

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Applicability domain of QSAR models

Iurii Sushko

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. Langosch

Prüfer der Dissertation:

1. Univ.-Prof. Dr. H.-W. Mewes
2. Univ.-Prof. Dr. K. Suhre
(Ludwig-Maximilians-Universität München)

Die Dissertation wurde am 24.11.2010 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 17.02.2011 angenommen.

Acknowledgements

I would like to express my gratitude to my research colleagues, who were always willing to help me with an advice and gave me an opportunity to work in a friendly and creative atmosphere: Anil Pandey, Robert Körner, Sergii Novotarskyi, Matthias Rupp, Simona Kovarich, Stefan Brandmaier, Wolfram Teetz, Eva Schlosser, Vlad Kholodovych and Ahmed Abdelaziz. I thank Benoit Mathieu for his advices.

I would like to thank my thesis advisor, Dr. Igor Tetko, for his help and creativeness, for having introduced me the scientific way of thinking and for his ideas that have significantly contributed to my thesis work.

I am very grateful to my supervisor Prof. Hans-Werner Mewes for giving me an opportunity to work on my thesis in Institute of Bioinformatics and Systems Biology and for supporting my research. I am also very grateful to Prof. Karsten Suhre for his interest in my work.

I would like to thank my family back in Ukraine, my mother Valentyna Sushko, my father Alexander Sushko and my brother Ievgenii Sushko for supporting me.

Iurii Sushko

Abstract

In recent decades, computational models have gained popularity for predictions of biological activities and physicochemical properties. This new and rapidly developing field of research is referred to as QSAR/QSPR (Quantitative Structure-Activity/Property Relationship) and is especially applicable in drug design and in environmental risk assessment (ecotoxicology), where screening of large datasets of compounds is required.

The major limiting point of computational models is questionable reliability of predictions. Computational models are not guaranteed to give equally accurate predictions on the whole chemical space; in other words, the computational models have limited domain of applicability. At present, the lack of a proper definition for the applicability domain (AD) of a model is one of the major issues restraining the practical application of computational models. The problem of the AD assessment is addressed in this work.

The work introduces the methodology for the AD assessment and conveys a comprehensive benchmarking analysis of existing and new approaches. The practical AD assessment is demonstrated in a number of studies on prediction of such properties as mutagenicity (Ames test), toxicity (inhibition growth concentration), lipophilicity and cytochromes inhibition. It is shown that the AD approaches allow to estimate the prediction accuracy for every compound individually and, thereby, to discriminate highly accurate predictions with the accuracy close to that of experimental measurements. All the introduced AD methods are implemented as a part of a new platform for chemical modeling (OCHEM) and are publicly available online at <http://ochem.eu>.

Table of Contents

1 Introduction.....	1
1.1 Motivation.....	1
1.2 Thesis roadmap.....	3
2 Methodology.....	5
2.1 QSAR research.....	5
2.1.1 Overview.....	5
2.1.2 Molecular descriptors.....	5
2.1.3 Machine learning methods.....	7
2.1.4 Meta-learning techniques.....	9
A. Model ensembles and bagging.....	9
B. LIBRARY model correction.....	9
2.1.5 Validation of models.....	9
2.1.6 Prediction accuracy.....	11
A. Regression models.....	11
B. Classification models.....	12
2.1.7 Detection of statistical significance.....	12
2.1.8 Representation of molecules.....	13
2.2 Applicability domain of QSAR models.....	14
2.2.1 Basic definitions.....	14
2.2.2 Distances to models.....	15
A. Leverage	16
B. Standard deviation of the ensemble predictions (STD).....	16
C. Tanimoto similarity.....	18
D. Correlation of prediction vectors (CORREL).....	18
E. Rounding effect (CLASS-LAG)	18
F. Concordance of a classification ensemble.....	19
G. Rounding effect and standard deviation combined (STD-PROB).....	20
H. Descriptor-based and property-based DMs.....	21
2.2.3 Analysis of prediction accuracy.....	22
A. Accuracy averaging.....	22
B. Estimation of prediction accuracy.....	23
2.2.4 Comparison of applicability domains.....	24
A. Discriminative power of DM.....	24
B. Fitness of probability distribution.....	26
2.2.5 Interpretation of applicability domains.....	27
2.3 Analyzed datasets.....	28
2.3.1 Datasets of experimental measurements.....	28
A. Ames test dataset.....	28
B. T. pyriformis toxicity dataset.....	29
C. Platinum complexes lipophilicity dataset.....	29
D. CYP450 inhibitors dataset.....	30
2.3.2 Datasets of chemical compounds.....	30
A. Enamine dataset.....	30
B. EINECS dataset.....	30
C. HPV dataset.....	30
2.4 Summary.....	31

3 Online chemical modeling environment – OCHEM.....	33
3.1 Motivation.....	33
3.2 The database of experimental measurements.....	34
3.2.1 Structure overview.....	34
3.2.2 Sources of information.....	35
3.2.3 Data access and management.....	36
3.3 Modeling framework.....	37
3.3.1 Overview.....	37
3.3.2 Calculation of models.....	37
3.3.3 Descriptors.....	39
3.3.4 Conditions of experiments.....	40
3.3.5 Configuration of the machine learning methods.....	40
3.3.6 Model calculation.....	41
3.3.7 Distributed calculations.....	41
3.3.8 Analysis and management of models.....	42
3.3.9 Application of models.....	44
3.3.10 Applicability domain assessment.....	45
A. DMs and accuracy averaging.....	45
B. Estimation of the prediction accuracy.....	46
3.4 Implementation aspects.....	47
3.5 Summary and outlook.....	47
4 Benchmarking studies.....	49
4.1 Prediction of Ames mutagenicity.....	49
4.1.1 Ames test and mutagenicity.....	49
4.1.2 Methods and datasets.....	49
A. QSAR approaches.....	49
B. Applicability domain assessment.....	51
C. Benchmarking criteria.....	53
4.1.3 Results and analysis.....	54
A. Comparison of distances to model.....	54
B. Analysis of the qualitative AD measures.....	59
C. Ability to estimate the prediction accuracy.....	61
D. Interpretation of the AD.....	62
E. Data variability analysis.....	62
F. Reliability of predictions vs. variability of experimental measurements.....	64
G. Reliable predictions for ENAMINE, EINECS and HPV databases.....	66
4.1.4 Summary.....	67
4.2 Toxicity against <i>T. Pyriformis</i>	69
4.2.1 Introduction.....	69
4.2.2 Methods.....	69
A. QSAR approaches.....	69
B. Applicability domain assessment.....	71
C. Benchmarking criteria.....	73
4.2.3 Results.....	74
A. Analysis of individual models.....	74
B. Comparison of distances to models.....	76
C. Ability to estimate the prediction accuracy.....	78
D. Interpretation of the AD.....	80
E. Reliable predictions for HPV, EINECS and ENAMINE databases.....	82
4.2.4 Summary.....	84

5 Applications.....	85
5.1 Lipophilicity of Pt complexes.....	85
5.1.1 Introduction.....	85
5.1.2 Methods.....	86
A. Dataset and the variability of measurements.....	86
B. QSAR approaches and AD assessment.....	86
5.1.3 Results.....	87
A. Comparison of the QSAR approaches.....	87
B. Assessment of prediction accuracy and applicability domain.....	89
C. Interpretation of the AD.....	90
5.1.4 Summary.....	91
5.2 Cytochrome P450 inhibition.....	93
5.2.1 Introduction and methods.....	93
5.2.2 Results.....	93
A. QSAR modeling.....	93
B. AD assessment.....	93
C. Interpretation of the AD.....	94
D. Reliable predictions for HPV, EINECS and ENAMINE datasets.....	96
5.2.3 Summary.....	98
6 Discussion.....	99
A. Prediction accuracy of QSARs is variable.....	99
B. Ensembles of models improve AD assessment.....	100
C. Property-based DMs instead of descriptor-based DMs.....	101
D. Distances to models are universal.....	101
E. Which compounds are well predicted?.....	102
F. Accuracy of experimental measurements is achievable with QSARs.....	103
G. More diverse measurements for better models.....	103
7 Conclusions and outlook.....	105
List of abbreviations.....	107
Alphabetical Index.....	109
List of Figures.....	111
List of Tables.....	117
References.....	119
Appendix.....	127
Curriculum vitae.....	135
Publication record.....	137

1 Introduction

1.1 Motivation

Pharmacology has made a considerable effect on our life. Drugs contributed to the increase of the length and the quality of life of billions people; pain relievers, vaccines, anti-cancer medicines and antibiotics can inhibit or completely cure the disorders that otherwise would have lead to a strong discomfort, severe health damages or even death.

However, modern pharmacology (and drug design in particular) faces serious challenges: a typical drug takes 10-12 years from the beginning of research to the availability of the drug on the market, while many drugs fail on the early development stages. The main reason for failure of potential drugs is their toxicity and pure pharmacokinetics, so called ADME/T (Absorption, Distribution, Metabolism, Excretion and Toxicity) [1]. Revealing a pure ADME/T profile on an early stage of the drug development can filter out unsuitable compounds as early as possible and, therefore, can save significant efforts and expenses thus making the drugs available at lower costs. An extremely cheap and fast method to screen chemical compounds for ADME/T and other properties, a method that does not require experimental measurements or even synthesis of a compound, is the prediction of properties of interest with computational models.

A rapidly developing field, that deals with prediction of physicochemical and biological properties of molecules with computational models, is referred to as QSAR (Quantitative Structure-Activity Relationship¹). In recent decades, there was a significant number of studies that proved the success of the QSAR approach for prediction of various properties, such as solubility, lipophilicity, toxicity, mutagenicity [2-5]. Nowadays, QSAR has proven to be an important tool in the workflow of the modern drug design.

QSAR applications	Drug design <ul style="list-style-type: none">• screening large number of compounds on early stage of drug development• assuring the desired properties <i>before</i> the compound is synthesized		Environmental toxicity <ul style="list-style-type: none">• screening of REACH compounds• reduction of animal testing
	Activities <ul style="list-style-type: none">• protein inhibition• protein activation	Toxicity <ul style="list-style-type: none">• mutagenicity (Ames test)• growth inhibition• cytochromes inhibition	Properties <ul style="list-style-type: none">• solubility, lipophilicity• ADME properties

Figure 1.1. An overview of QSAR: the applications and predicted properties.

¹ QSAR models are sometimes distinguished for QSPR (Quantitative Structure Property Relationship) models. The first type of models deals with prediction of biological activities, whereas the latter one deals with prediction of physicochemical properties. In this work, we will use the QSAR term to denote both the types of computational models.

The QSAR applications are not limited to drug design: a new European Community Regulation on safe use of chemicals, REACH (**R**egistration, **E**valuation, **A**uthorization and **R**estriction of **C**hemical substances) strongly encourages the use of alternative methods to determine toxicity of chemical compounds, the methods that allow to avoid animal testing or costly and time-consuming experimental measurements [6]. In the REACH context, QSAR models could be used to estimate the environmental hazard of chemicals.

The major problem restraining the practical application of QSAR models is the unassessed reliability of predictions. The computational models that have a good prediction accuracy for the compounds that were used to create and validate the model are not guaranteed to perform equally good on the new dissimilar compounds. There is no universal computational model that works equally well on the whole chemical space. However, this fact is often disregarded. The application of models for prediction of new compounds is often based on intuition. At present, there is no strict set of rules for determination of whether a computational model is applicable to a particular chemical compound. The failure to specify the chemical subspace, where the model is valid and is likely to give accurate predictions, i.e. the failure to specify the area of the model applicability, is the limiting point for the practical application of computational models.

The problem of uncertainty in the accuracy and the reliability of predictions is addressed in an emerging area of research, a subdomain of QSAR called the applicability domain (AD) research. A simple example on the following figure illustrates the problem of AD assessment.

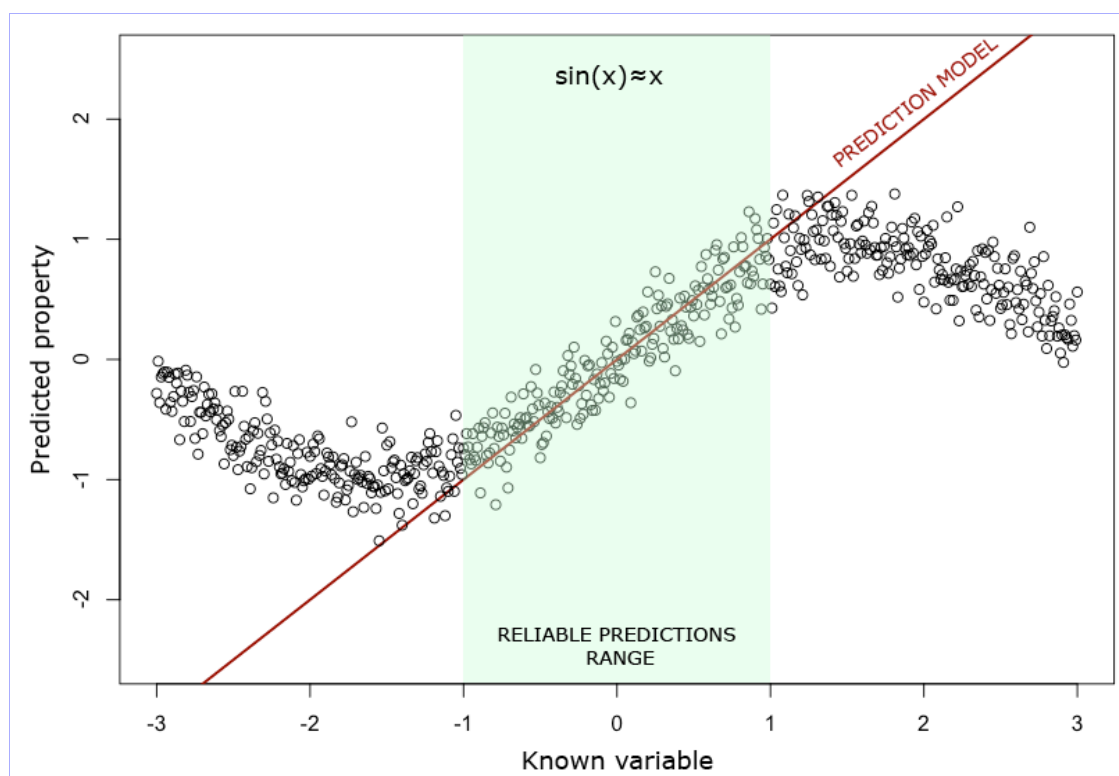


Figure 1.2. An illustrative example for the applicability domain problem. In the green region, the data are very well approximated with a linear model (red line). However, outside the green region, the approximation is not valid. Thus, the green region ([-1, 1] interval) defines the applicability domain of the linear model.

From the above figure, it is apparent that the data generated on the basis of sine function in the green region is very well approximated with a simple linear dependency

(red line). Indeed, in mathematics and physics, it is a known and widely used fact that the approximation $\sin(x) \sim x$ is accurate for small x but is completely invalid if x is large. In this example, it is of crucial importance that the application of the linear model is limited only to small x values, e.g. to the $[-1, 1]$ interval, which defines the model applicability domain.

The need for the AD assessment for QSAR models was clearly stated in the document adopted by OECD (Organization for Economic Cooperation and Development). On 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology, the OECD member countries adopted the five requirements for QSAR models used for regulatory purposes: a model should have a defined endpoint, an unambiguous algorithm, appropriate measures of goodness-of fit, a mechanistic interpretation and, importantly, a *defined domain of applicability*. Thus, an adequate AD assessment is a strict requirement for any QSAR model intended for regulatory purposes.

This work aims to create a rigid framework for determination of the model applicability domain and to verify this framework in practice using real QSAR problems. The methodology proposed in this study investigates the variability of the prediction accuracy in the chemical space and provides a set of rules that allow to determine whether a particular model can give a reliable prediction for a particular chemical compound. The proposed methods for AD assessment are verified on a number of QSAR studies for both biological and physicochemical properties such as toxicity, lipophilicity, mutagenicity and cytochrome inhibition.

Furthermore, the work addresses a second important problem – the absence of publicly available tools to perform QSAR research, estimate the applicability domain of QSAR models and publish the results. Nowadays, there are hundreds of predictive models that were published, forgotten and never used after the publication, since a significant effort is required to reproduce the published computational model. Indeed, to recreate the model, a researcher must recollect the dataset with the experimental measurements, obtain the software to calculate necessary molecular descriptors and train the model with the parameters used in the publication. Thus, it is often extremely tedious or even infeasible to reproduce the model. The work includes development of a unique novel online platform that allows both to perform QSAR research and to publish the results online. This platform, the Online Chemical Modeling Environment (OCHEM), includes the database of experimental measurements and the tools for creation of predictive models and, importantly, for estimation of applicability domains. The database includes already more than a million experimental measurements of physicochemical and biological properties and a dozen of published models. Additionally, all the AD-related methods and studies presented in this work are available in OCHEM. The developed platform is publicly available online at <http://ochem.eu>

The author hopes that the work will contribute to the practical application of reliable computational models in drug design.

1.2 Thesis roadmap

The structure of the work aims to outline the four aspects related to the problem of QSAR predictions and AD assessment: the methodology, the implementation, the benchmarking and the practical applications. The approaches introduced in the

“Methodology” section are further comprehensively benchmarked and applied for two QSAR studies in the “Benchmarking studies” and “Application” chapters, respectively. The implementation of the platform that served as the main tool for the research within this work is outlined in the “OCHEM – Online Chemical Modeling Environment” chapter.

The results are summarized in the “Discussion” chapter and presented as a number of key points.

2 Methodology

This chapter introduces the basic definitions, terminology and methodology of the QSAR research in general and of the applicability domain research in particular. The concepts and methods introduced here will be extensively used in the further chapters of the work.

2.1 QSAR research

2.1.1 Overview

The domain of QSAR, Quantitative Structure-Activity Relationship, encompasses computational predictions of various biological activities and physicochemical properties of molecules. The main assumption of QSAR is that *similar molecules have similar properties*. In other words, a “small” modification of molecular structure results into a “small” change of its biological activities and physicochemical properties.

An important question in QSAR studies is how to define the similarity of chemical compounds. When we define whether two molecules are similar, we may address e.g. similarities of the molecular graphs, presence or a number of particular functional groups, similarities of the shapes of 3D structures, similarities of the polar surfaces, etc. Speaking mathematically, the definition of similarity corresponds to the definition of a metrics in the space of chemical compounds. Thus, there is a vast variety of methods on how the similarity of molecules can be defined; the choice of an appropriate metrics is the key success point for QSAR predictions. Generally, the similarity of molecules is defined by representing a chemical compound with a set of numerical features, referred to as *molecular descriptors*. Simple examples of molecular descriptors are the molecular weight, the number of atoms of a particular type, the number of aromatic rings etc. Various types of molecular descriptors are described further in Section 2.1.2.

In QSAR, prediction of a particular biological activity of a physicochemical property is based on the information about the known property values for a set of molecules referred to as the *training set*, which usually contains the results of experimental measurements. This point is important and specific for QSAR: the prediction is not based solely on the basic laws of physics and chemistry (indeed, it is very difficult to explain particular property in terms of the essential laws such as laws of quantum chemistry), but uses information from a training set for prediction of new compounds, whose property values are yet unknown. The process of creation of such predictive model is referred to as supervised learning and is based on the machine learning methods described further in Section 2.1.3

2.1.2 Molecular descriptors

All machine learning methods are abstract mathematical methods; they operate with a particular numerical representation of chemical compounds. For most machine

learning methods, to build a predictive computational model, the chemical compounds must be ultimately represented as a set (a vector) of numerical features. These numbers, which describe a chemical compound in the mathematical way, are referred to as *molecular descriptors*. Obviously, molecules are complex objects and can be represented by numbers in a virtually unlimited number of ways. Thus, there is a need for an optimal choice of descriptors, specific for a particular prediction problem.

At present, there exist numerous types of molecular descriptors and their software implementations. Generally, molecular descriptors can be categorized as follows:

- linear or 1D (one-dimensional) descriptors, such as molecular weight, number of particular types of atoms or functional groups, number of fragments etc
- 2D descriptors, based on the graph of the molecular structure
- 3D descriptors, based on three-dimensional structure of a molecule. Such descriptors require to calculate the optimized stable 3D conformation(s) of a molecule

A more complete and elaborate review of different molecular descriptors can be found in an excellent book edited by Todeschini [7]. Here, we overview the descriptors used for the studies that are a part of this work.

E-State descriptors (electrotopological state descriptors). For every atom in a compound, E-State descriptors combine the information about electron richness (electronegativity) and the topological information. A detailed information on E-State indices can be found in the work by Hall and Kier [8]. These descriptors were used for most of the studies encompassed within this work and proved to provide good results for various prediction problems of both chemical and biological properties.

Molecular fragments counts (MFC). To calculate MFC descriptors, a molecule is split into sub-fragments of a particular size (for example 2-5 atoms in a sub-fragment). Thereafter, the appearances of every fragment in a molecule are counted. The software to calculate MFC descriptors used in our research was ISIDA Fragmentor utility[9]; therefore, these descriptors are also referred to ISIDA fragments.

LogP and LogS values (log of the octanol-water partition coefficient and aqueous solubility) are important in QSAR modeling, since they implicitly affect many other physicochemical and biological properties of molecules. Therefore, LogP and LogS are of particular interest as molecular descriptors. Most often, experimentally measured values for molecules used in modeling are not available and are substituted with their predicted values. To obtain predictions for LogP and LogS, we used ALogPS software. This program was recently top-ranked amid 18 competitors for logP prediction using > 96,000 *in house* molecules from Pfizer and Nycomed [10]. It was also reported to be “the best available off-the-shelf package for intrinsic aqueous solubility prediction” at F. Hoffmann-La Roche [11].

Dragon descriptors are named after the software developed by the group of Prof. Todeschini [12]. These descriptors encompass a vast variety of 1D, 2D and 3D descriptors separated into 20 logical blocks. Covering a vast variety of descriptor types, the Dragon descriptors are very popular and are often used for QSAR modeling of various properties. These descriptors were used by a number of the research groups that contributed their QSAR models investigated in the “Benchmarking studies” chapter.

2.1.3 Machine learning methods

The machine learning methods typically used for QSAR predictions are based on supervised learning. These methods, given the knowledge about a particular property or activity from a training set with known (usually measured experimentally) values of the predicted property, aim to generalize this knowledge in order to predict the property for new chemical compounds, for which the property is unknown. The most commonly used machine learning methods include: k-nearest neighbors, simple and ridge linear regression, neural networks and support vector machines (SVM). Additionally, the linear methods can be generalized using kernel techniques, which map the initial space of molecular descriptors to a space with a higher number of dimensions and, thereby, allow to apply linear methods to non-linear prediction problems. Kernel extensions of machine learning methods are very popular in QSAR modeling. A detailed description of the kernel techniques can be found in literature [13,14].

Here follows a brief summary of the aforementioned machine learning methods. A physicochemical property or biological property that needs to be predicted will be referred to as *target property*. We will use capital letter J for denoting a chemical compound, $x_i(J)$ – for denoting the i -th descriptor of a compound J , $y(J)$ and $\tilde{y}(J)$ for real and predicted values of the target property, M and N – for denoting the number of the used molecular descriptors and the number of molecules in the training set, respectively.

Linear regression is one of the simplest methods, which estimates a property as a linear combination of the input variables. The target property is calculated as

$$\tilde{y}(J) = w_0 + \sum_{i=1}^M w_i \cdot x_i(J) \quad (2.1)$$

where w_0 is the shift and w_i are the regression weights of each molecular descriptor. A linear model is completely described by the vector $w = \{w_0 \mid i = 0..N\}$, which is chosen to minimize the sum of squares of the prediction errors on the training set:

$$f(\vec{w}) = \sum_{i=1}^N (\tilde{y}_i(J_i, \vec{w}) - y(J_i))^2 \rightarrow \min \quad (2.2)$$

where $\tilde{y}_i(J_i, \vec{w})$ is the prediction value for compound J_i by the model defined by the regression weights \vec{w}

Ridge linear regression is a generalization of the linear regression, which minimizes a modified target function:

$$f(\vec{w}) = l \cdot |\vec{w}| + \sum_{i=1}^N (\tilde{y}_i(J_i, \vec{w}) - y(J_i))^2 \rightarrow \min \quad (2.3)$$

where the term l is an optimizable parameter. Because the term $l \cdot |\vec{w}|$ is also minimized, the absolute values of optimized regression weights tend to be reasonably low and the method is more stable than the simple linear regression, which minimizes only the sum of error squares (Expression 2.2) and does not have any limit on the regression weights. The simple linear regression involves inversion of a multivariate matrix, which can be ill-conditioned and thus result into the large values of the regression weights and, therefore, into a more complex model that has poor generalization ability. The $l \cdot |\vec{w}|$ component present in the optimization function of the ridge regression addresses the problem of inversion of the ill-conditioned matrix and helps to obtain a simpler model with a possibly lower fitting score on the training set but with a higher predictive ability.

K-nearest neighbors method (KNN) uses the mean of the target property values for K compounds, nearest to the predicted compound. The K nearest compounds are determined using any metric (usually the Euclidian distance) in the space of molecular descriptors. The advantage of this method is its simplicity and absence of any training process (except of selection of the optimal K - number of nearest neighbors). A model merely needs to store the molecular descriptors for the reference compounds from the training set.

Neural networks, or, more specifically, multilayered perceptrons, use a simplified mathematical model of a biological neuron, defined by Expression 2.4, to predict the target property. More precisely, a number of neurons is organized in consequent layers, where an input of every neuron from the next layer is an output of a neuron from the previous layer.

$$y(x_1, \dots, x_n) = f\left(\sum_{i=1}^n w_i \cdot x_i\right) \quad (2.4)$$

where $x_1..x_n$ are inputs, w_i are weights of a neuron and f is a non-linear response function. A neural network is completely defined by the set of neuron weights $W = \{w_{ij}, i = 1..L, j = 1..N_i\}$, where L is the number of layers, and N_i is the number of neurons in i -th layer. During the training of a network, the weights are optimized to minimize the sum of squares of errors (similarly to linear models, as in Expression 2.2). There are numerous methods for training of neural networks varying on the quality and the calculation speed. In this work, we used the Levenberg-Marquardt method, which is relatively slow and computationally demanding but allows extracting most of the information from the training set [15,16].

If a neural network is modified using the LIBRARY correction (explained in the next section), it is referred to as associative neural network (ASNN) [15]. Associative neural networks are extensively used in all the studies encompassed in this work.

Support vector machines (SVM) is the method originally intended for classification problem. SVM constructs a descriptor space hyperplane that separates the training set samples into two classes. The hyperplane is chosen to provide the maximum margin (i.e., maximum distance from the hyperplane to the samples of either class). The algorithm for constructing such a hyperplane is based on quadratic programming and can be found in the work by Boser and Vapnik [17]. In case of more than two classes, multiple hyperplanes are constructed.

If the separation of the classes by a hyperplane is impossible, the “soft margin” modification is used. This modification allows misclassified samples but includes a penalty component for them [18].

In QSAR modeling, SVM is often used with the kernel modification [14]. In brief, SVM is performed not in the original space but in the *feature space* obtained via a non-linear kernel transformation of the original space. The feature space has a higher dimensionality (often, it has an infinite number of dimensions), which makes it possible to separate classes that were not linearly separable in the original non-transformed space.

The algorithm was also generalized for regression problems [19]; in this case, it is referred to as support vector regression (SVR).

2.1.4 Meta-learning techniques

A. Model ensembles and bagging

It has been shown [20] that, instead of building a single model, it is often favorable to build a set of different models (referred to as model *ensemble*) and use the average value of their predictions to get a better performance. Such “average” model is referred to as *consensus model*. Obviously, the models within an ensemble should be different; to get different models based on the same training set, one can train multiple copies (often 100) of the same model with different training subsets replicated randomly on the basis of the complete training set. The replication of multiple training sets can be done using sampling with replacement. This approach, referred to as *bagging* (bootstrap aggregating)[20], was shown to improve the prediction accuracy in comparison to using single models. The effect of averaging of multiple predictions given by model ensembles is investigated in the benchmarking studies described in Chapter 4 of this work.

B. LIBRARY model correction

Given a predictive model and an additional set of new experimental measurements (referred to as “library”), it is possible to correct the model by taking into account these measurements. The process is called LIBRARY model correction (because we complement a model with a “library” of experimental measurements) and is performed as follows.

To obtain a corrected prediction for a molecule J , we calculate the original (non-corrected) prediction $\tilde{y}(J)$ given by the original model and find K molecules from the “library” $\{J_i, i = 1..K\}$ nearest to the molecule being predicted. The nearest compounds are defined by the correlation coefficient in space of model predictions. Then, we calculate the expected residual for the molecule J as the average residual for the K nearest compounds according to the following expression:

$$\Delta(J) = \frac{\sum_{i=1}^K (\tilde{y}(J_i) - y(J_i))}{K} \quad (2.5)$$

Finally, we correct the original prediction by subtracting the average residual:

$$\tilde{y}_{corr}(J) = \tilde{y}(J) - \Delta(J) \quad (2.6)$$

Thus, the LIBRARY correction assumes that for a new compound a model will behave similarly to compounds from the “library”. Such technique is especially useful in case if retraining of the original model is infeasible due to high computational complexity or unavailability of the original training set. The LIBRARY technique was introduced by Tetko and was shown to significantly increase the prediction accuracy for the lipophilicity and distribution coefficient models [10,21-23].

2.1.5 Validation of models

To estimate the accuracy of computational models, it is unreasonable to test a model on the data that were used for the model training. In fact, this approach can be misleading, because, given a sufficient number of the input variables (molecular

descriptors), it is possible to achieve perfect predictions on the training set. Such a model would merely “remember” the property for the compounds from the training set, thus providing the perfect prediction accuracy on the training set but having low predictive ability for other compounds. In machine learning, this type of models is referred to as *over-fitted* models. Over-fitted models have no practical use and should be avoided. Therefore, it is of crucial importance to estimate the real predictive ability of models.

There are several techniques typically used to estimate the predictive ability of a QSAR model. The simplest and most straight-forward method is to split the available data with experimental measurements into two parts: the training set and the validation set (also called external set). An improved validation approach, which allows to estimate a model's predictive ability for the whole dataset, is the *cross-validation*, an approach, where the data is randomly split into N (often 5) folds, one of which is used for validation while the others – for training of the model. Thus, N models are built, where each of the model has one of the folds “reserved” for validation. A special case of cross-validation is *Leave One Out* (LOO) validation, where the number of folds is equal to the number of compounds in a set; thus, in LOO, each of the validation models excludes one compound from the training set for validation purposes.

The validation of the whole dataset can also be done using the bagging approach, described above in Section 2.1.4. In more detail, when randomly replicating multiple copies (often 100) of the training set, in each replication case, a part of molecules will not be present in the training set and, therefore, can be used for the validation purposes. Having a sufficiently large number of validation sets, every molecules will be present in at least one (but usually more then one) validation set. In fact, given that the dataset is sufficiently large, every compound is assigned to the validation set with about 37% probability, which means that on average there will be 37 validation predictions for each compound. The average prediction values from these validation sets are used to estimate the predictive ability on the whole set.

All the three validation methods (external validation set, cross-validation and bagging-validation) can be combined. In QSAR, the most commonly used technique is cross-validation. However, in some cases the bagging validation is favorable since it (a) can improve prediction accuracy for new compounds [20] and (b) it provides multiple predictions, which can be used to calculate the statistical information for estimating the AD of the model. The main disadvantage of the bagging technique is its high computational complexity: it requires to train 100 models, whereas the cross-validation requires only 5-10 models. An external validation set is the optimal choice for testing a model's performance with compounds not present the training set. This method is often used for historical reasons, since it is considered a robust technique to test a model on compounds, not used for the model training. However, in many cases N-fold cross-validation is favorable since it uses the same principle but allows to estimate the model performance for the whole available dataset.

Whatever validation method is used, it is crucially important to ensure that the same compounds are not present in the training and validation sets simultaneously; ignoring this rule can provide misleading results, which do not estimate the real predictive ability of a model.

The correct validation is ensured in the implementation of the OCHEM platform, introduced further in the Chapter 3.

2.1.6 Prediction accuracy

To estimate the performance and applicability of the model, the prediction accuracy must be quantified. There are several numerical measures for the prediction accuracy, the most commonly used ones are described below. In the following expressions, N denotes the number of compounds in the set, for which the prediction accuracy is estimated; \tilde{y}_i and y_i denote predicted and real values of the predicted property for i -th compound in the set; $E(\tilde{y})$ and $E(y)$ are the means of the predicted and real property values; sigma (σ) denotes the standard deviation.

The measures of the prediction accuracy are different for regression and classification models.

A. Regression models

Root mean square error (RMSE) is calculated accordingly to the expression:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\tilde{y}_i - y_i)^2}{N}} \quad (2.7)$$

Mean absolute error (MAE) is calculated as follows:

$$MAE = \frac{\sum_{i=1}^N |\tilde{y}_i - y_i|}{N} \quad (2.8)$$

R-square (r^2 , square of the Pearson correlation coefficient) shows how well the variations of predictions are explained by the variations of actual values of a property. More precisely, the R-square is calculated as follows:

$$r^2 = \frac{\sum_{i=1}^N (\tilde{y}_i - E(\tilde{y}))(y_i - E(y))}{\sigma(\tilde{y}) \cdot \sigma(y)} \quad (2.9)$$

where $\sigma(\tilde{y})$ and $\sigma(y)$ are the standard deviations of predicted and observed property values in the investigated dataset. The ideal case is $r^2 = 1$, which signals a direct linear dependency between the predicted and observed values but, however, does not guarantee a perfect model, since r^2 does not take into account bias.

Coefficient of determination (denoted as q^2) is often reported in QSAR publications complementary to r^2 , it is calculated as follows:

$$q^2 = 1 - \frac{RMSE}{\sigma(y)} = 1 - \frac{\sum_{i=1}^N (y_i - \tilde{y}_i)^2}{\sum_{i=1}^N (y_i - E(y))^2} \quad (2.10)$$

q^2 can be interpreted as the percentage of the variance in the property, explained by the model. In case of a “mean” model, which always gives the mean value $E(y)$ as prediction, the q^2 is zero, whereas in case of a perfect error-free model (and only in this case), the q^2 equals to one. Theoretically, if the investigated model is less accurate than the “mean” model, q^2 can be negative.

The aforementioned accuracy measures, RMSE, MAE, r^2 and q^2 , describe the

model performance in different ways and should be used together. For example, MAE and RMSE are similar, but RMSE is more sensitive to high residuals. Because of being dimensionless, r^2 is convenient to compare performances of completely different models, even if they predict different properties. Indeed, r^2 ranges in the $[0, 1]$ interval and has a universal scale, whereas there is no universal scale for RMSE and MAE. However, the two latter measures provide a more meaningful interpretation, because they are measured in the units of the predicted property. In particular, the RMSE and MAE can be used to calculate the confidence interval.

B. Classification models

For classification problems, a commonly used measure of prediction accuracy is **Correct Classification Rate (CCR)**, which is simply the percentage of compounds correctly classified by a model. CCR (denoted as η) is appropriate for balanced sets, i.e. sets that have approximately same numbers of compounds from each class.

Sensitivity and **specificity** are measures of the prediction accuracy for binary classification problems that divide all compounds into two classes, “positive” and “negative” compounds. The properties that give rise to such classes are, for example, mutagenicity (positive compounds are mutagens), blood-brain permeability (compounds, that pass the barrier are positive), etc. Sensitivity is the percentage of actually positive compounds that are predicted as positive, whereas specificity is the percentage of actually negative compounds that are predicted as negative. A 100% sensitive model never misses an actually positive compound, but can give false positives. On the contrary, a 100% specific model will never give false positives, but can miss an actual positive and report it as negative.

The prediction accuracy measures used in most of the studies introduced further in this work are RMSE (for regression models) and CCR (for classification models).

2.1.7 Detection of statistical significance

Comparison of different computational approaches is usually based on some numerical criteria (e.g., RMSE, sensitivity, specificity etc). For example, to compare performances of QSAR models, we usually compare their prediction accuracies, as described in the section above. However, if we determine the superiority of one method over another one according to some criteria, we also have to check whether this superiority is not caused by a mere chance, but is significant in the statistical sense.

The commonly used approach for checking for significant difference involves determination of two hypotheses: the null-hypothesis, which claims that the performance of the two methods is same, and the alternative hypothesis, which claims the presence of a difference in performance. The approach is to calculate the probability of the null hypothesis given the data on hand and compare this probability (referred to as p -value) with a predefined level of significance (usually 0.05). If the p -value is less than the needed level of significance, the null -hypothesis is rejected and the alternative hypothesis (claiming that the difference is statistically significant) is accepted. In contrary, if the p -value is more than the level of significance, the null -hypothesis cannot be rejected, and we cannot claim the statistically significant difference.

There are several tests based on p -values subdivided into two main categories: parametric and non-parametric tests. Parametric tests rely on particular assumptions about the distribution of the analyzed data. Often, the data is supposed to be normally

distributed. For our purposes, we used the non-parametric tests that have fewer assumptions and, therefore, they are more universal and can be used in a wider context.

The main statistical test used in this work was the so called bootstrap test.

The bootstrap test is based on construction of many multiple replications of the same dataset by sampling with replacement. For example, to compare RMSE of two predictive models according to the bootstrap test, we sample N replicas from the original validation dataset, calculate RMSE of the models for all N replicated datasets and compare N pairs of RMSE values. If RMSE of the first method is less than that of the second method in 95% or more cases, than the first method is “better with 0.05 level of significance”.

2.1.8 Representation of molecules

In literature, a molecule is normally referenced by a name and represented by a 2D depiction, which is not suitable for computational purposes. In QSAR, the commonly used formats for representation of molecules are SMILES, SD-files (SDFs), MOL2-files and InChi.

SMILES (Simplified Molecular Input Line Entry Specification) is a compact, convenient and human-readable format of molecules. In this format, every molecule is represented as a single line. A molecule can have several different SMILES representations; however, the canonical SMILES, generated according to a specific rule, is always unique for a molecule. The SMILES is an open standard and the specification is available online at <http://www.opensmiles.org/spec/open-smiles.html>. The advantage of SMILES is its simplicity and being human-readable; the disadvantage is its inability to represent 3D structures and any supplementary information. The SMILES specification can be used to represent stereochemistry, but it does not keep 3D coordinates. Therefore, in SMILES format, the information about the exact conformation is lost.

Compound	SMILES code	Notes
Ethanol	CC	Hydrogens are implicit
Urea	C(=O)(N)N	brackets denote a branch in the graph
Benzene	c1ccccc1	small «c» denotes a carbon in an aromatic bond «1» denotes a connection in a cycle

Table 2.1. Examples of SMILES codes for simple molecules.

SDF (Structure Data Format), in comparison to SMILES, is more complex to read and generate for a human, but it provides the important possibility to specify 3D structure of molecule. Additionally, the SDF format allows to provide supplementary information, such as the name of a molecule and other properties. A more detailed description of the SDF format can be found at the EPA web-site <http://www.epa.gov/ncct/dsstox/MoreonSDF.html>.

MOL2 is a molecule format that is similar to SDF by contents but additionally provides a possibility to store partial charges of every atom. This format was developed by Tripos company for the SYBYL software and has gained popularity in chemoinformatics and in QSAR research. The format specification of this format can be found at <http://www.tripos.com/data/support/mol2.pdf>.

InChi is a relatively new canonical representation of molecules, which has gained popularity in the recent years [24,25]. A complement to this format, InChi-Key, is a 14

character hash-code, that is unique for every compound; that is, it does not depend on 3D coordinates, numbering of atoms and a particular way of depicting a compound. Thus, InChi-Key can serve as a canonical identifier of a molecule in databases and can be used to group same molecules, for example for purposes of a correct model validation. In particular, InChi-Key was used as the identifier for molecules in our modeling platform described in the third chapter of this work.

2.2 Applicability domain of QSAR models

2.2.1 Basic definitions

The general definition of applicability domain (AD) was formulated by Netzeva and colleagues within the 52th workshop of the European Centre for the Validation of Alternative Methods (ECVAM)[26]: “*The applicability domain of a QSAR model is the response and chemical structure space in which the model makes predictions with a given reliability*”.

To assess the applicability domain of the model, this work addresses another, more general problem, the assessment of the accuracy for every prediction. Typically, to assess the prediction accuracy, a QSAR model is validated against an external validation set (or a cross-validated training set) and the average prediction accuracy on this set is reported as the ultimate indicator of the model performance. There is a significant flaw in this reasoning, since the performance of the model on compounds that are dissimilar from the training and validation sets is likely to be different from the estimated accuracy. Moreover, even within the validation set, the accuracy may be inhomogeneous and variable. There can be clusters of compounds that are predicted with an accuracy that is significantly higher (or lower) than the average accuracy. Thus, considering only the average prediction accuracy for an inhomogeneous set does not reflect the complete information on the performance of the model and, therefore, can be misleading.

If we could find a way to discriminate predictions of high and low accuracy or, more generally, if we could estimate the prediction accuracy for every particular compound, then we would automatically assess AD using the simple rule: a compound is inside AD if its estimated prediction accuracy is within a predefined threshold and outside AD otherwise.

A synthetic example. To illustrate the idea of the accuracy discrimination, let us consider a simple synthetic example and predict an imaginary property y that is linearly dependent on the descriptor x_1 but has also an amount of noise that is partially dependent on the descriptor x_2 :

$$y = a \cdot x_1 + c + N(0, \sigma_1) + x_2 \cdot N(0, \sigma_2), \text{ where } x_2 > 0, \sigma_2 > \sigma_1 > 0 \quad (2.11)$$

where a and c are some constants, $N(0, \sigma_1)$ is the simulated normally distributed background noise and $N(0, \sigma_2)$ is the noise that depends on chemical structures. Given that the noise is unknown and unpredictable, the best model for prediction of the property y on basis of the descriptors $\{x_1, x_2\}$ is a simple linear model:

$$y = a \cdot x_1 + c \quad (2.12)$$

It is obvious that the compounds having lower values of x_2 will be predicted better because of a less amount of noise represented by the component $x_2 \cdot N(0, \sigma_2)$ in Expression 2.11. Indeed, after we generated 1,000 input samples on basis of (2.11) and

predicted it with the model (2.12), we observed that RMSE for compounds with $x_2 < 0.3$ was 0.51, for compounds with $x_2 > 0.3$, the RMSE was 0.77, whereas the average RMSE for all the compounds was 0.72. The scatter plots for the compounds with a higher (the green dots) and a lower (the red dots) prediction accuracy are shown in Figure 2.1.

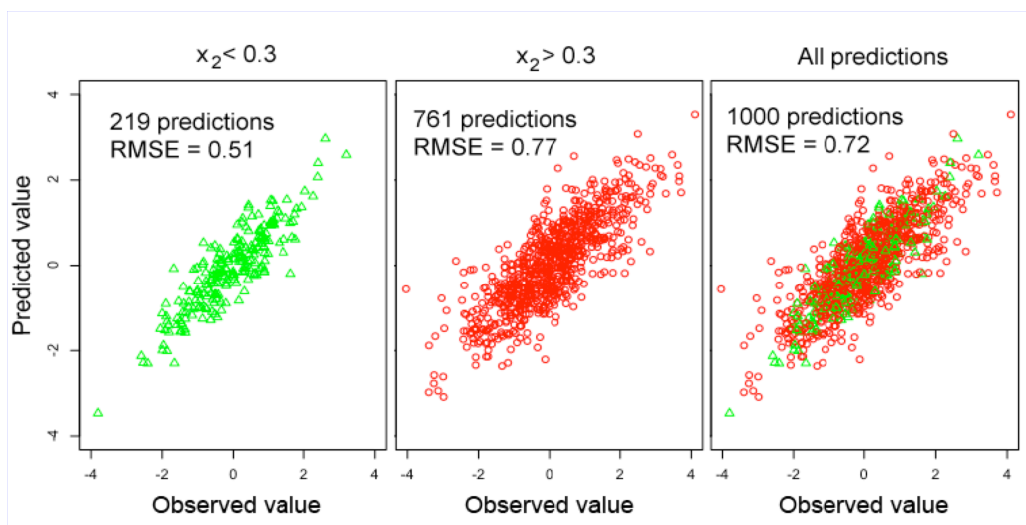


Figure 2.1. An example of the accuracy discrimination. As it can be seen on the scatter plots, the green compounds have higher prediction accuracy (the leftmost plot, RMSE 0.51) than the red compounds (the middle plot, RMSE 0.77). When mixed together, the compounds have RMSE of 0.72 (the rightmost plot).

This synthetic example demonstrates that prediction accuracy can be variable in the chemical space and, more importantly, can be estimated using a discriminating variable. Here, the discriminator of the accuracy was the descriptor x_2 , which controlled the amount of the compounds-dependent noise. The compounds with lower values of x_2 had a better prediction accuracy than the compounds with high x_2 values. In this work, all the numerical measures that possess this kind of discriminating ability will be referred to as *distances to models*. These measures form the core of the methods used for the AD assessment in this work and are defined more precisely in the following section.

2.2.2 Distances to models

The key abstract concept used in this work for assessment of AD is *distance to model* (DM), defined as follows:

Distance to a model is any numerical measure of the prediction uncertainty for a given compound by the model.

Distances to models were also used in earlier QSAR studies to estimate the AD of predictive models. However, the concept was firstly introduced by Tetko et al and clearly formalized, investigated and applied by the author of the current work in the published studies [27-30] and in this thesis work.

A distance to model assesses how “far” is the compound from the model. The compounds that are “further from the model”, which have larger values of DM, are by definition expected to have lower prediction accuracy than compounds that have smaller values of DM. It should be clearly stated that prediction accuracy correlates with DM only in average: for example, compounds with DM in range [0.5, 0.6] will on

average have higher prediction accuracy than compounds with DM in range [0.6, 0.7] but, nonetheless, the prediction errors for some compounds from the first interval can be bigger than for some compounds from the second interval. In other words, the key property of a DM is the discriminating ability, i.e. the ability to discriminate predictions of high and low accuracy.

Importantly, DMs estimate the *reliability* of predictions. While *accuracy* is an objective measure that has a rigid calculation procedure, *reliability* is subjective and can be estimated in numerous ways. Therefore, there is a number of different DMs that assess the reliability of predictions from different perspectives. Here, we briefly overview the DMs used for the AD assessment in this work.

A. Leverage

Leverage is one of the simplest DMs that corresponds Euclidian distance to the center of the training set in the space of molecular descriptors corrected by considering correlations between the descriptors. Leverage for a compound J is calculated as:

$$LEVERAGE(J) = \overrightarrow{x(J)} \cdot (X^T \cdot X)^{-1} \cdot \overrightarrow{x(J)}^T \quad (2.13)$$

where $\overrightarrow{x(J)}$ is a vector of molecular descriptors for the compound J , X is a matrix of descriptors for compounds from the training set. It can be seen that if descriptors are normalized, centered to zero and they do not correlate with one another, then the matrix $(X^T \cdot X)$ is identity and leverage corresponds to the Euclidian distance of the descriptor vector to the zero vector.

In linear modeling, the leverage, which is frequently notated as h , ranges between $1/N$ and 1 and averages $(K+1)/N$ for the N compounds in the learning data set, where K is the number of model variables. The residual of a compound a variance of $\sigma^2(1 - h)$ in the dataset and $\sigma^2(1 + h)$ for external compounds.

High leverage values signal that one starts extrapolating outside the training set range and it is no more guaranteed that the model is valid and applicable. Often, compounds with leverage exceeding a particular threshold h^* (referred to as the “warning leverage”, eq. 2.14) are considered outside of the AD of the model [31]:

$$h^* = \frac{3 \cdot (K+1)}{N} \quad (2.14)$$

An example of the leverage DM effect in a two-dimensional space of descriptors is shown in Figure 2.2

B. Standard deviation of the ensemble predictions (STD)

The standard deviation of the predictions obtained from an ensemble of models can be used as an estimator of model uncertainty. The general idea is that if different models yield significantly different predictions for a particular compound, then the prediction for this compound is more likely to be unreliable. The sample standard deviation can be used as an estimator of model uncertainty.

Assuming that $Y(J) = \{y_i(J), i=1..N\}$ is a set of predictions for a compound J given by a set of N trained models, the corresponding distance to model (STD) is:

$$d_{STD}(J) = stdev(Y(J)) = \sqrt{\frac{\sum (y_i(J) - \bar{y})^2}{N-1}} \quad (2.15)$$

Figure 2.3 shows three predictions with the same (zero) mean but different standard deviations. There are several subtypes of STD, depending on the type of an ensemble used to calculate the standard deviation: CONS-STD (consensus STD) for an ensemble of models based on different machine learning techniques, ASNN-STD for an ensemble of associative neural networks, BAGGING-STD for an ensemble of models created using the bagging technique.

The STD DM has been proven to provide excellent results for discrimination of highly accurate predictions in case of regression models [28,27,32] and is extensively used and benchmarked in the studies encompassed within this work.

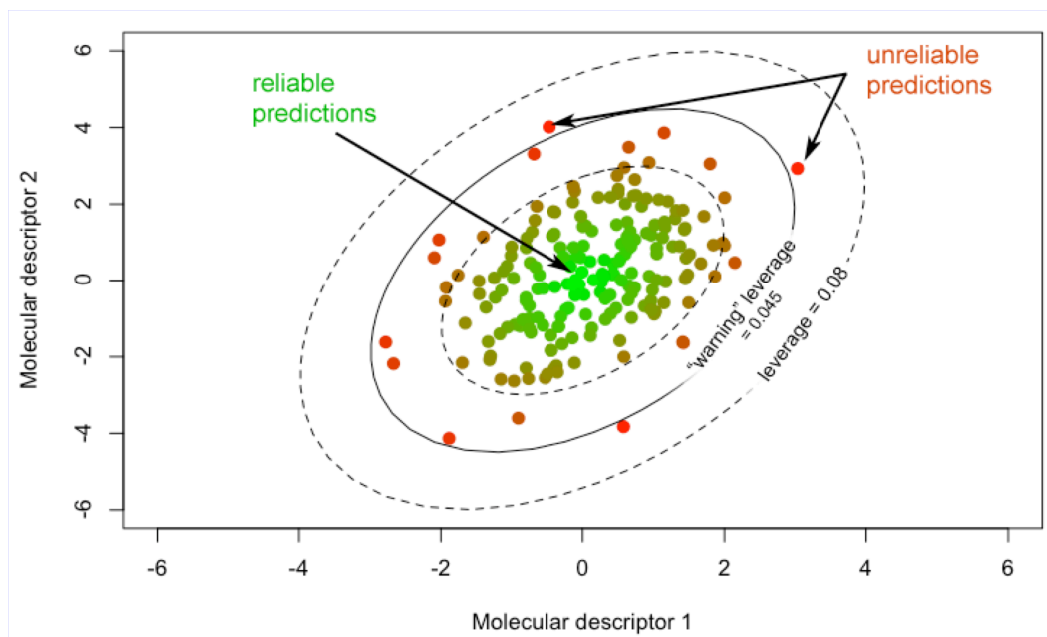


Figure 2.2. An illustrative example of the leverage DM. Leverage penalizes the compounds that are far from the center of the training set in the space of molecular descriptors. According to leverage, such compounds are unreliably predicted.

