Validation and Predictivity of QSAR Models

Hugo Kubinyi

University of Heidelberg, Germany

E-Mail kubinyi@t-online.de
http://home.t-online.de/home/kubinyi

15th EURO-QSAR Symposium
Istanbul, Turkey, Sept. 2004

University of Heidelberg

Validation of QSAR Models

Statistical Parameters
- r, s and F values, confidence intervals

Topliss criteria
- number of tested and included variables

Hansch-Unger criteria
- meaningful parameters, statistical significance,
- Ockham's razor, biophysical model

Crossvalidation: Q^2 and s_{PRESS} values

(Bootstrapping)

Lateral Validation

Y scrambling

External Predictions

A Common Situation

A chemist synthesizes about 30 compounds.

The biologist 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 Selwood Data Set

n = 31 compounds and k = 53 independent variables.

Theoretically, there are:

53	one-variable models,
1,378	two-variable models,
23,426	three-variable models,
292,825	four-variable models,
....,	
22,957,480	six-variable models,
....,	in total

7,160,260,814,092,303 regression models,

containing one to 29 variables,
selected from 53 X-variables.

The 53 X Variables of the Selwood Data Set

ATCH1 - ATCH10 = partial atomic charges
DIPV_X, DIPV_Y and DIPV_Z = dipole vectors
DIPMOM = dipole moment
ESDL1 - ESDL10 = electrophilic superdelocalizability
NSDL1 - NSDL10 = nucleophilic superdelocalizability
VDWVOL = van der Waals volume
SURF_A = surface area
MOFI_X, MOFI_Y and MOFI_Z = moments of inertia
PEAX_X, PEAX_Y and PEAX_Z = ellipsoid axes
MOL_WT = molecular weight
S8_1DX, S8_1DY and S8_1DZ = substituent dimensions
S8_1CX, S8_1CY and S8_1CZ = substituent centers
LOGP = partition coefficient
M_PNT = melting point
SUM_F and SUM_R = sums of the F and R constants

Ockham‘s Razor - Keep Things Simple !

only models with up to four variables are considered in the following simulations (317,682 different models = complete coverage)

Pluralitas non est ponenda sine necessitate
(≈ avoid complexity if not necessary)

Questions

Can we derive „good“ (statistically valid) models: yes

Do our models have internal predictivity (Q^2 values): yes

Are these models „better“ than models from scrambled or random data (y, x, y and x): yes

Are 53 X variables too many to select from: no, fine

Can our models predict a test set (r^2_{pred} value): not at all

Is there a relationship between internal and external predictivity: by no means

Results of PLS and Regression Analyses

a) PLS, all variables (5 components)

r = 0.929; s = 0.335; F = 31.58

Q^2 = 0.279; s_{PRESS} = 0.768

b) Regression (best 3-variable model)

r = 0.849; s = 0.460; F = 23.27

Q^2 = 0.647; s_{PRESS} = 0.518

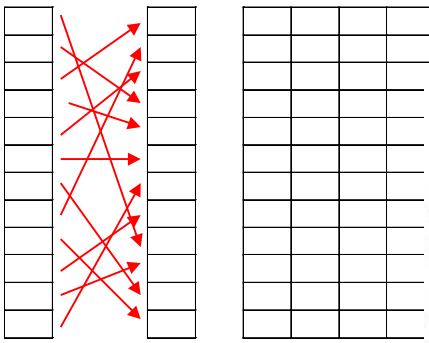
c) PLS, reduced variable set (5 components)

