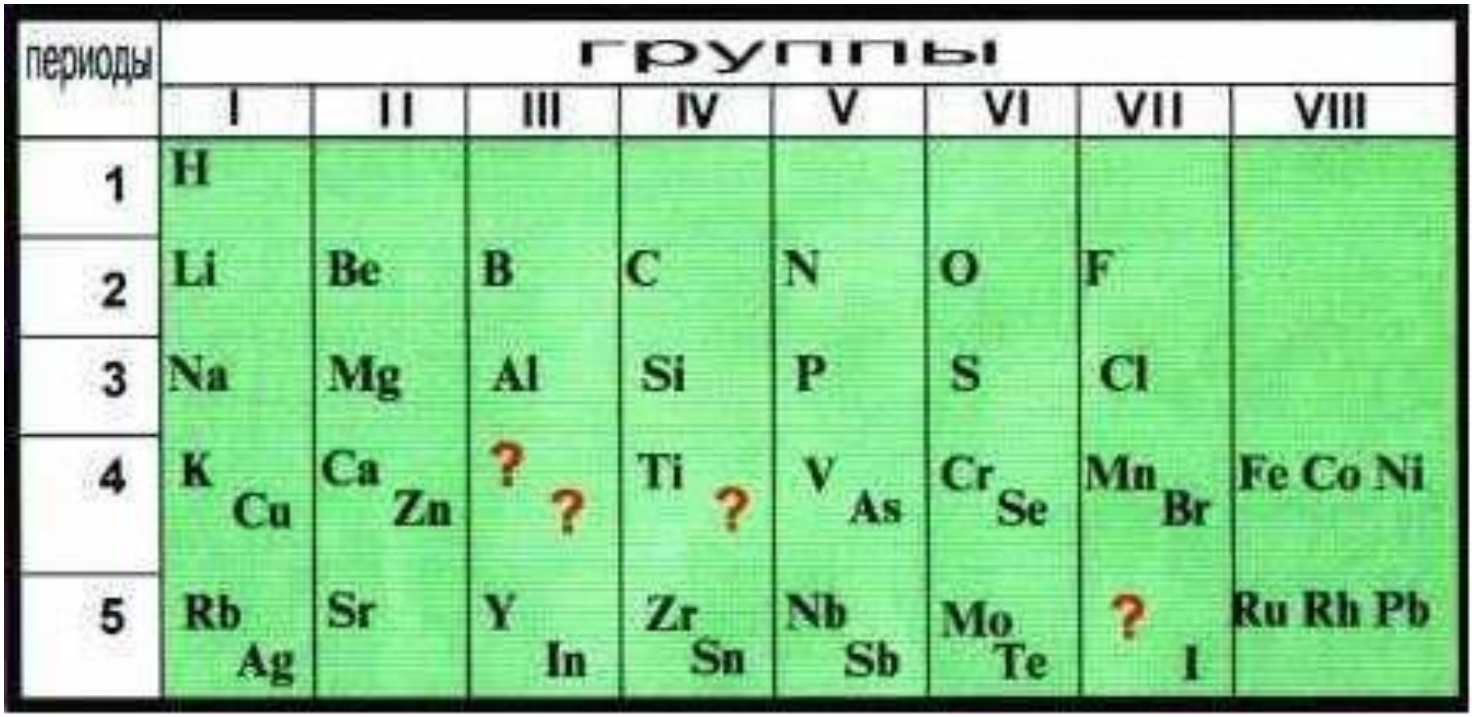
Structure-Property Modeling

Igor I. Baskin

Faculty of Physics

Moscow State University

Mendeleev's Periodic Law and Table

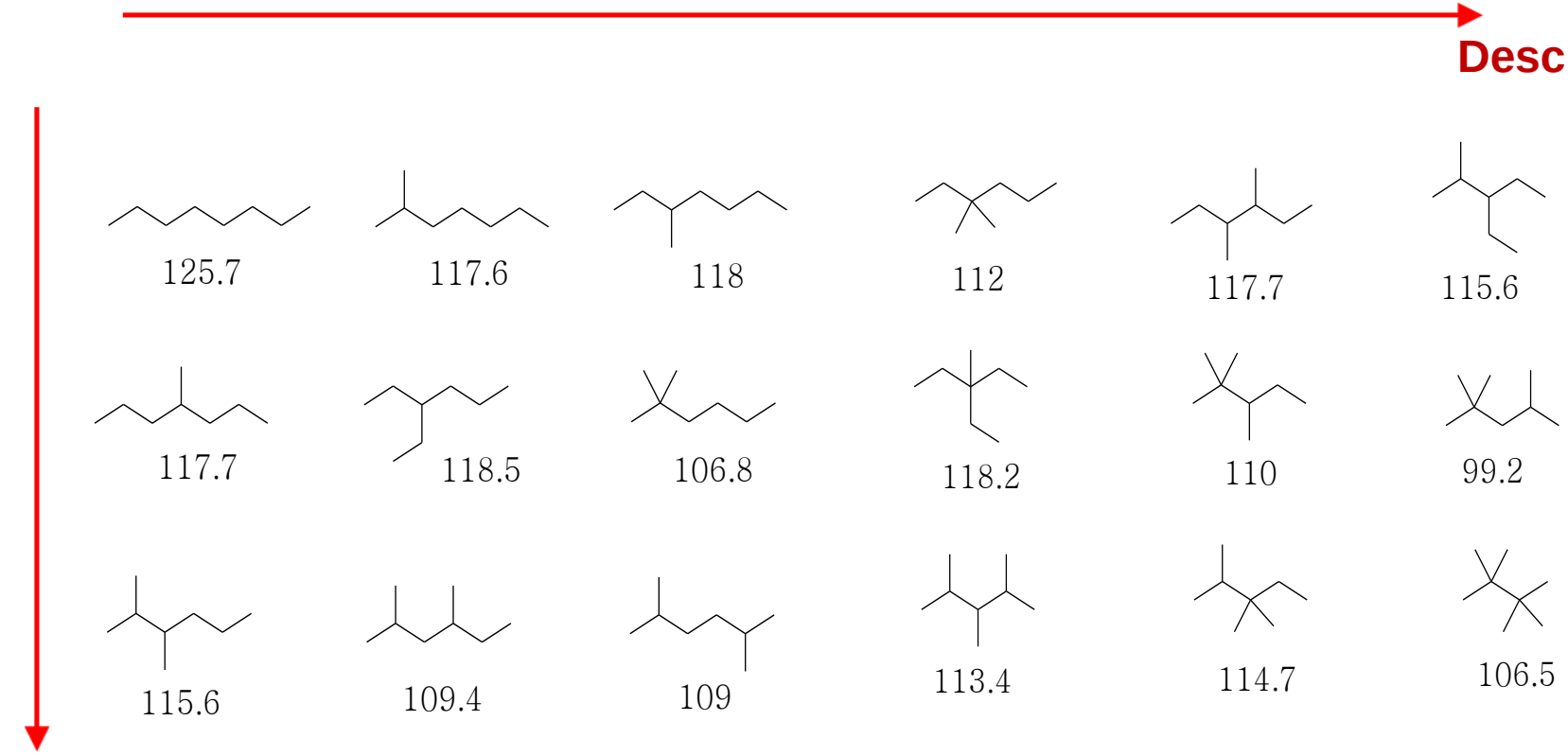


периоды	группы							
	I	II	III	IV	V	VI	VII	VIII
1	H							
2	Li	Be	B	C	N	O	F	
3	Na	Mg	Al	Si	P	S	Cl	
4	K Cu	Ca Zn	? ?	Ti ?	V As	Cr Se	Mn Br	Fe Co Ni
5	Rb Ag	Sr	Y In	Zr Sn	Nb Sb	Mo Te	? I	Ru Rh Pb

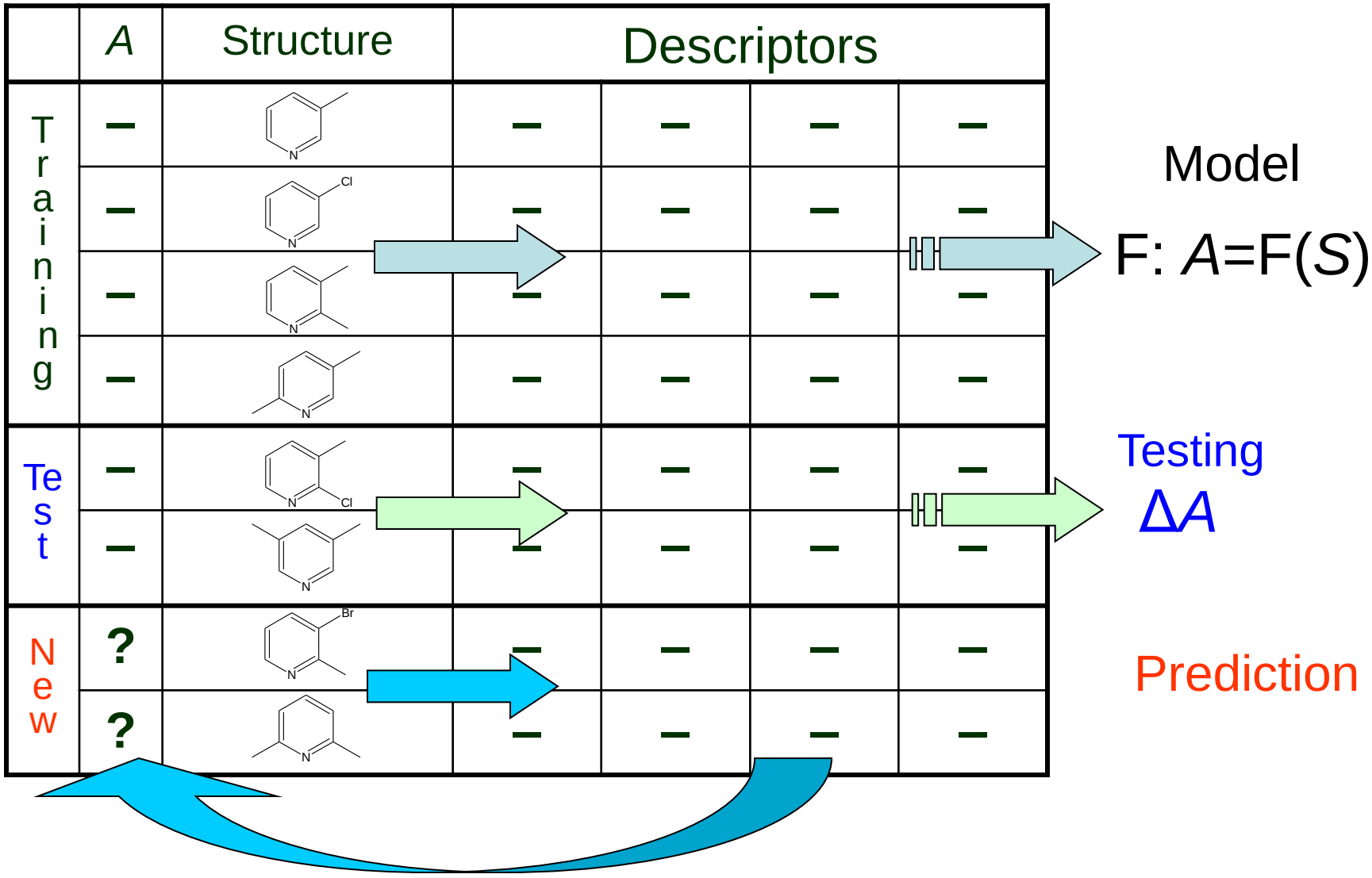
Properties of chemical elements change <u>smoothly</u> (i.e. in the <u>simplest</u> way) along the rows and columns. This enables to predict properties of unknown elements by <u>interpolation</u>.

