Variable Selection and Model Validation

Hugo Kubinyi

Germany

E-Mail kubinyi@t-online.de
HomePage www.kubinyi.de

Hugo Kubinyi, www.kubinyi.de

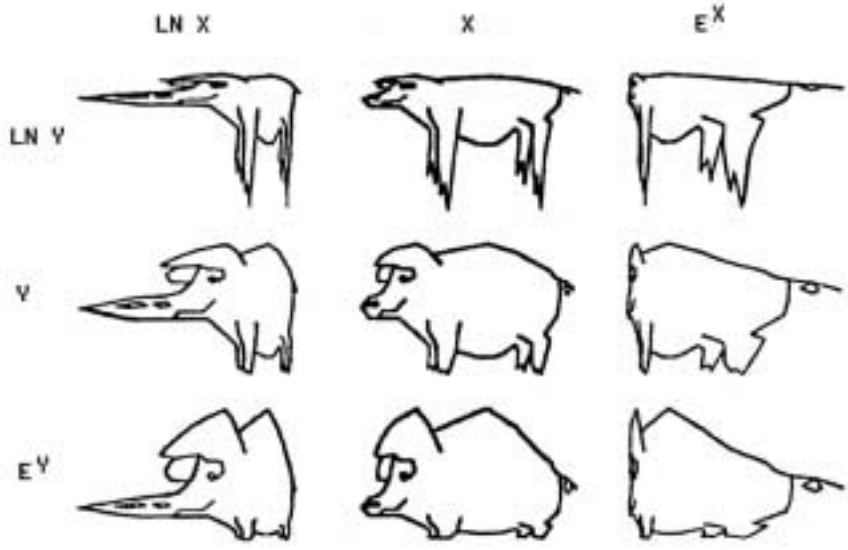
A Few Problems in Statistical Analyses

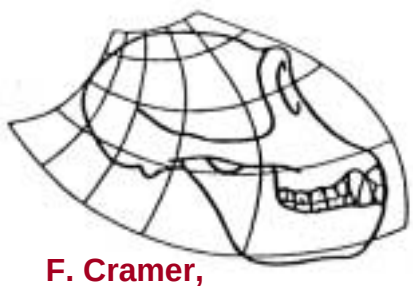
- inappropriate biological data
- wrong scaling of biological data
- data from different labs
- different binding modes
- mixed data (e.g. oral absorption and bioavailability)
- different mechanism of action (e.g. toxicity data)
- too few data points
- too many single points
- lack of chemical variation
- 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

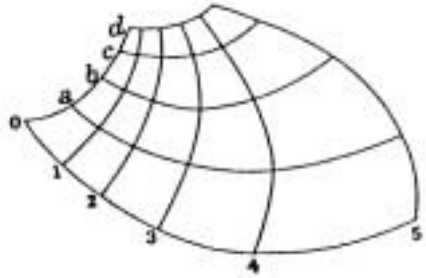
- inappropriate x variables
- too many x variables (Topliss)
  - a) in the model selection
  - b) in the final model
- x variable scaling in CoMFA fields
- interrelated x variables
- singular matrix
- elimination of variables that are significant only with others
- insignificant model (F test)
- insignificant 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, ...

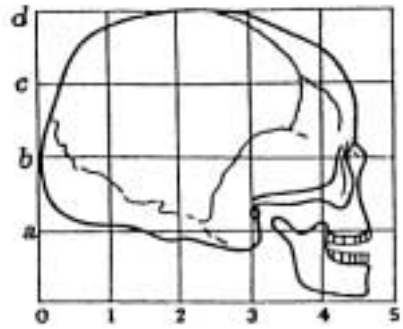
Scaling of Variables



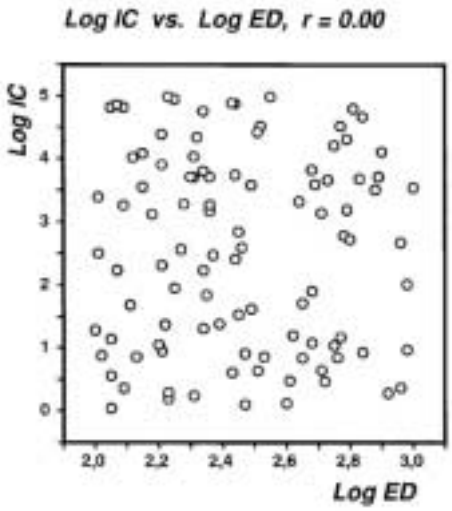


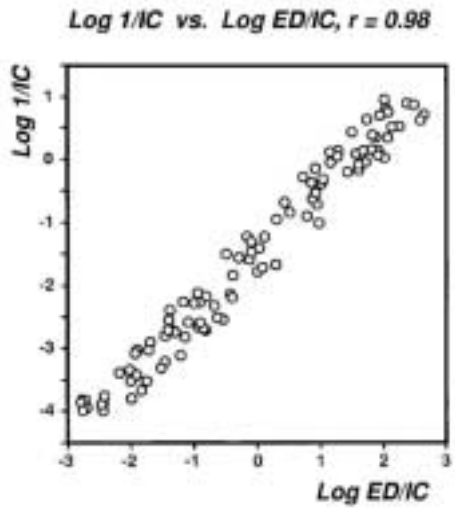
F. Cramer,
Chaos and Order





A Special Method for the Generation of „Good“ Correlations





Bailar's Laws of Data Analysis
(Clin. Pharmacol. Therapeutics, 1979)

- There are no "right" answers
- Statistics is not the only way to wisdom
- Rare events happen all the time
- No sample is ever large enough - so what?
- No analysis is ever perfect - so what?
- Something is always wrong with the data.

How to Lie With Statistics (Darrell Huff)
Lies, Damned Lies and Statistics (Benjamin Disraeli)

- All models are wrong - some may be useful
- The scaling of variables changes the result
- A diagram tells you more than thousand equations
- Validation - an extremely difficult problem.

Hugo Kubinyi, www.kubinyi.de

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 ...

Selection and Validation of QSAR Regression Models

- Careful selection of independent variables
- Significance of the variables (statistical parameters)
- Principle of parsimony (Occam's Razor)
- Minimum number of compounds per variable
- Importance of a qualitative (biophysical) model

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

Other References

S. Wold, Validation of QSAR's, Quant. Struct.-Act. Relat. 10, 191-193 (1991)

H. Mager and P. P. Mager, Validation of QSAR's: Some Reflections, Quant. Struct.-Act. Relat. 11, 518-521 (1992)

U. Thibaut et al., Recommendations for CoMFA Studies and 3D QSAR Publications, H. Kubinyi, Ed., 3D QSAR in Drug Design, ESCOM, Leiden, 1993.