r = 0.909; s = 0.376; F = 23.91

Q^2 = 0.671; s_{PRESS} = 0.519

University of Heidelberg

Y Scrambling - Random Permutations of Y Values

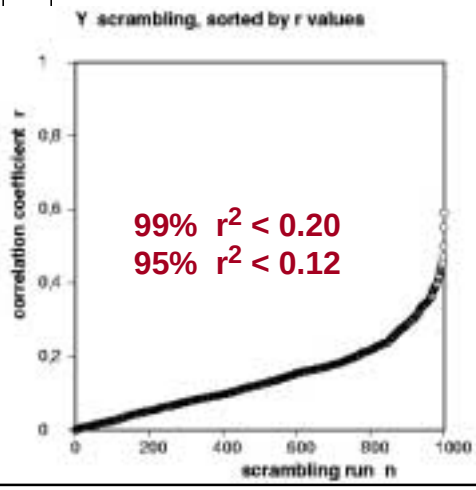


X block rem
scrambled y vect

scrambling

original y vector

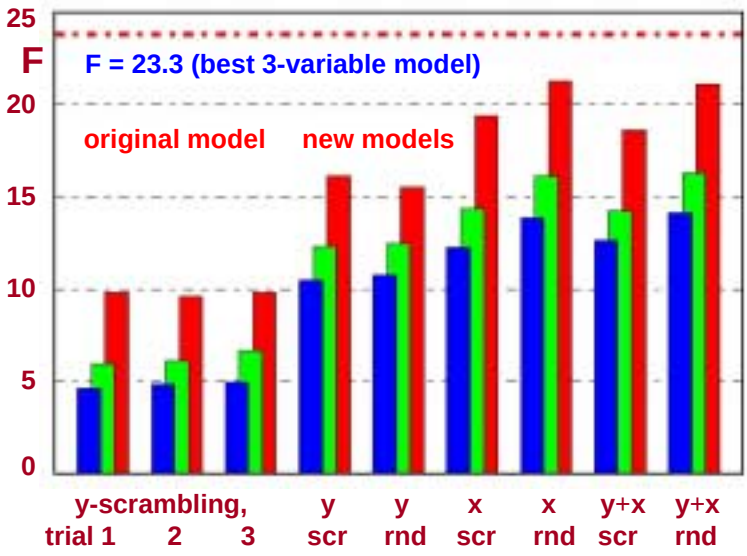
will y vs. y correlations disturb the result ?



99% $r^2 < 0.20$
95% $r^2 < 0.12$

University of Heidelberg

Scrambling and Random Y and X Values

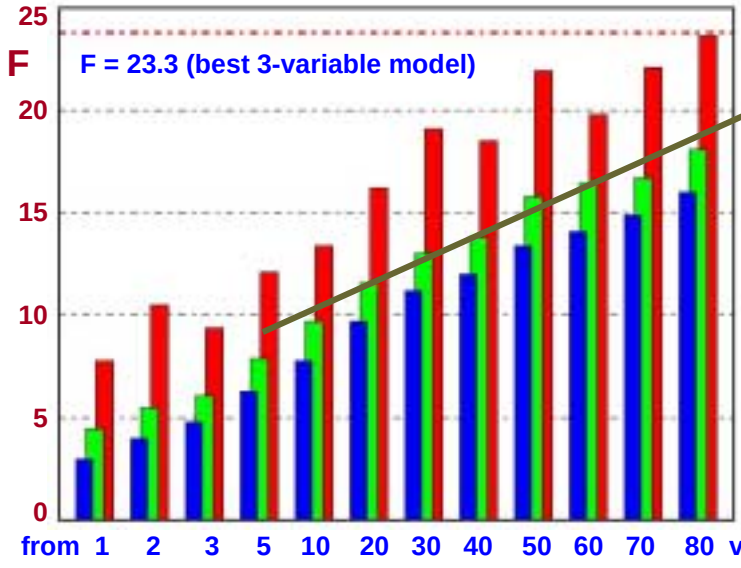


1000 runs for each group, best models.

90%
95%
99%
below F value

University of Heidelberg

Models Selected from Random X Variables



1000 runs for each group, best models.

90%

95%

99%

below F value

University of Heidelberg

The Real Situation

A chemist prepares some 20 compounds.

The biologist determines the activity values.

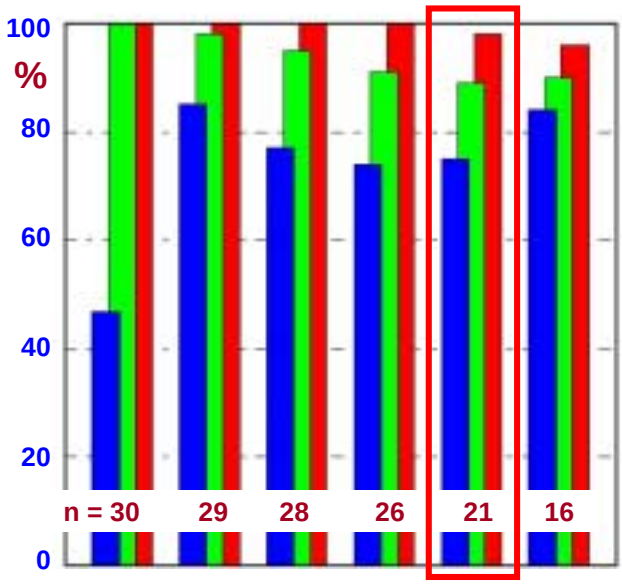
They both ask the chemoinformatician to derive a QSAR model.

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 cross-validation and y scrambling.

How good is the internal predictivity, how good is the external predicitivity for a test set of 10 compounds?

