QSAR modeling

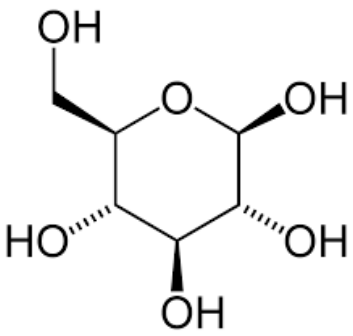
Guzel Minibaeva

Ph.D. student

Institute of Molecular and Translational Medicine
Faculty of Medicine and Dentistry
Palacky University

QSAR modeling workflow

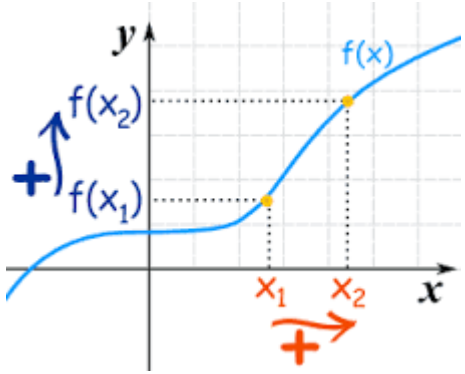
Structure



Descriptors (features)

D_1	D_2	D_3	D_4	D_5	D_6	...	D_N
1	0	9	0	11	1	...	1
4	0	1	0	0	0	...	1
0	0	0	0	0	4	...	6
0	2	3	6	0	0	...	3
...	...	...	...	...	...	...	...
4	0	0	0	1	2	...	1

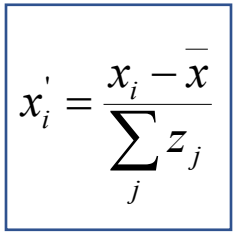
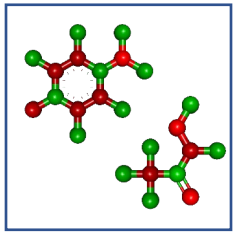
Model



Encoding
(represent structure with numerical features)

Mapping
(machine learning)

Overall QSAR workflow

Input data	Preprocessing	Feature engineering	Model training	Model validation	Interpretation
					
Bioassays	*Data normalization & curation*	*Feature selection*	*Classification*	*Cross-validation*	
Databases	*Feature extraction*	*Feature combination*	*Regression*	*Bootstrap*	
			Clustering	*Test set*	
				Applicability Domain	

OECD principles for the validation, for regulatory purposes, of (Q)SAR models

1) a defined endpoint
2) an unambiguous algorithm
3) a defined domain of applicability
4) appropriate measures of goodness-of–fit, robustness and predictivity
5) a mechanistic interpretation, if possible

Step 1. Data collection

Scientific literature and patents
Databases (ChEMBL, PubChem, BindingDB, etc)

Traditionally modeled compounds should have the same mechanism of action, however using of complex non-linear machine learning method allows to model data sets with mixed or even unknown mechanism of action with reasonable accuracy.

Conditions may substantially influence the results of bioassays (change in temperature, activators, detectors, etc)

Units checking

Step 2. Data curation (normalization)

J. Med. Chem. **2000**, *43*, 3233−3243

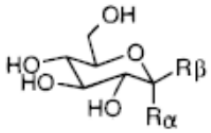
Option

GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors

Manuel Pastor,[†] Gabriele Cruciani,*[,†] Iain McLay,[§] Stephen Pickett,[§] and Sergio Clementi[†]

Laboratory on Chemometrics, Department of Chemistry, University of Perugia, Via Elce di Sotto 10, 06123 Perugia, Italy, and CADD Department, Rhone-Poulenc Rorer, Dagenham, Essex RM10 7XS, U.K.

Table 2. Series of 10 Glucose Analogue Inhibitors of Glycogen Phosphorylase

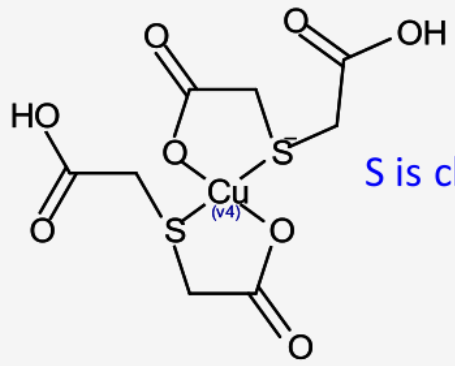


	substituent at C1 position		
no.	Rα	Rβ	pK_i (mM)
1	OH	H	2.77
2	C(=O)NH_2	H	3.43
3	H	C(=O)NH_2	3.36
4	H	$COOCH_3$	2.55
5	H	CH_2CN	2.05
6	H	NHC(=O)NH_2	3.85
7	C(=O)NH_2	$NHCOOCH_3$	4.80

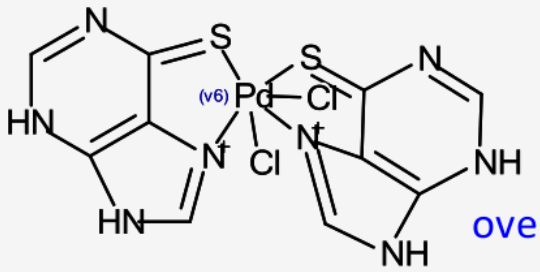
strange units

Step 2. Data curation (normalization)

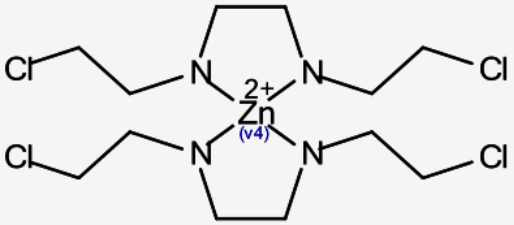
Data from NCI60

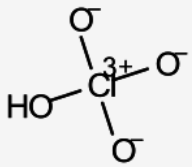


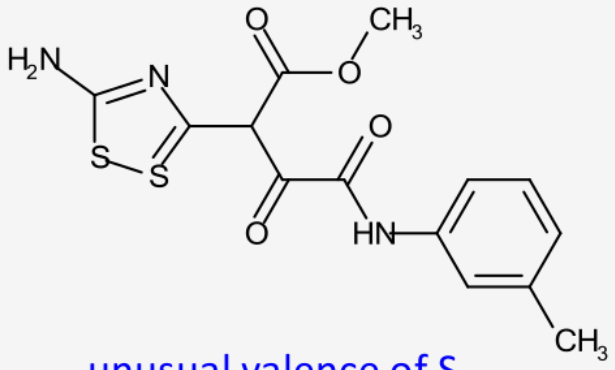
S is charged



overall charge is +2







unusual valence of S

overall charge is +2

HClO4 is represented with separated charges

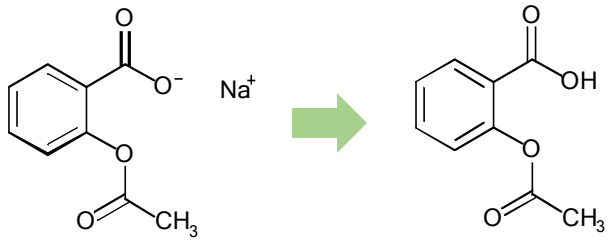
nitrogens are covalently bond to Zn?

wrong stoichiometry?