Hugo Kubinyi, www.kubinyi.de

Statistical Parameters, Fitness Criteria and Validation of QSAR Results

Regression coefficient values, t test
Statistical parameters r, s, Q^2, s_{PRESS}

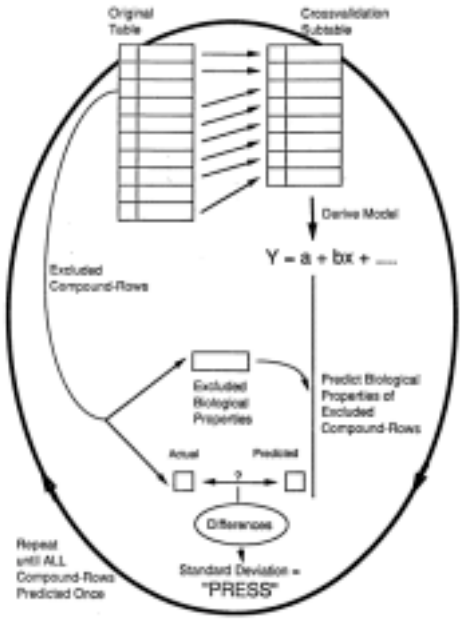
$$F = \frac{r^2 \cdot (n-k-1)}{k \cdot (1-r^2)} \qquad FIT = \frac{r^2 \cdot (n-k-1)}{(n+k^2) \cdot (1-r^2)}$$

Crossvalidation (group size?)
Bootstrapping
Biophysical model
Lateral validation
Y scrambling
Correct predictions (test set)

Jackknife Method

corresponds to LOO cross-validation; used for the estimation of confidence intervals of nonlinear parameters, like ß, and log P_o.

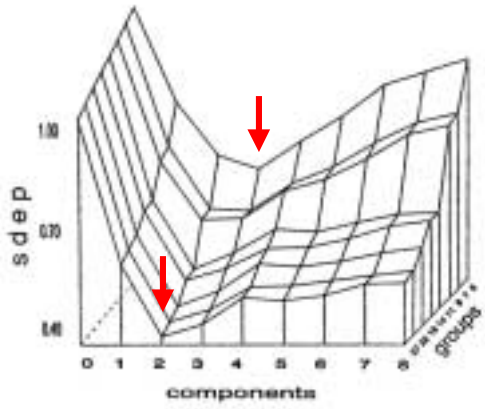
S. W. Dietrich, N. D. Dreyer, C. Hansch and D. L. Bentley, J. Med. Chem. 23, 1201-1205 (1980)



PLS Analysis Crossvalidation

In crossvalidation, many PLS runs are performed in which one ("leave-one-out" technique, LOO) or several objects (cross-validation in groups) are eliminated from the data set, randomly or in a systematic manner.
Only the excluded objects are predicted by the corresponding model.

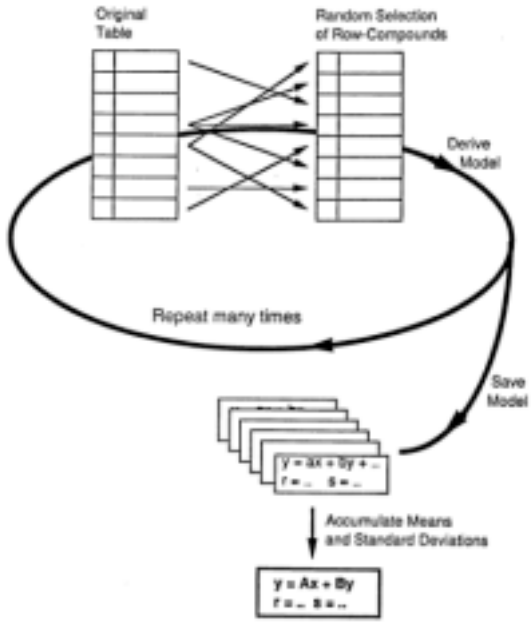
Problems of Crossvalidation
a) redundant data
b) data from a rigorous experimental design



crossvalidation in groups

n = 27, LOO and crossvalidation in 5, 7, 9, 11, 14, 18 and 22 groups.

sdep = $(\Sigma\Delta^2 / n)^{1/2}$



PLS Analysis „Bootstrapping“

In bootstrapping, several PLS runs are performed, in which one or several objects are randomly eliminated from the data set. The variance of the regression coefficients and the statistical parameters, which are derived from the different models, are an indidation for the stability of the model.

Lateral Validation of QSAR Models

Hydrolysis of X-C_6H_4OCO-$CH_2NHCOC_6H_5$ (I) and X-$C_6H_4OCOCH_2$-$NHSO_2CH_3$ (II); ρ Coefficients

Enzyme	Substrate	ρ	pH	Protease
Papain	I	0.57	6	Cysteine
Papain	II	0.55	6	Cysteine
Ficin	I	0.57	6	Cysteine
Ficin	II	0.62	6	Cysteine
Actinidin	I	0.74	6	Cysteine
Bromelain B	I	0.70	6	Cysteine
Bromelain B	II	0.68	6	Cysteine
Bromelain D	I	0.63	6	Cysteine
Subtilisin	I	0.49	7	Serine
Chymotrypsin	I	0.42	6.9	Serine
Trypsin	I	0.71	7	Serine

Hugo Kubinyi, www.kubinyi.de

Hald Example - Stepwise Regression Analysis

Y	X-1	X-2	X-3	X-4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

(N. R. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 1966, pp. 178 ff.)

Variable Selection in Regression Analysis

Forward Selection

- Risk of local minima

Backward Elimination

- Risk of local minima
- Not applicable if number of variables > objects

Stepwise Selection

- Risk of local minima with many variables
- Significance of the models !

Evolutionary and Genetic Algorithms

- Fast and reliable methods for the search of global optima (minima) - reproduction with mutation and crossover, „survival of the fittest“

The Hald Data Set: Forward Selection

Y vs. X-4	r = 0.821; s = 8.96; F = 22.80
Y vs. X-1 and X-4	**r = 0.986; s = 2.73; F = 176.63**

The Hald Data Set: Backward Elimination

Y vs. X-1 to X-4	r = 0.991; s = 2.45; F = 111.48
Y vs. X-1, X-2 and X-4	r = 0.991; s = 2.31; F = 166.83
Y vs. X-1 and X-2	**r = 0.989; s = 2.41; F = 229.50**

The Hald Data Set: Stepwise Selection of Variables

Y vs. X-4	r = 0.821; s = 8.96; F = 22.80
Y vs. X-1 and X-4	r = 0.986; s = 2.73; F = 176.63
Y vs. X-1, X-2 und X-4	r = 0.991; s = 2.31; F = 166.83
Y vs. X-1 and X-2	**r = 0.989; s = 2.41; F = 229.50**

The Hald Data Set, "Best" Model:

Y = 1.468 (±0.27) X-1 + 0.662 (±0.10) X-2 + 52.77 (±5.09)
(n = 13; r = 0.989; s = 2.41; F = 229.50)

A Common Situation

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 Selwood Data Set

n = 31 objects and k = 53 independent variables.

