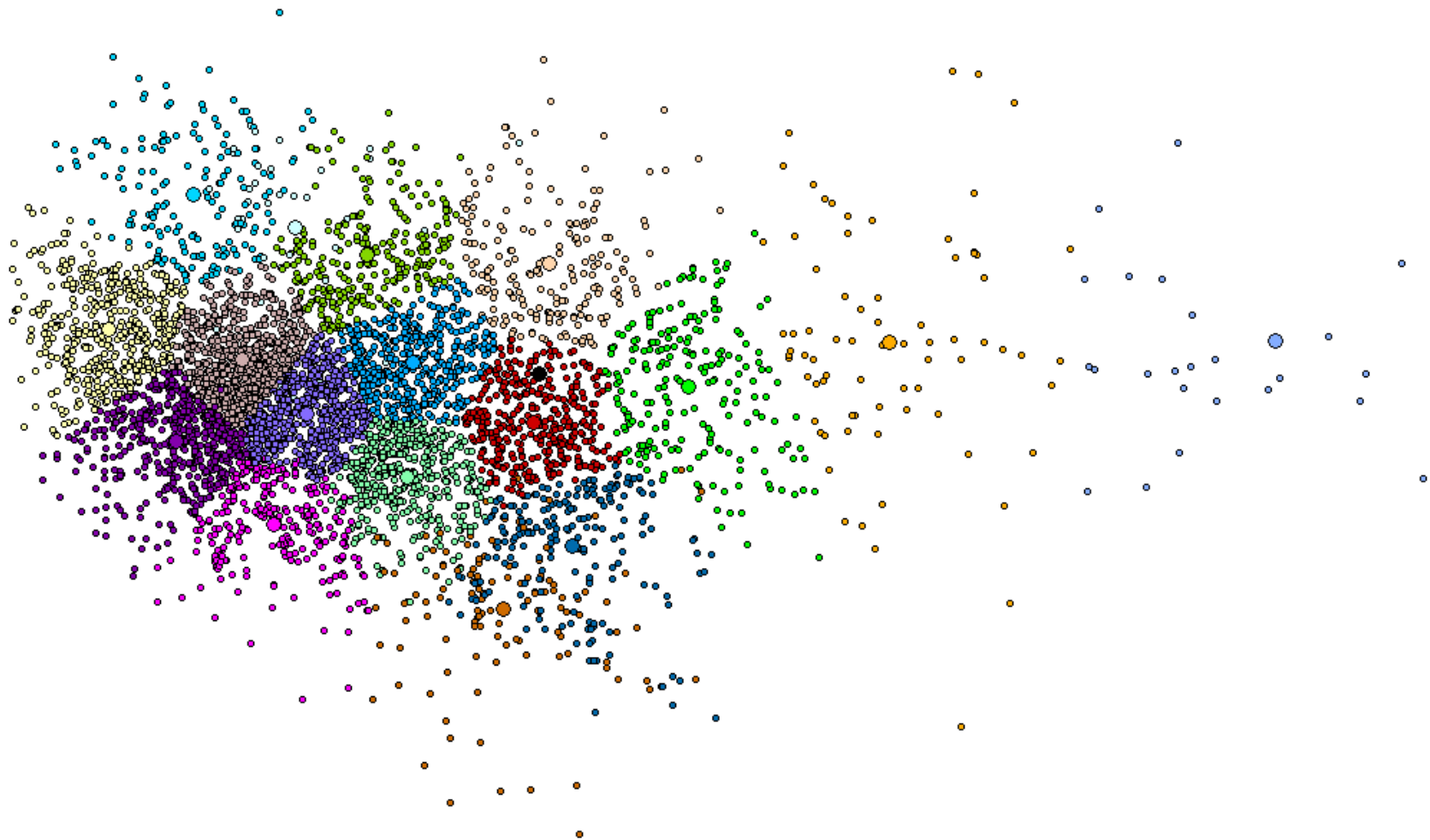


K-Means for Clustering

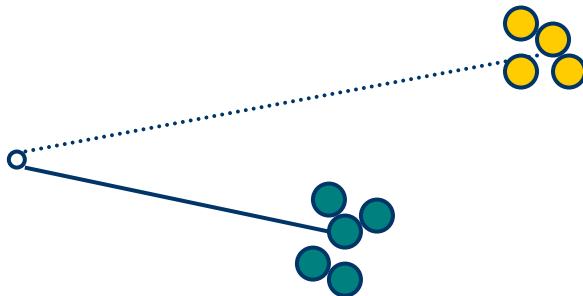
$$D_C = D_D + D_M + D_A - C$$

- Use 'clusters' to model the shape of the dataset
- *K*-Means algorithm iteratively adjusts partitioning into clusters to increase accuracy of the model
- Computationally feasible