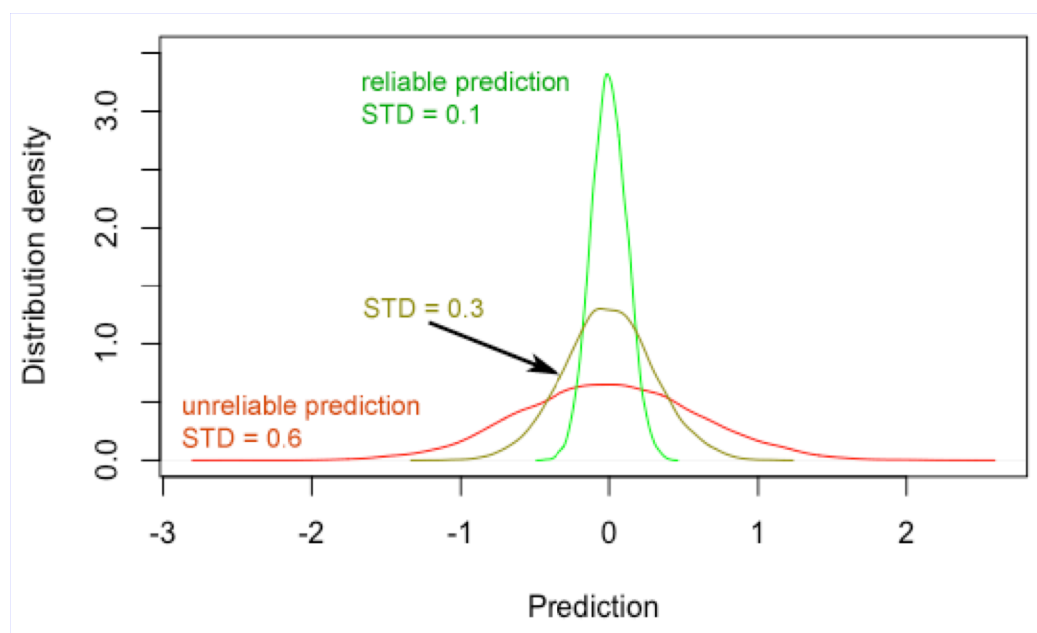


Figure 2.3. An example of three predictions with different standard deviations (STD). According to the STD DM, reliable predictions have a low prediction “spread”, which corresponds to the disagreement of individual predictions within an ensemble of models.

C. Tanimoto similarity

The Tanimoto index is a measure of similarity between two compounds based on the amount of common molecular fragments in these compounds. To calculate the Tanimoto similarity, we enumerate all unique fragments of a particular length in two compounds; then the Tanimoto similarity between the compounds J and I is defined as:

$$TANIMOTO(J, K) = \frac{\sum_{i=1}^N (x_{J,i} \cdot x_{K,i})}{\sum_{i=1}^N (x_{J,i} \cdot x_{J,i}) + \sum_{i=1}^N (x_{K,i} \cdot x_{K,i}) - \sum_{i=1}^N (x_{J,i} \cdot x_{K,i})} \quad (2.16)$$

where N is the number of unique fragments in both the compounds, $x_{J,i}$ and $x_{K,i}$ are the counts of the i -th fragment in the compounds J and K . Based on Expression 2.16, the distance between two compounds J and K is $1 - TANIMOTO(J, K)$ and the distance of a compound to a model is the minimum distance between the investigated compound and compounds from the training set of the model.

D. Correlation of prediction vectors (CORREL)

This measure is based on the correlation of vectors of ensemble's predictions for the target compound and compounds from the training set [15,33]. Similarly to the STD, this measure is applicable only for ensembles of models. More precisely, CORREL measure for the target compound J is calculated according to the following expression:

$$d_{CORREL}(J) = 1 - \max_{i=1..N} [corr(\overrightarrow{y(T_i)}, \overrightarrow{y(J)})] \quad (2.17)$$

where $\overrightarrow{y(T_i)}$ and $\overrightarrow{y(J)}$ are the vectors of ensemble's predictions for the training set compound T_i and the target compound J , $corr$ designates Spearman rank correlation coefficient between the two vectors and N is the number of compounds in the training set. The low value of CORREL (i.e., high Spearman correlation coefficient) indicates that for target compound J there is a compound T_k from training set for which predictions of the ensemble of models are strongly correlated. Indeed, if a compound T_k has the same descriptors as J , than predictions of models will be identical for both molecules and thus $CORREL(J) = 0$. Compounds with high correlation coefficient values are considered to be "near to the model". It has been shown [34], that the Spearman correlation outperformed several alternative measures, e.g. Euclidean distance and Pearson correlation coefficient, because of its higher accuracy and practical requirements to store ensemble predictions for all training set molecules.

E. Rounding effect (CLASS-LAG)

CLASS-LAG is a simple measure of prediction uncertainty, specific for binary classification problems. For such a problem, the labels (i.e. the values to be predicted by a computational model) are discrete and are selected as -1 and 1 for the two classes of compounds respectively. However, most machine learning methods give a numeric (continuous) number as a result of prediction, which is then rounded to the nearest label to identify the class of compound. The less amount of rounding is required, the more reliable the prediction is expected to be. This assumption is utilized by the CLASS-

LAG DM. Namely, the absolute value of difference between the prediction value and the nearest of the labels can be used as a measure of prediction uncertainty. This measure is calculated according to the following expression:

$$d_{CLASS-LAG}(J) = \min\{|-1 - y(J)|, |1 - y(J)|\} \quad (2.18)$$

Thus, this measure punishes deviations from target class labels $\{-1, 1\}$, both positive and negative deviations (i.e. both 1.2 and 0.8 predicted values have the same DM). Obviously, punishing negative deviations applies only to models that have prediction values outside of the $[-1, 1]$ interval.

The effect of CLASS-LAG is visually shown in Figure 2.4 with an example of a binary classification model that divides compounds into two classes, “active” and “inactive”, which have classification labels “1” and “-1”. Green and red dots represent reliable and unreliable predictions, respectively. The CLASS-LAG DM reaches its maximum in the borderline region, where the prediction value is close to zero and, therefore, the model is uncertain whether the compound is active or not.

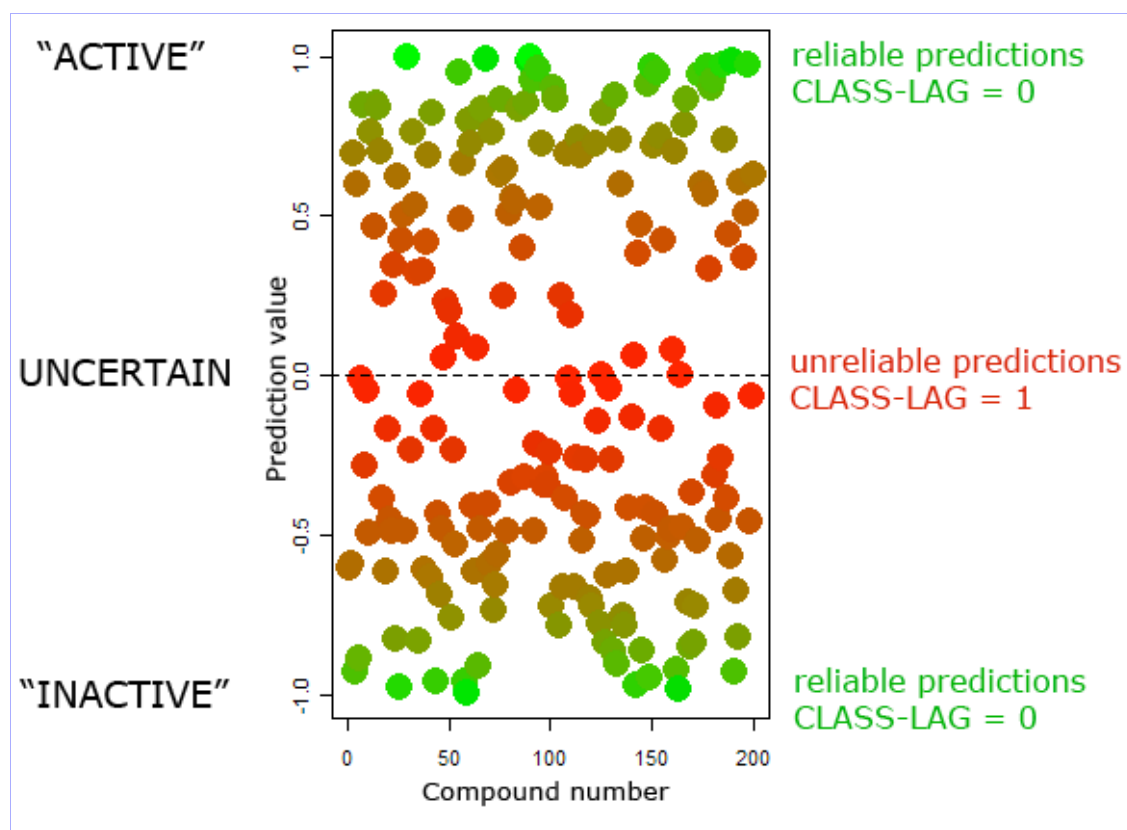


Figure 2.4. Graphical demonstration of the CLASS-LAG DM. According to this measure, the most unreliable predictions (i.e., the highest CLASS-LAG values) are near to the borderline that divides active and inactive compounds.

F. Concordance of a classification ensemble

The maximum number (or the respective percentage) of the models, that give the same prediction can be used as a measure of the concordance of an ensemble. For example if there are 5 models, that give predictions $\{1; -1; 1; 1; 1\}$, then the concordance is 4 (or 80%). The measure that is opposite to the concordance (i.e. 1-CONCORDANCE) can be used as a distance to model. The idea behind this DM is

similar to that of the standard deviation (STD) DM, but is adapted for classification models and qualitative predictions.

G. Rounding effect and standard deviation combined (STD-PROB)

The STD-PROB combines the two sources of the prediction uncertainty: (a) the uncertainty related to rounding of predictions and (b) the uncertainty the disagreement of different models. Instead of a “point” prediction, we consider a distribution of probabilities. We assume that the distribution is Gaussian with mean $y(J)$ and standard deviation that corresponds to the STD value. The suggested distance to model STD-PROB is calculated as follows:

$$d_{STD-PROB}(J) = \min \left\{ \begin{array}{l} \text{Probability}(c > 0 | N(y(J), d_{STD}(J))) \\ \text{Probability}(c < 0 | N(y(J), d_{STD}(J))) \end{array} \right\} \quad (2.19)$$

Or, more precisely:

$$d_{STD-PROB}(J) = \min \left\{ \begin{array}{l} \int_0^{+\infty} N(x, y(J), d_{STD}(J)) dx \\ \int_{-\infty}^0 N(x, y(J), d_{STD}(J)) dx \end{array} \right\} \quad (2.20)$$

where $N(x, y(J), d_{STD}(J))$ is the normal distribution density function with a mean $y(J)$ and a standard deviation $d_{STD}(J)$. Here $y(J)$ is the actual prediction of the analyzed model for a compound J and $d_{STD}(J)$ is the STD DM calculated according to equation (2.3).

This measure can be graphically illustrated as a part of the area under the curve of the normal distribution density function. Four exemplary prediction cases are shown in Figure 2.5, where the rounded prediction value is always fixed to “+1” but the numeric prediction values and the STD values are different. It is obvious that shifting the curve away from the center (decreasing CLASS-LAG) results into decrease of the filled area. The same effect appears when we make the curve less flat, i.e. decrease the STD value. Thus, STD-PROB combines information about uncertainty from both the measures: CLASS-LAG and STD.

The STD-PROB values have a simple interpretation: values close to 0.5 indicate equal probability to find given compound in either class, which means that the model cannot provide a reliable prediction. In contrary, values close to 0 indicate high probability to find the compound in one of the classes.

It should be clearly stated that, as well as all the other described DMs, STD-PROB is an *empirical* measure of the prediction reliability. This measure was introduced by the author of this work in a methodological study [29] and was benchmarked in the Ames test study described in Chapter 4.

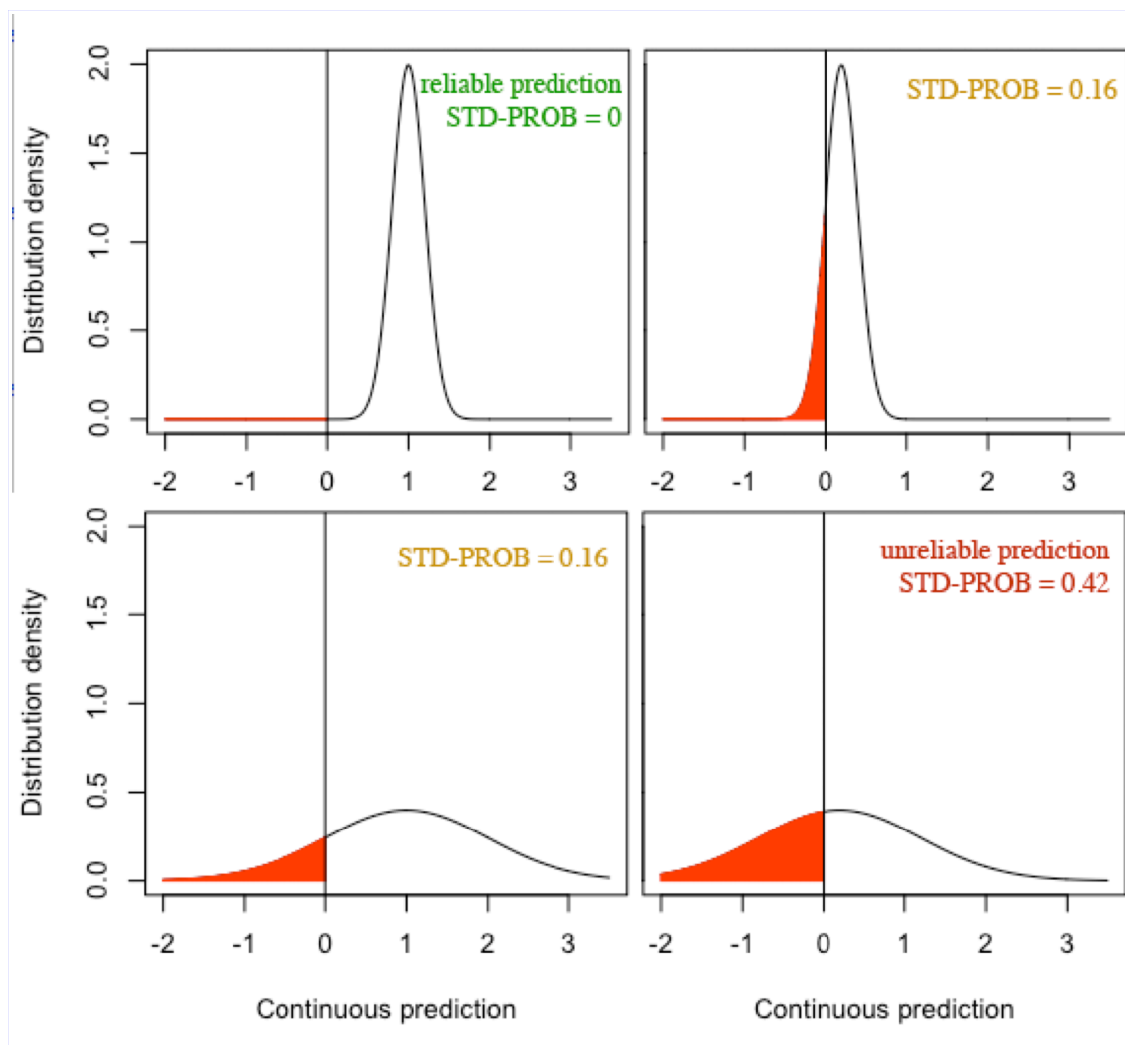


Figure 2.5. An example of four predictions with different reliability according to the STD-PROB DM. The reliability is affected by two factors: the standard deviation (the “flatness” of the curve) and the shift of the curve from the center. Ultimately, these two factors are combined into a single numerical representation, which corresponds to the filled area and is referred to as STD-PROB.

H. Descriptor-based and property-based DMs

The described above distances to models can be classified into two types: (a) the DMs in the space of descriptors and (b) the DMs in the space of properties. The DMs of the first type use only information about chemical structures represented by a number of molecular descriptors. Examples of such DMs are leverage and the Tanimoto similarity. On the contrary, the DMs of the second type use *outputs* of the models. Such DMs include the standard deviation (STD), the rounding effect (CLASS-LAG) and STD-PROB.

The descriptor-based DMs rely on the assumption that two compounds with similar molecular descriptors will have similar properties. These DMs are very popular and are nowadays explicitly or implicitly used in many QSAR studies [35-40]. However, recently an assumption was made [27] that the property-based DMs provide significantly better estimation of prediction accuracy, since they are based not only on descriptors, but also incorporate information about the analyzed model itself. This assumption is comprehensively investigated and validated in this work.

2.2.3 Analysis of prediction accuracy

A. Accuracy averaging

Nowadays, in many published QSAR studies, the prediction accuracy (represented by RMSE, MAE, R^2 etc) is averaged over the whole available set and reported as a single value. However, this approach does not reflect the complete information about the prediction accuracy. For a particular group of compounds, the prediction accuracy may be significantly higher than the average, while for some compounds the model may completely fail to predict the target property. In this work, we refer to this phenomenon as “accuracy variability”.

As it was mentioned in the previous section, the particular measures referred to as distances to models possess the ability to discriminate predictions of high and low accuracies. Thus, the average accuracy should be calculated while taking DM into account. Here, we introduce the DM-based approaches to the accuracy averaging.

Bin based averaging (BBA) splits DM values into non-overlapping intervals (bins) and averages the prediction accuracy for the compounds within each bin separately. By the nature of DM, the accuracy should not increase as the DM increases. Thus, if the average accuracy in a bin is higher than in a previous bin, the accuracy from the previous bin is used. The bin-based averaging results into a *BBA plot*, which shows dependency of the prediction accuracy from DM. This plot is often combined with a plot of residuals (so called *Williams plot*). An example is shown in Figure 2.6.

Sliding window averaging (SWA) aims to provide a continuous dependency and is performed on N adjacent compounds sorted by DM (where N is the window size, which controls smoothness). These N compounds form the averaging window. The advantage of sliding window averaging is that it is more stable to noisy data and, therefore, provides more smoothed dependencies.

According to the **cumulative averaging**, the accuracy is averaged over all the compounds with DMs less than a particular (variable) threshold. Often, it is more convenient to represent DM value not in the absolute scale but in the percentage scale related to a set of compounds. For example, according to the *percentage scale*, a DM value of “10%” would mean that 10% of the compounds from the set have DM values less than this DM value. The cumulative averaging in combination with the DM percentage scale results into a *cumulative accuracy plot*, which shows the prediction accuracy for best 10%, 20% etc compounds (see Figure 2.7). This cumulative averaging is easily interpretable and very stable against noise. In fact, it provides the smoothest dependencies when compared to the aforementioned SWA and BBA. However, due to the cumulative nature of this averaging, the result is dependent on the composition of the investigated set, whereas for the BBA and the SWA, the result is expected to be the same for different sets as long as the sets are sufficiently large.

The above examples were based on regression models, but the same approaches are readily applicable to classification models. The only difference for classification problems is that the correct classification rate is used instead of RMSE.

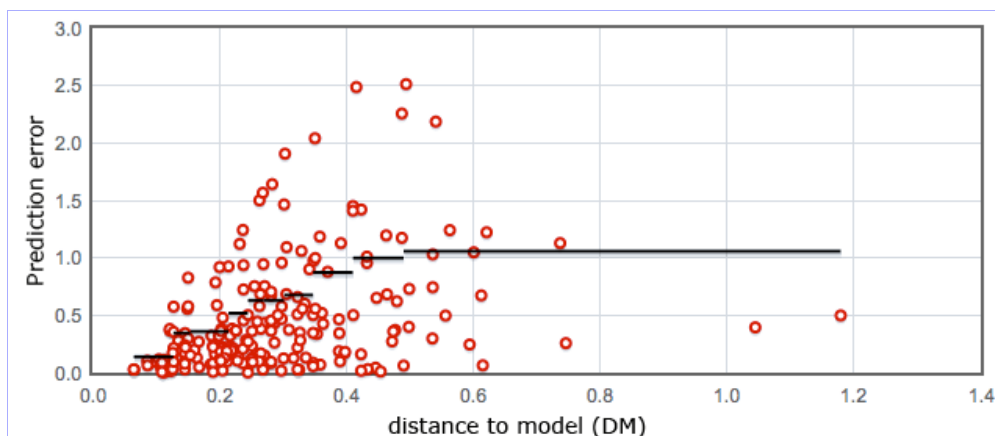


Figure 2.6. The bin-based averaging (BBA) of the prediction accuracy. The red dots represent the compounds from the investigated set; the black lines represent the averaged errors (RMSE) over different DM intervals (“bins”).

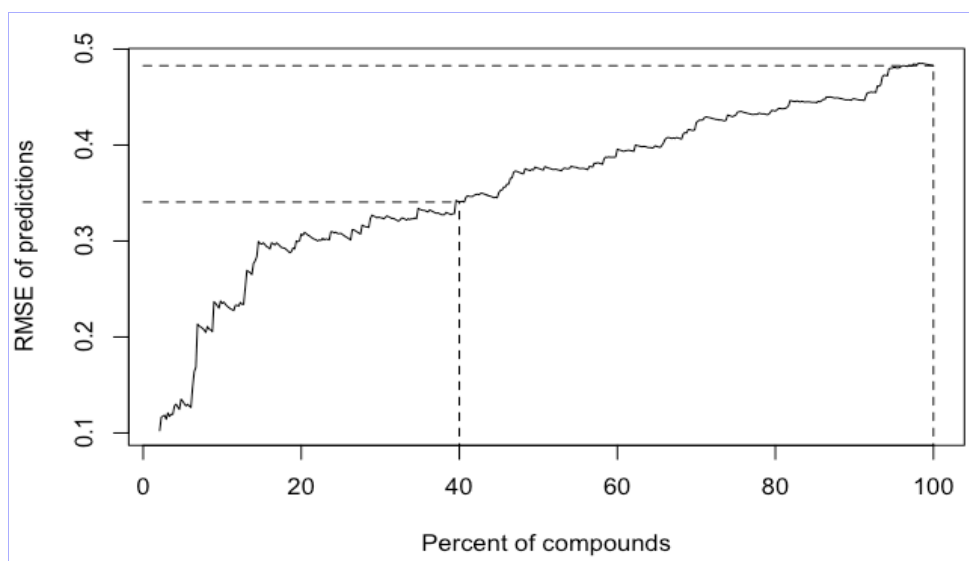


Figure 2.7. An example of a cumulative accuracy plot. This plot shows the RMSE of the predictions with DM less than a variable threshold. “100%” corresponds to the RMSE of all the predictions for the investigated set. Two percentages are highlighted: RMSE of 40% compounds of most reliable predictions is around 0.34, whereas RMSE of 100% compounds (the global RMSE) is around 0.49.

B. Estimation of prediction accuracy

Importantly, the accuracy averaging can be used to estimate the prediction accuracy for new compounds, which were not present in the training or validation sets. To obtain an accuracy estimate for a new compound, we calculate its DM value and define the corresponding accuracy using BBA or SWA plots (Figure 2.6). If an estimate for a set of compounds is required, it is calculated as an average of estimates for all compounds in the set as follows. For example, an estimate for RMSE of the set is calculated as:

$$\tilde{\sigma} = \sqrt{\frac{\sum_{i=1}^N (\tilde{\sigma}(J_i))^2}{N}} \quad (2.21)$$

where J_i is the i -th compound in the set, N is the size of the set, $\tilde{\sigma}(J_i)$ is the estimate of RMSE for the compound J_i , calculated according to the accuracy averaging. Similarly, an estimate for the correct classification rate (CCR) is determined as

$$\tilde{\eta} = \frac{\sum_{i=1}^N \tilde{\eta}(J_i)}{N} \quad (2.22)$$

The estimated accuracy can be explicitly used to sort compounds into the “inside AD” and “outside AD” groups. More precisely, if the estimated prediction accuracy for a compound is within the predefined threshold, the compound is considered to be within the AD of the model and outside of the AD otherwise. Thus, to define AD, we determine the maximum value of DM (referred to as *critical DM* and denoted as d_{AD}) that ensures the required prediction accuracy. All the compounds with DM less than d_{AD} are considered to be inside AD.

2.2.4 Comparison of applicability domains

A. Discriminative power of DM

The ability to discriminate predictions of high and low accuracy is different for every particular distance to model. Therefore, to compare different DMs and applicability domains of models, there is a need to quantify the performance of DMs. There are several ways to evaluate performance of a particular DM.

Accuracy coverage. This approach considers the percentage of the validation set compounds that are predicted with the predefined accuracy. This criterion, referred to as *accuracy coverage*, is useful to determine, for example, the coverage of the validation set that are predicted with the accuracy of experimental measurements. The expected prediction accuracy can be determined from one of the accuracy averaging procedures: the bin-based averaging, the sliding window averaging or the cumulative averaging, as described in Section 2.2.3 on page 22. An example of the accuracy coverage criterion in combination with a cumulative plot is shown in Figure 2.8. This plot is based on one of the QSAR models for the Ames test prediction, described further in this work in the “Benchmarking studies” chapter. For comparison, three different DMs (the red, green and blue curves) and two thresholds (85% and 90%) are shown.

There are two drawbacks of the aforementioned accuracy coverage criterion. First, as it is apparent from Figure 2.8, the coverage depends on the chosen accuracy threshold and different thresholds could possibly results into different ranks of the analyzed DMs. Second, the accuracy coverage depends not only on the ability of DM to separate highly accurate predictions, but also on the performance of the analyzed model. Indeed, the models having higher prediction accuracies will have higher accuracy coverages.

The AUC (area under curve) criterion. Another criterion for the DM performance that does not have the aforementioned drawbacks is the area under curve (AUC), calculated as the area of the square between the bin-based averaging curve (alternatively, the sliding window averaging curve) and the line of the average model performance. More formally, the AUC is calculated as integral:

$$AUC = \int_{x_{min}}^{x_{max}} |a(x) - a| \cdot d(p(x)) \quad (2.23)$$

where x is the DM value in the absolute scale and $p(x)$ in the percentage scale, $a(x)$ is the average prediction accuracy for this DM value, a is the global average accuracy of the model. Practically, $a(x)$ obtained using SWA or BBA provides a discrete dependency. Thus, the integral in the above expression is replaced with a normal sum.

On Figure 2.9, AUC corresponds to the area of the square between the SWA (red) curve and the horizontal line. AUC is higher for the DMs with better separation of compounds with higher and lower accuracies. In this work, we will use AUC as a complement for the accuracy coverage criterion.

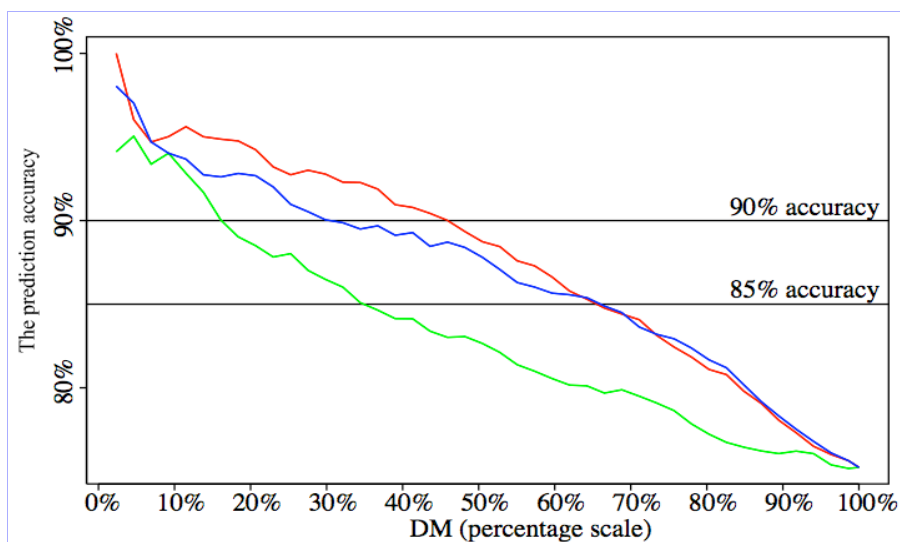


Figure 2.8. Identification of the accuracy coverage using the cumulative plot based on three DMs and the Ames test classification model. With 90% threshold, the “red” DM is superior to the others; it covers about 45% of the compounds from the validation set, while the other two DMs cover about 30% and 15% respectively. With 85% threshold, there is no difference between the “red” and “blue” DMs; both cover about 65% of the compounds.

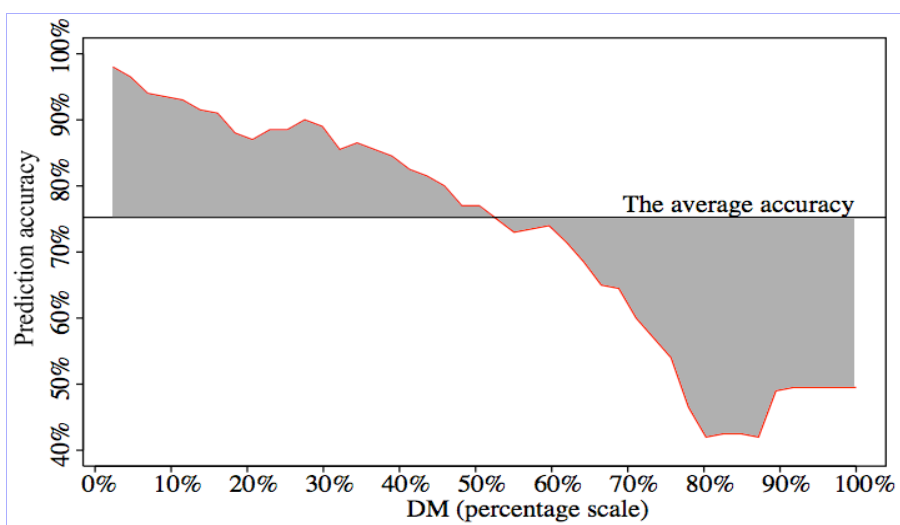


Figure 2.9. The area-under-curve (AUC) criterion corresponds to the filled area between the SWA plot and the average accuracy (the horizontal line).

B. Fitness of probability distribution

The mentioned above approaches estimate the ability of a DM to discriminate predictions of high and low accuracy. However, if the investigated dataset is homogeneous, such separation might not be possible and, namely, not because of a low discriminative power of a DM but because of the specifics of the dataset. For example, if the residuals of the model are distributed normally and do not depend on structures, then no separation of high and low accuracy predictions can be virtually possible. The alternative approach to address this problem considers how well the estimated distribution of residuals (EDR) suggested by a DM fits the actual distribution.

For regression models, the EDR can be approximated as a mixture of Gaussian distributions (MGD) with the same (zero) mean but with different standard deviations obtained from the bin-based averaging (BBA) procedure. In this case, the fitness of EDR is calculated as the likelihood criterion:

$$S = \sum_{i=1}^K \log N(0, \tilde{\sigma}(J_i), e_i) \quad (2.24)$$

where K is the number of the compounds in the set, N is the probability density function of the normal distribution, e_i is the residual (prediction error) for the i -th compound $\tilde{\sigma}(J_i)$ is the estimated RMSE for i -th compound in the investigated dataset; the RMSE can be estimated according to BBA (see section “Estimation of prediction accuracy“ on page 23). The likelihood score defined by (2.24) can be used to assess the quality of a DM.

Approval test. How can we evaluate whether the likelihood score is “sufficient” and, therefore, the DM can adequately estimate the accuracy? The score in expression 2.24 is the likelihood for the estimated MGD suggested by BBA procedure. If this estimated distribution is adequate, it should approximate the actual distribution significantly better than a much simpler single Gaussian distribution (SGD). To check whether MGD approximates the actual distribution better than SGD, we calculated the difference of likelihood scores for MGD and SGD:

$$D = S_{MGD} - S_{SGD} = \sum_{i=1}^K \log N(0, \tilde{\sigma}(J_i), e_i) - \sum_{i=1}^K \log N(0, \sigma_0, e_i) \quad (2.25)$$

where σ_0 is the standard deviation of the SGD corresponding to the standard deviation of the complete dataset. If this score was more than zero (and the difference was statistically significant), then the DM was approved. To evaluate the statistical significance, the bootstrap test was used. In more detail, using all the available residuals, we replicated 10,000 sets of residuals, calculated 10,000 scores according to expression 2.25 and identified the p-value as the percentage of scores less than zero. If the p-value was less than the required level of significance (usually 0.05) then the D score was higher than zero in the statistically significant sense and, therefore, the DM was approved.

Confidence consistency plot. To visually estimate how well the estimated distribution of residuals (EDR) approximates the actual distribution, we used an auxiliary plot referred to as *confidence consistency plot* and generated as follows. Every EDR has the area of confidence, i.e an area where we expect 90%, 80%, 70% etc of all residuals. The actual percentage of residuals in the area of confidence is usually different from the estimated percentage. The confidence consistency plot maps the estimated percentages against the actual ones. The ideal EDR will result into the identity line one the plot. An example of confidence consistency plots for two

hypothetical (single- and multi-gaussian) distributions is shown in Figure 2.10. This chart, generated on basis of a predictive model for growth inhibition concentration (reviewed in more detail in the “Benchmarking studies” chapter), shows that SGD is not a good approximation of the residuals distribution, whereas MGD fits significantly better and, therefore, the DM should be approved.

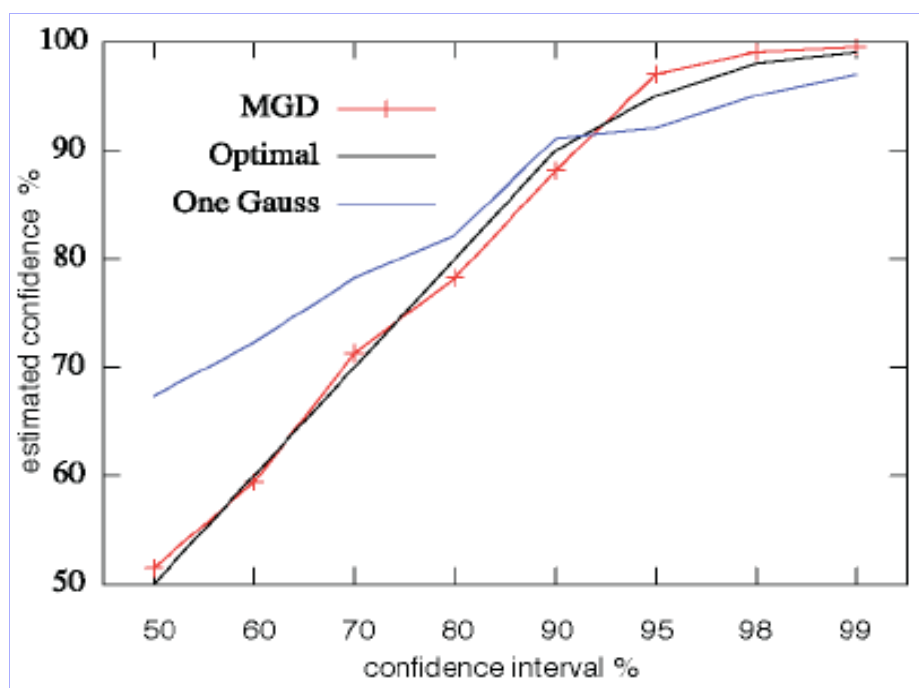


Figure 2.10. An example of confidence consistency plots. The black line is the optimal (identity) plot, the blue and red lines are based on SGD and MGD distributions; apparently, in this example the MGD approximates the optimal plot better. The scale is adjusted to highlight the higher percentages.

As there are no residuals for the classification models, the described above methods are relevant only for regression models. These methods are used in a benchmarking study for the prediction of the growth inhibition concentration for *T. Pyriformis* described further in this work in the “Benchmarking studies” chapter.

2.2.5 Interpretation of applicability domains

In this work, the applicability domain of a QSAR model has a mathematical definition, which relies on the measures of prediction uncertainty, so called “distances to models” (DMs, page 15). However, it is often more practical to interpret AD in the chemical sense. Basically, such an interpretation should provide a meaningful (rather than a mathematical) answer to the question: *which molecules are predicted accurately and why?*

Our approach for interpreting applicability domains is based on the substructure analysis of molecules, performed as follows. First of all, we split all the molecules from the analyzed dataset into molecular fragments. Secondly, for each unique fragment, we calculate the number of molecules containing it. The key point is that the number of the fragment-containing molecules is calculated separately for the most and the least reliably predicted compounds. The most (and the least) reliable predictions are selected accordingly to the lowest (and the highest) values of a particular DM. Thus, for each fragment, we obtain two numbers, which we will designate as $N_{\text{INSIDE_AD}}$ and $N_{\text{OUTSIDE_AD}}$. Finally, we determine whether $N_{\text{INSIDE_AD}}$ is greater (or less) than $N_{\text{OUTSIDE_AD}}$ in the

statistically significant sense, with the significance level of 0.05.

The statistically significant difference in $N_{\text{INSIDE_AD}}$ and $N_{\text{OUTSIDE_AD}}$ would signal that the investigated fragment is over-represented in (or outside of) AD, which in turn would mean that this fragment affects the prediction accuracy. Namely, if $N_{\text{INSIDE_AD}} \gg N_{\text{OUTSIDE_AD}}$, then the molecules containing this fragment tend to have high prediction accuracy and, in contrary, if $N_{\text{OUTSIDE_AD}} \ll N_{\text{INSIDE_AD}}$, the molecules with this fragment are not reliably predicted.

Complementary to the substructure analysis, one can perform a simple analysis of molecular weight to determine whether large (or small) molecules are predicted more reliably. Similarly, one can analyze how the prediction accuracy depends on solubility, lipophilicity, number of atoms and heteroatoms, etc. Additionally, it is possible to analyze whether the value (both observed and predicted) of the property affects the accuracy of predictions.

2.3 Analyzed datasets

2.3.1 Datasets of experimental measurements

This work used 4 datasets with experimental measurements, three of them for biological activities (inhibitory growth concentration, Ames test, CYP-450 inhibition) and one for a physicochemical property – octanol-water partition coefficient. These datasets were used for benchmarking of various QSAR and AD approaches (Chapter 4) and for the practical application of these approaches (Chapter 5). These four datasets are overviewed in Table 2.2 and are described in detail below.

Dataset	Size	Type of QSAR model	Measured property/activity
Ames test	6,542	classification	Mutagenicity according to the Ames test
<i>T. pyriformis</i> toxicity	1,093	regression	Growth inhibition concentration for the ciliated protozoan <i>T. pyriformis</i>
Pt complexes lipophilicity	178	regression	Octanol-water partition coefficient for Platinum complexes
CYP450 inhibitors	7,486	classification	Inhibition of cytochrome P450 (enzyme 1A2)

Table 2.2. An overview of the datasets of experimentally measured properties.

A. Ames test dataset

The Ames test is a biological assay used to identify the mutagenic activity of a chemical compound using histidine-dependent strains of *Salmonella Typhimurium*. The mutagenic activity of a compound determined by the Ames test may signal that the compound is a potential carcinogen [41].

There are different protocols for the Ames test: the test can be carried out with different bacteria strains and with or without the metabolic activation using liver cells. For this study, all such diverse data were pooled together. A molecule was considered as active if it demonstrated the mutagenic activity for at least one *Salmonella* strain. Since not all the molecules were tested with all strains, this could contribute to a significant variance of results. Moreover, the different authors used slightly different thresholds to decide whether a given molecule is active.

The dataset collected from literature (referred to as the “Ames dataset”) contained 6,542 compounds, 3,516 (54%) and 3,026 (46%) thereof with and without the

mutagenic activity respectively. The complete Ames dataset set was randomly divided into a training set and an external test set. The training set contained 4,361 compounds, including 2,344 (54%) mutagens and 2,017 (46%) non-mutagens. The external test set contained 2,181 compounds (1/3 of initial set) including 1,172 (54%) mutagens and 1,009 (46%) non-mutagens.

The accuracy of experimental measurements, expressed as the inter-laboratory agreement of Ames test measurements is reported to be 75-85%. In this work, we performed our own analysis of the Ames test variability, which was based on the more recent data. This analysis, described in more detail further in the work (page 62), identified the inter-laboratory concordance of 90%.

B. *T. pyriformis* toxicity dataset

The growth inhibition (IG) of the ciliated protozoan *Tetrahymena pyriformis* is an established toxicity screening tool developed by Schultz and colleagues [42-44]. The often employed quantitative representation of the IG toxicity is minus logarithm of 50% growth inhibition concentration, denoted as pIGC50.

In recent decades, the Schultz group has published the results of pICG50 measurements for more than a thousand compounds, which were used in this work to assess the AD of QSARs for IG predictions. The initial dataset collected from Schultz publications and the Tetratox website (<http://www.vet.utk.edu/>) contained 983 unique compounds and was randomly split into the training and validation sets containing 644 and 339 compounds respectively. The data from the most recent Schulz publication [42] contained pIGC50 measurements for additional 110 compounds that were not present in the initial dataset. These 110 compounds formed an additional external validation set.

Thus, the complete dataset contained 1,093 compounds split into 644 (the training set), 339 (the 1st validation set) and 110 compounds (the 2nd validation set). The pICG50 values within the complete set had mean of 0.23 and 95% of the values were within the [-1.8, 2.3] interval.

The accuracy of experimental measurements. The experimental analysis of reproducibility of toxicity against *T. pyriformis* was performed by Seward et al.[44] for 51 molecules. The authors divided all the molecules into two groups according to the expected mechanism of their toxicological action: the reactive and narcosis modes of action. The authors reported higher variability of measurements for the molecules with the reactive mode of action. Using the data collected by Seward et al[44], we estimated RMSE of 0.38, MAE of 0.24 and RMSE of 0.21, MAE of 0.13 for 27 and 24 molecules with the reactive and narcosis modes of toxicological action, respectively.

C. Platinum complexes lipophilicity dataset

We gathered a literature dataset containing 178 LogP (octanol/water) experimental measurements for 137 Platinum complexes: 145 measurements for 122 Pt(II) complexes and 33 measurements for 27 Pt(IV) complexes. The data were collected from 18 publications and were quite diverse, since they included measurements carried out with different measurement methods and different pH buffers. Namely, LogP_{o/w} was measured using the shake-flask (108 measurements), HPLC (9 measurements) and RP-HPLC (61 measurements) methods with extrapolated methanol 0% (24 measurements), 30% methanol (42 records), NaCl (43 measurements) and saline (5 measurements) used as pH buffers.

D. CYP450 inhibitors dataset

The CYP dataset contained experimentally determined inhibitors and non-inhibitors of cytochrome P450 (family 1A2) enzyme measured by NIH (National Institute of Health) Chemical Genomics Center and collected from NCBI web-site (<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=410>). The dataset contained 7,486 compounds; 4,016 (54%) thereof were active (inhibitors of cytochrome P450 1A2) and 3,470 non-active compounds. Thus, the dataset had little imbalance of active and non-active compounds.

2.3.2 Datasets of chemical compounds

In order to analyze the applicability of QSAR models investigated in this work to new data, we used a number of the industrial datasets with diverse chemical compounds. These datasets are summarized in Table 2.3 and are described below in more detail.

Dataset	Number of compounds	Description
Enamine	228,899	Drug-like compounds synthesized and screened by the Enamine company
EINECS	68,778	Compounds produced in Europe in amounts of 1 tone per year or more
HPV	2,355	Compounds produces in United States in amounts of 1 millions pounds per year or more

Table 2.3. A brief overview of the investigated industrial datasets of chemical compounds.

A. Enamine dataset

The Enamine dataset contained over 287,000 chemical compounds that were synthesized by the Enamine company (<http://www.enamine.net/>) in 2009. These compounds possessed improved ADMET profiles, in particular they had molecular weight not more than 350 and cLogP not more than 3, which made them significantly more water soluble and less lipophilic. Thus, the Enamine contained mostly drug-like compounds that can be used for screening purposes in drug design.

B. EINECS dataset

The EINECS (European INventory of Existing Commercial chemical Substances) dataset contained 68,779 unique chemical compounds that are produced or imported in Europe in amounts of more than one tone per year. These compounds are intended for the registration in REACH program and, therefore, they are of particular interest for the assessment of their environmental hazard.

C. HPV dataset

HPV (High Production Volume) dataset contained the chemicals produced or imported in the United States in quantities of 1 million pounds or more per year; collecting these compounds was a part of EPA (Environmental Protection Agency) HPV Challenge program (http://www.epa.gov/ncct/dsstox/sdf_hpvesi.html). After filtering out composite substances, stereoisomers and metals, 2,355 compounds remained. Being produced in high volumes, these compounds are interesting for estimation of their environmental harm. In this work, we investigated the applicability of QSAR models for the prediction of toxicity and mutagenicity to the HPV dataset.

2.4 Summary

In QSAR, prediction of biological and physicochemical properties of molecules is based on application of machine learning techniques to datasets with experimental measurements. To be able to apply the quantitative analysis to chemical compounds, they are represented as a set of numerical features, so called *molecular descriptors*. The knowledge extracted from datasets of experimental measurements is represented in a mathematical form with a help of machine learning techniques. Such mathematical representation of the knowledge is referred to as a *predictive model*, which aims to predict the investigated property for new compounds.

An important part of QSAR modeling is determination of the applicability domain (AD) of models, i.e. a subspace of the chemical space, where a model can be applied and give reliable predictions. In this work, a generalized approach to AD was used, which complements every prediction with an estimate of the prediction accuracy. Thus, the AD can be determined by taking only the molecules that have the estimated prediction accuracy higher than the predefined threshold. The assessment of prediction accuracy is done using auxiliary numerical measures referred to as *distances to models* (DM). Distances to models possess an important feature: they correlate with the prediction accuracy and, therefore, can be used for its estimation. The methodology for such estimation is based on the DM-based accuracy averaging. Additionally, the chapter described the methods for the chemical interpretation of AD based on the analysis of molecular sub-fragments.

3 Online chemical modeling environment – OCHEM

This chapter presents an online platform for QSAR research on the Web (OCHEM, accessible at <http://ochem.eu>). The platform allows to simplify and automate the process of the QSAR modeling. Moreover, it supports all the methods introduced in this work and makes their use publicly available online. The platform served as the main tool for most of the research performed in the scope of this thesis work.

3.1 Motivation

Creation of a predictive QSAR model involves a number of time-consuming and tedious steps including data search and preparation, selection and calculation of appropriate molecular descriptors, application of a particular machine learning method, evaluation of results and assessment of the model applicability domain. A particularly difficult step is collecting high quality experimental data. This step involves time-consuming work with scientific literature, manual extraction of experimental data from the literature and preparation of the data for further steps of the modeling process. On the next step, a researcher often uses external tools to calculate molecular descriptors for the data and to finally train a model using a machine learning method of choice. Further follows the evaluation of the model performance, the applicability domain assessment, the investigation of the outliers and irregularities in the data, the revision of the initial dataset and repetition of the whole process. To sum up, the process of modeling is tedious and iterative.

It is interesting that there are hundreds or possibly even thousands of models published every year (e.g., more than 50 models were estimated to be published only for lipophilicity, logP, and water solubility in 2005) [45,46]. However, for most models, a publication marks the end of their life cycle. Only seldom models continue as software tools and perform practical prediction of new data, i.e. they seldom serve the purpose that they were developed for. Thus, after spending a considerable effort on data preparation, development and publication, there is virtually no use of these models at the end of this endeavor. Attempts to reproduce the published models are not always successful and can be an art of their own.

One of the difficulties in the reimplementation of models is data availability. Indeed, models built with so-called memory-based approaches, such k-Nearest Neighbors (kNN), Support Vector Machines, Probabilistic Neural Networks, etc. require the initial data that was used to train the models. Nonetheless, many models are still being published without these data. Many published models may include only names of molecules or only calculated descriptors. Unfortunately, chemical names are often ambiguous, while calculated descriptors are subjected to the same implementation problems as aforementioned for models. Thus, substantial efforts could be necessary for reproducing published results.

There is a variety of online tools to store chemical information of small compounds (i.e. online databases) and tools to create predictive models. However, most of these tools lack essential facilities required for modeling. For example, some data providers (PubChem, ChemSpider, DrugBank, Chempedia, ChemExper [47-52]) provide storage of chemical information but lack modeling tools. Furthermore, some databases do not provide essential information required for data verification and modeling: the source of information and the conditions under which the experiments have been carried out. The quality of such data, which is very important to create a predictive model, cannot be easily verified. A number of online tools [53,54] provide modeling facilities, but lack integration with a chemical database and cannot provide a stable workflow of typical QSPR modeling steps.

This chapter presents a unique and innovative online platform, the Online Chemical Modeling Environment (OCHEM, <http://ochem.eu>), which allows to perform and partially automate the aforementioned steps of the QSAR modeling process. The platform includes two major subsystems: the database of experimental measurements and the modeling framework. The database subsystem includes the storage of experimental measurements and tools to efficiently introduce, search and manipulate chemical data. The modeling framework provides facilities to use these data in the modeling process and perform all the steps of a typical modeling workflow. Most importantly, the developed and published OCHEM models are publicly available (together with the data used to develop them) to the scientific community and can be freely used on the Web to predict new molecules.

The author of this work has made the key contribution to the development of OCHEM. More precisely, this contribution included the development of the global framework and concept, the structure of the database, the modeling framework and, importantly, the tools for applicability domain assessment.

3.2 The database of experimental measurements

3.2.1 Structure overview

The database contains results of experimental measurements of biological and physicochemical properties of small molecules together with the conditions under which the experiments have been carried out and the sources where this data was published. The structure of the database is schematically presented in Figure 3.1.

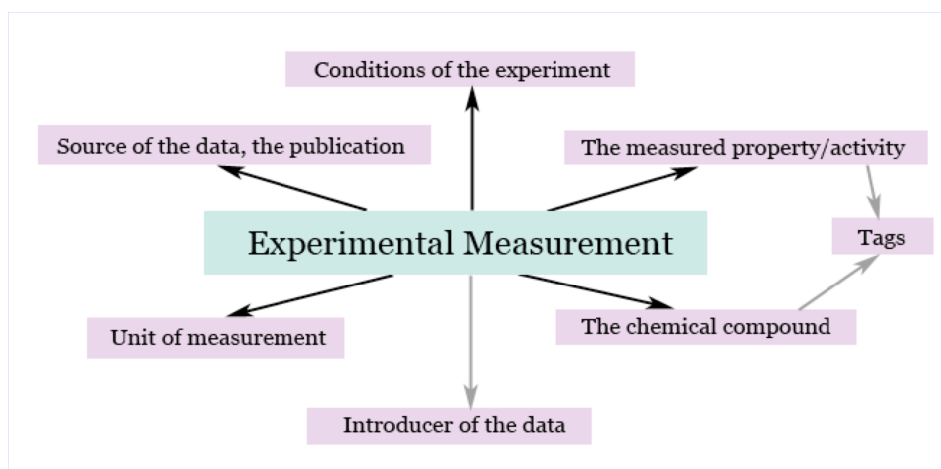


Figure 3.1. A schematic overview of the OCHEM database.

The *experimental measurements* are the central entities of the database, which combine all the information, related to the experiment, in particular the result of the measurement, which can be either numeric or qualitative, depending on the measured property.

An experimental measurement record includes information about the *property* that was measured and the *chemical compound*. The compounds and the properties can be marked with particular keywords, also known as *tags*, which allow convenient filtering and grouping of the data. Exemplary tags are “ADME properties”, “toxicity”, “perfluorinated compounds” etc.

For every measurement stored in the OCHEM, it is mandatory to specify the *source of the data*. This is usually a publication in a journal or a book. The OCHEM strict policy is to accept only the experimental records that have their source specified. This restriction improves the quality of the data and allows to verify it by checking the original publication.

The OCHEM database stores the original *units of measurements* (i.e., as provided in a publication) and can interconvert between different units, thus allowing to use data with different units for modeling. Units are grouped by categories (for example Kelvin, Celsius and Fahrenheit degrees belong to the “Temperature” category of units and can be automatically interconverted to each other). Each numeric property has a corresponding category of units, for example, the category of units for *Inhibition Concentration 50% (IC50)* is “Concentration”. Values in units within the same category can be automatically interconverted.

An important feature of the database, which is also unique among the other chemical databases, is the possibility to store *conditions of experiments*. This information is crucial for modeling: in many cases, it is useless to specify the result of an experimental measurement without specifying the conditions under which the experiment was carried out. For example, it does not make sense to specify boiling point without pressure. Such conditions can be indicated as obligatory, i.e., it is strictly required to specify them for every experimental measurement for this property. Similarly to values of the experimental measurements, the condition values can be numeric (with unit of measurement), qualitative or even textual.

3.2.2 Sources of information

One of the basic principles of the OCHEM database policy is a strict requirement to provide the source of information for every experimental measurement introduced to the database. Most chemical databases do not store this information, which makes it difficult to verify the data and to correct errors.