Theoretically, there are:

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

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

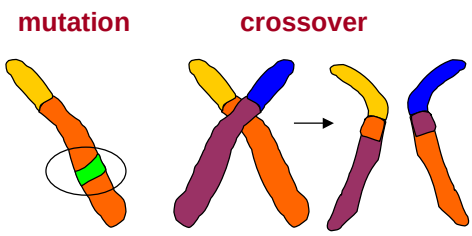
containing one to 29 variables,
selected from 53 X-variables.

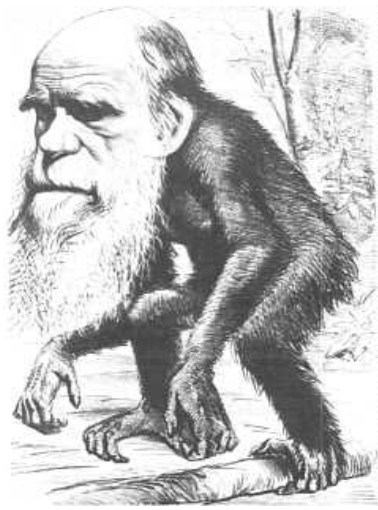
Variables of the Selwood Data Set

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

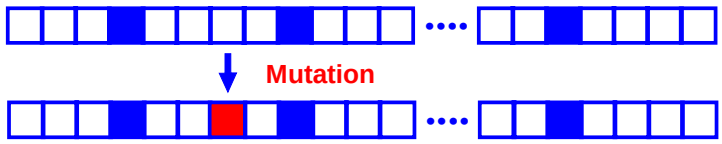
Evolutionary (EA) and Genetic Algorithms (GA)

are powerful optimization strategies which use mutation (EAs) and/or crossover (GAs) to find (near)optimal solutions

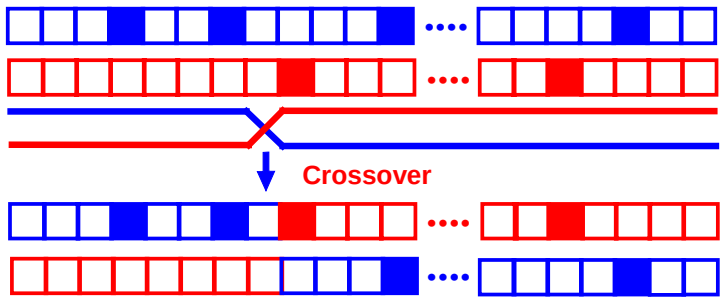




Evolutionary Algorithm

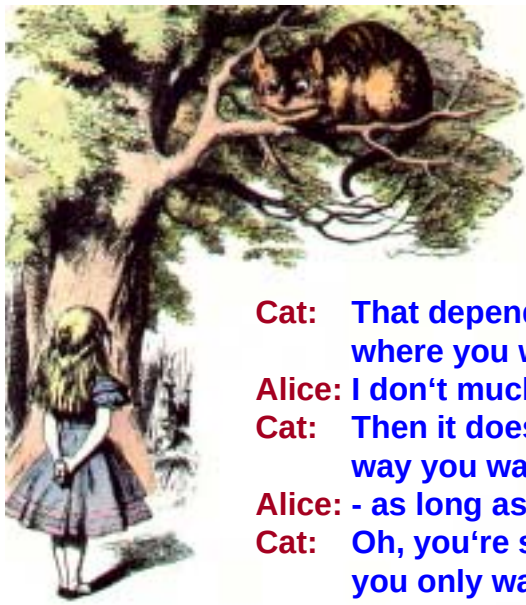


Genetic Algorithm



Hugo Kubinyi, www.kubinyi.de

Evolutionary Algorithms	Genetic Algorithms
start with one random model	start with several to many models (population)
mutation	mutation and crossover
linear path	parallel pathes
very fast (must be repeated)	slow (depends in the size of the population)
result: one or few models	result: several to many models
all variables have same chance	some variables may die out



Lewis Carroll

Alice in Wonderland

Alice: Would you tell me, please, which way I ought to walk from here?

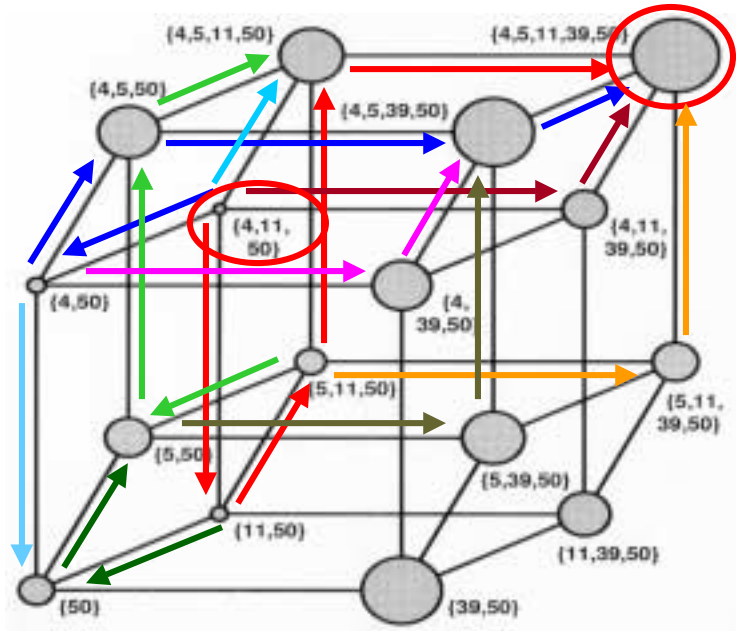
Cat: That depends a good deal on where you want to get to.

Alice: I don't much care where - .

Cat: Then it doesn't matter which way you walk.