QSAR/QSPR: Quantitative Structure-Activity or Structure-Property Relationships

Is it possible to apply the same approach to predict properties of new chemical compounds? **Yes**



QSAR/QSPR: Quantitative Structure-Activity or Structure-Property Relationships



Obtaining Models

QSAR/QSPR Models

SAR/QSAR/QSPR model is a functions ***f*** relating the value of some property ***y*** (which can be physicochemical property, biological activity, *etc*) to the values of descriptors $x_1,\ldots,x_M$ (which can represent chemical compounds, reactions, *etc*)

$$y \propto f(x_1,...,x_M)$$

$$y \propto F(c_1,...,c_P; x_1,...,x_M)$$

$$y \propto c_0 + c_1 x_1 + ... + c_M x_M$$

Continuous properties *y* are predicted by regression models

Discrete properties *y* are predicted by classification models

SAR/QSAR/QSPR models are obtained by finding the optimal values of model coefficients using statistical learning (or machine learning) algorithms

Expected and Empirical Risk Functions

Empirical risk function is a prediction error on the training set

$$R_{emp}(c_1,...,c_P) = \frac{1}{2N}\sum_{j=1}^{N}(y^j - F(c_1,...,c_P;x_1^j,...,x_M^N))^2$$ - squared loss function

Expected risk function is an **expectation** of a prediction error on any test set drawn from the same distribution as the training set

$$R(c_1,...,c_P) = E(\frac{1}{2N}\sum_{j=1}^{N}(y^j - F(c_1,...,c_P;x_1^j,...,x_M^N))^2)$$ - squared loss function

So, the expected risk function characterizes the predictive ability of model ***f***

Loss function ***l*(f(*x*),*y*)** is a measure of discrepancy between computed ***f***(***x***) and true property value ***y***

Squared loss function

$$l(f(x),y) = (f(x) - y)^2$$

The Optimal Set of Model Coefficients

The optimal set of model coefficients $c_1,\ldots,c_P$ should minimize the expected risk function and therefore provide a model with the highest predictive ability:

$$R(c_1,\ldots c_P) \rightarrow \min$$

How to perform such minimization?

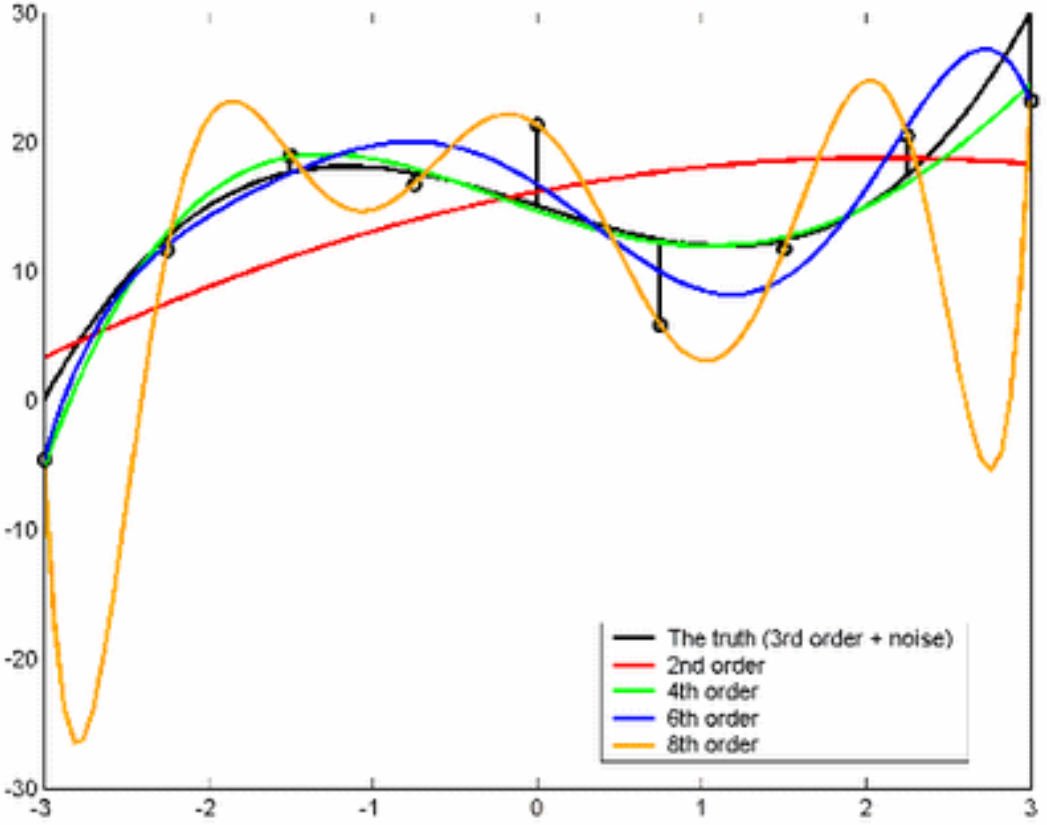
In <u>classical statistics</u> it is assumed that:

$$R(c_1,\ldots c_P) \rightarrow \min \quad \Leftrightarrow \quad R_{emp}(c_1,\ldots c_P) \rightarrow \min$$

<u>Is this correct</u>? Almost correct for big data sets and absolutely not correct for small data sets

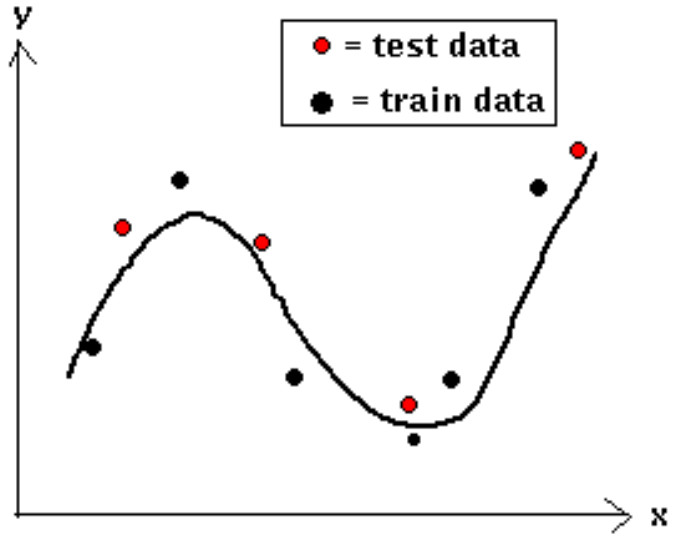
Incorrectness of Empirical Risk Minimization

Data approximation with polynoms of different order $Pol(x, p) = c_0 + \sum_{i=1}^{p} c_i x^i$

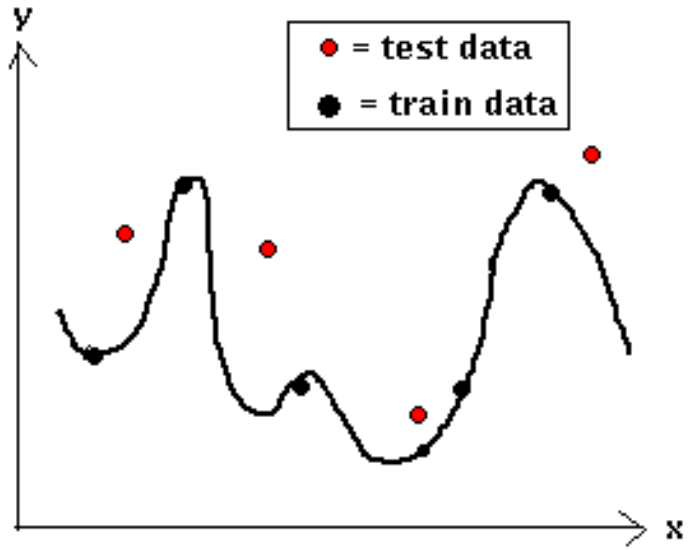


Minimization of the empirical risk function does not guaranties the best predictive performance of the model

Overfitting



Model is not overfitted



Model is overfitted

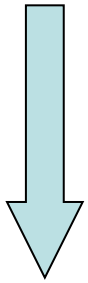
When the model is not overfitted (see at left), it fits the training (shown in green) and the testing (shown in red) data with equal quality. When the model is overfitted (see at right), the model perfectly matches the training data, but cannot predict the testing data.

Viewpoint of Statistical Learning Theory

$$R(c_1,...c_P) \to \min \quad \Leftrightarrow \quad R_{emp}(c_1,...c_P) + \lambda \cdot \Omega(F,N,\delta) \to \min$$

Ω(F,N,δ) – model complexity term

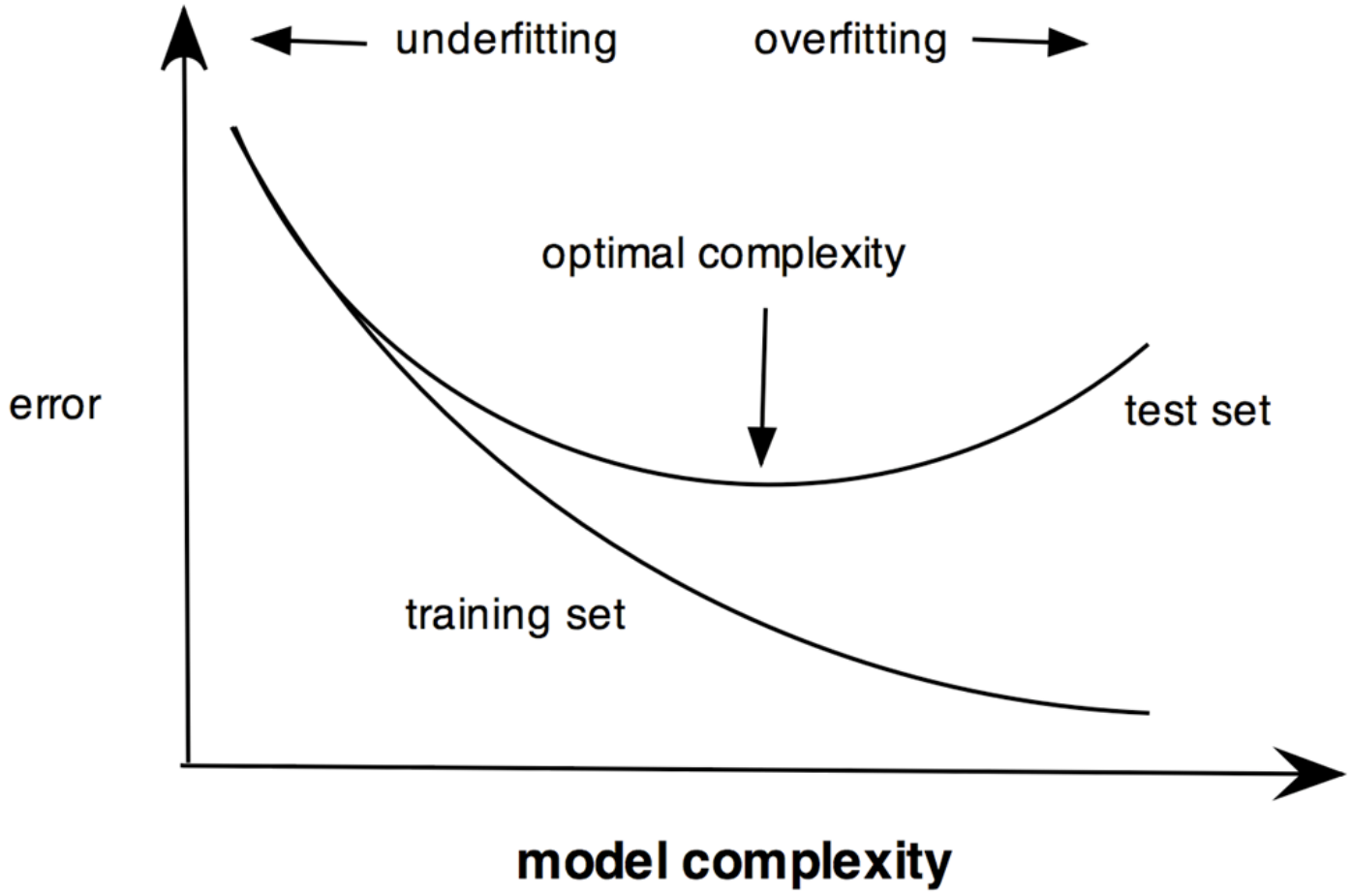
λ – tradeoff parameter



TEST_ERROR = TRAINING_ERROR + MODEL_COMPLEXITY

All QSAR/QSPR models should be both accurate and simple. Since these requirement usually contradict each other, one should always seek a trade-off between them.