Step 2. Data curation (normalization)

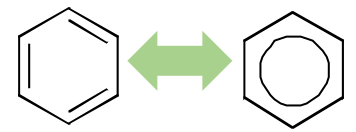
Removal of mixtures, inorganics, metalorganics, etc

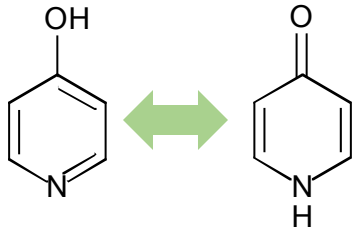
Strip of salts, counterions, etc

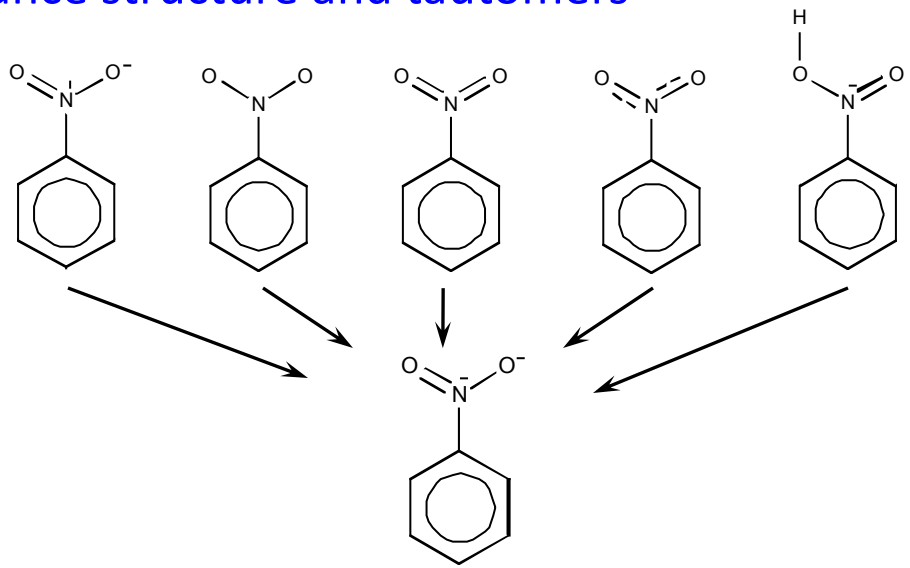


Ionization, if necessary (at the particular pH level)

Chemotype normalization, resonance structure and tautomers







Duplicates removal

Manual checking

Step 3. Descriptors: classification

Object type:

- **molecular** descriptors (single molecules)
- descriptors of molecular **ensemble** (mixtures, materials)
- **reaction** descriptors (reactions)

Descriptor origin:

- **calculated** from the structure
- **empirical** (Hammet constants, lipophilicity chemical shifts in NMR, etc)

Locality:

- **local** (atom charge)
- **global** (molecular weight, molecular volume, lipophilicity, etc)

Dimensionality:

- **1D** (number of methyl groups, molecular weight, etc)
- **2D** (topological indices, fragmental descriptors)
- **3D** (molecular volume, quantum chemical descriptors)
- **4D** (based on a set of conformers)

Calculation method:

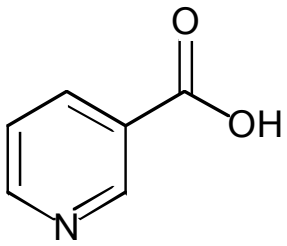
- **physico-chemical** (lipophilicity, etc)
- **topological** (invariants of molecular graph, Randic index, Wiener index, etc)
- **fragmental** (fingerprints, etc)
- **pharmacophore**
- **spatial** (moment of inertia, etc)
- **quantum-chemical** (energy of HOMO/LUMO, etc)
- etc.

R. Todeschini and V. Consonni Handbook of Molecular Descriptors, 2008

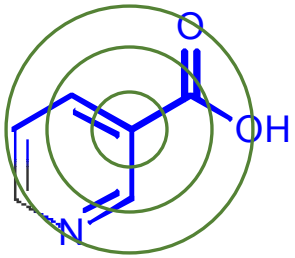
Atom-centric (augmented atoms) fingerprints

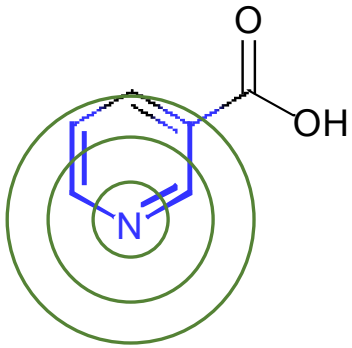
Generate substructures starting from each atom and considering all its neighbors up to the specified distance (radius or diameter).

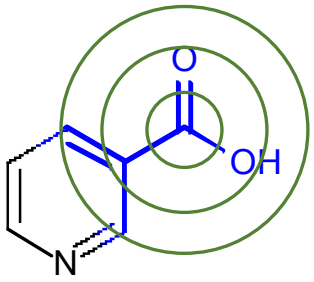
Morgan fingerprints, Extended-connectivity fingerprints (ECFP). Functional-class fingerprints (FCFP) , etc.



radius=2 (diameter=4)

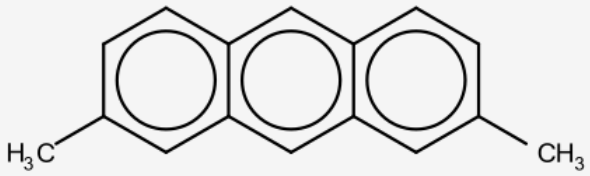


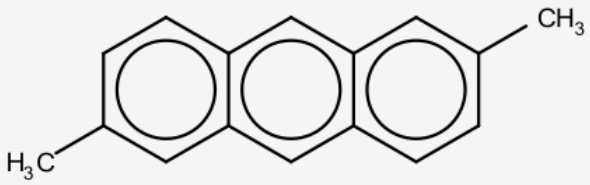




Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 50, 742-754 (2010)

