Why Models Fail

Hugo Kubinyi

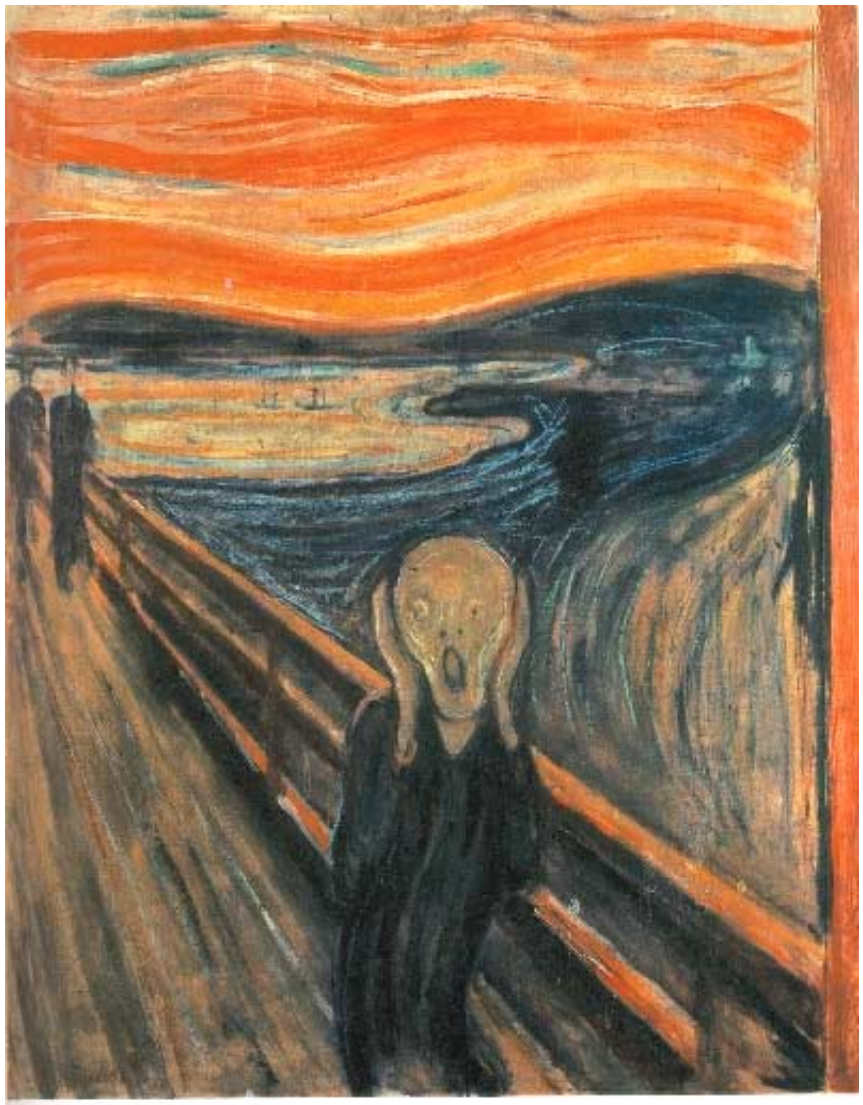
Germany

kubinyi@t-online.de

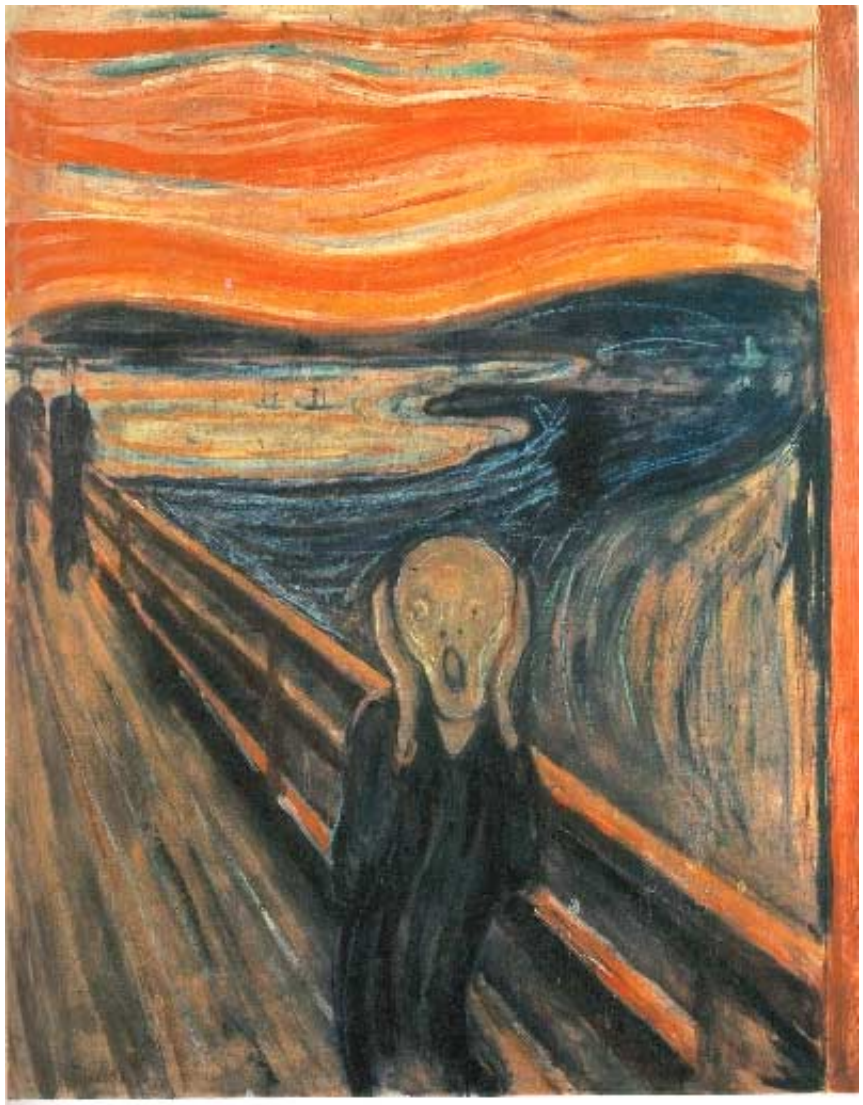
www.kubinyi.de

Some Problems in Statistical Analyses

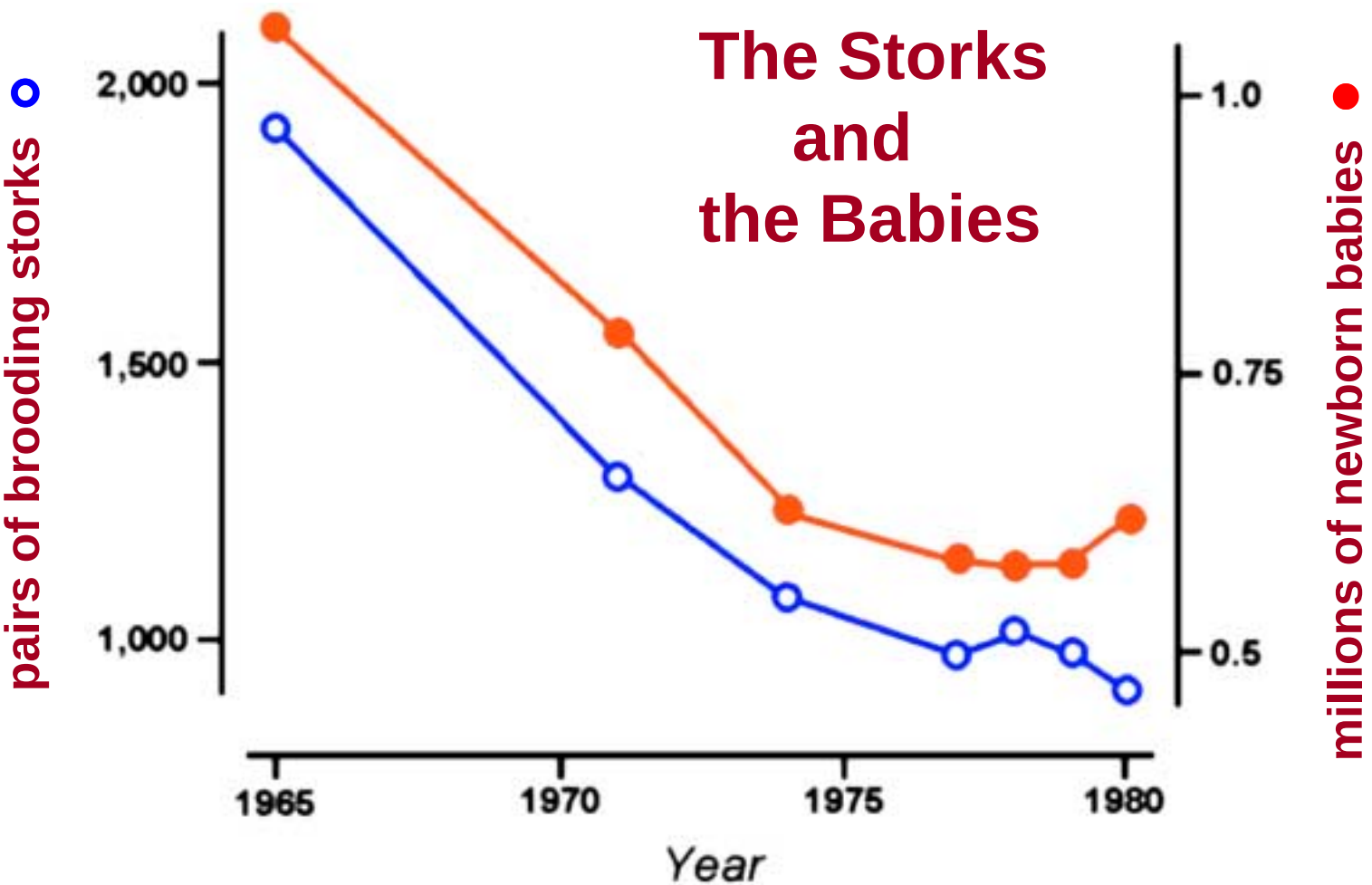
inappropriate biological data
wrong scaling of biological data
data from different labs
different binding modes
mixed data (e.g. oral absorption and bioavailability)
different mechanism of action (e.g. toxicity data)
too few data points
too many single points
lack of chemical variation
clustered data
small variance of y values
systematic error/s in y
too large errors in y values
outliers / wrong values
wrong model selection



Some More Problems in Statistical Analyses



inappropriate x variables
too many x variables (Topliss)
    a) in the model selection
    b) in the final model
wrong x variable scaling
interrelated x variables
singular matrix
elimination of variables that are