The OCHEM supports two types of sources: articles (publications in scientific journals) and books (or chapters of books). There is a number of supplementary fields for every type of source: title, abstract, journal, PubMed identifier, DOI identifier, ISBN number, web link etc. For every source, it is possible to store a PDF file, which makes it easier to verify the data later on. Articles and books can be either entered manually or uploaded automatically (from PubMed database [55], external file, by ISBN number etc)

All publications can be accessed from the *article browser*, which allows to search publications by author, PubMed ID, ISBN, title, journal etc. Additionally, from the

article browser, a user can navigate to the experimental measurements associated with a particular publication.

3.2.3 Data access and management

After data has been introduced to the OCHEM database, the experimental measurements can be navigated (by the measured property, by the conditions of experiment, by publication, etc.) and organized in the sets that can eventually be used to train a QSAR model.

Data search and filtering. Every entity in the OCHEM database has a corresponding dialog referred to as *browser*, where the records can be accessed, searched and modified. An example of the browser of experimental measurements can be found in Figure A1 on 133 in Appendix. In every browser, the filters are used to focus on a certain subset of the data, e.g. a training set to create a new predictive model. Records can be filtered by literature source (article or book where the data has been published), physicochemical property or experimental condition and structural information, e.g. molecule name or InChI key as well as by molecular sub-fragments. Additionally, there are the comprehensive filter options to find duplicated measurements, errors in molecular structures, the measurements with a particular range of property values etc.

A typical scenario of the dataset preparation is to filter data by a property (e.g., the octanol-water partition coefficient), by particular conditions of experiments (e.g., partition coefficient measured in 30% methanol solution) or by substructure (e.g., the measurements only for Platinum complexes).

Baskets. Experimental measurements can be combined in sets referred to as *baskets*. The typical use of a basket is to create a set for further use in the modeling process as a training or a validation set. In the basket profile, a user can see an overview of publications with experimental data used to create the basket, the number of unique compounds, etc.

Duplicates management. To ensure data consistency, it is essential to avoid redundancy in the database. Thus, there is a need for strict rules for definition of duplicates. In the OCHEM, two experimental records of a physicochemical or biological property are considered to be duplicates if they are obtained for the same compound under the same conditions, had the same measured value (with a precision up to 3 significant digits) and are published in the same article. We refer to these records as *strong duplicates*, as opposite to *weak duplicates*, for which only part of the information is the same. The OCHEM database does not forbid strong duplicates completely, but forces all the duplicates (except for the record introduced first) to be explicitly marked as error. This ensures that there are no strong duplicates among the valid (i.e, non-error) records.

Importantly, the uniqueness of chemical compounds is controlled by the special molecular hashes, referred to as Inchi-Keys.[56] Namely, for the determination of duplicated experimental measurements, two chemical structures are considered same if they have identical Inchi-keys.

The OCHEM allows weak duplicates (for example, completely identical experimental values, published in different articles) and provides facilities to find them. Moreover, in the modeling process, it is always automatically ensured that the same compounds in the training set appear only in one fold of the N-fold cross-validation

process.

3.3 Modeling framework

3.3.1 Overview

An essential part of the OCHEM platform is the modeling framework. Its goal is to provide facilities for building predictive QSAR models. The framework is integrated with the database of experimental data and supports all the necessary steps required to build a computational QSAR model: data preparation, calculation and filtering of molecular descriptors, application of machine learning methods, analysis of the model performance and, importantly, assessment of the applicability domain.

The modeling framework allows combining inhomogeneous experimental data measured under different conditions of experiments and represented in different measurement units. The conditions can be passed as additional inputs for calculation models, whereas data in different measurement units are automatically converted to the default unit of measurement.

The framework supports calculation both of classification models (i.e. qualitative predictions: active vs. non-active, mutagenic vs. non-mutagenic etc) and of regression models (predictions of numeric properties such as lethal dose or concentration, partition coefficient, melting point). Importantly, every experimental measurement used to train a model can be tracked and analyzed individually, which allows to identify the reasons for outliers and other irregularities in the data.

Finally, the OCHEM provides an assessment of applicability domain by complementing each prediction with an estimation of the prediction accuracy, which is based on the approaches introduced in the “Methodology” chapter.

This section gives an overview of these features and of the steps required to build a computational model in the OCHEM.

3.3.2 Calculation of models

Generally, to create a predictive model, it is necessary to specify the training and (optionally) validation set, to choose and configure the machine learning method, to select and configure descriptors and configure the filtration rules. Finally, the calculated model is reviewed based on various statistical measures (RMSE, MAE, r^2 , etc.) and on the prediction plots.

Training and validation datasets. One of the most important necessary steps to create a predictive model is the preparation of the input data, i.e. the training set that contains experimentally measured values of the predicted property. The property is identified automatically based on the contents of the training set. If the training set contains multiple properties, they will be predicted by the model simultaneously, which allows the knowledge about different (but related) properties to be combined into a single model. This feature is referred to as *multi-learning* [57]. Multi-learning was shown to significantly increase overall performance in comparison to models developed for each property separately.

For each predicted property, it is necessary to select the default measurement unit. All the input data will be automatically converted to this unit and the model will give

predictions in this unit. Conversion in a single unit allows combining inhomogeneous data with different units of measurements.

The screenshot shows the 'Model editor' interface with the following elements:

- Model editor**
Select model template and training set
- Training set (required): Pyriformis Zhu complete
Validation set (optional): [...]
- The model will predict this property:
log(IGC50-1) using unit: -log(mmol/L) [dropdown]
- Choose template for the model: ANN [dropdown]
- Model validation**
Validation method: N-Fold cross-validation [dropdown]
Number of folds: 5 [input field]
- Next>> [button]

Annotations on the right side of the interface:

- Select the training set
The predicted property is automatically identified on the basis of the dataset
- Select the unit of measurement
If the dataset contains measurements in multiple units, they are automatically converted into the selected unit. Automatically, the default unit for this property is selected
- Select the machine learning method
OCHEM supports neural networks, linear regression, kernel ridge regression and kernel PLS, support vector machines and k nearest neighbors (KNN)
- Select the validation protocol
currently, the OCHEM supports the cross-validation and the bagging validation
- Proceed to the selection of descriptors
Further steps include configuration of molecular descriptors, standardization of molecules and the parameters of the selected machine learning method

Figure 3.2. The first step of the model creation: the selection of training and validation sets, the machine learning method and the validation protocol.

The machine learning method. After having specified the training and validation sets (Figure 3.2), a user selects a machine learning method and the validation protocol. Currently, the OCHEM supports linear and kernel ridge regression, associative neural networks, kernel partial least squares (KPLS), correction-based LogP model, support vector machines (SVM), Fast Stage-wise Multivariate Linear Regression (FSMLR) and k nearest neighbors (kNN). These machine learning methods are overviewed in the “Methodology” chapter on page 7.

Validation of models. The OCHEM offers two possibilities to validate a model: N-fold cross-validation and bagging (refer to page 9 for the description). Although it is possible to build a model without validation, it is strongly recommended to always use one of the two validation options since the absence of a proper validation may result into misleading results and an over-fitted model [58,59]. Moreover, the bagging validation has an additional purpose: it is used to calibrate the estimation of the accuracy of the predictions, since it provides multiple predictions that can be used to calculate the standard deviation DM.

If cross-validation is chosen, the *whole* process of model development, including filtering of descriptors, is repeated N (by default 5) times with a different split of the initial set into training and validation sets. Only respective training set data are used in each step for model development.

In case of bagging validation, the system generates N (by default 100) training sets and builds N models, based on these sets. The N sets are generated from the original training sets by random sampling with replacement. The compounds that were not included in the training set of a particular model are used to validate the performance of this model; the final prediction for a compound is the mean prediction over all the models where this compound was in the validation set.

For both cross-validation and bagging, duplicated molecules (regardless of stereochemistry) are used either in training or validation sets, but never in both

simultaneously.

3.3.3 Descriptors

The OCHEM supports all of the molecular descriptors outlined in the “Methodology” chapter and a number of additional descriptor types. The available descriptors are grouped by the software that contributes them: ADRIANA.Code [60], CDK descriptors [61], Chirality codes [62-66], Dragon descriptors [7], E-State indices [8], ETM descriptors [67,68], GSFragment molecular fragments [69], inductive descriptors [70], ISIDA molecular fragments [9], quantum chemical MOPAC 7.1 descriptors [71], MERA descriptors [72-75], MolPrint 2D descriptors [76], ShapeSignatures [77] and logP and aqueous solubility calculated with ALOGPS program [10]. Many of the descriptor types have additional configuration options; for example, for ISIDA fragments, it is necessary to provide minimum and maximum length of the molecular sub-fragments taken into account; for Dragon descriptors, it is possible to individually select several of the 20 logical descriptor blocks, etc. The descriptor selection screen is shown in Figure 3.3.

Select descriptor blocks

Please select the MOLECULAR descriptors:

- E-state [W](#)
- ALogPS [W](#)
- MolPrint [W](#)
- GSFragment [W](#)
- Dragon [W](#)

[\[select all\]](#) [\[select none\]](#)

<input checked="" type="checkbox"/> constitutional descriptors	<input checked="" type="checkbox"/> topological descriptors
<input checked="" type="checkbox"/> walk and path counts	<input checked="" type="checkbox"/> connectivity indices
<input checked="" type="checkbox"/> information indices	<input checked="" type="checkbox"/> 2D autocorrelations
<input checked="" type="checkbox"/> edge adjacency indices	<input checked="" type="checkbox"/> BCUT descriptors
<input checked="" type="checkbox"/> topological charge indices	<input checked="" type="checkbox"/> eigenvalue-based indices
<input checked="" type="checkbox"/> Randic molecular profiles	<input checked="" type="checkbox"/> geometrical descriptors
<input checked="" type="checkbox"/> RDF descriptors	<input checked="" type="checkbox"/> 3D-MoRSE descriptors
<input checked="" type="checkbox"/> WHIM descriptors	<input checked="" type="checkbox"/> GETAWAY descriptors
<input checked="" type="checkbox"/> functional group counts	<input checked="" type="checkbox"/> atom-centred fragments
<input checked="" type="checkbox"/> charge descriptors	<input checked="" type="checkbox"/> molecular properties

- ISIDA fragments [W](#)

Fragments from to

Type of fragments

- MOPAC/Boinc-derived descriptors [W](#)
- ADRIANA.Code [W](#)
- CDK molecular descriptors [W](#)
- ShapeSignatures [W](#)
- 'Inductive' descriptors [W](#)
- MERA descriptors [W](#)

Please select the ATOMIC descriptors:

- E-state [W](#)
- MOPAC/Boinc-derived descriptors [W](#)

Figure 3.3. Choice and configuration of molecular descriptors.

3.3.4 Conditions of experiments

A unique feature of the OCHEM is the possibility to use conditions of experiments for modeling. The properties of chemical compounds usually depend on a number of conditions under which the experiment was carried out. Examples of such conditions are temperature, pressure, pH, measurement method, etc. The OCHEM allows to use these conditions in the modeling process as descriptors.

This feature makes it possible to combine the data measured under different conditions. For example, the boiling point data measured under different pressures can be combined into a single training set and used to develop a computational model. Another example is combining LogP values measured in different pH buffers, e.g. pure water, 30% methanol or saline. The case with different pH buffers occurred in one of the studies described further in this work, the prediction of octanol-water partition coefficient for Platinum complexes.

3.3.5 Configuration of the machine learning methods

There is a number of configuration options that are specific for each particular machine learning method. The available machine learning methods are associative neural networks (ASNN), k nearest neighbors (KNN), kernel ridge regression and kernel partial least squares (KRR and KPLS), the LogP model, multiple linear regression analysis and fast stage-wise multivariate linear regression (MLR and FSMLR) and support vector machines (SVM). These machine learning are described in detail in the section “Machine learning methods” on page 7.

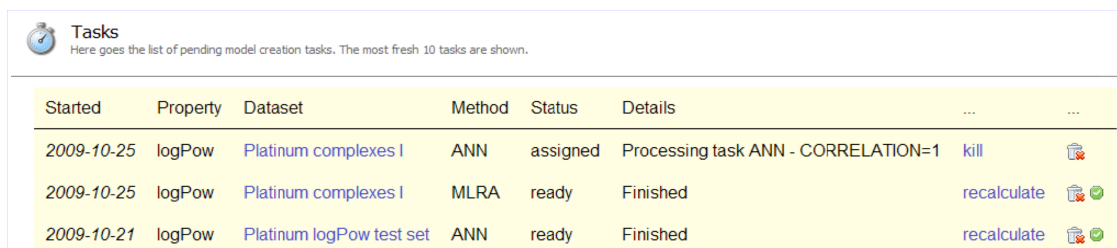
The basic configuration options and their effects on the modeling are summarized below in Table 3.1. In general, there are three types of effects of a parameter. First, a parameter can control how fast a model is calculated, whereas longer calculation times may result into better models. Second, a parameter can provide the data-specific effect and should be optimized for every problem individually. Third, the parameter can control the goodness of fit for predictions on the training set. A high fitness usually results in a more complex model, which can perform good on the training set but have a poor predictive ability in general. Basically, this parameter type controls over-fitting.

Machine learning method	Configurable parameter	Effect of the parameter
Associative neural networks (ASNN)	The number of neurons in the hidden layer	speed vs. quality
	The number of iterations	
	The training algorithm	
KNN	The metrics in the descriptors space	data-specific
	The number of neighbors	
KPLS and KRR	The kernel type	data-specific
	The grid-search optimization parameters	speed vs. quality
LogP	-	-
MLR	The significance level for variable selection	goodness of fit vs. generalization ability
FSMLR	The relative size of the internal validation set	generalization ability
	Shrinkage	
SVM	SVM type	data-specific
	Kernel type	
	The grid-search optimization parameters	speed vs. quality

Table 3.1. The configurable parameters of the machine learning methods in OCHEM and their effects.

3.3.6 Model calculation

After the modeling process has been initiated, the calculation task is posted to the system of distributed calculations (described further). In case of a moderately large model (based on several thousands of compounds in the training set and several hundreds of molecular descriptors), the calculation can take from several hours to several days depending on the utilized machine learning method. All the pending (in calculation) models are stored in a special registry (referred to as “pending tasks”, see a screenshot in Figure 3.4). After the training of a model has been completed, it can be fetched from the registry of pending tasks and saved for further use.



Tasks
Here goes the list of pending model creation tasks. The most fresh 10 tasks are shown.

Started	Property	Dataset	Method	Status	Details
2009-10-25	logPow	Platinum complexes I	ANN	assigned	Processing task ANN - CORRELATION=1	kill	
2009-10-25	logPow	Platinum complexes I	MLRA	ready	Finished	recalculate	
2009-10-21	logPow	Platinum logPow test set	ANN	ready	Finished	recalculate	

Figure 3.4. A screenshot of the registry of the pending QSAR models.

Importantly, multiple models can be trained simultaneously. For example, in case of the study for LogPow predictions described further in this work, we trained 20 models in parallel simultaneously. Moreover, as these models used bagging, each of these 20 models was in fact an ensemble of 100 individual models. Thus, there were 2,000 models that were trained in parallel.

3.3.7 Distributed calculations

QSAR modeling is often very intensive in terms of the required time and computational power. Both the calculation of molecular descriptors and training of models with thousands of compounds require intensive calculations. The time required for the calculation of ensembles of models is even higher: instead of one model, a hundred of models should be trained and afterwards applied for prediction of new compounds. The calculation of ensembles is essential for the calculation of the STD DM (page 16) and the assessment of the AD. For example, in case of the Ames test study described further in the work, the neural networks model alone had to be trained 100 times using a dataset with more than 6,000 compounds and then applied to about 300,000 compounds. Being performed on a single computer, these calculations could take up-to several months of calculations, which is infeasible for a regular research.

Thus, for QSAR research, there is a clear need for a distributed calculation system. This problem was addressed and in the OCHEM system. All the calculations are distributed to almost 200 CPUs, which include LSF (Load Sharing Facility) servers of Helmholtz Centre in Munich and a number of desktop computers (Figure 3.5). The calculation tasks are controlled by a central unit, the Metaserver, which receives, assigns and stores calculation tasks. The calculation task types distributed over the calculation servers include the machine learning algorithms (neural networks, support vector machines etc), the molecular descriptors and the management servers, which control the general flow of tasks in a typical QSAR modeling process.

The distributed calculation system was capable to reduce the calculation time up-to 100 times. The distribution of calculations was indispensable for the Ames test and cytochromes inhibition studies, which were based on the datasets with thousands of

measurements and ensembles with hundreds of individual models.

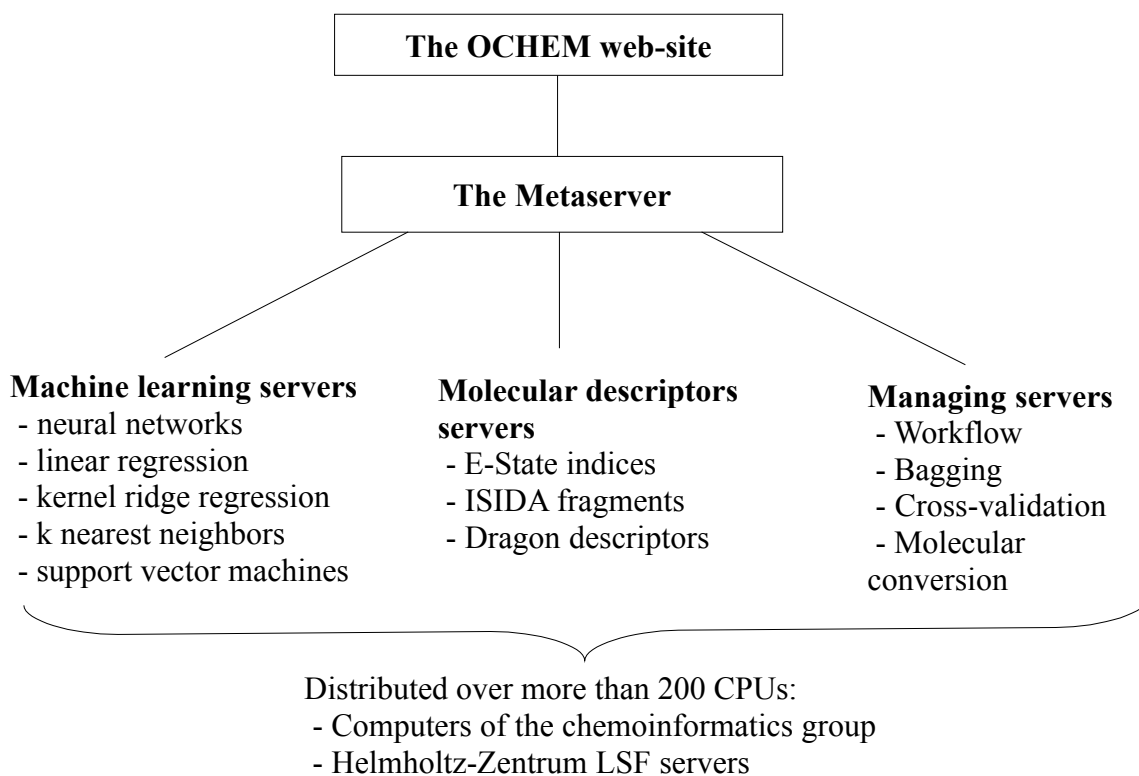


Figure 3.5. Distribution of calculations in OCHEM.

3.3.8 Analysis and management of models

The OCHEM provides with a variety of statistical instruments to analyze the performance of models, to find outliers in the training and validation sets, to discover the reasons for the outliers and to assess the applicability domain of the model. In this section, we briefly overview these instruments.

Regression models. The commonly used measures of a regression model performance are the root mean square error (RMSE), the mean absolute error (MAE) and the squared correlation coefficient (r^2). The OCHEM system calculates these statistical parameters for both the training and the validation sets.

For a convenient visual inspection of the results, the OCHEM is equipped with a graphical tool that allows to create the observed-vs-predicted chart (see Figure 3.6). This type of chart is traditionally used to visualize the model performance and to discover outliers. Each compound from the input dataset is represented as a dot on this chart, where the x-coordinate of the dot corresponds to the value of the property, observed experimentally and y-coordinate is the value, predicted by the model. Importantly, each dot on the chart is interactive; a click on the dot opens a window with the detailed information about the compound: the compound name, the measured and predicted property values, the publication, conditions of experiment, etc. The possibility to track each compound to the reference source is a very important feature for understanding of the reasons why the compound is considered to be an outlier. A user can quickly check why a bad prediction happened, whether it is due to an error in the dataset, differences in the experimental conditions or due to the failure of the model to predict the compound properly. This seemingly simple feature is a good example of the advantage of integrating the database with the modeling framework.

Classification models. In case of classification models, the performance measure is the average correct classification rate (in %). The accuracy is complemented with a confusion matrix, which shows the numbers of compounds classified correctly for every class, as well as details for misclassified compounds, e.g., how many compounds from a *class A* are classified as belonging to a *class B* (see Figure 3.7).

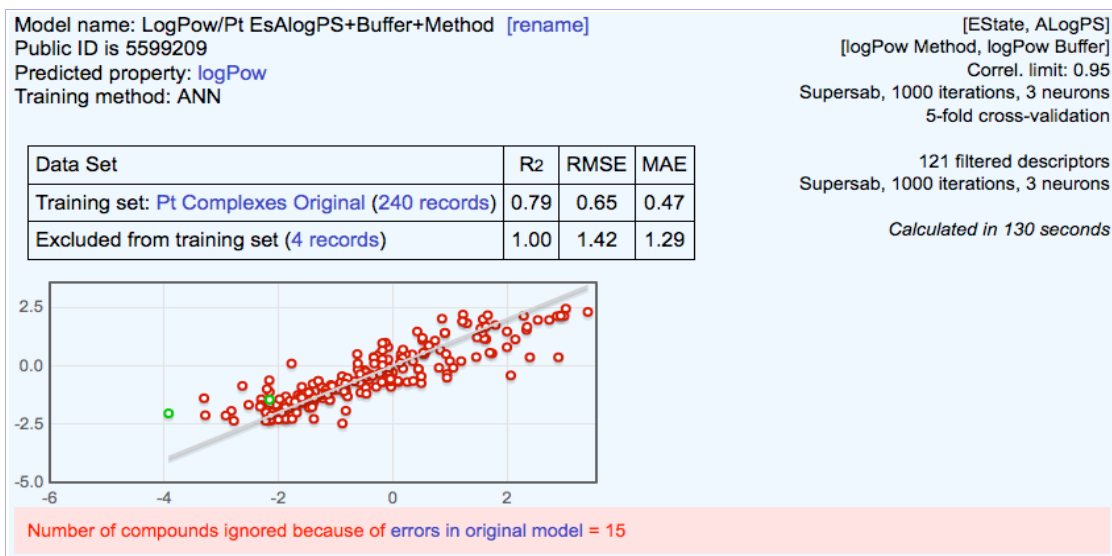


Figure 3.6. Basic statistics for a predictive model. The training set has a link that opens a browser of experimental records where a user can examine properties of all compounds used to in the model. A click on a dot in the observed-vs-predicted chart opens a similar browser information window for the corresponding compound.

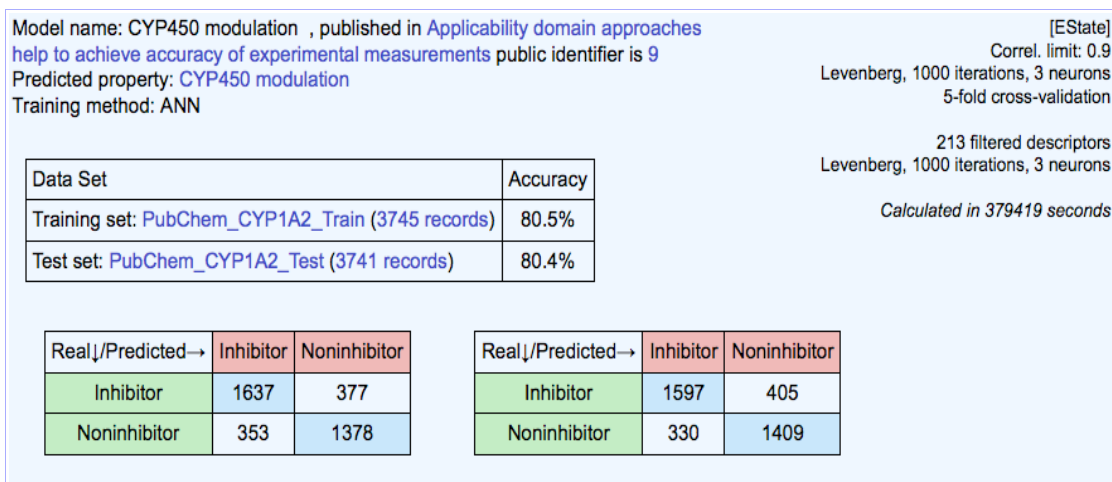


Figure 3.7. Statistics of a classification model. Summarized are the prediction accuracies and the confusion matrices for the training and test sets.

Comparison of multiple models. Often, it is useful to compare different models that are built on a basis of the same data but with different QSAR approaches, that is with different molecular descriptors and machine learning methods. The OCHEM supports a collective view of the models sharing the same training set. An example of such an overview is presented in Figure 3.8, which shows several models for prediction of the octanol/water partition coefficient, a study, which is reported further in the “Applications” chapter. Each point is clickable and allows to track every experimental measurement used in the models and, thereby, to investigate irregularities in the data and outliers.

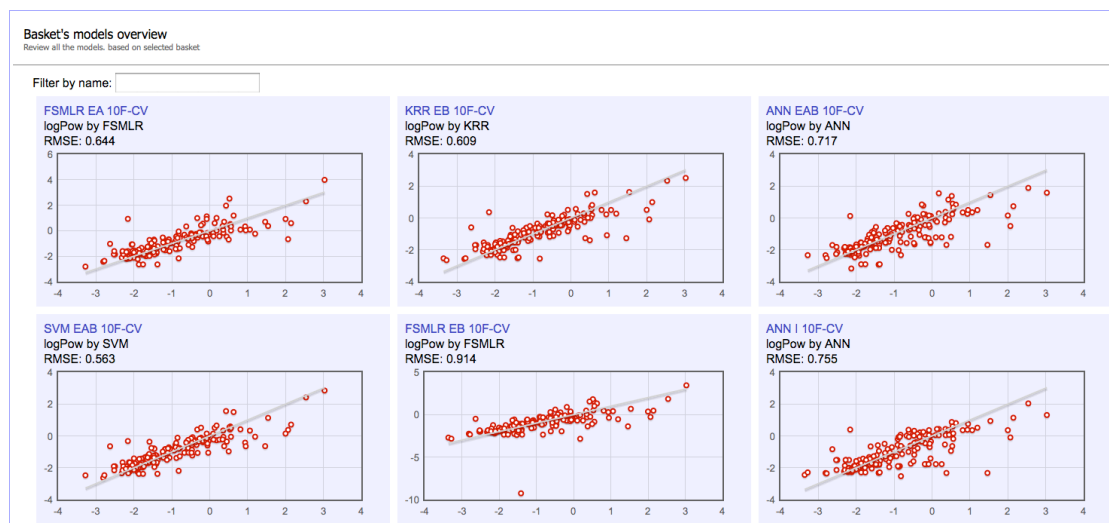


Figure 3.8. An overview of the models based on the same training set.

Automatic recalculation of models. Since the OCHEM is a public database that is populated by users, the data may contain errors. Therefore, data may be changed during verification and correction by other users over time. It may lead to a significant alteration of the training sets and to invalidation of the previously developed models. To address this problem, the OCHEM provides a possibility to recalculate existing model preserving the previous workflow parameters (e.g. by applying the same machine learning method with the same parameters and descriptors) and to compare new results with the original model.

The registry of models is a dialog that shows the previously trained and saved models (Figure 3.9). The dialog displays a brief summary of the model: the name, the predicted property, the training and validation sets and their sizes, the date of creation of the model. If necessary, it is possible to export a model in Excel or CSV format for further offline analysis. Here, the models can be published to the publicity so that every user on the Web can access the model. Such a model receives a unique web link, which makes it convenient to share models and reference them from publications.

Step 1. Select a model from the list				
Filter by model name: <input type="text" value="platinum"/>		and property name: <input type="text" value="logpow"/>		[refresh]
1 - 2 of 2				
Platinum complexes	predicts logPow using Platinum complexes I (190) validated by Platinum complexes II (36)	LogP	2009-09-18	csv excel print share
Platinum complexes Fragments	predicts logPow using Platinum complexes I (190) validated by Platinum complexes II (36)	ANN	2009-09-18	csv excel print share
1 - 2 of 2				

Figure 3.9. The registry of models in the OCHEM system.

3.3.9 Application of models

After a model has been successfully trained and saved, it can be applied to predict new compounds. The models to be applied are selected from the registry of models (Figure 3.9). The target compounds can be either provided in an SDF file or drawn manually in a molecule editor. Alternatively, the compounds can be selected from a prepared before basket.

After the molecules have been selected, a user submits a prediction job and is forwarded to the waiting screen. When the model calculation is completed, a user is

provided with the predictions by the selected models for all the target compounds. The predictions can be exported into an Excel file for offline analysis.

3.3.10 Applicability domain assessment

Importantly, the OCHEM modeling framework contains tools that allow to assess the applicability domain of the QSAR models and to estimate the prediction accuracy for each particular compound individually. The AD assessment is based on the methods developed in this thesis work and presented earlier in the “Methodology” chapter.

A. DMs and accuracy averaging

The heart of our AD assessment methodology is the concept of “distance to model” (DM). The DM concept and examples are discussed in detail on pages 15-22. The OCHEM supports a number of DMs: LEVERAGE, ASNN-STD, BAGGING-STD, CORREL, CLASS-LAG and STD-PROB. The use of each DM is limited by several (partially technical) restrictions summarized as follows.

1. The BAGGING-STD can be used with any model as long as it is validated using the bagging protocol, which provides an ensemble of the models required to calculate the standard deviation.
2. The LEVERAGE can be used if the size of the training set is bigger than the number of descriptors used for the modeling. The LEVERAGE cannot be used with LogP-LIBRARY model since this model does not involve calculation of molecular descriptors.
3. The CLASS-LAG can be used for any classification model based on neural networks and KNN methods. The STD-PROB can be used only for neural networks.
4. The ASNN-STD and CORREL can be used only with neural networks.

Once a QSAR model is trained and the chosen DMs are calculated, it is possible to analyze the prediction accuracy of the model based on a particular DM. This analysis is based on the accuracy averaging procedure (see section “Analysis of prediction accuracy” on page 22). In the OCHEM, the accuracy can be averaged using the bin-based accuracy averaging (BBA), the sliding window averaging (SWA) and the cumulative averaging. The BBA is used to estimate the prediction accuracy for new compounds, whereas the SWA – to inspect the dependency of the prediction accuracy from a DM visually. The cumulative accuracy averaging is useful for interpretation of the DM since it shows what percentage of the compounds from the training set in average are predicted with a particular accuracy. A user can interactively select the averaging type and recalculate the averaging dynamically.

To quantify the prediction accuracy, the OCHEM uses several measures: root mean square error (RMSE), mean absolute error (MAE), the Pearson correlation coefficient and the coefficient of determination for regression models and the correct classifications rate (CCR) for classification models. However, for AD assessment, only RMSE and CCR are used for regression and classification models, respectively. The

reason for the use of RMSE is that it allows an easy estimation of the confidence interval under assumption that the errors are distributed normally.

An example of BBA is shown in Figure 3.10, which is based on the BAGGING-STD DM and a KRR model for the prediction of the octanol-water partition coefficient, developed as a part of a study reported further in the “Applications” chapter. In this figure, the red dots represent the compounds from the training set, where the y-axis value corresponds to the residuals of predictions. The black “steps” represent the BBA itself, which is also summarized in Table 3.2. Thus, the values of DM less than 0.13 (the first “step” in the BBA) correspond to the highly accurate predictions with RMSE of 0.21, whereas the values of DM more than about 0.5 (the last “step” in the BBA) correspond to the predictions with a low accuracy and RMSE of 0.85.

The estimated RMSE values allow to estimate the confidence interval for a prediction. Namely, if the errors within a bin are distributed normally, then 95% of all residuals within this bin will be inside the $[-1.96\sigma, 1.96\sigma]$ interval, where σ is the standard deviation (see an example in Table 3.2).

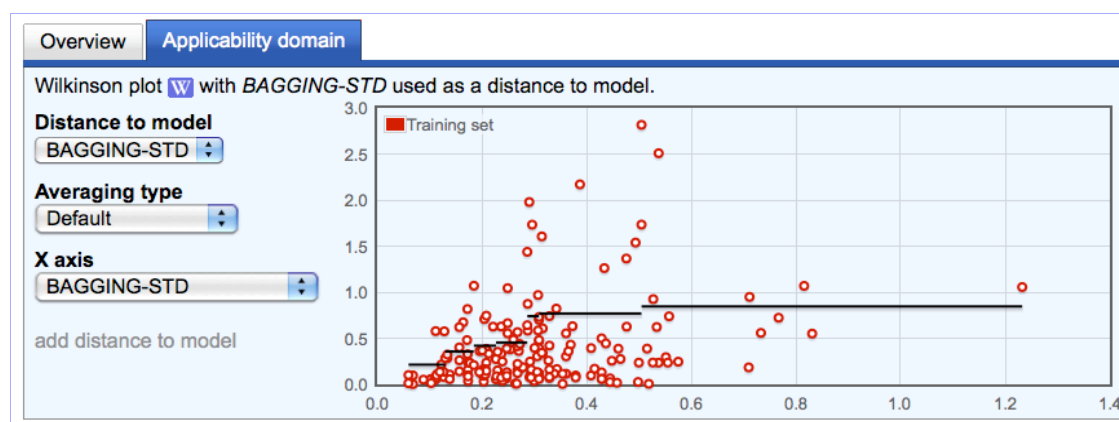


Figure 3.10. An example of the bin-based accuracy averaging, which is used in OCHEM for the estimation of prediction accuracy.

The interval of the DM	Estimated RMSE	95% confidence interval for the predictions
< 0.13	0,21	±0,42
(0.13, 0.18)	0,36	±0,70
(0.18, 0.23)	0,42	±0,82
(0.23, 0.29)	0,45	±0,89
(0.29, 0.31)	0,74	±1,45
(0.31, 0.51)	0,77	±1,51
> 0.51	0,85	±1,66

Table 3.2. The details of the bin-based average example.

B. Estimation of the prediction accuracy

Once the prediction procedure is completed, the OCHEM shows the estimated RMSE, the 95% confidence interval and the DM value for every prediction. For classification models, the correct classification rate (CCR) is used instead of RMSE. In order to discover the most reliable predictions, the compounds can be sorted by the prediction accuracy. An exemplary prediction of the inhibitory growth concentration ($\log(\text{IC}_{50}^{-1})$) by 3 different models is shown in Figure 3.11.

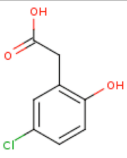
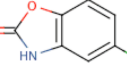
The predicted property	Model name	Prediction value	95% confidence interval	DM value
		$\log(\text{IGC50-1})(\text{Pyriformis Zhu}) = 0.36 -\log(\text{mmol/L}) \pm 0.82$	(ASNN-STDEV = 0.19, estimated RMSE = 0.42)	
		$\log(\text{IGC50-1})(\text{IG: ANN E Bag}) = 0.43 -\log(\text{mmol/L}) \pm 1.41$	(BAGGING-STD = 0.17, estimated RMSE = 0.72)	
		$\log(\text{IGC50-1})(\text{IG: SVM E Bag}) = 0.61 -\log(\text{mmol/L}) \pm 1.15$	(BAGGING-STD = 0.19, estimated RMSE = 0.59)	
		$\log(\text{IGC50-1})(\text{Pyriformis Zhu}) = 0.48 -\log(\text{mmol/L}) \pm 0.82$	(ASNN-STDEV = 0.26, estimated RMSE = 0.42)	
		$\log(\text{IGC50-1})(\text{IG: ANN E Bag}) = 0.5 -\log(\text{mmol/L}) \pm 1.41$	(BAGGING-STD = 0.19, estimated RMSE = 0.72)	
		$\log(\text{IGC50-1})(\text{IG: SVM E Bag}) = 0.33 -\log(\text{mmol/L}) \pm 1.15$	(BAGGING-STD = 0.33, estimated RMSE = 0.59)	

Figure 3.11. The prediction for new compounds and the estimation of the prediction accuracy in the OCHEM system.

3.4 Implementation aspects

The OCHEM is mainly based on the Java platform. The methods with high requirements to computational resources (ASNN, kNN, MLRA) were developed using C++ code. The data is stored in a MySQL database. All queries are executed using the Java Hibernate technology that provides an intermediate abstract layer between Java code and the database.

The JAXB library is used to create XML files and XSLT transformations to convert XML files to HTML web-pages. To connect design and functionality, we used the MVC methodology with the Java Spring framework. To make the client side dynamic and user friendly, we used Java-script and AJAX, which makes the system look more like a dynamic online application rather than a static Web site.

For chemistry-related features, we used the JME molecule editor[78], the CDK toolkits [79] and the ChemAxon (<http://www.chemaxon.com>) Standartizer. The CDK [79] is used for various chemoinformatics tasks such as preprocessing and fragmentation of molecules as well as calculation of descriptors. The visualization of molecules as well as interconversion of molecules between SDF, SMILES and MOL2 formats is done using the ChemAxon toolkit.

The OCHEM comprises about 100,000 lines of Java, C++, and shell script code. Several of its critical components, e.g., the task management system, were inspired by the Virtual Computational Chemistry Laboratory (VCCLAB, <http://www.vcclab.org>) [53].

3.5 Summary and outlook

The Online Chemical Modeling Environment (OCHEM) contains a set of tools for easy creation, publication and use of predictive models for physicochemical and biological properties. The user-contributed database allows upload of large amounts of experimental data. The database allows storing supplementary information, like conditions of experiments, units of measurements with automatic interconversions, sources of the data (scientific publications, books), etc.

The database is strongly integrated within the modeling framework; the data can be flexibly filtered and used for the training of predictive computational models. The modeling framework in the OCHEM supports all the typical steps of QSPR modeling:

data preparation, calculation and filtering of molecular descriptors, application of machine learning methods (both classification and regression), analysis of the model, assessment of the models domain of applicability and using the model to predict properties for new molecules. The complexity of the modeling process is hidden behind a convenient user interface. Models can be published on the Web and publicly used by others.

The OCHEM is available at <http://ochem.eu> and comes in two versions: the main database and the “sandbox”. The latter is intended for users to test and get acquainted with the system. Currently, the main database contains about 120,000 publicly available experimental measurements for about 300 properties. We developed tools that facilitate the migration of data from other databases and used them to introduce about 1,700,000 experimental measurements from the ChemBL database (<http://www.ebi.ac.uk/chembl/db>) to the “sandbox”. Recently, we have also uploaded more than 23,000 records from the ChemExper (<http://www.chemexper.com>) for physical properties such as boiling point, melting point and density. To keep the data up-to-date, the update server periodically uploads new data from the ChemExper database. A similar server is currently being implemented for automatic data retrieval from the ChemSpider (<http://www.chemspider.com>). The developed methodology can be used to incorporate any other database. The OCHEM modeling framework uses a cluster of more than 100 CPUs that allows calculations of models with large data sets in a reasonable amount of time.

On the practical aspect, the OCHEM was used to build the predictive models and estimate their AD for all the QSAR studies reported in this thesis work: the prediction of the Ames test, inhibition growth concentration (pIGC50) against *T. Pyriformis*, the octanol-water partition coefficient for platinum complexes and the inhibition of cytochromes.

Our intention is to make OCHEM the platform of choice to perform QSPR/QSAR studies online and share them with other users on the Web.

4 Benchmarking studies

The QSAR techniques presented in “Methodology” chapter were benchmarked in two studies, a qualitative modeling (classification) study and a quantitative modeling (regression) study. The first study involved the *in silico* prediction of mutagenicity (the Ames test); the second study – the prediction of the toxicity on *Tetrahymena pyriformis*. In both the studies, we benchmarked a number of DMs in combination with different QSAR approaches.

4.1 Prediction of Ames mutagenicity

4.1.1 Ames test and mutagenicity

Ames test is a rapid and inexpensive bacterial assay for the assessment of mutagenicity of chemical compounds. This assay uses genetically modified strains of *Salmonella Typhimurium* and checks whether a chemical compound can cause a reverse gene mutation [41]. As there is a strong evidence that a mutagenic compound is also likely to be carcinogenic [80], the Ames test can be used to identify potential carcinogens. The test is widely used in industry and pharmacology for screening and filtering out potential carcinogens on early stage before performing epidemiological surveys and expensive animal tests [81].

This study, based on the data from the Ames Challenge 2009 [82], aims to deeply investigate applicability domain of QSAR models for prediction of the Ames test results.

4.1.2 Methods and datasets

A. QSAR approaches

Twelve international teams submitted 29 models to the 2009 Ames mutagenicity challenge. All models were evaluated according to the 5-fold cross-validation procedure. Each group also developed models using the whole training set that were “blindly” applied to predict test set compounds. The consensus model was calculated by averaging predictions of all 29 individual models developed using whole training set data.

The descriptors used by the participating models included several types: Dragon descriptors, molecular fragments count (ISIDA fragments) and E-State indices (see section “Molecular descriptors” on page 5). Additionally, in this study, a number of research groups used two specific descriptor types: inductive electronegativity scale[83] (referred to as “inductive descriptors”, abbreviated as ID) and SiRMS[84] (**S**implex **R**epresentation of **M**olecular **S**tructure).

The machine learning methods included simple and associative neural networks (NN and ASNN), PLS, linear regression, decision tree, random forest, SVM, KNN and

naïve Bayes classifier (see section “Machine learning methods” on page 7). All the participating groups are summarized in Table 4.1 and the participating QSAR models in Table 4.2.

Group name	Abbreviation	The number of provided models	Models with numeric prediction values
University of Insubria	UI	2	
Technical University of Berlin	TUB	2	1
Lanzhou University	ULZ	2	
Linnaeus University	LNU	1	1
Helmholz-Zentrum Munich, OCHEM group	OCHEM	1	1
University of British Columbia	UBC	2	
Louis Pasteur University, Laboratory of Chemoinformatics	ULP	4	4
Moscow State University	MSU	2	2
Physico-Chemical Institute of the National academy of Sciences of Ukraine	PCI	3	3
University Milano-Bicocca	UMB	1	
University of North Carolina	UNC	7	6
Environmental Protection Agency, EPA	EPA	2	2
Total models		29	20

Table 4.1. The 12 international groups and their models for the Ames test predictions.

Model name	Descriptors used	Training method	Numeric predictions	Own DM
CONS	-	-	+	-
EPA_2D_FDA	PCID	FDA	+	ELLIPS
EPA_2D_NN	PCID	Neural networks	+	SCAvg
LNU_Drag_PLS	Dragon	PLS	+	
MSU_FRAG_LR	Fragments	Linear regression	+	
MSU_FRAG_SVM	Fragments	SVM	+	SVM1 AD
OCHEM_ESTATE_ANN	E-State indices	Associative neural networks	+	
PCI_Drag_RF	Dragon	Random forest	+	
PCI_SiRMS.Drag_RF	SiRMS	Random forest	+	
PCI_SiRMS_RF	SiRMS	Random forest	+	
TUB_3DDrag_RF	Dragon	Random forest		DA Index
TUB_3DDrag_SVM	Dragon	SVM		DA Index
UBC_ID_IWNN	Inductive descriptors	Weighted NN		
UBC_ID_NN	Inductive descriptors	NN		
UI_Drag_KNN	Dragon	KNN		
UI_Drag_LDA	Dragon	LDA		
ULP_ISIDA_NB	ISIDA Fragments	Naïve Bayes	+	Trust level
ULP_ISIDA_SQS	ISIDA Fragments	SQS	+	Trust level
ULP_ISIDA_SVM	ISIDA Fragments	SVM	+	Trust level
ULP_ISIDA_VP	ISIDA Fragments	Voted Perceptron	+	Trust level
ULZ_3DDrag_KNN	Dragon	KNN		
ULZ_3DDrag_SVM	Dragon	SVM		
UMB_Drag_DT	Dragon	Decision Tree		
UNC_Drag_KNN	Dragon	KNN		
UNC_Drag_RF	Dragon	Random forest	+	
UNC_Drag_SVM	Dragon	SVM	+	AD Mean
UNC_SiRMS.Drag_RF	SiRMS+Dragon	Random Forest	+	
UNC_SiRMS.Drag_SVM	SiRMS+Dragon	SVM	+	AD Mean
UNC_SiRMS_RF	SiRMS	Random forest	+	

Table 4.2. The models that participated in the Ames test AD benchmarking.

The chosen classifications labels were “-1” and “1” for non-mutagenic and mutagenic compounds, respectively. Numeric (continuous) prediction values were available only for 20 models including the consensus model, while the other 9 models had only discrete (“-1” and “1”) predictions available. As some of the investigated DM required continuous prediction values, only these 20 models were used in most of the further analysis. Several DMs that could be used with qualitative predictions were applied to all 30 models, including the consensus model (see Tables 4.1 and 4.2).

B. Applicability domain assessment

The AD assessment was based on a number of reliability measures referred to as distances to models (DMs). We compared a number of general DMs (STD, concordance, leverage, STD-PROB, CORREL, CLASS-LAG); these measures are described on pages 15-18 of the “Distances to models” section, which provides the general definition for DM as well as the description of every particular DM. Three STD measures were analyzed: the standard deviation of a neural networks ensemble (ASNN-STD) and the standard deviation of qualitative and numeric predictions (CONS-STD and CONS-STD-QUAL based on 20 and 29 models, respectively). Based on these three STD DMs, three corresponding STD-PROB DMs were analyzed: ASNN-STD-PROB, CONS-STD-PROB and CONS-STD-QUAL-PROB.

Additionally, we analyzed a number of individual DMs provided by participants of the Ames challenge separately:

DA-Index. The applicability domain employed by the TUB group is based on the kappa, gamma and delta indices introduced by Harmeling et al. [85]. The first two indices are heuristics that have been previously used in the chemoinformatics community: kappa and gamma are the maximum and the mean Euclidian distances to the k-nearest neighbors in the descriptor space. The delta index corresponds to the length of the mean vector to the k nearest neighbors. Since kappa and gamma are only based on distances, they do not explicitly indicate whether interpolation or extrapolation is expected for prediction. The delta index is capable of this distinction and indicates the degree of extrapolation. Input descriptors for all indexes were weighted following the development of Gaussian Process Classification model [86]. The arithmetic mean values of gamma and delta indices were used to estimate prediction confidence. A threshold value determined using training set was used to decide whether a test-compound was inside or outside of the AD. The output of this decision process was denoted DA-Index.

AD_MEAN values were provided by UNC group for SVM models that were developed using three sets of descriptors (SiRMS, Dragon and combined). AD_MEAN corresponds to the average Euclidean distance between a compound and its three nearest neighbours in the training set. The distances were calculated using the entire pool of descriptors. As AD_MEAN values were available for two models, UNC_SiRMS_SVM and UNC_Drag_RF, we investigated the two respective measures, AD_MEAN1 and AD_MEAN2.

ELLIPS values were calculated using the EPA_2D_FDA model. According to this DM, a prediction is within the applicability domain of the model if the test chemical is within the multidimensional ellipsoid defined by the ranges of descriptor values for the chemicals in the cluster (for the descriptors appearing in the cluster model). The model ellipsoid constraint is satisfied if the leverage of the test compound (h_{00}) is less than the maximum leverage value (h_{max}) for all the compounds used in the model [87]. The ratio

h_{00}/h_{max} was denoted as ELLIPS.

SCAvg (average similarity coefficient). The cosine similarity coefficient to the three nearest neighbours used in the EPA_2D_NN method was used as **SCAvg DM**.

Two groups (ULP and MSU) classified predicted molecules into several classes with different prediction quality as described below.

Trust level. The applicability domain for the models provided by ULP group was based on a measure referred to as *trust level*. This measure has values in the range of {1,2,..5}, where the "5" corresponds to the highest trust level ("optimal") and 1 - to the lowest trust level ("none"). The trust score for a particular compound is based on 3 factors: (i) the number of models having the compound in their local applicability domain, MINDIFF-OK, as described in [88], (ii) number of dissident predictors in the set (i.e., models that gave predictions, different from the mean prediction) and (iii) the average prediction value, where values close to 0.5 are considered as less reliable. The further details on the calculation of trust score are shown in Figure A2 in Appendix on page 134.

SVM1 AD. The Applicability Domain for the MSU group models was computed using the one-class classification approach (novelty detection) based on the 1-SVM method [89]. The parameters of 1-SVM method were chosen as follows: the RBF-kernel parameter γ was taken from the same value used for building classification SVM models, while the value of ν was fixed at 0.01.

The SVM1 AD procedure associates the applicability domain of QSAR/QSPR models with the area in the input descriptor space where density of training data points exceeds a certain threshold. The main assumption of this procedure is that the predictive performance of the models tends to be high for the test compounds inside the high density areas, since outside the high density area all test objects are located far from training objects, which makes interpolation of properties from training to test objects unreliable. Instead of searching a decision surface separating high and low density areas in the input space, the one-class classification 1-SVM approach looks for a hyperplane in the feature space associated with the RBF-kernel. The ability of novelty detection models to be used as AD of machine learning models was earlier demonstrated by Bishop [90]. The use of 1-class SVM novelty detection method to assess the applicability domain of models based on structured graph kernels has recently been suggested by Fechner et al [91].

In summary, 16 different DMs were analyzed in this study: 3 STD measures, 3 STD-PROB measures, CONCORDANCE, CLASS-LAG, CORREL, DA Index, AD_MEAN, ELLIPS, SCAvg, Trust level and SVM1-AD. Among these DMs, 5 were based on the space of descriptors and 11 – on the space of properties.

C. Benchmarking criteria

For most of the investigated models, the SWA and BBA averaging resulted into noisy dependencies (they do not fall monotonically, see an example in Figure 4.1) and, thus, BBA was inconvenient to identify the accuracy coverage. Therefore, to compare the performances of the DMs, we used the accuracy coverage criterion based on the cumulative averaging (see page 24 for the definition of the cumulative averaging).

As it was shown in the further analysis (page 62), the accuracy of inter-laboratory variations within the Ames challenge dataset was 90%. Therefore, the 90% accuracy threshold was selected for the accuracy coverage criterion. The accuracy coverage values for the training and test sets were denoted as $C_{TRAIN-90\%}$ and $C_{TEST-90\%}$.

Complementary to the accuracy coverage, we used the area under curve (AUC) criterion to validate the results.

