Data analysis in QSAR

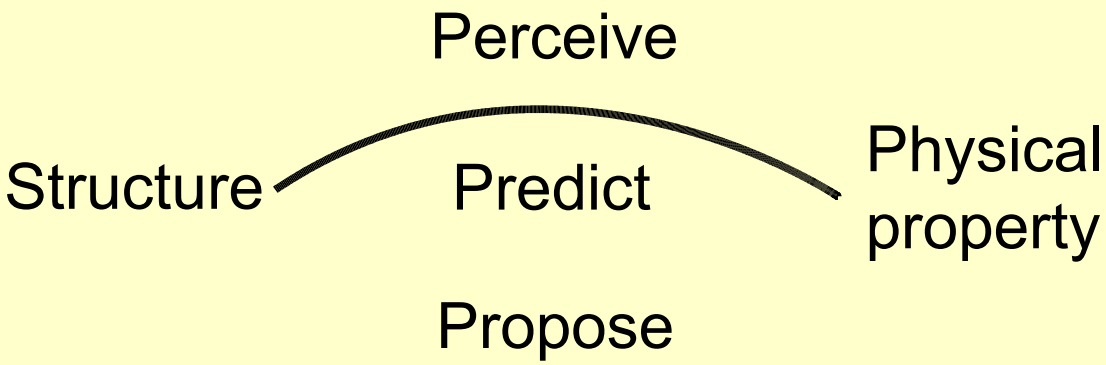
Noel O'Boyle

Dave Palmer, John Mitchell

Unilever Centre for Molecular Informatics, Cambridge

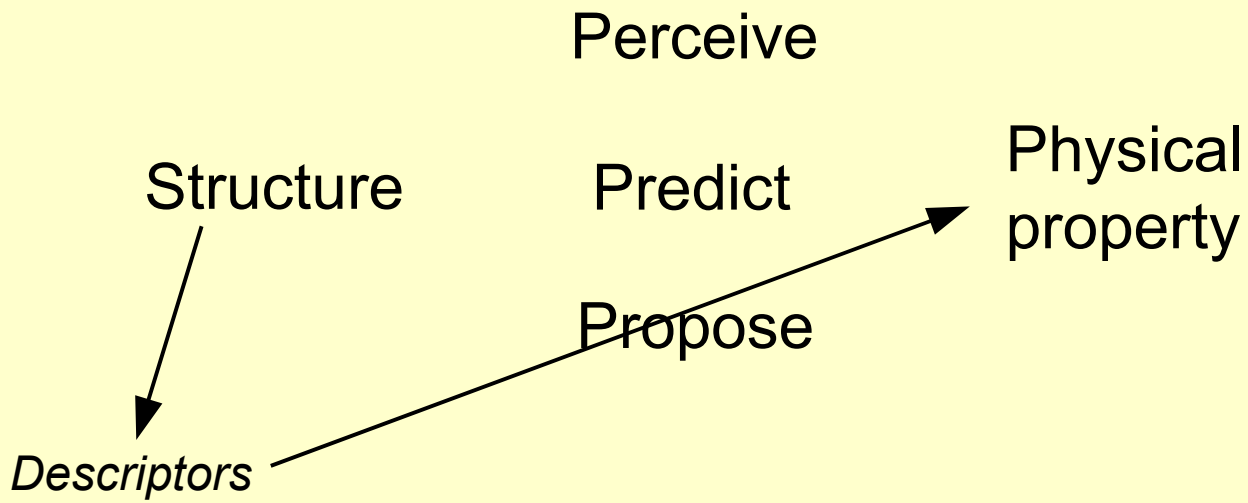
Quantitative Structure-Activity Relationship (QSAR)

Also QSPR (Property)



Quantitative Structure-Activity Relationship (QSAR)

Also QSPR (Property)



Molecular weight
No.of H-bonding groups
Surface area

MOE: 180 descriptors

Quantitative Structure-Activity Relationship (QSAR)

Optimise model on training set (2/3)
Test model on test set (1/3)

Need to use an estimate of the error of prediction (CV or bootstrap, but **not** resubstitution)

or

Build model on training set (2/3 of 2/3)
Optimise on test set (1/3 of 2/3)
Test model on validation set (1/3)

Using the error of prediction

Quantitative Structure-Activity Relationship (QSAR)

Optimise model on training set (2/3)
Test model on test set (1/3)

