

# Computational Chemistry - part III: Chemoinformatics

Tomas Öberg  
University of Kalmar, Department  
of Biology and Environmental  
Science

2003-02-14

# Content of the three seminars

- First meeting
  - Some concepts in molecular modelling
  - A brief introduction to *ab initio* modelling
  - Empirical force field models: Molecular mechanics
- Previous meeting
  - Energy minimisation
  - Optimisation methods
- Today
  - Chemoinformatics
  - Molecular descriptors
  - QSAR/QSPR, chemometrics

2003-02-14

# Literature

- Andrew R. Leach (2001): *Molecular modelling - principles and applications*, 2nd edition, Prentice-Hall, 744 pp.
- Mati Karelson (2000): *Molecular descriptors in QSAR/QSPR*, Wiley, 430 pp.
- Rene P. Schwarzenbach, Philip M. Gschwend and Dieter M. Imboden (2003): *Environmental organic chemistry*, 2nd edition, Wiley, 1313 pp.
- Harald Martens and Tormod Næs (1989): *Multivariate calibration*, Wiley, 419 pp.
- Öberg, T. *Env Sci Pollut Res* **9**, 405-411, 2002.

# Chemoinformatics

- A generic term that encompasses the <sup>1/</sup>design, <sup>2/</sup>creation, <sup>3/</sup>organisation, <sup>4/</sup>management, <sup>5/</sup>retrieval, <sup>6/</sup>analysis, <sup>7/</sup>dissemination, <sup>8/</sup>visualisation and <sup>9/</sup>use of chemical information.
- Some related terms are: Chemometrics, computational chemistry, and chemical informatics

2003-02-14

## Close link to drug discovery

*“The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization”.*

Frank Brown in Annual Reports of Medicinal Chemistry (1998)

# What is left?

- Previous two lectures have covered some of what is included in chemoinformatics, so what is left?
  - Organisation and retrieval of chemical information from databases
  - The analysis and use of structure information
- Today the focus will be on the last part:
  - Molecular descriptors
  - QSAR/QSPR
  - Chemometrics

2003-02-14

# Similarity

- Similarities are important when we assess, model and search for chemical substances
- It can relate to the whole molecule or some active part, so “activity” is another important concept
- How can similarity be defined and measured? What do we need?
  - Molecular descriptors, i.e. numbers that relate to structural properties of the molecule

2003-02-14

# Molecular descriptors

- Can you give some examples of molecular descriptors?
  - Molecular weight
  - Atom counts
  - Boiling point
  - Octanol/water partitioning
- What about the nature of these four examples, is there an obvious grouping?
  - Theoretical descriptors, also for compounds that not yet exist
  - Empirical descriptors, measured properties

# Classification of theoretical descriptors

- Constitutional descriptors
  - MW, atom and bond counts, etc.
- Topological descriptors (indices)
  - Atomic connectivity in the molecule
- Geometrical descriptors
  - From 3D-structure, e.g. van der Waals radii and charge distribution
- Mixed or combined descriptors

# Constitutional descriptors

- Reflect only chemical composition, without reference to the geometry
- Difficult to understand and calculate?
  - No, of course not. Constitutional descriptors are attractive because of the simplicity
- Historically, one of the first approaches of this type was suggested by Free and Wilson (1964)

$$P = P_0 + \sum_k c_k I_k$$

$P = \text{property}, I_k = \text{structural feature}$

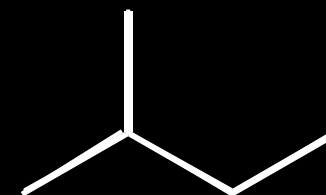
# Topological descriptors

- These descriptors are based on connection table representations of the structure and employ methods drawn from mathematical graph theory
- Topological descriptors are calculated from different combinations and weightings of atoms (vertices) and bonds (edges)
  - Path 1 molecular connectivity  ${}^1\chi_p$  (Randic) is an example of a simple topological descriptor

$${}^1\chi_p = \sum_{edges} \frac{1}{\sqrt{mn}}, \text{ } m \text{ and } n \text{ are the degrees of the adjacent vertices}$$

