More about QSAR...

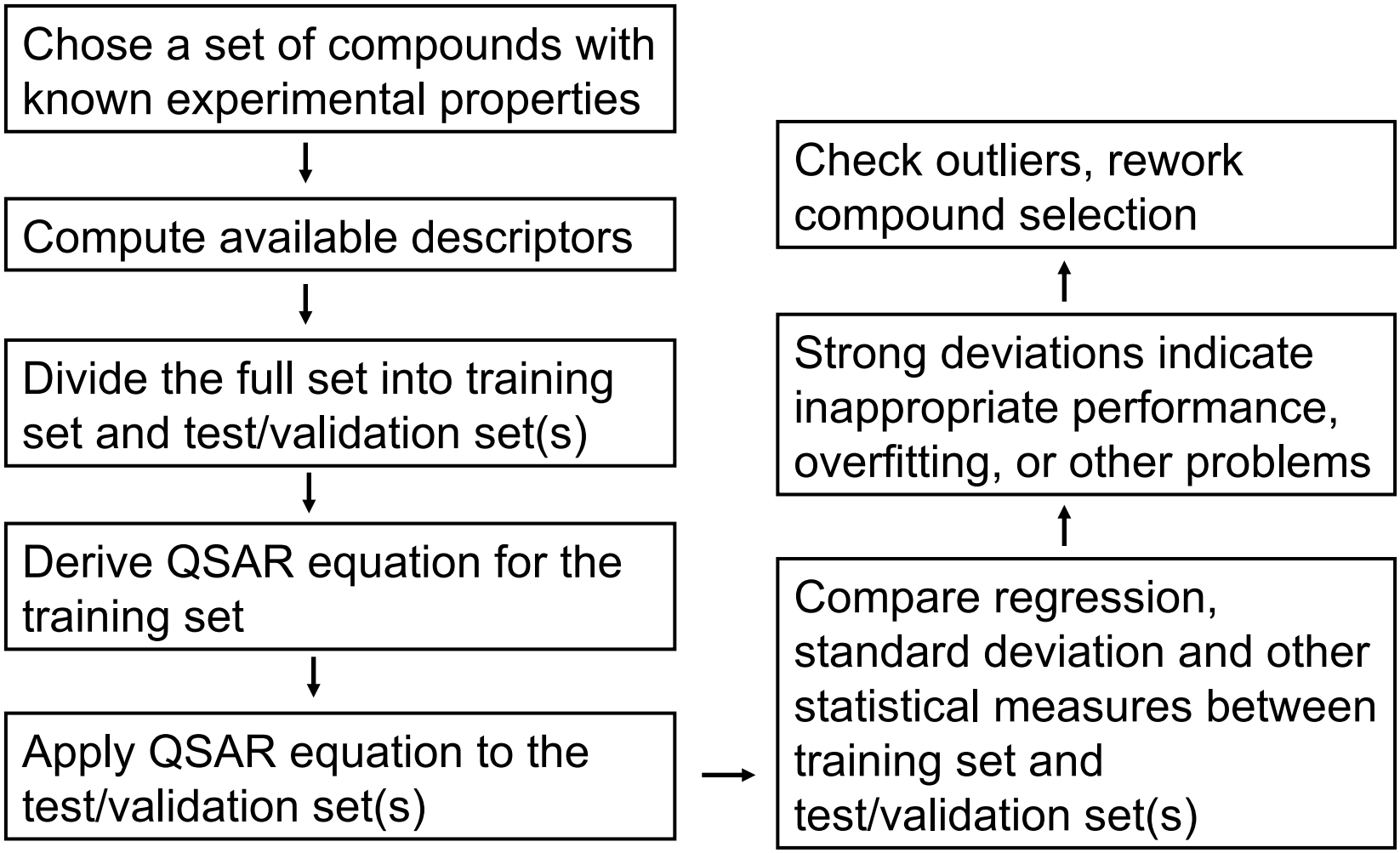
QSAR equations form a quantitative connection between chemical structure and (biological) activity.

$$\log(1/C) = k_1 \cdot P_1 + k_2 \cdot P_2 + \ldots + k_n \cdot P_n$$

Problems:

- Which and how many descriptors to use?
- How reliable are the predictions (applicability domain)?
- How to test/validate QSAR equations
  (continued from lecture 5)

Setting up and testing QSAR equations



Evaluating QSAR equations (1)

The most important **statistical measures** to evaluate QSAR equations are (preferred values given in parenthesis):

Correlation coefficient *r* (in squared from $r^2 > 0.75$)

Standard deviation *se* (small as possible, se < 0.4 units)

Fisher value *F* (level of statistical significance. Also a measure for the portability of the QSAR equation onto another set of data. Should be high, but decreases with increasing number of used variables/descriptors). Therefore only comparable for QSAR equations containing the same number of descriptors

t-test to derive the

probability value *p* of a single variable/descriptor.

Is a measure for coincidental correlation

p<0.05 = 95% significance
p<0.01 = 99%
p<0.001 = 99.9%
p<0.0001 = 99.99%

Evaluating QSAR equations (2)

Example output from OpenStat:

R	R2 (r^2)	F	Prob.>F	DF1	DF2
0.844	0.712	70.721	0.000	3	86

Adjusted R Squared = 0.702

Std. Error of Estimate = 0.427 (se)

Variable	Beta	B	Std.Error	t	Prob.>t
hbdon	-0.738	-0.517	0.042	-12.366	0.000
dipdens	-0.263	-21.360	4.849	-4.405	0.000
chbba	0.120	0.020	0.010	2.020	0.047

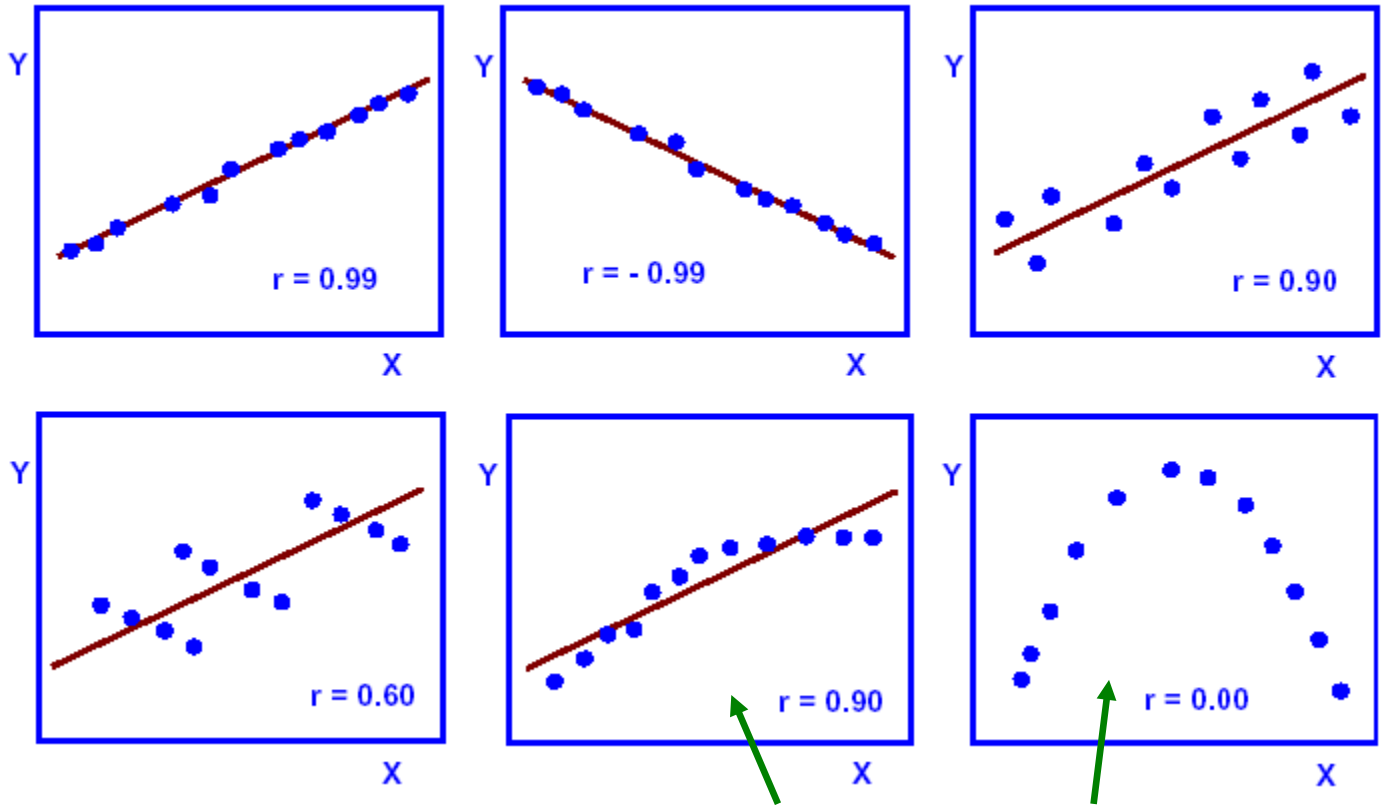
Constant = 0.621

$$\log(1/C) = -0.517 \cdot hbdon - 21.360 \cdot dipdens + 0.020 \cdot chbba + 0.621$$

Lit: William „Bill“ G. Miller, OpenStat Reference Handbook

Evaluating QSAR equations (3)

A plot tells more than numbers:



Shape of curve indicates non-linear correlation

Source: H. Kubinyi, Lectures of the drug design course
http://www.kubinyi.de/index-d.html

Evaluating QSAR equations (4)

Examples where statistical measures between training set and test set strongly deviate:

Training set n=15, r^2=0.91, se=0.27 (5 descriptors used)

Test set n=5, r^2=0.69, se=0.42

Obvious reason: too many descriptors used in QSAR eq. Therefore the training set becomes overfitted, correlation breaks down for the test set.
→ Limit number of used descriptors in the QSAR equation to 3.