Need to use an estimate of the error of prediction (CV or bootstrap, but **not** resubstitution)

or

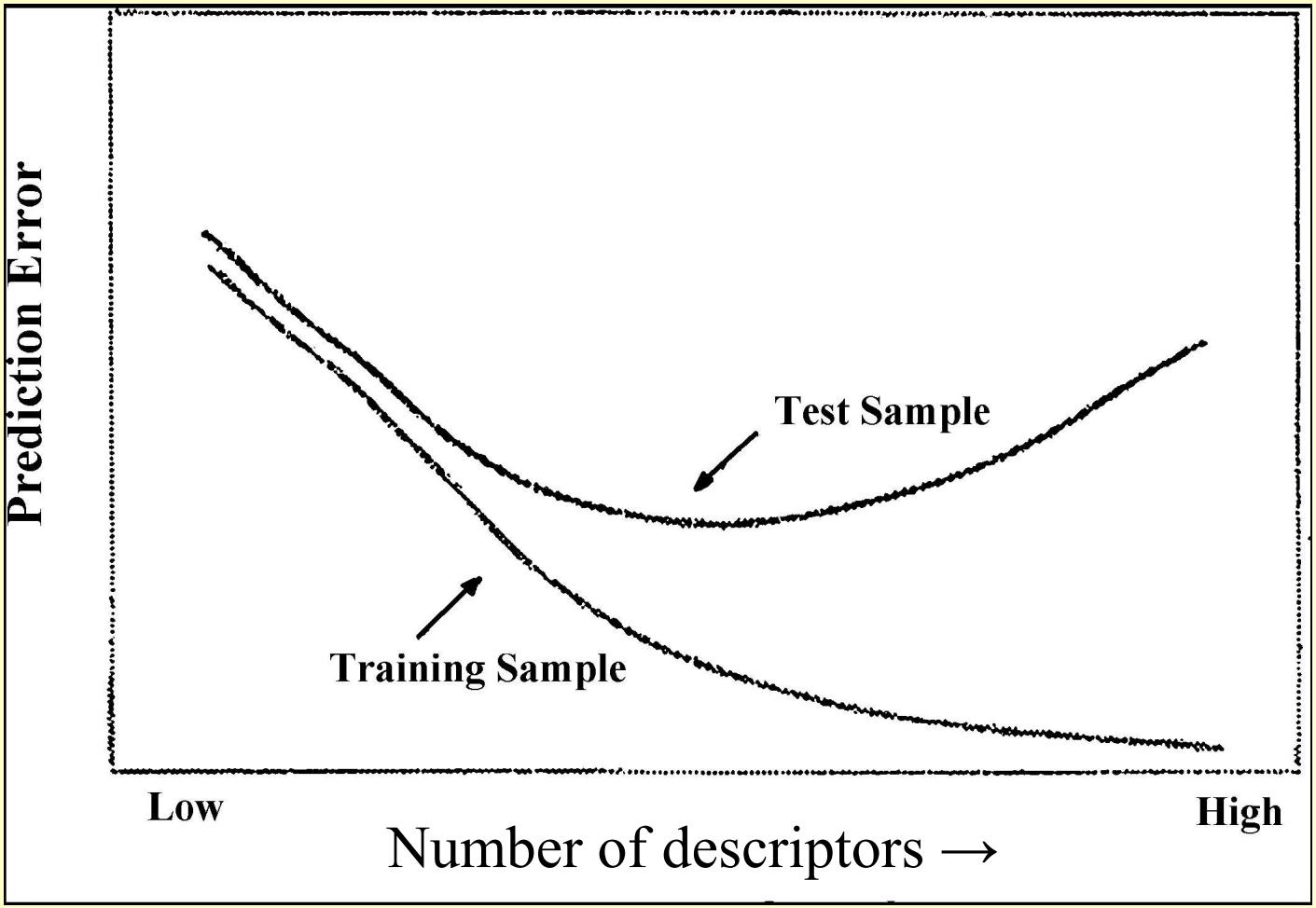
Build model on training set (2/3 of 2/3)
Optimise on test set (1/3 of 2/3)
Test model on validation set (1/3)

Using the error of prediction

Testing must be carried out on a hold-out set from the same distribution as the the set of molecules used for optimisation
=> don't optimise model on diverse set

Assessing Model Fit by Cross-Validation, JCICS, 2003, 43, 579.
Beware of q^2, J.Mol.Graph.Model., 2002, 20, 269.