Atom-centric (augmented atoms) fingerprints



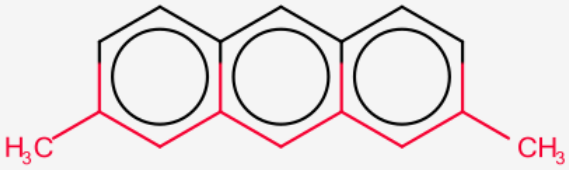


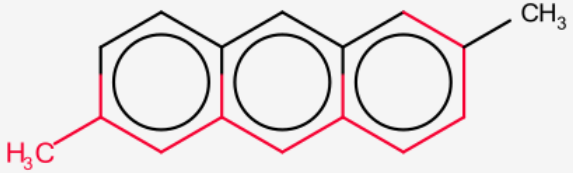
Morgan fingerprints
radius=2
(diameter=4)

identical fingerprints

Morgan fingerprints
radius=4
(diameter=8)

different fingerprints

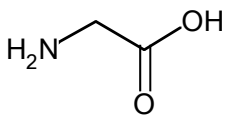
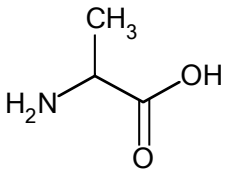
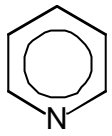




Fingerprints

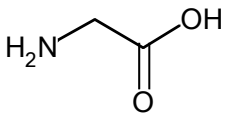
Each molecule has variable length set of substructures – variable length fingerprints

2-bond sequences

	N–C–C	C–C–O	C–C=O	C–C–C	C:C:C	C:C:N	C:N:C
	1	1	1	0	0	0	0
	1	1	1	1	0	0	0
	0	0	0	0	1	1	1

Hashed fingerprints

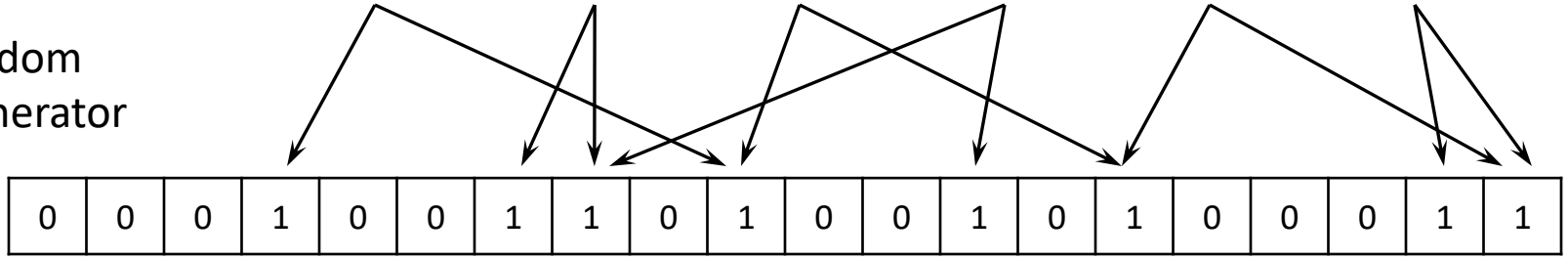
Have fixed length (usually 512, 1024 or 2048 bits)



N-C-C	C-C-O	C-C=O	C-C-C	C:C:C	C:C:N
13823	9740	37278	28478	874	283764

hash code

pseudo-random number generator



fixed-length bit string

Each substructure activates several bits (usually 4-5) to avoid collisions and produce bit string of enough density

Missing bits mean that certain substructures are not presented, Active bits mean that certain substructure may be present (but due to possible collisions one cannot be sure)

Step 4. Feature processing

Feature transformations:

linear and non-linear scaling

$$z_i = \frac{x_i - \bar{x}}{sd} \qquad sd = \sqrt{\frac{n}{n-1}\sum_{i=1}^{n}(x_i - x)^2}$$

$$z_i = \frac{1}{1 + e^{-x_i}}$$ ➡ range (0; 1)

Feature combinations:

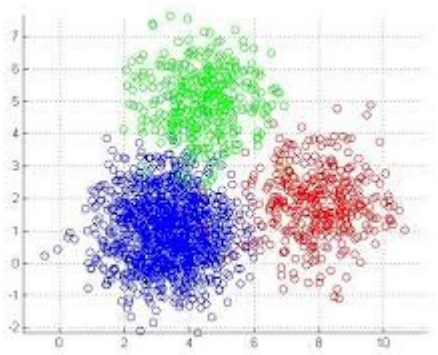
$$z_i = x_i^2$$ add quadratic term

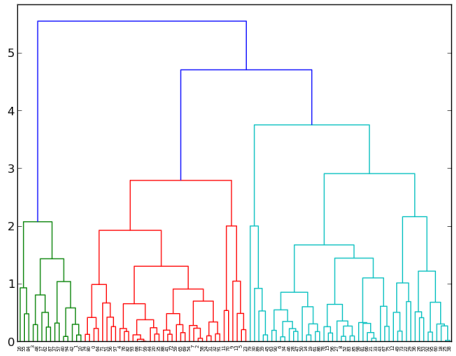
$$z_{ij} = x_i x_j$$

Step 5. Model building

Unsupervised

clustering



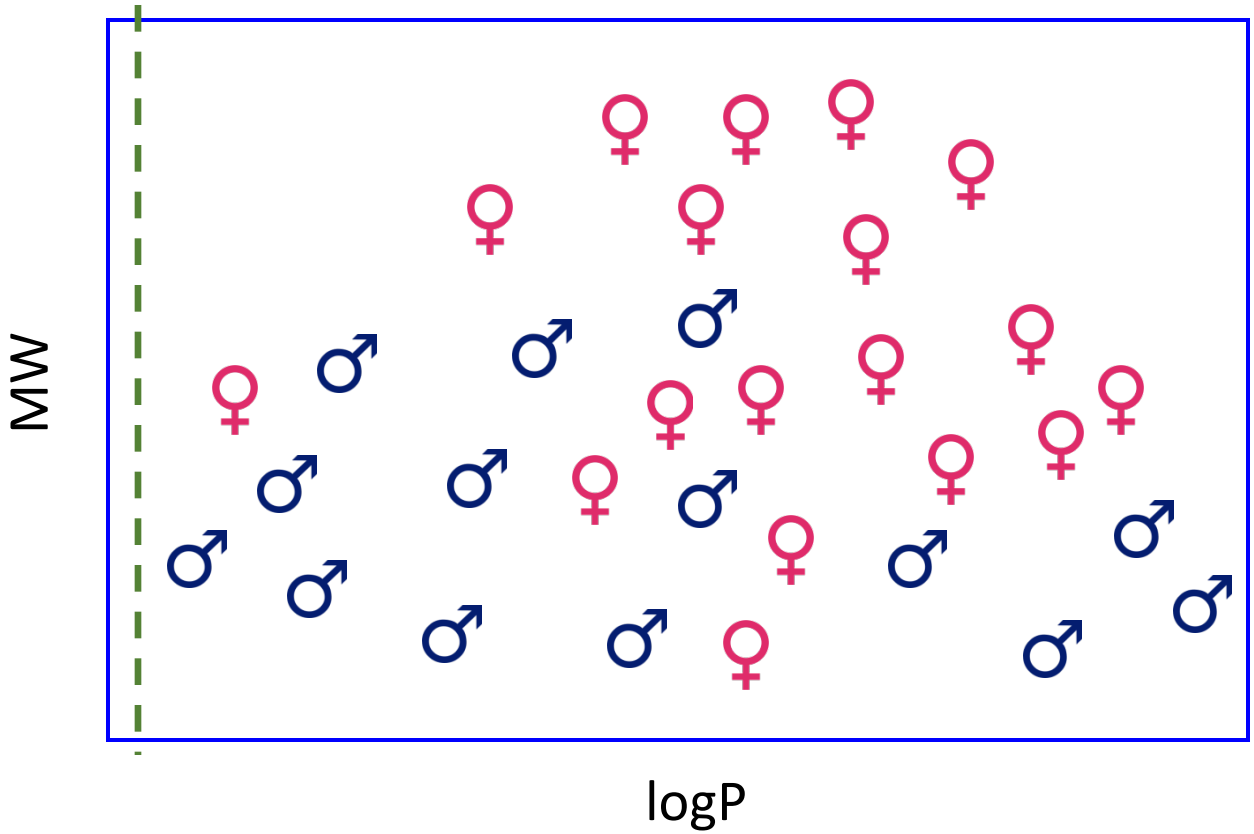


Supervised

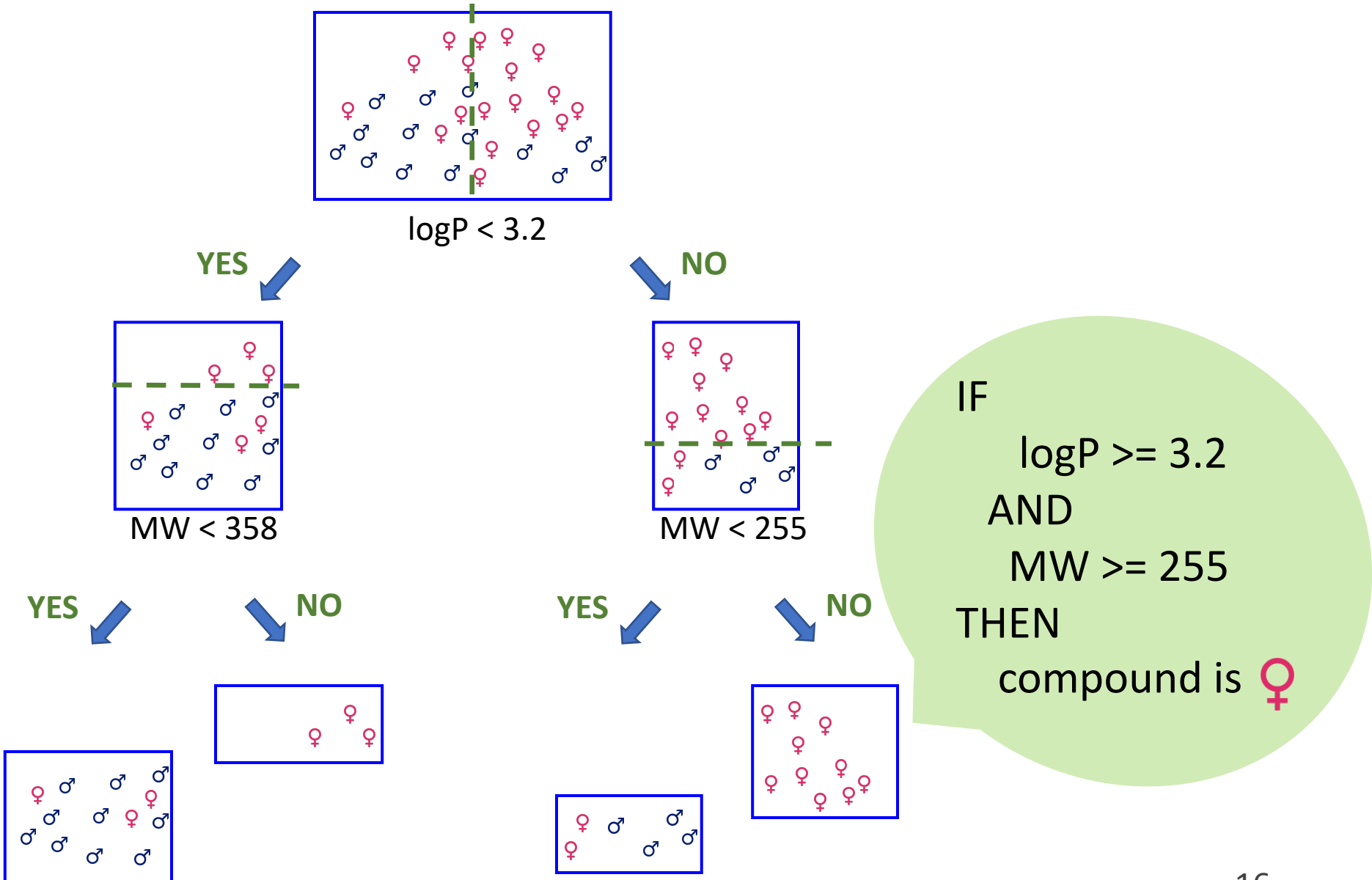
Regression	Classification
Multiple linear regression (MLR)	
Partial linear regression (PLS)	Logistic regression
Gaussian Process (GP)	Naïve Bayes (NB)
Decision trees (DT)	
Support vector machine (SVM)	
Neural nets (NN)	
Random forest (RF)	
k-Nearest neighbors (kNN)	