    significant only with others
insignificant model (F test)
insignificant x variables (t test)
no qualitative (biophysical) model
no causal relationship (the storks)
extrapolation too far outside of
    observation space
no validation method applied
wrong validation method, ...



Sir – There is concern in West Germany over the falling birth rate. The accompanying graph might suggest a solution that every child knows makes sense.

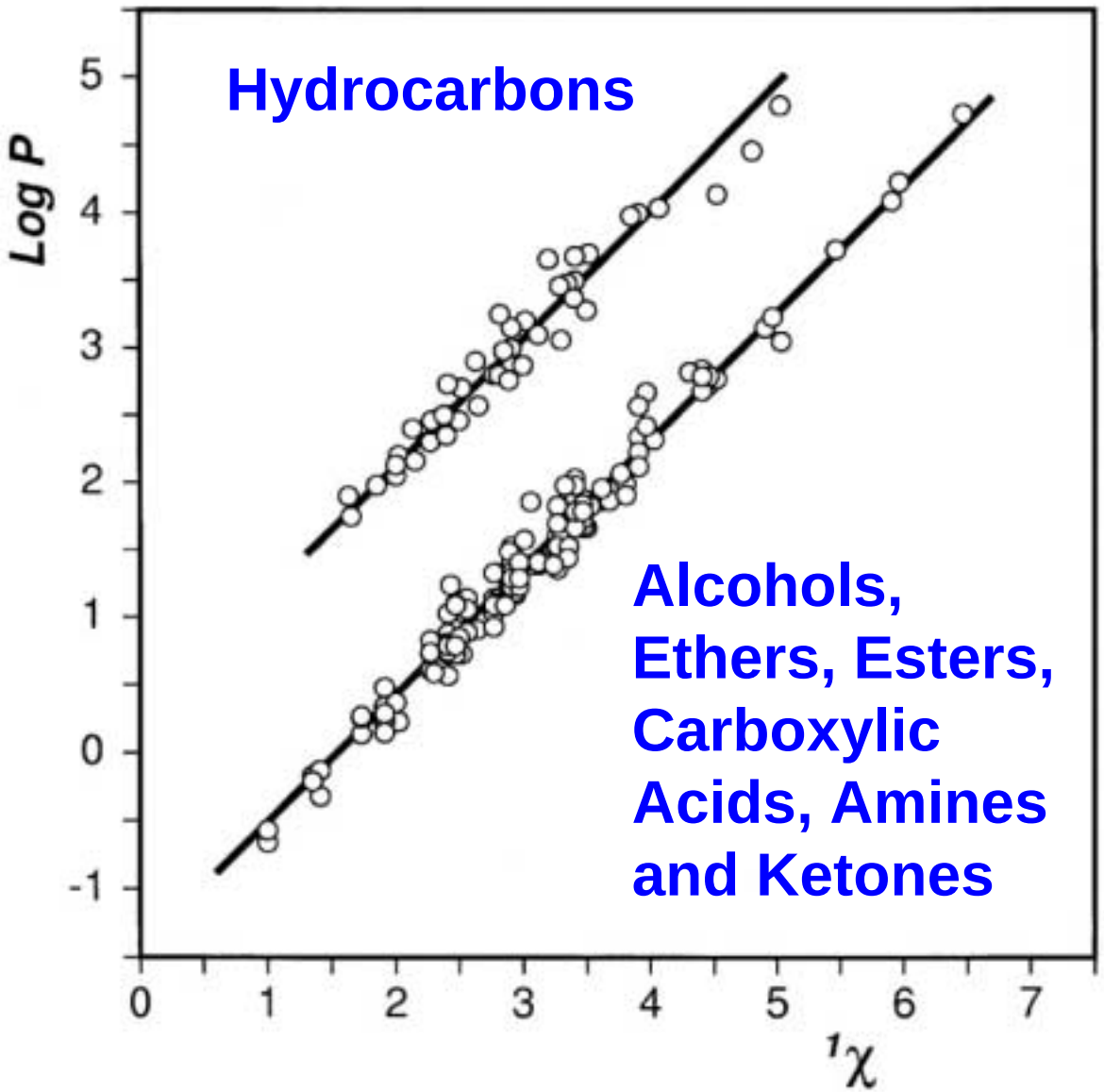
H. Sies, Nature 332, 495 (1988)

A Diagram Tells You More Than Thousand Equations

183 Hydrocarbons, Alcohols, Ethers, Esters, Carboxylic Acids, Amines and Ketones

MR *vs.* $^{1}\chi$	r = 0.908; s = 0.380; F = 855.26
MR *vs.* $^{2}\chi^{v}$	r = 0.826; s = 0.419; F = 389.58
log P *vs.* $^{1}\chi$	r = 0.719; s = 0.632; F = 193.36
log P *vs.* $^{2}\chi^{v}$	r = 0.635; s = 0.574; F = 122.33

Anil K. Saxena, Quant. Struct.-Act. Relat. 14, 142-150 (1995)



