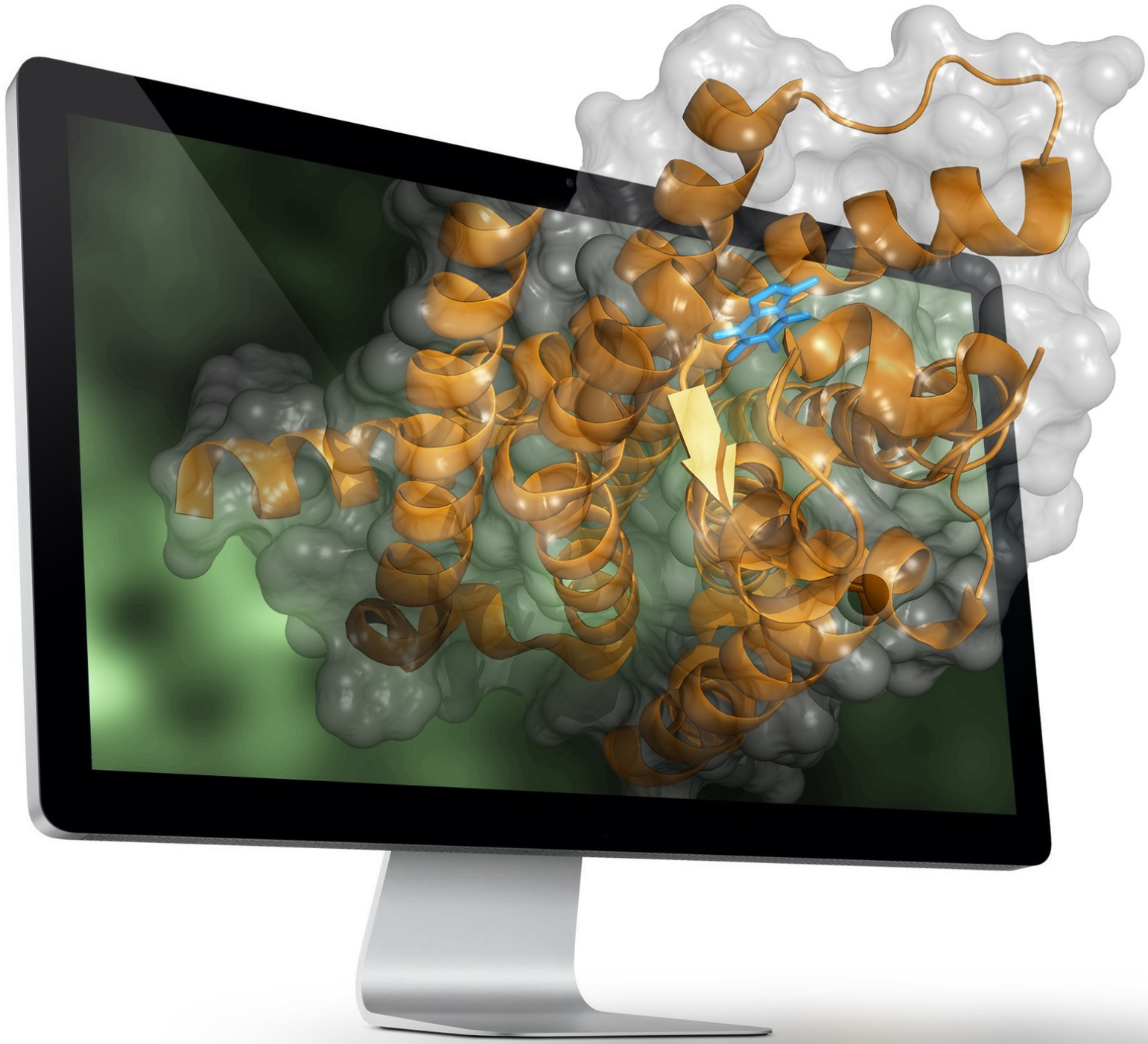


Pokročilá chemoinformatika

Úvod do QSAR/QSPR modelovaní
únor 2017

QSAR a QSPR modely – základní principy

(Q)SAR a QSPR

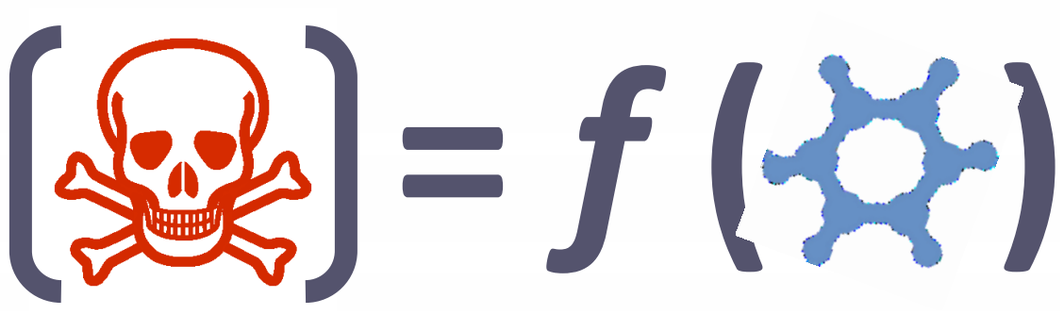
- SAR a QSAR
  - (Quantitative) Structure-Activity Relationship
  - modely pro predikci aktivity (účinnosti) chemických látek na konkretní protein nebo jinou biomolekulu
  - kvalitativní nebo kvantitativní modely
- QSPR
  - (Quantitative) Structure-Property Relationship
  - modely pro predikci fyzikálně chemických vlastností molekul

Ukázka QSAR – predikce toxicity

(Q)SAR

=

(Quantitative) Structure-Activity Relationship



IN SILICO

QSAR Modeling of Rat Acute Toxicity by Oral Exposure

Hao Zhu[†,‡], **Todd M. Martin**[§], **Lin Ye**[†,‡], **Alexander Sedykh**[‡], **Douglas M. Young**[§], and **Alexander Tropsha**[†,‡,*]