## Example: 2-Methylbutane

- Has 4 edges thus 4 terms, and for each term we just count the number of bonds attached to each atom participating in the bond



$${}^1\chi_p = \frac{1}{\sqrt{1 \cdot 3}} + \frac{1}{\sqrt{1 \cdot 3}} + \frac{1}{\sqrt{3 \cdot 2}} + \frac{1}{\sqrt{2 \cdot 1}} = 2.27$$

- This then encodes information about branching

# Geometrical descriptors

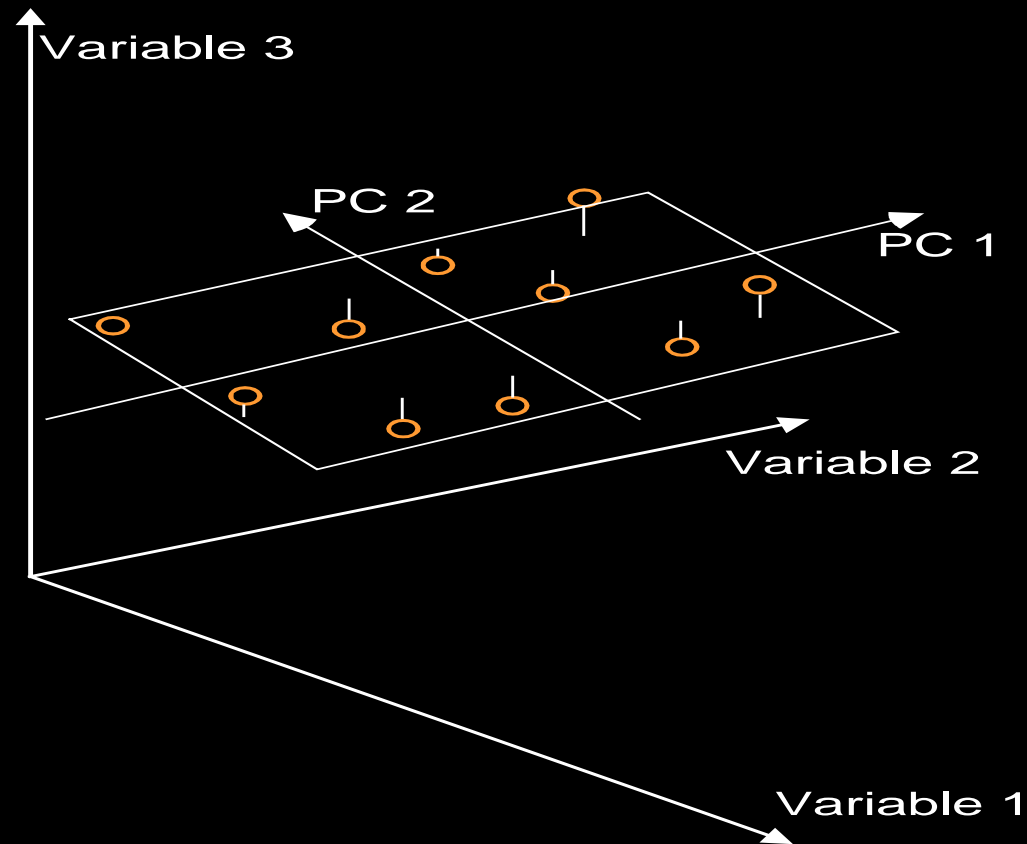
- Molecular surface area, molecular volume and gravitational indices (reflecting both shape and distribution of mass) are some examples of popular geometrical descriptors
  - The weighted holistic invariant molecular (WHIM) descriptors is another example and calculated by principal component analyses (PCA) on centred molecular coordinates using different atom weight schemes