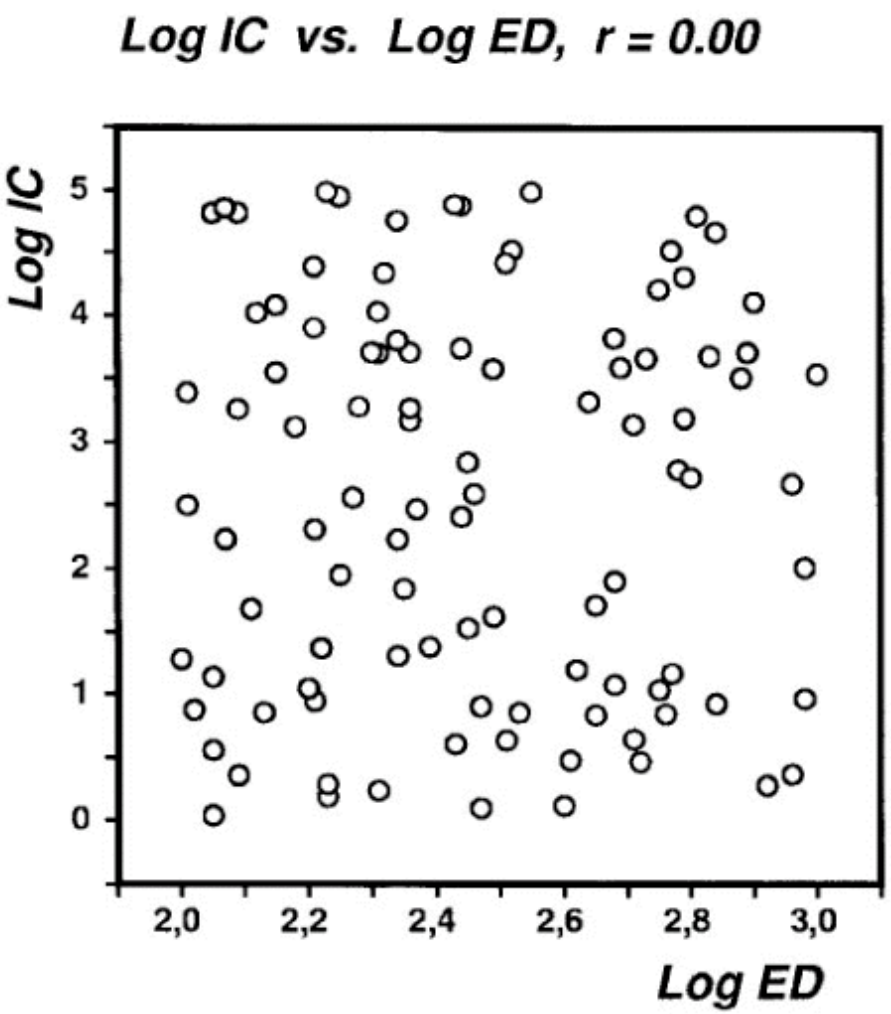
Rebuttal:

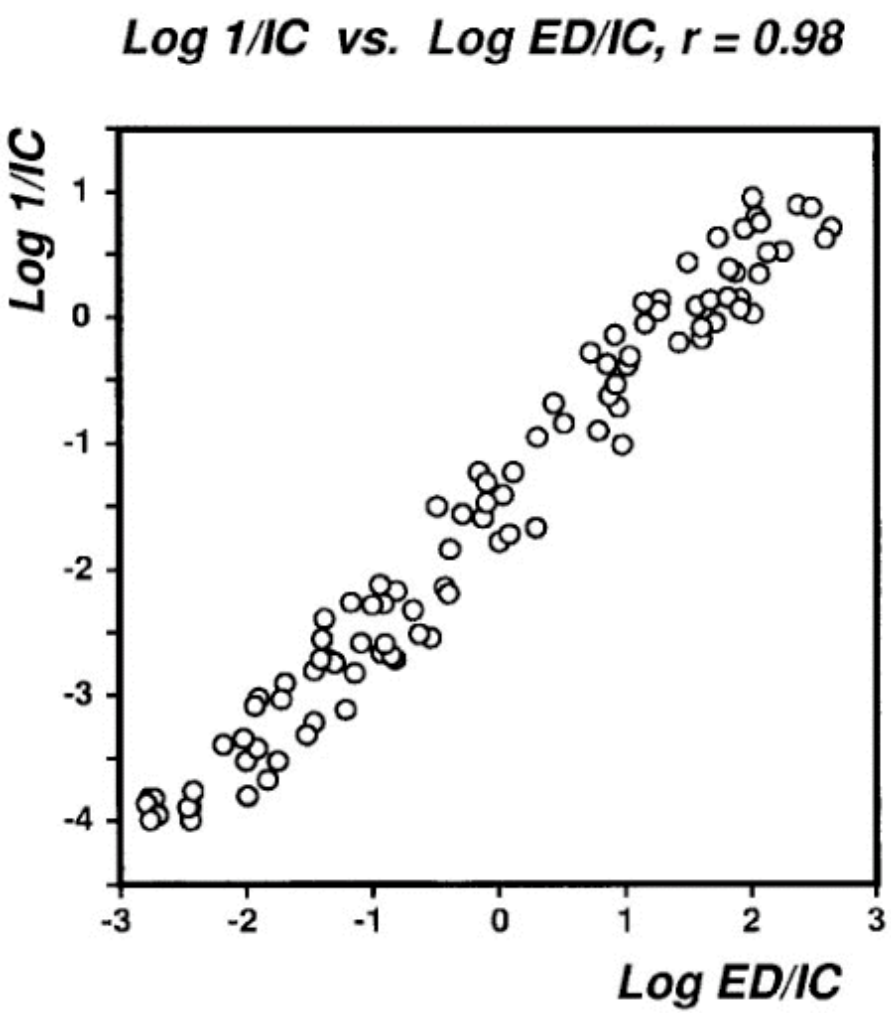
Log P =
0.941 (±0.02) $^1\chi$
- 1.693 (±0.05) I
+ 0.244 (±0.08)

(n = 183;
r = 0.990;
s = 0.150;
F = 4,633)

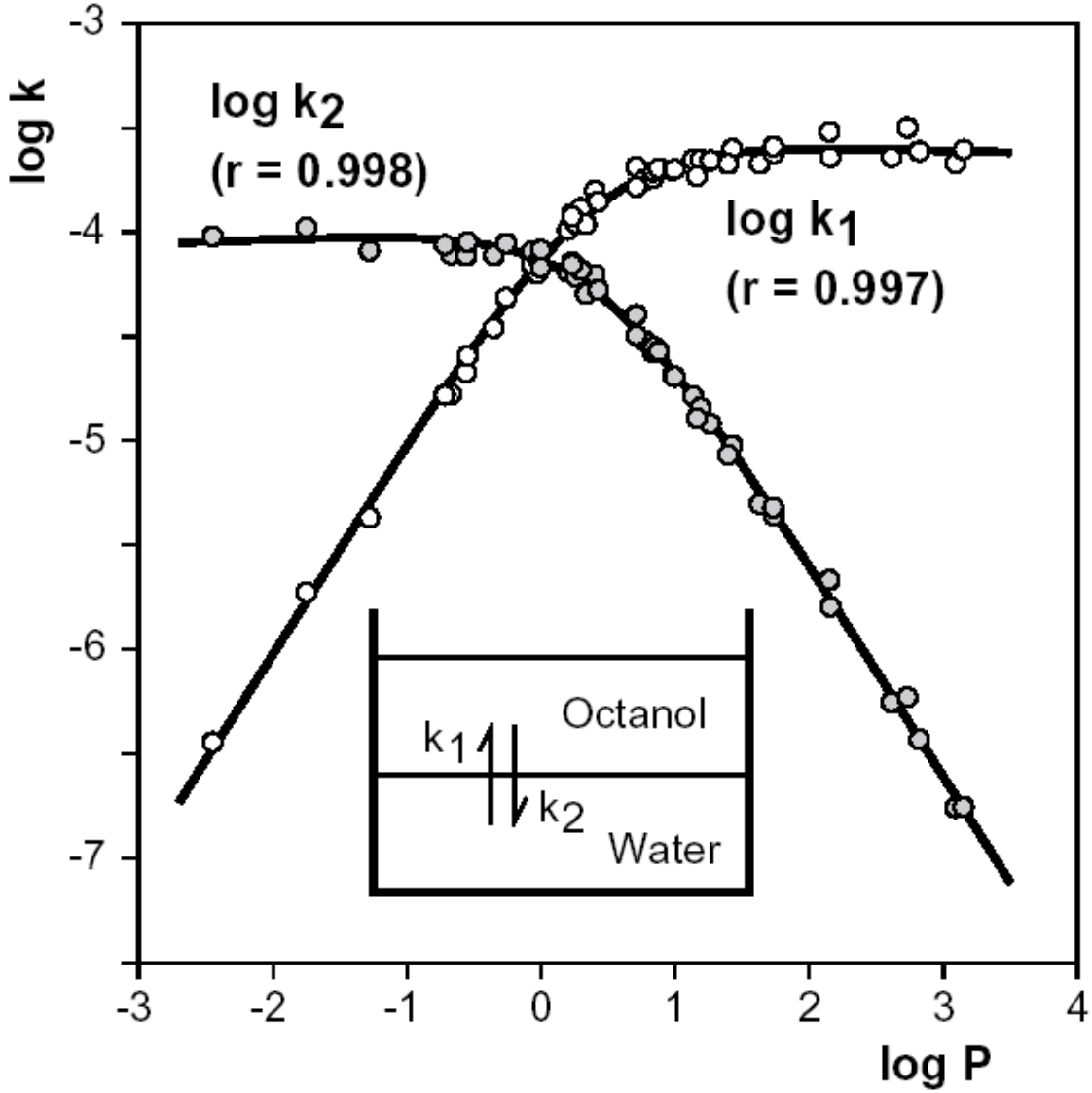
H. Kubinyi, Quant. Struct.-Act. Relat. 14, 149-150 (1995)