[†]Carolina Environmental Bioinformatics Research Center

[‡]Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy

[§]Sustainable Technology Division, National Risk Management Research Laboratory, Office of Research and Development, United States Environmental Protection Agency

Abstract

Few Quantitative Structure-Activity Relationship (QSAR) studies have successfully modeled large, diverse rodent toxicity endpoints. In this study, a comprehensive dataset of 7,385 compounds with their most conservative lethal dose (LD_{50}) values has been compiled. A combinatorial QSAR approach has been employed to develop robust and predictive models of acute toxicity in rats caused by oral exposure to chemicals. To enable fair comparison between the predictive power of models generated in this study *versus* a commercial toxicity predictor, TOPKAT (Toxicity Prediction by Komputer Assisted Technology), a modeling subset of the entire dataset was selected that included all 3,472 compounds used in the TOPKAT's training set. The remaining 3,913 compounds, which were not present in the TOPKAT training set, were used as the external validation set. QSAR models of five different types were developed for the modeling set. The prediction accuracy for the external validation set was estimated by determination coefficient R^2 of linear regression between actual and predicted LD_{50} values. The use of the applicability domain threshold implemented in most models generally improved the external prediction accuracy but expectedly led to the decrease in chemical space coverage; depending on the applicability domain threshold, R^2 ranged from 0.24 to 0.70. Ultimately, several consensus models were developed by averaging the predicted LD_{50} for every compound using all 5 models. The consensus models afforded higher prediction accuracy for the external validation dataset with the higher coverage as compared to individual constituent models. The validated consensus LD_{50} models developed in this study can be used as reliable computational predictors of *in vivo* acute toxicity.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796713/

The Practice of Structure Activity Relationships (SAR) in Toxicology

James D. McKinney; Ann Richard; Chris Waller; Michael C. Newman; Frank Gerberick

Toxicol Sci (2000) 56 (1): 8-17. **DOI:** https://doi.org/10.1093/toxsci/56.1.8
Published: 01 July 2000 **Article history** ▾

Abstract

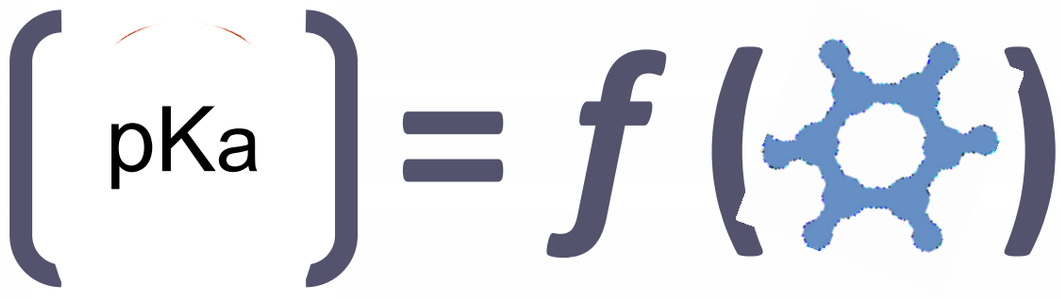
Both qualitative and quantitative modeling methods relating chemical structure to biological activity, called structure–activity relationship analyses or SAR, are applied to the prediction and characterization of chemical toxicity. This minireview will discuss some generic issues and modeling approaches that are tailored to problems in toxicology. Different approaches to, and some facets and limitations of the practice and science of, SAR as they pertain to current toxicology analyses, and the basic elements of SAR and SAR-model development and prediction systems are discussed. Other topics include application of 3-D SAR to understanding of the propensity of chemicals to cause endocrine disruption, and the use of models to analyze biological activity of metal ions in toxicology. An example of integration of knowledge pertaining to mechanisms into an expert system for prediction of skin

Ukázka QSPR – predikce disociační konstanty

QSPR

=

Quantitative Structure-Property Relationship



IN SILICO

RESEARCH ARTICLE Open Access

Predicting pK_a values from EEM atomic charges

Radka Svobodová Vařeková[1], Stanislav Geidl[1], Crina-Maria Ionescu[1], Ondřej Skřehota[1], Tomáš Bouchal[1], David Sehnal[1], Ruben Abagyan[2] and Jaroslav Koča[1*]

Abstract

The acid dissociation constant pK_a is a very important molecular property, and there is a strong interest in the development of reliable and fast methods for pK_a prediction. We have evaluated the pK_a prediction capabilities of QSPR models based on empirical atomic charges calculated by the Electronegativity Equalization Method (EEM). Specifically, we collected 18 EEM parameter sets created for 8 different quantum mechanical (QM) charge calculation schemes. Afterwards, we prepared a training set of 74 substituted phenols. Additionally, for each molecule we generated its dissociated form by removing the phenolic hydrogen. For all the molecules in the training set, we then calculated EEM charges using the 18 parameter sets, and the QM charges using the 8 above mentioned charge calculation schemes. For each type of QM and EEM charges, we created one QSPR model employing charges from the non-dissociated molecules (three descriptor QSPR models), and one QSPR model based on charges from both dissociated and non-dissociated molecules (QSPR models with five descriptors). Afterwards, we calculated the quality criteria and evaluated all the QSPR models obtained. We found that QSPR models employing the EEM charges proved as a good approach for the prediction of pK_a (63% of these models had $R^2 > 0.9$, while the best had $R^2 = 0.924$). As expected, QM QSPR models provided more accurate pK_a predictions than the EEM QSPR models but the differences were not significant. Furthermore, a big advantage of the EEM QSPR models is that their descriptors (i.e., EEM atomic charges) can be calculated markedly faster than the QM charge descriptors. Moreover, we found that the EEM QSPR models are not so strongly influenced by the selection of the charge calculation approach as the QM QSPR models. The robustness of the EEM QSPR models was subsequently confirmed by cross-validation. The applicability of EEM QSPR models for other chemical classes was illustrated by a case study focused on carboxylic acids. In summary, EEM QSPR models constitute a fast and accurate pK_a prediction approach that can be used in virtual screening.