Training set n=26, r^2=0.88, se=0.32, F=110.7 (3 descriptors used)

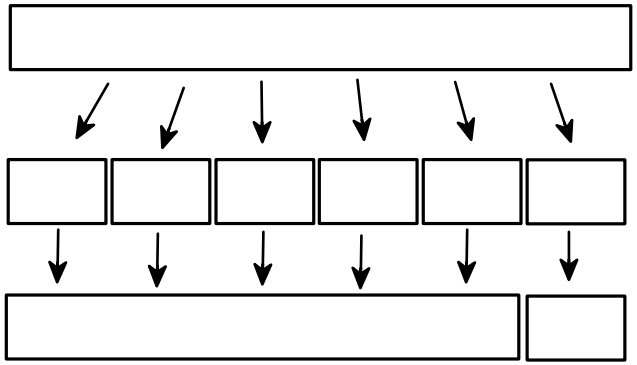
Test set n=7, r^2=0.75, se=0.38, F=66.5

Possible reason: Compounds in the test set are quite different compared to those in the training set.
→ Check compounds (and descriptor ranges) for similarity, redo compound selection for training and test set e.g. using cluster analysis

Evaluating QSAR equations (5)

(Simple) *k*-fold cross validation:

Partition your data set that consists of *N* data points into *k* subsets (*k* < *N*).



k times

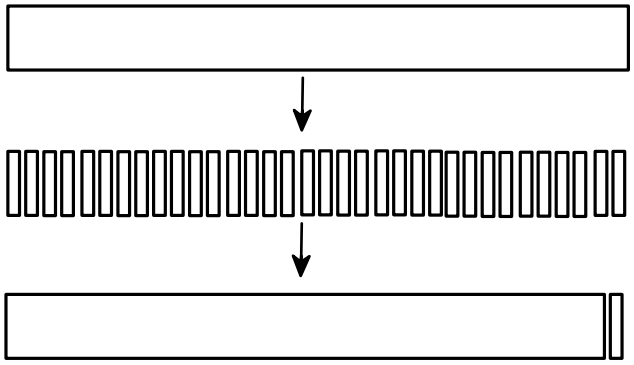
Generate *k* QSAR equations using a subset as test set and the remaining *k*-1 subsets as training set respectively. This gives you an average error from the *k* QSAR equations.

In practise *k* = 5 or *k* = 10 has shown to be reasonable (= 5-fold or 10-fold cross validation)

Evaluating QSAR equations (6)

Leave one out cross validation:

Partition your data set that consists of *N* data points into *k* subsets (*k* = *N*).



N times

Disadvantages:

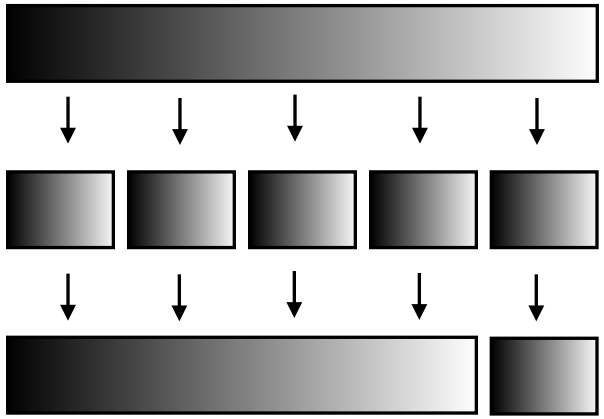
- Computationally expensive
- Partitioning into training and test set is more or less by random, thus the resulting average error can be way off in extreme cases.

Solution: (feature) distribution within the training and test sets should be identical or similar

Evaluating QSAR equations (7)

Stratified cross validation:

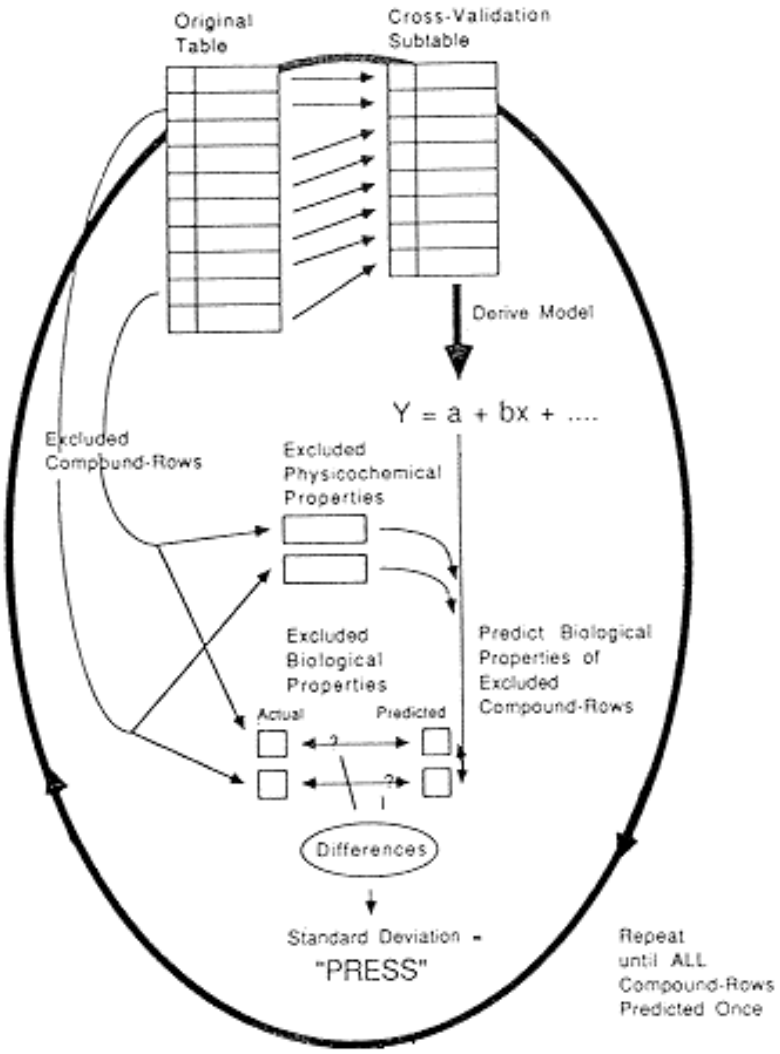
Same as *k*-fold cross validation but each of the *k* subsets has a similar (feature) distribution.



k times

The resulting average error is thus more prone against errors due to inequal distribution between training and test sets.

Evaluating QSAR equations (8)



alternative
Cross-validation and
leave one out (LOO)
schemes

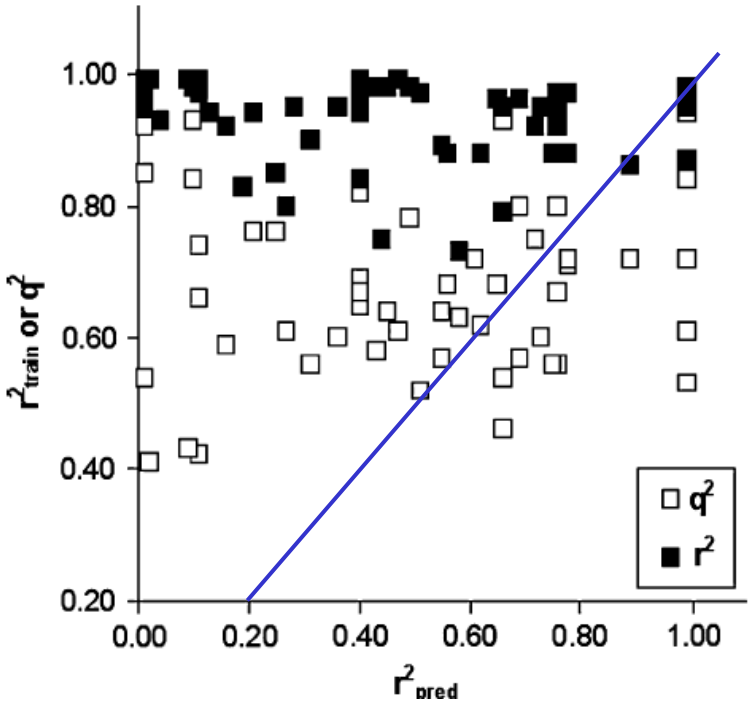
Leaving out one or more descriptors from the derived equation results in the cross-validated correlation coefficient q^2.

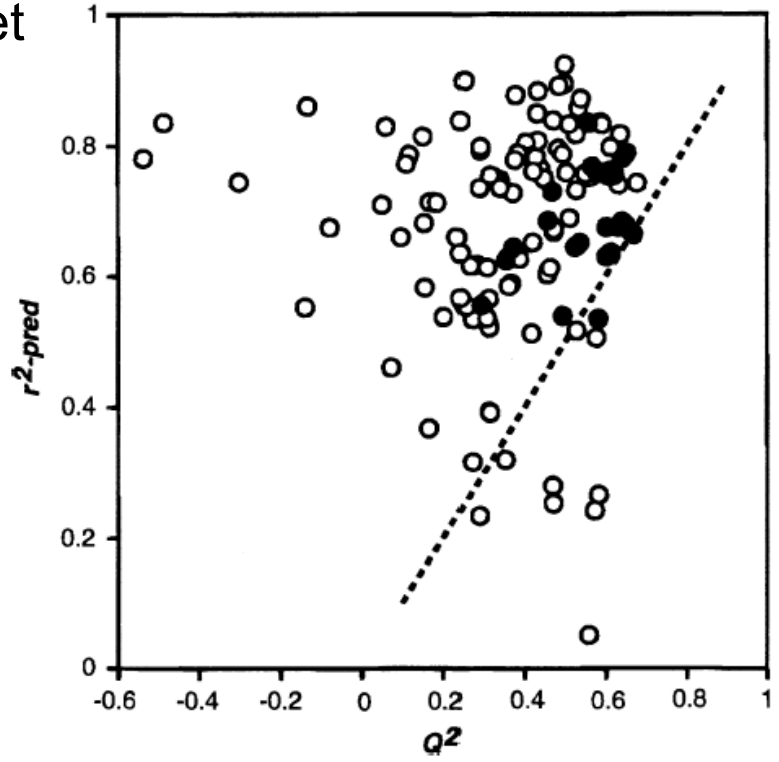
This value is of course lower than the original r^2.
q^2 being much lower than r^2 indicates problems...

Evaluating QSAR equations (9)

Problems associated with q^2 and *leave one out (LOO)*

→ There is no correlation between q^2 and test set predictivity, q^2 is related to r^2 of the training set





Kubinyi‘s paradoxon: Most r^2 of test sets are higher than q^2 of the corresponding training sets [due to manual selection?]

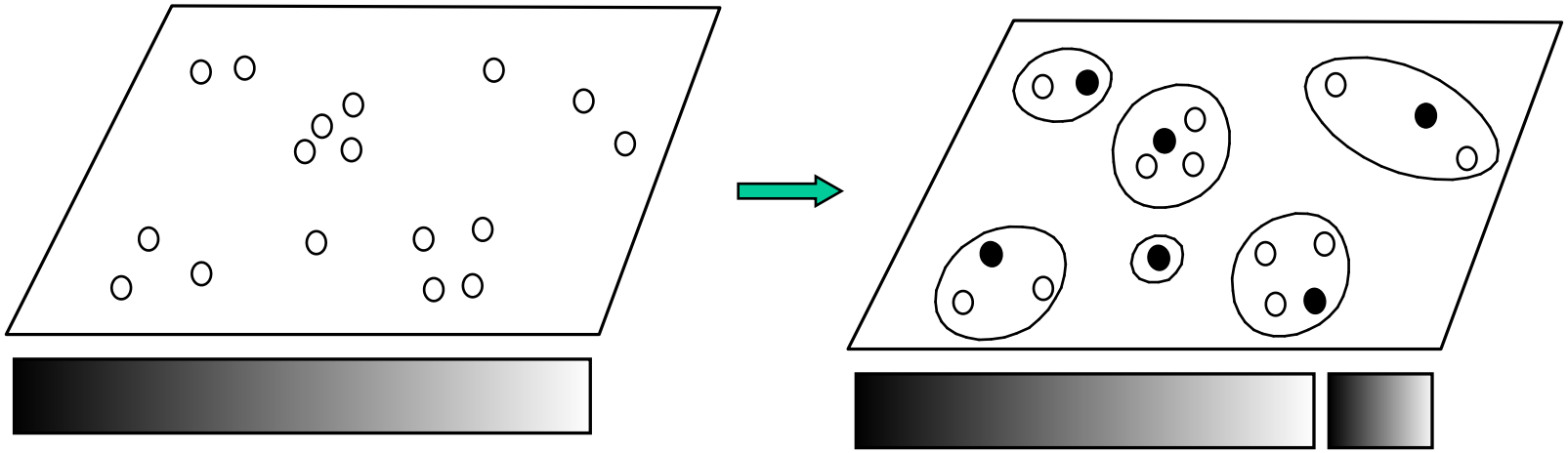
Lit: A.M.Doweyko *J.Comput.-Aided Mol.Des.* **22** (2008) 81-89.