Optimal Complexity of Model



Multiple Linear Regression

Y=CX

$$C = (X^T X)^{-1} X^T Y$$

M < N !!!

$$C = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_M \end{pmatrix} \qquad Y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_M^1 \\ 1 & x_1^2 & \cdots & x_M^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & \cdots & x_M^N \end{pmatrix}$$

Regression coefficients

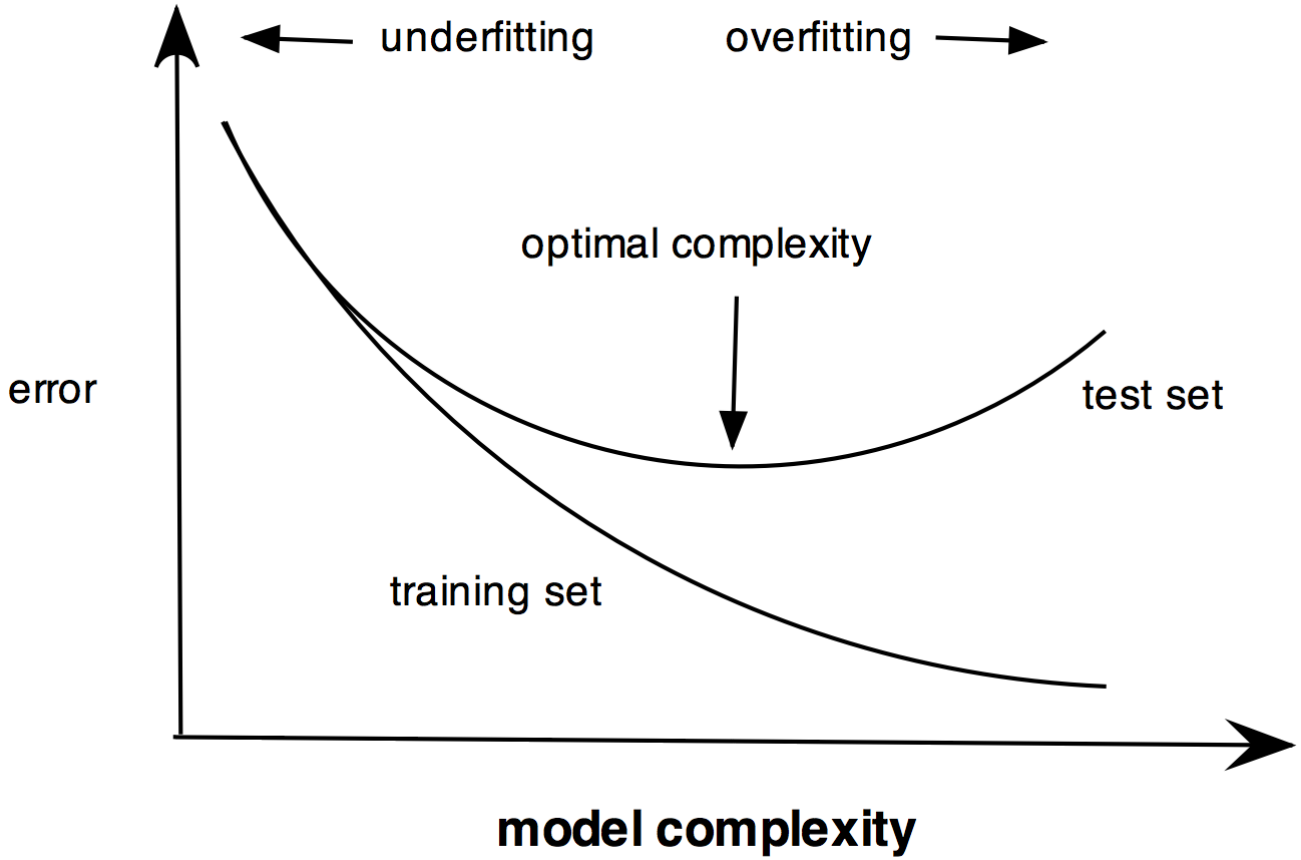
Experimental property values

Descriptor values

$$y = c_0 + c_1 x_1 + \ldots + c_M x_M$$

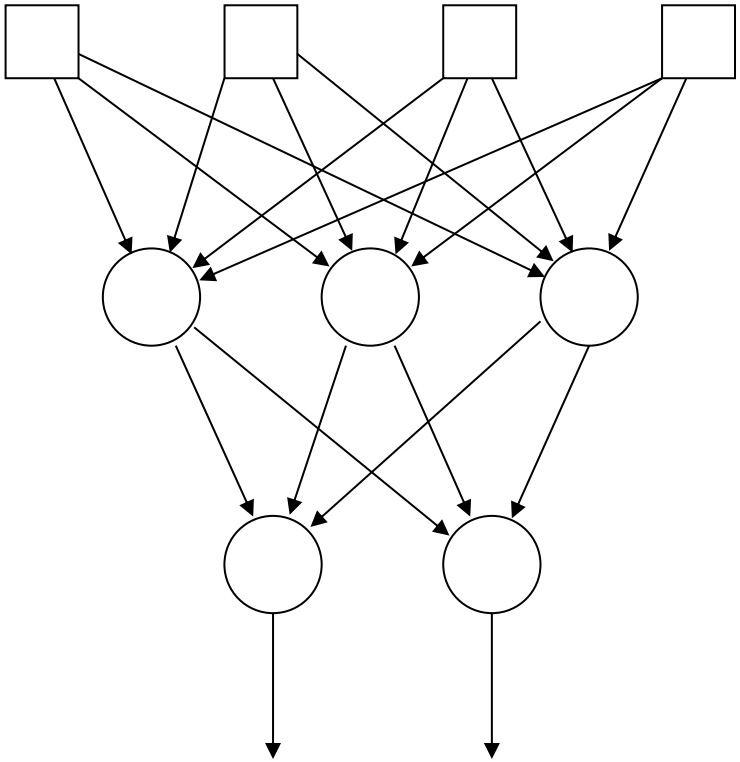
Topliss: N/M > 5 for good models. Mathematicians: eff(N)/eff(M) > 10

Overfitting in Multiple Linear Regression



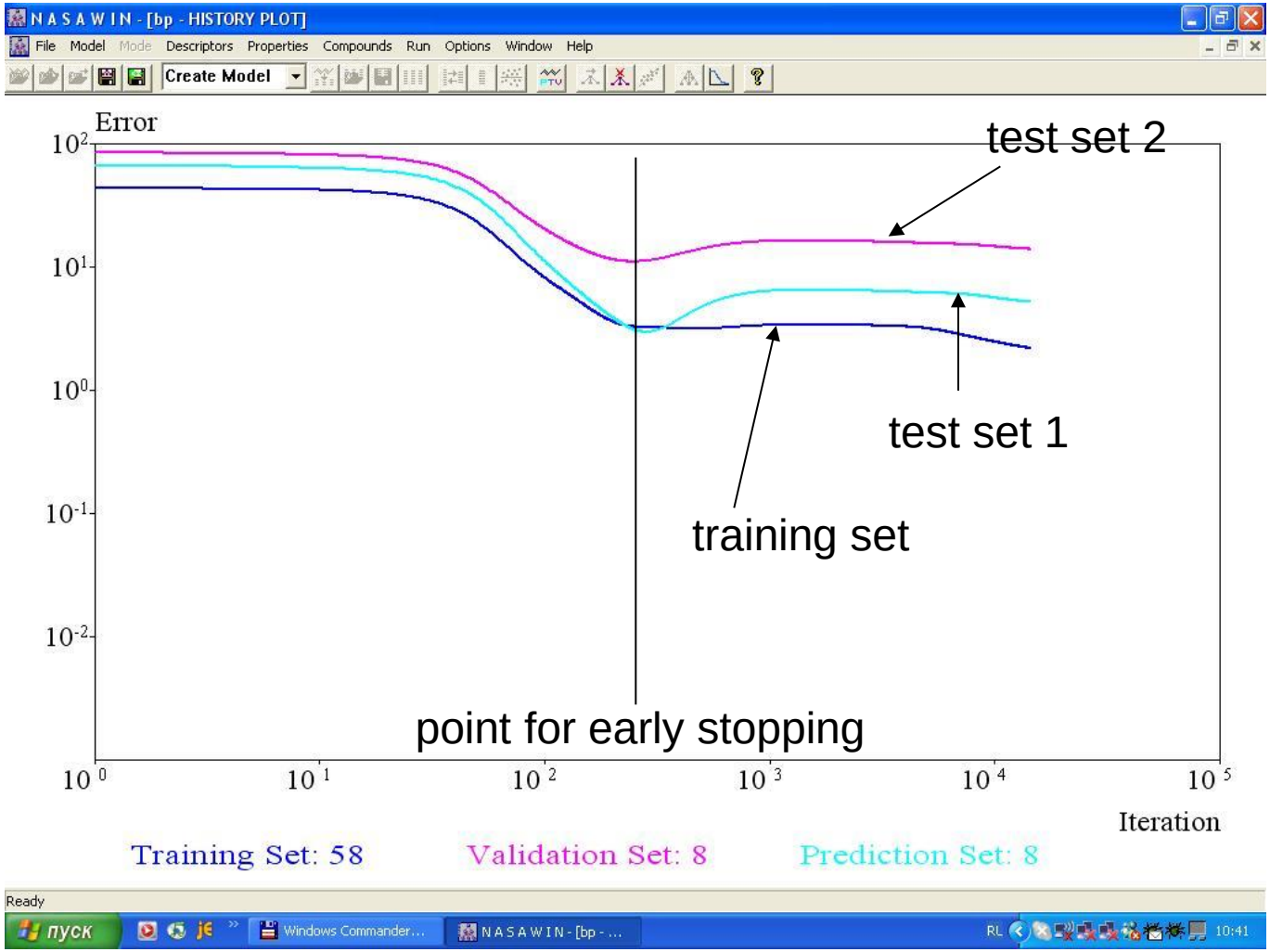
Model complexity ~ the number of descriptors

Neural Networks



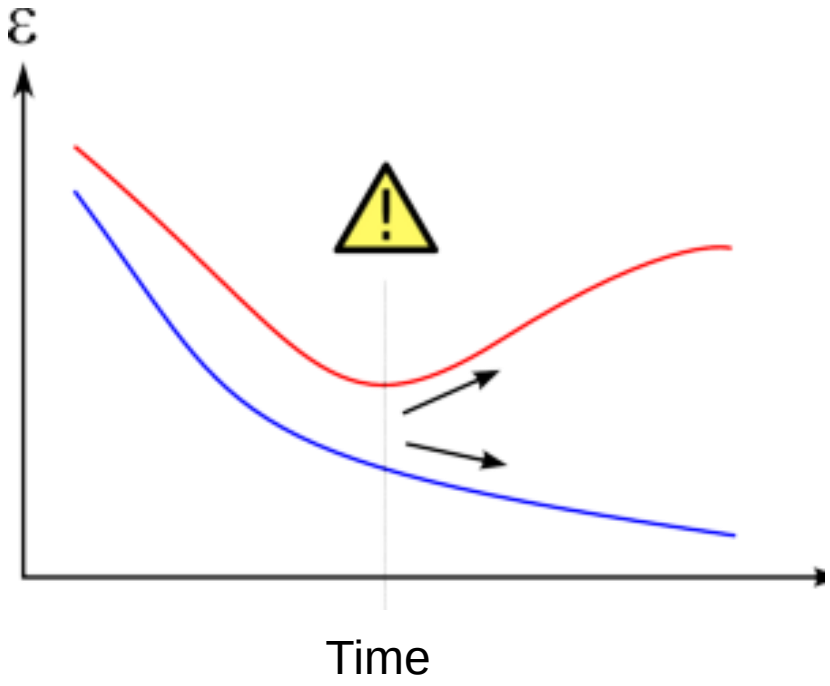
Neurons in the input layer correspond to *descriptors*, neurons in the output layer – to *properties* being predicted, neurons in the hidden layer – to *nonlinear latent variables.* Connection weights between neurons are adjustable parameters.