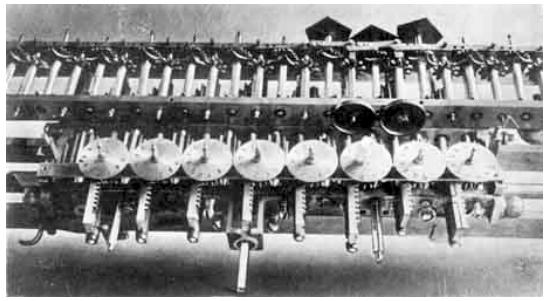
Alice: - as long as I get *somewhere.*

Cat: Oh, you're sure to do that, if you only walk long enough.



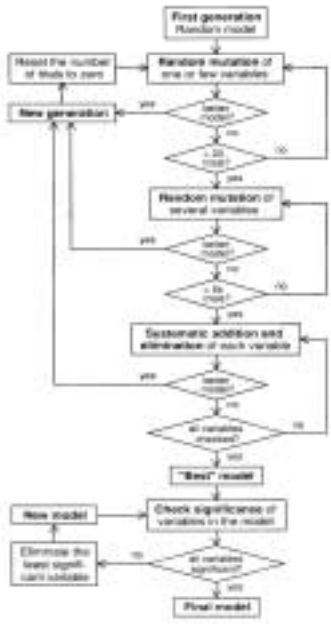
QSAR Models in Hyper-space

„It is unworthy for excellent men to lose hours like slaves in the labour of calculation which could safely be relegated to anyone else if machines were used“.



Gottfried W. Leibniz (1646-1716)

Hugo Kubinyi, www.kubinyi.de



The MUSEUM Program

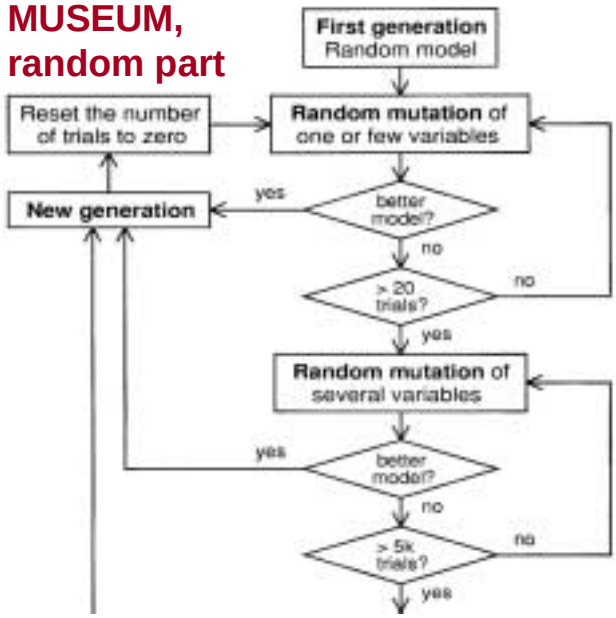
a) random mutation of one or few variables

b) random mutation of several variables

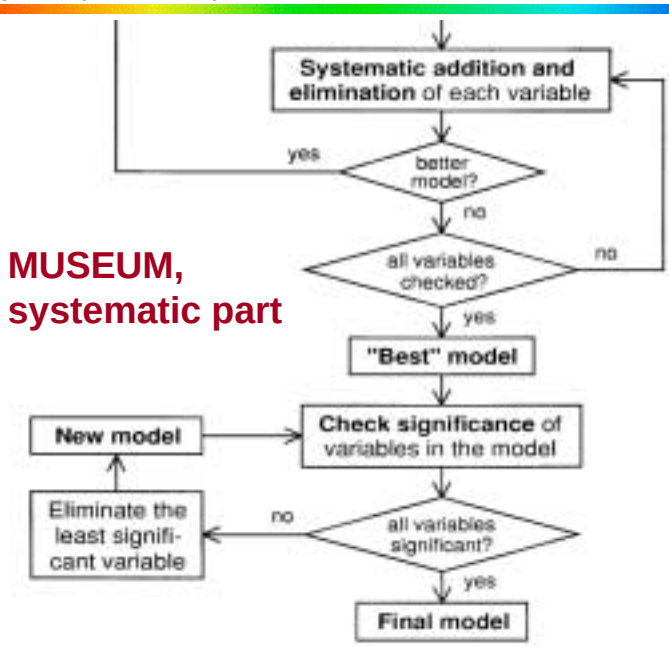
c) systematic variable addition / elimination

d) check the significance of variables and model

MUSEUM, random part



MUSEUM, systematic part



Evolution of a Model - F Criterion

9 generations, 111 models, 6 seconds

Variables	k	s	FIT	F
Start: 4, 17, 36	3	0.667	0.477	6.356
4, 17	2	0.666	0.519	9.084
17	1	0.697	0.411	13.142
5, 17, 36, 50	4	0.470	1.420	16.682
5, 36, 50	3	0.506	1.325	17.661
4, 5, 36, 50	4	0.452	1.576	18.520
4, 5, 11, 36, 50	5	0.415	1.676	18.775
4, 5, 11, 36, 39, 50	6	0.377_5	1.788	19.965
End: 4, 5, 11, 39, 50	5	0.377_1	2.127	23.818

Hugo Kubinyi, www.kubinyi.de

Evolution of a Model - FIT Criterion

8 generations, 129 models, 7 seconds

Variables	k	s	FIT	F
Start: 35, 52	2	0.695	0.412	7.212
52	1	0.683	0.467	14.934
11, 52	2	0.645	0.608	10.647
11, 39, 40, 50, 52	5	0.448	1.375	15.402
11, 39, 50, 52	4	0.449	1.614	18.964
39, 50, 52	3	0.462	1.720	22.935
4, 5, 39, 50	4	0.424	1.873	22.010
End: 4, 5, 11, 39, 50	5	0.377	2.127	23.818

MUSEUM: "Best" Models With Up to 6 Variables

Variables	r	s	F	Q^2	s_{PRESS}
4, 5, 11, 39, 50	0.909	0.377	23.818	0.696	0.499
4, 5, 11, 38, 50	0.909	0.377	23.781	0.696	0.499
38, 50, 52	0.849	0.460	23.267	0.647	0.518
4, 11, 38, 48, 50, 52	0.924	0.354	23.240	0.754	0.458
4, 11, 39, 48, 50, 52	0.924	0.354	23.233	0.751	0.461
4, 11, 38, 47, 50, 52	0.924	0.354	23.191	0.749	0.463
4, 11, 39, 47, 50, 52	0.923	0.355	23.087	0.746	0.466
17, 36, 50	0.848	0.462	23.040	0.644	0.520
39, 50, 52	0.847	0.462	22.935	0.643	0.520
4, 17, 35, 37, 50	0.905	0.385	22.709	0.676	0.515

Hugo Kubinyi, www.kubinyi.de

Comparison of Published "Best" Models

Variables	F	CSA	GFA	FIT-Cr
4, 5, 11, 39, 50	23.818		✓	✓
4, 5, 11, 38, 50	23.781	✓	✓	✓
38, 50, 52	23.267		✓	✓
4, 11, 38, 48, 50, 52	23.240			✓
4, 11, 39, 48, 50, 52	23.233			✓
4, 11, 38, 47, 50, 52	23.191			✓
4, 11, 39, 47, 50, 52	23.087			✓
17, 36, 50	23.040		✓	✓
39, 50, 52	22.935			✓
4, 17, 35, 37, 50	22.709		✓	✓

Is PLS Analysis Superior to Regression?