Evaluating QSAR equations (10)

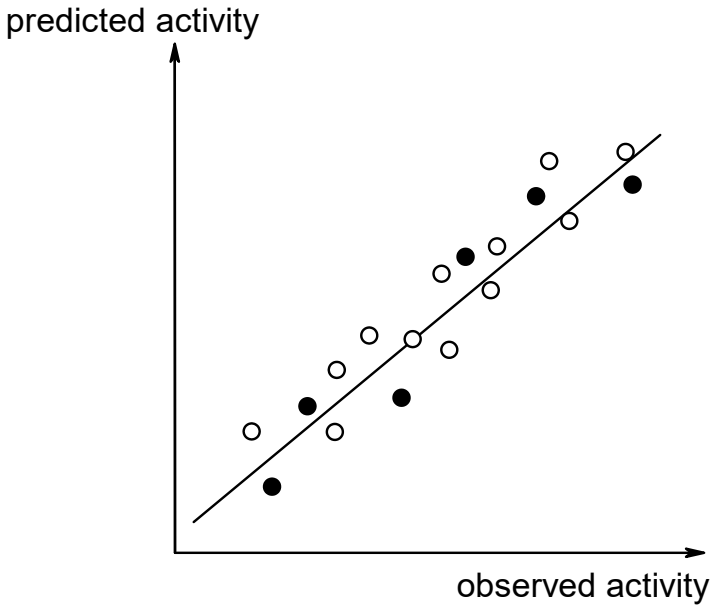
One of most reliable ways to test the performance of a QSAR equation is to apply an external test set.

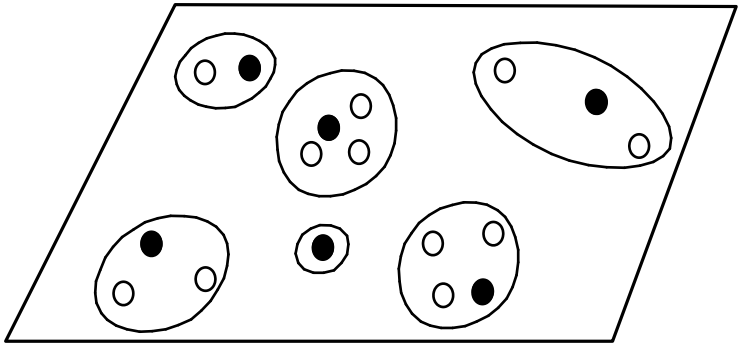
→ partition your complete set of data into training set (2/3) and test set (1/3 of all compounds, idealy)

Compounds of the test set should be representative (confers to a 1-fold stratified cross validation)
→ Cluster analysis using the descriptor values of each compound plus their activities.
→ Use cluster centroids as test set and the remaining compounds for the training set (these account for the diversity)



Evaluating QSAR equations (11)





Compounds of the test set must cover the same activity range as those of the training set

- ○ Training set
- • Test set

Evaluating QSAR equations (12)

Estimating the error (range) of predicted values is difficult.

Approaches to give a confidence range (as in statistics) or determining the applicability domain of the model:

• Distance based: similar to *k*-nearest neighbor; where is the predicted compound located in the descriptor space?
Close to one group or rather in between clusters?

Large training sets can be split into a further calibration set that is used for estimating the error of unseen data based on their similarity.

• Are there consistent outliers in the data set?
→ These are either to dissimilar or contain experimental errors

Lit: K. Roy et al. *Chemomet. Intell. Lab. Sys.* **145** (2015) 22-29.

Interpretation of QSAR equations

The kind of applied variables/descriptors should enable us to

- draw conclusions about the underlying physico-chemical processes
- derive guidelines for the design of new molecules by interpolation

$$\log(1/K_i) = +1.049 \cdot n_{fluorine} - 0.843 \cdot n_{OH} + 5.768$$

Higher affinity requires more fluorine, less OH groups

Some descriptors give information about the biological mode of action:

- A dependence of $(\log P)^2$ indicates a transport process of the drug to its receptor.
- Dependence from E_{LUMO} or E_{HOMO} indicates a chemical reaction

Evaluating QSAR equations (13)

Reduce the number of available descriptors before performing a regression analysis:

- More descriptors mean longer run times
- More descriptors raise the likelihood of accidental correlation (see also slides further below)
- Descriptors might be correlated to each other and thus do not provide more information
- Can you interpret what your descriptors mean?

→ get rid of the garbage

Correlation of descriptors

Other approaches to handle correlated descriptors and/or a wealth of descriptors:

Transforming descriptors to uncorrelated variables by

- *principal component analysis* (PCA)
- *partial least square* (PLS)

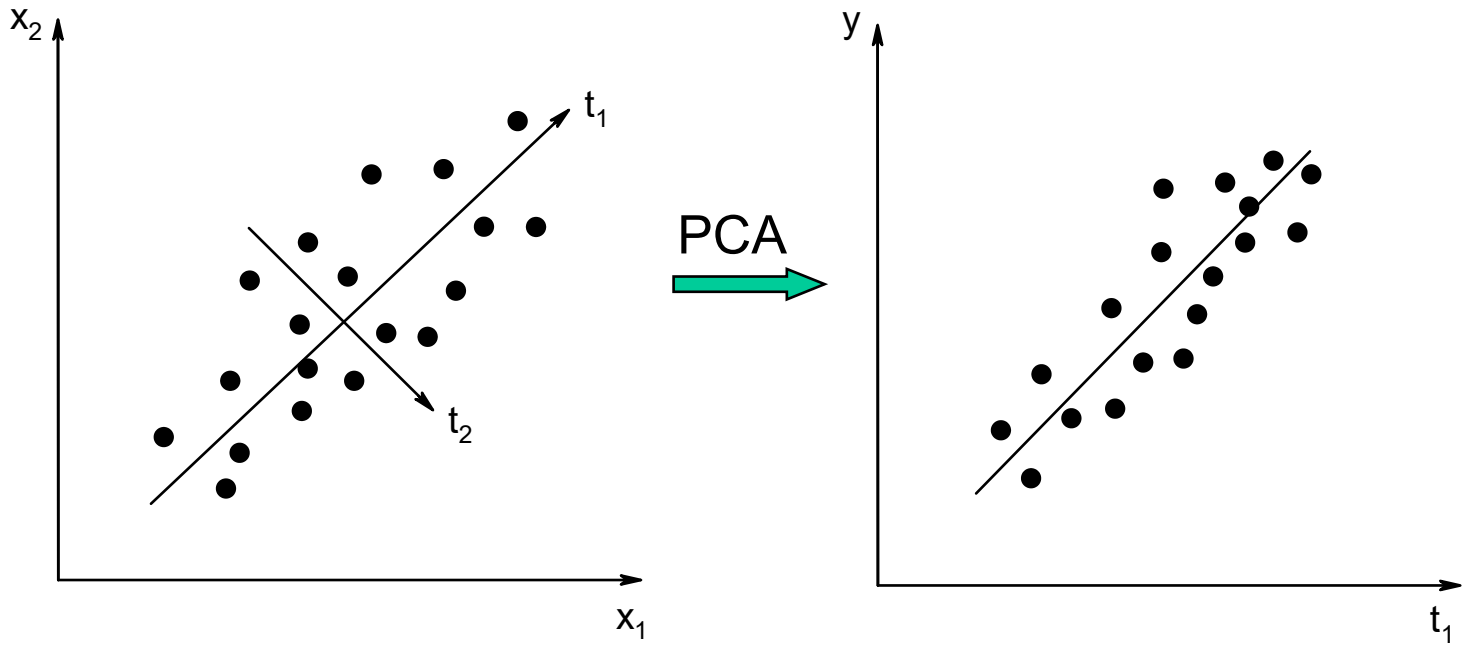
for example applied in *comparative molecular field analysis* (CoMFA), see below

Methods that intrinsically handle correlated variables

- *neural networks,* especially *deep learning networks*

Partial least square (I)

The idea is to construct a small set of latent variables t_i (that are orthogonal to each other and therefore uncorrelated) from the pool of inter-correlated descriptors x_i .



In this case t_1 and t_2 result as the normal modes of x_1 and x_2 where t_1 shows the larger variance.

Partial least square (II)

The predicted term *y* is then a QSAR equation using the latent variables t_i

$$y = b_1 t_1 + b_2 t_2 + b_3 t_3 + \ldots + b_m t_m$$

where

$$t_1 = c_{11} x_1 + c_{12} x_2 + \ldots + c_{1n} x_n$$
$$t_2 = c_{21} x_1 + c_{22} x_2 + \ldots + c_{2n} x_n$$
$$. \quad . \quad . \quad .$$
$$t_m = c_{m1} x_1 + c_{m2} x_2 + \ldots + c_{mn} x_n$$

The number of latent variables t_i is chosen to be (much) smaller than that of the original descriptors x_i.