Overtraining and Early Stopping

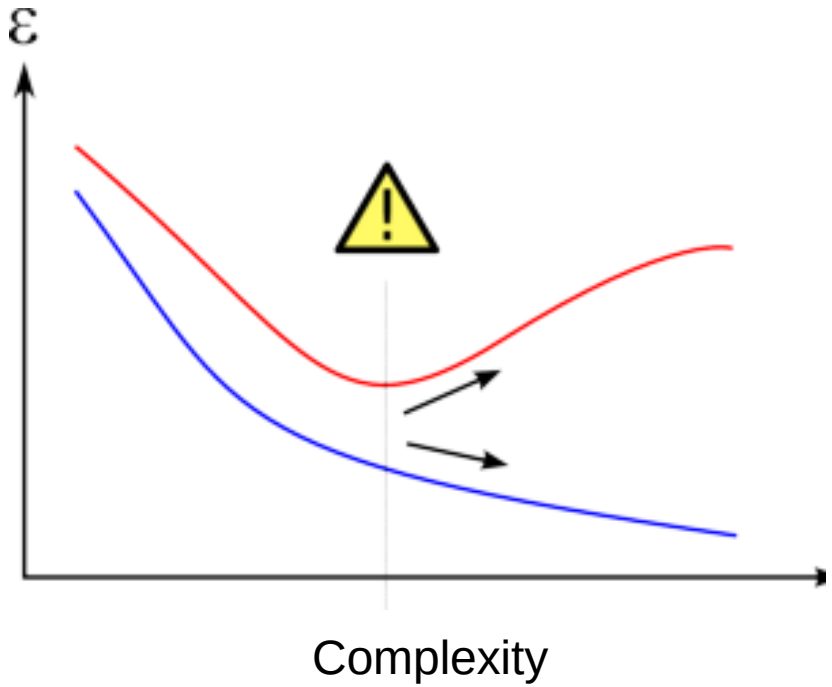


Overtraining *vs* Overfitting

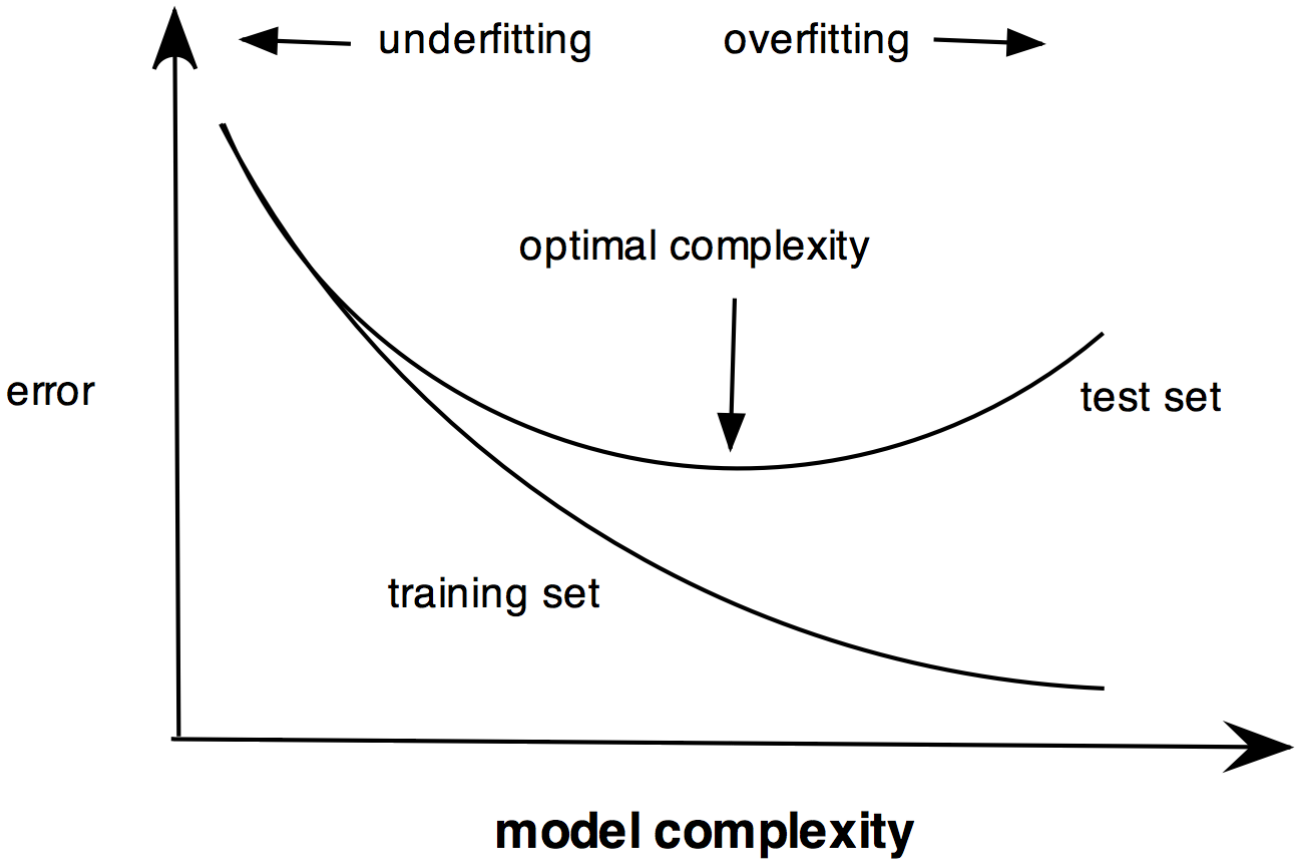
Overtraining



Overfitting

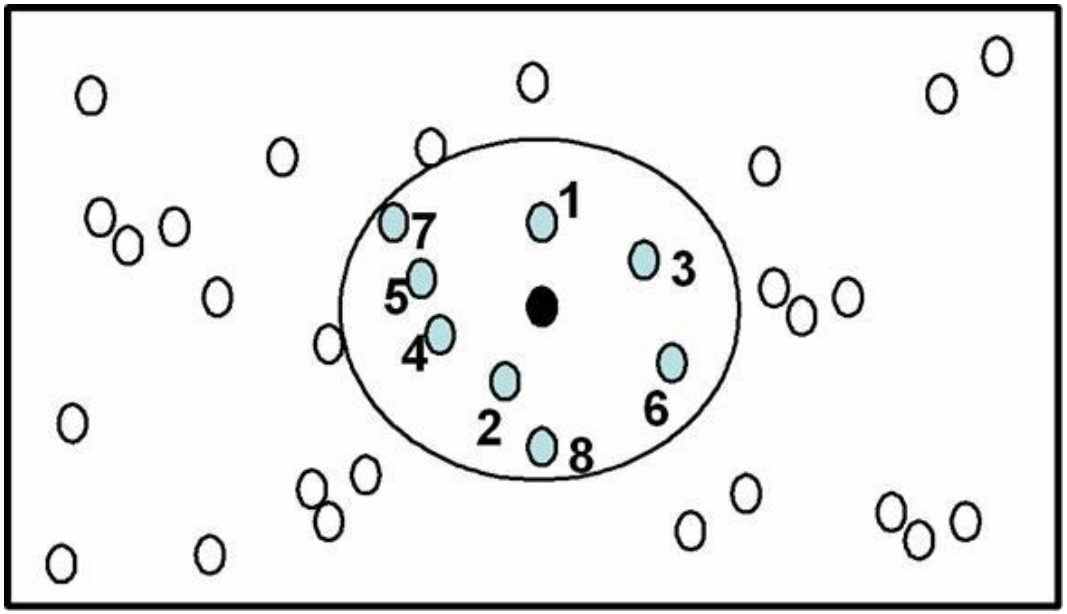


Overfitting in Neural Networks



Model complexity ~ number of iterations

K Nearest Neighbours



$$D_{ij}^{Euclid} = \sqrt{\sum_{k=1}^{M} (x_k^i - x_k^j)^2}$$

$$D_{ij}^{Manhattan} = \sum_{k=1}^{M} \left| x_k^i - x_k^j \right|$$

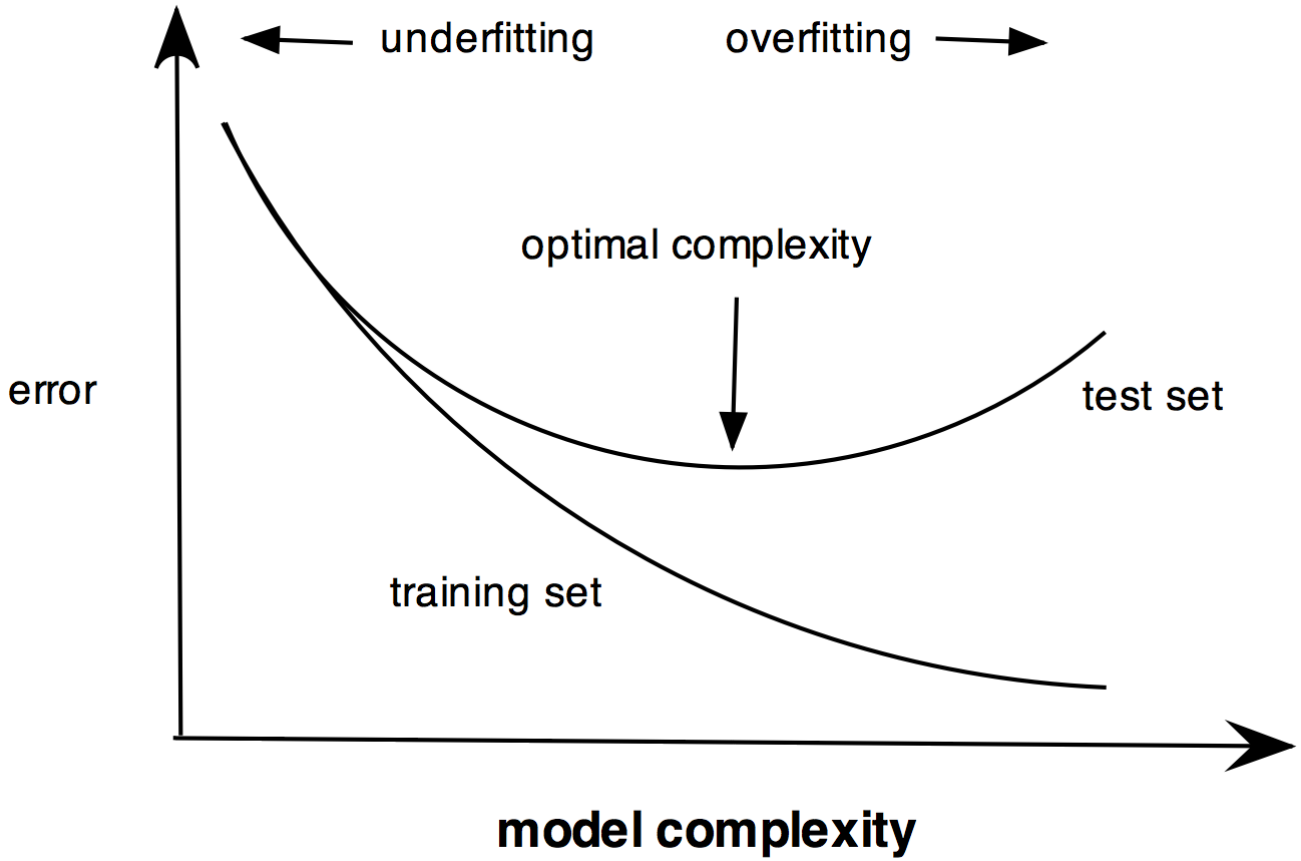
Non-weighted

$$y_i^{pred} = \frac{1}{\mathrm{k}} \sum_{j \in k-neighbours} y_j$$

Weighted

$$y_i^{pred} = \frac{1}{\sum_{j \in k-neighbours} \frac{1}{D_{ij}}} \cdot \frac{1}{D_{ij}} \sum_{j \in k-neighbours} y_j$$