2003-02-14

# Similarity based on molecular descriptors

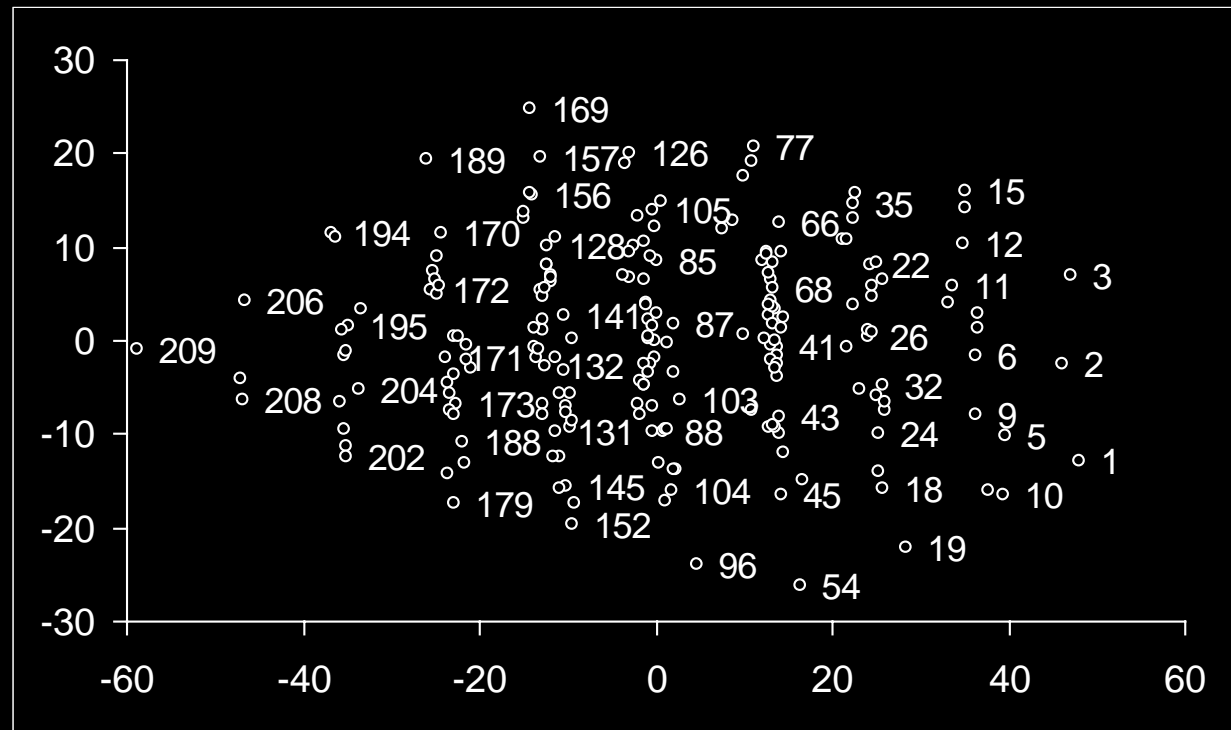
- How can we capture the similarities based on several molecular descriptors?
  - Euclidean distance, Hamming distance, etc.
  - Use a projection technique to reduce dimensionality
- Chemometric techniques, like PCA, are often used to reduce dimensionality and enable the simultaneous use of hundreds or thousands of descriptors
  - To visualise similarities and differences
  - To model the relationships between structure and properties, i.e. biological activity

# Principles of projection



2003-02-14

# First two principal components for PBDE



2003-02-14

# QSAR/QSPR

- Quantitative structure-activity and structure-property relationships are major application areas for molecular modelling and chemoinformatics.
- The first reports on QSAR however dates back more than a hundred years
  - At the end of the 19<sup>th</sup> century Meyer and Overton correlated potencies of narcotics with the partition coefficient between oil and water
  - In the 1940s Hammett recognised the effects of substituents on the dissociation of benzoic acids
  - But modern use of QSAR is usually attributed to Hansch (1960s)