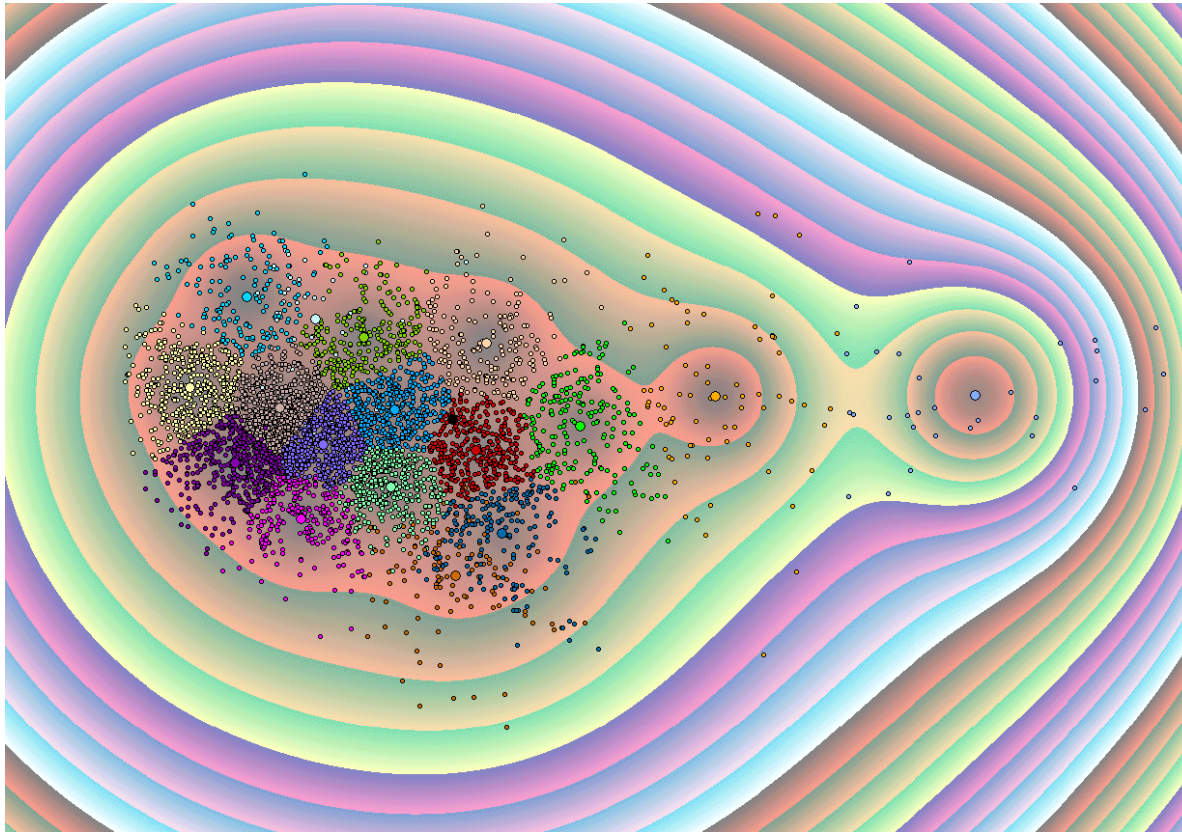


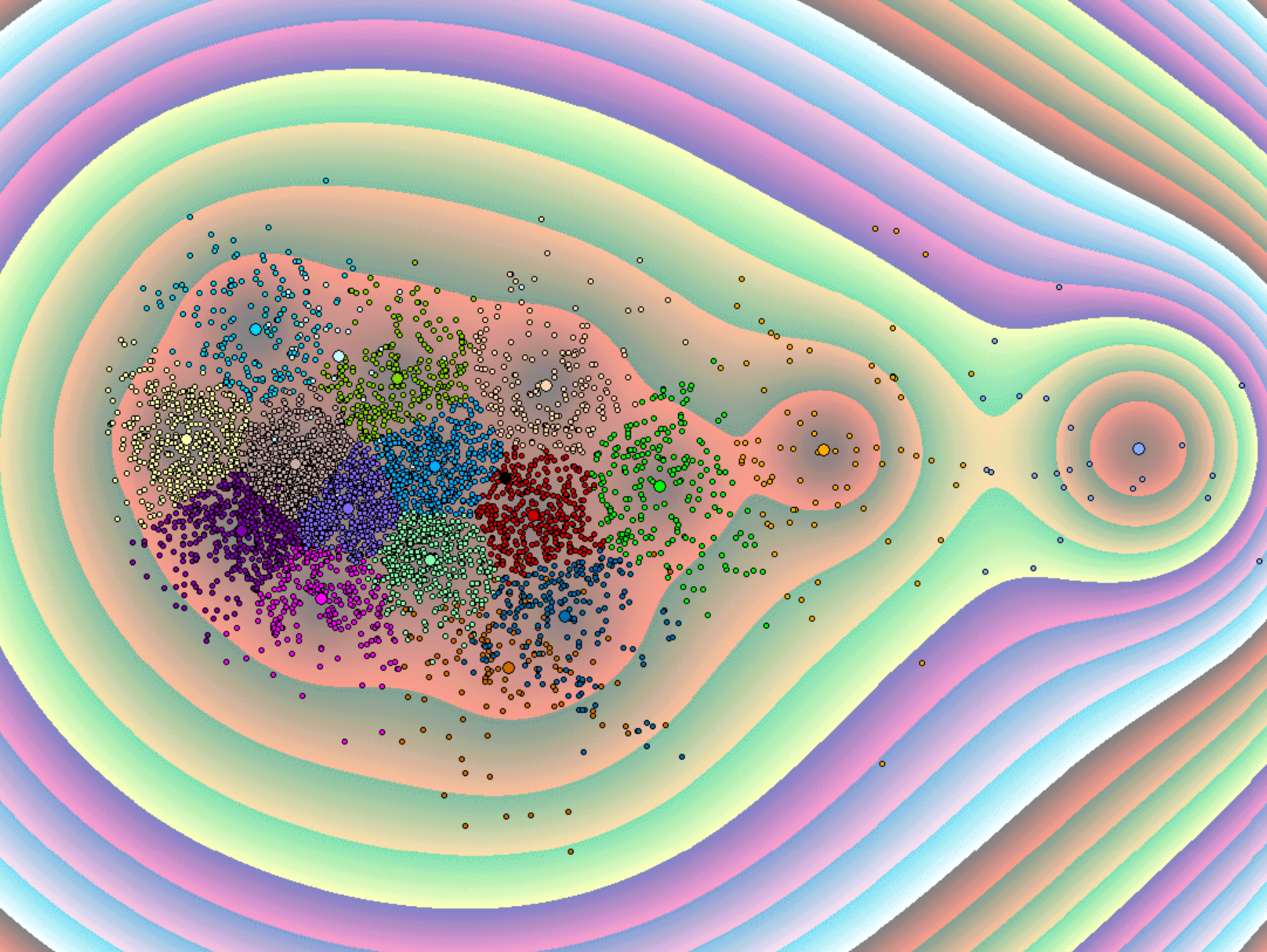
- Use the *K*-Means Model
 - Based on distances to cluster centroids
- Fuzzy cluster membership
- Weighted average of distances to cluster centroids, weighted according to cluster membership
- Computationally efficient