A Special Method for the Generation of „Good“ Correlations





Transport Rate Constants of Organic Compounds



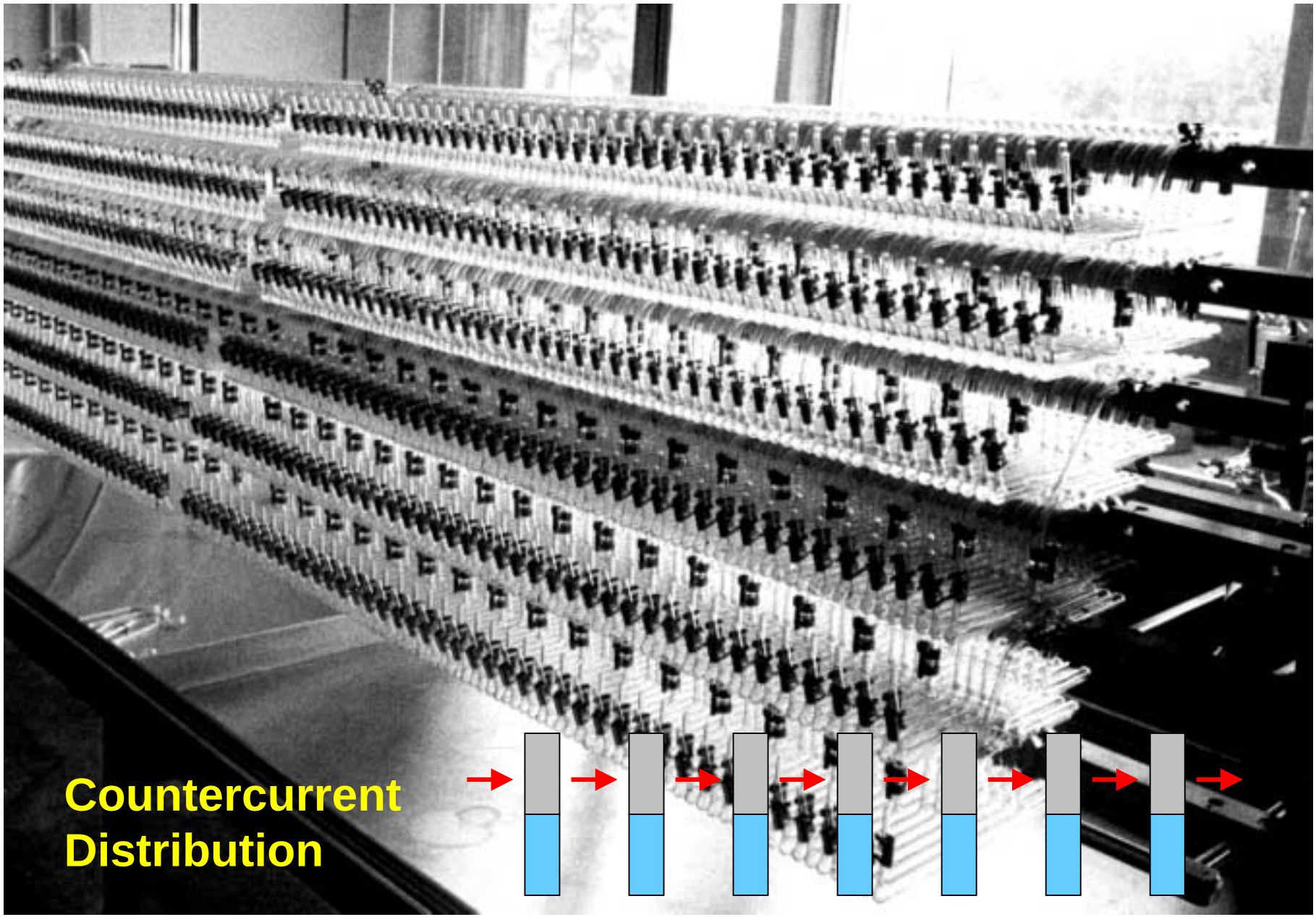
Quantitative models

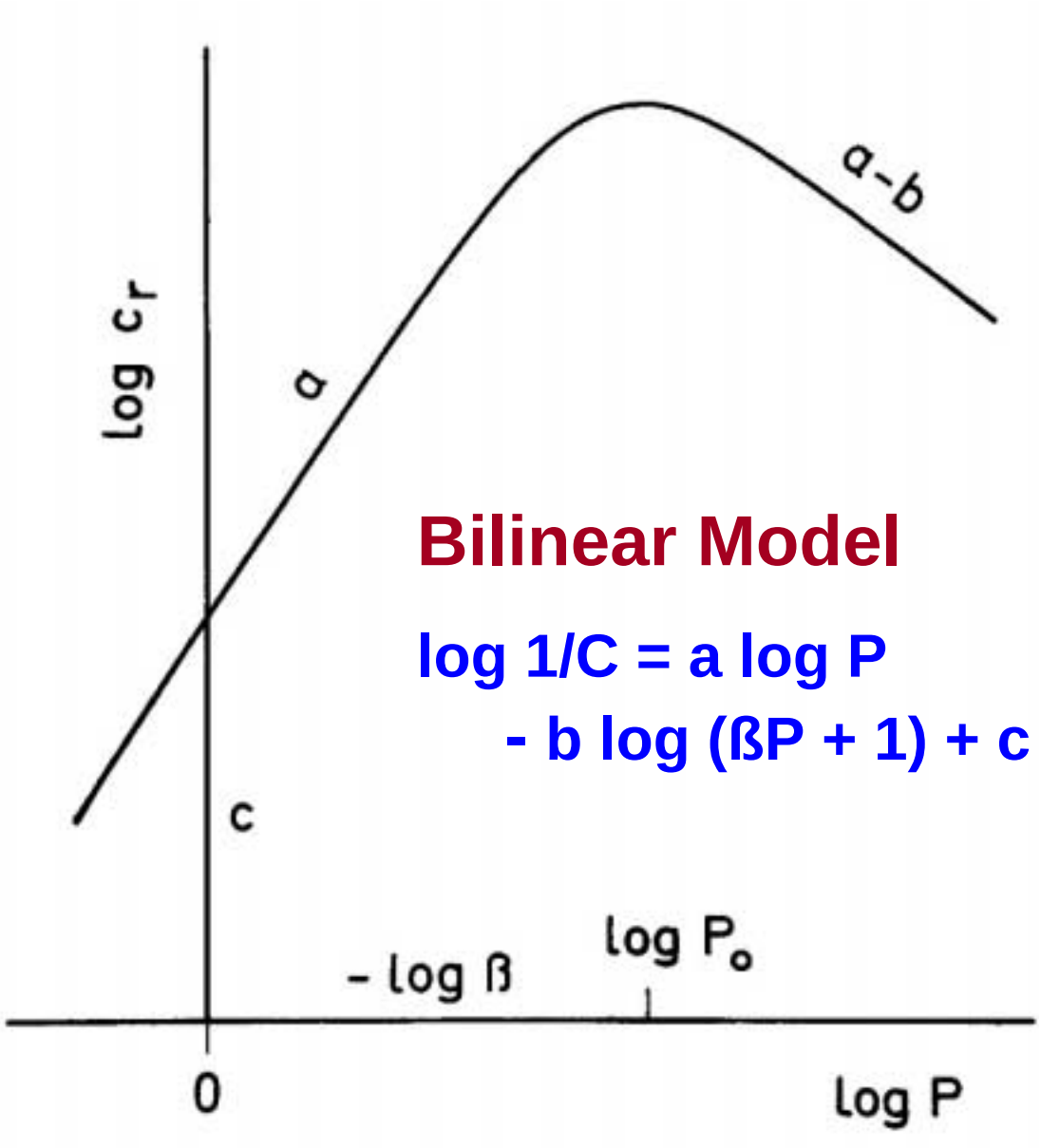
$\log k_1 = \log P - \log(\beta P+1) + c$