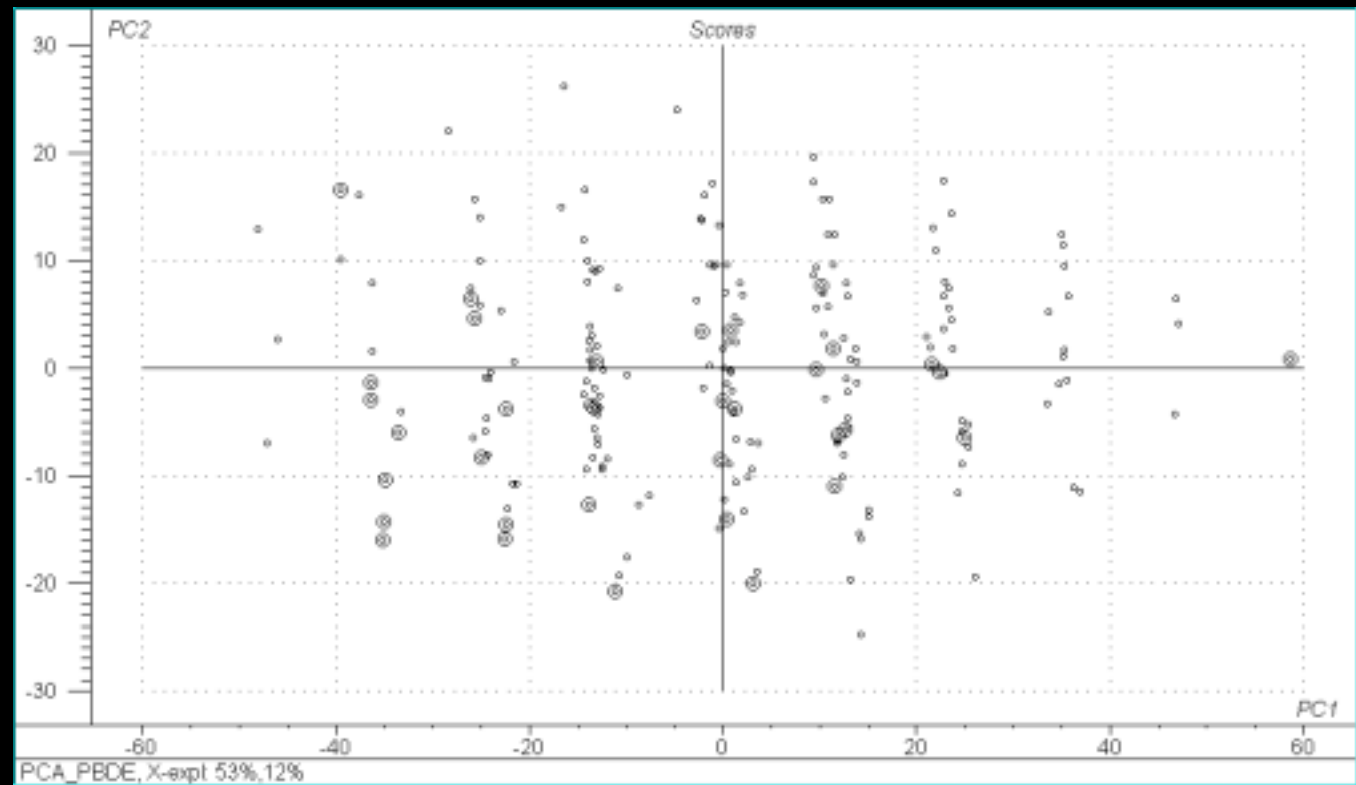
2003-02-14

# Developing QSAR/QSPR

- What is then needed to develop these relationships?
  - First we need reference compounds with experimentally determined activity
- Can molecular modelling assist us in the synthesis work?
  - Yes, prior modelling (PCA) and statistical experimental design can assist in developing candidates with enough structural variation
- What next?
  - Develop the models

2003-02-14

# An example of candidate selection: PBDE again



2003-02-14

# Linear free energy relationships - a theoretical justification?

- Many molecular properties are related to partitioning
- Often the free energy of transfer is unknown, one approach can then be to relate to the free energy of transfer in another system:

$$\Delta_{12}G_i = a \cdot \Delta_{34}G_i + \text{const}$$

- A more general approach is to assume that the free energy of transfer for the whole molecule can be expressed by a linear combination of terms that describe parts of the molecule:

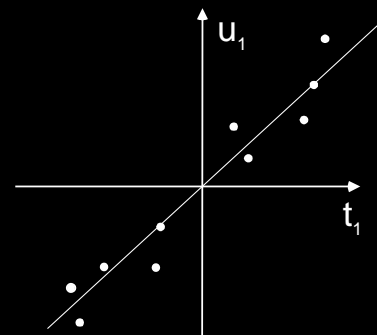
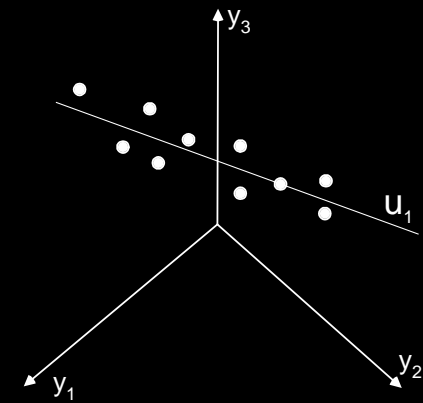
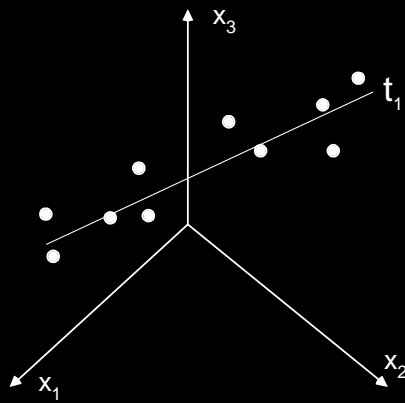
$$\Delta_{12}G_i = \sum \Delta_{12}G_{\text{part of } i} + \text{interaction terms}$$

- ***With this approach we can then estimate the property based solely on compound structure!***

# Some challenges in modelling

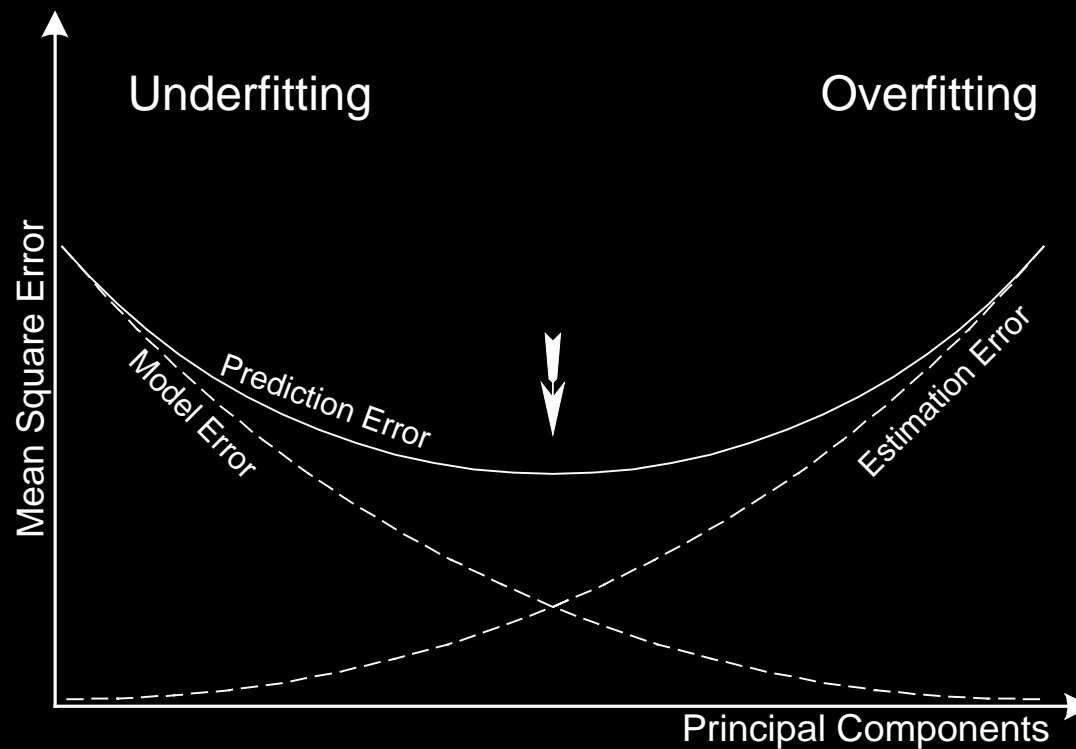
- Multiple linear regression cannot handle more variables (descriptors) than objects (compounds), how to deal with this?
  - Reduce the number of variables, selection strategy, any drawbacks?
  - Use redundancies in the data to reduce dimensionality, e.g. bilinear regression
- How can we decide if a model is good?
  - Test with new compounds/data?
  - Use existing data for validation, cross validation or test set