Vectors	r	s	F	Q^2	s_{PRESS}
1	0.687	0.611	25.93	0.201	0.751
2	0.814	0.497	27.52	-0.172	0.926
3	0.884	0.408	32.03	-0.419	1.038
4	0.909	0.371	30.77	0.198	0.795
5	0.929	0.335	31.58	0.279	0.768
6	0.949	0.292	35.98	0.238	0.806
7	0.953	0.285	32.67	0.251	0.817
8	0.959	0.274	31.32	-0.166	1.042

Hugo Kubinyi, www.kubinyi.de

Variable Selection: Best Three-Variable Models

Variables	r	s	F	Q^2	s_{PRESS}
38, 50, 52	0.849	0.460	23.267	0.647	0.518
17, 36, 50	0.848	0.462	23.040	0.644	0.520
39, 50, 52	0.847	0.462	22.935	0.643	0.520
17, 38, 50	0.838	0.476	21.153	0.604	0.548
17, 39, 50	0.835	0.479	20.708	0.601	0.551
17, 35, 50	0.830	0.486	19.877	0.596	0.553
40, 50, 52	0.830	0.486	19.863	0.598	0.552
4, 5, 11	0.829	0.487	19.827	0.612	0.543
36, 50, 52	0.829	0.487	19.769	0.586	0.560
17, 40, 50	0.827	0.490	19.411	0.589	0.559

PLS Analysis of Reduced Variable Set
(11 variables from from 10 best 3-variable models)

Vectors	r	s	F	Q^2	s_{PRESS}
1	0.729	0.576	32.83	0.284	0.711
2	0.826	0.507	25.91	0.519	0.593
3	0.889	0.399	33.86	0.658	0.509
4	0.902	0.384	28.25	0.665	0.514
5	0.909	0.376	23.91	0.671	0.519
6	0.913	0.377	19.97	0.618	0.571
7	0.918	0.375	17.57	0.532	0.646
8	0.919	0.380	14.99	0.558	0.642

Hugo Kubinyi, www.kubinyi.de

Comparison 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

Ockham‘s Razor - Keep Things Simple !

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

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

Questions

Can we derive „good“ (statistically valid) models ?

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

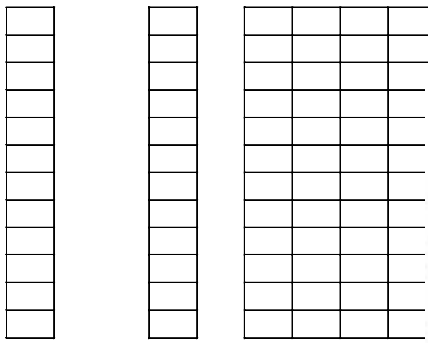
Are these models „better“ than models from scrambled or random data (y, x, y and x) ?

Are 53 X variables too many to select from ?

Can our models predict a test set (r^2_{pred} value) ?

Is there a relationship between internal and external predictivity ?

Y Scrambling - Random Permutation of Y Values



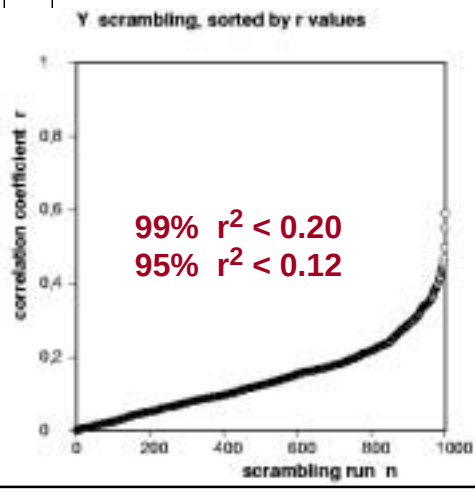
will y vs. y correlations disturb the result ?

X block rem

scrambled y vect

scrambling

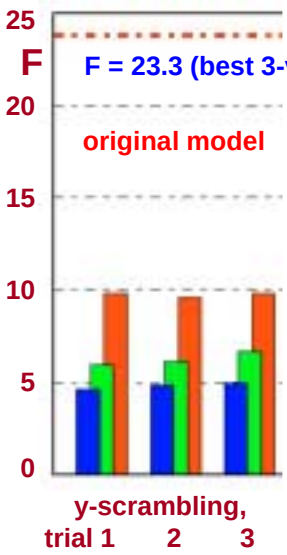
original y vector



99% $r^2 < 0.20$

95% $r^2 < 0.12$

Scrambling and Random Y and X Values



1000 runs for each group, best models.

90%

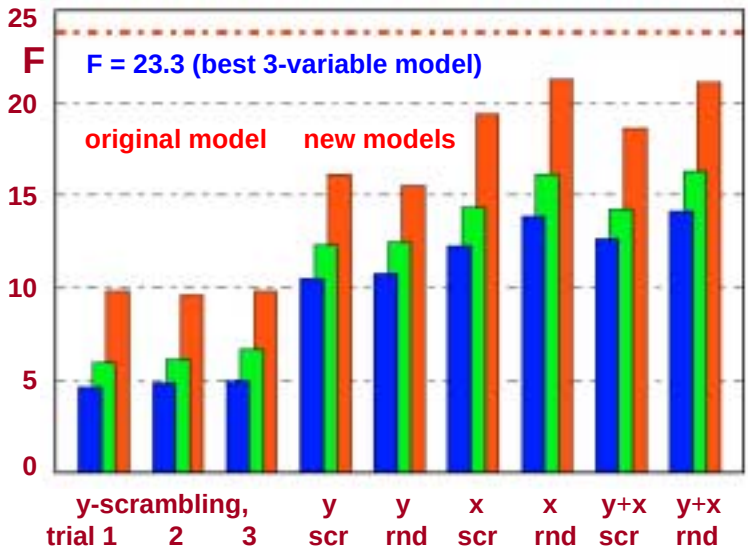
95%

99%

below F value

Hugo Kubinyi, www.kubinyi.de

Scrambling and Random Y and X Values



1000 runs for each group, best models.