$\log k_2 = - \log(\beta P+1) + c$

H. Kubinyi, J. Pharm. Sci. 67, 262-263 (1978)

(experimental data by Lippold and Schneider, 1976; van de Waterbeemd et al., 1980-1982)


Countercurrent
Distribution



Bilinear Model

$$\log 1/C = a \log P - b \log (\beta P + 1) + c$$

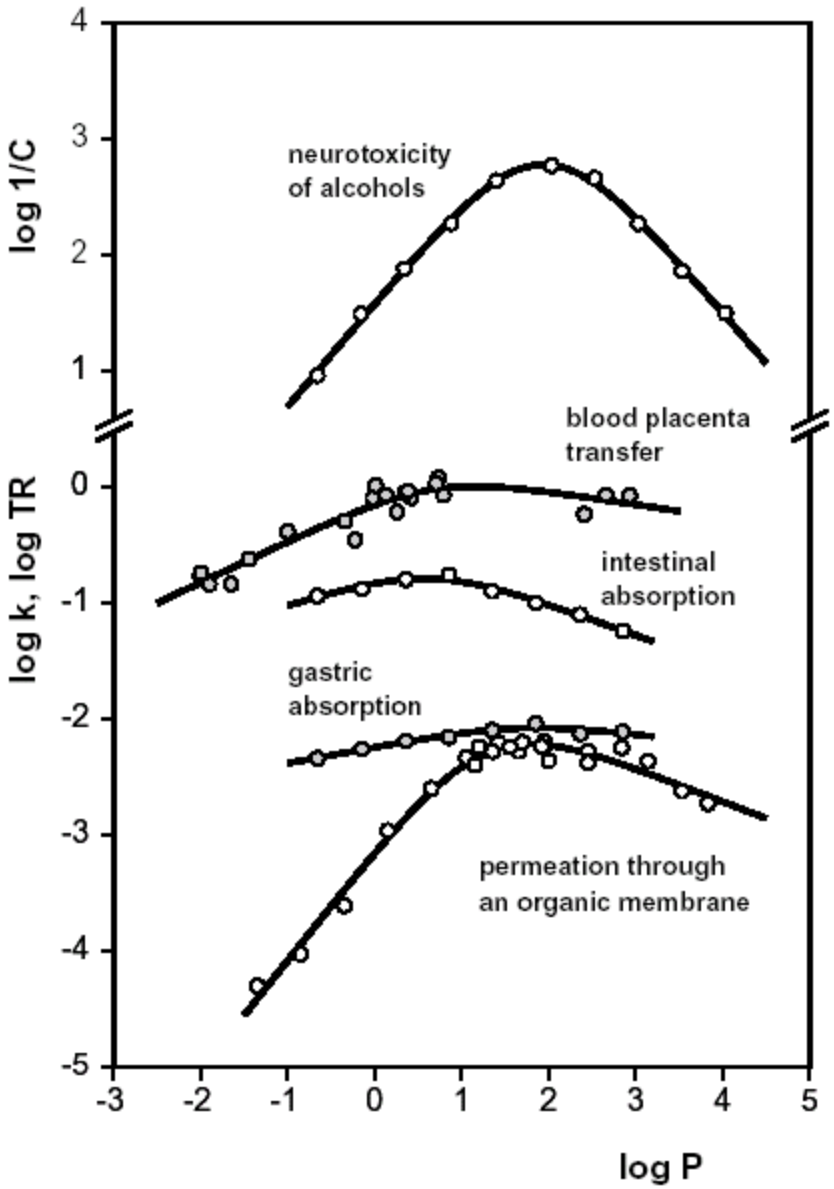
Advantages of the bilinear model

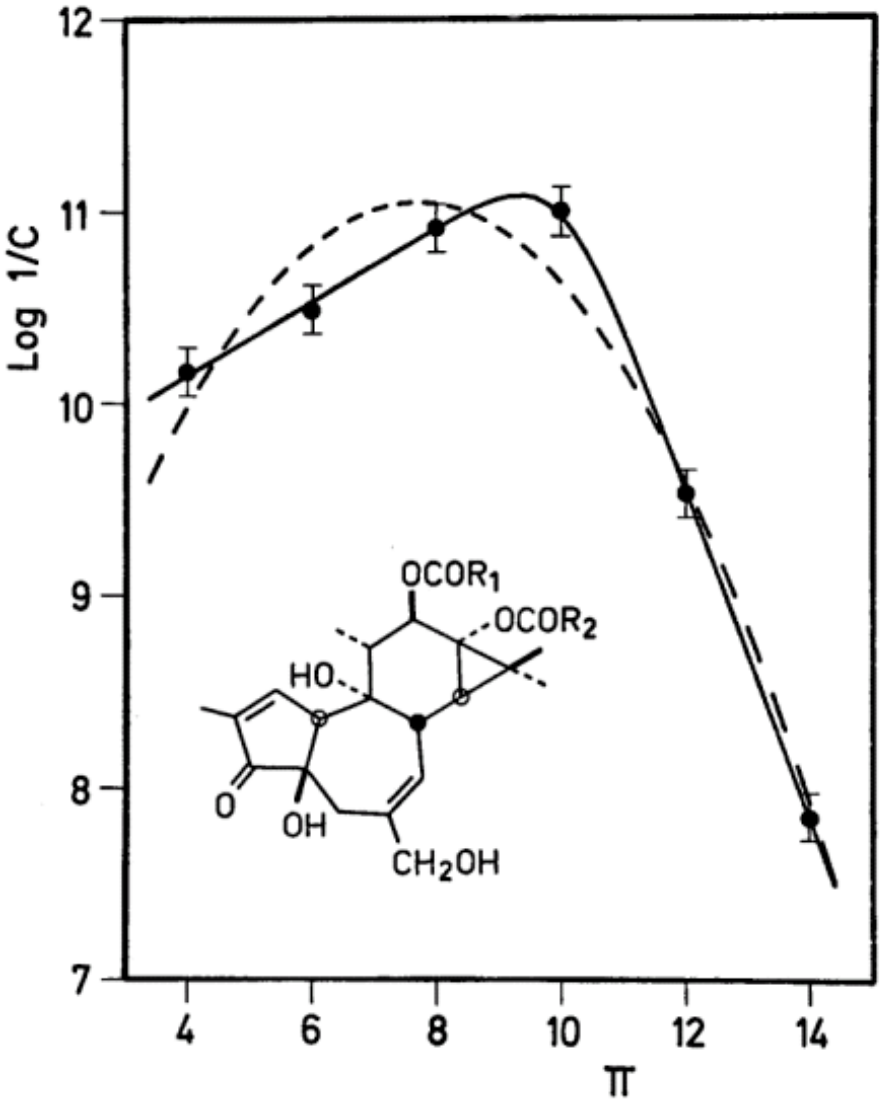
- better fit of the linear left and right sides
- better description of the lipophilicity optimum

Disadvantages of the bilinear model

- iterative estimation of the nonlinear parameter ß
- Loss of one degree of freedom (4 parameters, instead of 3)

Bilinear Relationships





Selecting the Right Model: The Zscheile Data Set

UV absorption of a mixture of adenylic acid (A), cytidylic acid (C), guanylic acid (G), and uridylic acid (U), at 36 different wavelengths

$$\varepsilon_{mixture} = c_A.\varepsilon_A + c_C.\varepsilon_C + c_G.\varepsilon_G + c_U.\varepsilon_U$$

Concentration errors up to 40% are observed due to, e.g., high intercorrelation between ε_A and ε_U (r = 0.96).