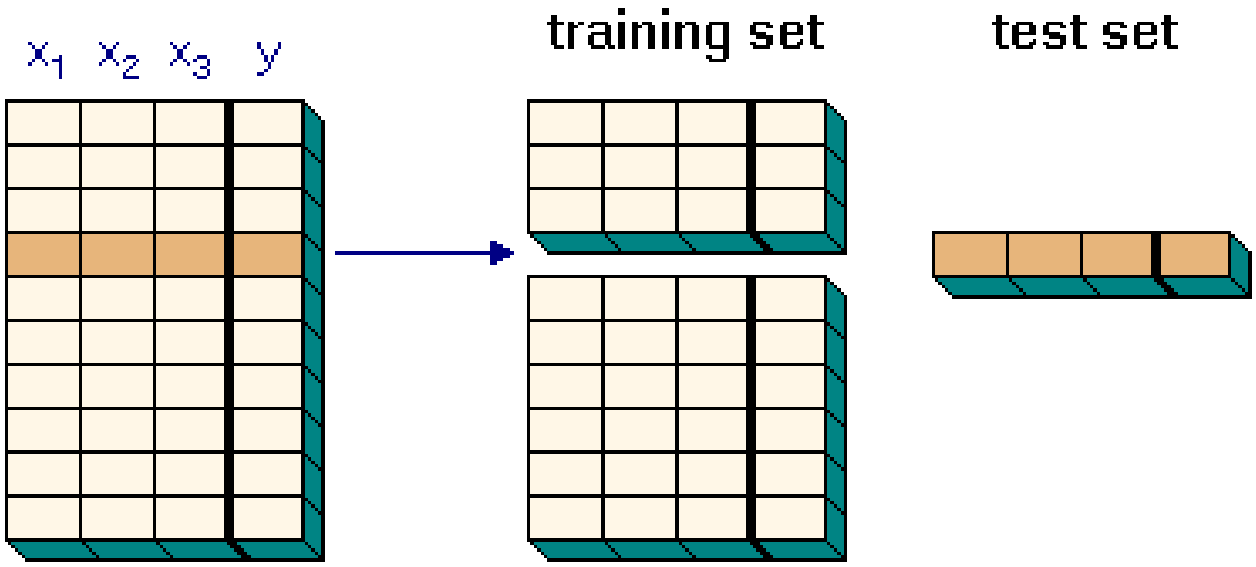
Overfitting for the *k* Nearest Neighbours



Model complexity ~ n/k

Validating Models

Validation of Models



The full data set is split into two mutually exclusive sets, a larger one (the 'training' set) and a smaller one (the 'test' set). The larger data set is used to obtain the model, while the smaller data set is used to validate the model.

Larson, S. C. (1931). *J. Edic. Psychol*., 22:45–55.

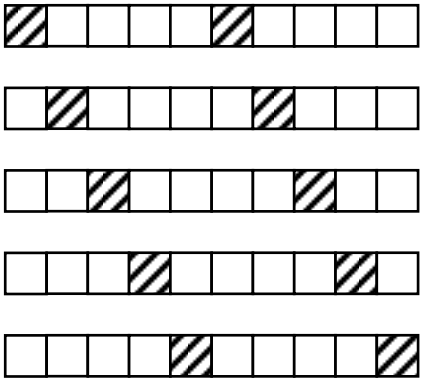
(Cross-)Validation

- Training set

- Test set

Hold-out (cross-)validation

(Devroye, L. and Wagner, T. J. (1979) *IEEE Transaction in Information Theory*, 25(5):601–604.)



- Training set

- Test set

V-Fold cross-validation

(Geisser, S. (1975) *J. Amer. Statist. Assoc.*, 70:320–328)

Leave-one-out (LOO) cross-validation

(Stone, M. (1974) *J. Roy. Statist. Soc. Ser. B,* 36:111–147.)

Arlot S., Celisse A. (2010) *Statistics Survers*, 4:40-79

Statistical Parameters for (Cross-)Validation

$$\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y^n$$

- property arithmetic mean

$$SS = \sum_{n=1}^{N} (y^n - \bar{y})^2$$

- variance; is the sum of squared deviations of experimental values

$$PRESS = \sum_{n=1}^{N} (\hat{y}^n - y^n)^2$$

- the predictive sum of squares of the differences between the experimental and computed property values

$$PRMSE = \sqrt{\frac{PRESS}{N}}$$

- predictive root-mean-square error

$$R^2 = \frac{SS - PRESS}{SS'}$$

- predictive determination coefficient for hold-out (cross-)validation

$$Q^2 = \frac{SS - PRESS}{SS}$$

- predictive determination coefficient for any other type of cross-validation

Methods of (Cross-)Validation:
Advantages and Disadvantages

Hold-out (cross-)validation
Computationally the most efficient
Estimations are largely biased and strongly depend on splitting.
The variance of estimations is the largest.

V-Fold cross-validation – trade-off between the hold-out validation and the leave-one-out cross-validation

Leave-one-out cross-validation
Estimations are almost unbiased and do not depend on splitting
The variance of estimations is the smallest
Requires very intensive computations

Limitations and Misuse of Cross-Validation

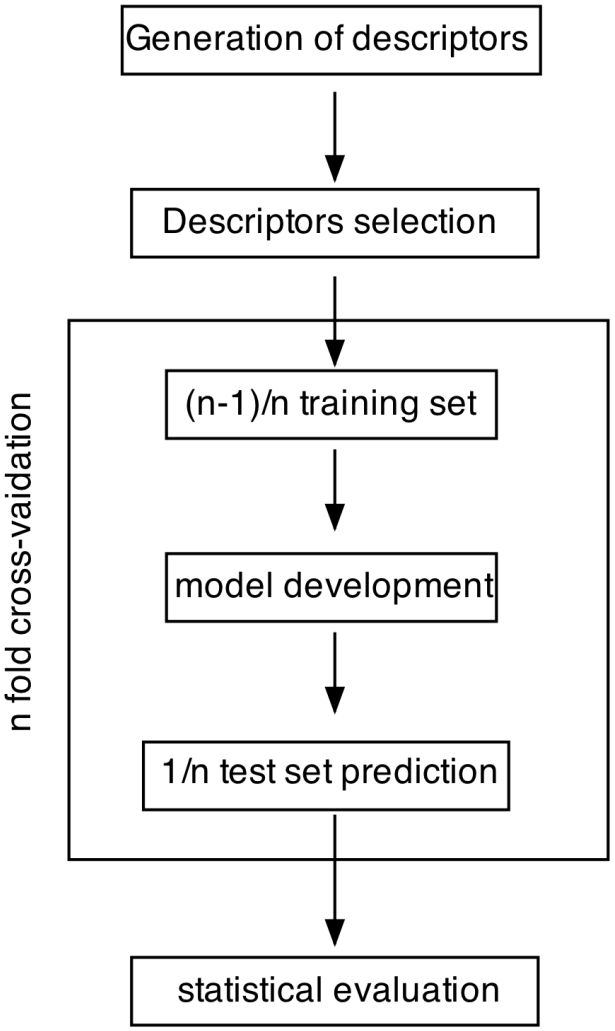
Cross-validation yields meaningful results only if the test set and training set are independently drawn from the same population.

One should avoid specially designed "optimal" splits into the training and test sets because they might become mutually dependent.

Training and test sets should not contain exactly the same or very close compounds because this makes them mutually dependent

Cross-validation should not be used for descriptor or model selection using the entire data set. Such selection should be carried out on every training set using an inner cross-validation loop.

Internal Cross-Validation

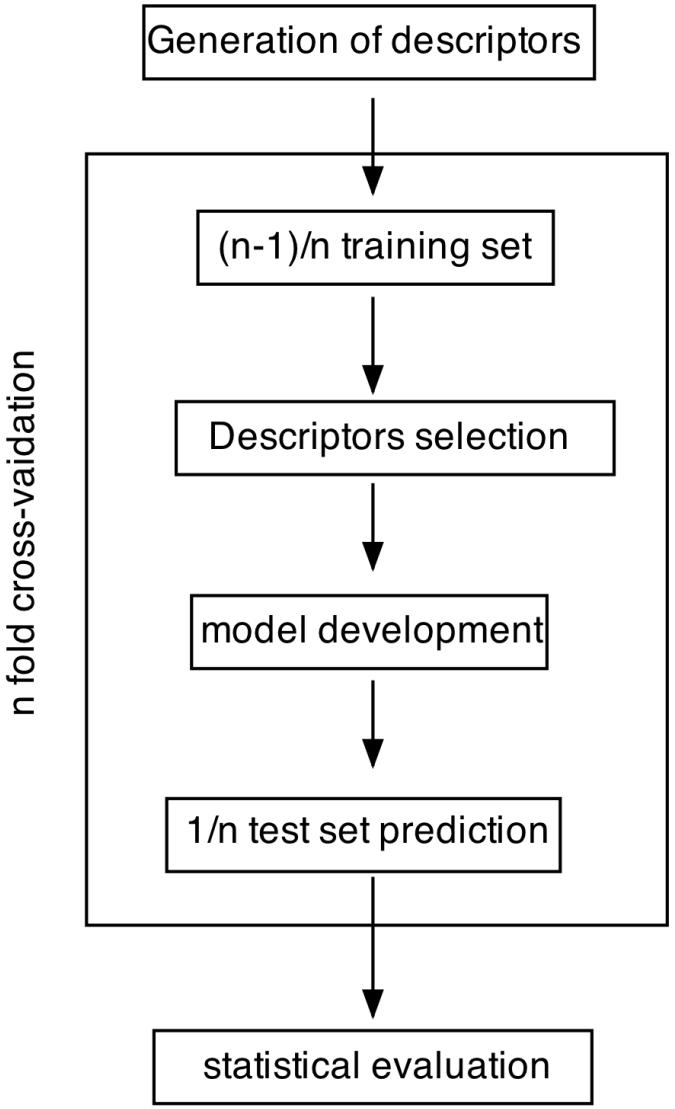


Internal cross-validation is performed after supervised descriptor or model selection in the inner (internal) cycle

Internal cross-validation can be used for selecting descriptors or models

Internal cross-validation cannot be used for validating models because predictions are not completely “blind”