Decision tree

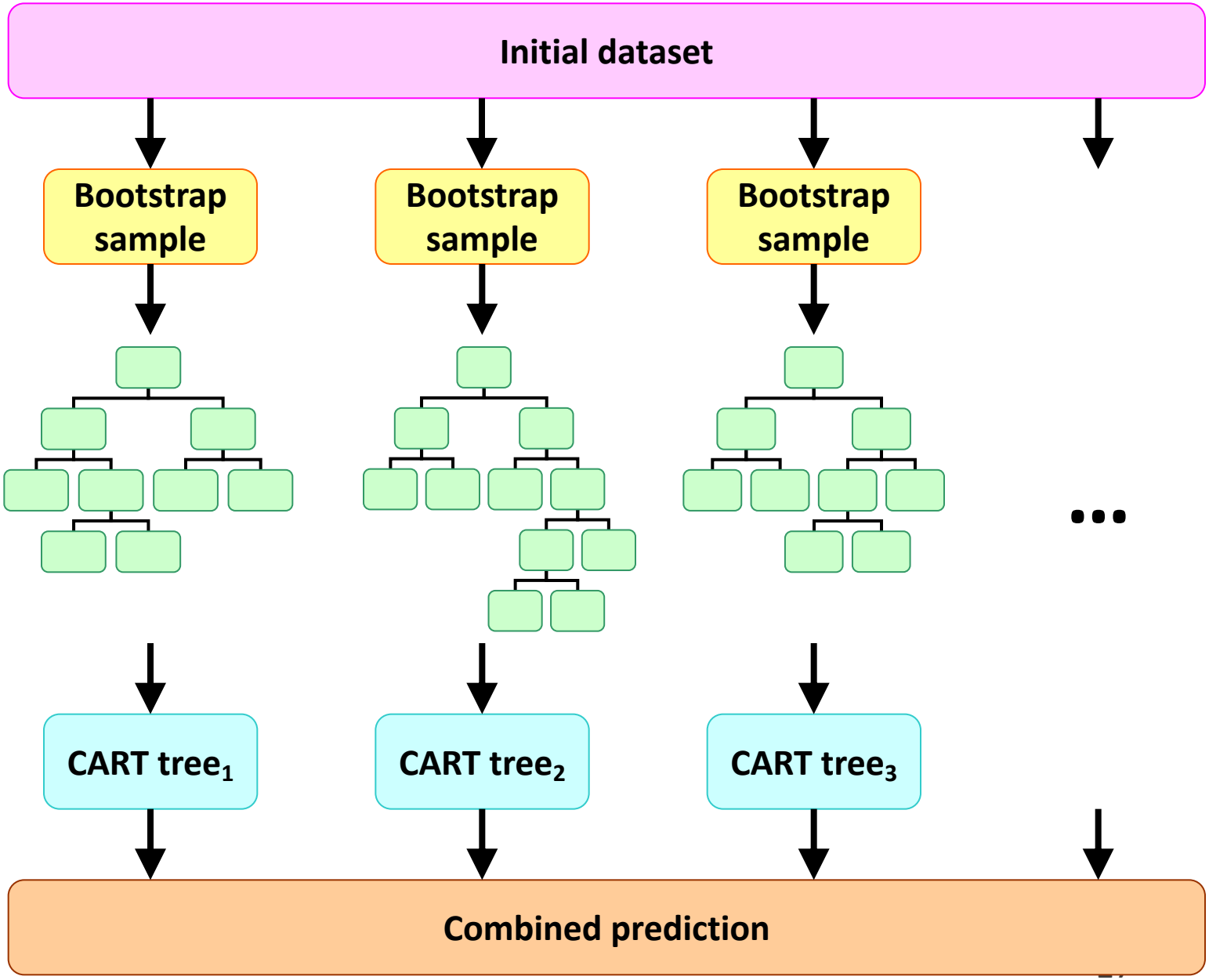
Simulated data set of actives and inactives with two descriptors – MW and logP



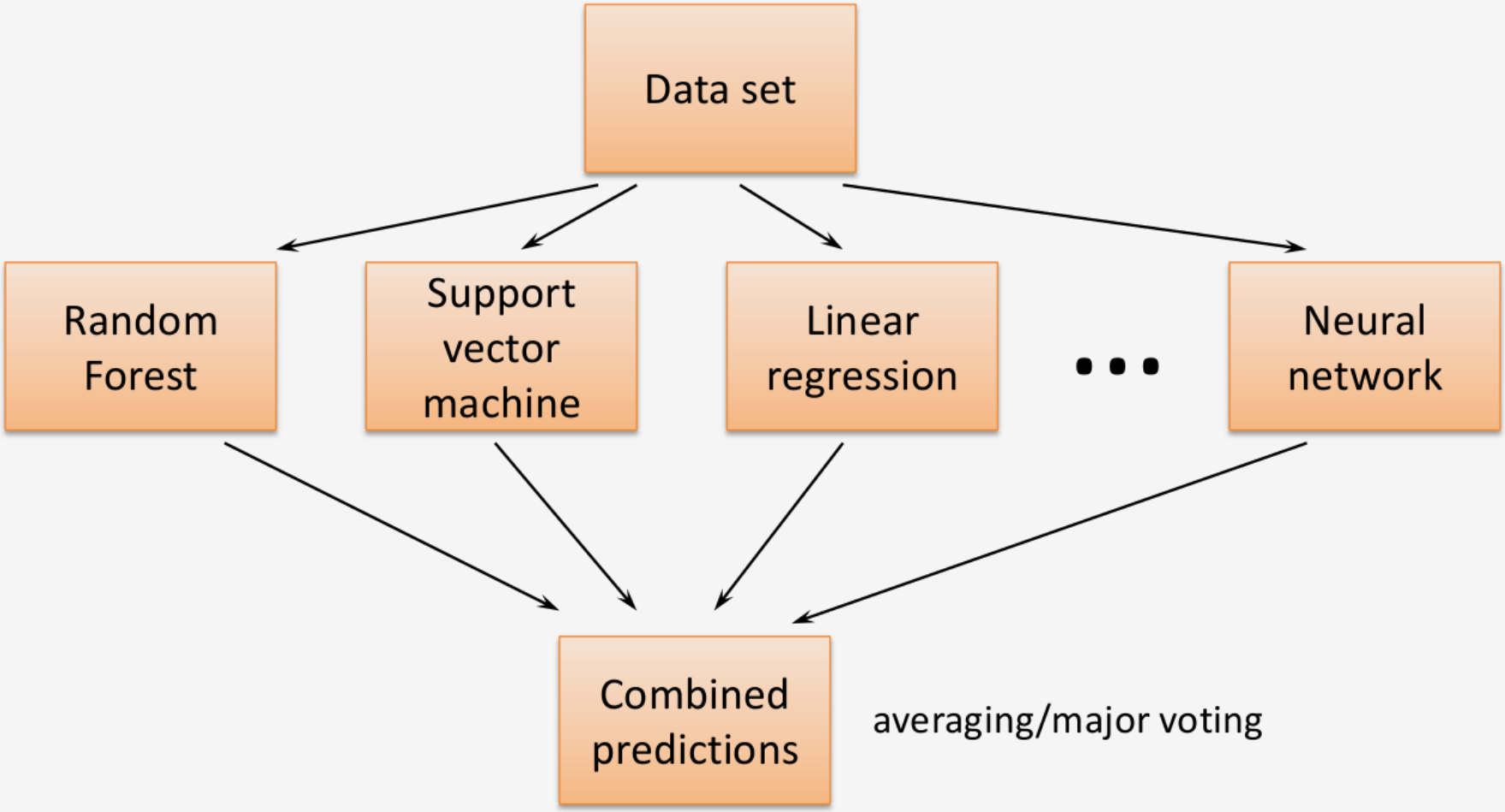
Decision tree



Random Forest



Consensus (ensemble) modeling



Models should be not correlated
(one may use different combination of descriptors and machine learning methods)