Keywords: Dissociation constant, Quantitative structure-property relationship, QSPR, Partial atomic charges, Electronegativity equalization method, EEM, Quantum mechanics, QM

Princip QSAR a QSPR modelů

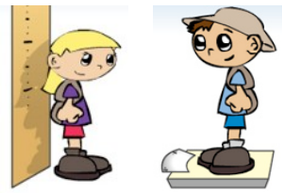
Chemoinformatický nástroj pro výpočet aktivity nebo vlastností na základě struktury

Obecné schéma:

struktura ⟶ **Číselný popis** (deskriptory) ⟶ **Aktivita (vlastnost) = funkce (deskriptory)**

Ilustrativní příklad ze života:

Tělo člověka



Výška a váha

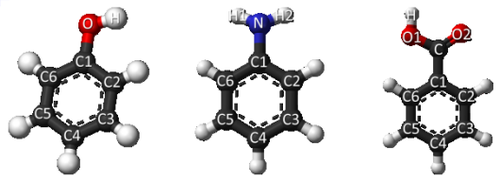
BMI = váha / výška²

BMI (Body Mass Index):

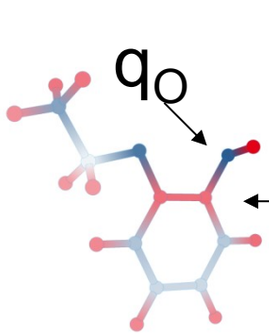
Podváha: BMI < 18,5

Obezita: BMI > 30

Příklad reálné aplikace:



3D struktura molekuly

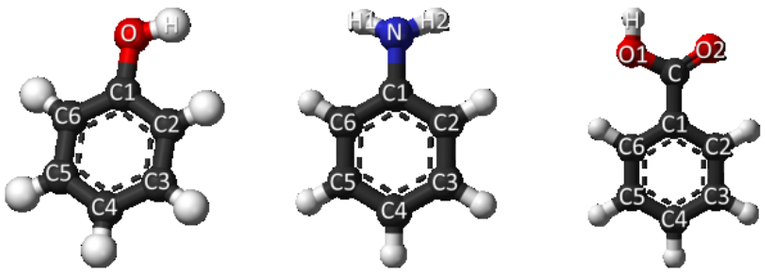


Náboje na atomech

$pK_a = c_H.q_H + c_O.q_O + c_{C1}.q_{C1}$

Kde c_H, c_O a c_{C1} jsou parametry modelu

Molekula, její struktura, deskriptor a model



molekula a její struktura

matematická, logická nebo statistická operace

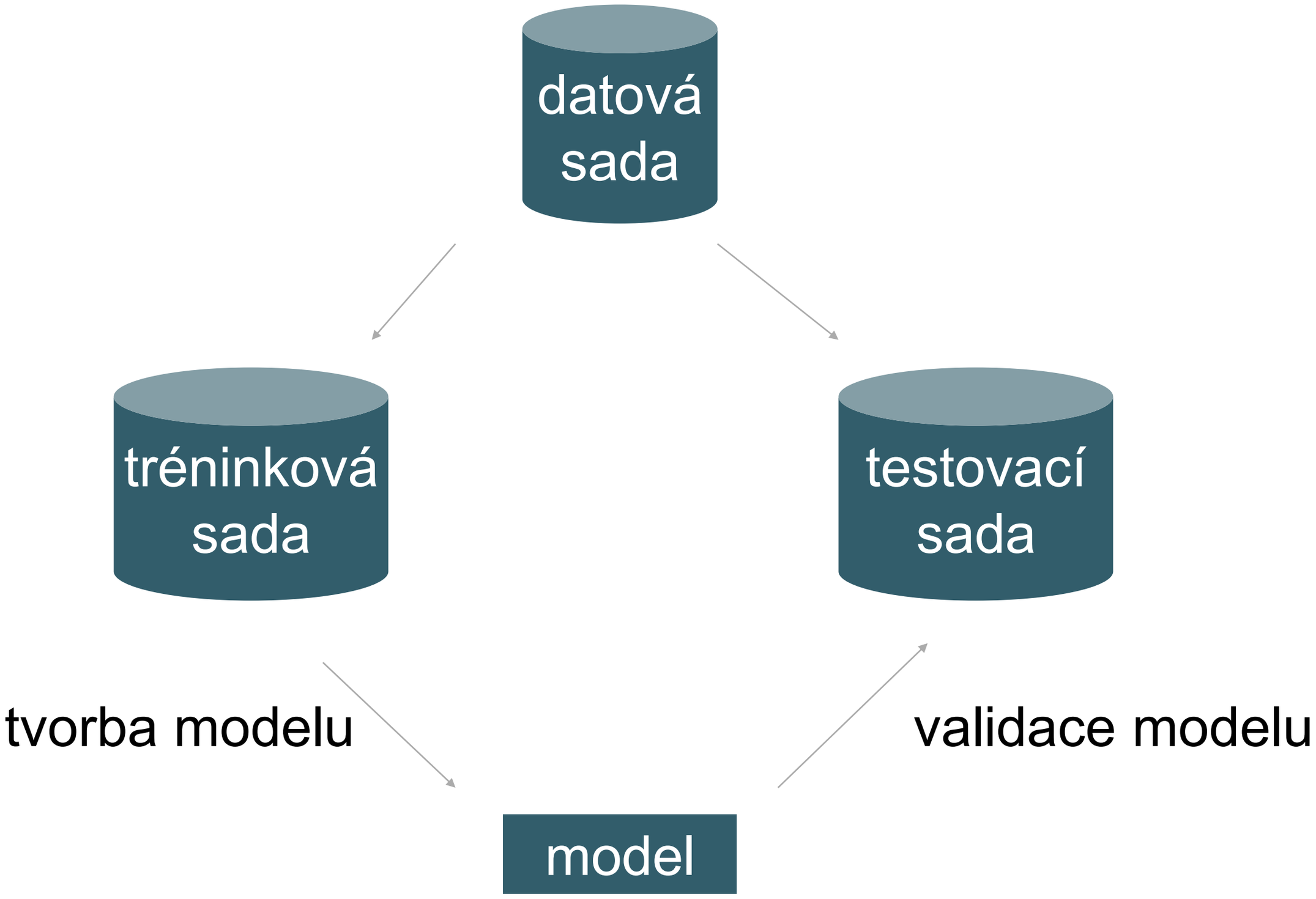
-100 4.6 8.2

deskriptory

black box

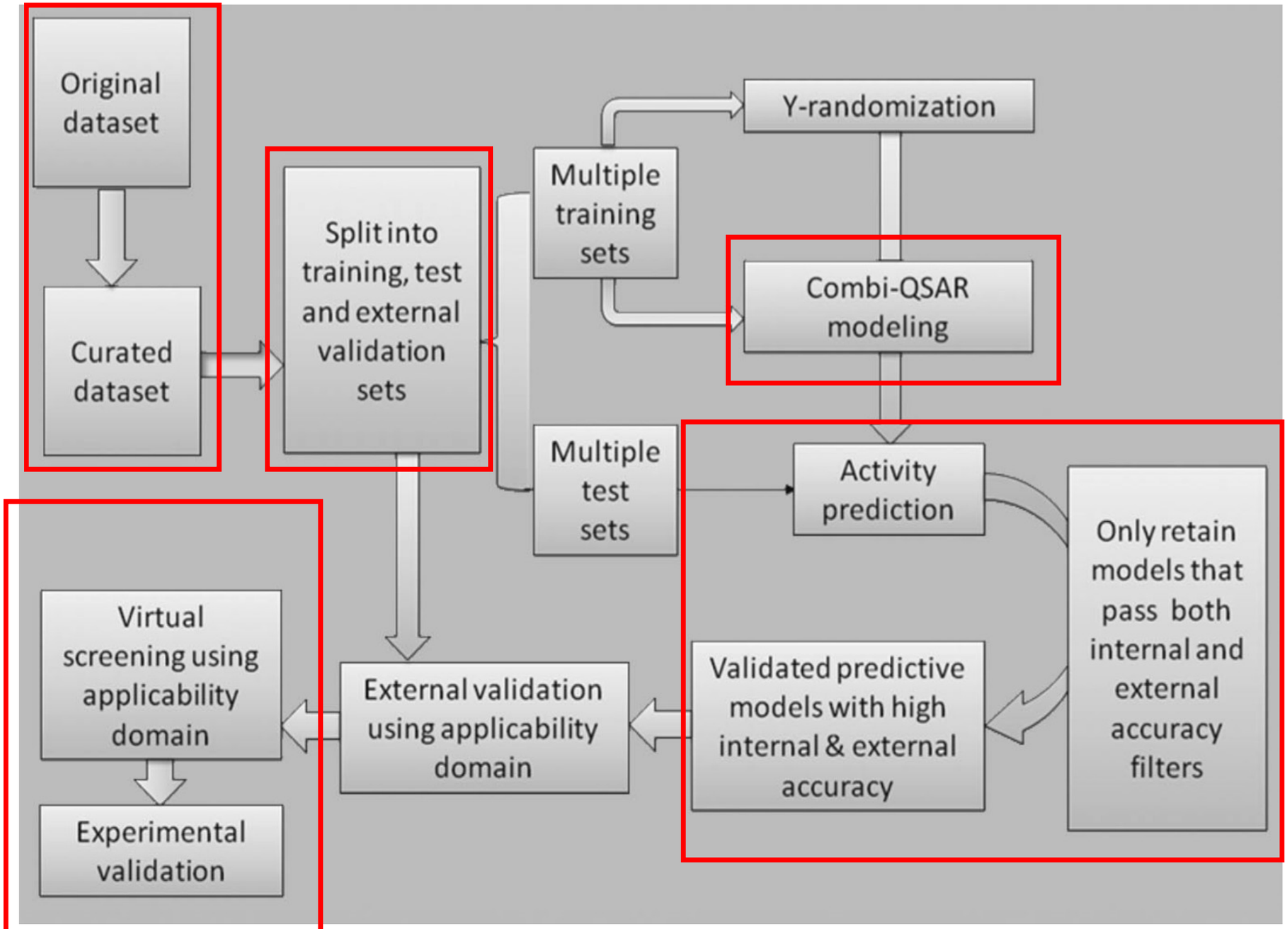
model

Tvorba modelu - schématicky



Tvorba modelu – best practicies

Studijní materiály > Články > QSAR_best_practices.pdf



http://onlinelibrary.wiley.com/doi/10.1002/minf.201000061/abstract

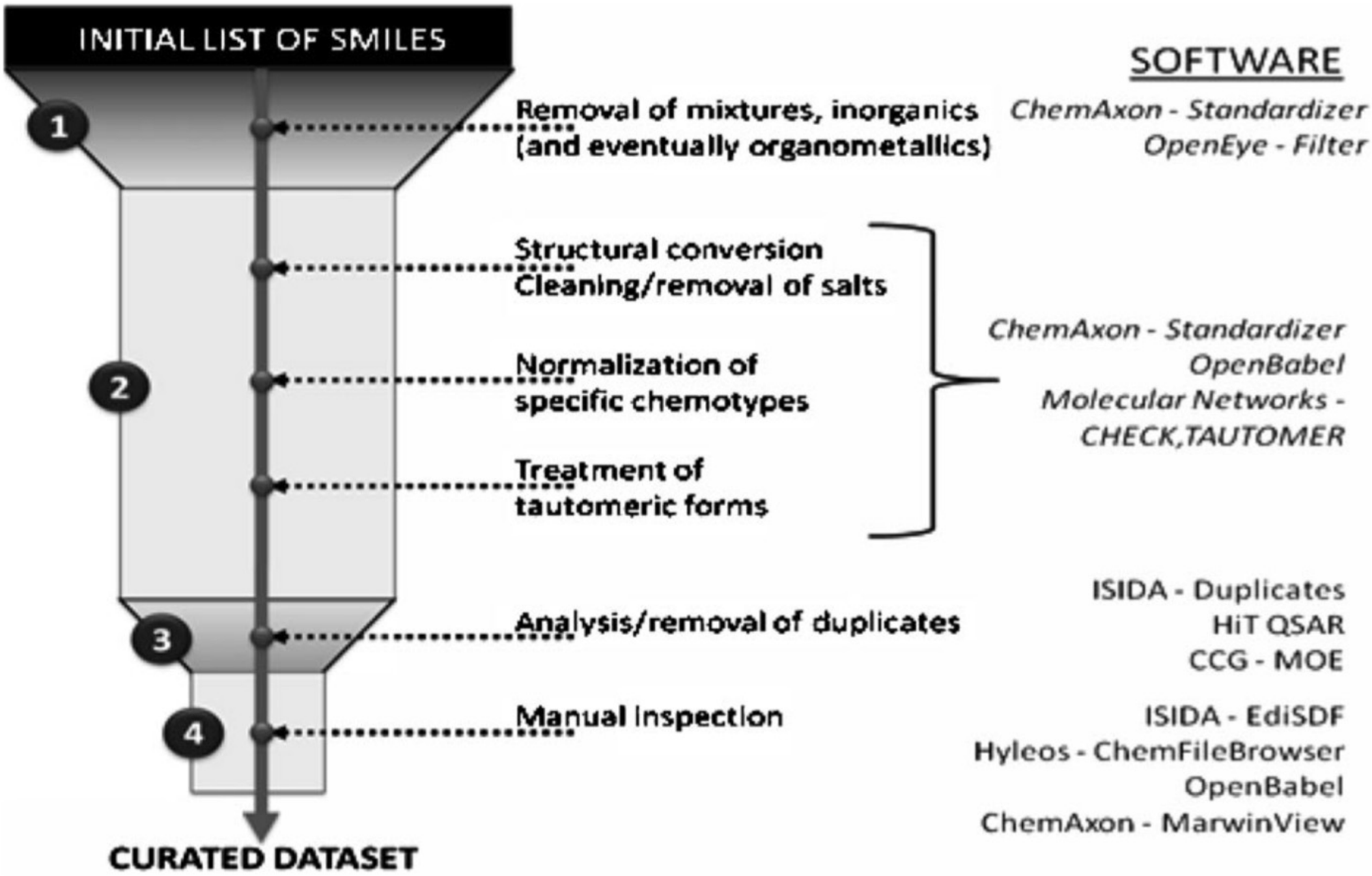
Datová sada

- molekuly a jejich experimentální aktivita nebo vlastnost, kterou hodláme predikovat
- struktura molekuly (nejméně SMILES)
- další kriteria na datovou sadu:
  - sada je dostatečně rozsáhlá
  - molekuly jsou dostatečně chemicky různorodé
  - hodnoty vlastnosti nebo aktivity musí být dostatečně různorodé

Základní aspekty čištění datové sady

- **kontola molekul**
  - odstranění duplicitních struktur
  - odstranění nadbytečných informací ze struktury
    - validace struktur (vazebné délky, počet vazeb, …)
- **kontrola vlastností**
  - správnost přiřazení, …

Čištění datové sady



Kriteria kvality modelů

Kvalita QSAR/QSPR modelů

- kvalitu modelu můžeme posuzovat podle dvou kritérií

① kvalitu modelu na tréninkové sadě dat

- **reprodukce –** data byla použita pro naučení modelu

- jak moc dobré modely jsme připravili?

② kvalitu modelu na testovací sadě dat

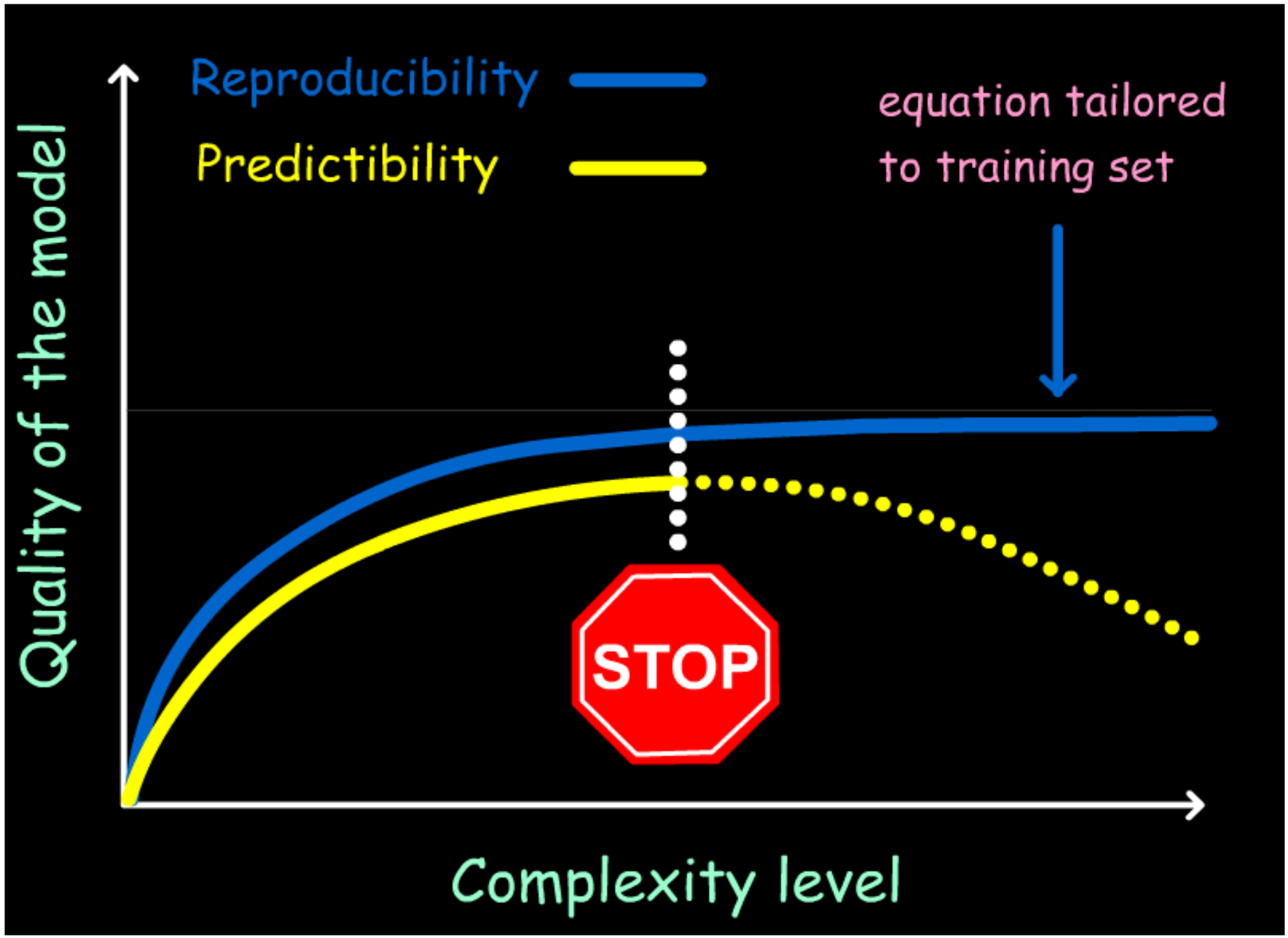
- **predikce** (na nových datech) – data nebyla použita na parametrizaci modelu