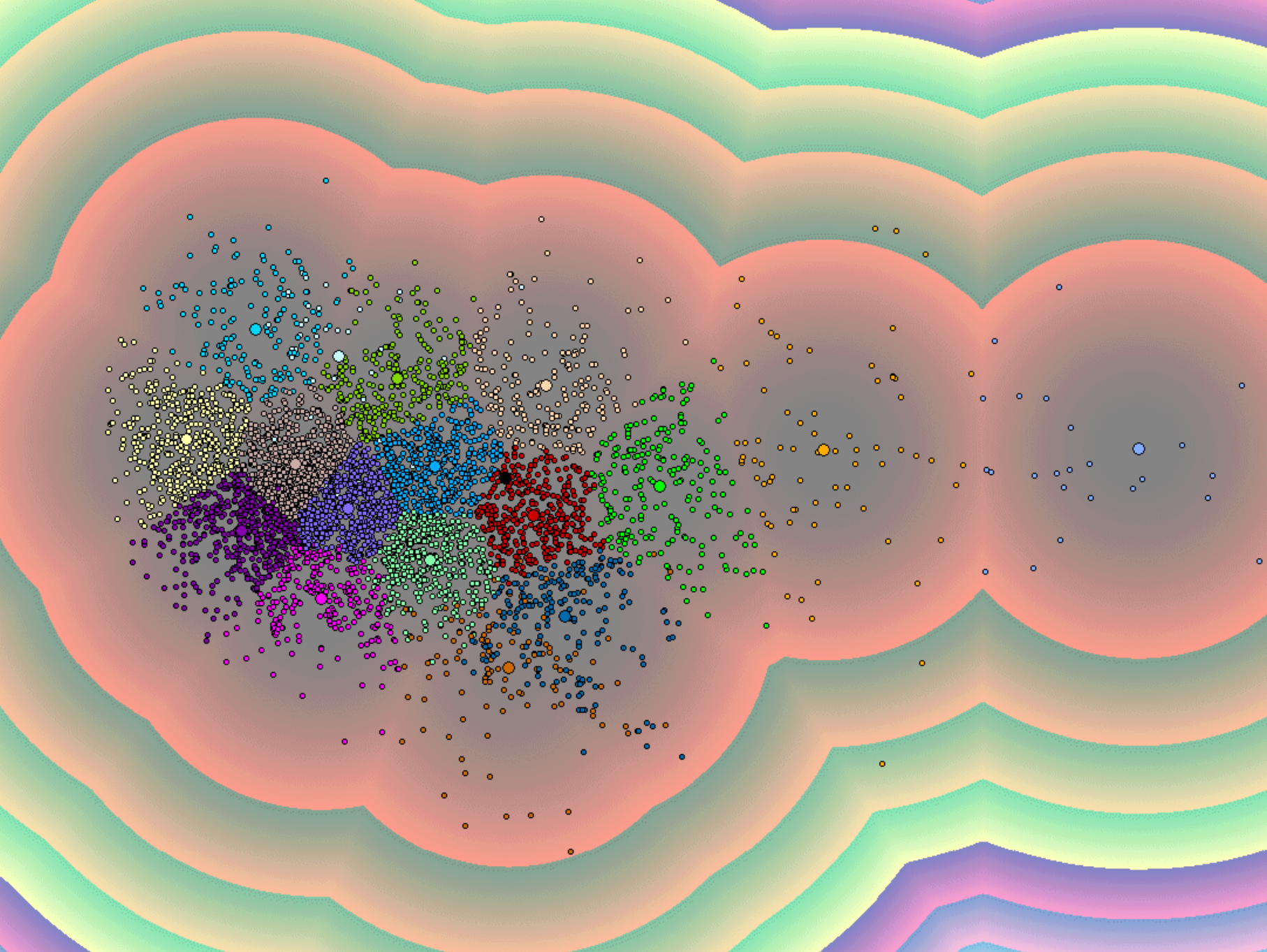


$$D_C = D_D + D_M + D_A - C$$

- Contour Plot
 - First contour defines boundary of applicability domain

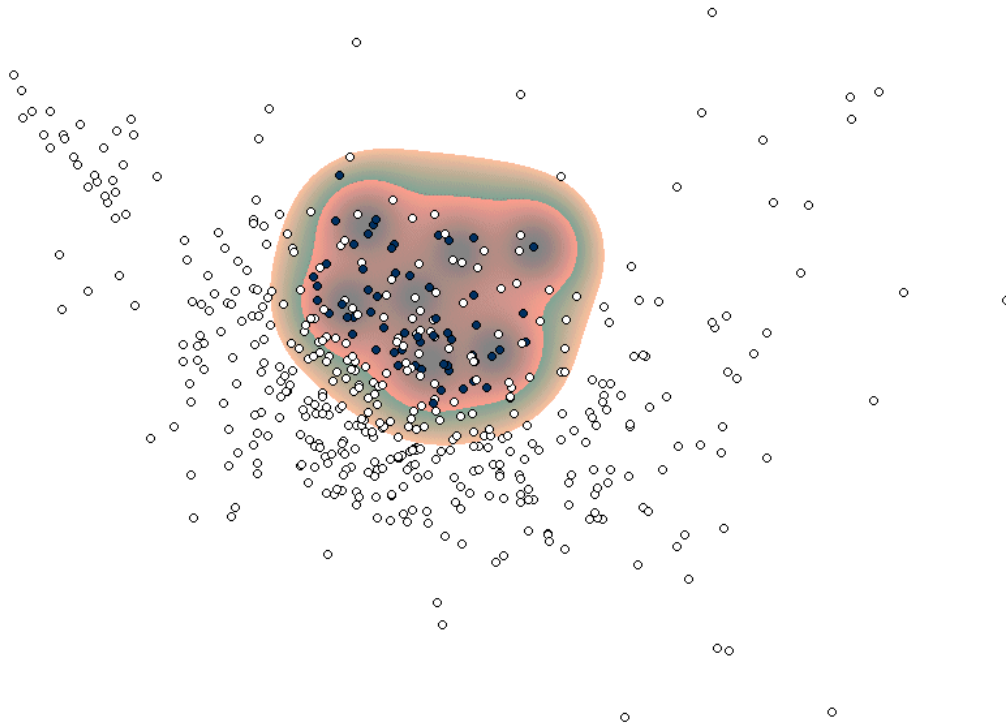






$$D_C = D_D + D_M + D_A - C$$

- Validate that distance to dataset measure correctly predicts that (only) similar structures are in the domain



Validation of Distance Measure

$$D_C = D_D + D_M + D_A - C$$

Iteration	P(false -ve)	P(false +ve)
1	6.8%	13.2%
2	2.4%	11.6%
3	4.3%	10.9%
4	8.0%	11.3%
5	5.0%	12.4%
6	6.3%	10.3%
7	5.1%	11.7%
8	3.7%	11.4%
9	6.8%	11.9%
10	6.8%	10.6%
Average	5.5%	11.5%

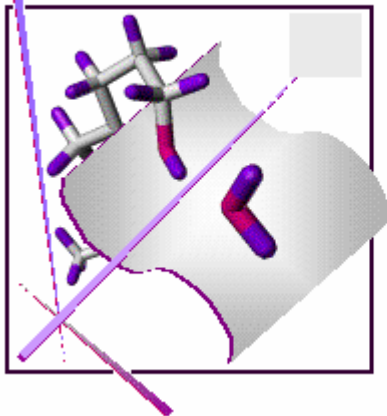
- Extraction of a Test Set
 - Test set must be in domain of applicability of training set
 - Training set should be in domain of applicability of test set, to ensure coverage
 - Measure this 'dual coverage' by aggregating distance-to-domain measurements
 - Use iterative algorithm to optimise 'dual coverage' measurement

- Need quantitative measure of applicability of a descriptor-based QSAR model to a structure
- Existing methods are all either too crude or too slow
- Our new method is computationally efficient, and copes well with non-convex domains

Acknowledgements

$$D_C = D_D + D_M + D_A - C$$

- Boris Mirkin, Birkbeck College
- Evgueni Kolossov, IDBS



UK QSAR & Chemoinformatics

Spring Meeting

14th April 2005