Step 6. Validation

Test set (usually 20-25% of the work set)

working set	1.2	1.3	1.7	2.0	2.2	2.8	3.1	3.2	3.2	3.6	4.7	5.7	5.8	6.4	7.2	8.1	9.0	9.1	9.2
random test set	1.2	1.3	1.7	**2.0**	2.2	2.8	3.1	3.2	**3.2**	3.6	**4.7**	5.7	5.8	**6.4**	7.2	8.1	**9.0**	9.1	9.2
stratified test set	1.2	1.3	1.7	**2.0**	2.2	2.8	3.1	**3.2**	3.2	3.6	4.7	**5.7**	5.8	6.4	7.2	**8.1**	9.0	9.1	9.2

Cross-validation

working set	1.2	1.3	1.7	2.0	2.2	2.8	3.1	3.2	3.2	3.6	4.7	5.7	5.8	6.4	7.2	8.1	9.0	9.1	9.2
fold 1	**1.2**	1.3	1.7	**2.0**	2.2	2.8	**3.1**	3.2	3.2	**3.6**	4.7	5.7	**5.8**	6.4	7.2	**8.1**	9.0	9.1	**9.2**
fold 2	1.2	**1.3**	1.7	2.0	**2.2**	2.8	3.1	**3.2**	3.2	3.6	**4.7**	5.7	5.8	**6.4**	7.2	8.1	**9.0**	9.1	9.2
fold 3	1.2	1.3	**1.7**	2.0	2.2	**2.8**	3.1	3.2	**3.2**	3.6	4.7	**5.7**	5.8	6.4	**7.2**	8.1	9.0	**9.1**	9.2

predictions of different folds are combined to calculate the final predictive measure

Step 6. Measures of predictive ability of models

Classification

Confusion matrix		Predicted	
		positive class (1)	negative class (0)
observed	positive class (1)	true positive (TP)	false negative (FN)
	negative class (0)	false positive (FP)	true negative (TN)

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{N}} \qquad [0; 1]$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \qquad [0; 1]$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad [0; 1]$$

$$\text{Balanced accuracy} = \frac{\text{Specificity}+\text{Sensitivity}}{2} \qquad [0; 1]$$

$$\text{Kappa} = \frac{\text{Accuracy} - \text{Baseline}}{1 - \text{Baseline}} \qquad [0; 1]$$

$$\text{Baseline} = \frac{(\text{TN}+\text{FP})(\text{TN}+\text{FN})+(\text{TP}+\text{FN})(\text{TP}+\text{FP})}{\text{N}^2}$$

$$\text{MCC} = \frac{\text{TP}*TN+FP*FN}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \qquad [-1; 1]$$

Step 6. Measures of predictive ability of models

Regression

Determination coefficient

$$Q^2 = 1 - \frac{\sum_i (y_{i,pred} - y_{i,obs})^2}{\sum_i (y_{i,pred} - \overline{y}_{obs})^2}$$

Root mean squared error

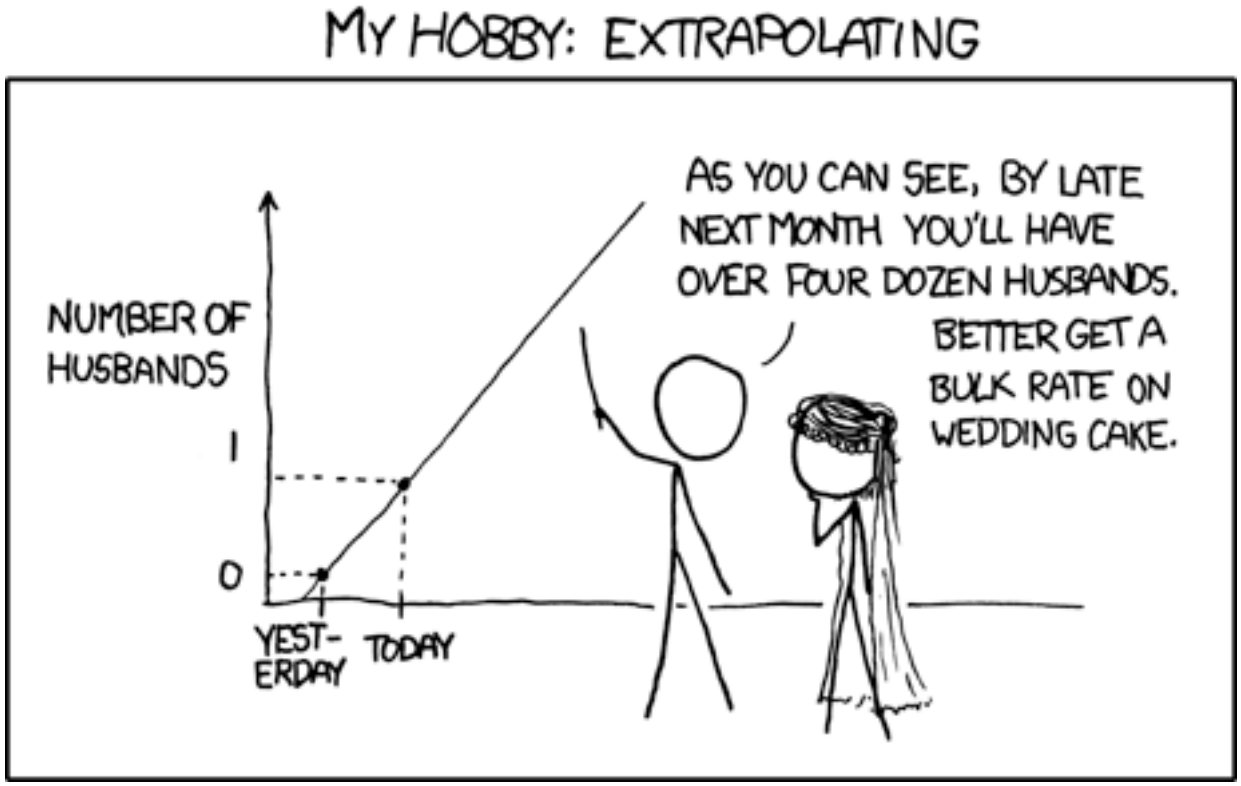
$$RMSE = \sqrt{\frac{\sum_i (y_{i,pred} - y_{i,obs})^2}{N - 1}}$$

Mean absolute error

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| y_{i,pred} - y_{i,obs} \right|$$

Step 7. Applicability domain (AD)

Extrapolation to very distant objects is dangerous



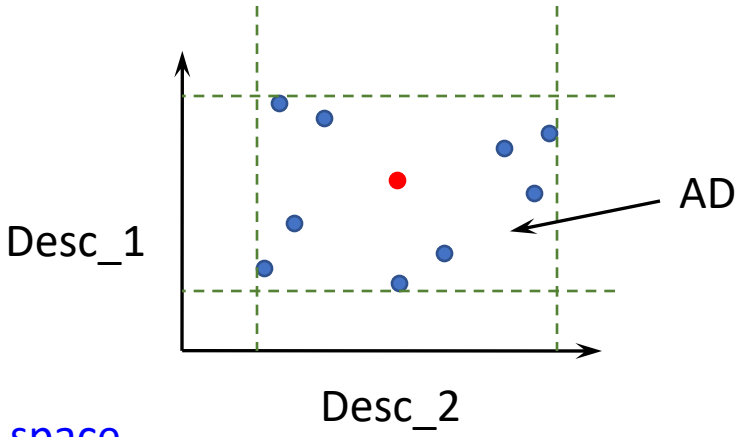
There is a need to define the domain where our model is reliable (models are not universal!)