90%

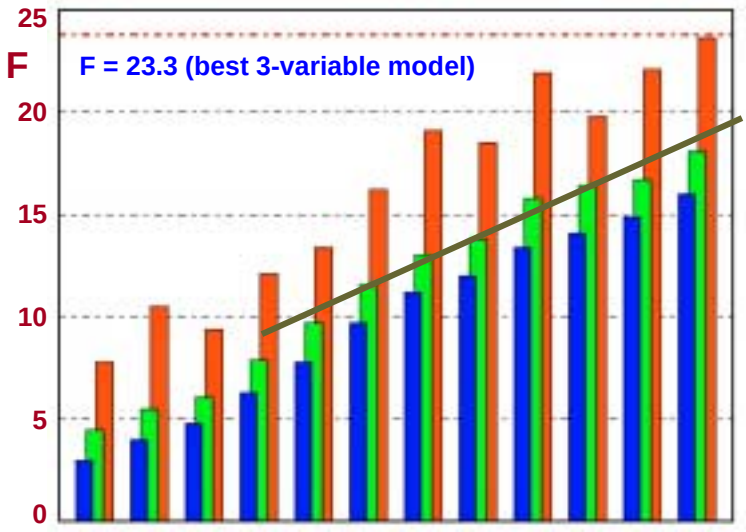
95%

99%

below F value

Hugo Kubinyi, www.kubinyi.de

Models Selected from Random X Variables



1000 runs for each group, best models.

90%

95%

99%

below F value

The Real Situation

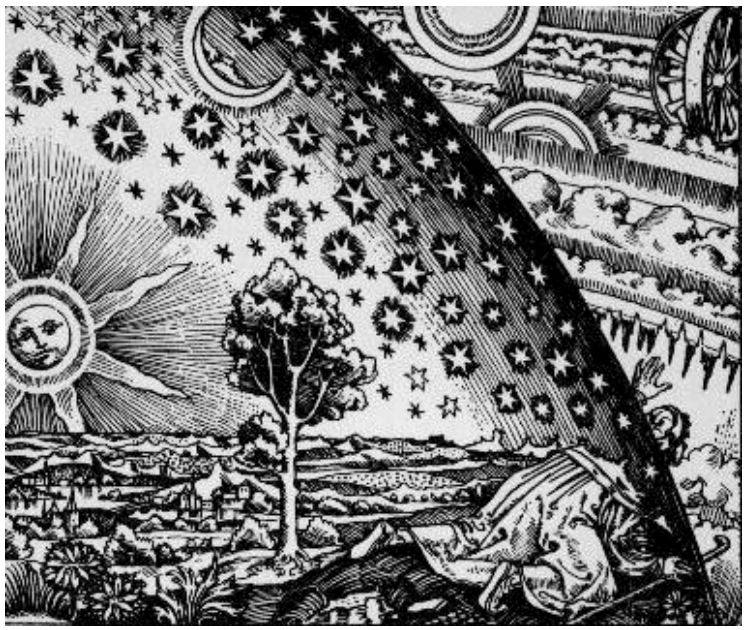
A chemist prepares some 20 compounds.

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 predictivity of the model for a test set of 10 compounds?



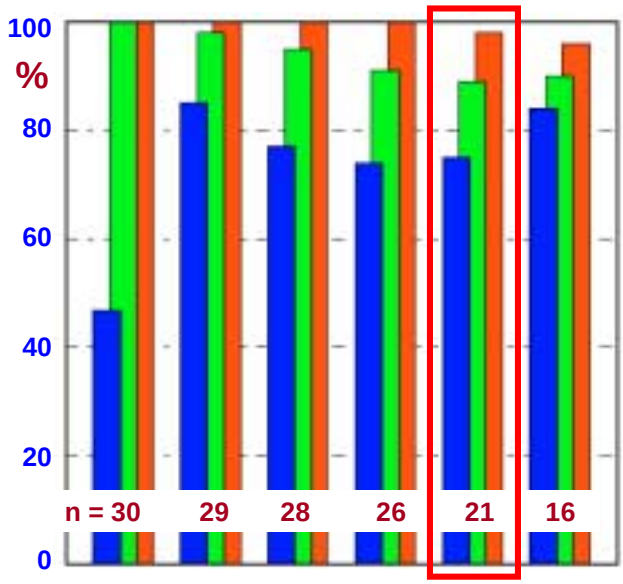
The Problem of Prediction

inside: trivial

outside: wrong

at the edge: 50/50

Training Sets, Internal Predictivity (LOO)



1000 runs for each group.

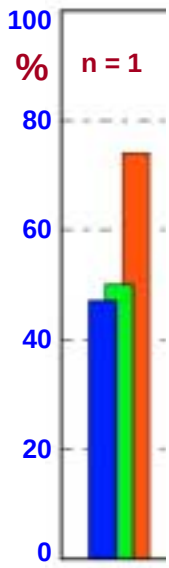
New model selected for every run = 1.9 billion runs

$Q^2 > 0.6$

$Q^2 > 0.5$

$Q^2 > 0$

Test Sets, External Predictivity



1000 runs for each group.

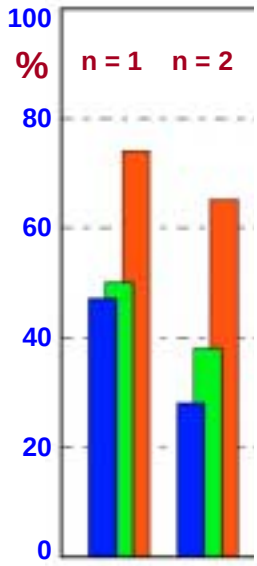
New model selected for every run.

$r^2_{pred} > 0.6$

$r^2_{pred} > 0.5$

$r^2_{pred} > 0$

Test Sets, External Predictivity



1000 runs for each group.

New model selected for every run.

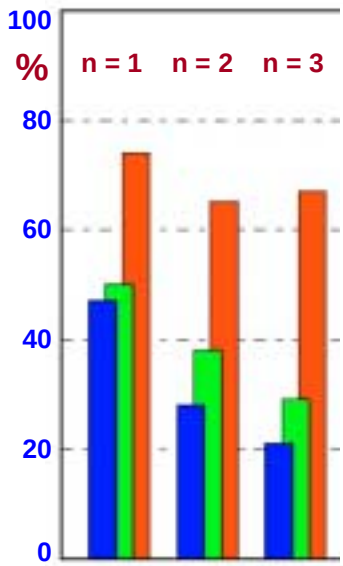
$r^2_{pred} > 0.6$

$r^2_{pred} > 0.5$

$r^2_{pred} > 0$

Hugo Kubinyi, www.kubinyi.de

Test Sets, External Predictivity



1000 runs for each group.