- jaká je predikční sada molekul?

Kvalita QSAR/QSPR modelů
reprodukce a predicke

	nekvalitní model na tréninkové sadě dat	kvalitní model na tréninkové sadě dat
nekvalitní model na testovací sadě dat	–	špatně rozdělené sady, “overfiting” neboli přeučení = použito příliš moc deskriptorů
kvalitní model na testovací sadě dat	–	KVALITNÍ MODEL

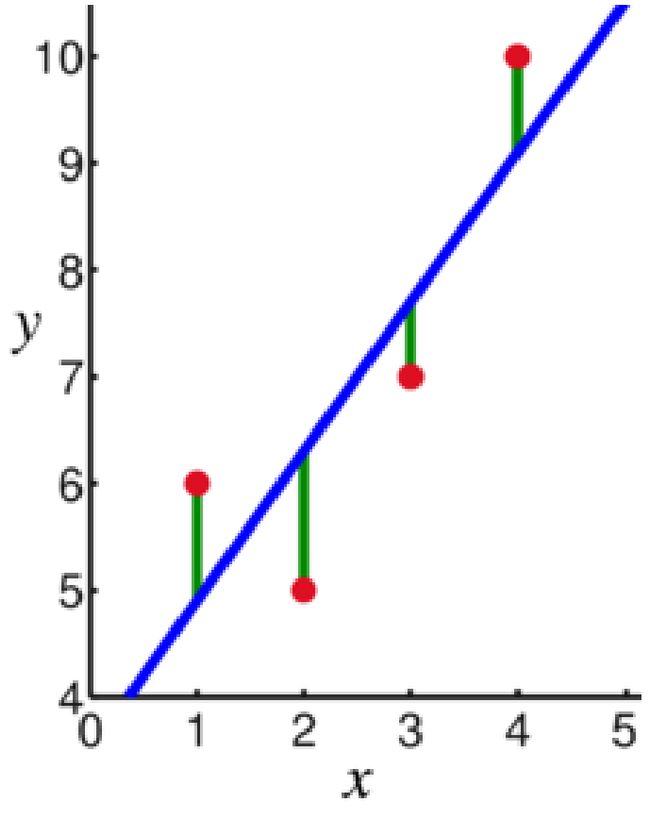
Kvalita QSAR/QSPR modelů – přeučení



Kvalita na základě chyb modelu

- chyby, rozdíly mezi predikovanou a naměřenou hodnotou = residua, nevysvětlitelná část modelu

P^{exp}	P^{calc}	$\text{error} = P^{exp} - P^{calc}$
...	...	...
$\text{p}K_a{}^{exp}$	$\text{p}K_a{}^{calc}$	error
10.0	10.1	-0.1
...	...	...