Only compounds which are similar to the training set compounds should be included in applicability domain of the model. One should estimate similarity of new compounds (test set, etc) to the training set compounds.

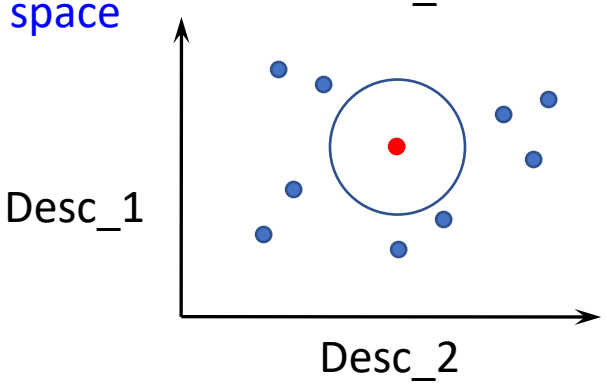
Step 7. Applicability domain (AD) measures

Bounding box - based on descriptor range

- internal regions are usually empty, especially if the number of descriptors is big
- it doesn't take into account descriptor correlation

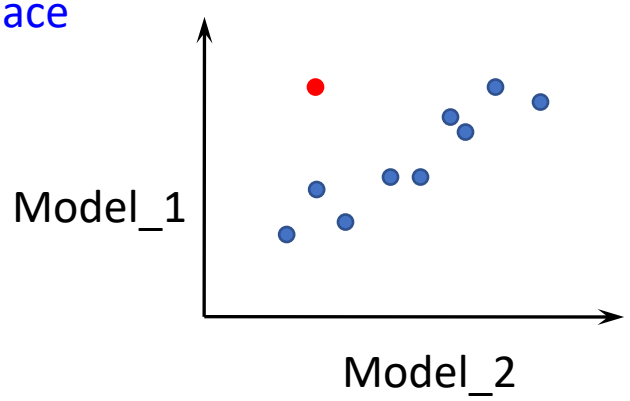


Distance from training set compounds in descriptor space



Distance from training set compounds in model space

Requires several models (e.g. consensus model, bootstrap models)



Step 8. Interpretation of QSAR models

Why interpretation is important?

Found active/inactive patterns which can be used for optimization of compound properties

Retrieve trends of stricture-activity relationships which can be used for knowledge-base model validation

Regulatory purposes

Step 8. Interpretation of QSAR models

Principles and issues

Model should be predictive

Interpretation is valid within the applicability domain of the model

Interpretation results are data set dependent

Step 8. Interpretation of QSAR models

plant growth inhibition activity of phenoxyacetic acids

$$1/C = 4.08\pi - 2.14\pi^2 + 2.78\sigma + 3.38$$

Hansch equation

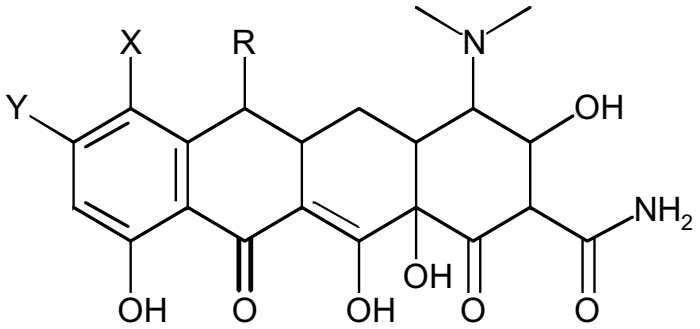
rate of penetration of membranes in the plant cell

electronic factors

$\pi = \log P_X - \log P_H$

σ - Hammet constant

Free-Wilson models



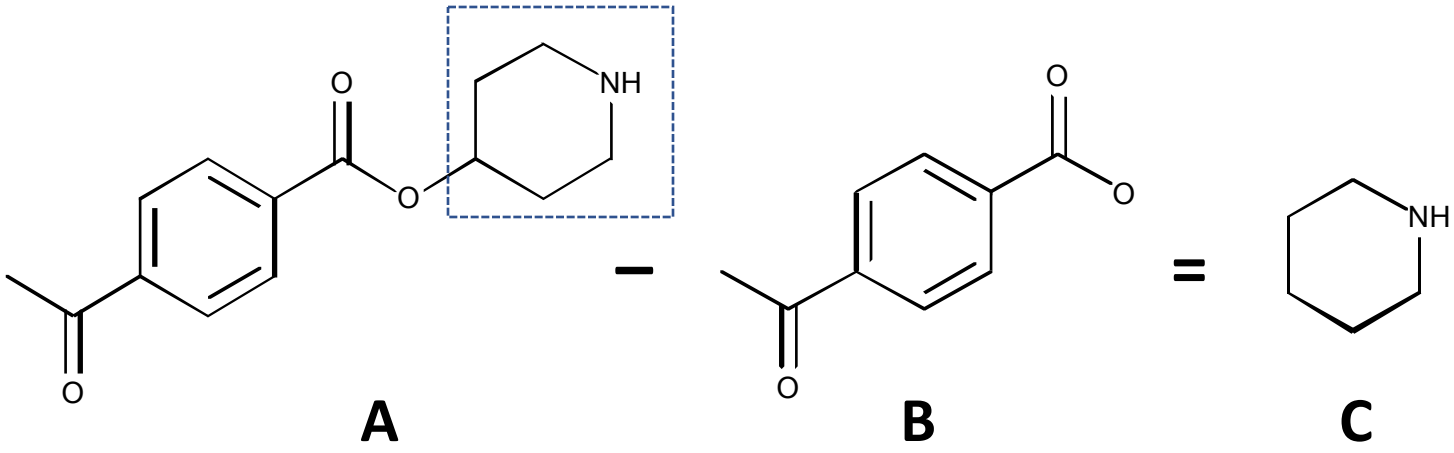
Inhibition activity of compounds against *Staphylococcus aureus*