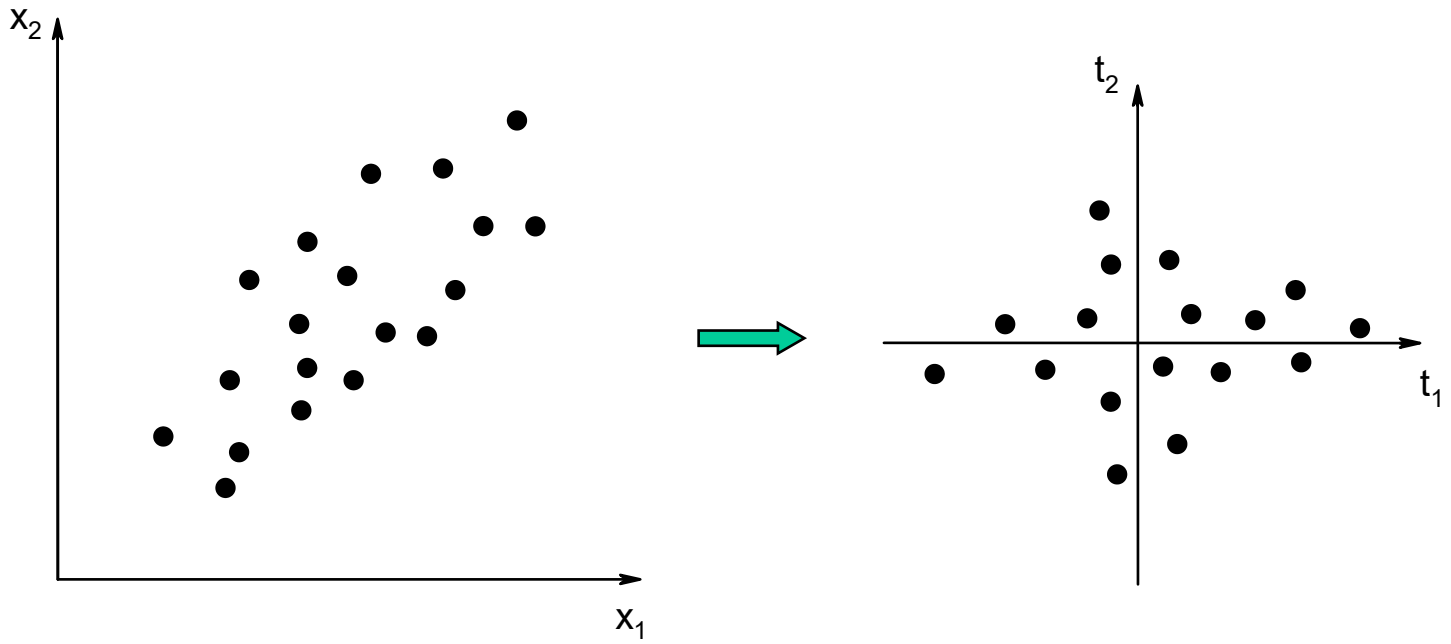
But, how many latent variables are reasonable?

→ plot r^2, se, q^2 and its fluctations against the number of latent variables and identify the minimal number of latent variables.

Principal Component Analysis PCA (I)

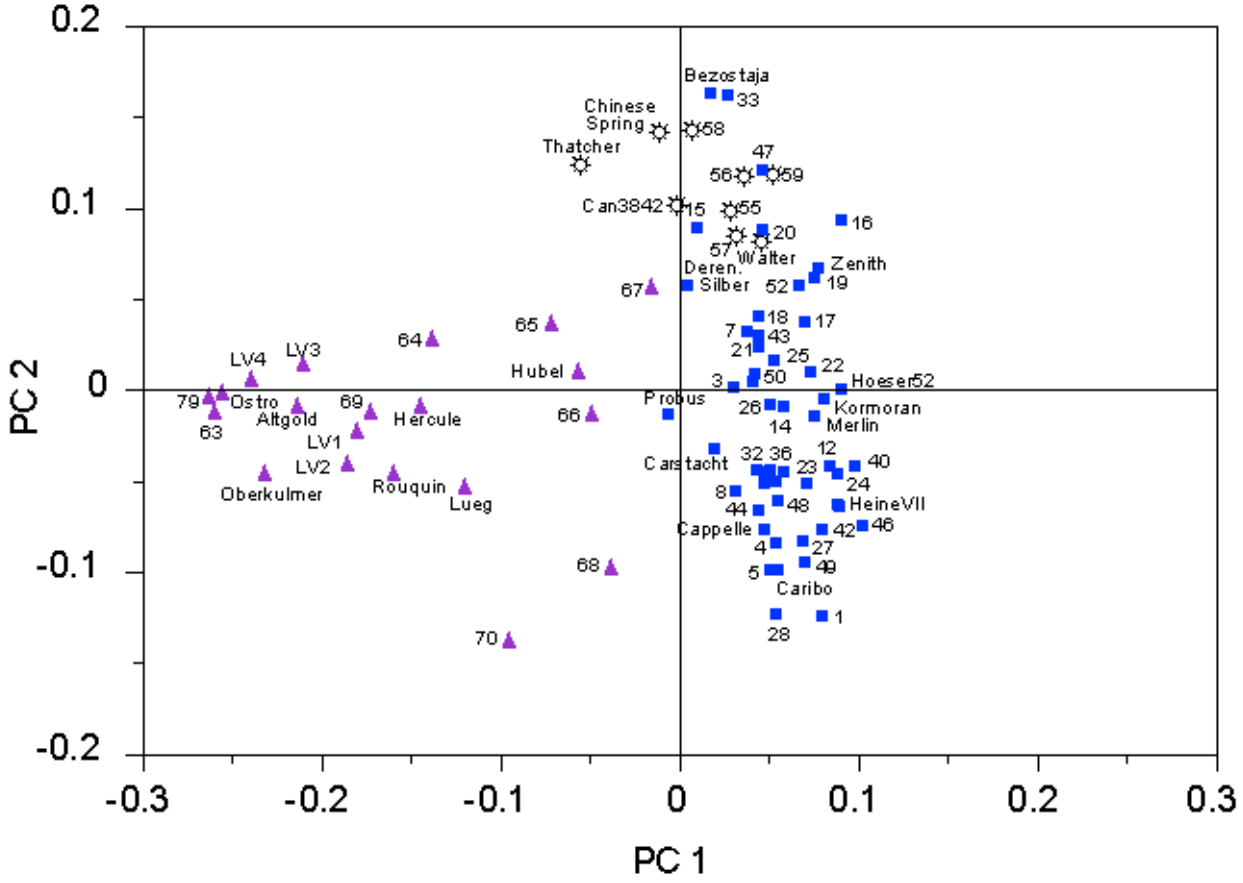
Problem: Which are the (decisive) significant descriptors ?

Principal component analysis determines the normal modes from a set of descriptors/variables.
This is achieved by a coordinate transformation resulting in new axes. The first principal component then shows the largest variance of the data. The second and further normal components are orthogonal to each other.



Principal Component Analysis PCA (II)

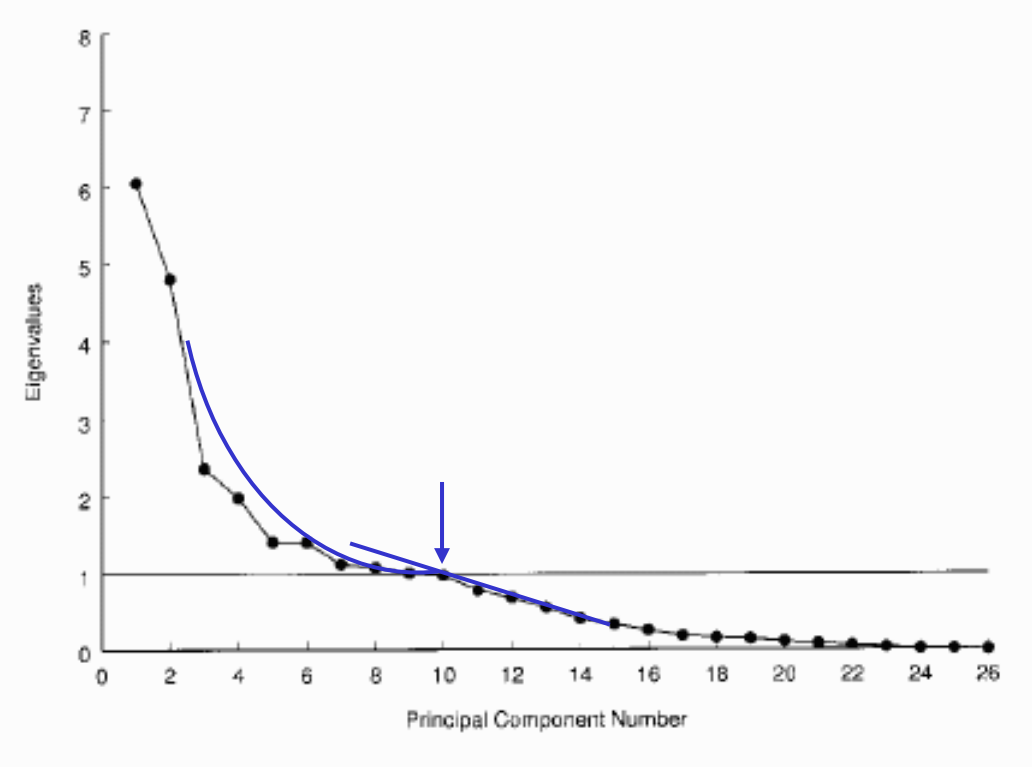
The first component (pc1) shows the largest variance, the second component the second largest variance, and so on.



Lit: E.C. Pielou: *The Interpretation of Ecological Data*, Wiley, New York, 1984

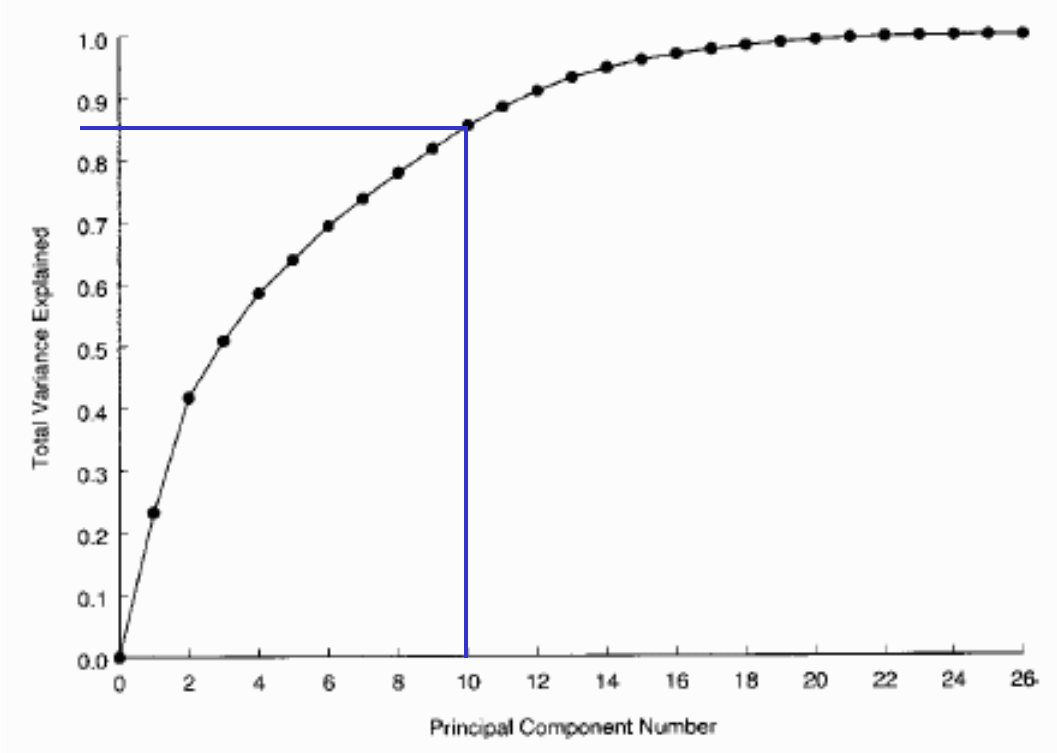
Principal Component Analysis PCA (III)