- vyjadřujeme pomocí R^2, adjR^2, RMSE, MAE a F

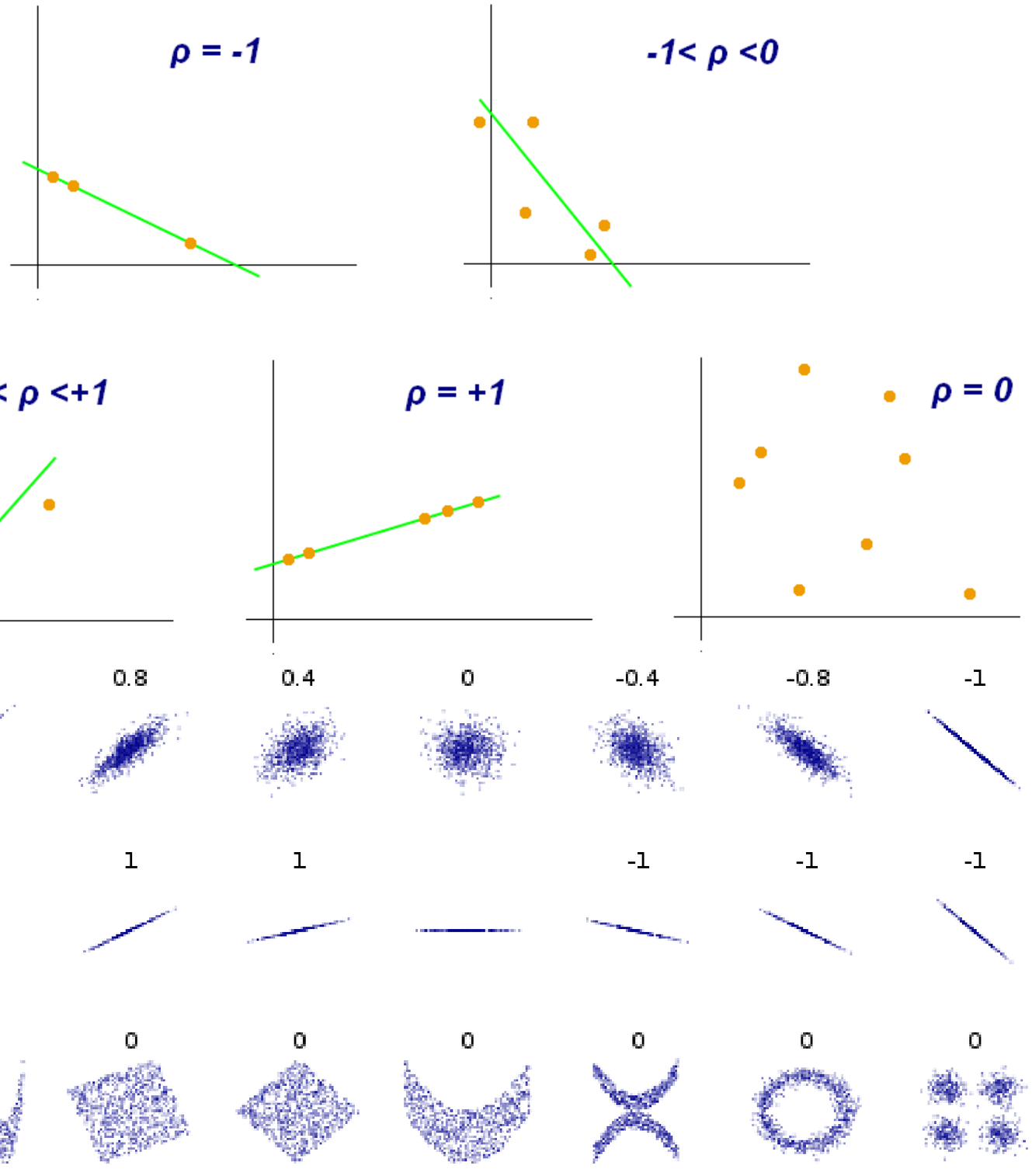
Pearsonův korelační koeficient – vzorec

$$R = \frac{\sum_{i=1}^{N} \left((P_i^{calc} - \bar{P}calc) \cdot (P_i^{exp} - \bar{P}exp) \right)}{\sqrt{\sum_{i=1}^{N} (P_i^{calc} - \bar{P}calc)^2 \cdot \sum_{i=1}^{N} (P_i^{exp} - \bar{P}exp)^2}}$$

$\bar{P}^{calc}$ průměrná vypočítaná hodnota,
$\bar{P}^{exp}$ průměrná experimentální hodnota

Nabývá hodnot od -1 do 1.

Pearsonův korelační koeficient – ukázka



Koeficient determinace R^2 – definice

- Leží v intervalu <0;1> a udává jaký podíl rozptylu v pozorování závislé proměnné se podařilo regresí vysvětlit (vetší hodnoty znamenají větší úspěšnost).

- Možná interpretace koeficientu R^2 je z kolika procent vysvětlují regresory (deskriptory) hodnotu závisle proměnné (predikované aktivity/vlastnosti).

Koeficient determinace R^2 – vzorec

Residual sum of squares:

$$RSS = \sum_{i=1}^{N} error^2 = \sum_{i=1}^{N} (P_i^{calc} - P_i^{exp})^2$$

Total sum of squares: $TSS = \sum_{i=1}^{N} (P_i^{exp} - \bar{P}^{exp})^2$

Explained sum of squares: $ESS = (P_i^{calc} - \bar{P}^{calc})^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$R^2 = \frac{\sum_{i=1}^{N} (P_i^{calc} - \bar{P}^{calc})^2}{\sum_{i=1}^{N} (P_i^{exp} - \bar{P}^{exp})^2} = 1 - \frac{\sum_{i=1}^{N} (P_i^{calc} - P_i^{exp})^2}{\sum_{i=1}^{N} (P_i^{exp} - \bar{P}^{exp})^2}$$

Korigovaný koeficient determinace adjR^2

- pokud do modelu přidáme deskriptor, hodnota R^2 nemůže klesnout, proto se někdy používá tzv. korigovaný koeficient determinace (**adjusted coefficient of determination**), který zohledňuje počet deskriptorů

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N - 1}{N - k - 1}$$

kde N je velikost sady, k počet deskriptorů

RMSE
root mean square error (deviation)

$$RMSE = \sqrt{mean(error^2)} = \sqrt{\frac{\sum error^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(P_i^{calc} - P_i^{epx})^2}{N}}$$

MAE
mean absolute error

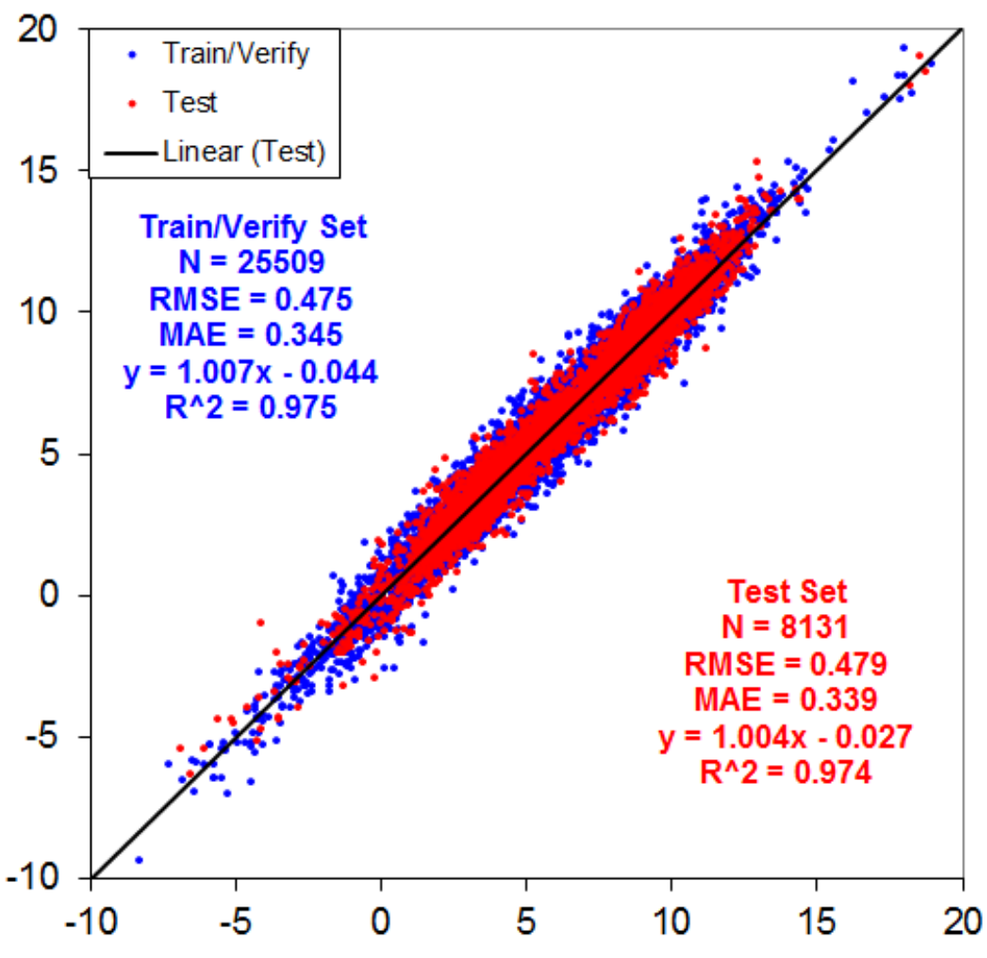
$$MAE = mean(abs(error)) = \frac{\sum |error|}{N} = \frac{\sum_{i=1}^{N} |P_i^{calc} - P_i^{epx}|}{N}$$