External Cross-Validation

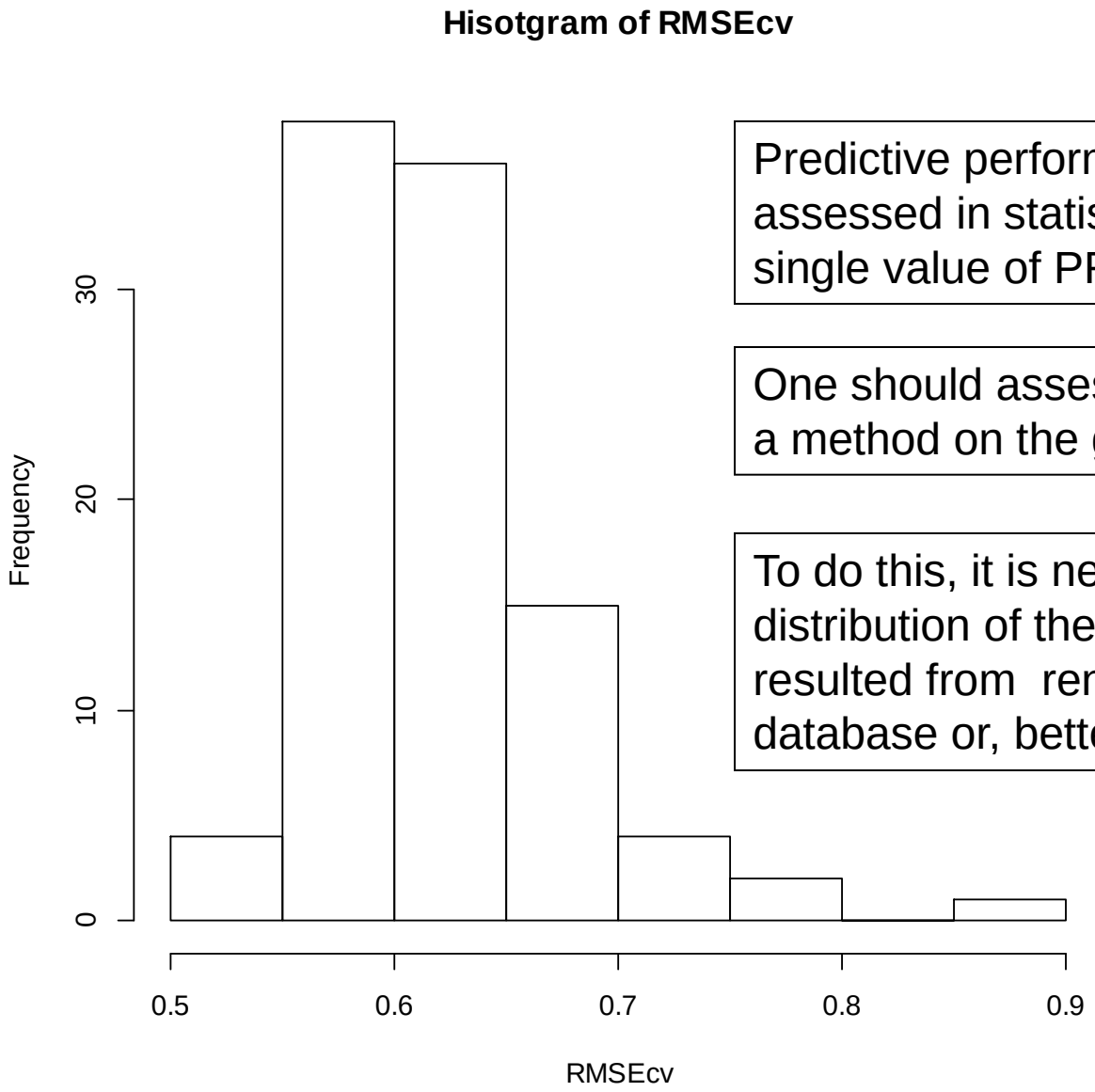


External cross-validation is performed before descriptor or model selection in the outer (external) cycle

External cross-validation can be used for validating models because predictions in it are completely “blind”

External cross-validation is the only correct way to assess the predictive performance of QSAR/QSPR models

Assessing Predictive Performance of Models



Predictive performance of a model cannot be assessed in statistically sound way from a single value of PRMSE or Q2.

One should assess predictive performance of a method on the given dataset

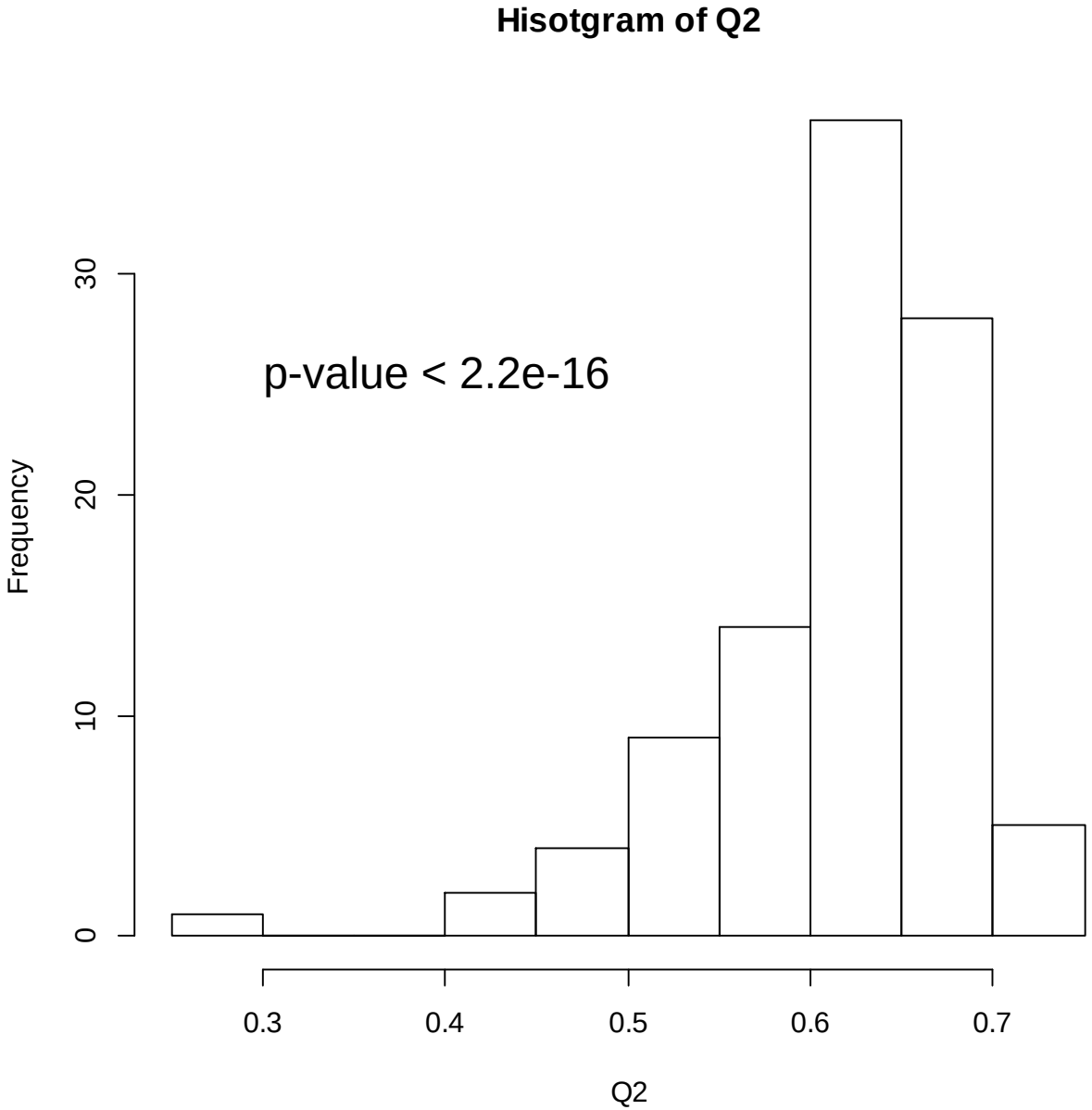
To do this, it is necessary to study the distribution of the values of PRMSE or Q2 resulted from renumbering of compounds in database or, better, from bootstrap.

3D-QSAR models, Method of Continuous Molecular Fields, 5-fold CV, pIC50 for protease inhibition

Assessing Statistical Significance of Models

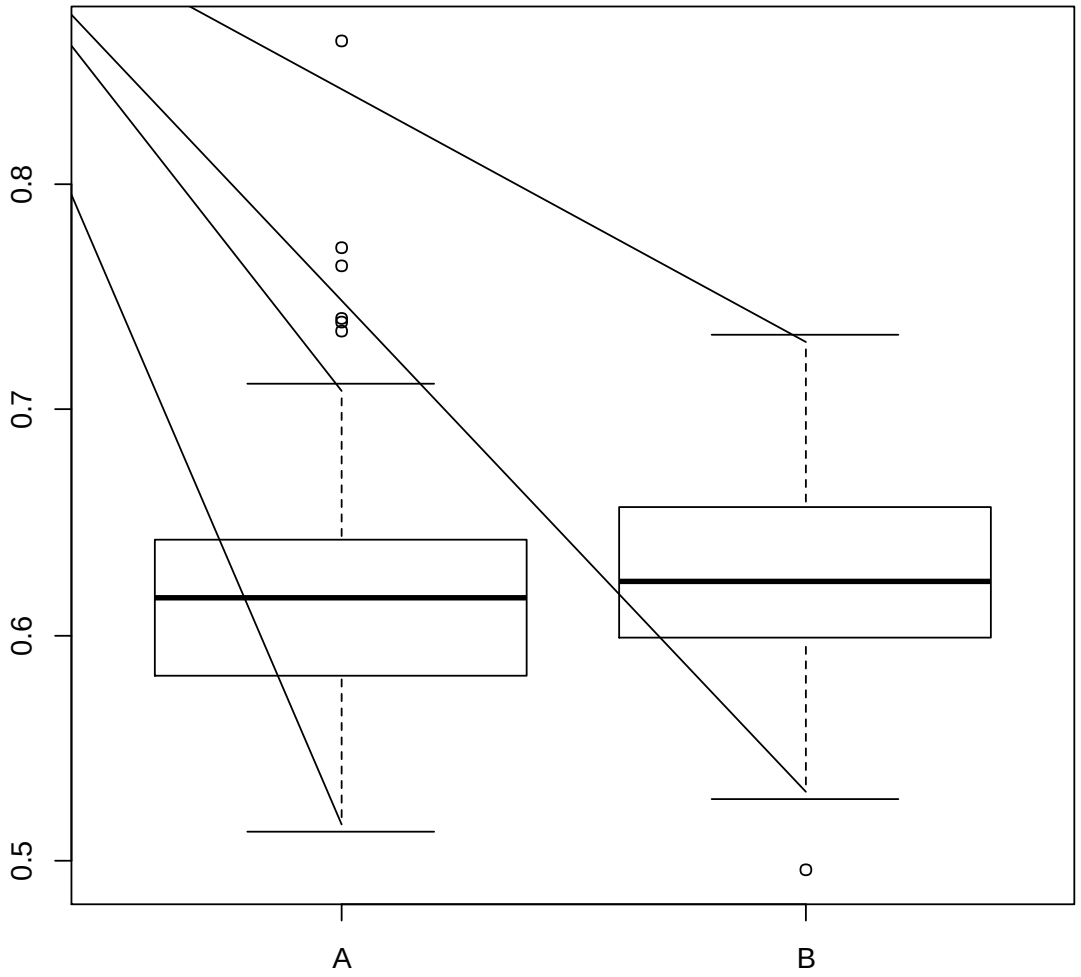
In order to prove the statistical significance of a regression model, it is necessary to test the statistical hypothesis that the mean of the distribution of Q2 is greater than zero.