The significant principal components usually have an Eigen value >1 (Kaiser-Guttman criterion). Frequently there is also a kink that separates the less relevant components (Scree test)



Principal Component Analysis PCA (IV)

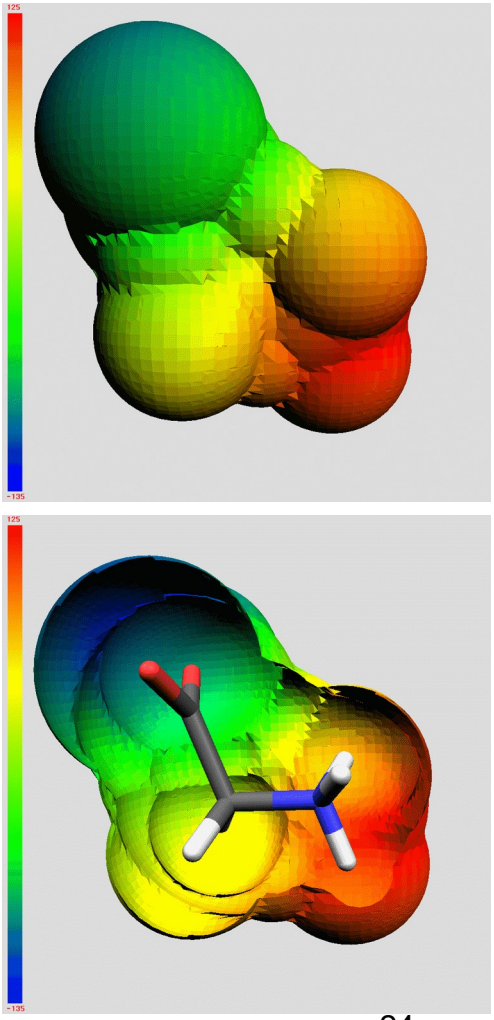
The obtained principal components should account for more than 80% of the total variance.



Principal Component Analysis (V)

Example: What descriptors determine the logP ?

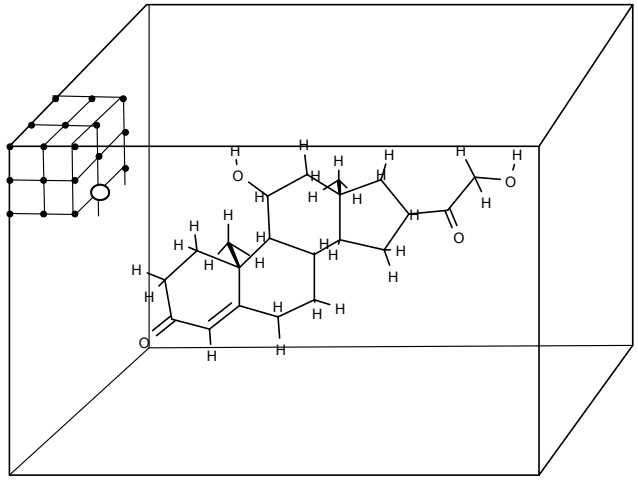
property	pc1	pc2	pc3
dipole moment	0.353		
polarizability		0.504	
mean of +ESP	0.397	-0.175	0.151
mean of –ESP	-0.389	0.104	0.160
variance of ESP 0.403		-0.244	
minimum ESP	-0.239	-0.149	0.548
maximum ESP	0.422		0.170
molecular volume		0.506	0.106
surface	0.519	0.115	
fraction of total variance	28%	22%	10%



Lit: T.Clark et al. *J.Mol.Model.* **3** (1997) 142

Comparative Molecular Field Analysis (I)

The molecules are placed into a 3D grid and at each grid point the steric and electronic interaction with a probe atom is calculated (force field parameters)



For this purpose the GRID program can be used:

P.J. Goodford
J.Med.Chem. **28** (1985) 849.

Problems: „active conformation“ of the molecules needed

All molecule must be superimposed (aligned according to their common scaffold)

Lit: R.D. Cramer et al. *J.Am.Chem.Soc.* **110** (1988) 5959.

Comparative Molecular Field Analysis (II)

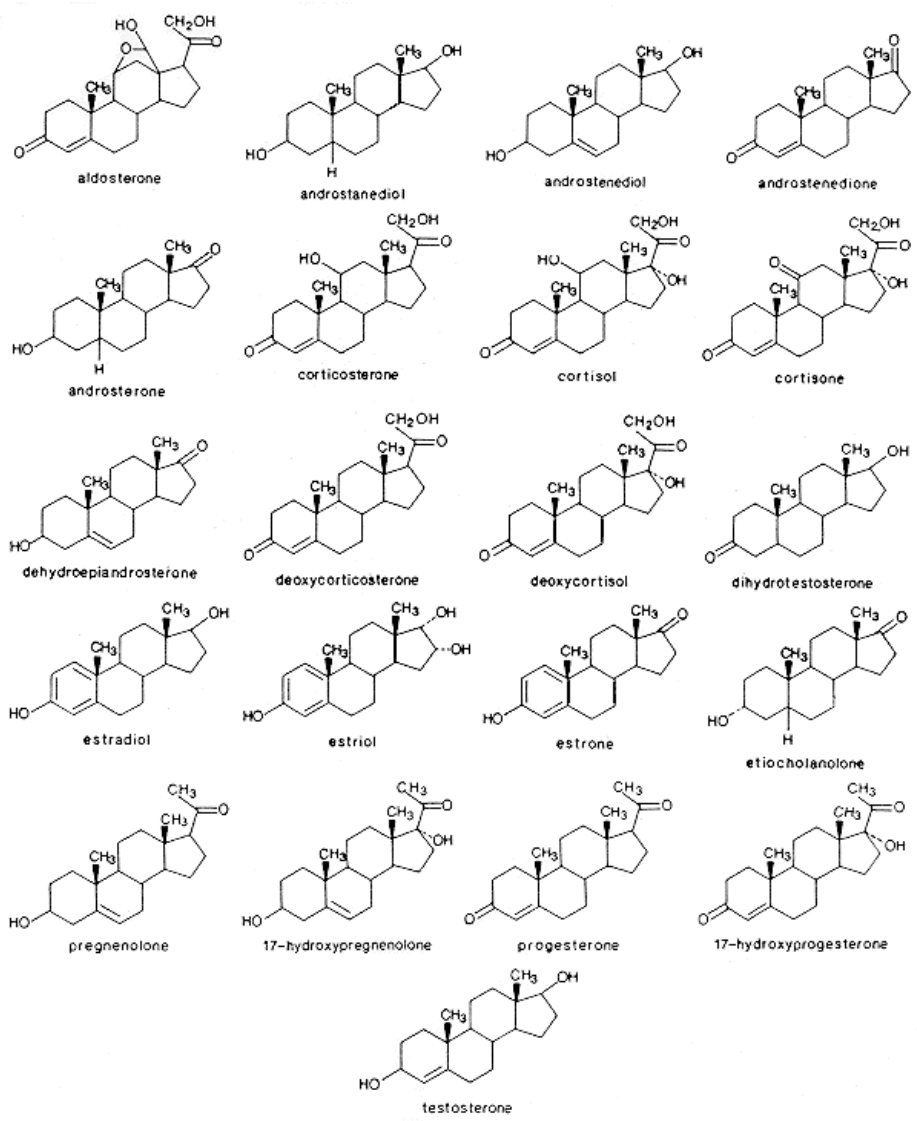
The resulting coefficients for the matrix S (N grid points, P probe atoms) have to determined using a PLS analysis.

compound	log (1/C)	S1	S2	S3	...	P1	P2	P3	...
steroid1	4.15								
steroid2	5.74								
steroid3	8.83								
steroid4	7.6								
...									

$$\Downarrow$$

$$\log(1/\mathrm{C}) = const + \sum_{i=1}^{N} \sum_{j=1}^{P} c_{ij} S_{ij}$$

Comparative Molecular Field Analysis (III)

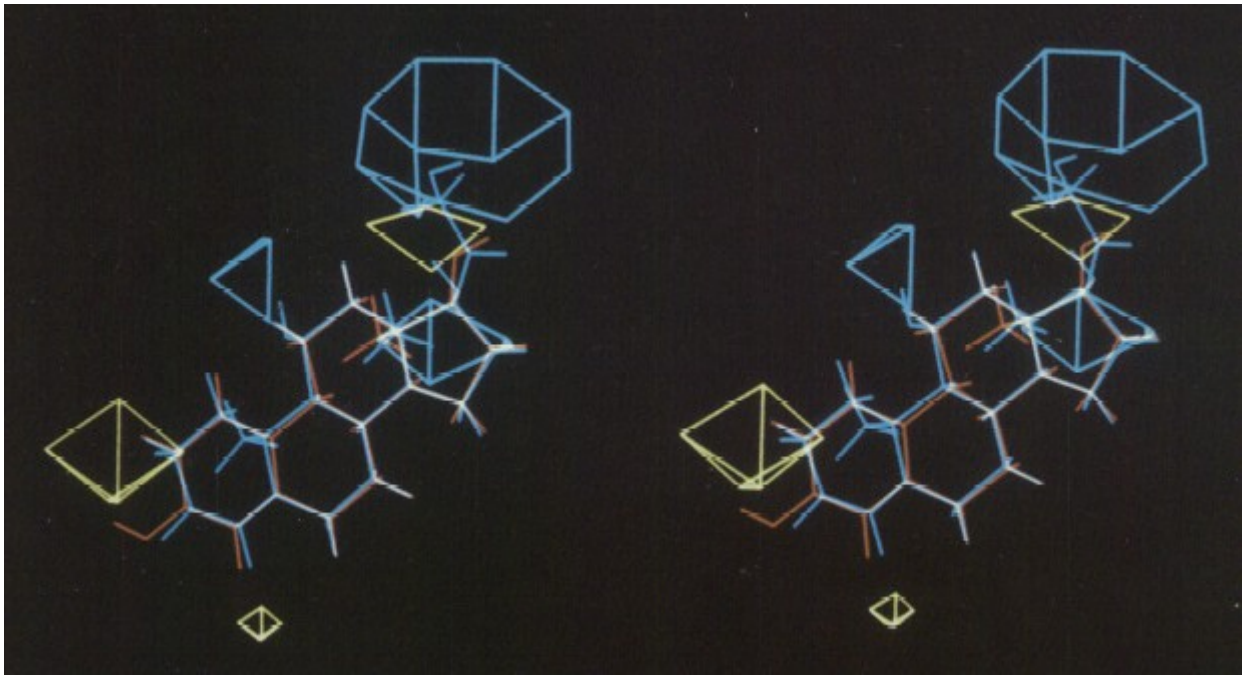


Application of CoMFA:
Affinity of steroids to the testosterone binding globulin

Lit: R.D. Cramer et al.
J.Am.Chem.Soc.
110 (1988) 5959.