University of Heidelberg

Training Sets, Internal Predictivity (LOO)



1000 runs for each group.

New model selected for every run = 1.9 billion runs

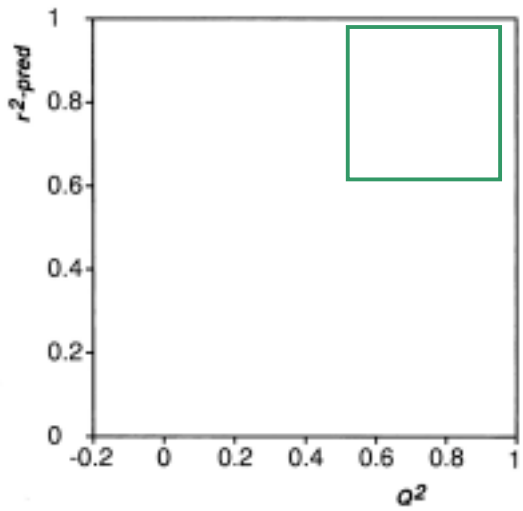
$Q^2 > 0.6$

$Q^2 > 0.5$

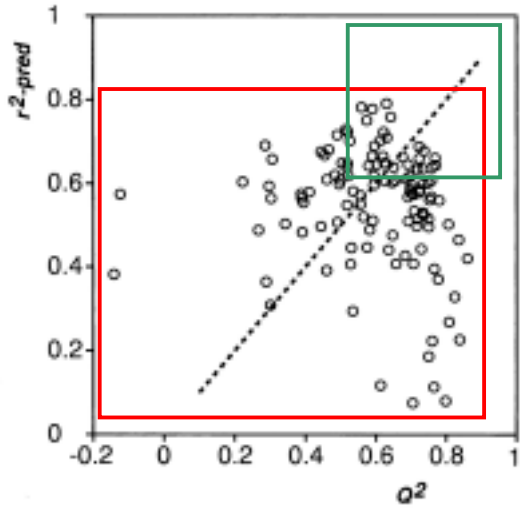
$Q^2 > 0$

University of Heidelberg

External vs. Internal Predictivity



External vs. Internal Predictivity

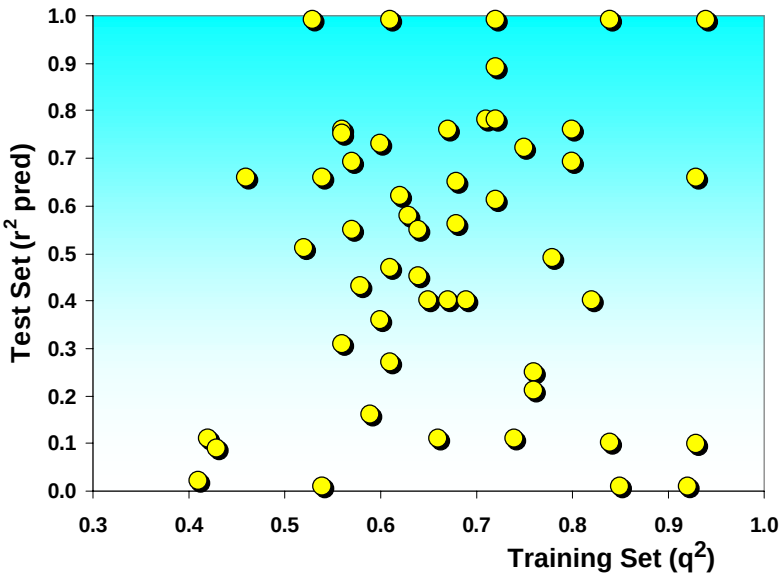


The „Kubinyi Paradox“

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.

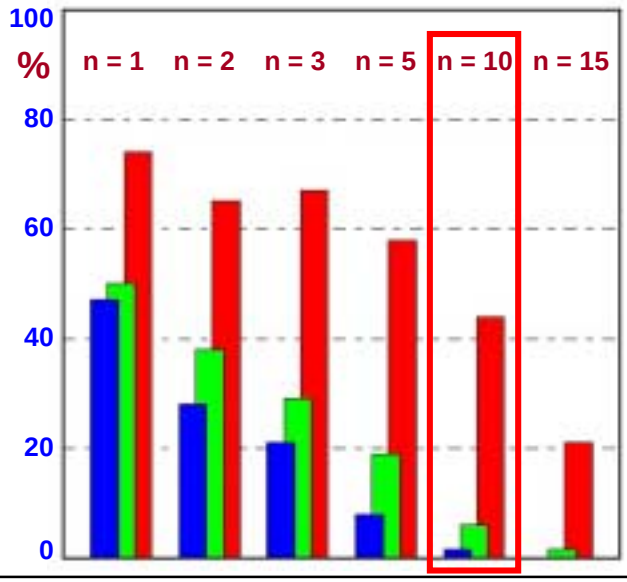
Data from H. Kubinyi et al., J. Med. Chem. 41, 2553-2564 (1998).