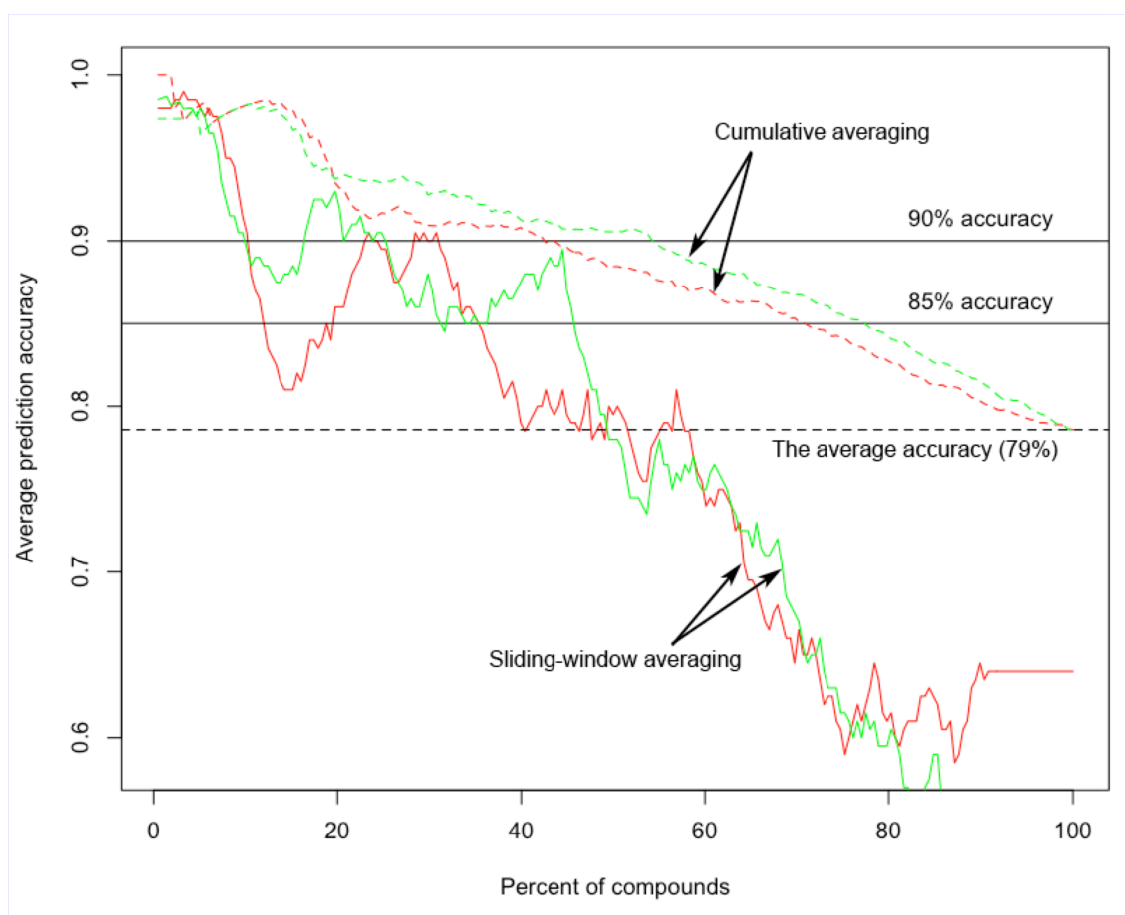


Figure 4.1. The prediction accuracy of the neural network model as a function of CONS-STD and CONS-STD-PROB. The solid lines (sliding-window averaging) show the averaged accuracy on the moving window with 200 compounds. Although there is a trend that the accuracy of prediction decreases with an increase of the DMs, the dependency is not smooth and there are significant fluctuations. The cumulative averaging (dashed lines) smooths the variations, which makes it more suitable for the comparison of DMs.

4.1.3 Results and analysis

A. Comparison of distances to model

As mentioned in the methods section above, the comparison of DMs was based on two criteria: the accuracy coverage ($c_{TRAIN-90\%}$ and $c_{TEST-90\%}$, for training and test set respectively) and the area under curve (AUC). Instead of comparing the absolute values of the criteria, we gave a rank to each DM for every model. Thus, the DM with the highest value of a criterion received rank “1”, the second-highest DM – rank “2” and so on. Thereafter, the ranks were averaged over all 20 models.

The results for $c_{TRAIN-90\%}$ and $c_{TEST-90\%}$ criteria are summarized in Table 4.3, where the DMs are sorted accordingly to their rank based on $c_{TEST-90\%}$ value. The complete set of $c_{TRAIN-90\%}$ and $c_{TEST-90\%}$ values can be found in Table A1 in Appendix, page 127. The CONS-STD-QUAL-PROB measure appeared to be the best one, considering averaged ranks over all models on the test set. According to the bootstrap test, the top 3 models (CONS-STD-QUAL-PROB, CONDCORDANCE and CONS-STD-PROB) were not significantly different from each other with $p > 0.05$ for both the training and test sets, but were significantly better ($p < 0.05$) than the other investigated measures.

The ranks based on the accuracy coverage (Table 4.3) were not significantly different from those based on the AUC (Table 4.4). More precisely, the ranks changed for the 4 lowest-ranked DMs (LEVERAGE, SCAvg, CORREL and AD_MEAN2), which were, however, not significantly different from each other. The single significant difference of the AUC ranks from the accuracy coverage ranks was that, according to the AUC criterion, the CLASS-LAG outperformed the ASNN-STD-PROB. Thus, the two criteria provided consistent results and, for all the further analysis, we used the accuracy coverage criterion ($c_{TRAIN-90\%}$ and $c_{TEST-90\%}$) because of its simpler and more intuitive interpretation.

Remarkably, the DMs based on the descriptor space (AD_MEAN1 and AD_MEAN2, ELLIPS, SCAvg and LEVERAGE) are in bottom of the ranking list according to both the criteria. It leads to the assumption that the property-based DMs (with an exception of CORREL) perform systematically better than the descriptor-based DMs.

According to the PCA (principal components analysis) plot on Figure 4.2, some of the models were quite similar, since they were based on the same descriptors and machine-learning methods, e.g. UNC_Drag_RF and PCI_Drag_RF, PCI_SiRMS_RF and UNC_SiRMS_RF. Indeed, agreement within these pairs of models was about 95%, whereas the average pairwise agreement of the models was only 83%. Therefore, it might be reasonable to combine these models when calculating the consensus-based DMs (CONS-STD, CONS-STD-PROB) to avoid redundancy. However, combining these four models in two by averaging their predictions did not affect the DM-based sorting of compounds. Therefore, the rankings of the DMs, given in Tables 4.3 and 4.4 were not affected after similar models were united.

Distance to model	Average rank (according to the accuracy coverage)	
	Training set	Test set
CONS-STD-QUAL-PROB	2,15	1,83
CONCORDANCE	1,65	2,15
CONS-STD-PROB	3,38	2,95
CONS-STD-QUAL	3,7	4,95
ASNN-STD-PROB	6,4	5,48
CONS-STD	4,88	5,75
CLASS-LAG	7,5	6,68
ASNN-STD	8,4	7,78
ELLIPS	9,15	8,98
AD_MEAN1	12,43	10,18
CORREL	10,35	11,65
SCAvg	11,08	11,85
AD_MEAN2	11,3	12,33
LEVERAGE	12,65	12,48

Table 4.3. The average ranks of the DMs considering their coverage of the 90% prediction accuracy.

Distance to model	Average rank (according to the AUC criterion)	
	Training set	Test set
CONS-STD-QUAL-PROB	2,15	1,95
CONCORDANCE	1,4	2,1
CONS-STD-PROB	3,4	2,75
CONS-STD-QUAL	3,8	4,9
CLASS-LAG	6	4,95
ASNN-STD-PROB	6,4	5,65
CONS-STD	5,3	6,1
ASNN-STD	8,05	7,9
ELLIPS	12,1	9,6
AD_MEAN1	10,9	11,25
LEVERAGE	12,85	11,3
SCAvg	11,6	11,7
CORREL	9,95	11,85
AD_MEAN2	11,1	13

Table 4.4. The averaged rankings of the DMs ranked by the AUC criterion.

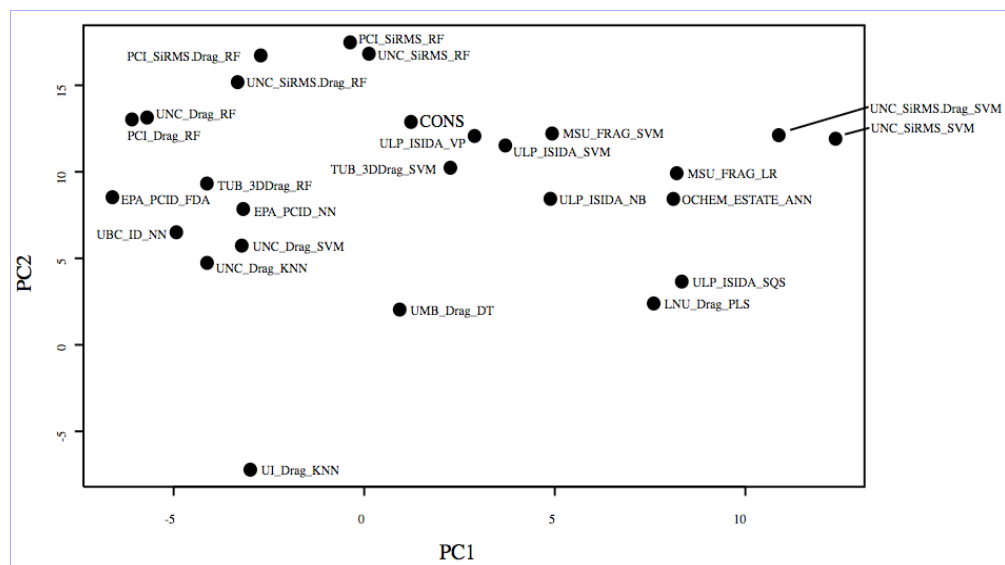


Figure 4.2. The PCA plot of the Ames challenge models based on the space of predictions for the test set. Four models (UI_Drag_LDA, UBC_ID_IWNN, ULZ_3DDrag_SVM, ULZ_3DDrag_KNN) are not shown, since they were apparent outliers.

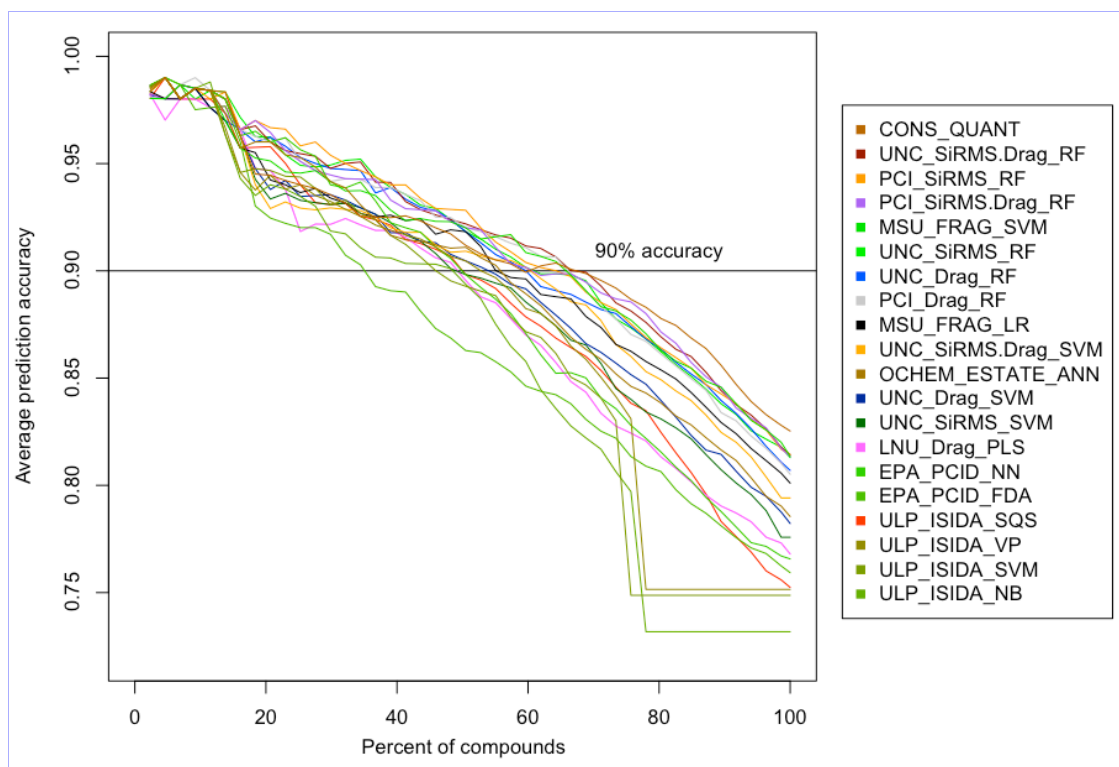


Figure 4.3. The cumulative accuracy-coverage plot for CONS-STD-PROB DM based on the test set predictions. The curves show the accumulative accuracy for a particular (variable) percentage of compounds. The curves clearly show that CONS-STD-PROB is highly correlated with the prediction accuracy.

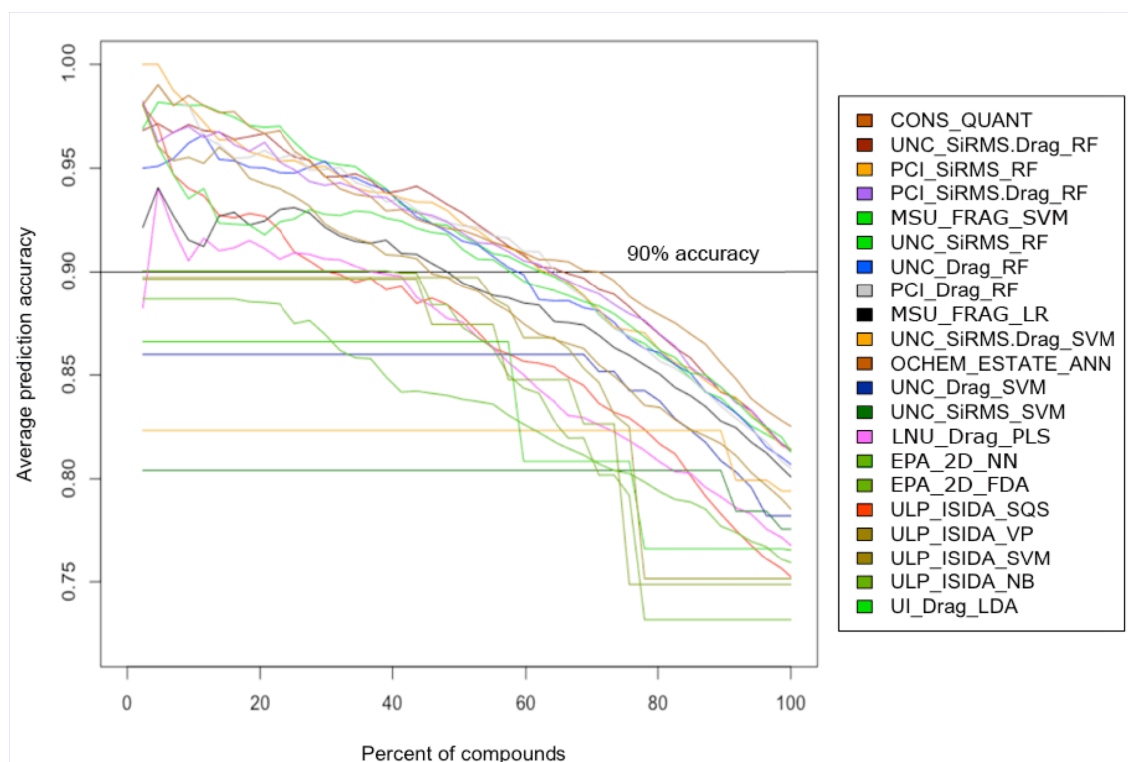


Figure 4.4. The cumulative accuracy-coverage plot for CLASS-LAG based on the test set predictions.

The dependency of the model performances from the CONS-STD-PROB DM is shown on the cumulative accuracy-coverage plot (Figure 4.3). The plot indicates that 35-70% of all compounds (depending on the model) are predicted with 90% accuracy.

The same kind of plot for the CLASS-LAG DM (Figure 4.4) reveals its poorer performance. The difference is visually apparent: for a part of the models the CLASS-LAG was not able to separate predictions with 90% accuracy. In Figure 4.4, these models correspond to the curves under 90% line. The visual difference in Figures 4.3 and 4.4 clearly indicates that combining CLASS-LAG with STD (resulting in CONS-STD-PROB) increased the quality of AD assessment.

As it can be observed in Figure 4.4, the performance of the CLASS-LAG DMI was very dependent on a model (see also the complete table of accuracy coverages in Table A1 on page 127 in Appendix). The poor performance of CLASS-LAG for a part of the models may be explained by the specifics of the distribution of their predictions. Indeed, the two histograms in Figure 4.5 reveal that the prediction values of UNC_SiRMS_SVM are similar to discrete values $\{-1, 1\}$. Therefore, these prediction values contained less information than predictions by PCI_SiRMS.Drag_RF, which were distributed more uniformly. Indeed, the CLASS-LAG DM failed for the first model ($c_{Test-90\%} = 0\%$ coverage), and yielded excellent results for the second one ($c_{Test-90\%} = 62\%$ coverage).

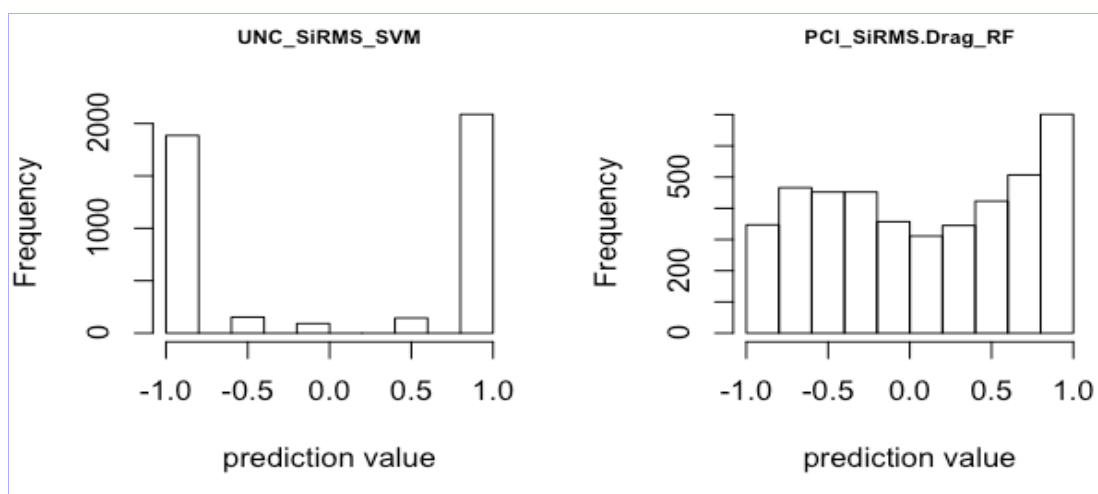


Figure 4.5. The distribution of the prediction values for two exemplary models. The prediction values of the model on the left chart resemble rounded discretized “-1” and “1” values, whereas the values on the right chart have a continuous distribution and, therefore, provide more information for the estimation of prediction reliability. This fact is confirmed in practice: CLASS-LAG of UNC_SiRMS_SVM (left chart) has poor performance (0% coverage of 90% accuracy) in contrary to PCI_SiRMS.Drag_RF (right chart), which separates 63% of compounds with 90% prediction accuracy.

The measures, on which DA Index was based (namely, delta-index and gamma-index), did not outperform DA Index itself. Therefore, they were not analyzed separately. The LEVERAGE DM could not separate 90% accuracy predictions for any model altogether ($c_{TEST-90\%} = c_{TRAIN-90\%} = 0$); therefore, it was not analyzed further.

Figure 4.6 demonstrates that AD_MEAN (solid lines) results that are apparently worse as compared to CONS-STD-PROB (dashed lines).

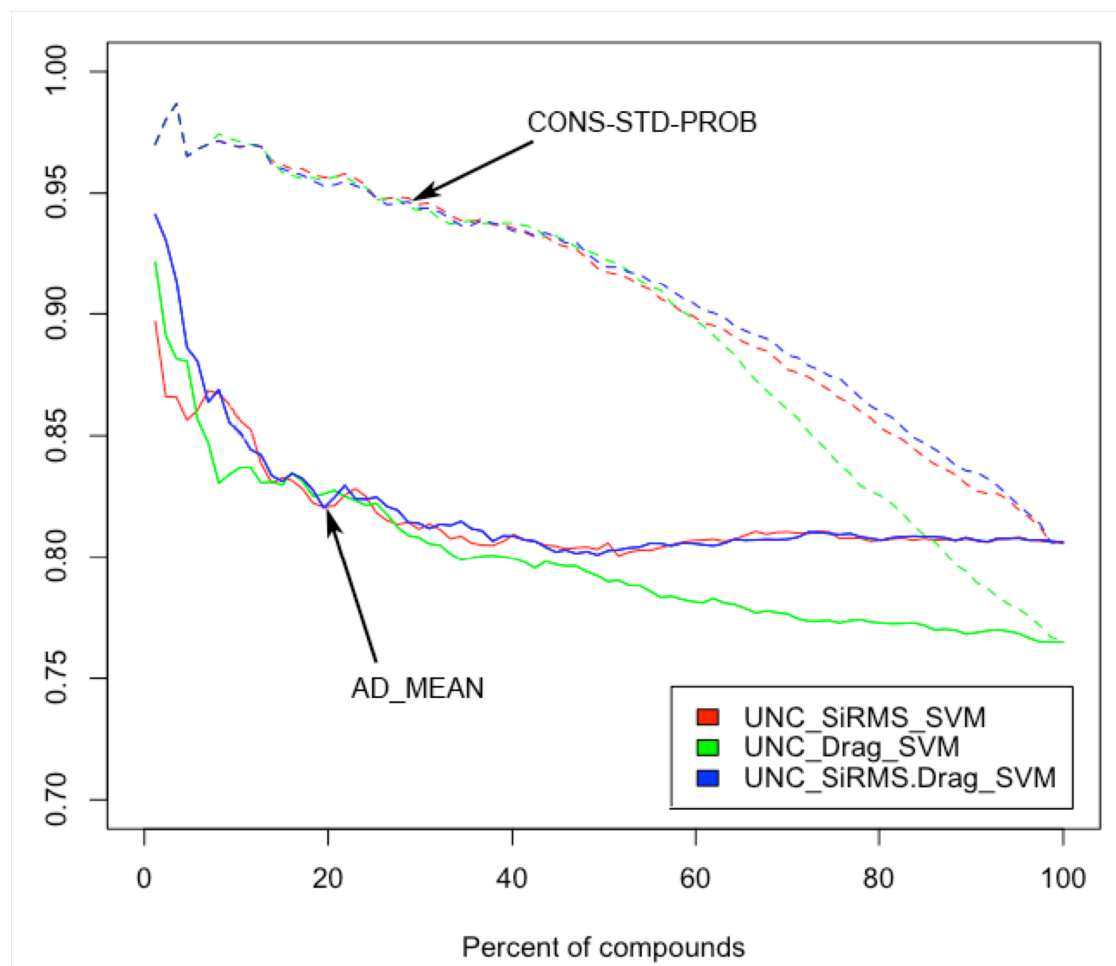


Figure 4.6. A comparison of AD_MEAN DM (solid lines) with CONS-STD-PROB DM (dashed lines) for the UNC SVM models. Apparently, CONS-STD-PROB provides a better separation of highly accurate predictions.

It was interesting to compare the performance of descriptor-based and property-based DMs. The DMs based on the descriptor space (LEVERAGE, DA Index, ELLIPS, SCAvg and AD_MEAN) identified only small percentages of molecules with > 90% accuracy. For example, the LEVERAGE and DA Index DMs completely failed to identify compounds 90% ($c_{TRAIN-90\%} = c_{TEST-90\%} = 0$)², whereas AD_MEAN could identify such highly accurate predictions only for a small part of the models. ELLIPS was successful for almost all the models (17 out of 20 models); however, its $c_{TRAIN-90\%}$ and $c_{TEST-90\%}$ values were significantly lower than that of the property-based DMs (e.g., 0%-16% for ELLIPS and 41%-69% for CONS-STD-PROB). Thus, the property based DMs performed significantly better than the descriptor-based DMs.

² For full details on the accuracy coverage for all models, refer to Table A1 in Appendix on page 127

A remarkable point is that the percentage of active (mutagenic) compounds within the area of 90% prediction accuracy is 51-55%, which is not significantly different from the percentage of active compounds in the whole test set (53%). Therefore, mutagenic compounds are neither overrepresented nor underrepresented in the applicability domain of the models. Moreover, prediction accuracy, sensitivity and specificity of all the models were not significantly different within the area of 90% prediction accuracy. Thus, a separate analysis of specificity and sensitivity would have been redundant and was not performed.

The PCA plot of the DMs (Figure 4.7), calculated using the DM-based rankings of Ames challenge compounds, reveals the high similarity of the five DMs, which are based on the global consensus model. Indeed, these models explore slightly different aspects of basically the same data and are strongly intercorrelated (see Table A2 in Appendix on page 128). The CONS-STD, CONS-STD-QUAL and CONCORDANCE DMs form one cluster within which the CONCORDANCE DM provided the best discrimination of the highly accurate predictions (Tables 4.3 and 4.4).

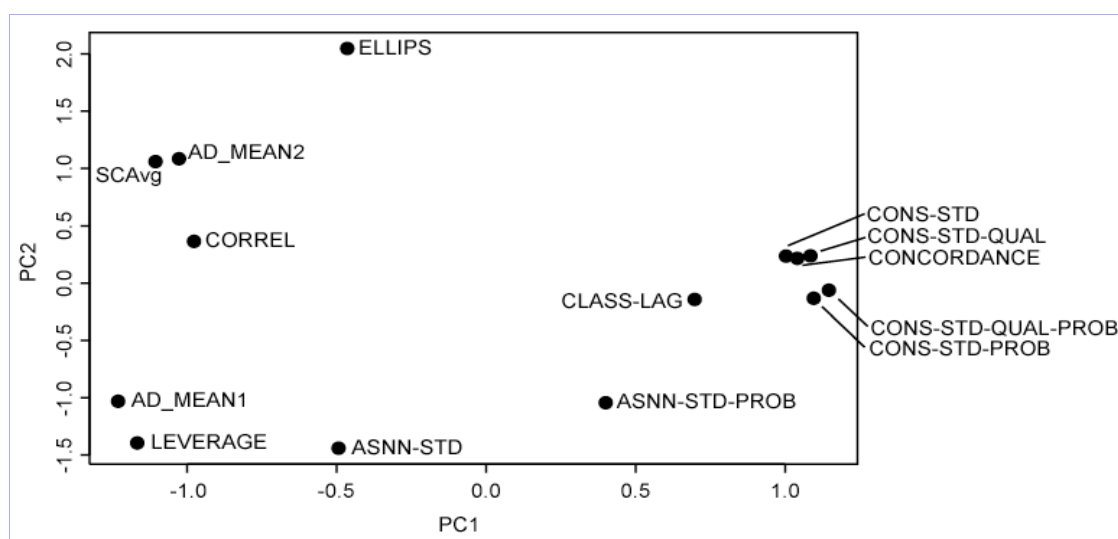


Figure 4.7. A principal components plot for the analyzed DMs. The PCA was based on the rankings that the DMs gave to the compounds from the training and test sets. Apparently, the 5 consensus-based DMs form two close clusters: CONS-STD, CONS-STD-QUAL and CONCORDANCE in the first cluster and CONS-STQ-QUAL-PROB and CONS-STD-PROB in the second one.

B. Analysis of the qualitative AD measures

As it was mentioned above, several groups provided qualitative AD measures for their models. Here, these measures are compared to the CONS-STD-PROB DM. For reference, the performance of CONST-STD-PROB for the relevant models binned on several intervals is shown in Table 4.5.

Number of compounds	The prediction accuracy				
	ULP_ISIDA_SQS	TUB_3DDrag_SVM	TUB_3DDrag_RF	MSU_FRAG_LR	MSU_FRAG_SVM
500	96%	93%	93%	94%	95%
500	86%	89%	90%	89%	90%
500	76%	79%	81%	80%	83%
500	53%	64%	65%	66%	68%
181	48%	61%	55%	54%	54%
2181	75%	80%	80%	80%	81%

Table 4.5. Accuracy of predictions according to the CONS-STD-PROB. For first 500 compounds, it achieved the accuracy as high as 93-96%.

TRUST LEVEL. This AD-related information provided by ULP group is a generic estimation of the degree of trust for the prediction of a particular compound. The trust level had values of OPTIMAL, GOOD, MEDIUM and POOR depending on the agreement of individual models and the number of models that had the compound in their ADs (see the detailed schema in Figure A2 on page 134 in Appendix). We grouped the test set compounds by trust level and computed de-facto prediction accuracy within each group. The results are summarized in Table 4.6.

Trust level	Number of compounds	Observed prediction accuracy	
		Trust level	CONS-STD-PROB
OPTIMAL	1,221	81%	89%
GOOD	512	79%	69%
MEDIUM	415	53%	46%
POOR (or less)	33	70%	45%
Overall test set	2,181		75%

Table 4.6. The CONS-STD-PROB and TRUST LEVEL juxtaposed for the ULP_ISIDA_SQS model.

Prediction accuracy apparently dropped with decrease of trust level (excluding the POOR trust level that included only 33 compounds, which may not be sufficient for the evaluation of prediction accuracy). Apparently, this measure had worse results than CONS-STD-PROB. 1,221 most reliable predictions had the accuracy of 89% according to CONS-STD-PROB and only 81% according to TRUST LEVEL.

One-class classification AD (SVM1 AD). This measure was provided by MSU group. Accuracies grouped by the SVM1 AD are summarized in Table 4.7. Majority of the compounds from training and test sets were predicted to be inside AD. The prediction accuracy for these compounds was on average 5% higher than those outside of AD. The CONS-STD-PROB method provided much better separation of molecules with differences up to 40% for reliable and non-reliable predictions (Table 4.5).

SVM1 AD	Number of compounds	Observed prediction accuracy			
		MSU_FRAG_LR		MSU_FRAG_SVM	
		SVM1 AD	CONS-STD-PROB	SVM1 AD	CONS-STD-PROB
<i>Training set</i>					
Inside (= 1)	4194	79%	80%	80%	81%
Outside (= -1)	167	75%	59%	79%	53%
Overall training set	4361		79%		80%
<i>Test set</i>					
Inside (= 1)	2046	81%	82%	81%	83%
Outside (= -1)	135	73%	53%	79%	55%
Overall test set	2181		80%		81%

Table 4.7. The CONS-STD-PROB and SVM1-AD DMs juxtaposed for the MSU models.

DA Index. The TUB group provided the DA-Index DM summarized in Table 4.8. The most compounds (1,819, or 83% of the test set) had DA-Index value of 0, which corresponds to the highest expected accuracy. However, the increase of the accuracy 2-6% was not significant for both TUB models, TUB_3DDrag_SVM and

TUB_3DDrag_RF. For the same models, 500 most accurately predicted compounds identified using CONS-STD-PROB had 93% classification accuracy for both models, Table 4.5.

DA Index	Number of compounds	Observed prediction accuracy			
		TUB_3DDrag_SVM		TUB_3DDrag_RF	
		DA Index	CONS-STD-PROB	DA Index	CONS-STD-PROB
0	1819	81%	83%	80%	84%
Between 0 and 1	183	75%	62%	78%	61%
1	179	75%	60%	80%	60%
Overall test set	2181	80%		80%	

Table 4.8. The CONS-STD-PROB and DA-Index DMs juxtaposed for the TUB models.

Tables 4.5-4.8 show that the CONS-STD-PROB could separate predictions of high and low accuracy better than any of the investigated qualitative DMs. For example, the 681 most unreliably predicted compounds (i.e., molecules with the largest CONS-STD-PROB values) had the accuracy as low as 52% (Table 4.5), i.e. almost the same as the accuracy of a random guess. None of the qualitative DM measures could identify predictions of such low accuracy.

Moreover, none of the qualitative DMs allowed to identify predictions with the accuracy of 90%, which corresponds to the inter-laboratory agreement of the Ames test. On the contrary, 500 of the most reliably predicted compounds according to CONS-STD-PROB DM had accuracy of as high as 95-96%.

In summary, the CONS-STD-PROB performed better than all the aforementioned qualitative DMs: TRUST LEVEL, SVM1-AD and DA-Index.

C. Ability to estimate the prediction accuracy

So far, we investigated the ability of DMs to distinguish accurate and inaccurate predictions. As the main criteria for this kind of performance, we used the percentage of compounds that were predicted with 90% accuracy on the training and test sets ($c_{TRAIN-90\%}$ and $c_{TEST-90\%}$) identified by the analyzed DM. However, it is also important to ensure that a DM is capable of estimating the prediction accuracy for new compounds.

Under assumption that a model is correctly cross-validated and the investigated DM is consistent, the prediction accuracy for the compounds from the training set and the test set within the same DM threshold should not be significantly different. Thus, the DM threshold selected for the training set should provide about the same prediction accuracy on the test set.

In order to check this assumption, we selected a DM threshold that provides the 90% prediction accuracy on the training set and calculated the prediction accuracy for the compounds within the same threshold on the test set. The comparison was performed for all the models in combination with all the investigated DMs. There were 20 models tested against 12 DMs, which resulted into $20 \times 12 = 240$ comparison cases. We found that the prediction accuracy for the training and test sets was consistent with significance level $p=0.01$. With the significance level of $p=0.05$, the estimated and observed accuracies were significantly different only for 2 cases, which does not exceed the statistically expected number of failures (for 240 comparison cases, 12 failures at 0.05 level of significance).

Thus, the estimated accuracy based on the training set was in agreement with the actual accuracy observed on the test set.

D. Interpretation of the AD

To determine which types of molecules are predicted accurately and, in contrary, which types of molecules have the lowest predictions accuracy, we analyzed the molecular sub-fragments which tend to induce high and low prediction accuracy. The methodology of this analysis is described in the methodological section “Interpretation of applicability domains” on page 27.

An overview of the significant fragments is presented in Figure 4.8. Apparently, the molecules containing long carbon chains, nitro-groups, thiophene-groups as well as acridine- and phenathrene derived molecules were overrepresented in the applicability domain. After a more detailed investigation, we found out that long non-saturated carbon chains were mostly presented in non-mutagenic compounds, whereas nitro- and thiophene-groups as well as acridine and phenathrene – in mutagenic compounds. For the prediction of such compounds, there was a high level of agreement between the models (and, therefore, such compounds had low values of CONS-STD-PROB). In contrary, the compounds containing halocarbons, sulfonate- and epoxide-groups were not reliably predicted by the QSARs investigated in this study.

Presumably, the low accuracy for particular sub-fragments was due to a particular mechanism of mutagenicity. For example, the mechanism of action of C-C-Halogen fragments can be explained by the electrophilic attack of a compound on the DNA backbone. A partial positive charge on a carbon atom is attracted to a negative charge of an oxygen or nitrogen atom in the DNA backbone, which results into covalent bonding to DNA with halogen release. Alkylated DNA is then prone to replication problems and/or impaired information transfer for protein synthesis, which ultimately lead to cell mutation or apoptosis.

As such fragments are very reactive, the aforementioned mechanism of mutagenicity can be affected by metabolism of the compound by liver cells, which makes the prediction particularly difficult. For the stable and less reactive compounds containing fused aromatic rings (i.e. acridines and phenathrenes), the metabolism is not likely to affect the structure and, thus, their mutagenic effect is stable and is easily predicted by the models.

E. Data variability analysis

There were several studies that analyzed the variability of Ames test experiments. Let us critically review them for a better understanding of the results of our modeling.

The first study by Benigni et al [92] assessed the Ames tests carried out by 12 laboratories for 42 compounds. For every pair of laboratories, we calculated the level of agreement as the number of the concordant measurements divided by the total number of measurements. The distribution of agreements of 66 lab-pairs is shown on Figure 4.9. The average pairwise agreement is only 75%. At the same time, Figure 4.9 reveals that agreement of results between some laboratories can be sometimes higher than 90%. This result was observed for 4 out of 66 pairs of laboratories (7% of all data). Moreover, it is possible to expect a higher agreement if the data is measured within same laboratory.

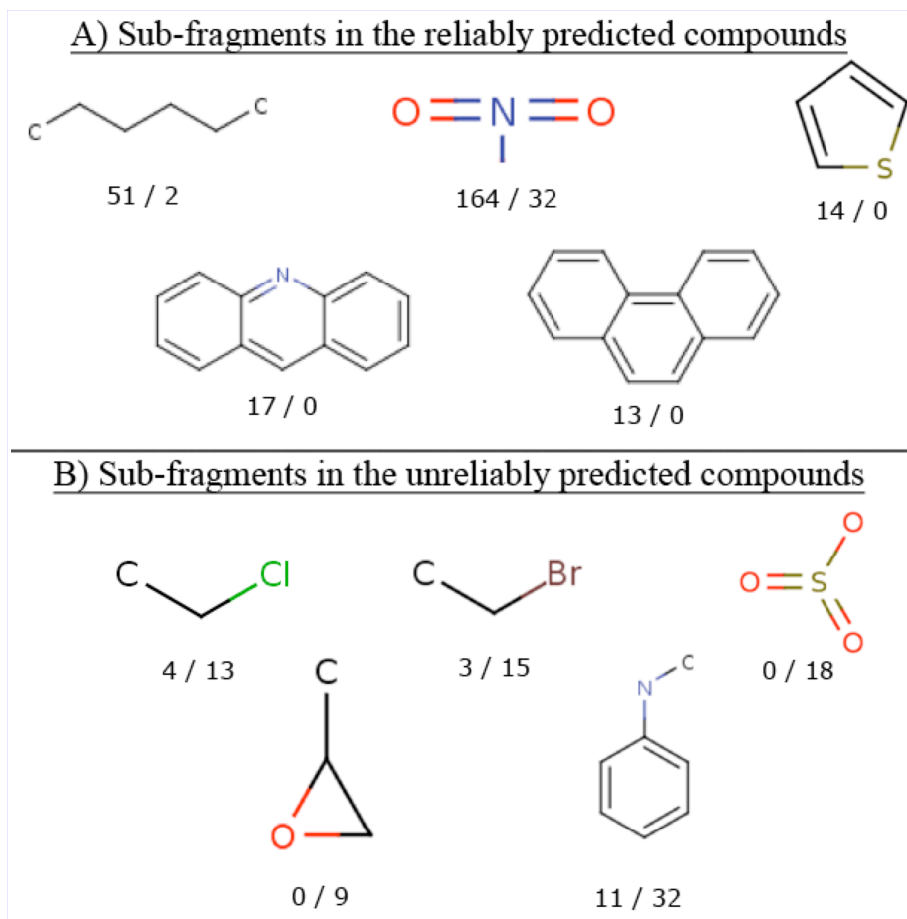


Figure 4.8. The molecular sub-fragments presented in the reliably and non-reliably predicted compounds. Shown are the sub-fragments that were significantly overrepresented in the molecules having the most and the least reliable 400 predictions (A and B) according to the CONS-STD-PROB DM. Below the fragments are the numbers of the relevant molecules with the most reliable (left of the dash) and the least reliable (right of the dash) predictions.

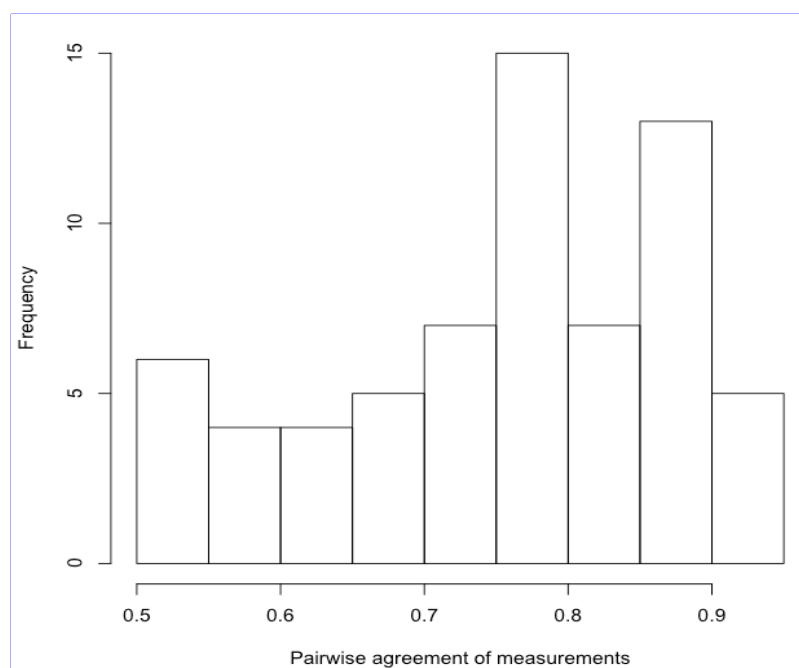


Figure 4.9. The distribution of the pairwise agreements of the Ames test measurements carried out by 12 laboratories. The 0.5 value on x-axis corresponds to the complete disagreement of two laboratories. The data for the plot was taken from a study by Benigni et al [92].

In the study by Piegorsch and Zeiger [93], the experimental concordance between different laboratories was reported in range of 70 to 87% for different definitions for the concordance measures. Each molecule in this set was measured in several experiments either in different labs or in the same lab but at different times. The outcomes of experiments were positive (+), weak positive (+W), negative (-) and questionable (?). Similar to as it was done in the analysis by the original authors, we considered positive and weak positive as AMES mutagens and ignored non-decisive experiments, which, are usually expected to be re-measured. Then, we defined the accuracy of a measurement of a compound as the maximum number of positive or negative tests divided by the total number of the decisive experiments. Such accuracy could be expected for our analysis, if we assume that molecules were tested on average just once. The average accuracy of the AMES test was 93% and 90% if we considered molecules with at least two (209 molecules) or three (49 molecules) decisive measurements, respectively.

We further explored this result by estimating the variability of the measurements used in our study. For this analysis, we used the Ames test data collected and publicly available at the OCHEM <http://ochem.eu>. The database contains results for 3,205 of the 6,542 Ames challenge compounds. We used the same definition of accuracy as above and calculated average accuracy of 94% for the compounds that had at least three measurements (1,680 compounds selected from 189 articles). The variation of the minimal number of measurements from 4 to 7 did not change this number for more than $\pm 0.3\%$. The 94% agreement is conformable with the achievable prediction accuracies of the models investigated in this study.

In this analysis, we mainly calculated the *intra-laboratory* variation, as compared to the inter-laboratory and mixture of the inter- and intra-laboratory variations estimated in works of Benigni et al [92] and Piegorsch&Zeiger [93], respectively. Unfortunately, it was impossible to do an inter-laboratory analysis in our study. First, there was a small overlap in molecules between different articles. Second, in the rare cases where it was possible, several authors (in particular Errol and Zeiger) contributed to majority of articles thus invalidating the goal of the inter-laboratory study. Therefore, for the comparison of the DMs, we selected the accuracy of 90% obtained in work of Piegorsch&Zeiger [93] as a conservative threshold for inter-laboratory comparison.

F. Reliability of predictions vs. variability of experimental measurements

Different subsets of molecules may differently behave in experiments: some of them may have easily reproducible results (either mutagenic or non-mutagenic) while the other molecules may show higher variability, e.g. because of various difficulties in experimental measurements such as metabolic stability, low solubility etc. It is interesting whether DMs can differentiate such chemicals.

We analyzed the variability of measurements for the molecules from the Piegorsch dataset [93]. The total set had 239 molecules, but 3 of them did not have structures defined and were excluded from our analysis. We developed a new ASNN model using all the Ames challenge molecules with an exception of these 236 molecules, which formed the test set. The reliability of predictions was determined using the ASNN-STD-PROB DM. Amid 50 compounds with the highest and the lowest prediction reliability, we selected the molecules that had at least three decisive measurements. There were 14 and 9 such molecules with an agreement of experimental measurements of 96% and 89% respectively. Moreover, there were also 13 and 21

compounds with questionable measurements within the same intervals.

We applied a similar analysis to the 1,680 Ames challenge compounds having at least three measurements. We found that 150 molecules with the highest and the lowest reliability of predictions had an agreement of experimental measurements of 97% and 91%, respectively. Thus the confidence of predictions determined by the DM correlated with the variability of experimental measurements: the molecules with higher confidence of predictions have better agreements of experimental measurements and vice versa.

Additionally to the above analysis, it is interesting to check out whether the molecules having an experimental uncertainty are differentiated by DMs. For such analysis, we selected all the Ames challenge compounds, which had at least one non-concordant measurement. Thus, e.g. if there were 10 “active” results and one “non-active” result, we considered the measurements for this compound as non-concordant. Obviously, such analysis can be done only for compounds with two or more measurements. A summary of such compounds is provided in Table 4.9.

Set	Total number of molecules	Thereof with multiple measurements	Thereof with non-concordant measurements
Training set	4,361	1,265	186
Test set	2,181	640	103

Table 4.9. A summary of non-concordant measurements.

Then, we calculated the DM values (CONS-STD-PROB) for such compounds using the percentage scale and plotted the distribution density (Figure 4.9). Apparently, for both the training and test sets, the non-concordant molecules tend to have larger than average values of the DM. This furthermore confirms the aforementioned assumption: the DM-based prediction reliability is affected by the uncertainty of experimental measurements.

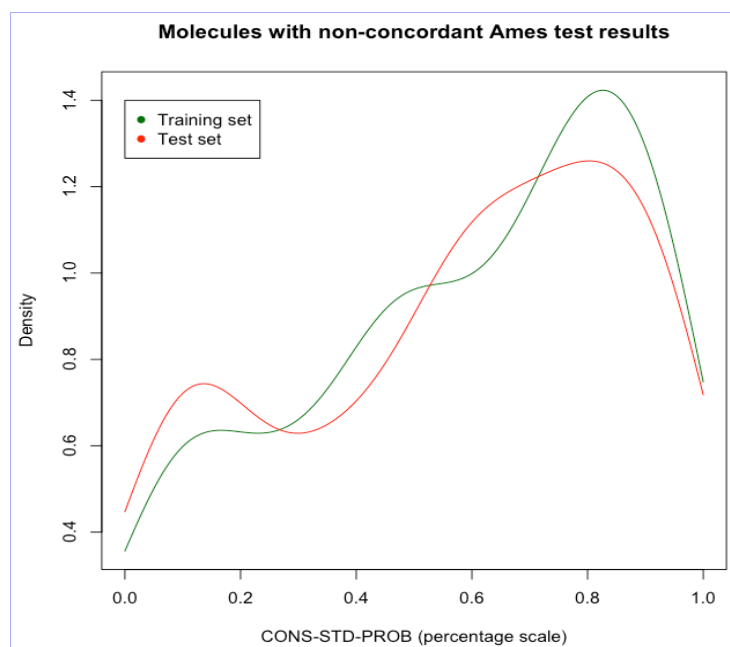


Figure 4.10. The distribution of CONS-STD-PROB (in percentage scale) for the molecules having at least one non-concordant (falling out) Ames test result. The green and red curves correspond to the training and test sets. Apparently, such molecules have bias towards larger values of CONS-STD-PROB. This fact further confirms the hypothesis: the prediction uncertainty determined by the DM is partially explained by the uncertainty of experiments.

G. Reliable predictions for ENAMINE, EINECS and HPV databases

In order to estimate the applicability of the investigated Ames test models to diverse chemical compounds, the OCHEM_ESTIMATE_ANN model was applied to the ENAMINE, EINECS and HPV databases (described in more detail in the section “Analyzed datasets” on page 28). Here, we will refer to the predictions having an estimated prediction accuracy of at least 90% as “reliable predictions”.

The accuracy of predictions was estimated using sliding-window averaging (SWA) based on ASNN-STD-PROB DM. For HPV and EINECS datasets, the percentages of reliable predictions were 30% and 16% respectively, which is close to the percentage in the original dataset, used for training (25%). However, the percentage of reliable predictions in ENAMINE dataset was only 4%, probably due to a higher chemical diversity of the ENAMINE compounds in comparison to the training set.

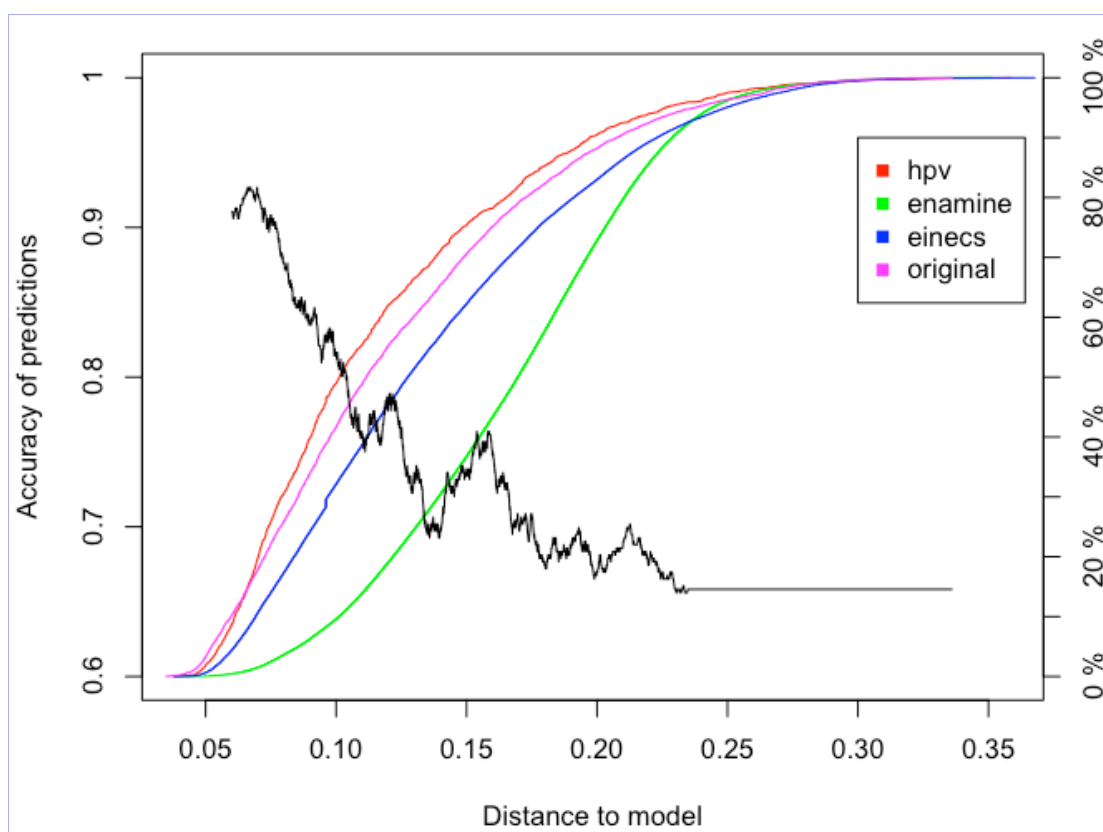


Figure 4.11. The estimated prediction accuracy for the original Ames challenge dataset, HPV, EINECS and ENAMINE datasets. The black curve, based on SWA, plots the prediction accuracy (left y-axis) against the ASNN-STD-PROB DM. The colored curves show the percentage of compounds from the 4 datasets (right y-axis), having DM not more than a threshold (x-axis).

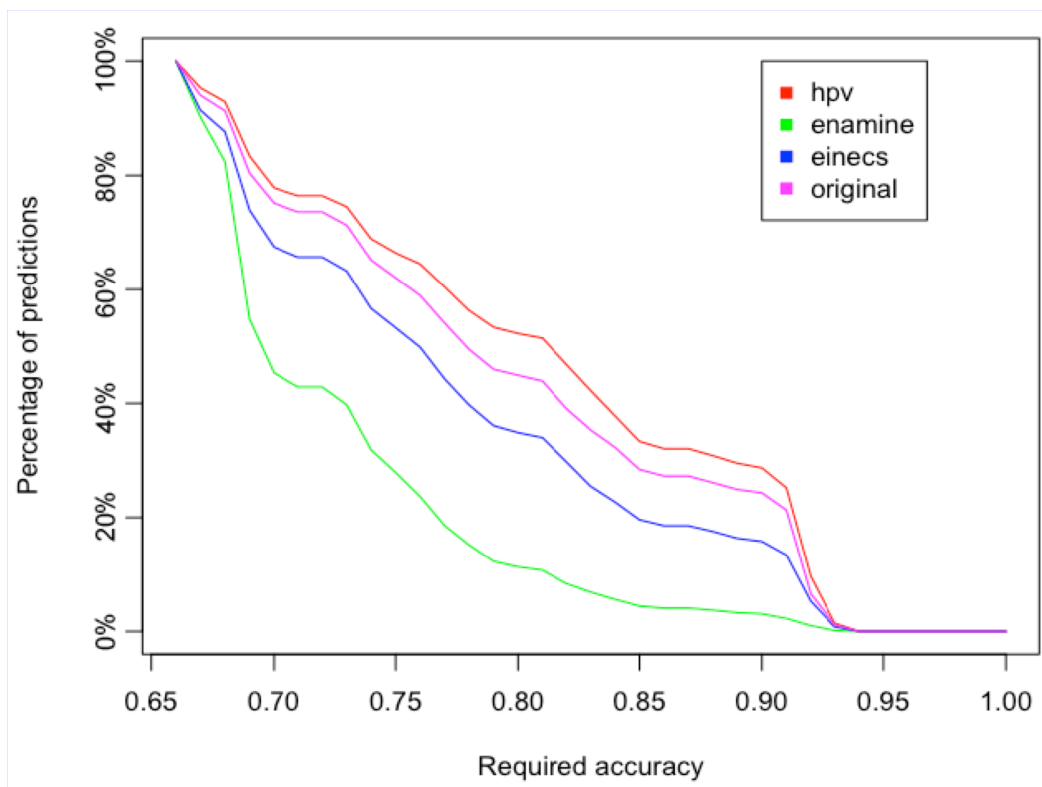


Figure 4.12. The percentage of compounds (y-axis) from the 4 datasets having the estimated prediction accuracy not less than a required accuracy (x-axis). This plot is based on the plot from Figure 4.11 with the DM-axis eliminated.