Comparative Molecular Field Analysis (IV)

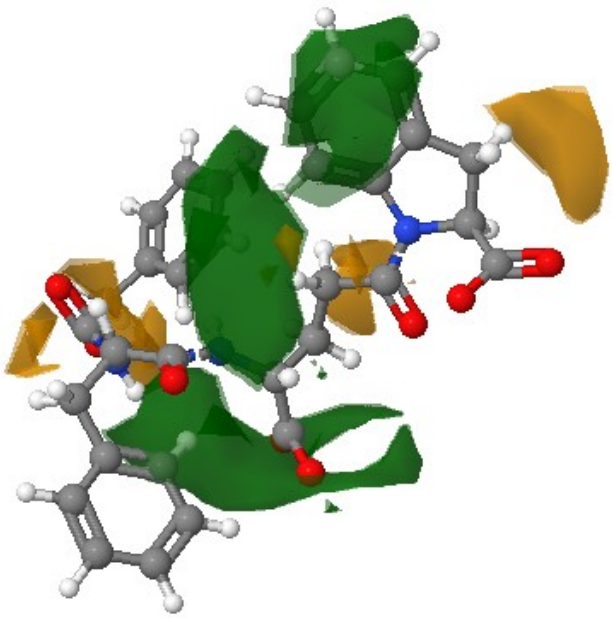
Analog to QSAR descriptors, the CoMFA variables can be interpreted. Here (color coded) contour maps are helpful



yellow: regions of unfavorable steric interaction
blue: regions of favorable steric interaction

Lit: R.D. Cramer et al. *J.Am.Chem.Soc.* **110** (1988) 5959

CoMFA (V) 3-D Database online:



www.3D-QSAR.com
THREE-DIMENSIONAL QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS SERVER

„A 3-D QSAR Models Database for Virtual Screening“

Compounds can be screened against a large set of precalculated models

Jmol

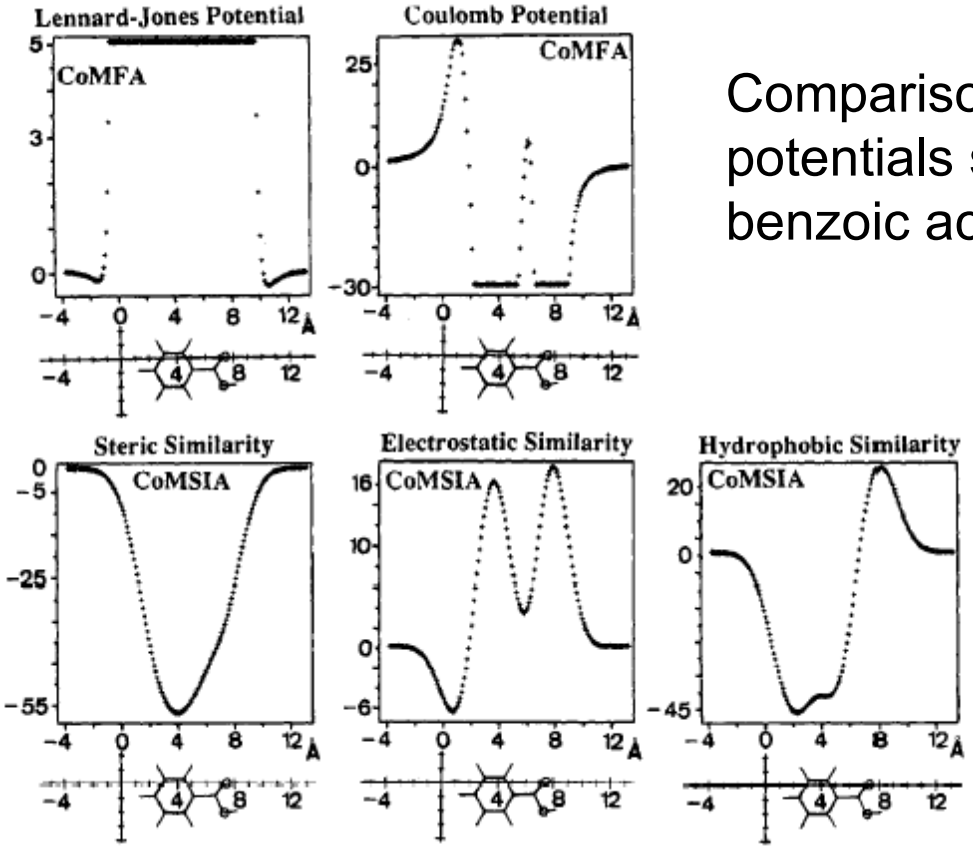
Maps Table

%	0	10	20	30	40	50	60	70	80	90	100
PLS_Coeff	◉	○	○	○	○	○	○	○	○	○	○
CoMFA_Maps	○	○	○	○	○	◉	○	○	○	○	○

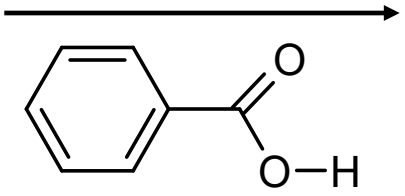
Rino Ragno et al. Università di Roma (Italy)

Comparative Molecular Similarity Indices Analysis (CoMSIA)

CoMFA based on similarity indices at the grid points



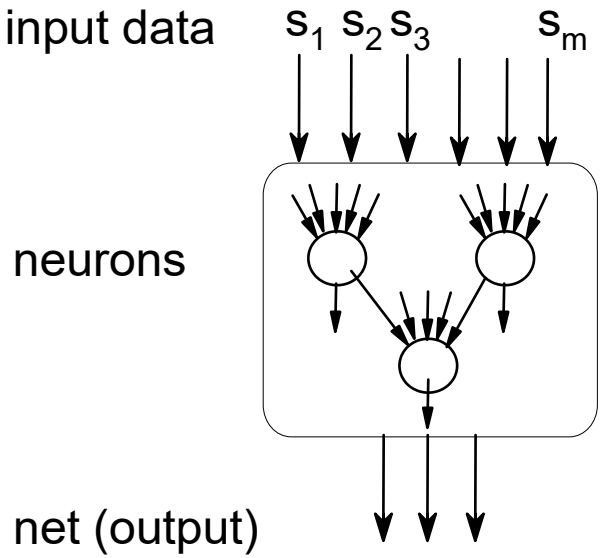
Comparison of CoMFA and CoMSIA potentials shown along one axis of benzoic acid

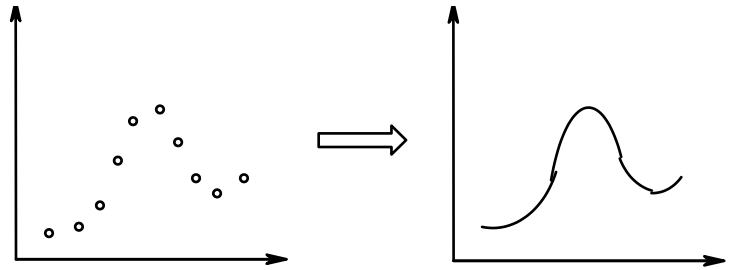


Lit: G.Klebe et al. *J.Med.Chem.* **37** (1994) 4130.

Neural Networks (I)

Neural networks can be regarded as a common implementation of artificial intelligence. The name is derived from the network-like connection between the switches (neurons) within the system. Thus they can also handle inter-correlated descriptors.



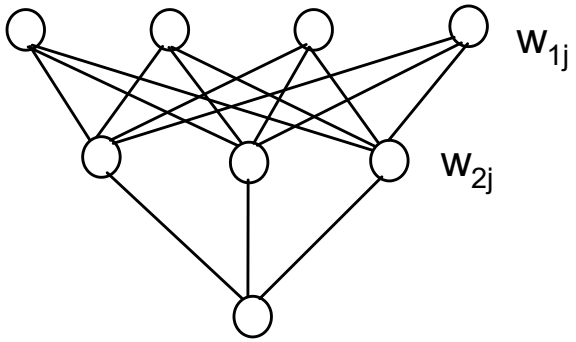


modeling of a (regression) function

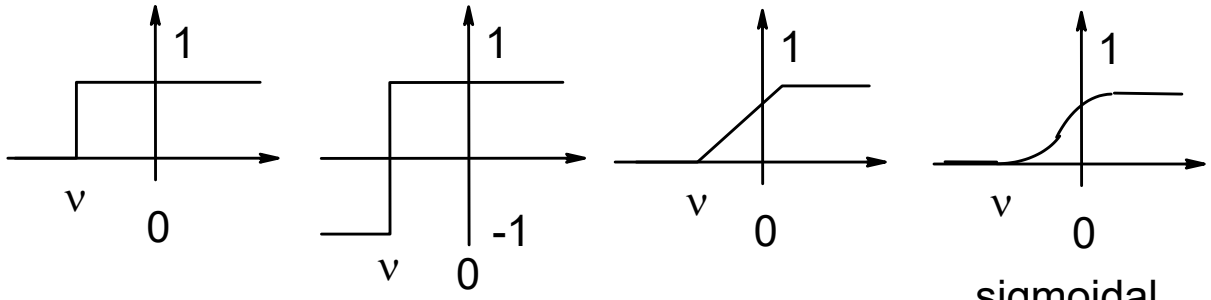
From the many types of neural networks, backpropagation and unsupervised maps are the most frequently used.

Neural Networks (II)

A typical backpropagation net consists of neurons organized as the *input layer*, one or more *hidden layers*, and the *output layer*



Furthermore, the actual kind of signal transduction between the neurons can be different:

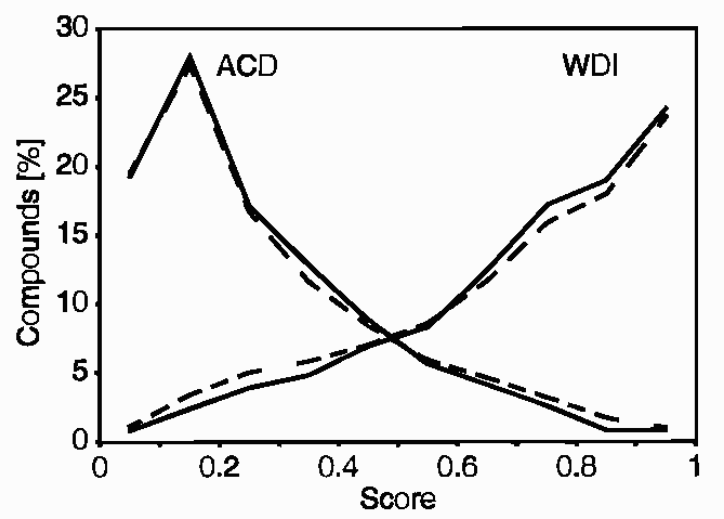


Recursive Partitioning

Instead of quantitative values often there is only qualitative information available, e.g. substrates versus non-substrates

Thus we need classification methods such as

- decision trees
- support vector machines
- (neural networks): partition at what score value ?