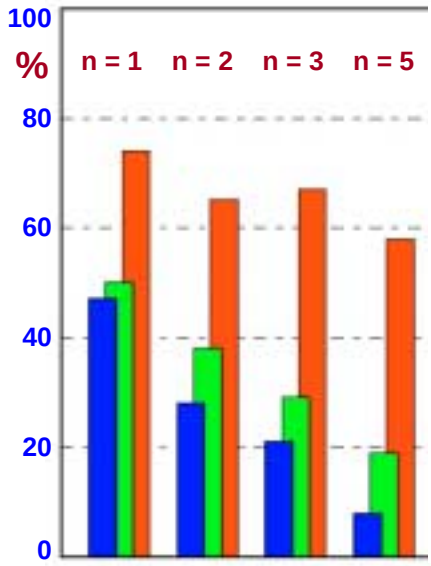
New model selected for every run.

$r^2_{pred} > 0.6$

$r^2_{pred} > 0.5$

$r^2_{pred} > 0$

Test Sets, External Predictivity



1000 runs for each group.

New model selected for every run.

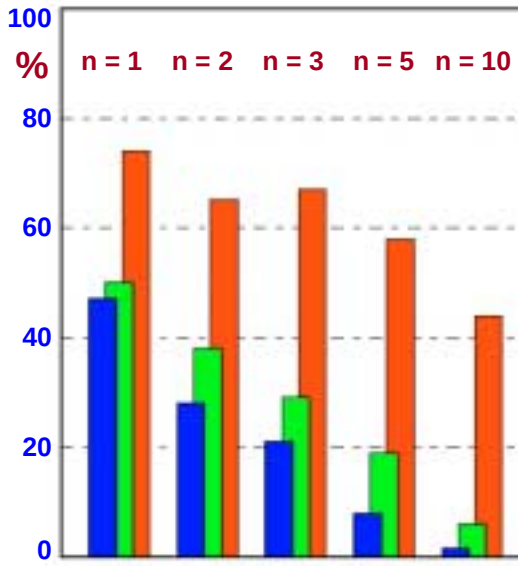
$r^2_{pred} > 0.6$

$r^2_{pred} > 0.5$

$r^2_{pred} > 0$

Hugo Kubinyi, www.kubinyi.de

Test Sets, External Predictivity



1000 runs for each group.

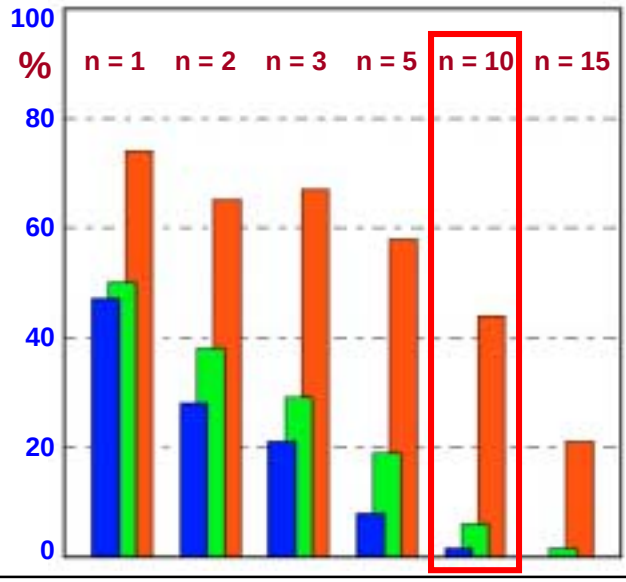
New model selected for every run.

$r^2_{pred} > 0.6$

$r^2_{pred} > 0.5$

$r^2_{pred} > 0$

Test Sets, External Predictivity



1000 runs for each group.

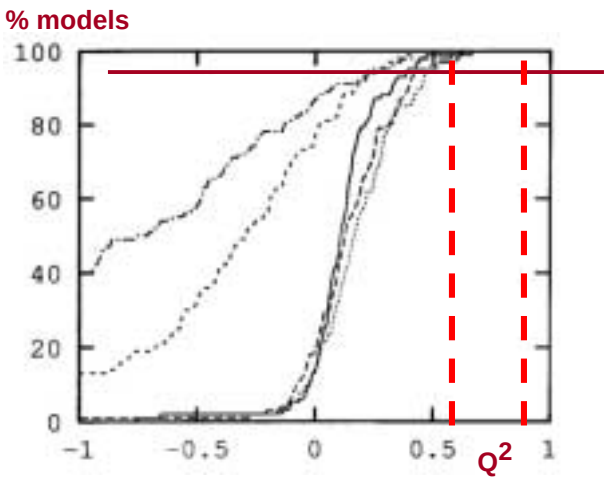
New model selected for every run.

$r^2_{pred} > 0.6$

$r^2_{pred} > 0.5$

$r^2_{pred} > 0$

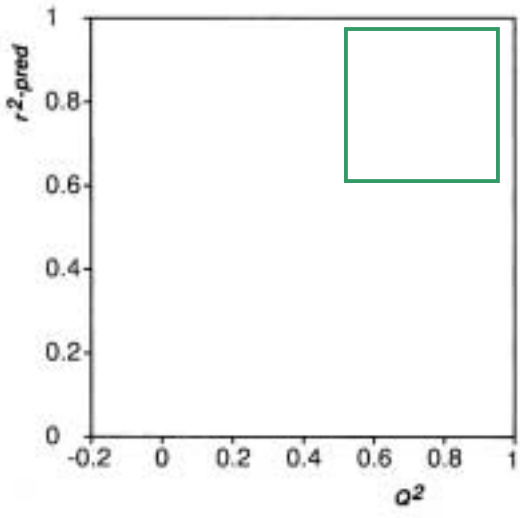
Validation by Random Shuffling of the Biological Data: „Y Scrambling“