# Partial least squares regression (PLSR)



2003-02-14

# Prediction error and model quality



2003-02-14

# Vapour pressures of 418 PXDE congeners - a practical example

- Polychlorinated and polybrominated diphenyl ethers (PCDE and PBDE) are substances of environmental concern
- The vapour pressure is an important physical property that influence transport and distribution in the environment
- There are 209 congeners of each substance group, but experimental data is only available for a minor portion of these

2003-02-14

# Purpose

- The purpose of the investigation was to characterise the different PCDEs and PBDEs using computationally derived molecular descriptors
- Multivariate calibration was subsequently used to investigate the relationship between the molecular descriptors and experimentally determined vapour pressures

2003-02-14

# Available experimental data

- Experimentally determined values for subcooled liquid vapour pressures at 25 °C were obtained from the literature
  - vapour pressures were reported for 106 PCDE congenersand
  - in two investigations for 23 respectively 14 PBDE congeners (overlapping for 7)

# Descriptor generation

- The chemical structure of each PCDE and PBDE congener was sketched on a PC using the software HyperChem
- Each congener was modelled using the force-field routine MM+, an extension by HyperCube of the standard MM2 force field
- The molecular structures were then used as input for the generation of 795 descriptors with the software Dragon v1.11 (Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milano, Italy)

2003-02-14

# Multivariate calibration

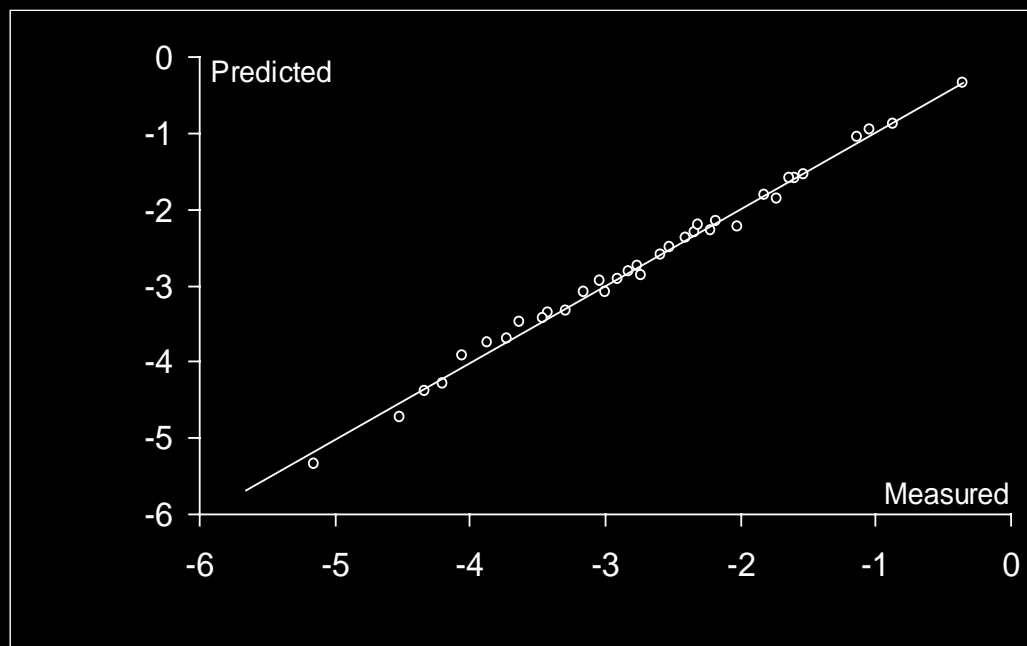
- The PLSR1-method was used to obtain calibration models with high accuracy and precision
- A further step to get parsimonious models was to assign zero weight to descriptor variables with minor influence in the regression (identified with a jack-knife method)
- All descriptor variables were pre-processed by auto scaling to zero mean and unit variance
- The vapour pressures were log-transformed prior to modelling