From the statistics in Table 4.10, it is evident that the number of predicted non-mutagens is significantly higher than of mutagens for all three datasets.

Predicted value	Enamine dataset		EINECS dataset		High production volume (HPV)	
	All	Reliable	All	Reliable	All	Reliable
Non-mutagens	171,758	12,696	55,143	14,271	1,945	805
Mutagens	57,141	153	13,635	1,350	410	85
Total	228,899	12,849	68,778	15,621	2,355	890

Table 4.10. Reliable Ames test predictions for ENAMINE, EINECS and HPV databases. * “reliable” predictions are those with the estimated prediction accuracy of at least 90%, which corresponds to the inter-laboratory variations.

Since the ENAMINE dataset contains the compounds with improved ADME profiles, this dataset is of particular interest in the context of drug discovery. The absence of mutagenic effects is one of essential requirements for a drug. In the ENAMINE dataset 12,696 from 228,899 compounds (5.5%) were reliably predicted as non-mutagens. Exactly these 12,696 compounds can be recommended for further screening. Moreover, this number can be increased by relieving the requirements for the prediction accuracy, e.g. to 85%, 80%, etc.

4.1.4 Summary

Based on 29 QSAR models used and the dataset in the Ames challenge 2009, we analyzed the problem of AD assessment for binary classification models.

The DM-based approaches allowed to distinguish the predictions of high and low prediction accuracy. The most reliable predictions of the Ames test achieved the experimental accuracy (ca 90%) while non-reliable predictions had the accuracy of the

random guess (50%). The predictions of the later compounds are useless; one should measure such compounds experimentally rather than rely on predictions.

The several DMs were benchmarked. The top-ranked DMs were based on the disagreement of the models: CONS-STD-PROB, CONS-STD-QUAL-PROB and CONCORDANCE, which were also strongly inter-correlated. Another simple measure, CLASS-LAG, was outperformed by the three aforementioned DMs but, nonetheless, was not significantly different for the consensus model. It is important to mention, that all three measures (CONS-STD-QUAL-PROB, CONS-STD-PROB, CONCORDANCE) implicitly use the predictions given by the consensus model. As the consensus model was the best of all 30 models, these DMs may have had an advantage because they incorporate information from the best (consensus) model. If we left out the consensus-based DMs, the best measures were the CLASS-LAG and the ASNN-STD-PROB. To sum up, the best performance was achieved by the STD-based DMs and the CLASS-LAG.

We discovered that, for all the models, the best separation of the reliable and non-reliable predictions was provided by the same DMs. In other words, the compounds having the best prediction accuracy were the same for all the models, regardless of the descriptors or/and the machine-learning technique used to develop them. This phenomenon attests to the “universality” of DMs: a DM that was developed for one model can be used for other models that are based on the same training set.

Another important result of this study is the discovery of a correlation between the prediction uncertainty and the variability of experimental measurements. Namely, we showed that the molecules with more reliable predictions had a higher agreement of experimental measurements and, vice versa, the molecules with less reliable predictions showed a higher disagreement of experimental measurements. Indeed, the molecules from the first group contributed “cleaner” training sets and, thus, allowed the models to achieve a higher prediction accuracy for their analogs.

Using the DM-based approaches, we estimated the prediction accuracy for three datasets with diverse chemical compounds: the EINECS, Enamine and HPV datasets. The accuracy of 90%, which is the estimated inter-laboratory variance, was achieved for 30% and 21% of HPV and EINECS databases of compounds using the ASNN model. However, for the larger and more diverse Enamine dataset, only 6% of compounds were predicted with such a high accuracy, presumably because of a higher chemical diversity of the Enamine collection. Thus, to increase the accuracy of predictions for such compounds, new experimental measurements are required.

The model developed using the OCHEM system (see Chapter 3) is publicly available at <http://ochem.eu/models/1>.

4.2 Toxicity against *T. Pyriformis*

4.2.1 Introduction

According to the requirements of the REACH program (**R**egistration, **E**valuation, **A**uthorisation and **R**estriction of **C**hemical substances), the compounds that are produced in Europe in amounts of more than 1 tone per year must be registered in order to estimate their environmental hazard. There are a number of the official human health and environmental endpoints that need to be reported, e.g., aquatic toxicity on fish [94]. The number of compounds that need to be registered by year 2018 is more than 140,000, the amount which is infeasible to test experimentally. Moreover, the REACH guidelines strongly encourage the usage of the alternative approaches for toxicity assessment [6]. One of such approaches is the *in-silico* virtual screening using QSAR models.

Thus, if QSAR models could provide reliable predictions with an accuracy comparable to that of experimental measurements, then such models would constitute a cheap, fast and reliable substitution for (or a complement to) experimental measurements.

Nowadays, the growth inhibition of the ciliated protozoan *Tetrahymena pyriformis* is an established screening tool for toxicity. This activity is often quantitatively expressed as the logarithm of the growth inhibitory concentration (pIGC50) and can be subjected to QSAR modeling. Although the toxicity on *Tetrahymena* is not an explicit REACH endpoint, it has been shown that there is a similarity in toxic potency of *T. pyriformis* and fish [95] and, therefore, pIGC50 on *T. pyriformis* can be useful in the REACH context as a surrogate endpoint.

This study, based on the dataset with more than a thousand of pIGC50 measurements (introduced on page 29 of this work), investigates the applicability domain of QSAR models for pIGC50 predictions.

4.2.2 Methods

For a better understanding of the methods used in this study, a reader may refer to a number of concepts introduced in the methodological section “Applicability domain of QSAR models“. The relevant concepts are *distance to model* (DM, page 15), *bin-based averaging* (BBA, page 22), *multi-gaussian distribution* (MGD, page 26), *approval test* for a DM (page 26).

A. QSAR approaches

In total, eleven QSAR models built by 5 international scientific groups were investigated. The models differed in the types of descriptors and modeling techniques. All the models were developed using the training set, which contained 644 compounds and were tested on 2 validation sets containing 339 and 110 compounds respectively. In more detail, the datasets are described in the section “Analyzed datasets” on page 29.

The models were based on 5 descriptor types: E-State indices, ISIDA fragments and descriptors calculated by Dragon (see section “Molecular descriptors” on page 5). Additionally, specifically for this study, two more software packages were involved for the calculation of descriptors: MolconnZ (<http://www.edusoft-lc.com/molconn/>) and

CODESSA-Pro (**C**Omprehensive **D**escriptors for Structural and Statistical Analysis, <http://www.codessa-pro.com/>).

Table 4.11 summarizes QSAR approaches used and Table 4.12 summarizes the statistical parameters for all models. Initially, all the eleven QSAR models were developed using only the training set and their accuracy was estimated using the Leave-One-Out (LOO) cross-validation. Following this analysis, we performed a “blind” prediction of them molecules from both the validation sets and calculated RMSE on these sets. The obtained RMSE values were used to compare the predictive ability of the model.

Consensus Model. Additionally to the eleven individual models, a consensus model was calculated as a simple non-weighted average the predictions given by the individual models listed in Table 4.11. The statistical parameters of both individual and consensus models are summarized in Table 4.12. The consensus model had a similar prediction ability to that of the Associative Neural Network (ASNN) model for all the three sets.

nn	group	modeling techniques	descriptors	abbreviation	distance to models	
					descriptor space	property-based space
1	UNC	ensemble of 192 kNN models	MolconnZ	kNN-MZ	EUCLID	STD
2	UNC	ensemble of 542 kNN models	Dragon	kNN-DR	EUCLID	STD
3	VCCLAB	ensemble of 100 neural networks	E-state indices	ASNN-ESTATE		CORREL, STD
4	ULP	kNN	ISIDA Fragments	kNN-FR	EUCLID, TANIMOTO	
5	ULP	MLR	ISIDA Fragments	MLR-FR	EUCLID, TANIMOTO	
6	UI	OLS	Dragon	OLS-DR	LEVERAGE	
7	UK	PLS	Dragon	PLS-DR	LEVERAGE	PLSEU
8	UNC	SVM	MolconnZ	SVM-MZ		
9	UNC	SVM	Dragon	SVM-DR		
10	ULP	SVM	ISIDA Fragments	SVM-FR		
11	ULP	MLR	Molecular properties (CODESSA-Pro)	MLR-COD		
12	Average of all the models		-	CONS		STD

Table 4.11. A summary of the analyzed QSAR approaches for pIGC50 prediction.

model abbreviation	training set				validation sets			
	internal LOO		5-CV		set 1		set 2	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
ASNN-ESTATE	0.84	0.42	0.82	0.44	0.85	0.41	0.66	0.52
kNN-DR	0.92	0.3	0.8	0.5	0.84	0.41	0.59	0.57
kNN-FR	0.77	0.51	0.73	0.55	0.71	0.56	0.37	0.71
kNN-MZ	0.91	0.32	0.76	0.53	0.83	0.43	0.49	0.64
MLR-COD	0.72	0.55	0.69	0.59	0.71	0.57	0.58	0.58
MLR-FR	0.94	0.26	0.74	0.55	0.49	0.56	0.43	0.67
OLS-DR	0.75	0.53	0.77	0.51	0.77	0.5	0.58	0.58
PLS-DR	0.88	0.36	0.79	0.48	0.81	0.46	0.59	0.57
SVM-DR	0.93	0.28	0.81	0.46	0.7	0.57	0.53	0.61
SVM-FR	0.95	0.24	0.8	0.48	0.76	0.51	0.38	0.7
SVM-MZ	0.89	0.35	0.77	0.51	0.77	0.5	0.58	0.58
CONS	0.92	0.31	0.83	0.44	0.85	0.4	0.67	0.51

Table 4.12. The statistical parameters of the investigated pIGC50 models.

B. Applicability domain assessment

The study analyzed the applicability domain of the 12 individual models based on the concept of distance to model (DM). The DM concept and examples can be found in the section “Distances to models” on page 15 of this work. Each participating group provided their own definitions of DMs; these DMs are analyzed in this study and are briefly overviewed below.

University of North Carolina at Chapel Hill in the United States (UNC). This group used the ensemble of variable selection k Nearest Neighbors (kNN) and Support Vector Machine (SVM) methods, which were applied to the descriptors calculated using the Dragon and MolconnZ software packages.

The AD for models derived using kNN approach was calculated from the distribution of similarities between each compound and its k nearest neighbors in the training sets. The similarities were defined as distances between a molecule i and a training set. They were computed as the average Euclidean distance to the k nearest neighbors of this molecule in the training set. The distances were calculated not with all variables, but only with a subset of variables identified by the modeling procedure as optimal. More precisely:

$$EU_M = \frac{\sum_{j=1}^k d_j}{k} \quad (4.1)$$

where d_j is the distance of a query compound to its k^{th} nearest neighbor and m is index of the model. The distribution of distances (pairwise similarities) between each compound and its k nearest neighbors in the training set was computed to produce an applicability domain threshold, D_T , calculated for each kNN model as follows:

$$D_T = \bar{y} + Z \sigma \quad (4.2)$$

Here, \bar{y} is the average Euclidean distance of the k nearest neighbors of each compound within the training set, σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. Typically, the default value of this parameter was set at 0.5, which formally places the boundary for which compounds will be predicted at one-half of the standard deviation (assuming a Boltzmann distribution of distances between each compound and its k nearest neighbors

in the training set). Thus, if the distance of an external compound from all of its nearest neighbors in the training set exceeded this threshold, the prediction was considered unreliable.

In total $M=192$ and $M=542$ individual models were calculated using MolconnZ and Dragon descriptors, respectively. The average values of the distances to each individual model $m=1, \dots, M$

$$EUCLID = EU_m = \frac{\sum_{j=1}^M EU_j}{M} \quad (4.3)$$

was used to estimate a distance of a molecule to the final ensemble of models. Notice, that the minimal value of EUCLID is observed when the training set model was built with $k=1$. The same definition of DM was also used for models built with the SVM method.

University of Louis Pasteur in France (ULP). This group used the kNN, SVM and Multiple Linear Regression (MLR) methods with the fragmental descriptors calculated using the ISIDA software [9]. Applicability domains for the ISIDA-MLR and ISIDA-kNN models were estimated with an approach similar to that of the UNC group (EUCLID DM) with an exception that only one ISIDA-MLR and one ISIDA-kNN model were calculated. Thus, no ensemble was built and there was no averaging over the models. For both the approaches, the distances were calculated using $k=3$, which was the optimal number of nearest neighbors for the kNN model.

Additionally, to define the applicability domain, the ULP group considered the minimal and maximal occurrences of fragments (which were selected by the regression) within compounds in the training set for the ISIDA-MLR model. These values defined an acceptable range for each fragment, resulting in a so called “descriptor bounding box”. The compounds outside the bounding box (i.e. having an unacceptable number of occurrences of a sub-fragment) were considered to be outside of the model's AD. For the validation set compounds, the distance to the training set was considered as infinite if at least one of its fragment descriptors was outside the corresponding range defined on the training set.

University of Insubria in Italy (UI). This group used an Ordinary Least Squares regression (OLS) and the genetic algorithm with the descriptors calculated using the Dragon software. To assess the AD for their model, the UI group used the leverage DM with a “warning” threshold calculated accordingly to Expression 2.14 (page 16). The compounds with DM that exceeded the “warning” leverage threshold were considered unreliable.

University of Kalmar in Sweden (UK). This group used Partial Least Squares (PLS) method and Dragon descriptors. To assess the applicability domain, the UK group used two DMs. The first DM was leverage, which was also employed by the UI group. However, since different descriptors were selected in OLS and PLS models, the nominal leverage values in both models were different. The second DM was the distance to the PLS model, PLSEU, which was calculated using the UNSCRAMBLER program as described in its manual or in the book [96]. In brief, PLSEU corresponds to the distance of the descriptor vector in to its projection on the hyperplane of latent variables (the PLS hyperplane). The idea behind PLSEU is that if the distance to the projected vector is relatively high, a part of the information is lost during the projection and, therefore, the prediction accuracy for this chemical compound may be

compromised.

Additionally, we applied the ASNN machine learning method with E-state indices. On the basis of this model, we calculated the CORREL DM (see page 18) and used the cut-off value of 0.7.

In addition to the 4 types of DMs that were provided by individual participants, we used two general DMs: the standard deviation (STD) based on a neural network ensemble (STD-ASNN) and a consensus ensemble (STD-CONS), and the Tanimoto similarity (TANIMOTO). A detailed description for these measures can be found in the section “Distances to models”.

Thus, our study included 14 DMs of 6 different types (EUCLID, LEVERAGE, PLSEU, CORREL, STD and TANIMOTO). The DM was named by combining its type (STD, EUCLID, etc) and abbreviation of the method (see Table 4.11) in which the DM was calculated.

C. Benchmarking criteria

Similarly to the Ames test study, we compared the DMs according to the cumulative accuracy coverage criterion and the AUC criterion (page 24 in “Methodology” chapter). As the accuracy threshold, we used the accuracy of experimental measurements for compounds with the narcosis mode of action (RMSE of 0.38).

Additionally to these two criteria, we performed the analysis of the residuals distribution. Namely, we checked how accurately the distribution of residuals is approximated by the estimated distribution of residuals, suggested by bin-based accuracy averaging (BBA, see section “Accuracy averaging” on page 22). To approximate distribution of residuals, we used the mixture of Gaussian distributions (MGD) with zero mean, but different standard deviations, which corresponded to the estimated RMSE values, evaluated from the BBA procedure. The goodness of fit was estimated quantitatively according to the likelihood score and visually according to confidence consistency plots (page 26).

The likelihood score was also used for the DM approval tests (see page 26 for definition). Namely, the likelihood score of the MGD was compared with the score of the single Gaussian distribution (SGD). If the score of MGD was significantly higher than the score of SGD, then the DM was considered as approved. To check whether the difference between MGD and SGD is not caused by a mere chance but is statistically significant, we calculated p-values on basis of the bootstrap test with 10,000 replicas. The p-values are reported in Tables A3 and A4 in Appendix on pages 129-131.

4.2.3 Results

A. Analysis of individual models

ASNN model. An example in Figure 4.13 demonstrates the variability of the prediction accuracy for the ASNN model.

First, a plot in Figure 4.13-D shows that the SGD was not a good approximation for the distribution of residuals. The MGD generated using bin-based averaging (BBA) over STD-CONS provided a significantly better approximation. The STD-CONS and STD-ASNN had the best likelihood scores for both the training and validation sets.

Second, Figures 4.13A,B demonstrate that the prediction accuracy was variable and correlated with the investigated DMs. The MGD calculated using STD-CONS DM allowed the best separation of molecules with small and large errors. For example, molecules from the training set with $STD-CONS < 0.19$ and $STD-CONS > 0.73$ had average errors of 0.19 and 0.78 log units, respectively. Thus, the most and least reliably predicted molecules had errors, which differed by the factor of four.

The other DMs performed worse as compared to STD-ASNN and STD-CONS. The EUCLID-kNN-MZ distance had a smaller likelihood score and provided a worse discrimination of molecules with small and large errors. The most reliable predictions according to this measure had the average error of 0.31 log units while the least reliable predictions had the average error of 0.57 log units for EUCLID-kNN-MZ values of < 0.23 and > 0.75 , respectively. Figures 4.13A,B demonstrate that the ASNN model errors correlated better with the STD-CONS distance and not with the EUCLID-kNN-MZ for the training set (red line). This difference, however, is not so obvious for the validation set (black line on Figures 4.13-A,B), for which both the DMs had similar performances. The confidence consistency plot for the STD-CONS was closer to the “optimal” plot compared to the EUCLID-kNN-MZ (Figure 4.13-D) thus indicating a higher discrimination ability of the former DM.

The LEVERAGE OLS, as well as several other DM (Table A3 on page 129), did not calculate MGD with a significant score and thus did not discriminate molecules with small and large errors for the training set. This result was also apparent from the absence of correlations between this DM and errors (Figure 4.13C).

OLS model. The OLS model included six Dragon descriptors

$$\log(IGC_{50}^{-1}) = -18(\pm 0.7) + 0.065(\pm 0.002) \cdot AMR - 0.50(\pm 0.04) \cdot O_{056} - 0.30(\pm 0.03) \cdot O_{058} - 0.29(\pm 0.02) \cdot nHAcc + 0.046(\pm 0.005) \cdot H_{046} + 16(\pm 0.7) \cdot Me \quad (4.4)$$

Figure 4.14 shows that, although the MGD based on the LEVERAGE DM discriminated molecules with low and large errors, the STD-CONS provided significantly better results. Indeed, the former DM identified predictions with $\tilde{\sigma} = 0.5$ and $\tilde{\sigma} = 0.66$ for molecules with lowest and largest errors from the joint validation set. For the same set, the STD-CONS DM had the minimum $\tilde{\sigma} = 0.36$ and the maximum $\tilde{\sigma} = 1.2$, respectively. Thus, the discriminative ability of the STD-CONS was significantly better as compared to the LEVERAGE.

Depending on the purpose of the analysis, the latter metric could be used to identify molecules that are predicted either accurately (e.g., registration within REACH) or inaccurately (e.g., selection of new molecules to extend the model AD).

The small discrimination power of the LEVERAGE DM did not allow performing such selection efficiently.

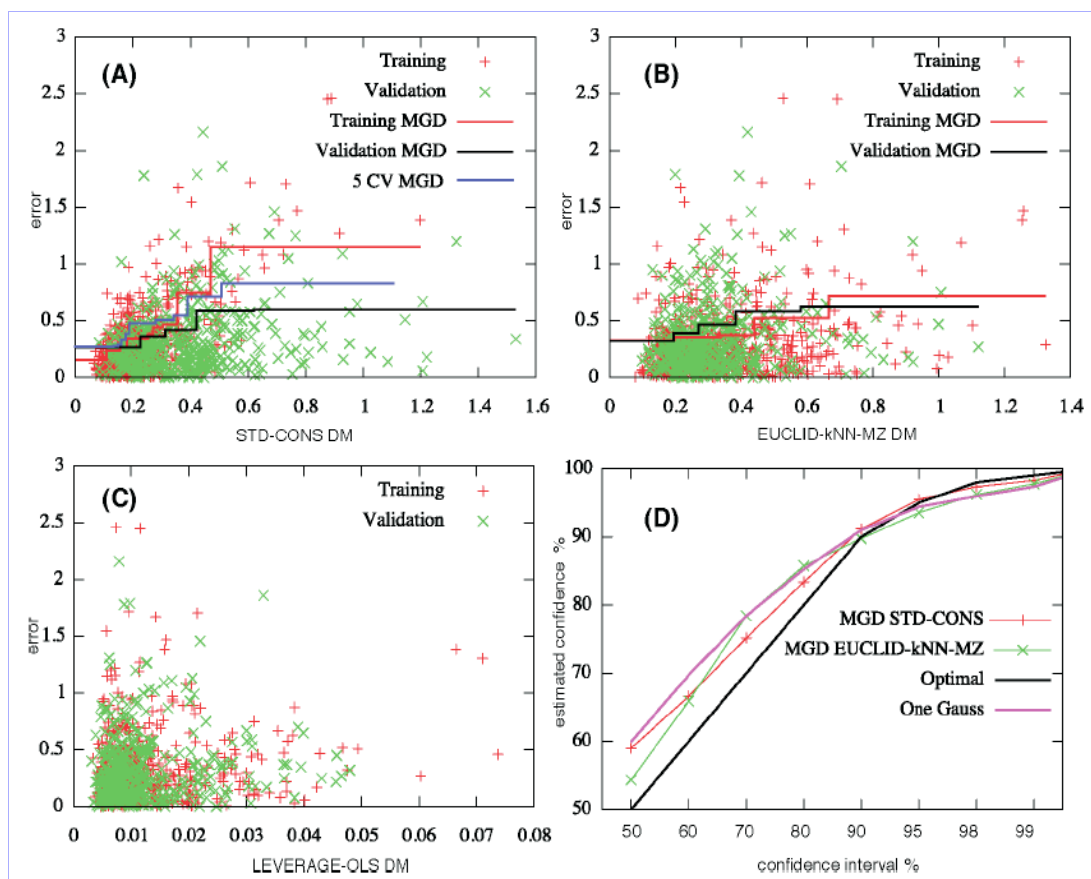


Figure 4.13. Analysis of the ASSN-ESTATE model. The MGD for the training and joint validation sets are shown for STD-CONS (A) and EUCLID-kNN-MZ (B) DMs. The MGDs are based on the bin-based averaging (BBA). As it is seen in (C), the distribution of errors for the LEVERAGE-OLS-DR did not calculate a significant MGD. The confidence consistency plot for two exemplary DMs is shown in (D).

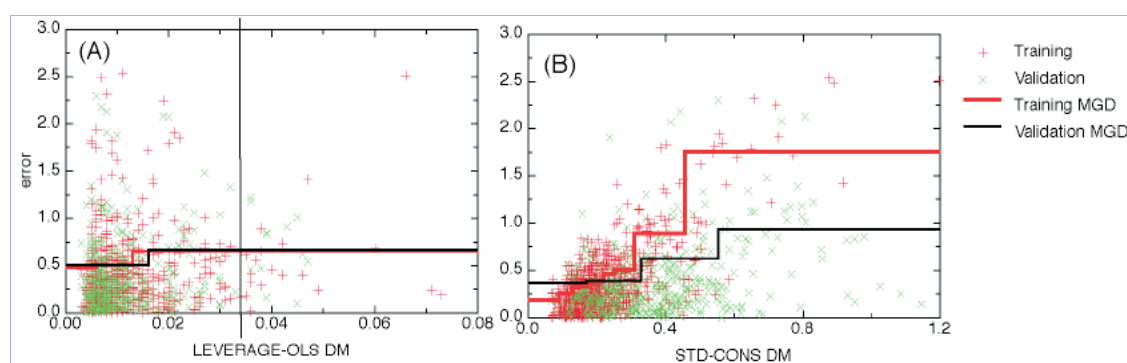


Figure 4.14. Analysis of the OLS-DR model given by eq 4.4. The STD-CONS DM (right plot) provides a better discrimination of molecules with low and large errors compared to that of LEVERAGE-OLS DM (left plot). The vertical line on the left plot corresponds to the leverage threshold $3(K+1)/N = 3 \cdot 7/664 = 0.033$ (the “warning” leverage).

Mechanism based model. Schultz *et al* [42] analyzed a simple model

$$\log(IGC_{50}^{-1})=0.545\cdot\log P+16.2\cdot A_{max}-5.91 \quad (4.5)$$

$$N=392, R^2=0.83, RMSE=0.31$$

which was developed using $N=384$ molecules (8 outlying molecules were excluded). This model was based only on two descriptors, the octanol-water partition coefficient ($\log P$) and Maximum Acceptor Superdelocalizability (A_{max}). This equation predicted molecules from the test set (the second validation set in our study) with $RMSE=0.54$ log units.

The authors of this model pointed out that the distance to the descriptor centroid did not allow them to differentiate molecules with low and large errors [42]. However, the BBA calculated using, e.g. STD-ASNN (Figure 4.15), successfully accomplished this goal for molecules from both the training and validation datasets.

Interestingly, five out of the eight outlying molecules (*Benzoyl isothiocyanate*, *Pentafluoronitrobenzene*, *Pentafluorobenzyl alcohol*, *a,a,a-4-Tetrafluoro-o-toluidine*, *4-Chloro-3,5-dinitrobenzotrile*, *1,5-Difluoro-2,4-dinitrobenzene*), which were excluded from the original equation, in fact had large STD-ASNN deviations (>0.27) and contributed to the Gaussian distribution with the largest $\tilde{\sigma}=0.49$. Thus, the low prediction ability of the mechanism-based model (eq. 4.5) for these five outlying molecules could be due to their structural diversity as compared to the other molecules in the training set. This structural diversity was successfully captured by the STD-ASNN DM, which identified that the predictions for the aforementioned compounds are unreliable.

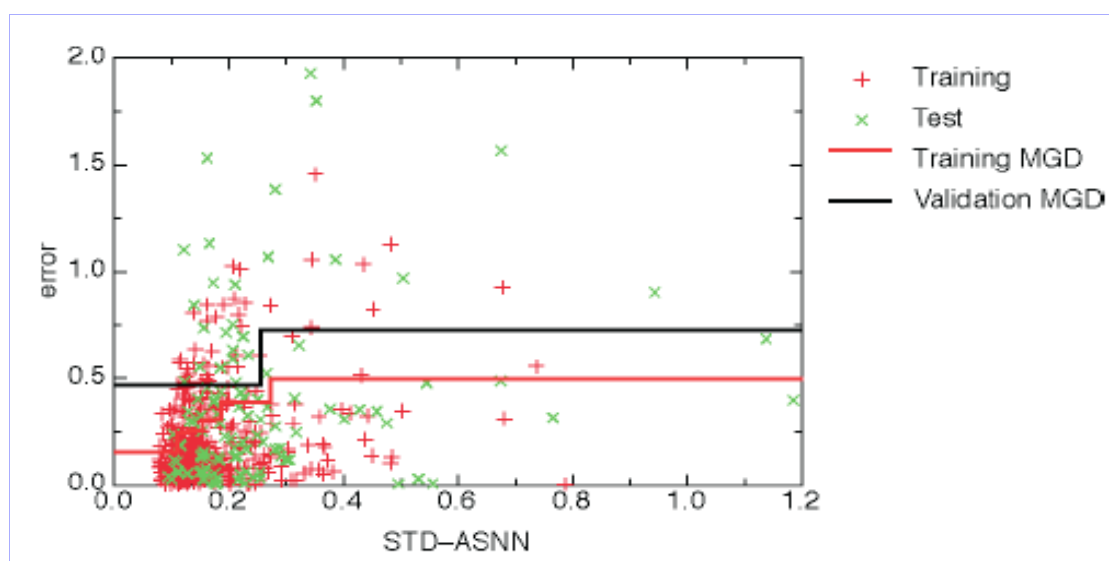


Figure 4.15. The BBA for the mechanism-based model (eq 4.5). The use of STD-ASNN DM allowed for the discrimination of molecules with low and large errors in both training and validation sets.

B. Comparison of distances to models

For all three comparison criteria, we ranked all the 14 DMs separately for each of the 13 models (the best DM received rank 1, the worst rank 14) and averaged the ranks over all the models. The averaged ranks were calculated separately for the training (5-fold cross-validation) set and the joint validation set. In case of the likelihood-score

criterion, the DMs that did not pass the approval tests were not used in the scoring. The results are summarized below in Tables 4.13, 4.14 and 4.15 for the accuracy coverage criterion, the AUC criterion and the likelihood-score criterion respectively.

Apparently, STD-ASNN and STD-CONS are in top according to all three criteria. We confirmed that superiority of these two DMs was statistically significant (*p-value* less than 0.05 according to the bootstrap test). Another STD-based measure, STD-KNN-MZ is ranked as third according to the accuracy coverage and AUC criteria and as fourth according to the likelihood-score criterion.

Although there were minor discrepancies in the ranks provided by the three used comparison criteria, the general picture is the same: the STD-based DMs outperform other DMs. This result is concordant with the results of the Ames test benchmarking: the STD-based DMs (based on ensembles of neural networks and the consensus model) are universal and work well for all the models. Thus, the compounds having the highest prediction accuracy were the same for all the investigated QSAR approaches.

Distance to model	Ranking (RMSE 0.35 coverage)	
	5FCV	Validation
STD-CONS	1,92	1,67
STD-ASNN	1,17	3,58
STD-KNN-MZ	11,25	4,54
EUCLID-KNN-MZ	8	5,42
EUCLID-KNN-FR	7,17	6,33
STD-KNN-DR	6,58	6,54
EUCLID-KNN-DR	7,58	7,04
AD-Si-PLS	9,04	7,5
TANIMOTO-MLR-FR	7,54	7,92
TANIMOTO-KNN-FR	5,96	9
LEVERAGE-OLS-DR	10,42	9,46
CORREL-ASNN	7,92	10,42
EUCLID-MLR-FR	7,83	12,25

Table 4.13. The averaged rankings of the DMs according to the accuracy coverage criterion (sorted by rankings based on the validation set)

Distance to model	Ranking (AUC)	
	5FCV	Validation
STD-CONS	2,67	2,25
STD-ASNN	1,17	3,75
STD-KNN-MZ	7,17	3,83
EUCLID-KNN-FR	3,33	5,08
EUCLID-KNN-DR	6,42	5,92
AD-Si-PLS	7	6,25
EUCLID-KNN-MZ	10	7,17
STD-KNN-DR	5,75	7,75
TANIMOTO-MLR-FR	9,67	7,75
TANIMOTO-KNN-FR	5,92	9
LEVERAGE-OLS-DR	12,83	11,17
CORREL-ASNN	10,58	11,25
AD-Hi-PLS	12,42	11,5
EUCLID-MLR-FR	10,08	12,33

Table 4.14. The average rankings of the DMs according to the AUC criterion (sorted by rankings based on the validation set)

Distance to model	average rank		failures of approval tests	
	5-CV	Valid.	5-CV	Valid.
STD-CONS	1,8	1,1		
STD-ASNN	1,2	2,5		
STD-kNN-DR	4,3	4,1		
STD-kNN-MZ	8,3	5,3		
EUCLID-kNN-DR	4,9	5,4		
LEVERAGE-PLS	5	6,3		
EUCLID-kNN-MZ	7,1	6,4		
TANIMOTO-kNN-FR	6,1	6,8		
TANIMOTO-MLR-FR	8,3	9		1
CORREL-ASNN	10,8	9,4		1
LEVERAGE-OLS-DR	12,6	11,1		2
EUCLID-MLR-FR	9,3	11,5		7
PLSEU-PLS	11,8	11,5		7

Table 4.15. The averaged rankings of the DMs according to the likelihood-score criterion and the number of models, for which the DMs failed to pass the approval tests

C. Ability to estimate the prediction accuracy

As described in Methods chapter (section “Estimation of prediction accuracy” on page 23), we used the bin-based averaging procedure performed on the training set to estimate the prediction accuracy for new compounds.

The BBA procedure calibrated on 5-fold cross-validation (5CV) residuals was used to predict the RMSEs for the molecules from the validation sets. An example of a BBA calculated using 5CV procedure is shown on Figure 4.13A as a blue line. This BBA-plot mapped the STD-CONS distances to the estimated RMSE values (denoted as $\tilde{\sigma}$, we use tilde to denote estimated values). For example, the STD-CONS distances in the range of [0, 0.15] corresponded to $\tilde{\sigma}=0.25$, while distances larger than 1.1 corresponded to $\tilde{\sigma}=0.80$. These ranges and values $\tilde{\sigma}$ were used to predict errors for molecules from the validation sets. To do this we, firstly, calculated STD-CONS for each new molecule and, secondly, estimated the prediction error using the BBA-plot based on the 5CV procedure. Thus, for a molecule with STD-CONS=0.1, which belongs to the [0, 0.15] interval, we predicted its average square of the error as $(\tilde{\sigma}(0.1))^2=0.25 \cdot 0.25=0.0625$. We made such predictions for all molecules from the validation set, $i=1, \dots, M$, and estimated the RMSE error for the validation set accordingly to Expression 2.21 (page 23), as a root mean of squares of estimated RMSEs for all the validation set compounds.

Table 4.16 reports a summary of the performances of the analyzed DMs (for the complete details, refer to Table A4 in Appendix on page 131, which reports the performance of analyzed DMs for all models).

First, all the DMs correctly recognized a higher complexity of the second validation set and predicted higher errors for this set as compared to the first validation set. Thus, all the DMs were useful to discriminate datasets of different complexity on the qualitative basis.

Second, the STD-ASNN DM was top-ranked, which is consistent with the results calculated on the training set (Tables 4.13-4.15). This means that STD-ASNN not only was able to discriminate the predictions of high and low accuracy on the training set,

but also provided a relatively accurate estimation for the prediction accuracy on the validation sets.

Distance to model	rank	calibrated on 5-CV set		on validation set 1			
		validation set 1		validation set 2			
		<i>RMSE</i> ¹	Δ err ²	<i>RMSE</i>	Δ err	<i>RMSE</i>	Δ err
STD-ASNN	5	0,53	0,06	0,62	0,05	0,58	0,07
LEVERAGE-PLS	5,7	0,5	0,04	0,54	0,07	0,52	0,09
EUCLID-kNN-MZ	7,9	0,45	0,05	0,51	0,1	0,57	0,06
EUCLID-kNN-DR	8,4	0,45	0,05	0,52	0,09	0,57	0,07
LEVERAGE-OLS-DR	10	0,5	0,04	0,52	0,09	0,51	0,09
TANIMOTO-kNN-FR	10,4	0,5	0,04	0,54	0,07	0,52	0,08
STD-kNN-DR	11,2	0,46	0,05	0,52	0,09	0,56	0,07
TANIMOTO-MLR-FR	11,2	0,51	0,04	0,53	0,07	0,52	0,09
CORREL-ASNN	11,4	0,49	0,04	0,54	0,07	0,53	0,08
STD-CONS	12	0,65	0,16	0,72	0,12	0,54	0,07
EUCLID-MLR-FR	12	0,49	0,04	0,52	0,09	0,52	0,09
PLSEU-PLS	12	0,49	0,04	0,5	0,1	0,5	0,11
STD-kNN-MZ	12	0,48	0,04	0,56	0,05	0,59	0,06
EUCLID-kNN-FR	12	0,5	0,04	0,52	0,09	0,5	0,1
average error ³		0,49		0,6		0,6	

Table 4.16. Estimated errors on the validation set. ¹average predicted RMSE (e.g., using STD-ASNN DM we predicted RMSE for all 12 analyzed models and averaged them). ² average absolute differences between predicted and actual RMSE for all methods (e.g., using STD-ASNN DM we predicted RMSE for all 12 models and calculated average absolute difference between predicted and RMSE errors for all models).

Remarkably, although the STD-CONS DM had an excellent discriminative power on the training set, it received a low rank when applied to predict errors on the validation set. This fact can be explained by the incorrect validation for a part of the individual models, which performed the variable selection procedure *before* the cross-validation and, therefore, may have provided over-fitted models. This assumption is furthermore confirmed with the fact that STD-ASNN, which was based on a properly validated model, was still on the top of the list (Table 4.16). Thus, the performance of the STD-ASNN DM was consistent on all the investigated datasets.

The case described above points out that a proper validation is important not only for avoiding over-fitted models but also for obtaining consistent DMs.

It was also possible to calibrate the BBA procedure on the first validation set to estimate the prediction accuracy on the second validation set. We performed such analysis and fitted the BBA using results calculated for the first validation set (Table A4 in Appendix on page 131). The errors predicted with these BBAs were similar to those fitted on 5-fold cross-validation results.

D. Interpretation of the AD

Substructural analysis. The substructural analysis was performed as described in the methodological section “Interpretation of applicability domains” on page 27, similarly to the analysis in the Ames test study (page 62). For this purpose, we re-developed the ASNN model using the complete set with 1,093 compounds. The model had the overall cross-validated RMSE of 0.47. As a result of the substructural analysis, we identified the molecular fragments for which the containing molecules had prediction accuracy significantly different than the average accuracy. Such fragments are summarized in Table 4.17.

Fragment name	Number of compounds	Accuracy	Significance, p-value
Prediction accuracy significantly higher than average			
C-O	447	0,35	<0.001
CCCC	355	0,38	<0.001
CCCCC	240	0,34	<0.001
CCCCCC	157	0,31	<0.001
Prediction accuracy significantly lower than average			
Sulfur	62	0,85	<0.001
S=O	17	1,21	<0.001
O=S=O	11	1,38	<0.001
S-C	26	0,98	0,001
Br	103	0,64	0,008
Br-C	59	0,76	0,002

Table 4.17. The compounds predicted by the ASNN model with the accuracy significantly higher (or lower) than the average accuracy of the model (0.47)

Apparently, the molecules that contained sulfur and bromine atoms were predicted relatively inaccurate and had RMSEs of 0.85 and 0.64 (compared to the average RMSE of 0.47). Furthermore, the sulfones (11 compounds) had even lower prediction accuracy with RMSE of 1.38. When we observed the sulfones in detail, we found out that, notwithstanding with a visual similarity of these compounds, their pIGC50 values have wide range, from -2.20 (methyl sulfone) to 1.41 (divinyl sulfone). Presumably, there might be a specific mechanism of toxicity, which was present in a part of sulfones and which was not captured by the model. This assumption was confirmed by investigations in a study by Seward et al (2001) [97], where the authors showed that large toxicity values are associated with electrophilic toxicants, whereas most of the investigated compounds have the neutral (narcotic) toxicity mechanism. A study by Schultz et al [43] confirms that, presumably, vinyl containing sulfones have the electrophilic toxicity mechanism. Importantly, the low prediction accuracy for sulfones was correctly captured by the STD-ASNN DM: all the 11 sulfones were among 25% of the highest DM values in the set and, therefore, they were correctly identified as unreliably predicted compounds.

In contrary, the molecules with long saturated carbon chains were predicted with the accuracy significantly higher than the average model accuracy. The molecules containing four subsequent carbon atoms (355 compounds) were predicted with RMSE of 0.38 and the molecules with six subsequent carbon atoms (157 compounds) – with RMSE of 0.31. In both the cases, the RMSEs are significantly different (in the statistical sense) than the average model RMSE 0.47.

Analysis of pIGC50 values. To check whether the toxicity (pIGC50) of a compound affects its prediction accuracy, we plotted the average RMSEs (over all 13 models) as a function of pIGC50 (the red curve Figure 4.16). Apparently, the more toxic compounds (compounds with pIGC50 > 1 (about 20% of compounds) had RMSE as high as 0.6-0.8, whereas the more toxic compounds with pIGC50 in the [-0.7, 1] had RMSE as low as 0.35-0.45. Thus, the investigated QSARs tend to have relatively poor prediction accuracy for highly toxic compounds. Presumably, this phenomenon can be explained by some unknown and rather strong mechanism of toxicity, which is different from the toxicity mechanism of the majority of the training set compounds. For example, as it was shown above with the sulfone-vinyl containing compounds, the highly toxicant compounds may have the electrophilic mechanism of toxicity, which was not captured by the model.

From a practical point of view, the usage of pIGC50 values for estimation of the prediction accuracy is infeasible, since pIGC50 values are not available for new predicted compounds. When we substituted the observed pIGC50 values with the predicted ones, the accuracy did not have any clear dependency from pIGC50 (the green curve on Figure 4.16).

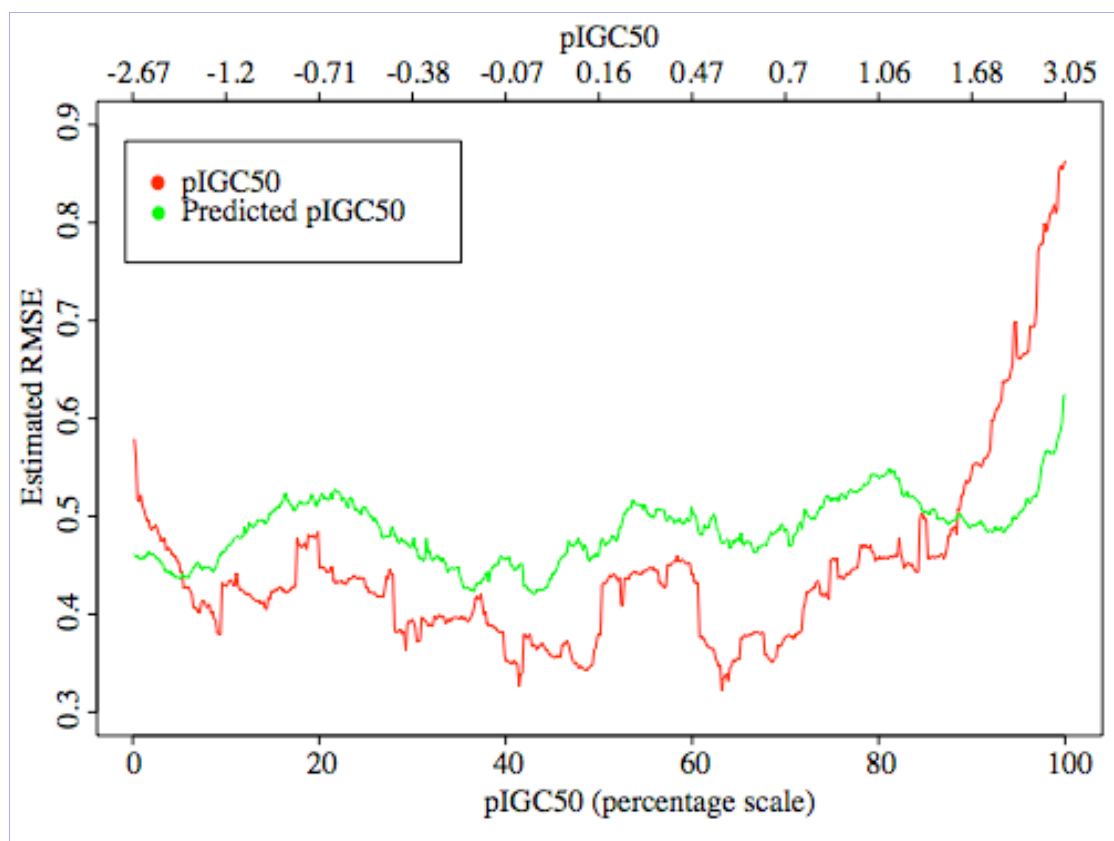


Figure 4.16. The average prediction accuracy (RMSE) depending on pIGC50 values. Apparently, predictions are more accurate for compounds with average or less than average pIGC50 values ([-0.7, 1.0]); such compounds have RMSE of as low as 0.35. Predictions with high (more than 1.0) pIGC50 values have low accuracy (RMSE up to 0.9). However, when the real pIGC50 values are substituted with the predicted ones (the green curve), they cannot discriminate predictions of high and low accuracy.

E. Reliable predictions for HPV, EINECS and ENAMINE databases

The ASNN-ESTATE model and STD-ASNN DM provided ones of the most accurate predictions and estimation of the errors. Therefore, we decided to evaluate the performance of this method for prediction of molecules from the three industrial databases, similarly to as it was done for the Ames test study. For this analysis we re-developed the ASNN model with the complete dataset (1,165 compounds).

To estimate the accuracy of predictions for new compounds, we used the bin-based accuracy averaging (the dashed curve on Figure 4.17). We estimated the percentage of predictions with RMSE not less than 0.38 and 0.21, which corresponds to the accuracy of experimental measurements for compounds with the reactive and narcosis modes of action. The results of the accuracy estimation are summarized in Table 4.18 and demonstrated visually on Figures 4.17 and 4.18.

There was an apparent dramatic difference in the performance of the model on the training set compared to the external sets. Namely, the percentage of predictions with RMSE of 0.38 was 92% for the training set and only 36%, 20% and 1% for HPV, EINECS and ENAMINE datasets. If we count only predictions with RMSE less than 0.21, the percentage was 10% for the training set in comparison to 2% for HPV dataset and close to 0% for EINECS and ENAMINE datasets. Thus, the number of accurate predictions is much less in external datasets in comparison to the training set. Therefore, the investigated model has a very limited domain of applicability.

Investigated set	Number of compounds, total	Thereof with error less than 0.38		Thereof with error less than 0.21	
		Num	%	Num	%
Original (T. Pyriformis)	1,165	1,071	92%	118	10%
HPV	2,355	856	36%	41	2%
EINECS	68,778	13,568	20%	170	0%
ENAMINE	228,899	1,504	1%	5	0%

Table 4.18. The percentages of accurate predictions in the original training set and 3 external sets. In comparison to the training set, the external sets have drastically low percentages of accurate predictions, which shows that the applicability domain of the model is very limited.

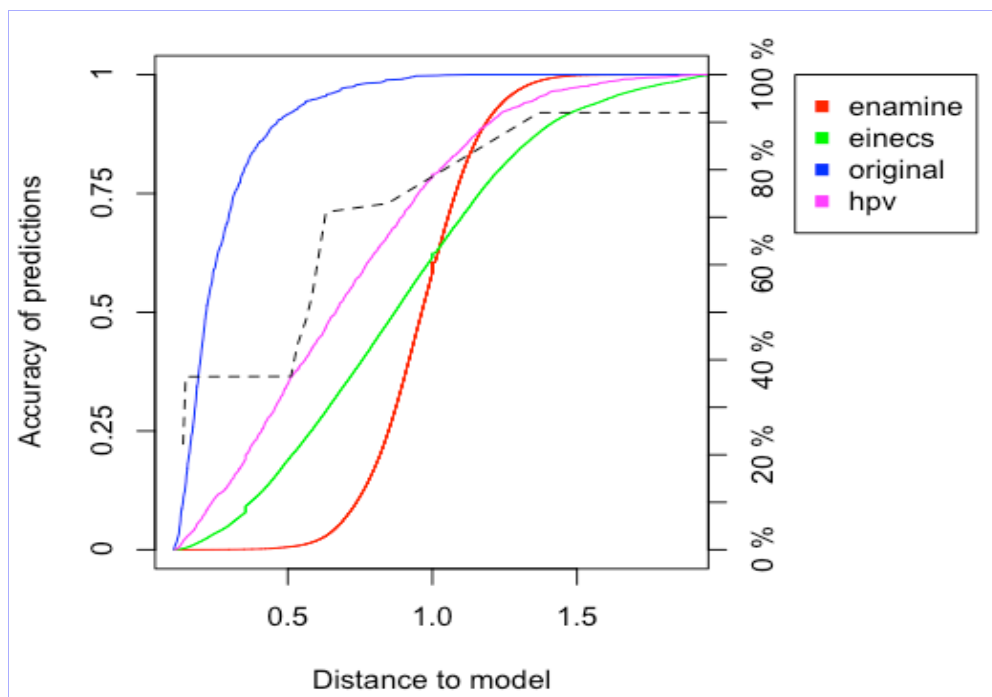


Figure 4.17. The estimated prediction accuracy for the original *T. Pyriformis* dataset (1,093 compounds), HPV dataset (2,355 compounds), EINECS (68,778 compounds) and ENAMINE (228,899 compounds) datasets. The black dashed curve, based on bin-based averaging, plots the prediction accuracy (left y-axis) against ASNN-STD DM. The colored curves show percentages of compounds from 4 datasets (right y-axis), having DM not more than a threshold (x-axis). Apparently, the distributions of the external datasets dramatically differ from the original dataset distribution.

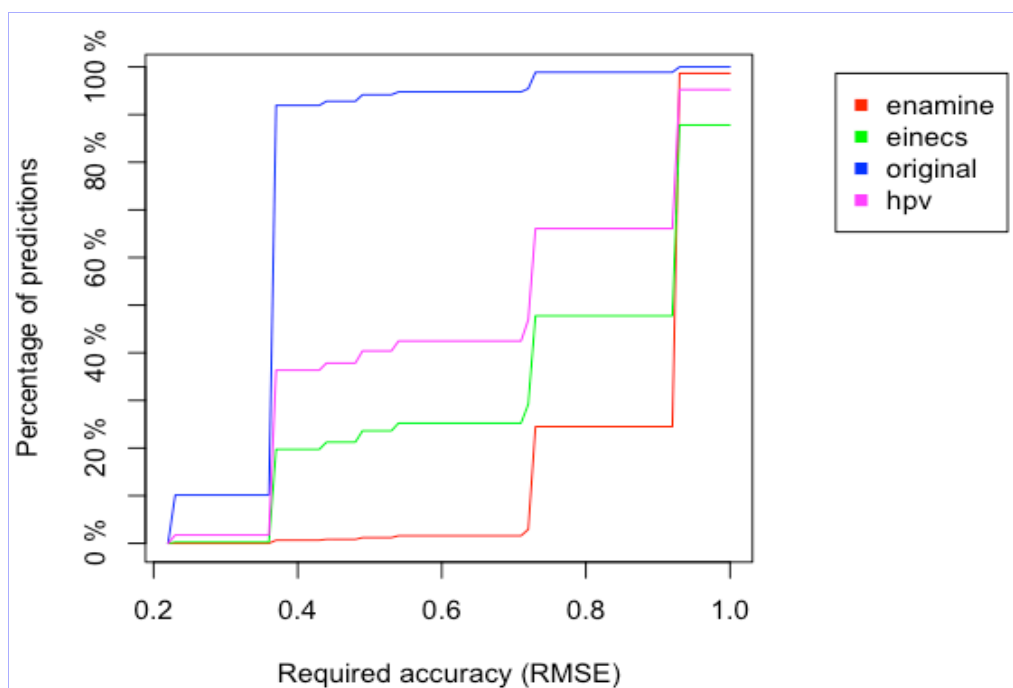


Figure 4.18. The percentages of compounds having a particular prediction accuracy. The plot is based on the previous figure. There is a dramatic difference in the training set and the external sets: while about 90% of the original compounds have RMSE of 0.4, a very low percentage (about 1%) of compounds from EINECS reach this accuracy.

4.2.4 Summary

This study furthermore confirmed that the prediction accuracy is variable in the chemical space: there are clusters of compounds that are predicted with relatively low and high accuracies. Again, the prediction reliability could be assessed using the concept of “distances to models” (DMs).