95% confidence level for chance correlations

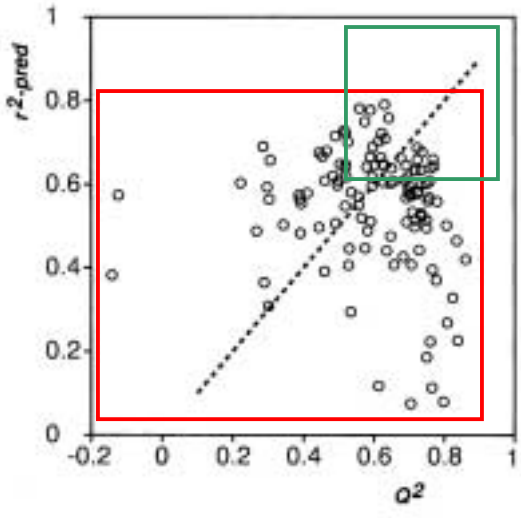
original models: Q^2 = 0.879, 0.608, 0.862, 0.880 and 0.791 (from the left to the right; same data)

External vs. Internal Predictivity



Hugo Kubinyi, www.kubinyi.de

External vs. Internal Predictivity

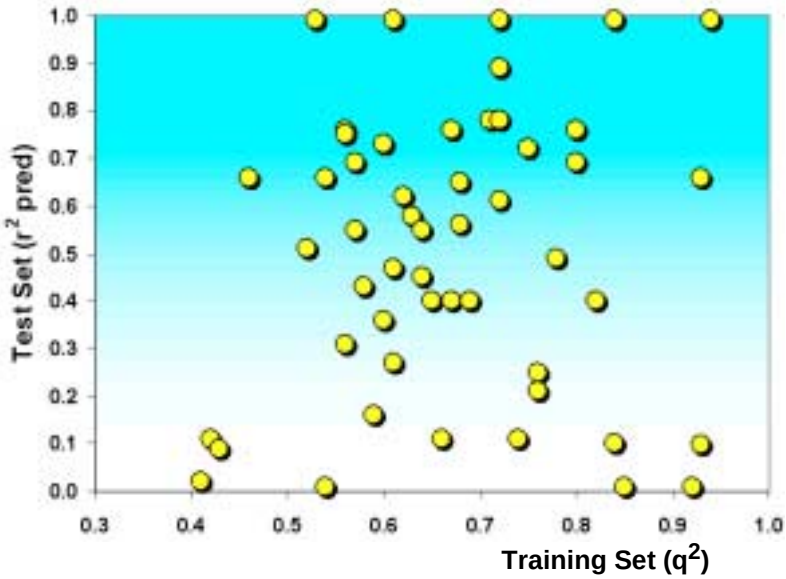


The „Kubinyi Paradox“

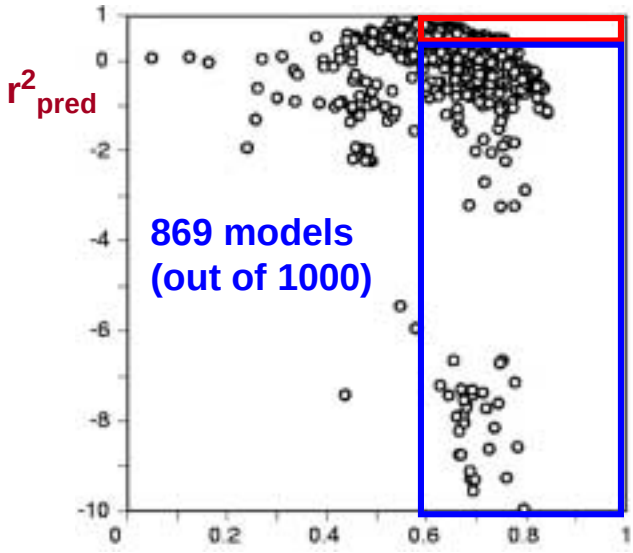
J. H. van Drie, Curr. Pharm. Des. 9, 1649-1664 (2003); J. H. van Drie, in: Computational Medicinal Chemistry for Drug Discovery, P. Bultinck et al., Eds., Marcel Dekker, 2004, pp. 437-460.

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

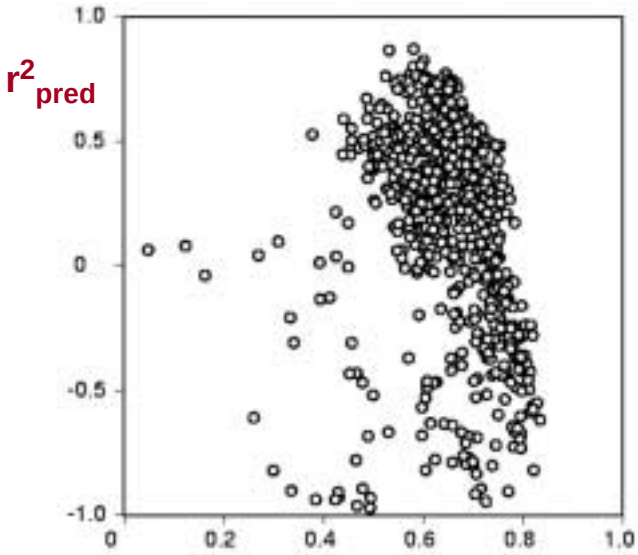


Test set (n = 10),
„best“ models

red area
56 models
blue area
688 models

Q^2

External vs. Internal Predictivity, Selwood Data



The „best fit“ models are not the best ones for prediction !

781 out of 1000 models

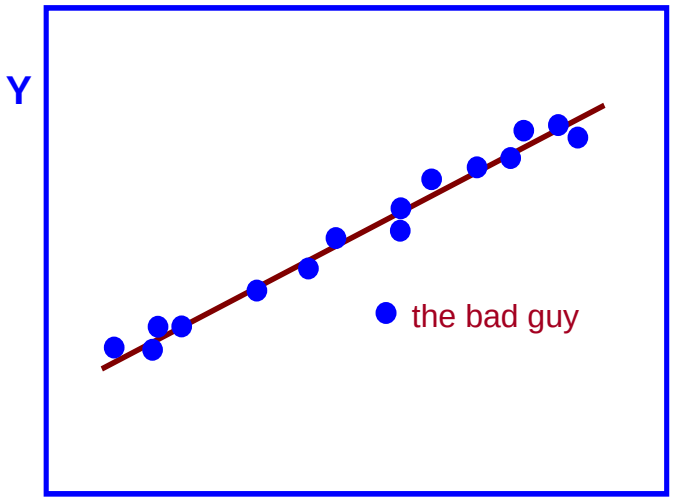
Hugo Kubinyi, www.kubinyi.de

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?

„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

Hugo Kubinyi, www.kubinyi.de

Summary, Conclusions and Recommendations

Cave !

A Model is (only) a Model

La Trahison des Images (The Perfidy of Images)

R. Magritte

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