2003-02-14

# Model validation

- Cross-validation was used to establish the rank of the calibration models (number of latent variables),
- An external test set was used to estimate the prediction error
- Each calibration model was characterised by the standard error of calibration (SEC), explained variance in the calibration data ( $R^2$ ), standard error of prediction (SEP) and explained variance in the validation data ( $Q^2$ )

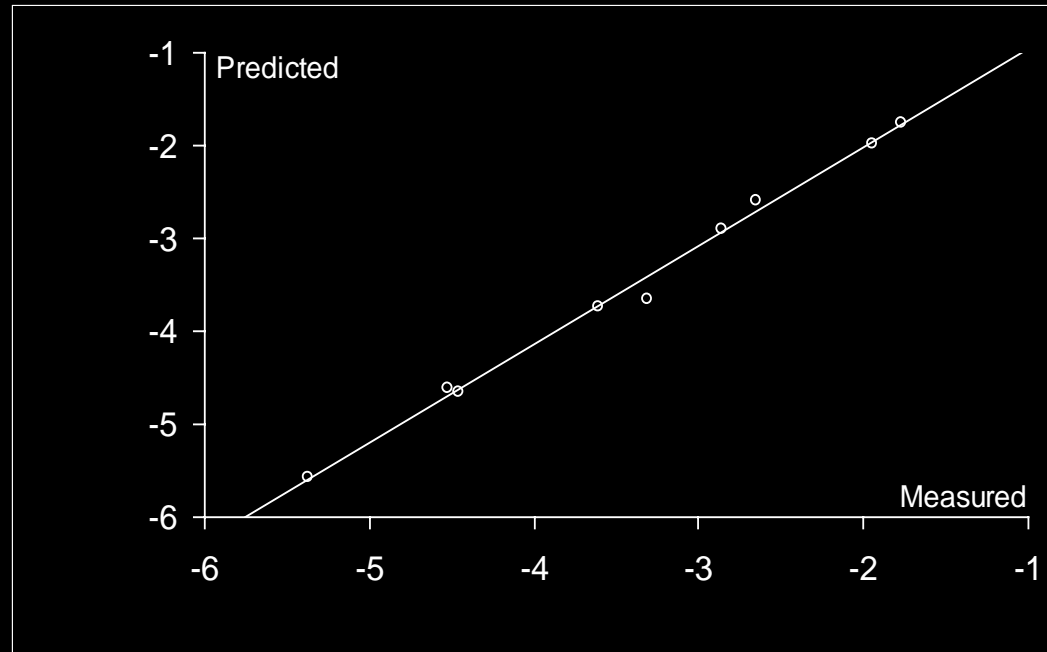
## Results - PCDE (external test set with 35 congeners)



The model performance parameters were SEC 0.069 (log Pa),  $R^2_{\text{Cal}}$  0.996, SEP 0.090 (log Pa) and  $Q^2_{\text{Ext}}$  0.994. The standard error of prediction was less than half of the reported uncertainty for the experimental results

2003-02-14

## Results - PBDE (external test set with 9 congeners)



The model performance parameters were SEC 0.16 (log Pa),  $R^2_{\text{Cal}}$  0.992, SEP 0.13 (log Pa) and  $Q^2_{\text{Ext}}$  0.994. The standard error of prediction was less than half of systematic difference between the two experimental investigations

2003-02-14

## Reliable predictions?

- A basic assumption about similarity underlies the use of a prediction model
- Prediction results are only reported for those congeners where the descriptor variables were adequately described by the model
- The model adequately described all 209 PCDE congeners, and 165 PBDE congeners.