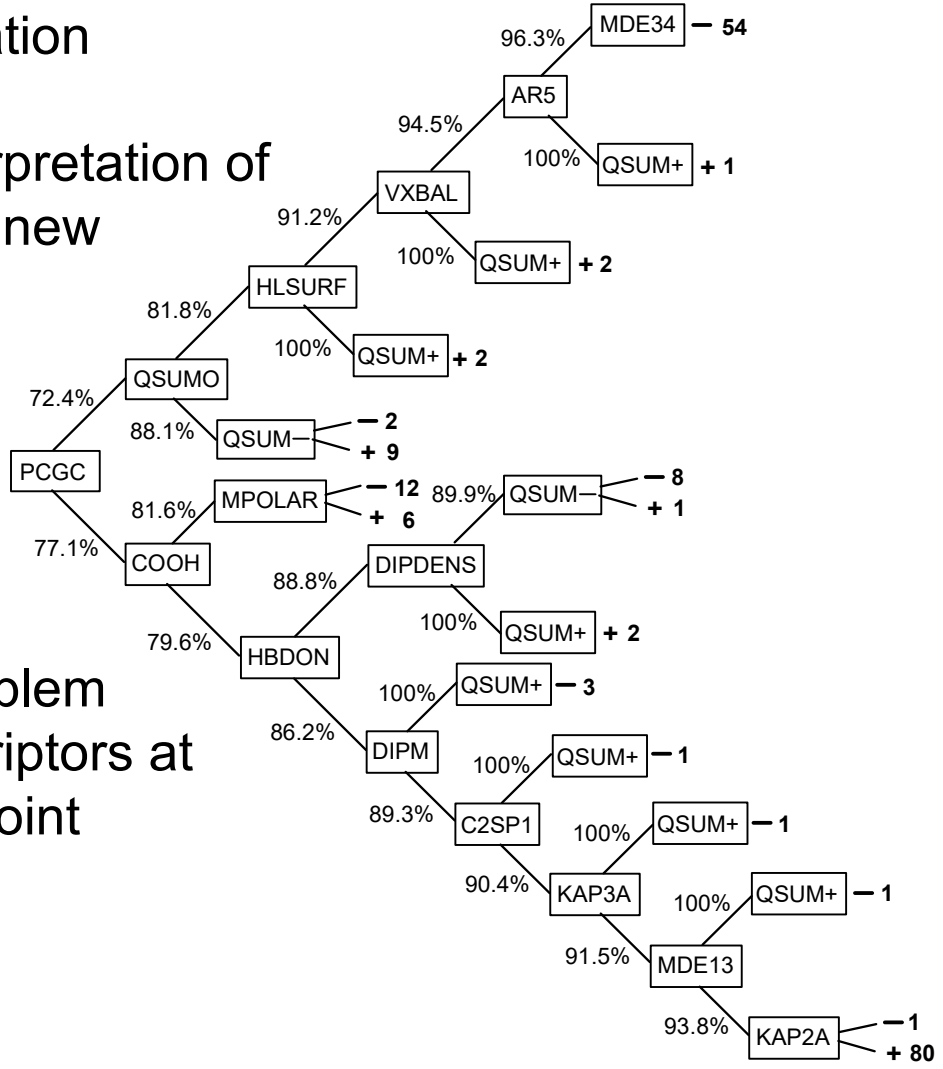
Picture: J. Sadowski & H. Kubinyi *J.Med.Chem.* **41** (1998) 3325.

Decision Trees

Iterative classification

Advantages: Interpretation of results, design of new compounds with desired properties

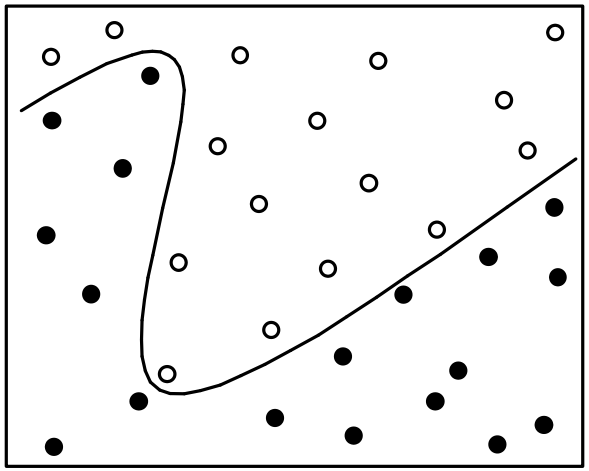
Disadvantage: Local minima problem chosing the descriptors at each branching point



Lit: J.R. Quinlan *Machine Learning* **1** (1986) 81.

Support Vector Machines

Support vector machines generate a hyperplane in the multi-dimensional space of the descriptors that separates the data points.



Advantages: accuracy, a minimum of descriptors (= support vectors) used

Disadvantage: Interpretation of results, design of new compounds with desired properties, which descriptors for input

Property prediction: So what ?

Classical QSAR equations: small data sets, few descriptors that are (hopefully) easy to understand

CoMFA: small data sets, lots of descriptors } easy visual interpretation of resulting interaction regions

Partial least square: small data sets, many descriptors

black box methods:

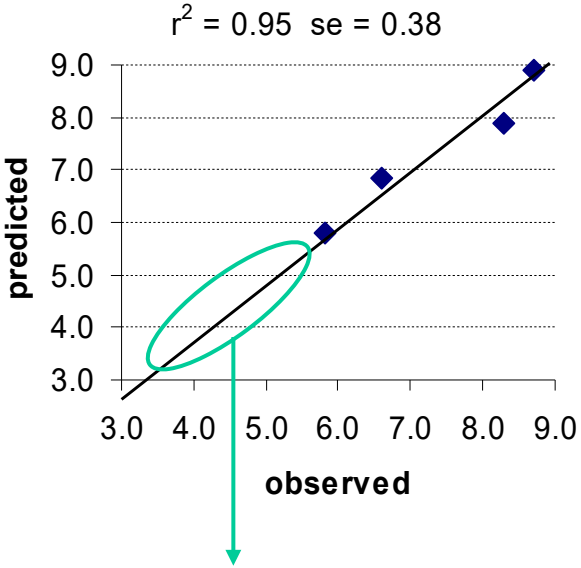
Neural nets: large data sets, some, preselected descriptors

Support vector machines: large data sets, many descriptors

} interpretation of results often difficult

Interpretation of QSAR equations

Caution is required when extrapolating beyond the underlying data range. Outside this range no reliable predicitions can be made



Beyond the black stump ...

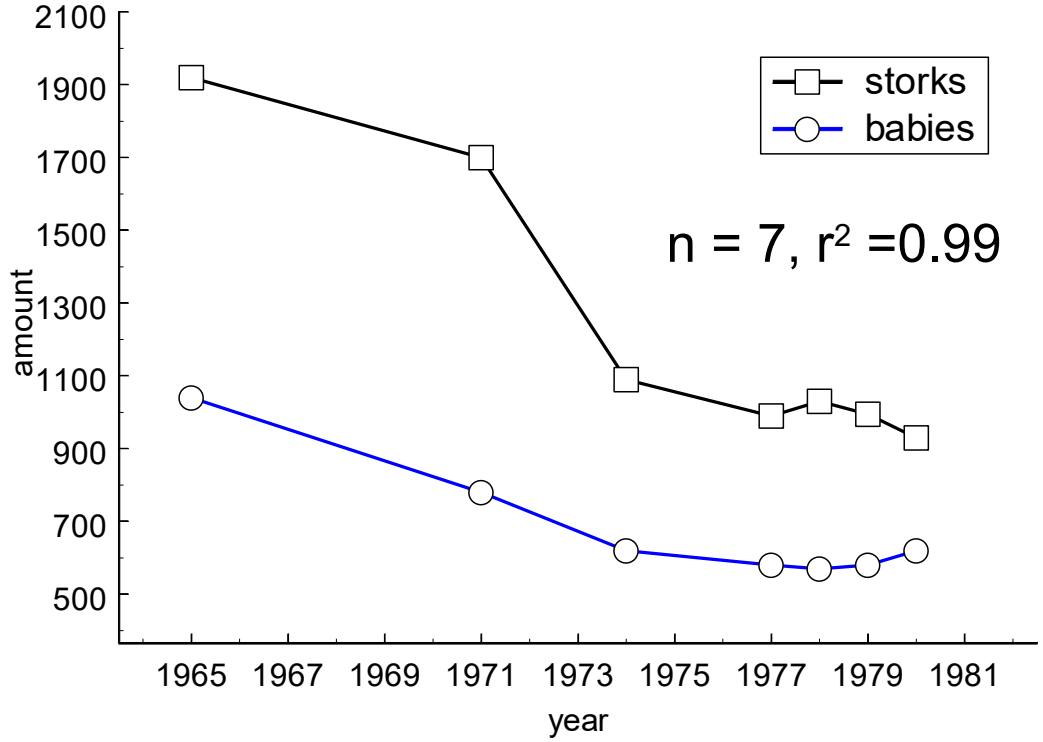
Kimberley, Western Australia

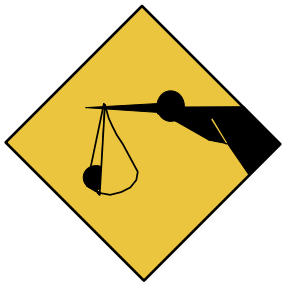
Interpretation of QSAR equations

There should be a reasonable connection between the used descriptors and the predicted quantity.

Example: H. Sies *Nature* **332** (1988) 495.

Scientific proof that babies are delivered by storks

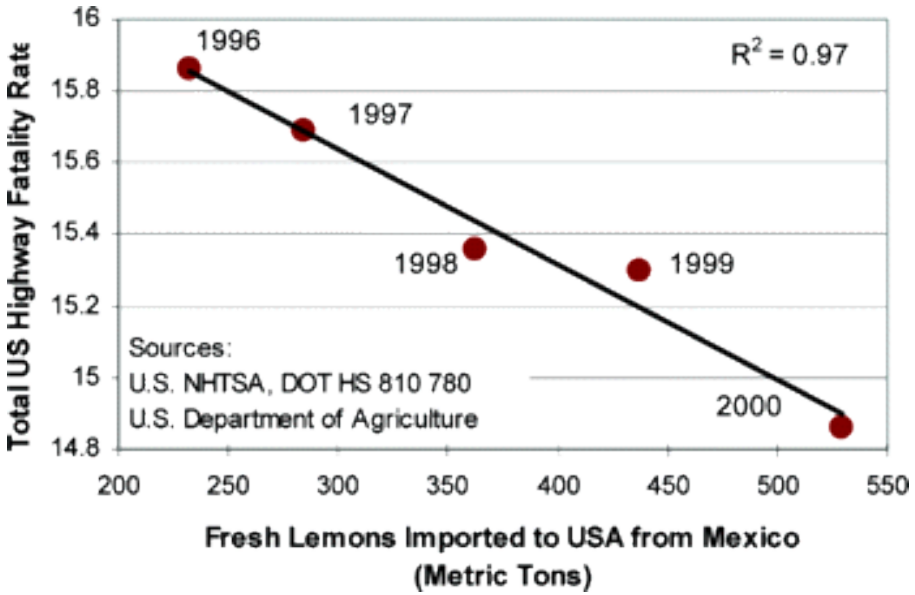




Interpretation of QSAR equations

Another striking correlation

„QSAR has evolved into a perfectly practiced art of logical fallacy“



n = 5, r^2 =0.97

very small data set

S.R. Johnson *J.Chem.Inf.Model.* **48** (2008) 25.