One can use one sample z-test or t-test for this purpose



3D-QSAR models, Method of Continuous Molecular Fields, 5-fold CV, pIC50 for protease inhibition

Comparing Predictive Performance of Models



Two-Sample t-Test

t = -0.8905
df = 187.926
p = 0.3744

Difference between mean values of RMSEcv is not significant

3D-QSAR models, Method of Continuous Molecular Fields, 5-fold CV, pIC50 for protease inhibition

Appling Models

Applicability Domain

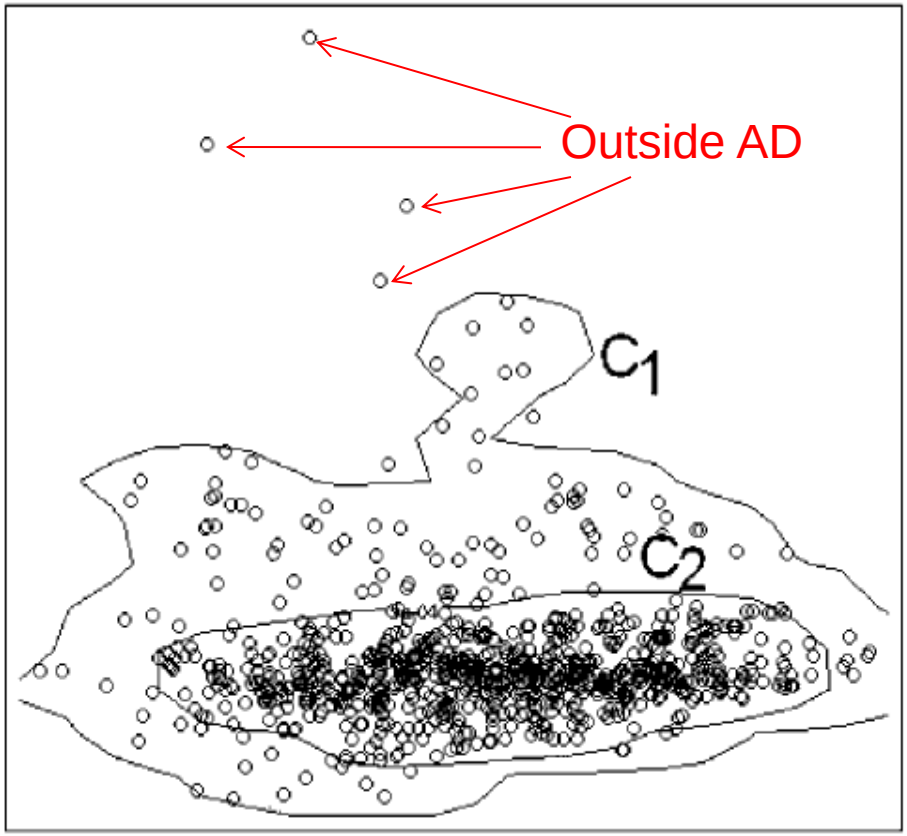
The applicability domain of a QSAR model is "the response and chemical structure space in which the model makes predictions with a given reliability"

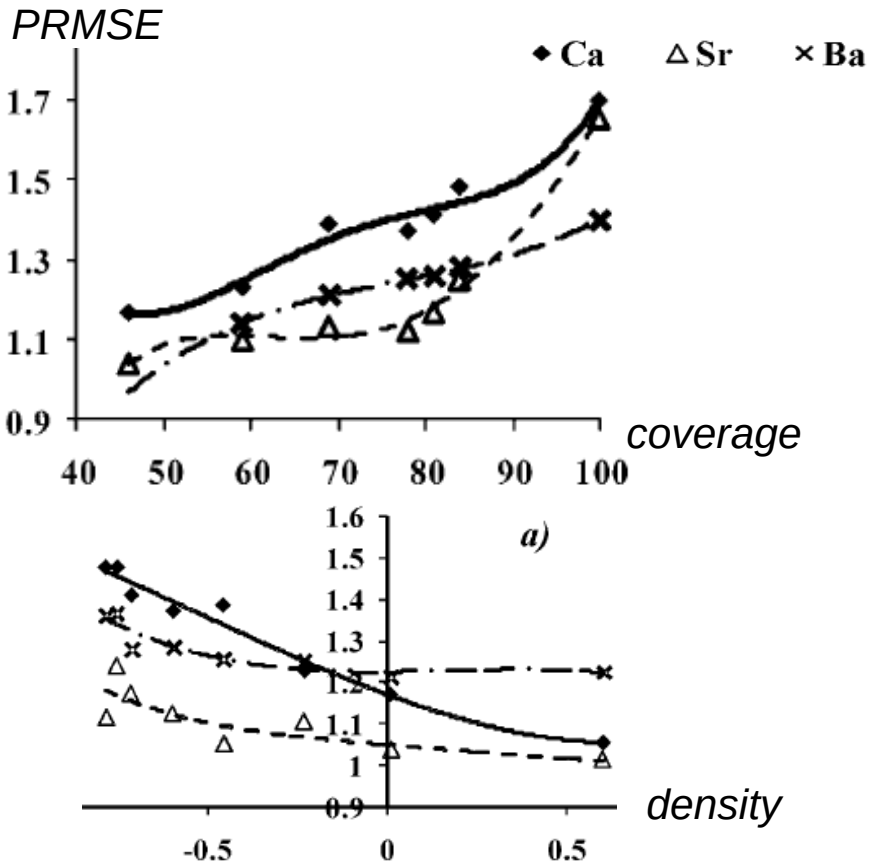
Netzeva, T. I. et al. *ALt. Lab. Anim.* ***2005, 33, 155–173.***

The AD of a (Q)SAR is the physico-chemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds. The AD of a (Q)SAR should be described in terms of the most relevant parameters i.e. usually those that are descriptors of the model. Ideally, the (Q)SAR should only be used to make predictions within that domain by interpolation not extrapolation.

The Setubal Workshop report

Applicability Domain (AD)





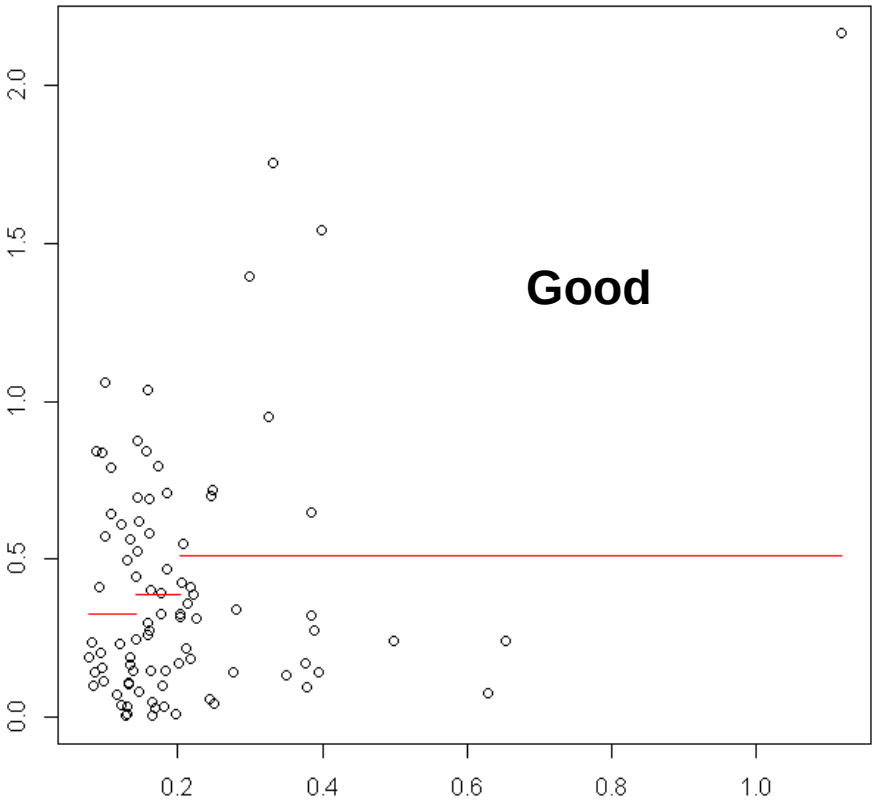
Prediction errors tend to increase with the increase of the distance from the training set (*distance to model*) and the decrease of the point distribution density. The more similar is a test compound to the training set, the more reliable are predictions for it. Outside AD predictions are made by *extrapolation*, inside AD – by *interpolation.*

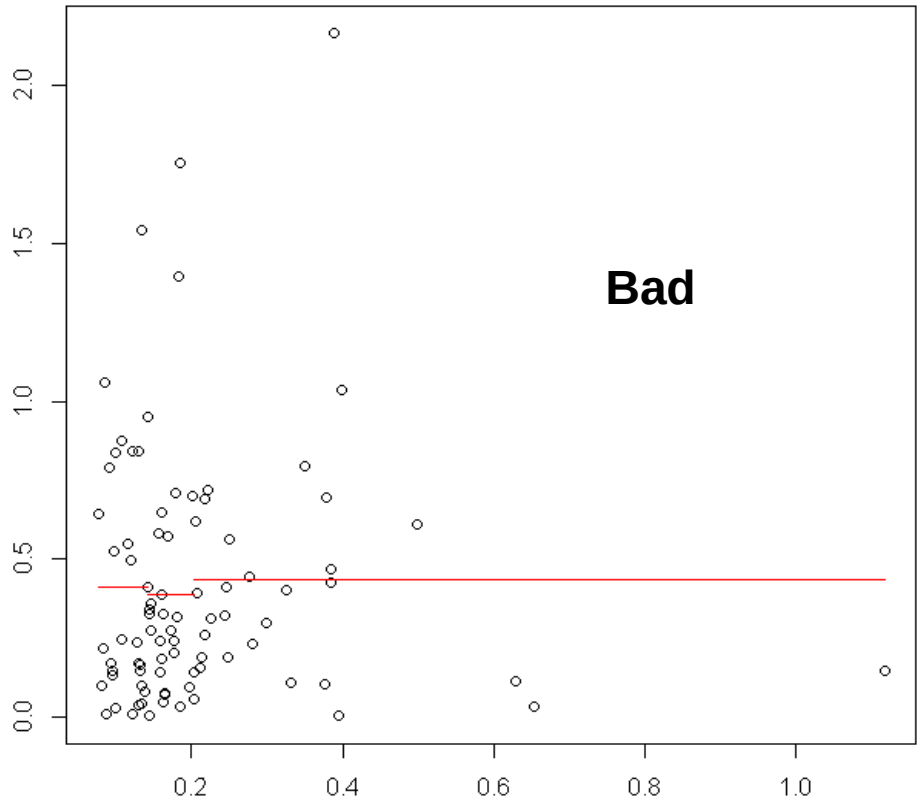
Methods for Defining Applicability Domain

- Based on descriptor ranges
- Based on the distance to the training set in descriptor space
- Based on prediction variance (distance in model space)
- Based on probability density in descriptor space
- Based on algorithms of one-class classification
- Based on conditional and unconditional probabilities

Netzeva, T.I., et al., Atla - Alternatives to Laboratory Animals, 2005. 33(2): p. 155-173.
Jaworska J., et al., Atla - Alternatives to Laboratory Animals, 2005. 33(5): p. 445-459.
Tetko, I.V., et al., J. Chem. Inf. Model., 2008. 48(9): p. 1733-1746.
Sushko, I., et al., J. Chem. Inf. Model., 2010. 50(12): p. 2096-2111.