In accordance with the Ames test benchmarking, the results of this study indicated that the standard deviation of the models in an ensemble provided the best estimation of the prediction accuracy of models for toxicity on *T. pyriformis*. For example, although the average prediction accuracies were not high (RMSE of 0.44-0.59 for different QSAR approaches), STD-ASNN and STD-CONS could identify the predictions of a high accuracy, which was close to the accuracy of experimental measurements (RMSE of 0.21 and 0.38, depending on the mode of action of the compound). When we considered the threshold of 0.21, these DMs could identify such highly accurate predictions for up-to 26% of the training set and 7% of the validation set compounds. In case of the threshold of 0.38, the percentages are 90% and 72%. If the prediction accuracy is estimated in the traditional way, i.e. by averaging the accuracy over the training or validation set, it is infeasible to identify predictions of such high accuracies.

However, the situation was dramatically different for the prediction of the external diverse datasets of compounds: HPV, EINECS and ENAMINE. Only 36%, 20% and 1% of these 3 sets were estimated to have the prediction RMSE of 0.38 or better. When we further increased the accuracy requirements and considered the threshold of 0.21, the percentages decreased to 2% for HPV dataset and was close to zero for EINECS and ENAMINE datasets. Thus, these datasets contain a very limited number of compounds that can be reliably predicted by the investigated QSAR approaches.

We have also shown that a DM developed with one method and one set of descriptors could be also used to estimate the accuracy of models developed with a different set of descriptors or/and machine learning methods. For example, the DMs developed with neural networks, STD-ASNN, k- Nearest Neighbors (STD-kNN-DR) or the consensus model (STD-CONS), in most cases provided better discrimination of molecules with low and large errors for all analyzed models, even if these models were developed with different sets of descriptors and different machine learning methods. Moreover, we have also demonstrated that the STD-ASNN successfully discriminated molecules with low and large errors for the mechanism-based model based on logP and the Maximum Acceptor Superdelocalizability descriptors. Considering that the distance to the descriptors centroid did not allow the authors of the mechanism-based model to differentiate molecules with low and large errors, our approach can complement the methods based on the mechanism of action by estimating the prediction errors of such models for each chemical compound. This could be particularly useful for the prediction of new scaffolds of molecules, for which determination of the mechanism can be difficult.

5 Applications

The two further QSAR studies demonstrate the practical application of the methods introduced in this work and benchmarked in the previous chapter. The first study aims to predict the octanol-water partition coefficient for platinum complexes, which are nowadays established as important anti-cancer drugs. The second study investigates QSAR models for the identification of cytochrome inhibitors.

In the previous chapter, it was shown that the DMs based on the standard deviation (STD) provide the best estimation of prediction accuracy and, therefore, the highest quality of the AD assessment. For this reason, the analysis in the following two studies was performed using only the STD-based approaches.

5.1 Lipophilicity of Pt complexes

5.1.1 Introduction

In recent decades, platinum complexes proved to be promising medicines for cancer treatment. One of the most known and most efficient drugs from this family is cis-platin [98]. However, there are several issues that limit the usage of cis-platin and other existing platinum-based anticancer drugs: these are their toxicity and the resistance (inherent and acquired) to these compounds. Therefore, there is a need for new active platinum complexes that possess the anticancer activity but at the same time are less toxic and have less resistance [99].

Besides the anticancer activity, the candidate platinum complexes must be capable of entering the cell. As it has been shown in a number of studies [100,101], the cellular uptake of platinum complexes strongly correlates with their lipophilicity or, more precisely, with the octanol-water partition coefficient ($\text{LogP}_{o/w}$). For large numbers of compounds, many of which are possibly virtual, it is time consuming and expensive to synthesize and measure $\text{LogP}_{o/w}$ for every compound. This problem can be partially addressed with a help of *in silico* (QSAR) predictions, which can identify a low lipophilicity of a compound before it is synthesized and, thereby, filter out the virtual platinum complexes that have low cellular uptake.

There is a number of existing computational models for $\text{LogP}_{o/w}$ [34,102,103,101]. A part of the models were based on the training sets that did not contain any Platinum complexes. The usage of such models for the prediction of Pt complexes is limited. Additionally, in many cases there is no information on AD of the models, i.e there is no clear rule to identify whether $\text{LogP}_{o/w}$ for a particular compound can be predicted by the model with a desired accuracy.

This study aimed to investigate performance of various QSAR approaches based on a literature set containing $\text{LogP}_{o/w}$ experimental measurements for more than 200 Platinum complexes. The study made a particular focus on the reliability of predictions and the AD assessment.

5.1.2 Methods

A. Dataset and the variability of measurements

The dataset used to create the QSAR models contained 178 measurements for 137 unique compounds. The dataset is described on page 29. As the dataset contained several compounds with multiple measurements, it was possible to estimate the variability of $\text{LogP}_{\text{o/w}}$ experimental measurements. For this analysis, we selected the compounds with at least 3 measurements (there were 8 such compounds). The statistical parameters of experimental measurements for these compounds are reported in Table 5.1. The results suggest that the standard deviation of $\text{LogP}_{\text{o/w}}$ measurements is 0.15-0.38 log units and the average (root mean square) standard deviation is 0.26. Since we had only a limited number of compounds with multiple measurements, the 0.26 figure is only an approximate estimate. We used this value as a reference to evaluate performance of the analyzed QSARs.

Compound	Number of measurements	Mean LogPow	Standard deviation
cis-platin	7	-2.20	0.29
oxaliplatin	5	-1.61	0.15
ormaplatin	5	-1.17	0.23
JM216	5	-0.02	0.19
carboplatin	5	-1.76	0.38
dichloroethylenediamineplatinum(II)	4	-2.26	0.21
cis-dichlorobis(pyridine)platinum(II)	3	-0.40	0.32
ethylenediaminemalatoplatinum(II)	3	-1.67	0.26
Average (root mean square)			0.26

Table 5.1. The variability of $\text{LogP}_{\text{o/w}}$ measurements for 8 compounds that had at least 3 available measurements.

B. QSAR approaches and AD assessment

For modeling, we used E-States and ISIDA molecular fragments (see section “Molecular descriptors” on page 5). As additional descriptors, we used LogP and LogS predictions provided by AlogPS software, which was benchmarked as a top-ranked software for LogP predictions [21,22], but did not contain any platinum complexes in the training set. Therefore, the AlogPS software was not directly applicable in this study (for the investigated dataset, we estimated RMSE of as high as 0.84) and was used implicitly as an additional model input.

These types of descriptors were tried in various combinations. Importantly, the pH buffer was also used for the modeling as a qualitative descriptor. To incorporate information about pH buffer into the models, we used 5 additional inputs, that corresponds to the number of the used buffers. During the training, for every input sample one of the inputs was set to 1 and others to zero, depending on the pH buffer of the experimental measurement, associated with the given sample. This preprocessing was done automatically by the OCHEM platform. In names of the models, we denoted E-States as “E”, ALogPS as “A”, pH buffer as “B” and ISIDA fragments as “I”; e.g. “ASNN EA” referred to an ASNN model, based on E-states and ALogPS descriptors. Thus, we used 5 different sets of descriptors denoted as E, EB, EA, EAB and I.

To train the models, we used associative neural networks (ASNN), support vector machines (SVM), kernel ridge regression (KRR), fast stage-wise multivariate linear regression (FSMLR) and ALogPS model, trained using the so called LIBRARY

correction mode (refer to pages 7-9 for detailed description of these machine learning techniques). The AlogPS model was based on E-State indices while the first 4 methods were tried with all five sets of descriptors, which resulted in $4 \times 5 + 1 = 21$ models. Additionally, we calculated the consensus model as the average of all the 21 individual models.

For the validation, we used two protocols: the N-fold cross-validation technique and the bagging technique (page 9). In the benchmarking studies, the bagging protocol could not be employed for all QSAR models, since they were provided “as is” by a number of international groups and we could not replicate multiple copies of the models. Here, we investigate whether the bagging validation protocol is superior to the N-fold cross-validation. In contrast to the previous studies described in this work, we did not use external validation set because of a small dataset size (only 178 measurements in comparison to 6,542 and 1,093 measurements for the Ames test and T. Pyriformis toxicity datasets, respectively).

For some compounds, there were several experimental values per compound (there were 178 measurements for 137 compounds) since the measurements were carried out using different methods, different pH buffers and in different experiments. Importantly, the measurements for the same compound were included either in the training set or in the validation set but never in both simultaneously. This ensures that the model is not over-fitted.

As it was shown in the chapter “Benchmarking studies”, the standard-deviation (STD) based DMs provide the best separation of accurate and inaccurate predictions and, thereby, provide AD assessment of the highest quality. For this reason, we did not analyze other DMs here but used the STD DM exclusively. To obtain STD values, we used the ensemble of models created according to the bagging procedure³; we denoted the DM obtained in such a way as BAGGING-STD. This procedure was applied to each QSAR approach, which resulted into 21 different BAGGING-STD DMs. Additionally, we calculated the standard deviation of predictions given by 21 models, a DM referred to as STD-CONS. We compared performances of these 22 DMs for all 22 models according to the accuracy coverage and the AUC criteria. As a threshold for the accuracy criteria, we used the average RMSE of experimental measurements, which was estimated above as 0.26.

5.1.3 Results

A. Comparison of the QSAR approaches

We built 21 models using one linear (FSMLR) and 3 non-linear (ASNN, SVM and KRR) machine learning methods with different combination of descriptors and ALogPS model in the LIBRARY mode. Each model was build and validated two times: using bagging validation and 10-fold cross-validation. The RMSEs of the models are reported in Tables 5.2 and 5.3 for cross-validation and bagging, respectively.

³ In the two benchmarking studies described in the previous chapter, the bagging validation was infeasible since the models were provided “as is” by a number of international groups

Descriptors / Method	ASNN	SVM	KRR	FSMLR	AlogPS LIBRARY	Lowest RMSE
Estate	0.61	0.54	0.62	0.89	0.56	0.54
Estate + Buffer	0.65	0.55	0.61	0.91	-	0.55
Estate + AlogPS	0.68	0.58	0.59	0.64	-	0.58
Estate + AlogPS + Buffer	0.72	0.56	0.61	0.64	-	0.56
ISIDA	0.76	0.60	0.62	0.69	-	0.60
Lowest RMSE	0.61	0.54	0.59	0.64	0.56	0.54

Table 5.2. The cross-validated RMSEs of the 21 investigated models

Descriptors / Method	ASNN	SVM	KRR	FSMLR	AlogPS LIBRARY	Lowest RMSE
Estate	0.56	0.53	0.61	0.70	0.56	0.53
Estate + Buffer	0.57	0.54	0.60	0.71	-	0.54
Estate + AlogPS	0.55	0.52	0.59	0.64	-	0.52
Estate + AlogPS + Buffer	0.55	0.52	0.59	0.63	-	0.52
ISIDA	0.58	0.69 (*)	0.60	0.63	-	0.58
Lowest RMSE	0.55	0.52	0.59	0.66	0.56	0.52
Consensus model RMSE						0.50

Table 5.3. The bagging-validated RMSEs of the 21 investigated models and the consensus model

The lowest RMSE was 0.52 for bagging validation and 0.54 for cross-validation. Additionally, we calculated the lowest RMSE achieved by each set of descriptors and each machine learning method. The lowest RMSEs were 0.52-0.58 when grouped by descriptors and 0.52-0.63 when grouped by machine learning method. The bootstrap statistical test showed that neither of the descriptor sets and neither of the non-linear machine learning methods was superior with a statistically significant difference. Remarkably, in the consensus (average) model was better than all of 21 models. This model had RMSE of as low as 0.50, while the best of 21 individual models, SVM-EA-Bag had RMSE of 0.52. This result is similar to that of the benchmarking studies described in the previous chapter: the consensus model systematically showed a better performance than any of the individual models.

A specific behavior was observed with the linear FSMLR method. Namely, this method performed best when the descriptor set contained AlogPS descriptors. This suggests that the $\text{LogP}_{o/w}$ cannot be accurately approximated as a linear combination of only E-State indices and ISIDA fragment counts; this dependency is rather non-linear. However, adding AlogPS descriptors increases the accuracy of the linear approximation. In contrary, the non-linear methods could provide accurate predictions using E-States and ISIDA, e.g ASNN-E-Bag model had RMSE of 0.56 and was among the best models. Thus, as one might have expected, linear methods were limited to particular descriptors and, therefore, the non-linear methods are more universal and should be favored.

The performance of the ALogPS model in the LIBRARY mode (RMSE 0.56) was not significantly different from the best achieved performance (RMSE of 0.52 for SVM-EA-Bag model).

In general, the above results suggest that the performance of the $\text{LogP}_{o/w}$ predictions depends mostly on the size, diversity and quality of the training dataset rather than on choice of particular QSAR approach.

A remarkable point is that, as compared to the cross validation, the bagging validation provided better results for almost all the models (compare Tables 5.2 and 5.3) except of a few exceptions. The lowest RMSE achieved using bagging (0.52) was also better than without bagging (0.54). This result is concordant with the observations of the author of the bagging method Leo Breiman, who showed that bagging reduces the noise (non-systematic error) of predictions [20].

Surprisingly, the use of the pH buffer as a descriptor did not significantly increase

the performance of the models. Probably, the QSARs were not able to apprehend the dependency of LogPo/w from pH buffers due to a low number of compounds (only 19 of 137) that were measured with multiple buffers.

B. Assessment of prediction accuracy and applicability domain

Similarly to the Ames and pIGC50 QSAR studies described in previous chapters, we used the average ranks to compare different STD-based DMs. Namely, using the accuracy coverage and the AUC criteria, we ranked the 9 DMs for every model separately and averaged the ranks over all 21 models. The results are reported in Table 5.4. Apparently, according to both the criteria, the consensus-based standard deviation outperformed all the other DMs. The CONS-STD could identify up-to 53% highly accurate predictions, whereas for the other DMs the percentage was 12%-36%. About a half of all DMs (10) failed to identify any highly accurate predictions (0% coverages in Table 5.4)

DM	Accuracy coverage rank	Maximum coverage	AUC rank
STD-CONS	1.48	53%	1.04
STD-KRR-EAB-Bag	4.39	31%	9.87
STD-KRR-EA-Bag	5.89	36%	6.09
STD-KRR-EB-Bag	6.46	33%	8.61
STD-KRR-E-Bag	7.17	23%	7.65
STD-SVM-EA-Bag	8.48	20%	14
STD-FSMLR-EB-Bag	8.85	20%	9.35
STD-SVM-EAB-Bag	9.48	19%	6.74
STD-ASNN-I-Bag	10.63	15%	13.43
STD-FSMLR-EAB-Bag	14.54	15%	21.3
STD-FSMLR-EA-Bag	15.02	12%	19.26
STD-ASNN-EB-Bag	15.04	15%	4.48
STD-ASNN-EA-Bag	15.09	12%	9.61
STD-KRR-I-Bag	15.35	0%	4.43
STD-FSMLR-I-Bag	15.35	0%	13.57
STD-ASNN-E-Bag	15.35	0%	15.3
STD-SVM-I-Bag	15.35	0%	13.04
STD-SVM-E-Bag	15.35	0%	10.3
STD-ASNN-EAB-Bag	15.35	0%	12.78
STD-SVM-EB-Bag	15.35	0%	19.74
STD-FSMLR-E-Bag	15.35	0%	16.04
STD-LIBRARY-Bag	15.35	0%	18.7

Table 5.4. The average rankings of the DMs generated by different models. The consensus-based standard deviation (STD-CONS) significantly outperformed all other DMs.

To demonstrate the ability of the DMs to separate highly accurate predictions, we plotted cumulative accuracy plots for three DMs: STD-CONS (the top-ranked DM), STD-KRR-E (a middle-ranked DM) and STD-LogP (one of the lowest-ranked DMs); the plots are shown on Figure 5.1. For the purpose of comparison, the plots are shown for two models: the consensus model and the LogP model.

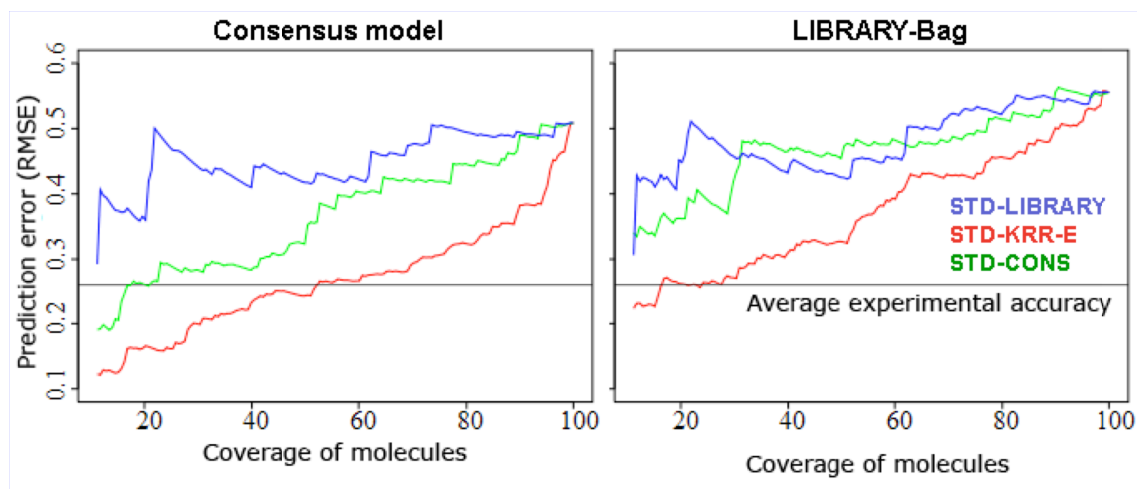


Figure 5.1. The cumulative accuracy plots based on three selected DMs for the consensus and LogP models. The STD-CONS (red curves) was able to separate highly accurate predictions for both the models, STD-KRR-E (green curves) – only for the consensus model, and STD-LIBRARY – for none altogether.

For STD-CONS (the red curve) and STD-KRR-E (the green curve), there is a clear increasing dependency; the prediction error increases with the increase of the DM. The STD-LogP (the blue curve), on the contrary, provided a noisy dependency, which does not always show an increasing trend. The three analyzed DMs differed also in their ability to separate highly accurate predictions, i.e. predictions with RMSE of 0.26, corresponding to the average accuracy of experimental measurements. Thus, the CONS-STD could identify the highly accurate predictions for both the models, about 53% for the consensus model and 24% for the LogP-Bag model; the STD-KRR could identify such predictions (21%) only for the consensus model, whereas the STD-LogP failed to identify highly accurate predictions altogether.

C. Interpretation of the AD

Analysis of $\text{LogP}_{o/w}$ values. The Figure 5.2 shows the dependency of the prediction accuracy (RMSE) from the lipophilicity (observed and predicted $\text{LogP}_{o/w}$ values). Apparently, the highest prediction accuracy was achieved for non-lipophilic compounds; the compounds with LogP_{ow} between -2 and 0 had RMSE of as low as 0.2-0.4. On the contrary, the highly lipophilic compound had a dramatically worse prediction accuracy with RMSE up-to 1.2 (the red curve in the right [0, 1] region). This figure is based on KRR-E-Bag model, but the dependency is similar for all the 21 models.

Obviously, the measured LogP_{ow} values for new compounds are unknown and cannot be used to estimate the prediction accuracy on practice. If we substitute the measured values with predicted ones, the dependency is distorted and noisy (dashed curve on Figure 5.2); the accurately predicted compounds have predicted values in ranges [-2, -1.5] and [-0.8; -0.2]. Moreover, the dependencies for the other investigated QSARs are different. Thus, the predicted values cannot be reliably used for accuracy

estimation. The rule of a thumb is that the highly lipophilic compounds are likely to have low prediction accuracy, which can be explained by the deficiency of such compounds in the training set (82% of all the available measurements have LogPow less than zero).

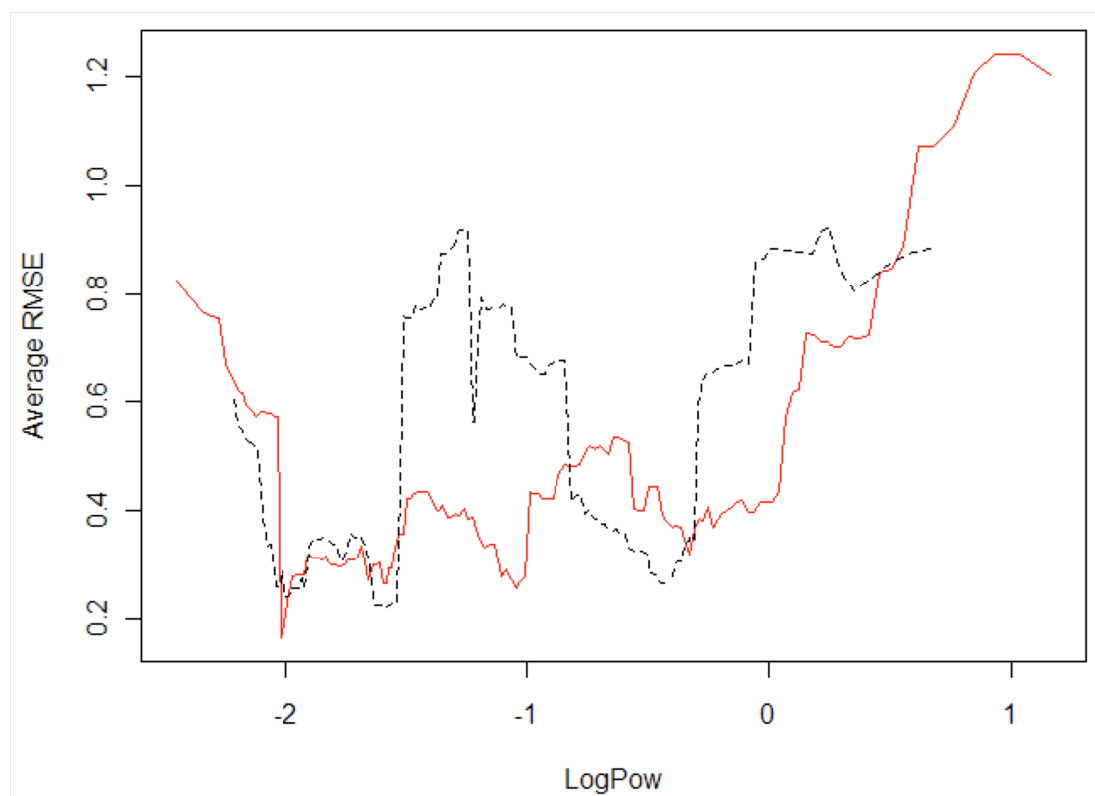


Figure 5.2. The dependency of the prediction accuracy of KRR-E-Bag from observed (the red curve) and predicted (dashed curve) values of $\text{LogP}_{o/w}$.

Substructural analysis. In contrary to the Ames test and the *T. Pyriformis* studies, for the $\text{LogP}_{o/w}$ QSAR we could not identify any molecular fragments that were over-represented inside or outside of the AD. In other words, there were no particular fragments that affected the prediction accuracy of the investigated models. A possible reason for this is significantly smaller size of the training dataset in comparison to the aforementioned studies. Thus, the interpretation of the AD in this study is based solely on lipophilicity of the predicted compounds.

5.1.4 Summary

This study investigated the problem of the AD assessment of QSAR models for prediction of the octanol-water partition coefficient ($\text{LogP}_{o/w}$) for Platinum complexes, which are nowadays promising anti-cancer agents. In total, 21 QSAR models based on different descriptors and machine learning techniques were created and compared. The models were trained on the biggest publicly available dataset of LogPow measurements for platinum complexes, which contained 178 measurements for 137 compounds. Based on this dataset, we estimated the variability of $\text{LogP}_{o/w}$ measurements: the average standard deviation was 0.26 log units. We also compared the DMs based on the standard deviation derived from a bagging ensemble of each of the investigated QSARs.

Remarkably, the bagging approach used in this study was not only helpful for STD calculation, but also provided the consensus models with higher accuracies than if

the models were trained and validated using simple cross-validation. This fact furthermore confirms the hypothesis stated by Leo Breiman, the inventor of the bagging method: ensembles of models tend to be superior to individual models.

Similarly to the previous studies, the consensus (average) model based on 21 different QSAR approaches had the lowest prediction error (RMSE of 0.50), which was lower than the errors of all the 21 individual models (RMSE 0.52-0.69). Moreover, the DM, based on the consensus standard deviation (STD-CONS) was significantly better than all the other STD-based DMs. This DM was able to identify the highly accurate predictions for about a half (53%) of the dataset. Thus, consensus models can be helpful not only for increasing prediction accuracy but also for the applicability domain assessment. This result is consistent with the benchmarking studies described in Chapter 4.

We found out that the prediction accuracy was strongly dependent on the lipophilicity of the predicted compounds. The highest prediction (RMSE as low as 0.2) was achieved for the hydrophilic compounds having $\text{LogP}_{o/w}$ between -2 and 0, whereas the highly lipophilic compounds ($\text{LogP}_{o/w} > 1$) had the highest prediction errors (RMSE of 0.8 and higher). This dependency was correctly captured by the investigated DMs, which had relatively high values for highly lipophilic compounds.

5.2 Cytochrome P450 inhibition

5.2.1 Introduction and methods

Cytochromes P450 are a family of enzymes that actively participate in the catalysis of metabolisation of both endogenous and exogenous substances. These enzymes include three big families, where the enzymes of the first family (including 1A1, 1A2 and 1B1) are very important in the first phase of metabolism of many xenobiotic compounds [104]. Many modern medicines interact with cytochromes, which is usually considered as a negative side effect that should be avoided. *In silico* predictions could assist to identify the compounds that are likely to inhibit cytochromes even before the compounds are synthesized. Thus, *in silico* models could filter out the compounds with the unwanted effects related to the inhibition of cytochromes on the earliest stage of the drug development.

This study investigated a QSAR model for the prediction of CYP1A2 inhibition activity. The model was based on the dataset that contained 7,486 compounds; 4,016 thereof were active (inhibitors of CYP1A2) and 3,470 non-active compounds. In more detail, the dataset is described in section “Analyzed datasets” on page 30. For the modeling purposes, the dataset was randomly split into training and validation sets, which contained 3,745 and 3,741 compounds respectively. As the validation set was selected randomly, the balance of active and non-active compounds in was similar to that of the original set.

For this study, we used only one QSAR approach, which was based on E-State indices and the neural networks (ASNN), since this approach showed a good performance in the benchmarking studies described in the previous chapters of this work.

For the prediction accuracy estimation and the AD assessment, we used the STD-PROB DM based on the standard deviation provided by a bagging ensemble.

5.2.2 Results

A. QSAR modeling

The performance of ASNN model validated the bagging approach is summarized in Table 5.5. The model provided the correct classification rate of 81% for both the training and validation sets. On both the sets, the model had a little higher sensitivity, which could be caused by a minor imbalance of active and inactive compounds in the training set.

Dataset	Accuracy	Sensitivity	Specificity
The training set	81%	83%	79%
The validation set	81%	83%	80%

Table 5.5. The performance of the ASNN model for prediction of p450 inhibition.

B. AD assessment

Figure 5.3 shows the cumulative accuracy plot and sliding window averaging plot (SWA-plot) based on the STD-PROB DM. The plots were built using the prediction accuracy of the ASNN model for the validation set compounds. Apparently, the highest

accuracy was 100% and was achieved for about 8% of the validation set compounds with the lowest values of STD-PROB. The accuracies of 95% and 90% were achieved for 33% and 62% of the validation set compounds, respectively. The reasons for such a high accuracy is discussed further in the interpretation section.

From the dashed curve (sliding window accuracy averaging), it is apparent that the accuracy of the most unreliable predictions was close to the accuracy of a random classifier (50%). Obviously, the compounds with such predictions are outside of the applicability domain of the model.

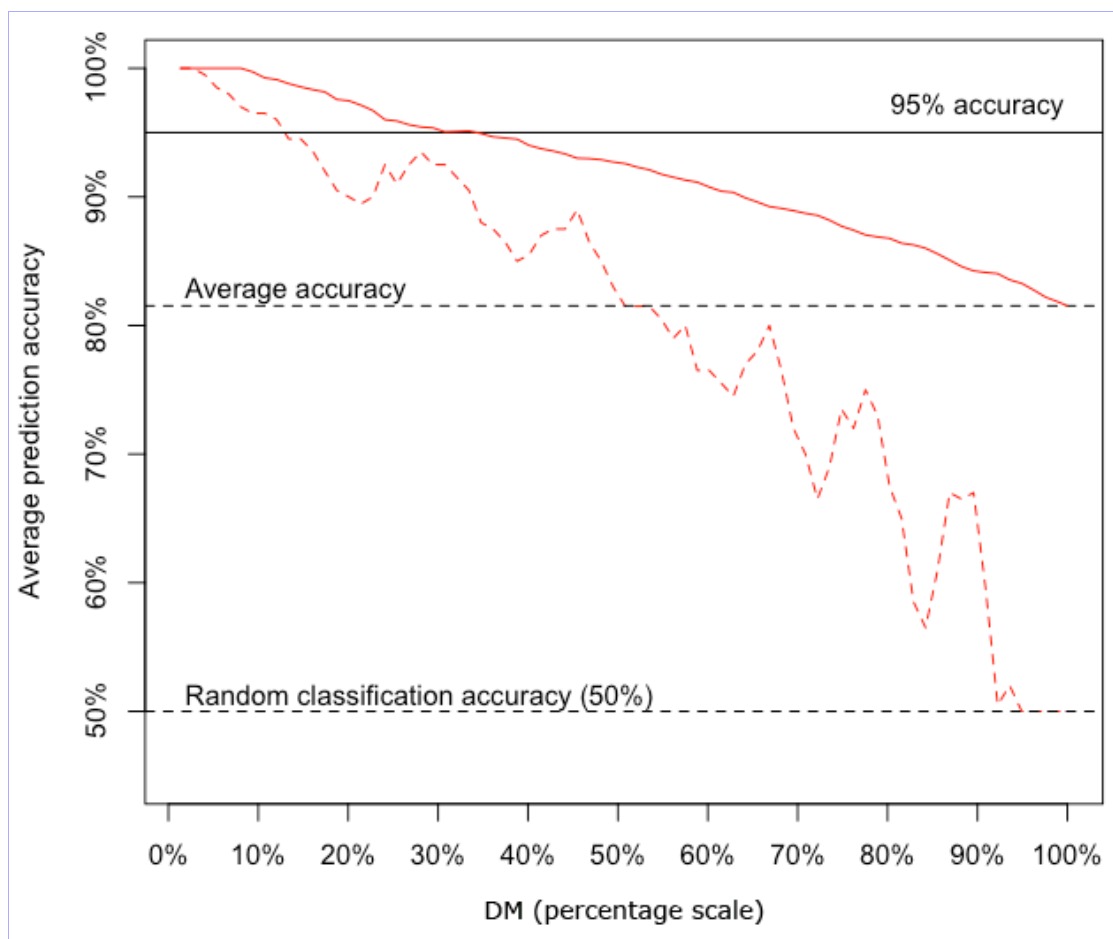


Figure 5.3. The prediction accuracy of the ASNN model for prediction of p450 inhibition depending on the STD-PROB DM (in percentage scale). The solid curve shows the cumulative accuracy for a different percentages of the validation set compounds, whereas the dashed curve shows the sliding window accuracy averaging.

C. Interpretation of the AD

Substructural analysis. We analyzed the molecular sub-fragments that were significantly over-represented in the most reliably predicted compounds. For this analysis, we considered the compounds with the highest and the lowest values of the STD-PROB DM (200 + 200 compounds) and considered significant fragments as described in the “Methodology” chapter (refer to page 27 for details). The sub-fragments that were present in the most reliably predicted compounds are shown in Figure 5.4; these compounds are quinazoline, pteridinone and 4-Pyrimidinamine derivatives, all having pyrimidine as a common sub-fragment. The correct classification rate for the molecules derived from these sub-fragments had high prediction accuracies (92%-98% of correct classifications). When we analyzed the inhibition profile of such

compounds, we discovered that majority (more than 90%) were CYP inhibitors. Furthermore, almost all such compounds were classified by the model as inhibitors, which resulted into the 100% sensitivity and the correct classification rate of 92%-100%. However, most of the rarely occurring non-inhibitors among pteridinone, quinazoline and 4-Pyrimidinamine derivatives were misclassified as inhibitors, thus resulting into a near zero specificity (0% for quinazoline-, 7% for pteridinone- and 27% for 4-Pyrimidinamine derivatives).

The reliability of the predictions for such compounds was captured by the STD-PROB DM, which is apparent from the distribution of DM values in Figure 5.5. Thus, a very high rate of correct classifications for pyrimidine-containing compounds is due to the tendency of such compounds for CYP inhibition, which was correctly apprehended both by the investigated QSAR model and DM.

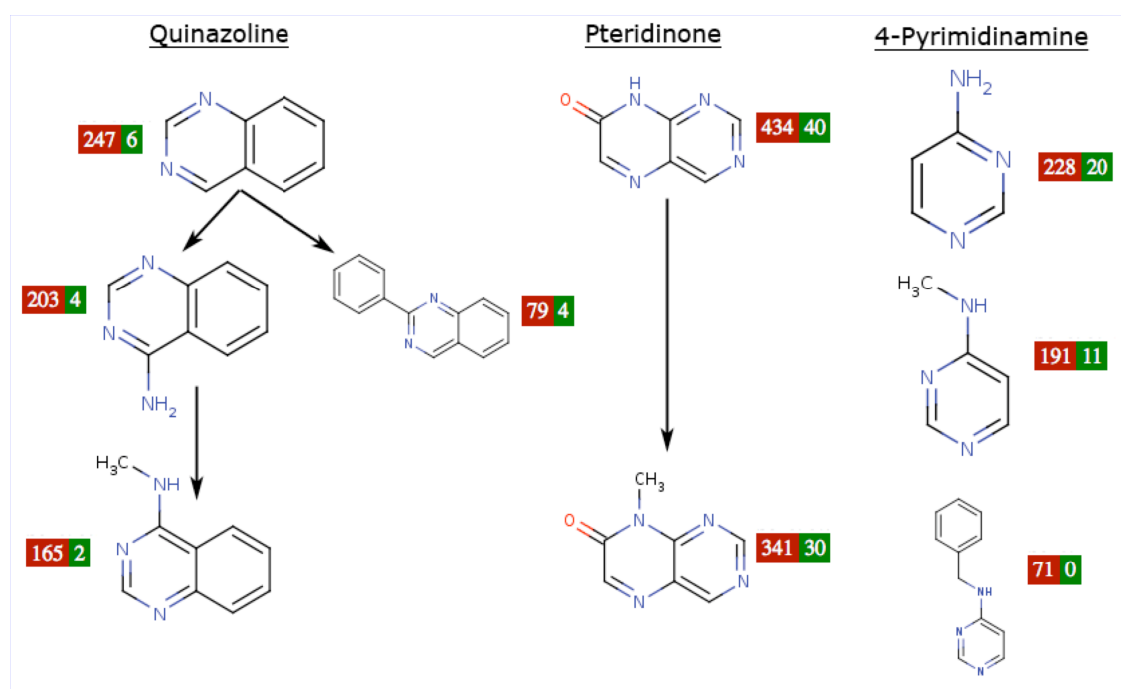


Figure 5.4. The sub-fragments that were significantly over-represented in the most reliable predictions of CYP inhibition. Most of the highly reliable predictions contained two fused aromatic rings containing nitrogens. The most of such compounds were CYP inhibitors. Green and red numbers show inhibitors and non-inhibitors containing the sub-fragment.

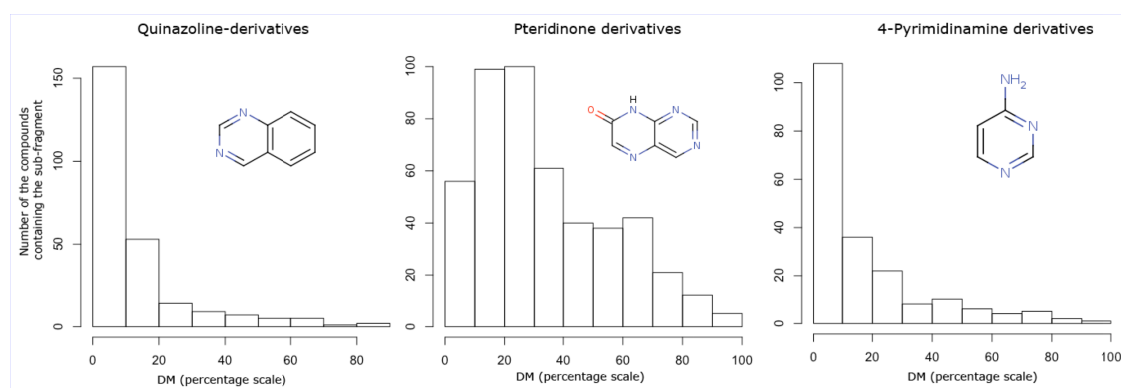


Figure 5.5. The distribution of DM values (in the percentage scale) of quinazoline, pteridinone and 4-Pyrimidinamine derivatives. Apparently, such compounds tend to have low values of DM, which means they have high reliability of predictions.

Remarkably, despite the sensitivity of the compounds with the aforementioned fragments is high (100%), for the rest of compounds the sensitivity of the model is only 69%.

Analysis of active and non-active compounds. As it was shown above, the active compounds (inhibitors) dominated among the most reliably predicted compounds. This is confirmed by the explicit analysis: among first 10% of the most reliably predicted compounds (374 compounds with the lowest DM values), 91% (332 compounds) thereof are inhibitors. Thus, a very high correct classification rate of the model for these compounds (98%) is due to a high sensitivity (99%), whereas the specificity of the model is not higher than average (80%).

D. Reliable predictions for HPV, EINECS and ENAMINE datasets

Similarly to the Ames test study (page 66), to investigate applicability of the CYP model to diverse chemical compounds, the model was applied to the ENAMINE, EINECS and HPV datasets (page 28). The results are summarized in Table 5.6 and depicted on Figures 5.6 and 5.7 (which are analogues of Figures 4.11 and 4.12 for Ames test study, page 66). Interestingly, the percentage of reliable predictions (i.e. having the estimated prediction accuracy of at least 90%) for all 3 datasets is almost equal (11-13%), whereas for the training set the percentage is 45%. This phenomenon can also be observed on Figures 5.6 and 5.7: the “original” curve is apparently higher than the curves for ENAMINE, EINECS and HPV datasets. The low percentage of reliable predictions in the external datasets attests to a limited applicability domain of the CYP model, presumably due to a lack in chemical diversity in the training set.

Predicted value	Enamine dataset		EINECS dataset		High production volume (HPV)	
	All	Reliable	All	Reliable	All	Reliable
Non-inhibitors	131390	19516	30192	4160	1073	121
Inhibitors	97509	6588	38586	4496	1282	169
Total	228899	26104	68778	8656	2355	290

Table 5.6. The reliable predictions of CYP450 inhibitors for the ENAMINE, EINECS and HPV datasets.

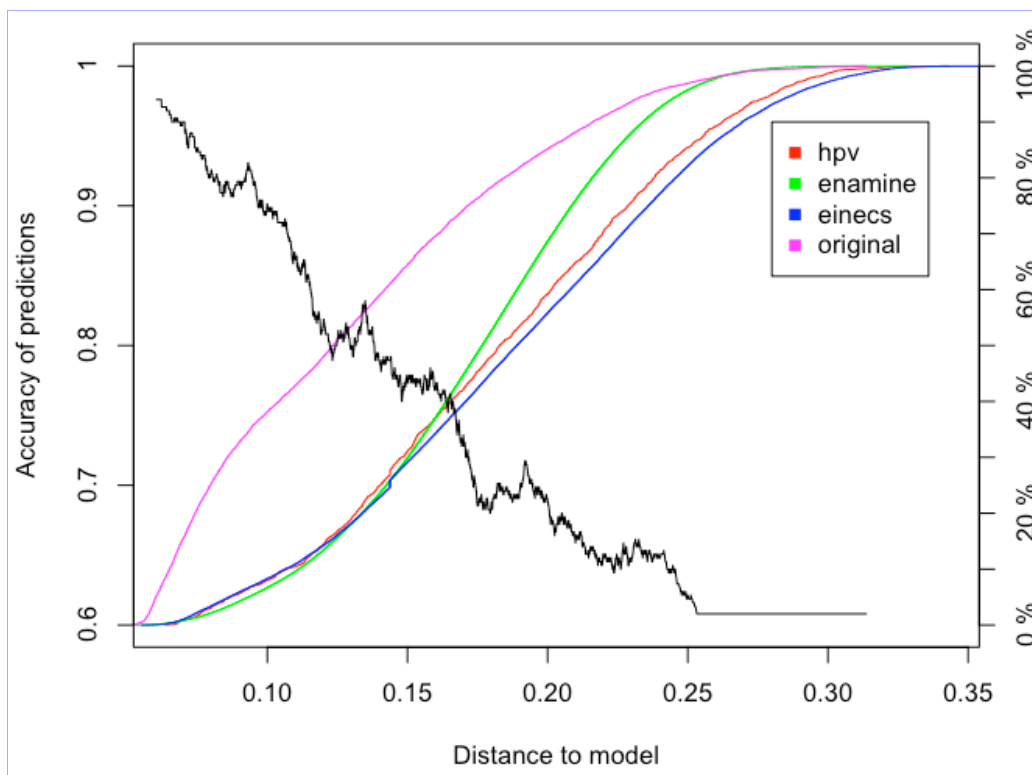


Figure 5.6. Estimated prediction accuracy for the original p450 inhibitors dataset in comparison to the HPV, EINECS and ENAMINE datasets. Black curve, based on SWA, plots the prediction accuracy (left y-axis) against ASNN-STD-PROB DM. Colored curves show percentages of compounds from 4 datasets (right y-axis), having DM not more than a threshold (x-axis).

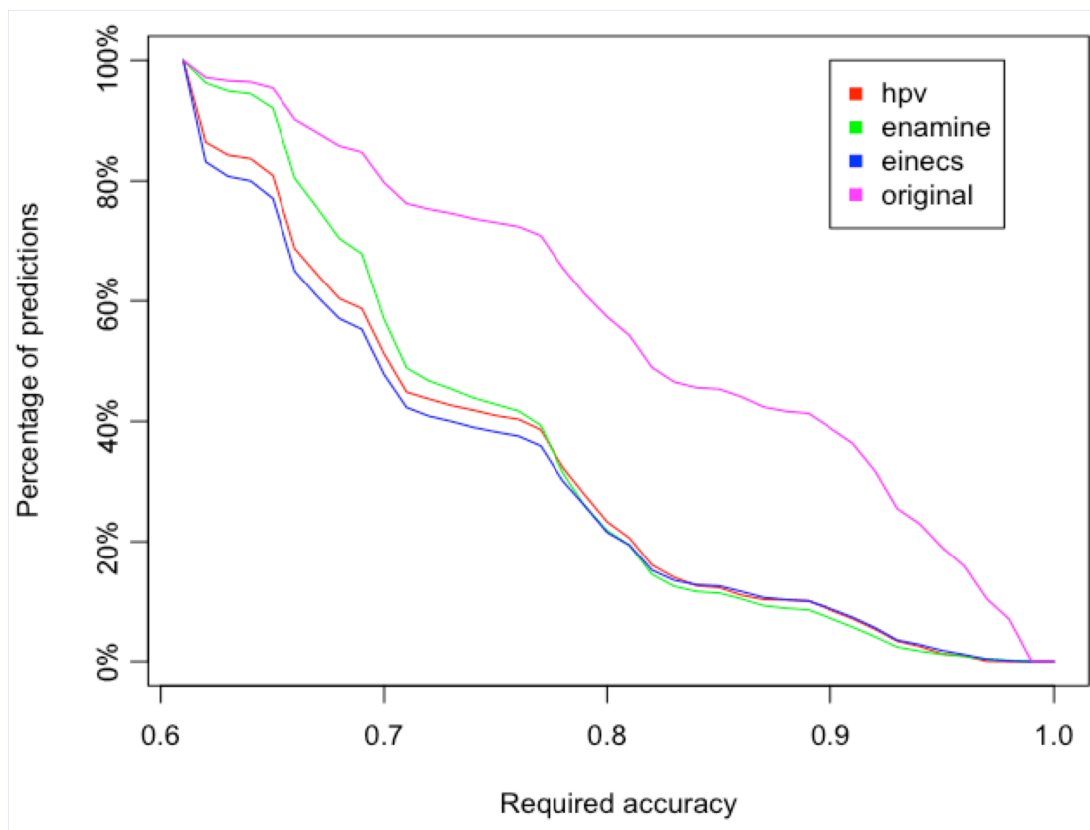


Figure 5.7. Percentages of compounds (y-axis) in 4 datasets having the estimated prediction accuracy not less than a required accuracy (x-axis). This plot is based on the plot from Figure 5.6 with DM-axis eliminated.

5.2.3 Summary

The study investigated the applicability domain of a classification model for the prediction of cytochrome P450 (CYP1A2) inhibition.

In general, the model had a good predictive ability providing 81% of correct classifications on the validation set. Furthermore, by taking only the most reliable predicted compounds identified by the STD-PROB DM, the accuracy could be increased up-to 95% for 33% of the validation set compounds and even up-to 100% for 8% of the compounds. At the same time, the DM allowed to identify unreliable predictions: the lowest detected accuracy was close to the accuracy of a random classifier (50%). Thus, similarly to the Ames test study, the STD-PROB was successful for the discrimination of accurate and inaccurate predictions.

Importantly, we could interpret such a high achievable accuracy. We showed that most of the highly reliable predictions were derivatives of quinazoline, pteridinone and 4-Pyrimidamine, the majority of which were inhibitors of CYP1A2. Due to the abundance of such compounds in the training set, the model captured this behavior and predicted most of such compounds as inhibitors, thus providing a 100% sensitivity and about 95% of correct classifications. The STD-PROB DM successfully captured the high reliability of such predictions.

Application of the investigated model to the HPV, EINECS and ENAMINE datasets showed that applicability of the model to these diverse chemical datasets is limited. Indeed, while the model could achieve 85% accuracy for at least 45% of the training and validation sets compounds, for the HPV, EINECS and ENAMINE datasets the percentage was only 11-13%. Thus, to enlarge the applicability domain of QSARs, it is necessary to perform more experimental measurements for diverse chemical compounds.

6 Discussion

A. Prediction accuracy of QSARs is variable

The current work has shown that the accuracy of QSAR models monotonically decreases with increase of the abstract measures of uncertainty, referred to as distances to models (DMs). The phenomenon of accuracy variability was apparent both quantitatively and visually (e.g., in Figures 4.1, 4.3, 4.15, 5.3 on pages 53, 56, 76 and 94).

Particularly, the variability of the accuracy was apparent for the classification problems. In this work, we investigated two classification problems: the identification of mutagenic compounds (Ames test) and of cytochrome P450 inhibitors. While the average classification accuracy for both the modeling problems was 80-81%, the accuracy of highly reliable predictions was 95-100% and, on the contrary, the accuracy of non-reliable predictions was close to 50%, which is the accuracy of a random guess (Table 6.1). Traditionally, these predictions are not discriminated and only the average accuracy (in this case 80-81%) is reported, which ignores the fact of the accuracy variability. The DM approach suggested and investigated in this work allowed to discriminate the predictions of high and low accuracy and to estimate the prediction accuracy for every particular chemical compound individually.

Classification problem	Classification accuracy of investigated QSARs		
	Average (all compounds)	10% most reliable predictions	10% least reliable predictions
Mutagenicity (Ames test)	75-81%	97%	60%
CYP inhibition	81%	100%	55%

Table 6.1. The accuracy variability identified by a DM for the classification QSARs.

A similar tendency was observed for the two regression QSARs investigated in this work (Table 6.2). For example, for the growth inhibition QSAR, the average root mean square error (RMSE) was 0.43; however, the RMSEs of 10% most and 10% least reliable predictions identified by a DM were 0.28 and 0.78, respectively. Thus, there was almost three times difference between the RMSEs of the reliable and non-reliable predictions. For the second regression problem, the prediction of lipophilicity of Platinum complexes, the difference was even more drastic: the RMSEs of the most and least reliable predictions were 0.12 and 1.12, which makes almost ten times difference.

Regression problem	RMSEs of investigated QSARs		
	Average (all compounds)	10% most reliable predictions	10% least reliable predictions
Growth inhibition for T. Pyriofmis (pICG50)	0.43	0.28	0.78
Pt complexes lipophilicity	0.51	0.12	1.12

Table 6.2. The accuracy variability for the regression QSARs.

The ability of DMs to estimate the prediction accuracy can be directly applied to the problem of AD assessment. On the contrary to previous approaches to AD, which use strict separation of the chemical space into compounds “inside” and “outside” of the AD, the DM-based methodology provides a more flexible separation, which depends on

the required prediction accuracy. For example, the study with predictions of toxicity against *T. pyriformis* showed, that only 6% of HPV database can be predicted with RMSE of 0.24 log units; however, if higher prediction errors are acceptable, the AD can be extended to 27% resulting into RMSE of 0.48. Thus, our definition of AD is based on a tradeoff between the prediction accuracy and the coverage of compounds.

B. Ensembles of models improve AD assessment

Ensembles of models were useful both for obtaining high prediction accuracies and for AD assessment.

First, it has been shown that ensembles can increase the prediction ability of QSAR models. Namely, when we took several different predictive models and created the average model, this model (so called *consensus* model) had a better prediction accuracy than any of the individual models. This phenomenon was observed for all the studies that involved the consensus model, the studies for predictions of the Ames test, the growth inhibition concentration and the octanol-water partition coefficient of platinum complexes (Table 6.3). The consensus model had lower RMSEs for the regression models and higher percentages of correct classifications for the classification models. Only for pIGC50, one of the models had the accuracy close to the consensus model (RMSE of 0.44); but that accuracy was achieved by the ASNN model, which itself is based on an ensemble of neural networks. Thus, based on these results, our general recommendation is to always create an ensemble of diverse models, which can be based either on different training sets (the bagging approach) or on completely different QSAR methods (the consensus approach).

Predicted property	Performance measure	Performance	
		Individual models	Consensus model
The Ames test	Correct classifications	75%-81%	83%
Lipophilicity of Pt complexes	RMSE	0,52-0,70	0,5
Growth inhibition concentration (pICG50)	RMSE	0,44-0,59	0,44

Table 6.3. Performance of consensus models versus individual models

Second, ensembles of models provided a way to estimate the prediction reliability. More precisely, the standard deviation (STD) of predictions given by an ensemble of models, which indicates how well the individual models agree on a particular prediction, was shown to strongly correlate with the prediction accuracy. Higher levels of agreement (and, thus, lower STD values) corresponded to higher prediction accuracies. This feature of STD allowed us to use it as a distance to model (DM) and, thereby, to estimate the prediction accuracy. Moreover, in the benchmarking studies encompassed within this work, the STD and the other STD-based DMs (STD-PROB, CONCORDANCE) were shown to be superior to all other investigated DMs.

The superiority of the STD-based DMs was proven with different tests. In the benchmarking study for *T. Pyriformis* pIGC50 QSAR, the test for the fitness of probability distribution showed that STD of ensemble of neural networks provided the best approximation for the distribution of residuals. Moreover, in the pIGC50 study the STD showed the best ability to estimate prediction accuracy. Finally, in the Ames study, the STD-PROB could provide the widest applicability domain: up to 60% of the Ames validation set could be predicted with the accuracy of inter-laboratory variation, which was estimated as 90%.

The ensemble of models can be created either using the bagging technique with different training sets or the consensus technique with different QSAR approaches. The

bagging serves 3 goals simultaneously: (a) it improves the prediction accuracy by reducing non-systematical errors (b) it can be used as a validation technique instead of N-fold cross validation and finally (c) it provides multiple predictions for every compounds, which allows to calculate the standard deviation. The consensus technique utilizes more diverse models, which are based on different descriptor sets and machine learning methods. Furthermore, the consensus technique can be combined with bagging, whereby an individual model from the consensus ensemble is created using bagging.

Our benchmarking studies showed that both the bagging STD and the consensus STD are good estimators of the prediction accuracy.