→ the more descriptors are available, the higher is the chance of finding some that show a chance correlation

Interpretation of QSAR equations

The scientific proof that chocolate makes you smarter....

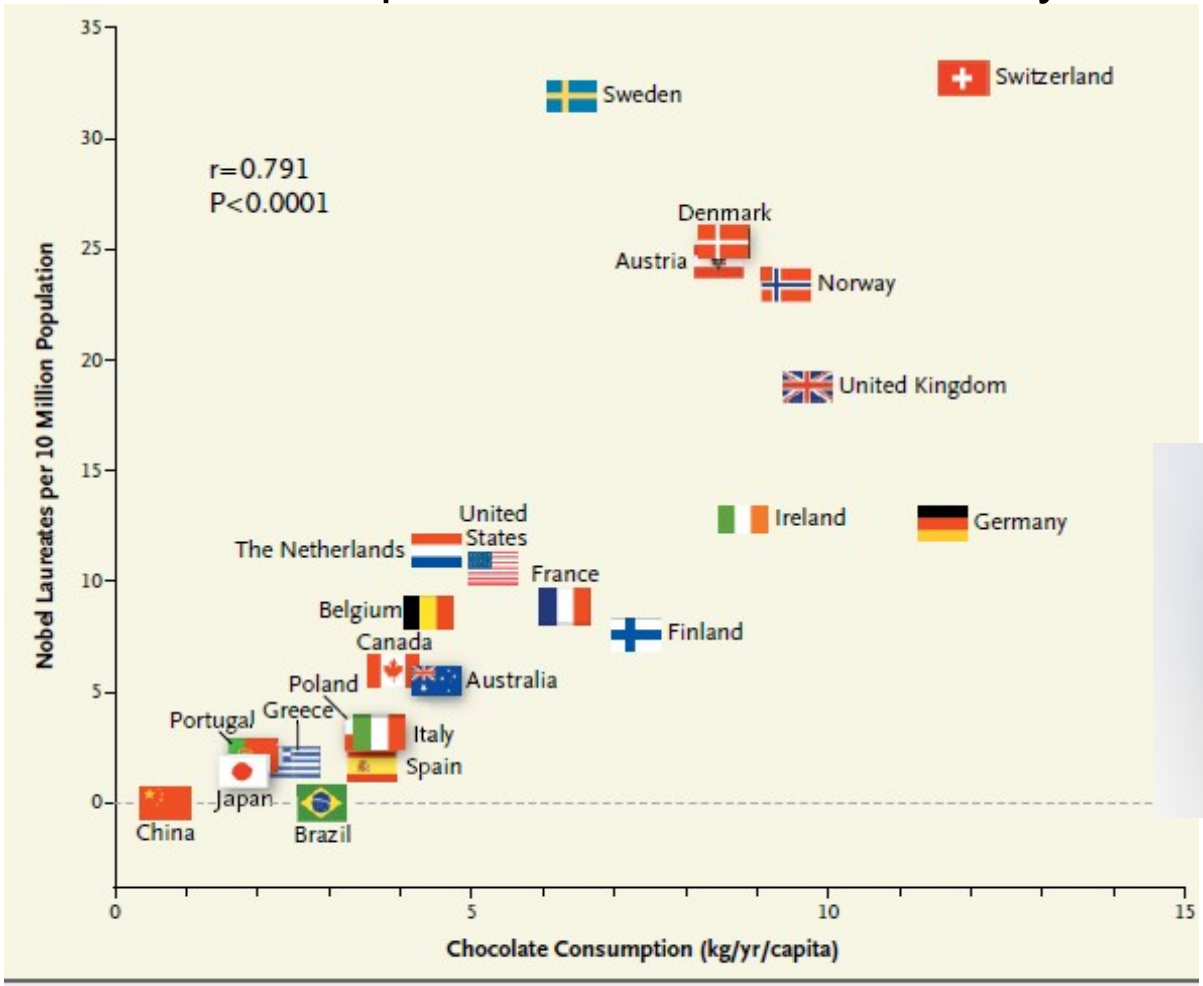


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

$n = 22, r^2 = 0.63$

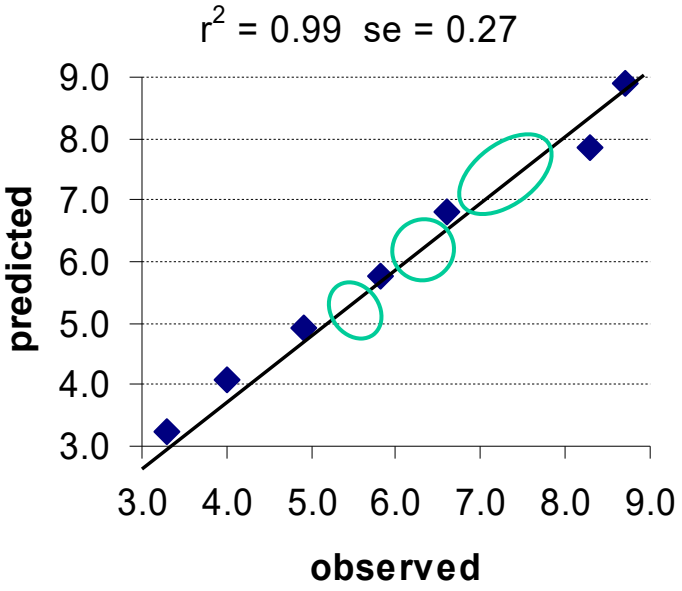
small data set

F.H. Messerli *New England J. Med.* Oct.10, 2012 DOI:10.1056/NEJMon1211064

Interpretation of QSAR equations

Predictivity of QSAR equations in between data points.

The hypersurface is not smooth: activity islands vs. activity cliffs



Bryce Canyon National Park, Utah

Lit: G.M. Maggiora *J.Chem.Inf.Model.* **46** (2006) 1535.

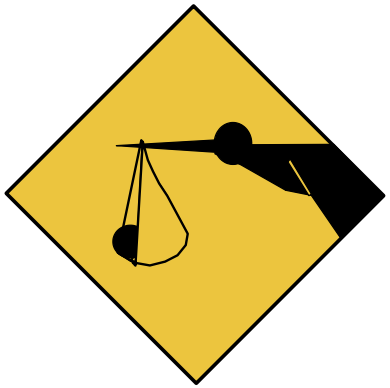
S.R. Johnson *J.Chem.Inf.Model.* **48** (2008) 25.

Interpretation of QSAR equations

Which QSAR performance is realistic?

- standard deviation (se) of 0.2–0.3 log units corresponds to a typical 2-fold error in experiments („soft data“). This gives rise to an upper limit of
- r^2 between 0.77–0.88 (for biological systems)

→ obtained correlations above 0.90 are highly likely to be accidental or due to overfitting (except for physico-chemical properties that show small errors, e.g. boiling points, logP, NMR ^{13}C shifts)

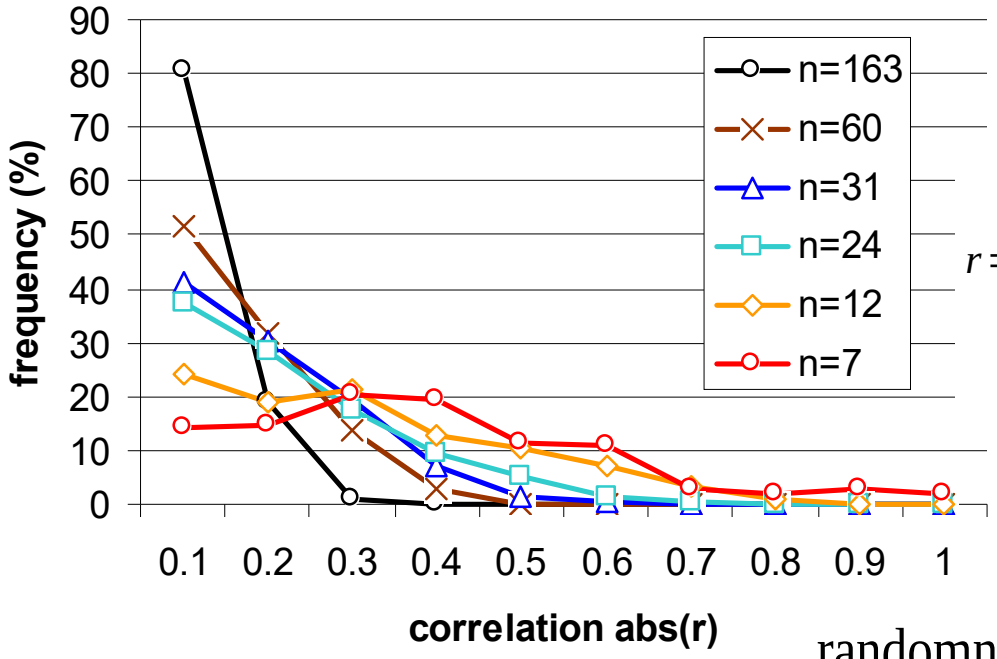


But: even random correlations can sometimes be as high as 0.84

Lit: A.M.Doweyko *J.Comput.-Aided Mol.Des.* **22** (2008) 81-89.

Interpretation of QSAR equations

Accidental correlation of a single descriptor (1000 random descriptors)



n = number of data points

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)}} \in [-1 \ldots 1]$$

$$\text{randomness}\ (\%) = \frac{10\sqrt{2n}}{\sqrt{3}} \exp\left(\frac{-n \cdot r^2}{3}\right)$$

→ Dismiss unsuitable variables from the pool of descriptors.

Lit: M.C.Hutter *J.Chem.Inf.Model.* (2011) DOI: 10.1021/ci200403j

Interpretation of QSAR equations

According to statistics more people die after being hit by a donkey than from the consequences of an airplane crash.

„An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts – for support rather than for illumination“

Andrew Lang (1844 – 1912)

further literature: R.Guha *J.Comput.-Aided Mol.Des.* **22** (2008) 857-871.