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



University of Heidelberg

Test Sets, External Predictivity



1000 runs for each group.

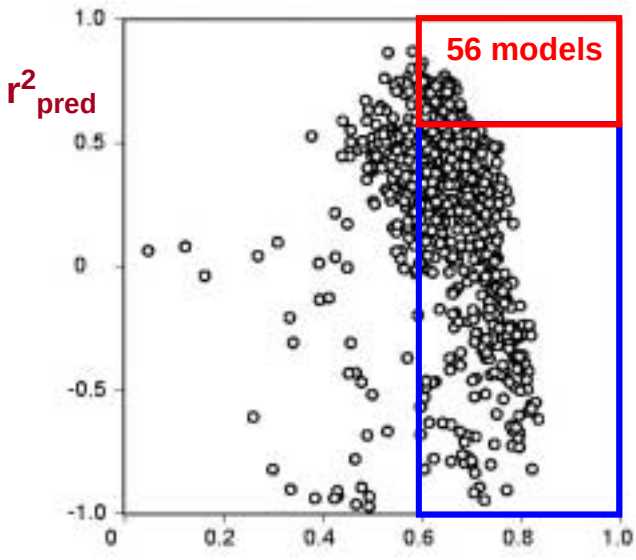
$r^2_{pred} > 0.6$

$r^2_{pred} > 0.5$

$r^2_{pred} > 0$

University of Heidelberg

External vs. Internal Predictivity, Selwood Data



Test set (n = 10) predictions from best training set models (n = 21)

red area:
56 models
(out of 1000)

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

Q^2

Answers to Our Questions

1) We can derive „good“ (statistically valid) models
2) The models have „good“ internal predictivity
3) These models are significantly „better“ than models from scrambled or random data (y, x, y and x)
4) 53 X variables are not too many to select from
5) The models have no external predictivity at all !
6) There is no relationship between internal and external predictivity

Reasons? Explanations? Help?

University of Heidelberg

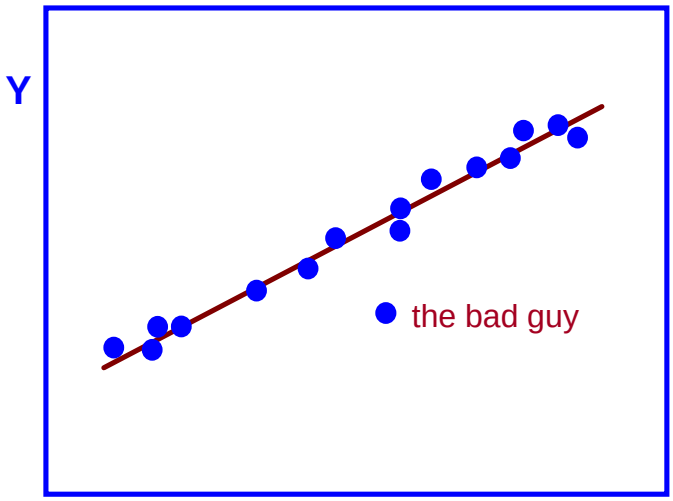
S. H. Unger and C. Hansch
J. Med. Chem. 16, 745-749 (1973)

One must rely heavily on statistics in formulating a quantitative model but, at each critical step in constructing the model, 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.

… it has recently become all too clear that one can correlate a set of dependent variables using random numbers as dependent variables. Such correlations meet the usual criteria of high significance ...

„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

Summary, Conclusions and Recommendations

Apply the Unger and Hansch recommendations:

1. Selection of meaningful variables
2. Elimination of interrelated variables
3. Justification of variable choices by statistics
4. Principle of parsimony (Ockham's Razor)
5. Number of variables to choose from
6. Number of variables in the model
7. Qualitative biophysical model

Additional recommendations:

8. Beware of Q^2 (Alex Tropsha)
9. Search for outliers in the test set
10. Do not expect your model to be predictive

University of Heidelberg

Summary, Conclusions and Recommendations

La Trahison des Images (The Perfidy of Images)

R. Magritte

"All Models Are Wrong But Some Are Useful."
George E. P. Box, 1979