Measures for Defining Applicability Domain



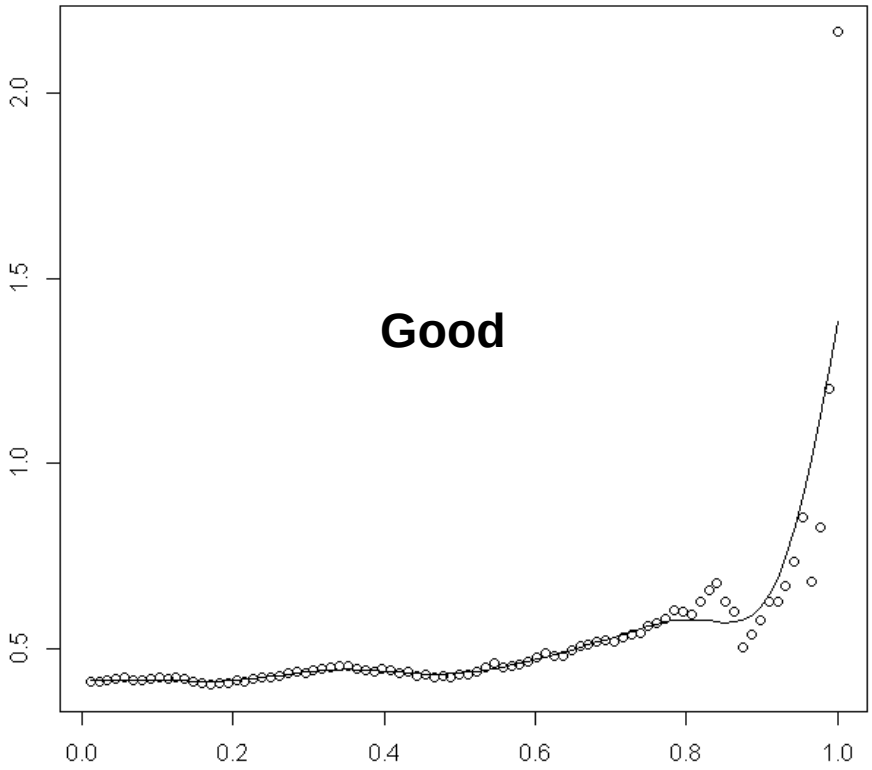


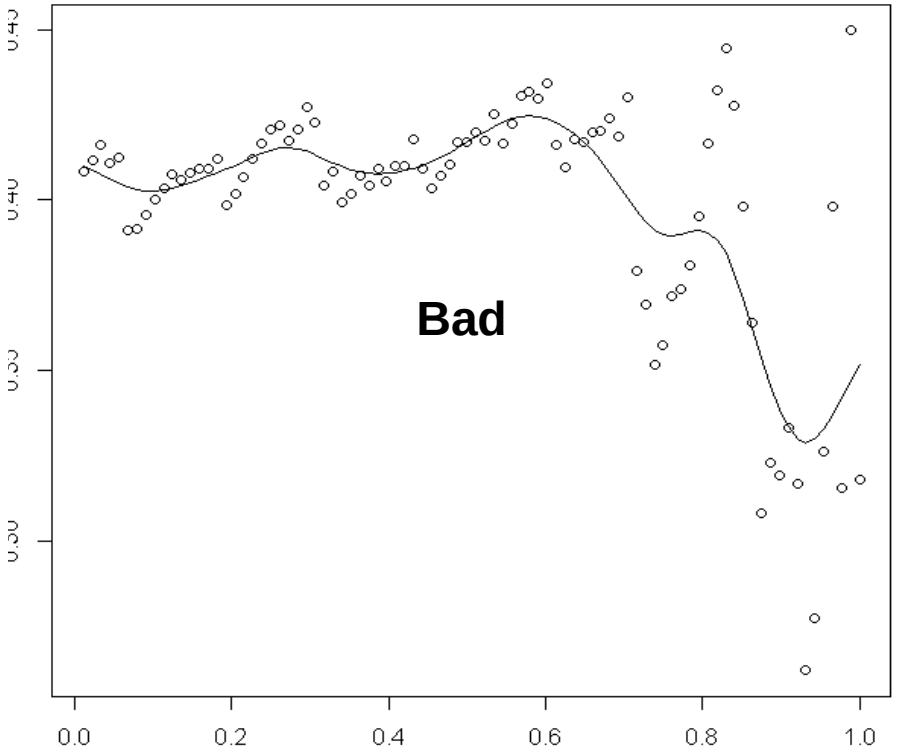
X – distance to model; **Y** – absolute prediction error; **------** - mean absolute prediction error for 1/K-part of points

If the measure is good, the step level should increase with the increase of the distance to model

3D-QSAR model, Method of CMF, prediction variance AD, pIC50 for protease inhibition

Measures for Defining Applicability Domain





X – coverage (quota or percentage of compounds inside applicability domain);
Y – mean absolute prediction error for all compounds outside applicability domain

If the measure is good, the mean absolute prediction error should increase with the increase of coverage

3D-QSAR model, Method of CMF, prediction variance AD, pIC50 for protease inhibition

Assessment of Predictive Error

Are predictions made for compounds outside AD always unreliable?

No

Are predictions made for compounds inside AD always reliable?

In addition to AD, it is necessary also to assess predictive errors

Method 1 for assessing predictive error:
- Calculate uncertainty of model coefficients
- Propagate this uncertainty through the model using Monte-Carlo simulation (or analytically for simple linear models):

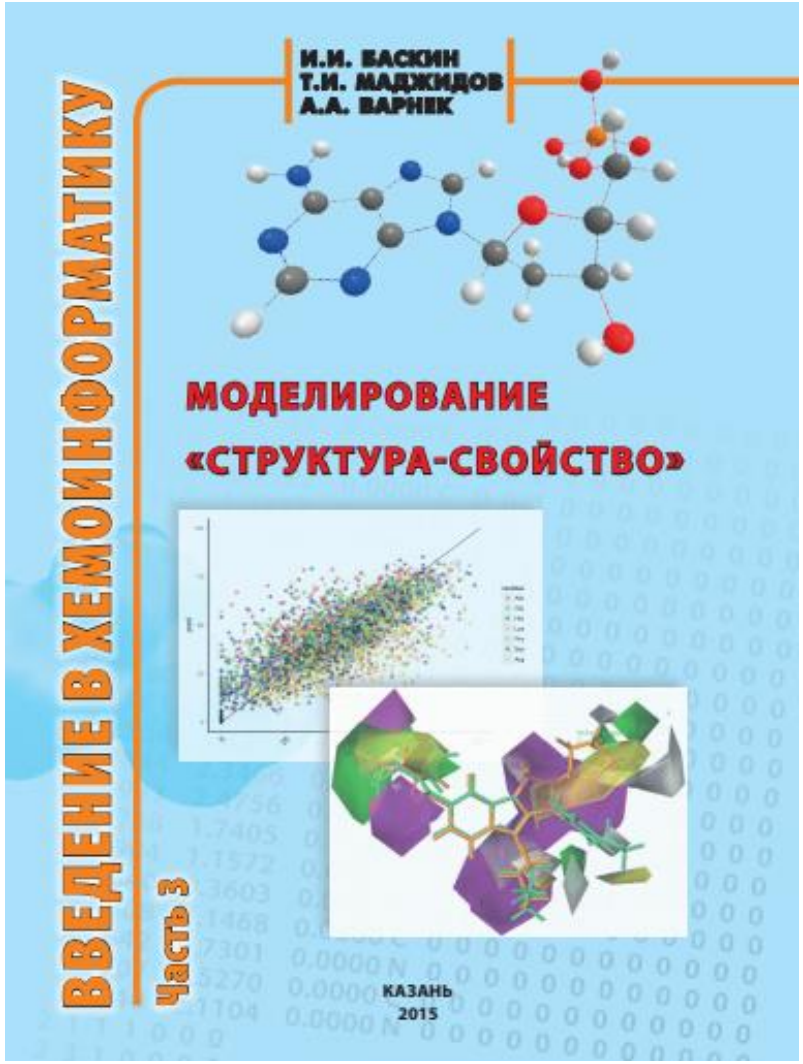
Method 2 for assessing predictive error:
- Create an ensemble of training sets from the initial data set using a resampling procedure (such as bootstrap)
- Build a model for each of the training sets
- Make predictions using ensemble of models
- Analyze prediction distribution (its variance estimates prediction error)

OECD Principles for the Validity of QSAR Models

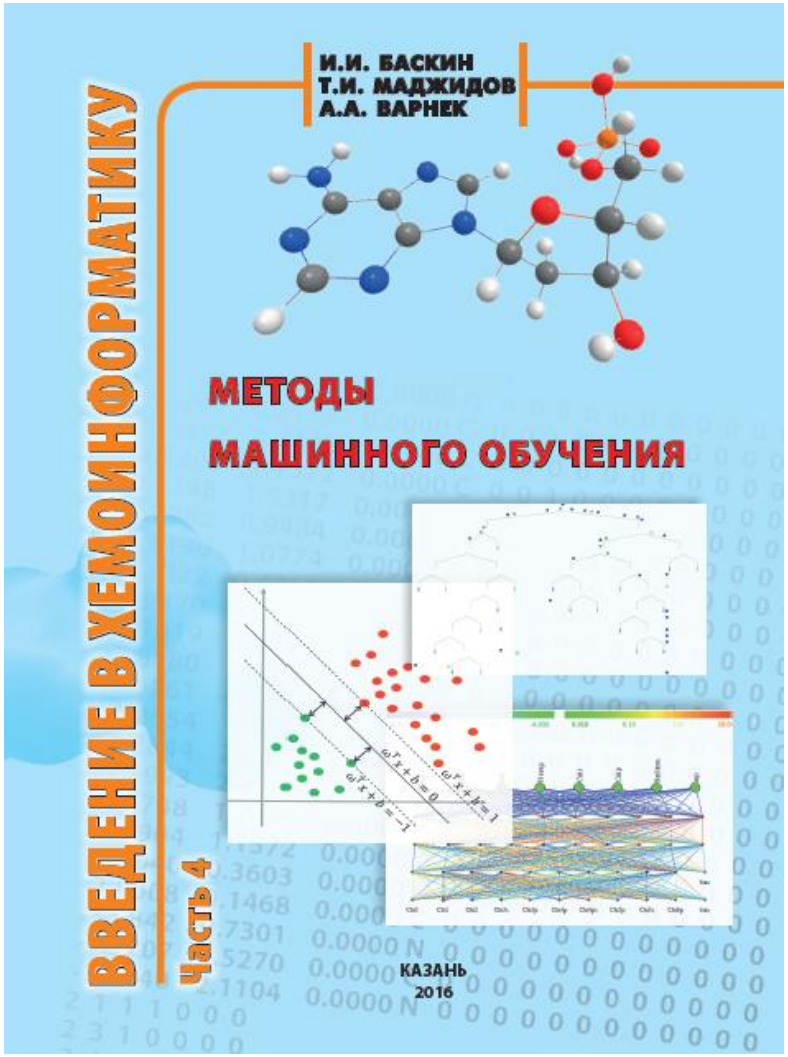
- A defined endpoint
- An unambigious algorithm
- A defined domain of applicability
- Appropriate measures of goodness-of-fit, robustness and predictivity
- A mechanistic interpretation, if possible

37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology in November 2004

Gramatica, P., *Principles of QSAR models validation: internal and external.* QSAR & Combinatorial Science, 2007. **26**(5): p. 694-701.



I.I.Baskin, T.I.Madzhidov, A.A.Varnek, Introduction to Chemoinformatics, Vol. 3, Structure-Property Modeling, Kazan, 2015



I.I.Baskin, T.I.Madzhidov, A.A.Varnek, Introduction to Chemoinformatics, Vol. 4, Methods of Machine Learning, Kazan, 2016

Questions?