Test významnosti modelu F

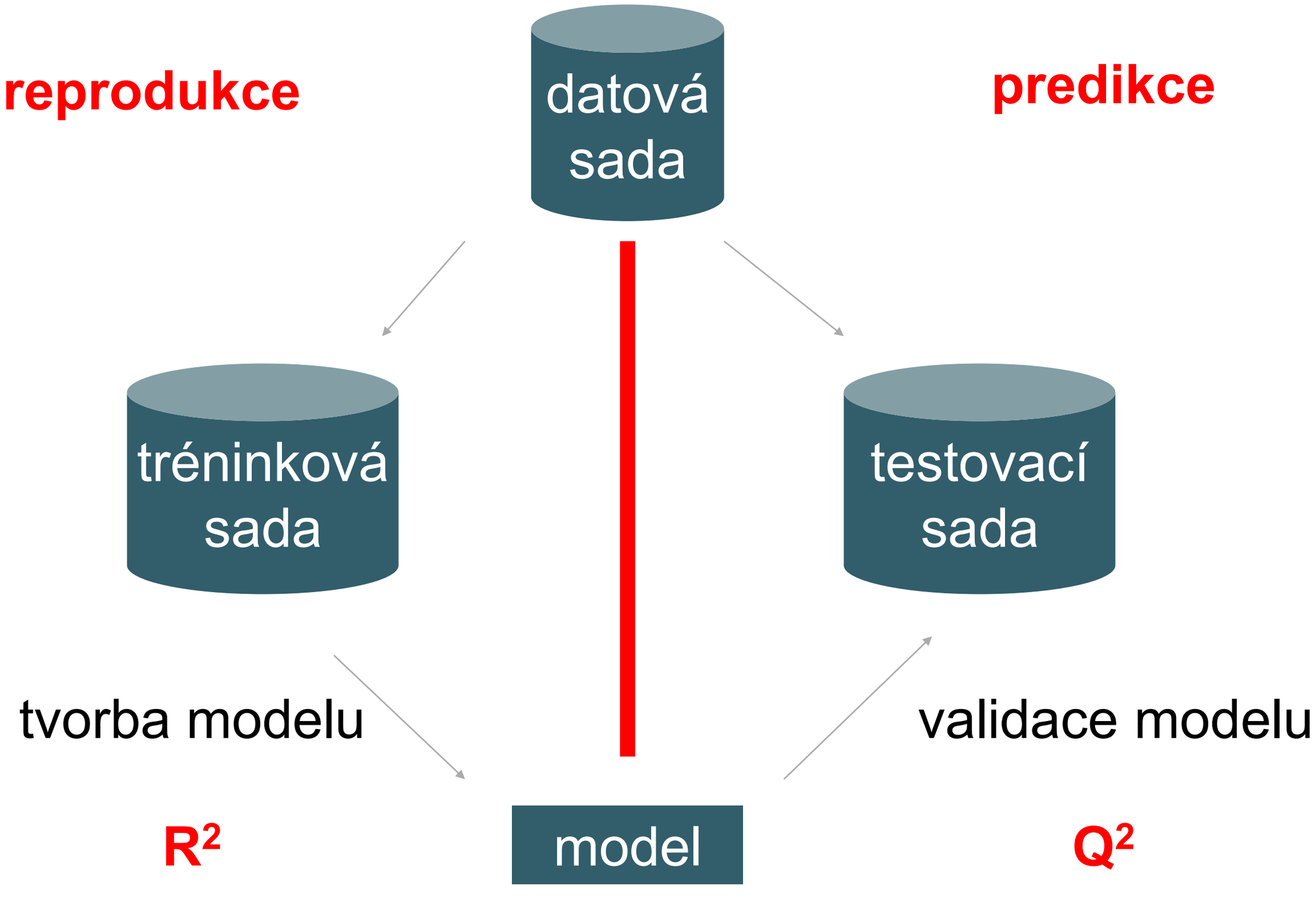
$$F \sim \frac{N-k+1}{k} \frac{RSS-TSS}{TSS} = \frac{N-k+1}{k} \frac{R^2}{1-R^2}$$

Předpokládané hodnoty pro kvalitní modely

- Kvalitní model by měl splňovat tato kritéria:
  - vysoké hodnoty R^2 (>0.8) a F
  - nízké hodnoty RMSE a MAE



Rozdělění datové sady - terminologie



Křížová validace
Cross validation

- v případě menší sady molekul

- nejčastěji se používá tzv. *k*-fold cross validation; příklad 5-fold:

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
compounds 1-10	Test set	Training set	Training set	Training set	Training set
compounds 11-20	Training set	Test set	Training set	Training set	Training set
compounds 21-30	Training set	Training set	Test set	Training set	Training set
compounds 31-40	Training set	Training set	Training set	Test set	Training set
compounds 41-50	Training set	Training set	Training set	Training set	Test set