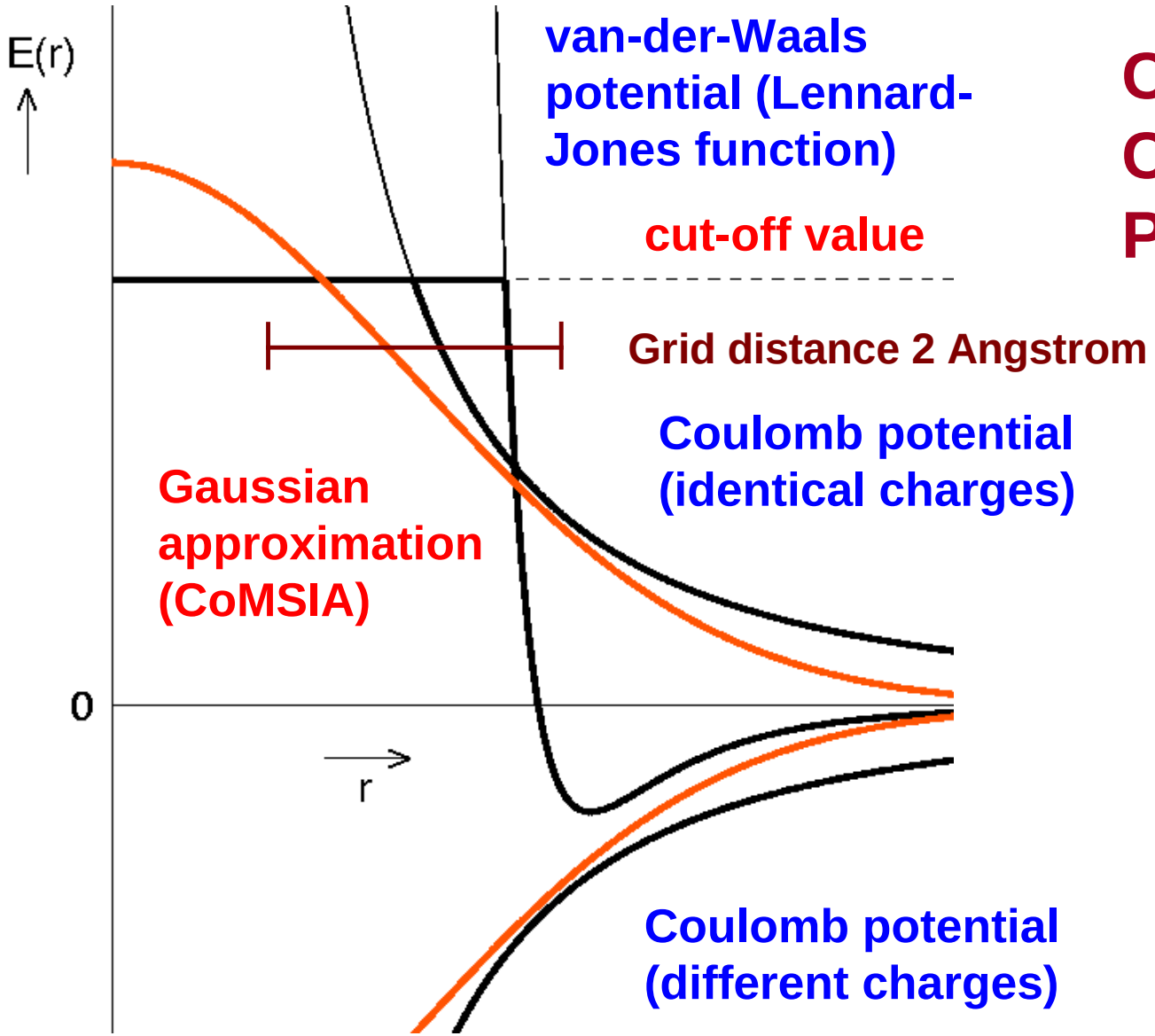
However, adding a constant term to

$$\varepsilon_{mixture} = c_A.\varepsilon_A + c_C.\varepsilon_C + c_G.\varepsilon_G + c_U.\varepsilon_U + const.$$

reduced the errors to < 10%.

H. Kubinyi, Trends Anal. Chem. 14, 199-201 (1995)

CoMFA and CoMSIA Potentials



G. Klebe et al., J. Med. Chem. 37, 4130-4146 (1994)

The Problem of Prediction

inside: trivial

outside: wrong

at the edge: 50/50

A Common Situation (e.g. the Selwood data set)

A chemist synthesizes about 30 compounds.

The biologists determines the activity values.

Both ask the chemoinformatician to derive a QSAR model.

The chemoinformatician loads 1500 variables (e.g. from the program DRAGON, Roberto Todeschini) and derives a QSAR model, containing only a few variables, which meets all statistical criteria.

Chemist, biologist and chemoinformatician publish the results. Everybody is happy.

The Real Situation (e.g. the Selwood data set)

A chemist prepares some 20 compounds.

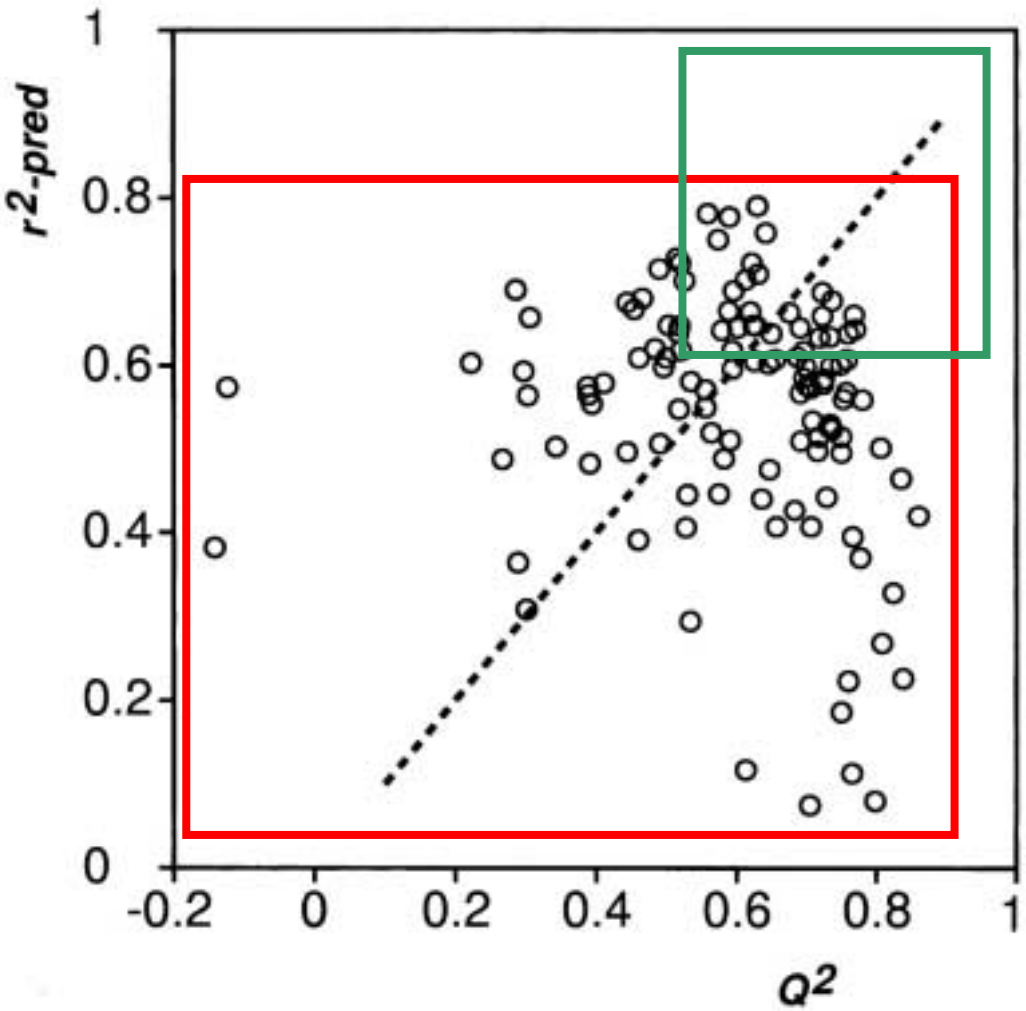
The biologist determines the activity values.