C. Property-based DMs instead of descriptor-based DMs

Importantly, the distances to models that rely on the model *outputs* (prediction values) were proven to systematically outperform DMs that rely on model *inputs* (molecular descriptors). The first category (referred to as DMs in the property space) includes such DMs as STD, STD-PROB, CORREL and CLASS-LAG. Interestingly, in many QSAR studies [35-40], mostly descriptor-based DMs were used for the AD assessment. Many of these studies used Euclidian or leverage distances in the space of descriptors. Although this approach is simple and intuitive, it takes into account only the descriptor values but not the actual predictive model, which can lead to inadequate estimation of the prediction accuracy.

What is a possible reason for the poor performance of descriptor-based DMs? The similarity of structures, i.e. similarity of molecular descriptors of two chemical compounds does not guarantee similarity of their properties. The dependency of a property from descriptors can be complex and non-linear, where a small change in descriptors can result into a significant change in the property. This phenomenon is referred to as “activity cliff” [105], i.e. a significant and poorly predictable change in the property/activity of a compound with a small change of its molecular descriptors. The “activity cliffs” are invisible in the descriptor space but can be visible in the space of predictions. The predictive model integrates information about the property dependency and is likely to have a high uncertainty (or disagreement) of predictions for the compounds on “activity cliffs”. This explains the superiority of such measures as STD and CLASS-LAG over the descriptor-based DMs.

To sum up, our benchmarking analysis confirmed the hypothesis stated by Tetko [27] that the property space DMs are superior to the descriptor-based DMs for both regression and classification models.

D. Distances to models are universal

As it was demonstrated in Chapter 4 in both the benchmarking studies, the DMs developed using one model can be successfully used with other models, based on the same training set. This phenomenon can be interpreted as follows: the most accurately predicted compounds are the same for different models based on different QSAR approaches but the same training set.

For example, the *T. pyriformis* growth inhibition QSAR study showed that the standard deviation of a neural networks ensemble (STD-ASNN) was a good estimator of prediction accuracy for the models based on the other machine learning methods, e.g. partial least squares, linear regression, k-nearest neighbors and support vector

machines. Similarly, in the study for the prediction of lipophilicity of Platinum complexes, the STD based on ensembles of kernel ridge regression models (STD-KRR) was among the best DMs for all the other models, based on such machine learning methods as support vector machines, neural networks and linear regression.

The universality of DMs suggests that a compound that had an accurate prediction by a particular QSAR approach is likely to have relatively accurate predictions with other QSAR approaches. This result leads to an important conclusion: the prediction accuracy for a particular compound mostly depends not on the modeling approach (i.e. molecular descriptors, a machine learning method) but on the training dataset or, more precisely, on the relation of this compound to the other compounds in the training set.

E. Which compounds are well predicted?

Importantly, the work not only delivered the mathematical framework for estimation of prediction accuracy but also suggested a methodology for the interpretation of the results.

The main interpretation method was the analysis of molecular sub-fragments. Such analysis allowed identifying the molecular sub-fragments of the compounds that tended to induce high (or low) prediction accuracy. For example, in the study for the mutagenicity prediction, we discovered that a significant part the compounds that were predicted with the highest accuracy contained nitro-groups, long non-saturated carbon chains, thiophene groups, acridines and phenathrenes. A more detailed analysis showed that mutagenicity of such compounds could be deduced using simple rules. Namely, the compounds containing long non-saturated carbon chains were mostly non-mutagens and, on the contrary, the compounds containing nitro groups and the compounds derived from thiophene, acridine and phenathrene were mutagenic. These simple rules were correctly apprehended by the QSAR models, which resulted into a high prediction accuracy for such compounds. A similar case was observed with the study for the cytochrome P450 inhibition: the most accurately predicted compounds (with up-to 100% of correct classifications) were derivatives of pteridinone, 4-pyrimidinamine and quinazoline, most of which were active (inhibitors of P450). Thus, a high prediction accuracy was achieved for the compounds containing a particular sub-fragment that has a simple prediction rule.

On the other side, there was a number of the “bad” sub-fragments, which contributed to a low prediction accuracy. This phenomenon was observed in the studies for the prediction of mutagenicity and growth inhibition concentration. A detailed analysis revealed that such “bad” sub-fragments contributed to a particular mechanism of action that was dissimilar from the majority of compounds. For example, halogen containing fragments like C-C-Cl or C-C-Br had the mutagenicity mechanism based on the electrophilic attack. In the study for the prediction of growth inhibition concentration, the vinyl-containing sulfones had the electrophilic mechanism of toxicity, which, by its nature, is dissimilar from the narcotic mode of action that was the “default” mode assumed for the majority of the training set compounds. These compounds had a low prediction accuracy. Thus, in both the studies, the reason for low prediction accuracy of the “bad” fragments was a specific mechanism of action.

Another important cause of the low prediction accuracy is the uncertainty of experimental measurements. The mutagenicity study showed that the compounds with high disagreement of measurements (carried out in different experiments) tend to have

low prediction accuracies. Namely, we discovered that the average agreement of the Ames tests for 150 compounds with the highest confidence of predictions was 97%, while for the 150 compounds with the lowest confidence of predictions was only 91%. The difference between these two figures was statistically significant. Thus, the accuracy of predictions is affected by the uncertainty of experimental measurements.

F. Accuracy of experimental measurements is achievable with QSARs

It was shown that, for some part of chemical compounds, QSAR models can deliver the prediction accuracy comparable to the accuracy of experimental measurements. Despite the average model accuracy for a particular validation set may be relatively low, by using the accuracy assessment techniques described in this work, it is often possible to separate highly accurate predictions with an accuracy comparable to the that of experimental measurements.

The percentage of highly reliable predictions was estimated for the datasets of diverse chemical compounds: HPV, EINECS and ENAMINE datasets (Table 6.4). For example, in the study for prediction of mutagenicity (the Ames test) based on more than 6,000 compounds, it was possible to achieve the accuracy that corresponds to the inter-laboratory agreement (90%) for up-to 25% of the 2,181 validation set compounds, 30% of 2,355 HPV set compounds, 18% of 68,778 EINECS set compounds and 4% of 228,899 ENAMINE set compound.

Predicted property	Accuracy of experimental measurements	Average prediction accuracy	% of compounds predicted with experimental accuracy			
			Validation set	HPV set	EINECS set	ENAMINE set
Ames test	90%*	75%-81%*	25%	30%	18%	4%
pIGC50	0,38**	0,44-0,59**	92%	36%	20%	1%
LogPow for Platinum complexes	0,26**	0,50-0,70**	53%	-	-	-

Table 6.4. The percentage of compounds predicted with the average accuracy of experimental measurements.

Remarks: *The percentage of correct classifications (for accuracy of experimental measurements – the inter-laboratory agreement of measurements). ** The standard deviation (root mean square error, RMSE).

For the chemicals predicted with a high accuracy, *in silico* predictions can be used to avoid the experimental measurements. For example, it was estimated that Ames mutagenicity test QSARs can deliver the predictions with an accuracy corresponding to inter-laboratory agreement for more than 13,000 out of 229,000 compounds from the ENAMINE set, which contains drug-like compounds. These 13,000 compounds are non-mutagens with 90% accuracy and can be used for further filtering with other QSARs or high-throughput screening techniques.

G. More diverse measurements for better models

The above analysis of highly reliable predictions (Table 6.4) also indicates that the AD of the investigated models is limited. Indeed, although the models could reliably predict a significant part of the validation set compounds (25%-92%), of the the HPV compounds (30-36%) and of the EINECS compounds (about 20%) but, nonetheless, the percentage of reliable predictions for a larger and more diverse ENAMINE dataset was significantly lower. For example, only 1% of the ENAMINE compounds were

estimated to have reliable predictions, while for the original validation dataset the percentage was as high as 92%. Presumably, the lack of accurate predictions is caused by a limited diversity of the training set compounds, which results into a poor predictive ability for the majority of ENAMINE compounds. Moreover, this assumption attests to our previously mentioned claim: the prediction accuracy depends not on the modeling approach but on the relation of the predicted compound to the training set compounds.

Thus, these results deliver an important practical conclusion: to broaden applicability domains of QSAR models, it is necessary to ensure a high diversity of compounds within the datasets with experimental measurements used for model training. To make QSAR models more universal, experimentalists should focus not on making more measurements, but on measuring more diverse compounds.

The AD approaches introduced in this work can be used to aid the experimental design. Namely, valuable measurements would be for the compounds outside of AD of existent models. Therefore, our recommendation for experimental design is to measure the compounds that have the lowest estimated prediction accuracy, i.e. the compounds having the largest values of DMs. For example, in case of using STD as a DM, it is recommended to measure the compounds that have the highest level of disagreement in predictions given by different models. Such compounds would significantly extend applicability domains of existent QSAR models.

7 Conclusions and outlook

A model can be used successfully only if its limitations are known. This rule applies to all kind of models: for example, in physics, the Newtonian mechanics is valid only for speeds that are significantly lower than the speed of light. No model describes the reality ultimately, every model has its limitations, its domain of applicability. The current thesis work showed that this fact is especially true for QSAR models: based on the knowledge collected from a limited set of chemicals, QSAR models can reliably predict only a part of the chemical space and under no circumstances are guaranteed to give accurate predictions for the whole chemical space.

In this work, I introduced the methodology and developed the practical tools to build QSAR models and to assess their domain of applicability. The methodology is based on abstract measures of the prediction uncertainty, referred to as distance to models (DMs). The DM approach allowed to estimate the accuracy of every prediction individually, thus allowing to restrict the applicability domain (AD) of the model to the compounds predicted with the required accuracy.

The work provides not only the methodology but also a robust implementation. Namely, all the AD methods introduced in this work as well as the already established QSAR techniques were integrated into the novel online platform for QSAR research, the Online Chemical Modeling Environment – OCHEM, which supports all the steps of a typical QSAR research: collection and preparation of data, calculation of molecular descriptors, application of machine learning methods and, finally, the AD assessment. The database integrated within this platform already contains more than 160,000 experimental measurements of more than 300 biological and physicochemical properties. These data have a significant value for the potential QSAR studies. The OCHEM is open on the Web and is intensively used for various QSAR studies.

The introduced methods were benchmarked with and applied to a number of practical studies, which involved predictions of both biological and physicochemical properties such as mutagenicity, lipophilicity, toxicity and CYP450 inhibition. The studies confirmed that the prediction accuracy is variable in the chemical space and, more importantly, it can be estimated. For all the modeling studies, it was possible to identify highly accurate predictions with the accuracy comparable to that of experimental measurements. On the other side, it was possible to identify unreliably predicted compounds, which had accuracy of a random guess. Thus, the studies attested to the irrelevance of the question “*Is the model accurate?*” and provided the answer to the appropriate question “*Can the model accurately predict this particular compound?*”.

The limited applicability domain of the computational models does not invalidate them and does not prevent their successful practical application. In the main fields of QSAR research, drug design and environmental toxicity assessment, QSAR models can be used successfully, but only if their limitations are made clear. Precisely this problem, that is the determination of the restrictions of QSAR models, has been addressed in this work.

Outlook. A problem that is interesting for the future work concerns a methodology for the interpretation of prediction accuracy. For using the proposed approaches in industry, it is not sufficient just to be aware that the prediction for a particular compound is unreliable. It is of crucial importance to understand the reason for a low prediction accuracy. Why the activity of an investigated compound cannot be explained by the model? Is it caused by a particular mechanism of action that was not captured by the model or is it just an inaccurate measurement? The “black box” approach cannot provide an ultimate answer to this question and should be complemented with a comprehensive methodology for the interpretation of results.

The proposed approaches for AD assessment are promising in the area of experimental design. Often, there is a need to estimate a particular property for a large number of compounds, while the number of possible experimental measurements is limited with budget and time. The AD approaches can identify which compounds cannot be reliably predicted with existent QSAR models and, thereby, can help to identify valuable experimental measurements. The application of the proposed approaches in experimental design is a promising direction of research in future.

The author hopes that this thesis work will contribute to the widespread use of computational QSAR models in the drug design and ecotoxicity assessment fields.

List of abbreviations

AD	Applicability domain
ADME	Absorption, distribution, metabolism, excretion
ASNN	Associative neural network
AUC	Area under curve
BBA	Bin-based averaging
CCR	Correct classification rate
CONS	Consensus model
CV	Cross-validation
CYP	Cytochrome p450
DM	Distance to model
ECHA	European Chemicals Agency
EINECS	European INventory of Existing Commercial chemical Substances
EPA	Environmental Protection agency
ERD	Estimated distribution of residuals
FSMLR	Fast stagewise multivariate linear regression
HPV	High production volume
IGC	Growth inhibition concentration
InChi	International chemical identifier
KNN	K-nearest neighbors
KRR	Kernel ridge regression
LOO	Leave one out (a method for model validation)
MAE	Mean absolute error
MFC	Molecular fragments counts
MGD	Multi-gaussian distribution
MLR	Multivariate linear regression
NIH	National Institute of Health
OCHEM	Online chemical modeling environment
PCA	Principal components analysis
QSAR	Quantitative structure-activity analysis
QSPR	Quantitative structure-property analysis
REACH	Regulation on Registration, Evaluation, Authorization and Restriction of Chemicals
RMSE	Root mean square error
SDF	Structure data file
SGD	Single-gaussian distribution
SMILES	Simplified molecular input line entry specification
STD	Standard deviation
STD-PROB	Standard deviation and probability based DM

Alphabetical Index

Accuracy averaging.....	22	Linear regression.....	7
Accuracy coverage.....	24	Mean absolute error (MAE).....	11
Applicability domain.....	14	Mixture of Gaussian distributions (MGD).....	26
Approval test for DM.....	26	Model ensemble.....	9
AUC.....	24	MOL2.....	13
Bagging.....	9	Molecular descriptors.....	6
Bin based averaging (BBA).....	22	Neural networks.....	8
Bootstrap test.....	13	OCHEM.....	34
CLASS-LAG.....	18	Over-fitted models.....	10
Coefficient of determination.....	11	Percentage scale.....	22
Concordance.....	19	R-square.....	11
Confidence consistency plot.....	26	Ridge linear regression.....	7
Consensus model.....	9	Root mean square error (RMSE).....	11
Correct Classification Rate (CCR).....	12	SDF.....	13
CORREL.....	18	Sensitivity.....	12
Critical DM.....	24	Sliding window averaging (SWA).....	22
Cross-validation.....	10	SMILES.....	13
Cumulative averaging.....	22	Specificity.....	12
Distance to model.....	15	Standard deviation.....	16
Estimated distribution of residuals.....	26	STD-PROB.....	20
InChi	13	Substructure analysis.....	27
K-nearest neighbors method (KNN).....	8	Support vector machines (SVM).....	8
Leave One Out (LOO).....	10	Tanimoto similarity.....	18
Leverage.....	16		
LIBRARY model correction.....	9		
Likelihood criterion.....	26		

List of Figures

Figure 1.1. An overview of QSAR: the applications and predicted properties.....	1
Figure 1.2. An illustrative example for the applicability domain problem. In the green region, the data are very well approximated with a linear model (red line). However, outside the green region, the approximation is not valid. Thus, the green region ([-1, 1] interval) defines the applicability domain of the linear model.....	2
Figure 2.1. An example of the accuracy discrimination. As it can be seen on the scatter plots, the green compounds have higher prediction accuracy (the leftmost plot, RMSE 0.51) than the red compounds (the middle plot, RMSE 0.77). When mixed together, the compounds have RMSE of 0.72 (the rightmost plot).....	15
Figure 2.2. An illustrative example of the leverage DM. Leverage penalizes the compounds that are far from the center of the training set in the space of molecular descriptors. According to leverage, such compounds are unreliably predicted.....	17
Figure 2.3. An example of three predictions with different standard deviations (STD). According to the STD DM, reliable predictions have a low prediction “spread”, which corresponds to the disagreement of individual predictions within an ensemble of models.	17
Figure 2.4. Graphical demonstration of the CLASS-LAG DM. According to this measure, the most unreliable predictions (i.e., the highest CLASS-LAG values) are near to the borderline that divides active and inactive compounds.....	19
Figure 2.5. An example of four predictions with different reliability according to the STD-PROB DM. The reliability is affected by two factors: the standard deviation (the “flatness” of the curve) and the shift of the curve from the center. Ultimately, these two factors are combined into a single numerical representation, which corresponds to the filled area and is referred to as STD-PROB.....	21
Figure 2.6. The bin-based averaging (BBA) of the prediction accuracy. The red dots represent the compounds from the investigated set; the black lines represent the averaged errors (RMSE) over different DM intervals (“bins”).....	23
Figure 2.7. An example of a cumulative accuracy plot. This plot shows the RMSE of the predictions with DM less than a variable threshold. “100%” corresponds to the RMSE of all the predictions for the investigated set. Two percentages are highlighted: RMSE of 40% compounds of most reliable predictions is around 0.34, whereas RMSE of 100% compounds (the global RMSE) is around 0.49.....	23
Figure 2.8. Identification of the accuracy coverage using the cumulative plot based on three DMs and the Ames test classification model. With 90% threshold, the “red” DM is superior to the others; it covers about 45% of the compounds from the validation set, while the other two DMs cover about 30% and 15% respectively. With 85% threshold, there is no difference between the “red” and “blue” DMs; both cover about 65% of the compounds.....	25

Figure 2.9. The area-under-curve (AUC) criterion corresponds to the filled area between the SWA plot and the average accuracy (the horizontal line).....	25
Figure 2.10. An example of confidence consistency plots. The black line is the optimal (identity) plot, the blue and red lines are based on SGD and MGD distributions; apparently, in this example the MGD approximates the optimal plot better. The scale is adjusted to highlight the higher percentages.....	27
Figure 3.1. A schematic overview of the OCHEM database.....	34
Figure 3.2. The first step of the model creation: the selection of training and validation sets, the machine learning method and the validation protocol.....	38
Figure 3.3. Choice and configuration of molecular descriptors.....	39
Figure 3.4. A screenshot of the registry of the pending QSAR models.....	41
Figure 3.5. Distribution of calculations in OCHEM.....	42
Figure 3.6. Basic statistics for a predictive model. The training set has a link that opens a browser of experimental records where a user can examine properties of all compounds used to in the model. A click on a dot in the observed-vs-predicted chart opens a similar browser information window for the corresponding compound.....	43
Figure 3.7. Statistics of a classification model. Summarized are the prediction accuracies and the confusion matrices for the training and test sets.....	43
Figure 3.8. An overview of the models based on the same training set.....	44
Figure 3.9. The registry of models in the OCHEM system.....	44
Figure 3.10. An example of the bin-based accuracy averaging, which is used in OCHEM for the estimation of prediction accuracy.....	46
Figure 3.11. The prediction for new compounds and the estimation of the prediction accuracy in the OCHEM system.....	47
Figure 4.1. The prediction accuracy of the neural network model as a function of CONS-STD and CONS-STD-PROB. The solid lines (sliding-window averaging) show the averaged accuracy on the moving window with 200 compounds. Although there is a trend that the accuracy of prediction decreases with an increase of the DMs, the dependency is not smooth and there are significant fluctuations. The cumulative averaging (dashed lines) smooths the variations, which makes it more suitable for the comparison of DMs.....	53
Figure 4.2. The PCA plot of the Ames challenge models based on the space of predictions for the test set. Four models (UI_Drag_LDA, UBC_ID_IWNN, ULZ_3DDrag_SVM, ULZ_3DDrag_KNN) are not shown, since they were apparent outliers.....	55
Figure 4.3. The cumulative accuracy-coverage plot for CONS-STD-PROB DM based on the test set predictions. The curves show the accumulative accuracy for a particular (variable) percentage of compounds. The curves clearly show that CONS-STD-PROB is highly correlated with the prediction accuracy.....	56

Figure 4.4. The cumulative accuracy-coverage plot for CLASS-LAG based on the test set predictions.....	56
Figure 4.5. The distribution of the prediction values for two exemplary models. The prediction values of the model on the left chart resemble rounded discretized “-1” and “1” values, whereas the values on the right chart have a continuous distribution and, therefore, provide more information for the estimation of prediction reliability. This fact is confirmed in practice: CLASS-LAG of UNC_SiRMS_SVM (left chart) has poor performance (0% coverage of 90% accuracy) in contrary to PCI_SiRMS.Drag_RF (right chart), which separates 63% of compounds with 90% prediction accuracy.....	57
Figure 4.6. A comparison of AD_MEAN DM (solid lines) with CONS-STD-PROB DM (dashed lines) for the UNC SVM models. Apparently, CONS-STD-PROB provides a better separation of highly accurate predictions.....	58
Figure 4.7. A principal components plot for the analyzed DMs. The PCA was based on the rankings that the DMs gave to the compounds from the training and test sets. Apparently, the 5 consensus-based DMs form two close clusters: CONS-STD, CONS-STD-QUAL and CONCORDANCE in the first cluster and CONS-STQ-QUAL-PROB and CONS-STD-PROB in the second one.....	59
Figure 4.8. The molecular sub-fragments presented in the reliably and non-reliably predicted compounds. Shown are the sub-fragments that were significantly overrepresented in the molecules having the most and the least reliable 400 predictions (A and B) according to the CONS-STD-PROB DM. Below the fragments are the numbers of the relevant molecules with the most reliable (left of the dash) and the least reliable (right of the dash) predictions.....	63
Figure 4.9. The distribution of the pairwise agreements of the Ames test measurements carried out by 12 laboratories. The 0.5 value on x-axis corresponds to the complete disagreement of two laboratories. The data for the plot was taken from a study by Benigni et al [92].....	63
Figure 4.10. The distribution of CONS-STD-PROB (in percentage scale) for the molecules having at least one non-concordant (falling out) Ames test result. The green and red curves correspond to the training and test sets. Apparently, such molecules have bias towards larger values of CONS-STD-PROB. This fact further confirms the hypothesis: the prediction uncertainty determined by the DM is partially explained by the uncertainty of experiments.....	65
Figure 4.11. The estimated prediction accuracy for the original Ames challenge dataset, HPV, EINECS and ENAMINE datasets. The black curve, based on SWA, plots the prediction accuracy (left y-axis) against the ASSN-STD-PROB DM. The colored curves show the percentage of compounds from the 4 datasets (right y-axis), having DM not more than a threshold (x-axis).....	66
Figure 4.12. The percentage of compounds (y-axis) from the 4 datasets having the estimated prediction accuracy not less than a required accuracy (x-axis). This plot is based on the plot from Figure 4.11 with the DM-axis eliminated.....	67
Figure 4.13. Analysis of the ASSN-ESTATE model. The MGD for the training and joint validation sets are shown for STD-CONS (A) and EUCLID-kNN-MZ (B) DMs. The MGDs are based on the bin-based averaging (BBA). As it is seen in (C), the distribution	

of errors for the LEVERAGE-OLS-DR did not calculate a significant MGD. The confidence consistency plot for two exemplary DMs is shown in (D).....	75
Figure 4.14. Analysis of the OLS-DR model given by eq 4.4. The STD-CONS DM (right plot) provides a better discrimination of molecules with low and large errors compared to that of LEVERAGE-OLS DM (left plot). The vertical line on the left plot corresponds to the leverage threshold $3(K+1)/N = 3*7/664 = 0.033$ (the “warning” leverage).....	75
Figure 4.15. The BBA for the mechanism-based model (eq 4.5). The use of STD-ASNN DM allowed for the discrimination of molecules with low and large errors in both training and validation sets.....	76
Figure 4.16. The average prediction accuracy (RMSE) depending on pIGC50 values. Apparently, predictions are more accurate for compounds with average or less than average pIGC50 values ([-0.7, 1.0]); such compounds have RMSE of as low as 0.35. Predictions with high (more than 1.0) pIGC50 values have low accuracy (RMSE up-to 0.9). However, when the real pIGC50 values are substituted with the predicted ones (the green curve), they cannot discriminate predictions of high and low accuracy.....	81
Figure 4.17. The estimated prediction accuracy for the original T. Pyriformis dataset (1,093 compounds), HPV dataset (2,355 compounds), EINECS (68,778 compounds) and ENAMINE (228,899 compounds) datasets. The black dashed curve, based on bin-based averaging, plots the prediction accuracy (left y-axis) against ASNN-STD DM. The colored curves show percentages of compounds from 4 datasets (right y-axis), having DM not more than a threshold (x-axis). Apparently, the distributions of the external datasets dramatically differ from the original dataset distribution.....	83
Figure 4.18. The percentages of compounds having a particular prediction accuracy. The plot is based on the previous figure. There is a dramatic difference in the training set and the external sets: while about 90% of the original compounds have RMSE of 0.4, a very low percentage (about 1%) of compounds from EINECS reach this accuracy.....	83
Figure 5.1. The cumulative accuracy plots based on three selected DMs for the consensus and LogP models. The STD-CONS (red curves) was able to separate highly accurate predictions for both the models, STD-KRR-E (green curves) – only for the consensus model, and STD-LIBRARY – for none altogether.....	90
Figure 5.2. The dependency of the prediction accuracy of KRR-E-Bag from observed (the red curve) and predicted (dashed curve) values of LogPo/w.....	91
Figure 5.3. The prediction accuracy of the ASNN model for prediction of p450 inhibition depending on the STD-PROB DM (in percentage scale). The solid curve shows the cumulative accuracy for a different percentages of the validation set compounds. whereas the dashed curve shows the sliding window accuracy averaging.....	94
Figure 5.4. The sub-fragments that were significantly over-represented in the most reliable predictions of CYP inhibition. Most of the highly reliable predictions contained two fused aromatic rings containing nitrogens. The most of such compounds were CYP inhibitors. Green and red numbers show inhibitors and non-inhibitors containing the sub-fragment.....	95

Figure 5.5. The distribution of DM values (in the percentage scale) of quinazoline, pteridinone and 4-Pyrimidinamine derivatives. Apparently, such compounds tend to have low values of DM, which means they have high reliability of predictions.....95

Figure 5.6. Estimated prediction accuracy for the original p450 inhibitors dataset in comparison to the HPV, EINECS and ENAMINE datasets. Black curve, based on SWA, plots the prediction accuracy (left y-axis) against ASNN-STD-PROB DM. Colored curves show percentages of compounds from 4 datasets (right y-axis), having DM not more than a threshold (x-axis).....97

Figure 5.7. Percentages of compounds (y-axis) in 4 datasets having the estimated prediction accuracy not less than a required accuracy (x-axis). This plot is based on the plot from Figure 5.6 with DM-axis eliminated.....97

List of Tables

Table 2.1. Examples of SMILES codes for simple molecules.....	13
Table 2.2. An overview of the datasets of experimentally measured properties.....	28
Table 2.3. A brief overview of the investigated industrial datasets of chemical compounds.....	30
Table 3.1. The configurable parameters of the machine learning methods in OCHEM and their effects.	40
Table 3.2. The details of the bin-based average example.....	46
Table 4.1. The 12 international groups and their models for the Ames test predictions.	50
Table 4.2. The models that participated in the Ames test AD benchmarking.....	50
Table 4.3. The average ranks of the DMs considering their coverage of the 90% prediction accuracy.....	55
Table 4.4. The averaged rankings of the DMs ranked by the AUC criterion.....	55
Table 4.5. Accuracy of predictions according to the CONS-STD-PROB. For first 500 compounds, it achieved the accuracy as high as 93-96%.....	60
Table 4.6. The CONS-STD-PROB and TRUST LEVEL juxtaposed for the ULP_ISIDA_SQS model.....	60
Table 4.7. The CONS-STD-PROB and SVM1-AD DMs juxtaposed for the MSU models.....	60
Table 4.8. The CONS-STD-PROB and DA-Index DMs juxtaposed for the TUB models.	61
Table 4.9. A summary of non-concordant measurements.....	65
Table 4.10. Reliable Ames test predictions for ENAMINE, EINECS and HPV databases. * “reliable” predictions are those with the estimated prediction accuracy of at least 90%, which corresponds to the inter-laboratory variations.....	67
Table 4.11. A summary of the analyzed QSAR approaches for pIGC50 prediction.....	70
Table 4.12. The statistical parameters of the investigated pIGC50 models.....	71
Table 4.13. The averaged rankings of the DMs according to the accuracy coverage criterion (sorted by rankings based on the validation set).....	77
Table 4.14. The average rankings of the DMs according to the AUC criterion (sorted by rankings based on the validation set).....	77

Table 4.15. The averaged rankings of the DMs according to the likelihood-score criterion and the number of models, for which the DMs failed to pass the approval tests	78
Table 4.16. Estimated errors on the validation set. 1 average predicted RMSE (e.g., using STD-ASNN DM we predicted RMSE for all 12 analyzed models and averaged them). 2 average absolute differences between predicted and actual RMSE for all methods (e.g., using STD-ASNN DM we predicted RMSE for all 12 models and calculated average absolute difference between predicted and RMSE errors for all models).....	79
Table 4.17. The compounds predicted by the ASNN model with the accuracy significantly higher (or lower) than the average accuracy of the model (0.47).....	80
Table 4.18. The percentages of accurate predictions in the original training set and 3 external sets. In comparison to the training set, the external sets have drastically low percentages of accurate predictions, which shows that the applicability domain of the model is very limited.....	82
Table 5.1. The variability of LogPo/w measurements for 8 compounds that had at least 3 available measurements.....	86
Table 5.2. The cross-validated RMSEs of the 21 investigated models.....	88
Table 5.3. The bagging-validated RMSEs of the 21 investigated models and the consensus model.....	88
Table 5.4. The average rankings of the DMs generated by different models. The consensus-based standard deviation (STD-CONS) significantly outperformed all other DMs.....	89
Table 5.5. The performance of the ASNN model for prediction of p450 inhibition.....	93
Table 5.6. The reliable predictions of CYP450 inhibitors for the ENAMINE, EINECS and HPV datasets.....	96
Table 6.1. The accuracy variability identified by a DM for the classification QSARs..	99
Table 6.2. The accuracy variability for the regression QSARs.....	99
Table 6.3. Performance of consensus models versus individual models.....	100
Table 6.4. The percentage of compounds predicted with the average accuracy of experimental measurements.	103

References

1. Dimasi JA. Risks in new drug development: approval success rates for investigational drugs. *Clin. Pharmacol. Ther.* 2001 May;69(5):297-307.
2. Kubinyi H. QSAR and 3D QSAR in drug design Part 2: applications and problems. *Drug Discovery Today.* 1997 Dec;2(12):538-546.
3. Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA, Wikel JH. Progress in predicting human ADME parameters in silico. *Journal of Pharmacological and Toxicological Methods.* Jul;44(1):251-272.
4. Perkins R, Fang H, Tong W, Welsh WJ. Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.* 2003 Aug;22(8):1666-1679.
5. Butina D, Segall MD, Frankcombe K. Predicting ADME properties in silico: methods and models. *Drug Discovery Today.* 2002 May 6;7(11):S83-S88.
6. Worth AP, Van Leeuwen CJ, Hartung T. The prospects for using (Q)SARs in a changing political environment--high expectations and a key role for the european commission's joint research centre. *SAR - and QSAR - in Environmental Research.* 2004;15(5):331.
7. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics.* Wiley-VCH; 2009.
8. Hall LH, Kier LB, Brown BB. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comp. Sci.* 1995 Nov 1;35(6):1074-1080.
9. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov', ev V, Hoonakker F, Tetko IV, Marcou G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Current Computer - Aided Drug Design.* 2008 Sep;4:191-198.
10. Tetko IV, Poda GI, Ostermann C, Mannhold R. Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chem. Biodivers.* 2009 Nov;6(11):1837-1844.
11. Aqueous solubility of drug-like compounds [Internet]. [cited 2010 Apr 30]; Available from: <http://deposit.ddb.de/cgi-bin/dokserv?idn=979726220>
12. Consonni V, Pavan M, Todeschini R, Mauri. Dragon software: An easy approach to molecular descriptor calculations. *56(2):237-248.*
13. Müller K, Mika S, Rättsch G, Tsuda K, Schölkopf B. An introduction to kernel-

- based learning algorithms. *IEEE TRANSACTIONS ON NEURAL NETWORKS*. 2001;12(2):181--201.
14. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. 1st ed. Cambridge University Press; 2000.
 15. Tetko IV. Associative neural network. *Methods Mol. Biol.* 2008;458:185-202.
 16. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. Cambridge University Press; 1992.
 17. Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. *PROCEEDINGS OF THE 5TH ANNUAL ACM WORKSHOP ON COMPUTATIONAL LEARNING THEORY*. 1992;:144--152.
 18. Cortes C. Support-Vector Networks. *Machine learning*. 1995;20(3):273.
 19. Drucker H, Burges CJC, Kaufman L, C CJ, Kaufman BL, Smola A, Vapnik V. Support Vector Regression Machines. 1996 [cited 2010 Oct 17]; Available from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.4845>
 20. Breiman L. Bagging predictors. *Mach. Learn.* 1996;24(2):123-140.
 21. Tetko IV, Poda GI. Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds. *J. Med. Chem.* 2004 Nov 4;47(23):5601-5604.
 22. Tetko IV, Bruneau P. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *J Pharm Sci.* 2004 Dec;93(12):3103-3110.
 23. Tetko IV, Tanchuk VY, Villa AE. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J Chem Inf Comput Sci.* 2001 Oct;41(5):1407-1421.
 24. Prasanna MD, Vondrasek J, Wlodawer A, Bhat TN. Application of InChI to curate, index, and query 3-D structures. *Proteins.* 2005 Jul 1;60(1):1-4.
 25. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y. Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem.* 2005 May 21;3(10):1832-1834.
 26. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts D, Schultz T, Stanton DW, van de Sandt JJM, Tong W, Veith G, Yang C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern Lab Anim.* 2005 Apr;33(2):155-173.
 27. Tetko IV, Bruneau P, Mewes H, Rohrer DC, Poda GI. Can we estimate the

accuracy of ADME-Tox predictions? Drug Discovery Today. 2006 Aug;11(15-16):700-707.

28. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model*. 2008 Sep;48(9):1733-1746.
29. Sushko I, Novotarskyi S, Körner R, Pandey AK, Kovalishyn VV, Prokopenko VV, Tetko IV. Applicability domain for in silico models to achieve accuracy of experimental measurements. *Journal of Chemometrics*. 2010;24(3-4):202-208.
30. Tetko IV, Poda G, Ostermann C, Mannhold R. Accurate In Silico log P Predictions: One Can't Embrace the Unembraceable. *QSAR & Combinatorial Science*. 2009;28(8):845-849.
31. Tropsha A, Gramatica P, Gombar V. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science*. 2003;22(1):77, 69.
32. Manallack DT, Tehan BG, Gancia E, Hudson BD, Ford MG, Livingstone DJ, Whitley DC, Pitt WR. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J Chem Inf Comput Sci*. 2003 Apr;43(2):674-679.
33. Tetko IV. Neural network studies. 4. Introduction to associative neural networks. *J Chem Inf Comput Sci*. 2002 Jun;42(3):717-728.
34. Tetko IV, Tanchuk VY. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J Chem Inf Comput Sci*. 2002 Oct;42(5):1136-1145.
35. Luilo GB, Cabaniss SE. Quantitative structure-property relationship for predicting chlorine demand by organic molecules. *Environ. Sci. Technol*. 2010 Apr 1;44(7):2503-2508.
36. Alvarez-Ginarte YM, Crespo-Otero R, Marrero-Ponce Y, Noheda-Marin P, Garcia de la Vega JM, Montero-Cabrera LA, Ruiz García JA, Caldera-Luzardo JA, Alvarado YJ. Chemometric and chemoinformatic analyses of anabolic and androgenic activities of testosterone and dihydrotestosterone analogues. *Bioorg. Med. Chem*. 2008 Jun 15;16(12):6448-6459.
37. Papa E, Gramatica P. Externally validated QSPR modelling of VOC tropospheric oxidation by NO₃ radicals. *SAR QSAR Environ Res*. 2008;19(7-8):655-668.
38. González-Díaz H, Vilar S, Santana L, Podda G, Uriarte E. On the applicability of QSAR for recognition of miRNA bioorganic structures at early stages of organism and cell development: embryo and stem cells. *Bioorg. Med. Chem*. 2007 Apr 1;15(7):2544-2550.
39. Gramatica P, Pilutti P, Papa E. Approaches for externally validated QSAR

- modelling of Nitrated Polycyclic Aromatic Hydrocarbon mutagenicity. SAR QSAR Environ Res. 2007 Mar;18(1-2):169-178.
40. Oberg T. A QSAR for baseline toxicity: validation, domain of application, and prediction. Chem. Res. Toxicol. 2004 Dec;17(12):1630-1637.
 41. Ames BN, Lee FD, Durston WE. An improved bacterial test system for the detection and classification of mutagens and carcinogens. Proc. Natl. Acad. Sci. U.S.A. 1973 Mar;70(3):782-786.
 42. Schultz TW, Hewitt M, Netzeva T, Cronin M. Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action. QSAR & Combinatorial Science. 2007;26(2):238-254.
 43. Schultz TW, Sinks GD, Miller LA. Population growth impairment of sulfur-containing compounds to *Tetrahymena pyriformis*. Environ. Toxicol. 2001;16(6):543-549.
 44. Seward JR, Sinks GD, Schultz TW. Reproducibility of toxicity across mode of toxic action in the *Tetrahymena* population growth impairment assay. Aquat. Toxicol. 2001 Jun;53(1):33-47.
 45. Balakin KV, Savchuk NP, Tetko IV. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions. Curr. Med. Chem. 2006;13(2):223-241.
 46. Tetko I, Livingstone D. Rule-Based Systems to Predict Lipophilicity [Internet]. In: Comprehensive Medicinal Chemistry II. Oxford: Elsevier; 2007 [cited 2010 Apr 6]. p. 649 - 668. Available from: <http://www.sciencedirect.com/science/article/B8F9N-4MWJ66G-48/2/c3032bf110110456332c0ea5a0b7d6a8>
 47. Kaiser J. SCIENCE RESOURCES: Chemists Want NIH to Curtail Database. Science. 2005 May 6;308(5723):774a.
 48. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucl. Acids Res. 2006 Jan 1;34(suppl_1):D668-672.
 49. Williams AJ. Internet-based tools for communication and collaboration in chemistry. Drug Discovery Today. 2008 Jun;13(11-12):502-506.
 50. The Chempedia project [Internet]. Available from: <http://chempedia.com/>
 51. Patiny L. Sharing Product Physical Characteristics over the Internet. Internet Journal of Chemistry. 2000;3(2).
 52. Pubchem database [Internet]. Available from: <http://pubchem.ncbi.nlm.nih.gov/>
 53. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV. Virtual

computational chemistry laboratory--design and description. *J. Comput. Aided Mol. Des.* 2005 Jun;19(6):453-463.

54. The OpenTox project [Internet]. Available from: <http://www.opentox.org/>
55. The Pubmed database [Internet]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/>
56. McNaught A. The IUPAC international chemical identifier : InChI-A new standard for molecular informatics. *Chemistry International*. 2006;28(6):12-15.
57. Varnek A, Gaudin C, Marcou G, Baskin I, Pandey AK, Tetko IV. Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients. *J Chem Inf Model*. 2009 Jan;49(1):133-144.
58. Livingstone D, Manallack D, Tetko I. Data modelling with neural networks: Advantages and limitations. *Journal of Computer-Aided Molecular Design*. 1997 Mar 1;11(2):135-142.
59. Tropsha A, Gramatica P, Gombar V. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science*. 2003;22(1):69-77.
60. Adriana.Code web-page [Internet]. Available from: <http://www.molecular-networks.com/products/adrianacode/>
61. Steinbeck C[, Hoppe C[, Kuhn S[, Floris M[, Guha R[, Willighagen EL. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design*. 2006 Jun;12:2111-2120.
62. Aires-de-Sousa J, Gasteiger J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *Journal of Chemical Information and Computer Sciences*. 2001 Mar 1;41(2):369-375.
63. Aires-de-Sousa J, Gasteiger J. Prediction of enantiomeric selectivity in chromatography. Application of conformation-dependent and conformation-independent descriptors of molecular chirality. *J. Mol. Graph. Model*. 2002 Mar;20(5):373-388.
64. Zhang Q, Aires-de-Sousa J. Physicochemical Stereodescriptors of Atomic Chiral Centers†. *Journal of Chemical Information and Modeling*. 2006 Nov 1;46(6):2278-2287.
65. Aires-de-Sousa J, Gasteiger J. Prediction of Enantiomeric Excess in a Combinatorial Library of Catalytic Enantioselective Reactions. *Journal of Combinatorial Chemistry*. 2005 Mar 1;7(2):298-301.
66. Aires F, Prigent C, Rossow WB. Neural Network Uncertainty Assessment Using Bayesian Statistics: A Remote Sensing Application. *Neural Computation*. 2004

67. Dimoglo A, Shvets N, Tetko I, Livingstone D. Electronic-Topological Investigation of the Structure - Acetylcholinesterase Inhibitor Activity Relationship in the Series of N-Benzylpiperidine Derivatives. Quantitative Structure-Activity Relationship. 2001;20(1):31-45.
68. Dimoglo A. Compositional approach to electronic structure description of chemical compounds, oriented computer analysis of structure-activity relation. *Khim. Farmaz. Zh.* 1985;4:438-444.
69. Skvortsova MI, Baskin II, Skvortsov LA, Palyulin VA, Zefirov NS, Stankevich IV. Chemical graphs and their basis invariants. *Journal of Molecular Structure: THEOCHEM.* 1999 Jun 25;466(1-3):211-217.
70. Cherkasov A, Ban F, Santos-Filho O, Thorsteinson N, Fallahi M, Hammond GL. An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. *J. Med. Chem.* 2008 Apr 10;51(7):2047-2056.
71. Stewart JJP. Optimization of parameters for semiempirical methods I. *Method. Journal of Computational Chemistry.* 1989;10(2):209-220.
72. Potemkin VA, Grishina MA. A new paradigm for pattern recognition of drugs. *J Comput Aided Mol Des.* 2008 3;22(6-7):489-505.
73. Grishina MA, Bartashevich EV, Potemkin VA, Belik AV. Genetic Algorithm for Predicting Structures and Properties of Molecular Aggregates in Organic Substances. *Journal of Structural Chemistry.* 2002 Nov 1;43(6):1040-1044.
74. Potemkin VA, Pogrebnoy AA, Grishina MA. Technique for Energy Decomposition in the Study of "Receptor-Ligand" Complexes. *Journal of Chemical Information and Modeling.* 2009 Jun 22;49(6):1389-1406.
75. Potemkin VA, Bartashevich EV, Belik AV. New approaches to prediction of thermodynamic parameters of substances using molecular data. *Rus. J. Phys. Chem.* 1996 Mar;70(3):411-416.
76. Bender A, Mussa HY, Glen RC, Reiling S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *Journal of Chemical Information and Computer Sciences.* 2004 Jan 1;44(1):170-178.
77. Zauhar RJ, Moyna G, Tian L, Li Z, Welsh WJ. Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* 2003 Dec 18;46(26):5674-5690.
78. Ertl P. Molecular structure input on the web. *J Cheminf.* 2010;2(1):1.
79. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL. The Blue Obelisk- interoperability in chemical

- informatics. *J Chem Inf Model.* 2006 Jun;46(3):991-998.
80. McCann J, Ames BN. Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals: discussion. *Proc Natl Acad Sci U S A.* 1976 Mar;73(3):950-954.
 81. Mortelmans K, Zeiger E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* 2000 Nov 20;455(1-2):29-60.
 82. Muratov E, Fourches D, Artemenko A, Kuzmin V, Zhao G, Golbraikh A, Polischuk P, Gramatica P, Martin T, Hormozdiari F, Dao P, Sahinalp C, Cherkasov A. *Combi-QSAR Modeling of Ames Mutagenicity.* Obernai, France: 2010.
 83. Cherkasov A. Inductive electronegativity scale. Iterative calculation of inductive partial charges. *J Chem Inf Comput Sci.* 2003 Dec;43(6):2039-2047.
 84. Kuz'min V, Artemenko A, Muratov E. Hierarchical QSAR technology based on the Simplex representation of molecular structure. *Journal of Computer-Aided Molecular Design.* 2008 Jun 1;22(6):403-421.
 85. Harmeling S, Dornhege G, Tax D, Meinecke F, Müller K. From outliers to prototypes: Ordering data. *Neurocomputing.* 2006 Aug;69(13-15):1608-1618.
 86. Schwaighofer A, Schroeter T, Mika S, Hansen K, Ter Laak A, Lienau P, Reichel A, Heinrich N, Müller K. A probabilistic approach to classifying metabolic stability. *J Chem Inf Model.* 2008 Apr;48(4):785-796.
 87. Montgomery D. *Introduction to linear regression analysis.* New York: Wiley; 1982.
 88. Dragos H, Gilles M, Alexandre V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *Journal of Chemical Information and Modeling.* 2009 Jul 27;49(7):1762-1776.
 89. Schölkopf B, Smola AJ. *Learning with kernels.* MIT Press; 2002.
 90. Bishop CM. Novelty Detection and Neural Network Validation. 1994 [cited 2010 Aug 26]; Available from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.9127>
 91. Fechner N, Jahn A, Hinselmann G, Zell A. Estimation of the applicability domain of kernel-based machine learning models for virtual screening. *J Cheminf.* 2010;2(1):2.
 92. Benigni R, Giuliani A. Computer-assisted analysis of interlaboratory Ames test variability. *Journal of Toxicology and Environmental Health, Part A: Current Issues.* 1988;25(1):135.
 93. Piegorsch W, Zeiger E. Measuring intra-assay agreement for the Ames Salmonella assay. *Lecture Notes in Medical Informatics.* 43:35-41.

94. Guidance on information requirements and chemical safety assessment. Chapter R.7A – Endpoint specific guidance [Internet]. 2006; Available from: http://echa.europa.eu/about/reach_en.asp
95. Schultz TW. TETRATOX: Tetrahymena Pyriformis population growth impairment endpointa surrogate for fish lethality. *Toxicology Mechanisms and Methods*. 1997;7(4):289-309.
96. Eriksson L, Umetrics AB. Multi- and megavariate data analysis. 2nd ed. Umeå Sweden;; Umetrics AB;; 2006.
97. Seward JR, Sinks GD, Schultz TW. Reproducibility of toxicity across mode of toxic action in the Tetrahymena population growth impairment assay. *Aquat. Toxicol*. 2001 Jun;53(1):33-47.
98. Lebwohl D, Canetta R. Clinical development of platinum complexes in cancer therapy: an historical perspective and an update. *Eur. J. Cancer*. 1998 Sep;34(10):1522-1534.
99. Wang D, Lippard SJ. Cellular processing of platinum anticancer drugs. *Nat Rev Drug Discov*. 2005 Apr;4(4):307-320.
100. Ghezzi A, Aceto M, Cassino C, Gabano E, Osella D. Uptake of antitumor platinum(II)-complexes by cancer cells, assayed by inductively coupled plasma mass spectrometry (ICP-MS). *J. Inorg. Biochem*. 2004 Jan;98(1):73-78.
101. Platts JA, Hibbs DE, Hambley TW, Hall MD. Calculation of the hydrophobicity of platinum drugs. *J. Med. Chem*. 2001 Feb 1;44(3):472-474.
102. Meylan WM, Howard PH. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J Pharm Sci*. 1995 Jan;84(1):83-92.
103. Kodaka M, Dohta Y, Rekonen P, Okada T, Okuno H. Physicochemical factors for cytotoxic activity in platinum dinuclear complexes with pyrimidine and imide ligands. *Biophys. Chem*. 1998 Dec 14;75(3):259-270.
104. Lewis DF, Lake BG, George SG, Dickins M, Eddershaw PJ, Tarbit MH, Beresford AP, Goldfarb PS, Guengerich FP. Molecular modelling of CYP1 family enzymes CYP1A1, CYP1A2, CYP1A6 and CYP1B1 based on sequence homology with CYP102. *Toxicology*. 1999 Nov 29;139(1-2):53-79.
105. Maggiora GM. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *Journal of Chemical Information and Modeling*. 2006 Jul 1;46(4):1535.

Appendix

Table A1. The 90% accuracy coverages of the compounds from the Ames test training and validation datasets. The coverages are shown for all the analyzed DMs.

Model name	ASNN-STD		CONS-STD		CLASS-LAG		CORREL		CONS-STD-QUAL		CONCORDANCE		CONS-STD-PROB	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
ULP_ISIDA_SQS	20%	16%	47%	39%	31%	30%	0%	0%	52%	44%	63%	57%	49%	48%
UNC_SiRMS_RF	20%	25%	53%	46%	62%	62%	3%	0%	61%	53%	65%	64%	67%	60%
UNC_Drag_RF	24%	21%	57%	48%	55%	58%	5%	0%	67%	53%	65%	58%	61%	60%
LNU_Drag_PLS	20%	25%	39%	39%	10%	37%	0%	5%	42%	44%	62%	57%	42%	48%
UNC_SiRMS.Drag_RF	21%	25%	56%	48%	59%	64%	3%	0%	67%	61%	65%	59%	64%	67%
UNC_SiRMS_SVM	9%	18%	50%	46%	0%	0%	3%	5%	52%	44%	69%	62%	50%	48%
UNC_Drag_SVM	7%	16%	46%	37%	0%	0%	0%	0%	52%	33%	63%	57%	55%	53%
UNC_SiRMS.Drag_SVM	20%	18%	52%	48%	0%	0%	2%	5%	61%	53%	64%	63%	56%	60%
PCI_SiRMS_RF	17%	23%	53%	48%	59%	62%	3%	0%	61%	53%	64%	64%	63%	64%
PCI_Drag_RF	24%	23%	56%	48%	57%	64%	3%	0%	61%	53%	65%	58%	61%	64%
PCI_SiRMS.Drag_RF	22%	23%	57%	48%	61%	62%	3%	0%	67%	61%	65%	59%	65%	60%
MSU_FRAG_LR	22%	25%	52%	39%	48%	48%	3%	5%	52%	44%	64%	63%	60%	55%
MSU_FRAG_SVM	23%	23%	53%	53%	49%	57%	3%	5%	61%	61%	64%	64%	58%	64%
EPA_2D_NN	5%	7%	41%	37%	0%	0%	3%	0%	42%	33%	61%	60%	50%	48%
EPA_2D_FDA	0%	5%	38%	37%	15%	0%	0%	0%	30%	33%	66%	60%	41%	34%
ULP_ISIDA_NB	0%	2%	46%	39%	0%	41%	0%	0%	42%	44%	56%	52%	42%	48%
ULP_ISIDA_SVM	0%	2%	48%	41%	0%	0%	0%	0%	42%	44%	61%	52%	47%	46%
ULP_ISIDA_VP	0%	2%	54%	46%	0%	0%	0%	0%	42%	44%	62%	58%	48%	53%
OCHEM_ESTATE_ANN	24%	25%	44%	37%	44%	44%	3%	5%	42%	33%	62%	62%	55%	55%
CONS_QUANT	23%	28%	60%	53%	70%	71%	10%	5%	67%	61%	65%	59%	69%	67%

Table A1 (continuation).