R is H or CH_3;
X is Br, Cl, NO_2 and
Y is NO_2, NH_2, $NHC(=O)CH_3$

$$Act = 75R_H - 112R_{CH3} + 84X_{Cl} - 16X_{Br} - 26X_{NO2} + 123Y_{NH2} + 18Y_{NHC(=O)CH3} - 218Y_{NO2}$$

Step 8. Interpretation of QSAR models

Universal approach



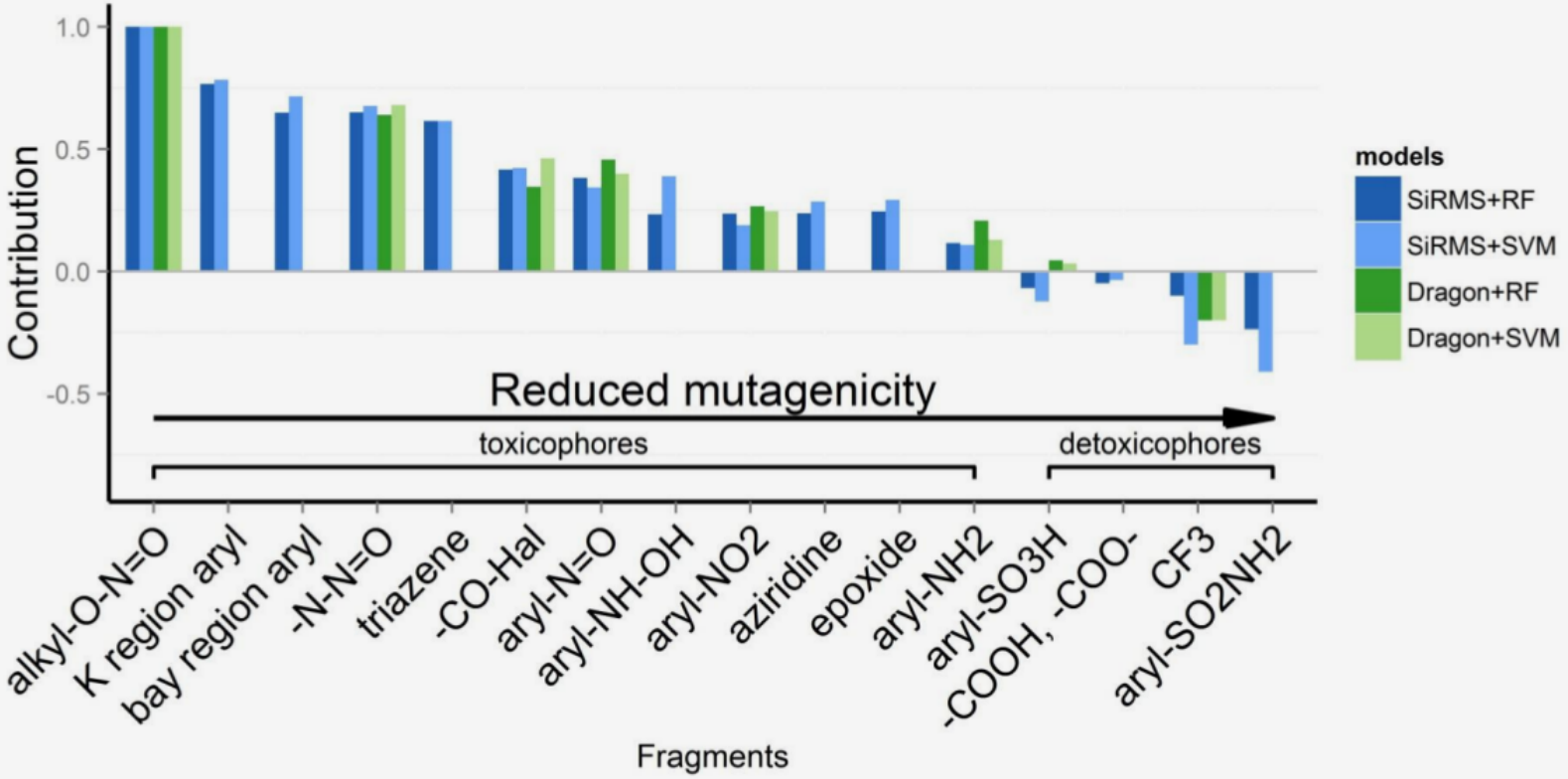
$Activity_{pred}(A)$	$Activity_{pred}(B)$	Contribution(C)
$f(A) = x$	$f(B) = y$	$W(C) = x - y$

Polishchuk, P. G.; Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Universal Approach for Structural Interpretation of QSAR/QSPR Models. *Molecular Informatics* **2013,** *32*, 843-853

Step 8. Interpretation of QSAR models

5-fold external cross validation results

Descriptors	Algorithm	Balanced Accuracy
SiRMS	RF	0.817
	SVM	0.800
Dragon	RF	0.816
	SVM	0.793



Polishchuk P.G. et al. *Molecular Informatics*, **2013**, 843-853

Step 8. Interpretation of QSAR models

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Perspective

Cite This: J. Chem. Inf. Model. 2017, 57, 2618-2639

pubs.acs.org/jcim

Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future

Pavel Polishchuk*

Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University and University Hospital in Olomouc, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic

Table 7. Applicability of Interpretation Approaches to QSAR Models

Models	Descriptors	
	interpretable	non-interpretable
linear regression	regression coefficients (Hansch, Free-Wilson)	universal structural interpretation, similarity maps, computational matched molecular pairs and series
PLS (OPLS, O2PLS, etc)	regression coefficients, X- and Y-scores, variable importance	
decision trees	logical rules	
NN	variable importance based on weights and biases, variable contributions	
RF	variable importance based on permutation, variable contributions	
NN, SVM, RF	rule extraction	
any model including consensus ones	partial derivatives, variable importance based on permutation, sensitivity analysis	
Interpretation paradigm	model → descriptors → (structure) or model → structure	model → structure

Step 8. Interpretation of QSAR models

Interpretation results of valid predictive models should converge independent of:

- interpretation approach
- descriptors
- machine learning method

All models are interpretable but not all end-points

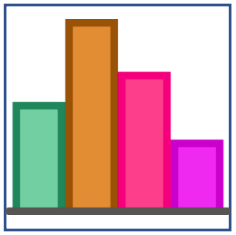
Overall QSAR workflow

Input data

Bioassays

Databases

Preprocessing



Data normalization

Feature extraction

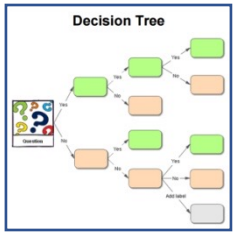
Feature engineering

$$x_i^{'} = \frac{x_i - \bar{x}}{\sum_j z_j}$$

Feature selection

Feature combination

Model learning



Classification

Regression

Clustering

Model validation

Cross-validation

Bootstrap

Test set

Applicability Domain

Interpretation