They both ask the chemoinformatician to derive a QSAR model.

The resulting model does not contain more than four variables, is selected from about fifty variables and is validated by all statistical criteria, including LOO/LMO cross-validation and y scrambling.

How good is the predictivity of the model for a test set of 10 compounds?

External vs. Internal Predictivity

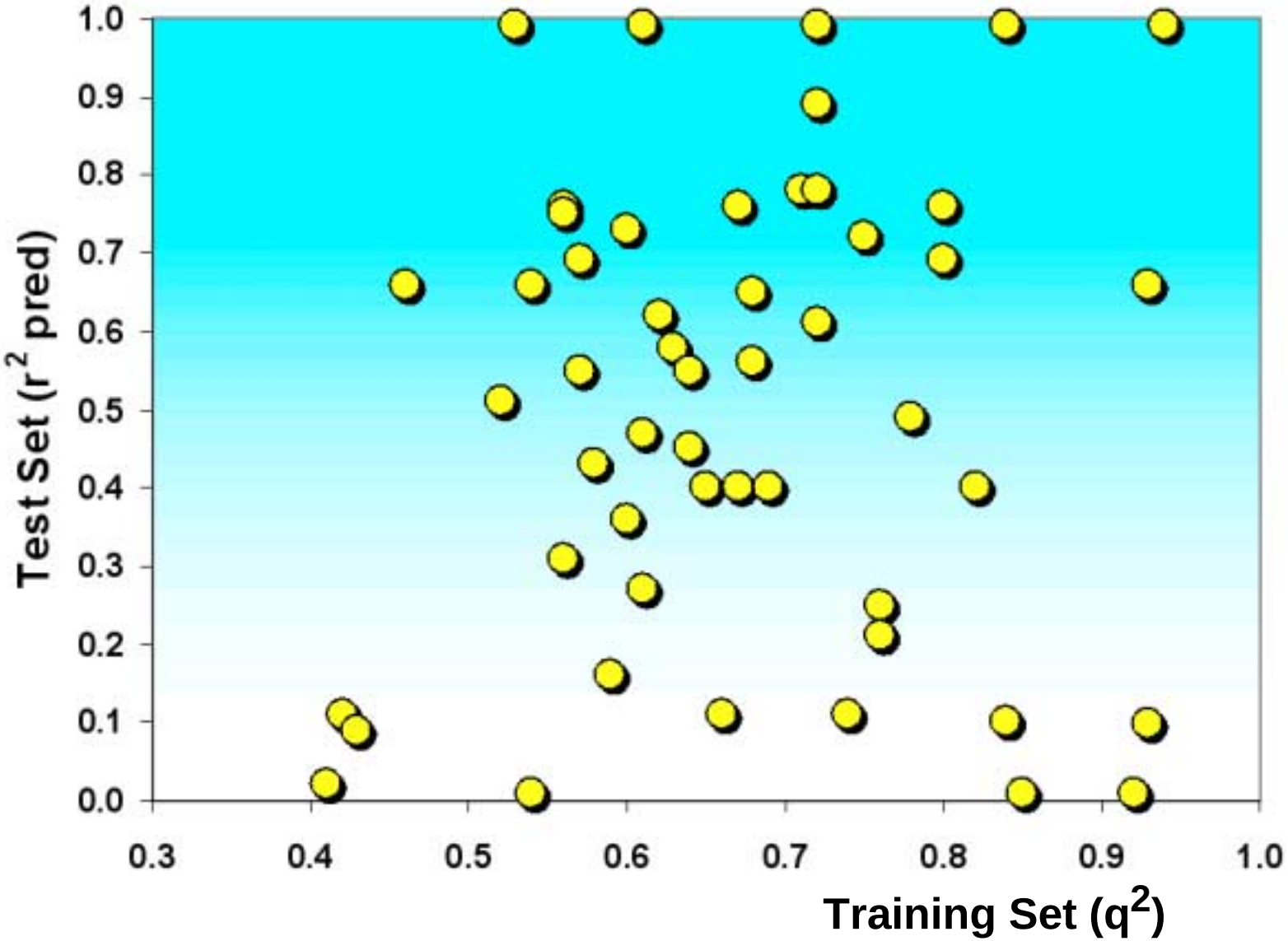


The „Kubinyi Paradox“

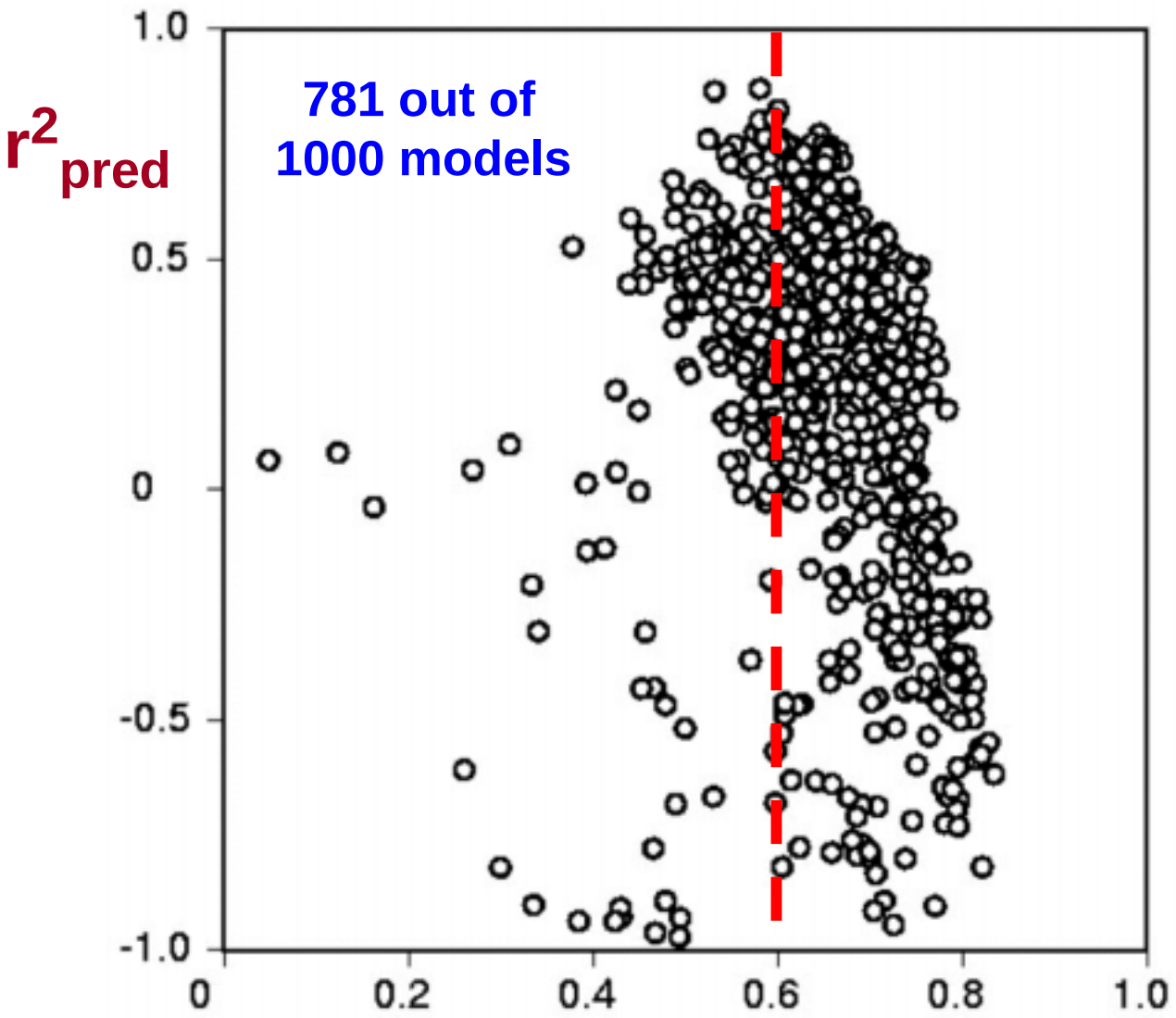
J. H. van Drie, Curr. Pharm. Des. 9, 1649-1664 (2003);
J. H. van Drie, in: Computational Medicinal Chemistry for Drug Discovery, P. Bultinck et al., Eds., Marcel Dekker, 2004, pp. 437-460.