Feature selection

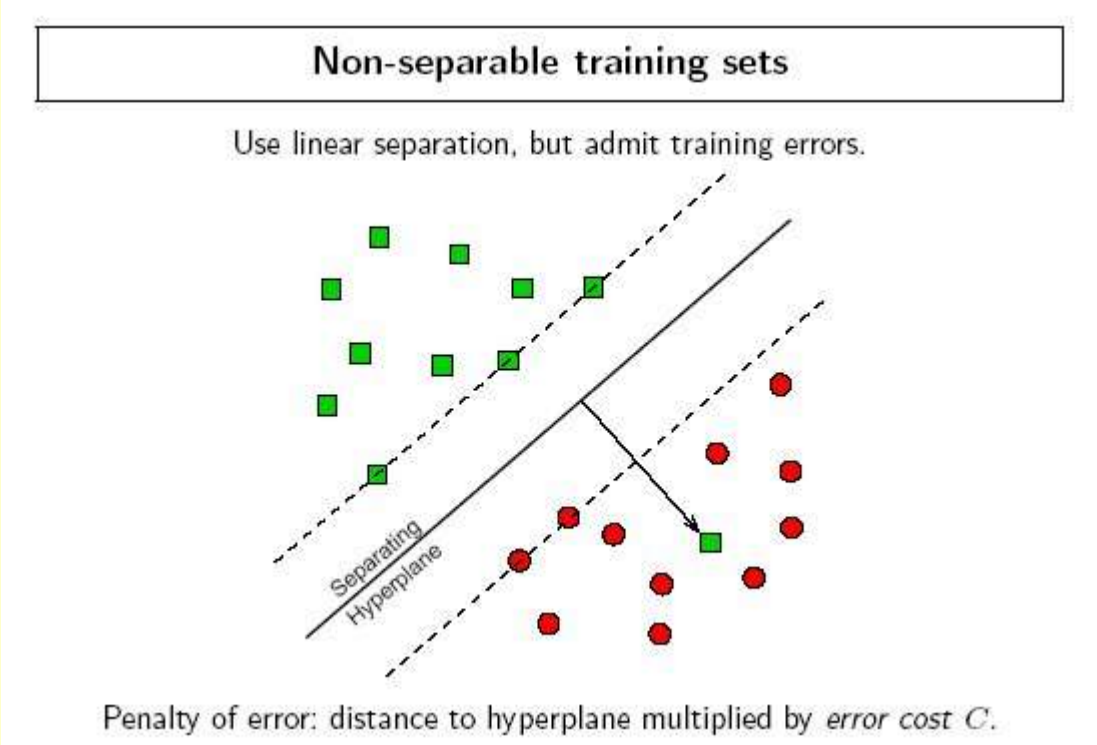


Features = descriptors

2^n different subsets

Parameter optimisation

- Models have parameters which should be tuned
  - Support vector machines
    - gamma, cost, epsilon

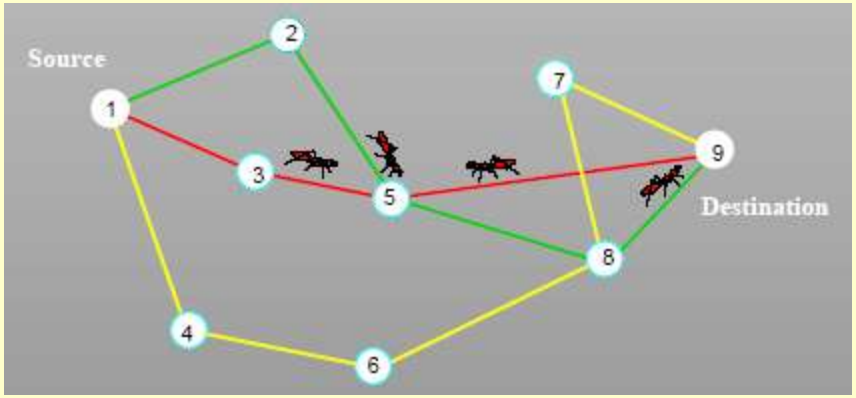


QSAR: Feature selection and parameter optimisation

- **Aim**: To find a robust method of simultaneously optimising the parameters and performing feature selection
- Prediction of solubility for drug-like molecules
  - David Palmer
  - Support vector machines
    - gamma, cost, epsilon
  - 127 descriptors
- Large search space
  - stochastic algorithm required

Ant Colony Optimisation (ACO)

- Insipired by behaviour of ants foraging for food
- Ants lay down pheromones, which influence the path taken by other ants. Meanwhile pheromones are evaporating.
- Ants' trails converge to shortest path between nest and food