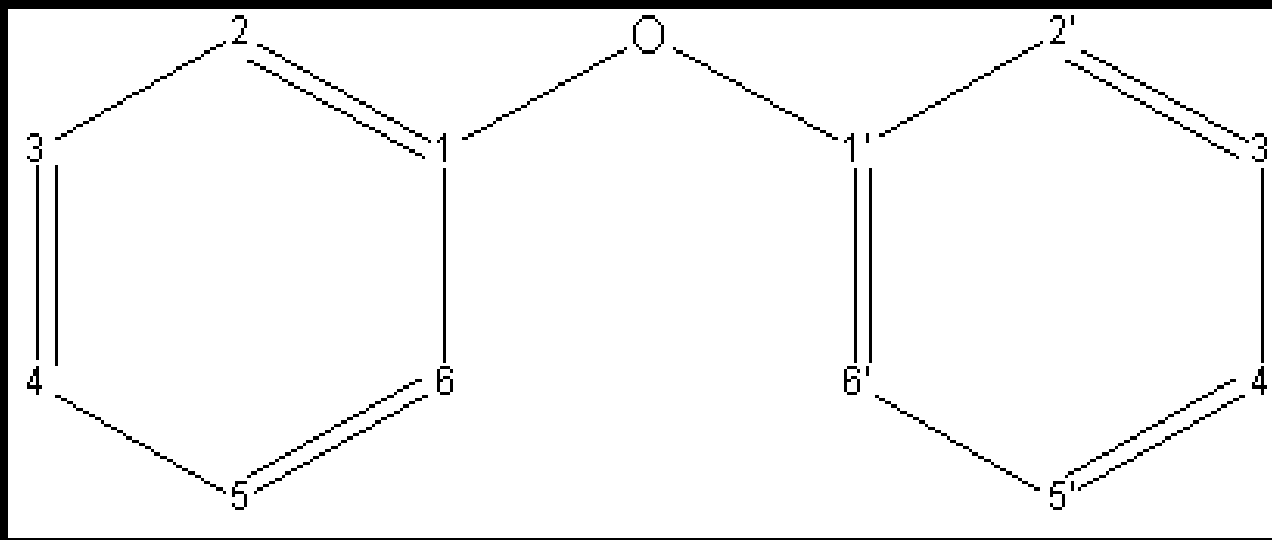
2003-02-14

# The relationship between structure and vapour pressure

- The type and number of halogen atoms in the molecule are the main factors influencing the vapour pressures of halogen-substituted diphenyl ethers
- The third important factor is the substitution pattern, where non-ortho substituted congeners have a significantly lower vapour pressure compared to those with di-, tri- or tetra-ortho substitution
- The variation in vapour pressures (pooled data) due to these three factors was evaluated with analysis of variance (ANOVA)

2003-02-14

# Structure - numbering



- Congeners were characterised by the type and number of halogens, and the number of halogens in the ortho-positions (2,2',6,6'-)

# ANOVA

Source	SS	DF	MS	F	Prob.
Model	210.2	3	70.1	963.9	<0.0001
Error	9.6	132	0.07		
Total	219.8	135	1.6		
Factors					
Br/Cl	38.1	1	38.1	524.4	<0.0001
No. halogen	119.4	1	119.4	1643	<0.0001
No. ortho	1.1	1	1.1	15.3	0.0001

2003-02-14

# Recommendations and outlook

- Quantitative structure-property relationships (QSPR) is a viable approach to estimate physical properties for halogenated diphenyl ethers
- The quality of the experimental data is of crucial importance in developing calibration models, and experimental investigators might consider giving priority to measurements that span the molecular descriptor space as efficiently as possible
- QSPR models can also be used to validate experimental data

2003-02-14

## In summary

- Measures of similarity are important when we assess, model and search for chemical information
- Similarity is often defined using molecular descriptors, i.e. numbers that define structural properties
- Quantitative structure-activity and structure-property relationships are major application areas for molecular modelling and chemoinformatics
- Chemometric methodology assist this modelling work

2003-02-14

# Demonstrations

- With HyperChem®
  - Conformation searching of 2,2',3,3',4,6'-hexabromodiphenylether
- With Unscrambler®
  - Overview of the multivariate descriptor space and development of a calibration model