Data from H. Kubinyi et al., J. Med. Chem. 41, 2553-2564 (1998).

Test vs. Training Set Predictivity (A. Doweyko, ACS 2004)



External vs. Internal Predictivity, Selwood Data

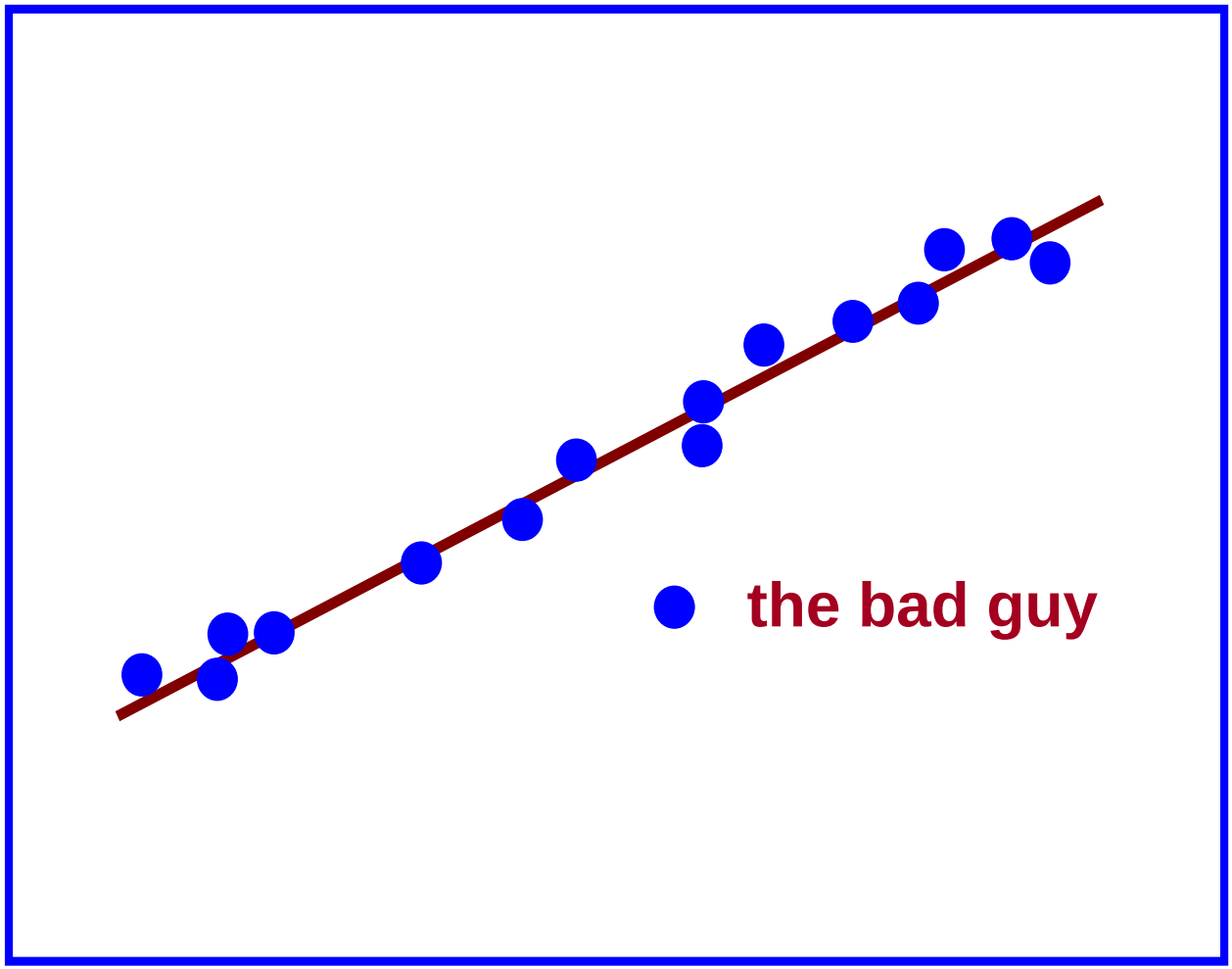


Training sets:
n = 21
Test sets:
n = 10

The „best fit“ models are <u>not</u> the best ones in external prediction !

H. Kubinyi, Proc. 15th EuroQSAR, 2006, pp. 30-33

„Good“ and „Bad“ Guys in Regression Analysis



outlier in the test set

r^2, Q^2 good
r^2_{pred} poor

outlier in the training set

r^2, Q^2 poor
r^2_{pred} good

External vs. Internal Predictivity

Corticosteroid-binding globulin affinities of steroids

log 1/CBG = 1.861 (±0.46) [4,5 >C=C<] + 5.186 (±0.36)

(n = 31; r = 0.838; s = 0.600; F = 68.28;

Q^2 = 0.667; s_{PRESS} = 0.634)

Training set # 1-21; test set # 22-31

Q^2 = 0.726; r^2_{pred} = 0.477; s_{PRED} = 0.733

Training set # 1-12 and 23-31; test set # 13-22

Q^2 = 0.454; r^2_{pred} = 0.909; s_{PRED} = 0.406

H. Kubinyi, in: Computer-Assisted Lead Finding and Optimization van de Waterbeemd, H., Testa, B., and Folkers, G., Eds.; VHChA and VCH, Basel, Weinheim, 1997; pp. 9-28

A Simple Explanation of the Prediction Paradox

Even in the absence of real outliers, external prediction will be worse than fit: the model tries to „fit the error“.

Accordingly, external predictions contain the model error and the experimental error.

Variable selection in QSAR and CoMFA

No independent variable selection is performed in the crossvalidation runs; correspondingly, variables that were included to “explain the error” remain in the model and cause wrong predictions.

Chemical vs. Biological Landscapes

“Activity landscapes are not continuous, they contain cliffs, like the Bryce Canyon”

rem: applies also to scoring functions !

G. M. Maggiora, On outliers and activity cliffs - why QSAR often disappoints, J. Chem. Inf. Model. 46, 1535 (2006)

One must rely heavily on statistics ..... but, at each critical step ..... one must set aside statistics and ask questions.

... without a qualitative perspective one is apt to generate statistical unicorns, beasts that exist on paper but not in reality.

… one can correlate a set of dependent variables using random numbers ..... such correlations meet the usual criteria of high significance ...

S. H. Unger and C. Hansch
J. Med. Chem. 16, 745-749 (1973)

Summary, Conclusions and Recommendations

Apply the Unger and Hansch recommendations:

- Select meaningful variables
- Eliminate interrelated variables
- Justify variable selection by statistics
- Principle of parsimony (Ockham‘s Razor)
- Number of variables to choose from (John Topliss)
- Number of variables in the model (John Topliss)
- Qualitative biophysical model

Additional recommendations:

- Search for outliers in the training and test sets
- Beware of Q^2 (Alex Tropsha)
- Do not overrely in y scrambling
- Do not expect your model to be predictive !

Many thanks to

Erich Hecker, MPI Biochemistry and DKFZ

Ott-Hermann Kehrhahn, KNOLL

Reinhard Neudert, BASF (now at Wiley)

Hans-Joachim Böhm, BASF
(now at Roche)

Gerhard Klebe, BASF
(now at Univ. Marburg)

Jens Sadowski, BASF (now at AstraZeneca)

as well as many other colleagues of the BASF Drug Design Group