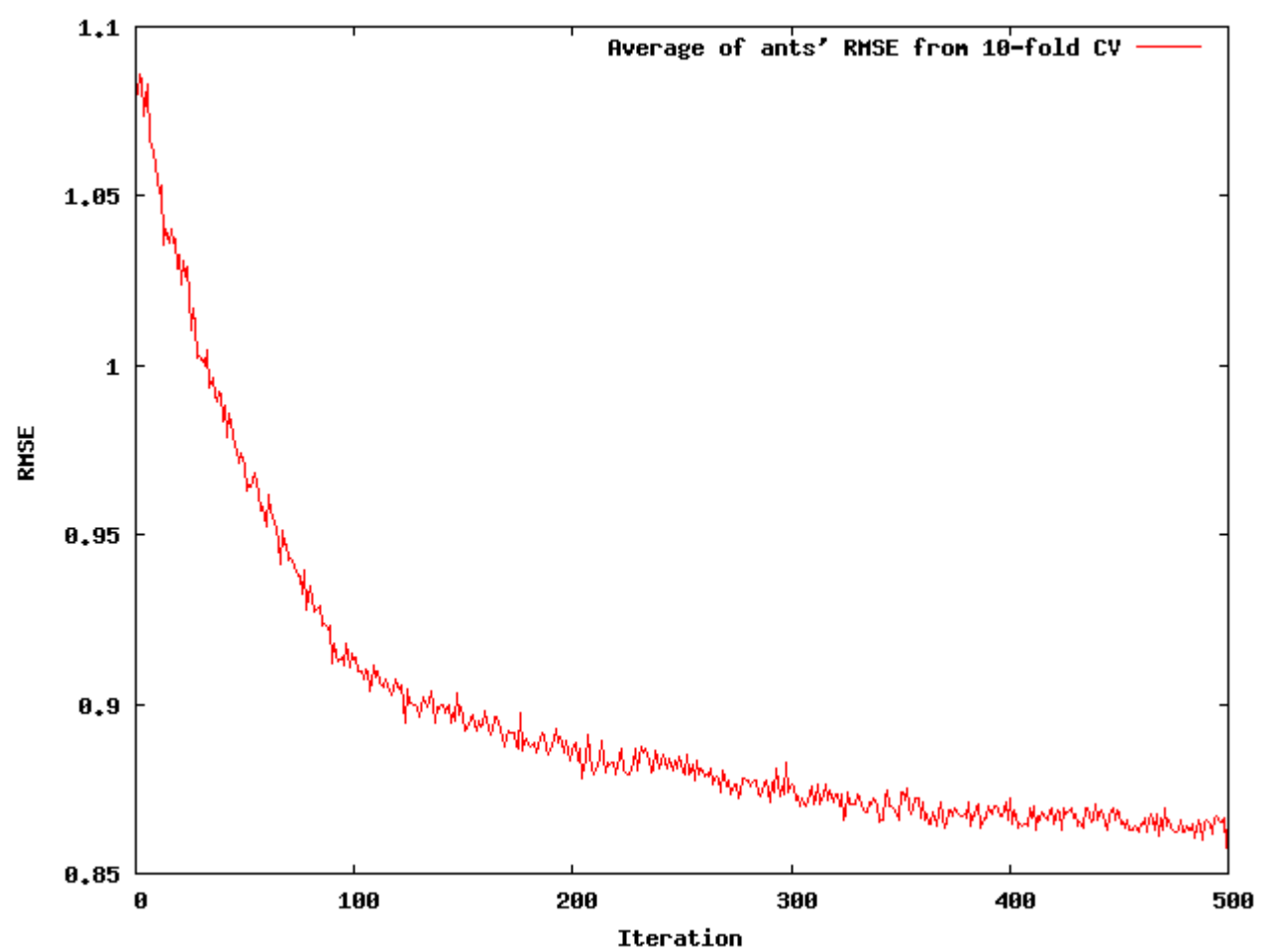


- Ant Colony Optimisation (ACO) – Marco Dorigo, PhD Thesis, 1992.
- ACO for feature selection – Shen et al, JCIM, 2005, 45, 1024.
- We have extended it to perform simultaneous parameter optimisation

Ant Colony Optimisation (ACO)

- population of ants (typically 50 to 100)
  - each ant represents a model – i.e. a subset of descriptors and values for the parameters
  - each ant has a fitness score, e.g. 10-fold cross validation rmse
- it is more likely that an ant will choose a particular descriptor/parameter value in the next iteration if
  - many ants have chosen it in this iteration (local search), or
  - many ants have chosen it in their best models to date (global search)
- for descriptors/parameter values that are not chosen in the current or best models, the probability that they will be chosen decreases (evaporation)

Does it work?


Average of ants' RMSE from 10-fold CV
RMSE
Iteration
1.1
1.05
1
0.95
0.9
0.85
0
100
200
300
400
500