Model name	CONS-STD-QUAL-PROB		ASNN-STD-PROB		LEVERAGE		AD_MEAN1		AD_MEAN2		ELLIPS		SCAvg	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
ULP_ISIDA_SQS	53%	53%	36%	46%	0%	0%	2%	3%	1%	0%	1%	7%	1%	0%
UNC_SiRMS_RF	67%	67%	58%	57%	0%	0%	0%	14%	2%	0%	5%	14%	1%	5%
UNC_Drag_RF	63%	62%	54%	55%	0%	0%	0%	7%	1%	0%	7%	12%	1%	0%
LNU_Drag_PLS	48%	50%	33%	37%	0%	0%	0%	5%	0%	0%	5%	7%	0%	0%
UNC_SiRMS.Drag_RF	67%	67%	54%	60%	0%	0%	0%	14%	2%	0%	8%	16%	0%	0%
UNC_SiRMS_SVM	59%	51%	20%	21%	0%	0%	0%	5%	0%	0%	7%	7%	0%	5%
UNC_Drag_SVM	58%	55%	24%	32%	0%	0%	0%	5%	1%	0%	3%	10%	1%	0%
UNC_SiRMS.Drag_SVM	59%	59%	24%	21%	0%	0%	0%	3%	0%	0%	7%	10%	0%	5%
PCI_SiRMS_RF	65%	64%	53%	55%	0%	0%	0%	5%	1%	0%	5%	12%	1%	0%
PCI_Drag_RF	63%	67%	52%	55%	0%	0%	0%	7%	1%	0%	8%	12%	1%	0%
PCI_SiRMS.Drag_RF	65%	62%	53%	60%	0%	0%	0%	7%	1%	0%	9%	12%	0%	0%
MSU_FRAG_LR	62%	60%	50%	44%	0%	0%	0%	5%	1%	2%	8%	12%	1%	0%
MSU_FRAG_SVM	62%	67%	47%	53%	0%	0%	0%	3%	1%	0%	6%	12%	2%	0%
EPA_2D_NN	53%	52%	38%	41%	0%	0%	0%	0%	1%	0%	6%	7%	1%	0%
EPA_2D_FDA	48%	37%	7%	12%	0%	0%	0%	5%	1%	0%	6%	12%	1%	0%
ULP_ISIDA_NB	45%	51%	17%	32%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ULP_ISIDA_SVM	50%	51%	34%	30%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ULP_ISIDA_VP	51%	58%	33%	35%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
OCHEM_ESTATE_ANN	57%	57%	34%	41%	0%	0%	0%	7%	1%	2%	2%	12%	1%	0%
CONS_QUANT	69%	71%	60%	62%	0%	0%	0%	7%	5%	0%	8%	12%	4%	5%

Table A2. The rank-based correlation coefficients between the DMs. The correlation coefficients are based NOT on the absolute values of the DMs, but on the ranks, given by them to the Ames challenge compounds

	ASNN-STD	CONS-STD	CLASS-LAG	CORREL	CONS-STD-QUAL	CONCORDANCE	CONS-STD-PROB	CONS-STD-QUAL-PROB	ASNN-STD-PROB	LEVERAGE	AD_MEAN1	AD_MEAN2	ELLIPS	SCAvg
ASNN-STD	1	0.42	0.36	0.37	0.42	0.42	0.44	0.44	0.75	0.73	0.46	0.29	0.1	0.28
CONS-STD	0.42	1	0.49	0.24	0.95	0.86	0.8	0.82	0.55	0.2	0.13	0.16	0.15	0.16
CLASS-LAG	0.36	0.49	1	0.23	0.53	0.59	0.85	0.8	0.82	0.15	0.13	0.14	0.12	0.18
CORREL	0.37	0.24	0.23	1	0.22	0.23	0.27	0.25	0.34	0.36	0.4	0.32	0.11	0.5
CONS-STD-QUAL	0.42	0.95	0.53	0.22	1	0.92	0.81	0.87	0.58	0.18	0.12	0.15	0.16	0.16
CONCORDANCE	0.42	0.86	0.59	0.23	0.92	1	0.81	0.87	0.63	0.18	0.13	0.17	0.16	0.19
CONS-STD-PROB	0.44	0.8	0.85	0.27	0.81	0.81	1	0.96	0.83	0.2	0.15	0.16	0.14	0.19
CONS-STD-QUAL-F	0.44	0.82	0.8	0.25	0.87	0.87	0.96	1	0.79	0.19	0.14	0.15	0.15	0.18
ASNN-STD-PROB	0.75	0.55	0.82	0.34	0.58	0.63	0.83	0.79	1	0.47	0.32	0.23	0.12	0.25
LEVERAGE	0.73	0.2	0.15	0.36	0.18	0.18	0.2	0.19	0.47	1	0.58	0.34	0.08	0.24
AD_MEAN1	0.46	0.13	0.13	0.4	0.12	0.13	0.15	0.14	0.32	0.58	1	0.13	0.04	0.33
AD_MEAN2	0.29	0.16	0.14	0.32	0.15	0.17	0.16	0.15	0.23	0.34	0.13	1	0.1	0.46
ELLIPS	0.1	0.15	0.12	0.11	0.16	0.16	0.14	0.15	0.12	0.08	0.04	0.1	1	0.1
SCAvg	0.28	0.16	0.18	0.5	0.16	0.19	0.19	0.18	0.25	0.24	0.33	0.46	0.1	1

Table A3. T. Pyriformis toxicity study: MGD scores and probabilities calculated for all the analyzed models

LOO (training set)				5-fold cross-validation (training set)				join validation set			
DM	Score	p-value	ns	DM	Score	p-value	ns	DM	Score	p-value	ns
Model: ASNN				Model: ASNN				Model: ASNN			
STD-CONS	188	0		STD-ASNN	250	0		STD-ASNN	225	0	
STD-ASNN	236	0		STD-CONS	317	0		STD-CONS	232	0	
TANIMOTO-kNN-FR	301	0		TANIMOTO-kNN-FR	335	0		STD-kNN-MZ	240	0	
TANIMOTO-MLR-FR	304	0		STD-kNN-DR	336	0		CORREL-ASNN	243	0.01	
STD-kNN-DR	307	0		TANIMOTO-MLR-FR	338	0		EUCLID-kNN-MZ	244	0	
EUCLID-MLR-FR	312	0		EUCLID-kNN-DR	340	0		TANIMOTO-kNN-FR	244	0.01	
EUCLID-kNN-DR	317	0		STD-kNN-MZ	341	0		STD-kNN-DR	246	0	
EUCLID-kNN-MZ	322	0		EUCLID-kNN-MZ	342	0		LEVERAGE-PLS	249	0.01	
LEVERAGE-PLS	323	0.01		LEVERAGE-PLS	343	0		EUCLID-kNN-DR	250	0.01	
STD-kNN-MZ	323	0		CORREL-ASNN	347	0		TANIMOTO-MLR-FR	252	0.04	
CORREL-ASNN	331	0.01		EUCLID-MLR-FR	368	0.01		LEVERAGE-OLS	260	0.06	1
PLSEU-PLS	344	0.07	1	LEVERAGE-OLS	378	0.04		EUCLID-MLR-FR	265	0.2	1
LEVERAGE-OLS	345	0.05	1	EUCLID-kNN-FR	379	0.06	1	PLSEU-PLS	266	0.18	1
EUCLID-kNN-FR	346	0.11	1	PLSEU-PLS	380	0.03		EUCLID-kNN-FR	267	0.14	1
SGD score	358			SGD score	391			SGD score	269		
Model: kNN-Dr				Model: kNN-Dr				Model: kNN-Dr			
STD-CONS	16	0		STD-ASNN	335	0		STD-CONS	235	0	
STD-ASNN	50	0		STD-CONS	367	0		STD-ASNN	242	0	
EUCLID-MLR-FR	90	0		EUCLID-kNN-DR	396	0		STD-kNN-DR	252	0	
EUCLID-kNN-MZ	98	0		STD-kNN-DR	397	0		STD-kNN-MZ	254	0	
STD-kNN-DR	100	0		EUCLID-kNN-MZ	402	0		EUCLID-kNN-MZ	254	0	
EUCLID-kNN-DR	101	0		LEVERAGE-PLS	403	0		EUCLID-kNN-DR	256	0	
TANIMOTO-MLR-FR	104	0		TANIMOTO-kNN-FR	405	0		LEVERAGE-PLS	263	0.01	
LEVERAGE-PLS	106	0		STD-kNN-MZ	411	0		CORREL-ASNN	263	0.02	
TANIMOTO-kNN-FR	109	0		EUCLID-MLR-FR	412	0		TANIMOTO-kNN-FR	265	0.03	
CORREL-ASNN	113	0.03		TANIMOTO-MLR-FR	416	0		TANIMOTO-MLR-FR	266	0.03	
LEVERAGE-OLS	118	0.02		PLSEU-PLS	425	0		LEVERAGE-OLS	274	0.04	
EUCLID-kNN-FR	119	0.05		CORREL-ASNN	430	0		EUCLID-MLR-FR	276	0.09	1
PLSEU-PLS	121	0.06	1	LEVERAGE-OLS	432	0		PLSEU-PLS	284	0.13	1
SGD score	132			EUCLID-kNN-FR	439	0.01		EUCLID-kNN-FR	285	0.43	1
				SGD score	462			SGD score	286		
Model: kNN-Fr				Model: kNN-Fr				Model: kNN-Fr			
STD-CONS	321	0		STD-CONS	425	0		STD-CONS	345	0	
STD-ASNN	414	0		STD-ASNN	443	0		STD-ASNN	360	0	
STD-kNN-DR	434	0		EUCLID-kNN-DR	462	0		STD-kNN-DR	365	0	
STD-kNN-MZ	434	0		STD-kNN-DR	467	0		EUCLID-kNN-DR	368	0	
EUCLID-kNN-DR	436	0		LEVERAGE-PLS	475	0		TANIMOTO-kNN-FR	369	0	
LEVERAGE-PLS	436	0		EUCLID-kNN-MZ	481	0		LEVERAGE-PLS	375	0	
EUCLID-kNN-MZ	438	0		CORREL-ASNN	482	0		TANIMOTO-MLR-FR	378	0.01	
CORREL-ASNN	439	0		EUCLID-MLR-FR	489	0		EUCLID-kNN-MZ	380	0	
EUCLID-MLR-FR	447	0		PLSEU-PLS	490	0		CORREL-ASNN	395	0.01	
TANIMOTO-kNN-FR	449	0.01		STD-kNN-MZ	490	0		STD-kNN-MZ	396	0.01	
PLSEU-PLS	450	0.01		TANIMOTO-kNN-FR	500	0		LEVERAGE-OLS	397	0.09	1
TANIMOTO-MLR-FR	451	0.01		TANIMOTO-MLR-FR	503	0		PLSEU-PLS	400	0.11	1
LEVERAGE-OLS	452	0.01		LEVERAGE-OLS	505	0		EUCLID-MLR-FR	405	0.31	1
EUCLID-kNN-FR	459	0.06	1	EUCLID-kNN-FR	515	0.01		EUCLID-kNN-FR	408	0.32	1
SGD score	477			SGD score	533			SGD score	410		
Model: kNN-MZ				Model: kNN-MZ				Model: kNN-MZ			
STD-CONS	64	0		STD-ASNN	387	0		STD-CONS	266	0	
STD-ASNN	110	0		STD-CONS	415	0		STD-ASNN	278	0	
EUCLID-kNN-DR	144	0		LEVERAGE-PLS	443	0		EUCLID-kNN-DR	279	0	
EUCLID-MLR-FR	147	0		EUCLID-kNN-DR	445	0		STD-kNN-DR	279	0	
STD-kNN-DR	147	0		STD-kNN-DR	448	0		EUCLID-kNN-MZ	279	0	
EUCLID-kNN-MZ	148	0		EUCLID-kNN-MZ	453	0		STD-kNN-MZ	279	0	
LEVERAGE-PLS	149	0		TANIMOTO-kNN-FR	455	0		LEVERAGE-PLS	290	0	
TANIMOTO-kNN-FR	152	0.01		EUCLID-MLR-FR	458	0		TANIMOTO-kNN-FR	294	0.02	
CORREL-ASNN	154	0.01		TANIMOTO-MLR-FR	461	0		CORREL-ASNN	295	0.01	
TANIMOTO-MLR-FR	155	0		STD-kNN-MZ	464	0		TANIMOTO-MLR-FR	296	0.03	
STD-kNN-MZ	159	0.01		PLSEU-PLS	467	0		LEVERAGE-OLS	299	0.01	
LEVERAGE-OLS	162	0.03		LEVERAGE-OLS	471	0.01		EUCLID-MLR-FR	302	0.01	
EUCLID-kNN-FR	165	0.04		CORREL-ASNN	474	0		PLSEU-PLS	311	0.06	1
SGD score	168	0.1	1	EUCLID-kNN-FR	492	0.02		EUCLID-kNN-FR	315	0.36	1
	176			SGD score	506			SGD score	315		
Model: CODESSA				Model: CODESSA				Model: CODESSA			
STD-CONS	390	0		STD-CONS	466	0		STD-CONS	344	0	
STD-ASNN	468	0		STD-ASNN	475	0		STD-ASNN	356	0	
EUCLID-kNN-MZ	500	0		LEVERAGE-PLS	529	0		STD-kNN-DR	367	0	
EUCLID-kNN-DR	507	0		EUCLID-kNN-MZ	530	0		LEVERAGE-PLS	367	0	
LEVERAGE-PLS	507	0		STD-kNN-DR	536	0		EUCLID-kNN-DR	369	0	
CORREL-ASNN	510	0		EUCLID-kNN-DR	538	0		TANIMOTO-MLR-FR	371	0.02	
EUCLID-MLR-FR	511	0		CORREL-ASNN	540	0		TANIMOTO-kNN-FR	372	0	
STD-kNN-MZ	511	0		TANIMOTO-kNN-FR	540	0		EUCLID-kNN-MZ	374	0.02	
TANIMOTO-kNN-FR	513	0		STD-kNN-MZ	540	0		LEVERAGE-OLS	374	0.02	
STD-kNN-DR	514	0		TANIMOTO-MLR-FR	542	0		STD-kNN-MZ	374	0.01	
PLSEU-PLS	515	0.01		EUCLID-MLR-FR	543	0		EUCLID-MLR-FR	377	0.1	1
TANIMOTO-MLR-FR	515	0		EUCLID-kNN-FR	544	0		PLSEU-PLS	378	0.03	
EUCLID-kNN-FR	517	0.01		PLSEU-PLS	547	0		CORREL-ASNN	378	0.1	1
LEVERAGE-OLS	520	0.03		LEVERAGE-OLS	553	0.02		EUCLID-kNN-FR	379	0.09	1
SGD score	534			SGD score	574			SGD score	384		
Model: MLR-Fr				Model: MLR-Fr				Model: MLR-Fr			
STD-CONS	17	0		STD-ASNN	355	0		STD-CONS	302	0	
STD-ASNN	35	0.03		STD-CONS	380	0		TANIMOTO-kNN-FR	339	0	
TANIMOTO-MLR-FR	40	0.1	1	LEVERAGE-PLS	429	0		STD-ASNN	340	0	
CORREL-ASNN	41	0.18	1	STD-kNN-DR	441	0		STD-kNN-DR	344	0	
STD-kNN-MZ	41	0.11	1	STD-kNN-MZ	445	0		STD-kNN-MZ	344	0	
TANIMOTO-kNN-FR	42	0.12	1	TANIMOTO-kNN-FR	449	0		EUCLID-kNN-DR	347	0	
LEVERAGE-PLS	43	0.13	1	EUCLID-kNN-DR	450	0		LEVERAGE-PLS	349	0	
PLSEU-PLS	43	0.15	1	EUCLID-kNN-MZ	453	0		TANIMOTO-MLR-FR	356	0.01	
EUCLID-MLR-FR	43	0.25	1	TANIMOTO-MLR-FR	458	0		EUCLID-kNN-MZ	359	0	
EUCLID-kNN-MZ	45	0.4	1	EUCLID-MLR-FR	465	0		EUCLID-MLR-FR	359	0.01	
LEVERAGE-OLS	45	0.3	1	CORREL-ASNN	468	0		CORREL-ASNN	361	0.01	
EUCLID-kNN-FR	46	0.48	1	LEVERAGE-OLS	495	0.02		LEVERAGE-OLS	378	0.02	
STD-kNN-DR	46	1	1	PLSEU-PLS	500	0.01		PLSEU-PLS	395	0.16	1
EUCLID-kNN-DR	46	1	1	EUCLID-kNN-FR	504	0.02		EUCLID-kNN-FR	402	0	
SGD score	46			SGD score	528			SGD score	402		

Table A3 (continuation)

	LOO (training set)			5-fold cross-validation (training set)			join validation set				
DM	Score	p-value	ns	DM	Score	p-value	ns	DM	Score	p-value	ns
Model: OLS				Model: OLS				Model: OLS			
STD-CONS	302	0		STD-ASNN	388	0		STD-CONS	297	0	
STD-ASNN	386	0		STD-CONS	397	0		STD-ASNN	302	0	
TANIMOTO-kNN-FR	444	0		LEVERAGE-PLS	432	0		LEVERAGE-PLS	313	0	
LEVERAGE-PLS	447	0		STD-kNN-DR	443	0		EUCLID-kNN-DR	325	0	
TANIMOTO-MLR-FR	454	0		EUCLID-kNN-DR	444	0		STD-kNN-DR	326	0	
STD-kNN-DR	454	0		EUCLID-MLR-FR	450	0		EUCLID-kNN-MZ	327	0	
EUCLID-kNN-DR	457	0		TANIMOTO-kNN-FR	452	0		STD-kNN-MZ	329	0.01	
EUCLID-MLR-FR	458	0		STD-kNN-MZ	453	0		PLSEU-PLS	331	0.01	
EUCLID-kNN-MZ	463	0		EUCLID-kNN-MZ	453	0		TANIMOTO-kNN-FR	331	0.02	
PLSEU-PLS	465	0		PLSEU-PLS	455	0		TANIMOTO-MLR-FR	332	0.07	1
CORREL-ASNN	475	0		TANIMOTO-MLR-FR	456	0		CORREL-ASNN	334	0.04	
STD-kNN-MZ	480	0.01		CORREL-ASNN	463	0.02		LEVERAGE-OLS	337	0.04	
EUCLID-kNN-FR	485	0.07	1	LEVERAGE-OLS	467	0.02		EUCLID-MLR-FR	343	0.13	1
LEVERAGE-OLS	490	0.04		EUCLID-kNN-FR	471	0.05		EUCLID-kNN-FR	345	0.18	1
SGD score	504			SGD score	484			SGD score	347		
Model: PLSR				Model: PLSR				Model: PLSR			
STD-CONS	247	0		STD-ASNN	353	0		STD-CONS	259	0	
STD-ASNN	288	0		STD-CONS	370	0		STD-ASNN	262	0	
EUCLID-kNN-DR	322	0		STD-kNN-DR	392	0		STD-kNN-MZ	274	0	
STD-kNN-DR	335	0		EUCLID-kNN-DR	400	0		STD-kNN-DR	280	0	
LEVERAGE-PLS	335	0		TANIMOTO-kNN-FR	400	0		EUCLID-kNN-MZ	282	0	
EUCLID-kNN-MZ	335	0		LEVERAGE-PLS	403	0		EUCLID-kNN-DR	282	0	
STD-kNN-MZ	338	0		STD-kNN-MZ	404	0		LEVERAGE-PLS	282	0	
EUCLID-MLR-FR	340	0		EUCLID-kNN-MZ	404	0		TANIMOTO-kNN-FR	284	0	
TANIMOTO-kNN-FR	343	0		TANIMOTO-MLR-FR	408	0		TANIMOTO-MLR-FR	288	0.01	
LEVERAGE-OLS	347	0		EUCLID-MLR-FR	413	0		LEVERAGE-OLS	291	0.01	
PLSEU-PLS	347	0		CORREL-ASNN	418	0		CORREL-ASNN	300	0.01	
TANIMOTO-MLR-FR	350	0.01		LEVERAGE-OLS	423	0.02		EUCLID-MLR-FR	301	0.04	
CORREL-ASNN	355	0		EUCLID-kNN-FR	425	0.02		PLSEU-PLS	302	0.03	
EUCLID-kNN-FR	364	0.06	1	PLSEU-PLS	428	0.03		EUCLID-kNN-FR	307	0.15	1
SGD score	382			SGD score	442			SGD score	312		
Model: SVM-Dr				Model: SVM-Dr				Model: SVM-Dr			
STD-CONS	35	0		STD-ASNN	311	0		STD-CONS	284	0	
STD-ASNN	78	0		STD-CONS	338	0		STD-kNN-DR	292	0	
TANIMOTO-kNN-FR	90	0.03		STD-kNN-DR	360	0		STD-ASNN	293	0	
EUCLID-MLR-FR	91	0.03		EUCLID-kNN-DR	364	0		EUCLID-kNN-DR	296	0	
STD-kNN-MZ	92	0.02		TANIMOTO-kNN-FR	367	0		STD-kNN-MZ	300	0	
TANIMOTO-MLR-FR	94	0.07	1	TANIMOTO-MLR-FR	371	0		EUCLID-kNN-MZ	301	0	
STD-kNN-DR	94	0.04		STD-kNN-MZ	372	0		LEVERAGE-PLS	317	0	
EUCLID-kNN-MZ	95	0.12	1	EUCLID-kNN-MZ	373	0		CORREL-ASNN	341	0	
EUCLID-kNN-FR	96	0.32	1	LEVERAGE-PLS	374	0		EUCLID-MLR-FR	343	0	
EUCLID-kNN-DR	96	0.12	1	EUCLID-MLR-FR	381	0		PLSEU-PLS	348	0	
CORREL-ASNN	99	0.3	1	CORREL-ASNN	390	0		TANIMOTO-kNN-FR	351	0	
PLSEU-PLS	102	1	1	LEVERAGE-OLS	392	0		LEVERAGE-OLS	355	0	
LEVERAGE-OLS	102	1	1	PLSEU-PLS	395	0.02		TANIMOTO-MLR-FR	363	0.02	
LEVERAGE-OLS	102	1	1	EUCLID-kNN-FR	402	0.07	1	EUCLID-kNN-FR	386	0.21	1
SGD score	102			SGD score	415			SGD score	393		
Model: SVM-Fr				Model: SVM-Fr				Model: SVM-Fr			
STD-CONS	-63	0		STD-ASNN	280	0		STD-CONS	282	0	
STD-ASNN	-41	0.02		STD-CONS	316	0		STD-kNN-DR	311	0	
TANIMOTO-MLR-FR	-35	0.03		LEVERAGE-PLS	349	0		EUCLID-kNN-DR	312	0	
LEVERAGE-PLS	-25	0.09	1	EUCLID-kNN-DR	359	0		STD-kNN-MZ	313	0	
TANIMOTO-kNN-FR	-25	0.08	1	STD-kNN-DR	360	0		STD-ASNN	315	0	
CORREL-ASNN	-20	0.16	1	EUCLID-kNN-MZ	373	0		TANIMOTO-kNN-FR	316	0	
STD-kNN-MZ	-16	0.18	1	TANIMOTO-kNN-FR	376	0		LEVERAGE-PLS	324	0	
PLSEU-PLS	-15	0.21	1	TANIMOTO-MLR-FR	381	0		EUCLID-kNN-MZ	325	0	
STD-kNN-DR	-14	0.39	1	STD-kNN-MZ	385	0		TANIMOTO-MLR-FR	329	0	
EUCLID-kNN-MZ	-13	0		EUCLID-MLR-FR	390	0		CORREL-ASNN	338	0.01	
EUCLID-MLR-FR	-13	1	1	CORREL-ASNN	393	0		EUCLID-MLR-FR	341	0.01	
LEVERAGE-OLS	-13	1	1	PLSEU-PLS	397	0		LEVERAGE-OLS	354	0.02	
EUCLID-kNN-FR	-13	0.49	1	EUCLID-kNN-FR	423	0.07	1	PLSEU-PLS	368	0.11	1
EUCLID-kNN-DR	-13	0		LEVERAGE-OLS	423	0.05		EUCLID-kNN-FR	378	0.36	1
SGD score	-13			SGD score	438			SGD score	379		
Model: SVM-MZ				Model: SVM-MZ				Model: SVM-MZ			
STD-CONS	122	0		STD-ASNN	371	0		STD-CONS	272	0	
STD-ASNN	184	0		STD-CONS	380	0		STD-kNN-MZ	276	0	
TANIMOTO-MLR-FR	207	0		TANIMOTO-MLR-FR	414	0		EUCLID-kNN-MZ	280	0	
TANIMOTO-kNN-FR	213	0.02		TANIMOTO-kNN-FR	414	0		STD-ASNN	282	0	
EUCLID-kNN-DR	218	0.06	1	LEVERAGE-PLS	415	0		STD-kNN-DR	287	0	
STD-kNN-DR	219	0.03		STD-kNN-DR	423	0		LEVERAGE-PLS	288	0	
PLSEU-PLS	220	0.04		EUCLID-kNN-DR	424	0		EUCLID-kNN-DR	288	0	
EUCLID-MLR-FR	223	0		EUCLID-kNN-MZ	430	0		TANIMOTO-kNN-FR	296	0	
LEVERAGE-PLS	224	0.06	1	STD-kNN-MZ	433	0		TANIMOTO-MLR-FR	302	0.01	
EUCLID-kNN-MZ	233	0.14	1	PLSEU-PLS	438	0		LEVERAGE-OLS	306	0.01	
CORREL-ASNN	235	0.16	1	EUCLID-MLR-FR	439	0		CORREL-ASNN	311	0.01	
STD-kNN-MZ	240	0.18	1	LEVERAGE-OLS	446	0		EUCLID-MLR-FR	312	0.06	1
EUCLID-kNN-FR	241	0.37	1	CORREL-ASNN	453	0.02		PLSEU-PLS	315	0.02	
LEVERAGE-OLS	243	0.28	1	EUCLID-kNN-FR	455	0.02		EUCLID-kNN-FR	326	0.21	1
SGD score	245			SGD score	483			SGD score	343		
Model: CONS				Model: CONS				Model: CONS			
STD-CONS	33	0		STD-ASNN	247	0		STD-CONS	214	0	
STD-ASNN	87	0		STD-CONS	296	0		STD-ASNN	220	0	
TANIMOTO-kNN-FR	125	0		TANIMOTO-kNN-FR	322	0		TANIMOTO-kNN-FR	232	0	
EUCLID-MLR-FR	126	0		STD-kNN-DR	323	0		STD-kNN-MZ	235	0	
TANIMOTO-MLR-FR	126	0		LEVERAGE-PLS	326	0		TANIMOTO-MLR-FR	235	0.01	
STD-kNN-DR	132	0.01		EUCLID-kNN-DR	328	0		LEVERAGE-PLS	236	0	
EUCLID-kNN-MZ	134	0.01		TANIMOTO-MLR-FR	328	0		STD-kNN-DR	236	0	
LEVERAGE-PLS	134	0		EUCLID-MLR-FR	332	0		EUCLID-kNN-DR	237	0	
EUCLID-kNN-DR	135	0.02		EUCLID-kNN-MZ	333	0		EUCLID-kNN-MZ	239	0	
STD-kNN-MZ	137	0.02		STD-kNN-MZ	338	0		CORREL-ASNN	241	0.02	
CORREL-ASNN	137	0.04		CORREL-ASNN	346	0		LEVERAGE-OLS	249	0.04	
PLSEU-PLS	142	0.05	1	PLSEU-PLS	353	0.01		EUCLID-MLR-FR	257	0.16	1
EUCLID-kNN-FR	144	0.13	1	LEVERAGE-OLS	358	0.01		PLSEU-PLS	257	0.08	1
LEVERAGE-OLS	146	0.12	1	EUCLID-kNN-FR	362	0.03		EUCLID-kNN-FR	261	0.26	1
SGD score	151		29	SGD score	381		2	SGD score	261		12

Table A4. T. Pyriformis toxicity study: Predicting RMSE for the validation sets

DM	MGD calibrated on the 5-CV set				MGD calibrated on the 5-CV set					MGD calibrated on the validation set 1									
	set 1	set 2	set 2	set 2	Model: ASNN	scores	RMSE	RMSE pred	p	Model: ASNN	scores	RMSE	RMSE pred	p	Model: ASNN	scores	RMSE	RMSE pred	p
Model: ASNN	163	0.41	0.47	0.06	Model: ASNN	80	0.52	0.47	0.23	Model: ASNN	82	0.52	0.45	0.44	Model: ASNN	82	0.52	0.45	0.44
STD-ASNN	168	0.41	0.41	0.08	TANIMOTO-MLR-FR	81	0.52	0.46	0.37	EUCLID-kNN-DR	82	0.52	0.49	0.42	EUCLID-kNN-DR	82	0.52	0.49	0.42
EUCLID-kNN-MZ	169	0.41	0.45	0.03	STD-kNN-DR	81	0.52	0.5	0.3	STD-kNN-MZ	85	0.52	0.47	0.71	STD-kNN-MZ	85	0.52	0.47	0.71
LEVERAGE-PLS	169	0.41	0.44	0.13	EUCLID-kNN-DR	82	0.52	0.46	0.45	EUCLID-kNN-MZ	83	0.52	0.46	0.55	EUCLID-kNN-MZ	86	0.52	0.46	0.75
TANIMOTO-kNN-FR	177	0.41	0.44	0.17	TANIMOTO-kNN-FR	82	0.52	0.48	0.43	TANIMOTO-kNN-FR	82	0.52	0.48	0.43	TANIMOTO-kNN-FR	85	0.52	0.44	0.66
EUCLID-kNN-FR	177	0.41	0.45	0.32	EUCLID-kNN-FR	83	0.52	0.46	0.55	CORREL-ASNN	86	0.52	0.45	0.66	CORREL-ASNN	86	0.52	0.45	0.66
TANIMOTO-MLR-FR	177	0.41	0.4	0.3	LEVERAGE-PLS	85	0.52	0.47	0.62	STD-kNN-DR	86	0.52	0.46	0.75	STD-kNN-DR	86	0.52	0.46	0.75
EUCLID-kNN-DR	177	0.41	0.44	0.48	STD-CONS	85	0.52	0.6	0.65	TANIMOTO-MLR-FR	88	0.52	0.43	0.78	TANIMOTO-MLR-FR	88	0.52	0.43	0.78
STD-kNN-MZ	178	0.41	0.42	0.35	LEVERAGE-OLS	86	0.52	0.45	0.73	LEVERAGE-OLS	88	0.52	0.42	0.85	LEVERAGE-OLS	88	0.52	0.42	0.85
EUCLID-MLR-FR	181	0.41	0.44	0.48	PLSEU-PLS	86	0.52	0.45	0.77	LEVERAGE-PLS	88	0.52	0.44	0.79	LEVERAGE-PLS	88	0.52	0.44	0.79
STD-kNN-DR	181	0.41	0.4	0.49	EUCLID-kNN-FR	91	0.52	0.45	0.96	PLSEU-PLS	89	0.52	0.41	0.87	PLSEU-PLS	89	0.52	0.41	0.87
LEVERAGE-OLS	183	0.41	0.44	0.68	CORREL-ASNN	96	0.52	0.47	0.88	EUCLID-MLR-FR	95	0.52	0.43	0.94	EUCLID-MLR-FR	95	0.52	0.43	0.94
PLSEU-PLS	184	0.41	0.44	0.8	EUCLID-MLR-FR	100	0.52	0.45	0.84	EUCLID-kNN-FR	98	0.52	0.42	0.98	EUCLID-kNN-FR	98	0.52	0.42	0.98
STD-CONS	185	0.41	0.55	0.62	STD-ASNN	109	0.52	0.58	0.94	STD-CONS	98	0.52	0.46	0.77	STD-CONS	98	0.52	0.46	0.77
CORREL-ASNN	190	0.41	0.44	0.64	One Gauss, S(GO)	83				STD-ASNN	106	0.52	0.49	0.94	STD-ASNN	106	0.52	0.49	0.94
One Gauss, S(GO)	181				One Gauss, S(GO)	83				One Gauss, S(GO)	83				One Gauss, S(GO)	83			
Model: kNN-Dr	162	0.41	0.52	0.06	Model: kNN-Dr	87	0.57	0.51	0.14	Model: kNN-Dr	92	0.57	0.49	0.33	Model: kNN-Dr	92	0.57	0.49	0.33
STD-ASNN	165	0.41	0.45	0.06	TANIMOTO-MLR-FR	90	0.57	0.53	0.11	STD-kNN-MZ	94	0.57	0.46	0.51	STD-kNN-MZ	94	0.57	0.46	0.51
EUCLID-kNN-MZ	168	0.41	0.43	0.12	EUCLID-kNN-MZ	94	0.57	0.5	0.5	EUCLID-kNN-DR	97	0.57	0.46	0.69	EUCLID-kNN-DR	97	0.57	0.46	0.69
EUCLID-kNN-DR	174	0.41	0.48	0.22	STD-kNN-MZ	94	0.57	0.55	0.56	STD-kNN-DR	98	0.57	0.48	0.75	STD-kNN-DR	98	0.57	0.48	0.75
EUCLID-MLR-FR	175	0.41	0.44	0.27	TANIMOTO-kNN-FR	95	0.57	0.54	0.61	EUCLID-kNN-MZ	103	0.57	0.43	0.86	EUCLID-kNN-MZ	103	0.57	0.43	0.86
STD-kNN-DR	179	0.41	0.49	0.4	EUCLID-kNN-DR	96	0.57	0.49	0.61	TANIMOTO-MLR-FR	107	0.57	0.41	0.97	TANIMOTO-MLR-FR	107	0.57	0.41	0.97
TANIMOTO-kNN-FR	180	0.41	0.46	0.44	CORREL-ASNN	96	0.57	0.53	0.59	PLSEU-PLS	108	0.57	0.44	0.95	PLSEU-PLS	108	0.57	0.44	0.95
STD-kNN-MZ	182	0.41	0.49	0.51	LEVERAGE-PLS	97	0.57	0.54	0.65	TANIMOTO-kNN-FR	109	0.57	0.44	0.94	TANIMOTO-kNN-FR	109	0.57	0.44	0.94
LEVERAGE-OLS	188	0.41	0.49	0.71	STD-CONS	99	0.57	0.7	0.73	CORREL-ASNN	109	0.57	0.42	0.98	CORREL-ASNN	109	0.57	0.42	0.98
LEVERAGE-PLS	189	0.41	0.5	0.7	PLSEU-PLS	103	0.57	0.5	0.89	EUCLID-kNN-FR	109	0.57	0.42	0.98	EUCLID-kNN-FR	109	0.57	0.42	0.98
TANIMOTO-MLR-FR	189	0.41	0.49	0.83	LEVERAGE-OLS	103	0.57	0.5	0.92	LEVERAGE-PLS	110	0.57	0.43	0.97	LEVERAGE-PLS	110	0.57	0.43	0.97
EUCLID-kNN-FR	189	0.41	0.49	0.69	EUCLID-kNN-FR	104	0.57	0.51	0.99	LEVERAGE-OLS	111	0.57	0.43	0.98	LEVERAGE-OLS	111	0.57	0.43	0.98
PLSEU-PLS	196	0.41	0.48	0.98	STD-ASNN	117	0.57	0.6	0.93	STD-CONS	113	0.57	0.47	0.77	STD-CONS	113	0.57	0.47	0.77
STD-CONS	199	0.41	0.64	0.79	EUCLID-MLR-FR	125	0.57	0.51	0.83	STD-ASNN	121	0.57	0.49	0.95	STD-ASNN	121	0.57	0.49	0.95
One Gauss, S(GO)	182				One Gauss, S(GO)	94				EUCLID-MLR-FR	122	0.57	0.44	0.99	EUCLID-MLR-FR	122	0.57	0.44	0.99
One Gauss, S(GO)	182				One Gauss, S(GO)	94				One Gauss, S(GO)	94				One Gauss, S(GO)	94			
Model: kNN-Fr	255	0.56	0.57	0.01	Model: kNN-Fr	114	0.71	0.59	0.22	Model: kNN-Fr	111	0.71	0.61	0.21	Model: kNN-Fr	111	0.71	0.61	0.21
STD-ASNN	266	0.56	0.56	0.03	TANIMOTO-kNN-FR	115	0.71	0.61	0.36	STD-CONS	112	0.71	0.65	0.06	STD-CONS	112	0.71	0.65	0.06
TANIMOTO-MLR-FR	269	0.56	0.55	0.02	CORREL-ASNN	116	0.71	0.6	0.4	EUCLID-kNN-DR	113	0.71	0.65	0.2	EUCLID-kNN-DR	113	0.71	0.65	0.2
TANIMOTO-kNN-FR	278	0.56	0.55	0.15	LEVERAGE-PLS	117	0.71	0.57	0.44	EUCLID-kNN-MZ	115	0.71	0.65	0.31	EUCLID-kNN-MZ	115	0.71	0.65	0.31
LEVERAGE-PLS	279	0.56	0.73	0.38	STD-kNN-DR	117	0.71	0.63	0.41	STD-kNN-DR	116	0.71	0.58	0.42	STD-kNN-DR	116	0.71	0.58	0.42
STD-CONS	284	0.56	0.49	0.38	STD-ASNN	120	0.71	0.66	0.57	TANIMOTO-kNN-FR	118	0.71	0.67	0.49	TANIMOTO-kNN-FR	118	0.71	0.67	0.49
EUCLID-kNN-MZ	286	0.56	0.55	0.46	STD-kNN-MZ	117	0.71	0.63	0.41	STD-ASNN	119	0.71	0.61	0.74	STD-ASNN	119	0.71	0.61	0.74
LEVERAGE-OLS	287	0.56	0.53	0.45	STD-ASNN	120	0.71	0.66	0.57	STD-kNN-MZ	119	0.71	0.6	0.63	STD-kNN-MZ	119	0.71	0.6	0.63
STD-kNN-MZ	290	0.56	0.48	0.53	TANIMOTO-MLR-FR	120	0.71	0.57	0.71	LEVERAGE-PLS	119	0.71	0.6	0.63	LEVERAGE-PLS	119	0.71	0.6	0.63
EUCLID-kNN-DR	294	0.56	0.55	0.7	STD-CONS	122	0.71	0.8	0.65	TANIMOTO-MLR-FR	119	0.71	0.58	0.62	TANIMOTO-MLR-FR	119	0.71	0.58	0.62
CORREL-ASNN	298	0.56	0.55	0.93	EUCLID-kNN-DR	123	0.71	0.56	0.73	EUCLID-kNN-FR	124	0.71	0.56	0.92	EUCLID-kNN-FR	124	0.71	0.56	0.92
EUCLID-kNN-FR	305	0.56	0.49	0.86	EUCLID-kNN-FR	127	0.71	0.57	0.93	LEVERAGE-OLS	124	0.71	0.59	0.82	LEVERAGE-OLS	124	0.71	0.59	0.82
PLSEU-PLS	324	0.56	0.54	0.89	LEVERAGE-OLS	129	0.71	0.56	0.9	PLSEU-PLS	126	0.71	0.58	0.9	PLSEU-PLS	126	0.71	0.58	0.9
EUCLID-MLR-FR	346	0.56	0.53	0.7	LEVERAGE-PLS	145	0.71	0.56	0.96	CORREL-ASNN	127	0.71	0.57	0.84	CORREL-ASNN	127	0.71	0.57	0.84
One Gauss, S(GO)	287				EUCLID-MLR-FR	148	0.71	0.55	0.96	EUCLID-MLR-FR	158	0.71	0.56	0.83	EUCLID-MLR-FR	158	0.71	0.56	0.83
One Gauss, S(GO)	287				One Gauss, S(GO)	117				One Gauss, S(GO)	117				One Gauss, S(GO)	117			
Model: kNN-MZ	177	0.43	0.56	0.1	Model: kNN-MZ	95	0.63	0.55	0.02	Model: kNN-MZ	104	0.63	0.5	0.39	Model: kNN-MZ	104	0.63	0.5	0.39
STD-ASNN	183	0.43	0.47	0.16	STD-kNN-DR	99	0.63	0.54	0.11	STD-kNN-MZ	107	0.63	0.47	0.64	STD-kNN-MZ	107	0.63	0.47	0.64
EUCLID-kNN-MZ	186	0.43	0.47	0.11	EUCLID-kNN-DR	100	0.63	0.56	0.07	EUCLID-kNN-DR	108	0.63	0.48	0.65	EUCLID-kNN-DR	108	0.63	0.48	0.65
EUCLID-kNN-DR	189	0.43	0.48	0.27	TANIMOTO-MLR-FR	101	0.63	0.55	0.23	EUCLID-kNN-MZ	109	0.63	0.47	0.73	EUCLID-kNN-MZ	109	0.63	0.47	0.73
STD-kNN-DR	189	0.43	0.51	0.33	EUCLID-kNN-MZ	101	0.63	0.55	0.23	STD-kNN-DR	109	0.63	0.47	0.73	STD-kNN-DR	109	0.63	0.47	0.73
EUCLID-MLR-FR	196	0.43	0.53	0.58	STD-CONS	102	0.63	0.74	0.25	LEVERAGE-PLS	117	0.63	0.45	0.93	LEVERAGE-PLS	117	0.63	0.45	0.93
TANIMOTO-kNN-FR	196	0.43	0.53	0.58	LEVERAGE-PLS	102	0.63	0.57	0.29	STD-CONS	119	0.63	0.48	0.73	STD-CONS	119	0.63	0.48	0.73
LEVERAGE-PLS	197	0.43	0.53	0.69	STD-kNN-MZ	103	0.63	0.58	0.21	TANIMOTO-MLR-FR	119	0.63	0.45	0.94	TANIMOTO-MLR-FR	119	0.63	0.45	0.94
LEVERAGE-OLS	198	0.43	0.52	0.7	LEVERAGE-OLS	103	0.63	0.58	0.21	TANIMOTO-kNN-FR	120	0.63	0.45	0.96	TANIMOTO-kNN-FR	120	0.63	0.45	0.96
STD-kNN-MZ	200	0.43	0.52	0.76	STD-kNN-MZ	103	0.63	0.58	0.21	CORREL-ASNN	123	0.63	0.45	0.98	CORREL-ASNN	123	0.63	0.45	0.98
CORREL-ASNN	204	0.43	0.52	0.78	TANIMOTO-MLR-FR	104	0.63	0.57	0.37	LEVERAGE-OLS	124	0.63	0.45	0.99	LEVERAGE-OLS	124	0.63	0.45	0.99</

Table A4 (continuation)

DM	MGD calibrated on the 5-CV set				MGD calibrated on the 5-CV set				MGD calibrated on the validation set 1					
	set 1				set 2				set 2					
Model: OLS					Model: OLS					Model: OLS				
STD-ASNN	217	0.5	0.55	0	STD-KNN-DR	94	0.58	0.53	0.37	EUCLID-KNN-DR	93	0.58	0.56	0.09
LEVERAGE-PLS	230	0.5	0.52	0.01	TANIMOTO-MLR-FR	95	0.58	0.55	0.36	STD-KNN-DR	94	0.58	0.55	0.28
EUCLID-KNN-MZ	237	0.5	0.47	0.02	TANIMOTO-KNN-FR	96	0.58	0.55	0.46	PLSEU-PLS	95	0.58	0.51	0.45
PLSEU-PLS	239	0.5	0.5	0.1	LEVERAGE-PLS	97	0.58	0.55	0.59	TANIMOTO-MLR-FR	96	0.58	0.53	0.56
TANIMOTO-KNN-FR	244	0.5	0.51	0.27	CORREL-ASNN	97	0.58	0.54	0.62	LEVERAGE-PLS	96	0.58	0.53	0.53
EUCLID-KNN-DR	244	0.5	0.47	0.23	STD-KNN-MZ	97	0.58	0.55	0.56	CORREL-ASNN	100	0.58	0.53	0.73
LEVERAGE-OLS	244	0.5	0.51	0.2	EUCLID-KNN-DR	98	0.58	0.53	0.67	STD-ASNN	100	0.58	0.61	0.71
STD-KNN-MZ	245	0.5	0.48	0.27	PLSEU-PLS	98	0.58	0.52	0.66	EUCLID-MLR-FR	101	0.58	0.52	0.85
EUCLID-KNN-FR	245	0.5	0.51	0.15	STD-CONS	99	0.58	0.71	0.72	LEVERAGE-OLS	102	0.58	0.52	0.88
EUCLID-MLR-FR	249	0.5	0.51	0.51	EUCLID-KNN-MZ	100	0.58	0.52	0.8	TANIMOTO-KNN-FR	102	0.58	0.54	0.83
STD-KNN-DR	249	0.5	0.47	0.46	EUCLID-KNN-FR	102	0.58	0.53	0.96	EUCLID-KNN-MZ	103	0.58	0.58	0.86
TANIMOTO-MLR-FR	250	0.5	0.51	0.58	LEVERAGE-OLS	102	0.58	0.53	0.86	STD-CONS	104	0.58	0.56	0.69
CORREL-ASNN	253	0.5	0.51	0.7	STD-ASNN	104	0.58	0.64	0.85	STD-KNN-MZ	104	0.58	0.58	0.65
STD-CONS	259	0.5	0.65	0.63	EUCLID-MLR-FR	111	0.58	0.53	0.88	EUCLID-KNN-FR	107	0.58	0.52	0.99
One Gauss, S(G0)	249				One Gauss, S(G0)	96				One Gauss, S(G0)	96			
Model: PLSR					Model: PLSR					Model: PLSR				
STD-ASNN	182	0.46	0.51	0	STD-KNN-MZ	82	0.57	0.54	0	STD-KNN-MZ	86	0.57	0.55	0
TANIMOTO-KNN-FR	200	0.46	0.48	0.07	TANIMOTO-MLR-FR	87	0.57	0.52	0.13	EUCLID-KNN-DR	87	0.57	0.51	0.05
LEVERAGE-OLS	201	0.46	0.48	0.02	STD-ASNN	88	0.57	0.59	0.27	STD-ASNN	90	0.57	0.54	0.33
EUCLID-KNN-DR	201	0.46	0.43	0.03	EUCLID-KNN-DR	91	0.57	0.48	0.42	EUCLID-KNN-MZ	92	0.57	0.53	0.43
LEVERAGE-PLS	202	0.46	0.48	0.01	LEVERAGE-PLS	91	0.57	0.51	0.34	STD-CONS	92	0.57	0.5	0.44
EUCLID-KNN-MZ	207	0.46	0.44	0.14	TANIMOTO-KNN-FR	92	0.57	0.53	0.44	STD-KNN-DR	93	0.57	0.51	0.5
TANIMOTO-MLR-FR	208	0.46	0.48	0.2	EUCLID-KNN-MZ	93	0.57	0.48	0.48	PLSEU-PLS	94	0.57	0.46	0.61
EUCLID-KNN-FR	209	0.46	0.48	0.1	LEVERAGE-OLS	93	0.57	0.5	0.47	LEVERAGE-PLS	94	0.57	0.47	0.53
EUCLID-MLR-FR	210	0.46	0.47	0.31	CORREL-ASNN	95	0.57	0.5	0.66	TANIMOTO-KNN-FR	94	0.57	0.49	0.54
STD-KNN-DR	210	0.46	0.44	0.29	STD-CONS	97	0.57	0.66	0.68	LEVERAGE-OLS	94	0.57	0.48	0.61
STD-KNN-MZ	212	0.46	0.46	0.37	STD-KNN-DR	98	0.57	0.5	0.74	TANIMOTO-MLR-FR	96	0.57	0.48	0.61
CORREL-ASNN	213	0.46	0.48	0.45	EUCLID-KNN-FR	98	0.57	0.5	0.84	CORREL-ASNN	98	0.57	0.48	0.82
STD-CONS	214	0.46	0.6	0.48	PLSEU-PLS	101	0.57	0.48	0.67	EUCLID-MLR-FR	100	0.57	0.48	0.82
PLSEU-PLS	225	0.46	0.47	0.76	EUCLID-MLR-FR	106	0.57	0.49	0.81	EUCLID-KNN-FR	104	0.57	0.47	0.95
One Gauss, S(G0)	214				One Gauss, S(G0)	93				One Gauss, S(G0)	93			
Model: SVM-Dr					Model: SVM-Dr					Model: SVM-Dr				
STD-ASNN	219	0.57	0.49	0	EUCLID-KNN-MZ	94	0.61	0.47	0.11	EUCLID-KNN-DR	86	0.61	0.72	0.03
EUCLID-KNN-MZ	248	0.57	0.42	0	EUCLID-KNN-DR	96	0.61	0.47	0.23	STD-KNN-MZ	90	0.61	0.77	0.1
EUCLID-KNN-DR	253	0.57	0.41	0.01	LEVERAGE-PLS	96	0.61	0.5	0.25	EUCLID-KNN-MZ	90	0.61	0.7	0.1
LEVERAGE-PLS	260	0.57	0.46	0.02	STD-KNN-MZ	97	0.61	0.49	0.22	LEVERAGE-PLS	92	0.61	0.64	0.16
TANIMOTO-KNN-FR	264	0.57	0.46	0.1	CORREL-ASNN	99	0.61	0.49	0.35	STD-CONS	93	0.61	0.64	0.27
EUCLID-MLR-FR	280	0.57	0.45	0.17	PLSEU-PLS	101	0.61	0.47	0.48	TANIMOTO-MLR-FR	95	0.61	0.62	0.14
PLSEU-PLS	282	0.57	0.45	0.23	TANIMOTO-MLR-FR	101	0.61	0.49	0.47	PLSEU-PLS	96	0.61	0.59	0.26
TANIMOTO-MLR-FR	287	0.57	0.47	0.39	LEVERAGE-OLS	103	0.61	0.47	0.6	LEVERAGE-OLS	99	0.61	0.6	0.36
CORREL-ASNN	288	0.57	0.45	0.39	TANIMOTO-KNN-FR	106	0.61	0.49	0.71	TANIMOTO-KNN-FR	99	0.61	0.62	0.29
STD-KNN-MZ	290	0.57	0.44	0.46	STD-ASNN	106	0.61	0.58	0.63	STD-ASNN	100	0.61	0.76	0.46
EUCLID-KNN-FR	305	0.57	0.46	0.71	EUCLID-KNN-FR	112	0.61	0.47	0.9	EUCLID-KNN-FR	101	0.61	0.58	0.48
LEVERAGE-OLS	312	0.57	0.46	0.72	EUCLID-MLR-FR	128	0.61	0.47	0.96	STD-KNN-DR	103	0.61	0.72	0.55
STD-KNN-DR	344	0.57	0.41	0.86	STD-KNN-DR	128	0.61	0.48	0.85	CORREL-ASNN	105	0.61	0.68	0.69
STD-CONS	365	0.57	0.58	0.71	STD-CONS	150	0.61	0.63	0.74	EUCLID-MLR-FR	105	0.61	0.65	0.64
One Gauss, S(G0)	291				One Gauss, S(G0)	101				One Gauss, S(G0)	101			
Model: SVM-Fr					Model: SVM-Fr					Model: SVM-Fr				
LEVERAGE-PLS	224	0.51	0.48	0.1	STD-CONS	115	0.7	0.72	0.43	EUCLID-KNN-DR	111	0.7	0.61	0.16
STD-KNN-DR	225	0.51	0.42	0.1	TANIMOTO-MLR-FR	117	0.7	0.51	0.56	STD-KNN-MZ	112	0.7	0.65	0.07
EUCLID-KNN-MZ	227	0.51	0.43	0.04	STD-KNN-MZ	119	0.7	0.54	0.63	STD-CONS	118	0.7	0.56	0.55
TANIMOTO-MLR-FR	229	0.51	0.48	0.06	TANIMOTO-KNN-FR	119	0.7	0.53	0.63	TANIMOTO-MLR-FR	119	0.7	0.54	0.64
STD-ASNN	229	0.51	0.51	0.15	STD-KNN-DR	138	0.7	0.5	0.89	STD-KNN-DR	121	0.7	0.61	0.74
TANIMOTO-KNN-FR	232	0.51	0.47	0.12	LEVERAGE-OLS	139	0.7	0.49	0.96	TANIMOTO-KNN-FR	121	0.7	0.54	0.72
EUCLID-MLR-FR	236	0.51	0.47	0.2	CORREL-ASNN	141	0.7	0.52	0.96	CORREL-ASNN	128	0.7	0.58	0.85
EUCLID-KNN-DR	243	0.51	0.41	0.31	EUCLID-KNN-MZ	141	0.7	0.49	0.97	EUCLID-KNN-MZ	129	0.7	0.6	0.89
LEVERAGE-OLS	243	0.51	0.47	0.19	STD-ASNN	145	0.7	0.59	0.93	EUCLID-KNN-FR	129	0.7	0.51	0.94
PLSEU-PLS	251	0.51	0.47	0.43	EUCLID-KNN-FR	145	0.7	0.49	0.99	PLSEU-PLS	129	0.7	0.51	0.93
STD-KNN-MZ	258	0.51	0.45	0.56	PLSEU-PLS	146	0.7	0.48	0.96	LEVERAGE-OLS	132	0.7	0.52	0.91
EUCLID-KNN-FR	264	0.51	0.48	0.81	EUCLID-KNN-DR	154	0.7	0.48	0.95	EUCLID-MLR-FR	143	0.7	0.56	0.94
CORREL-ASNN	300	0.51	0.47	0.84	LEVERAGE-PLS	158	0.7	0.53	0.98	STD-ASNN	145	0.7	0.58	0.94
STD-CONS	374	0.51	0.65	0.97	EUCLID-MLR-FR	175	0.7	0.49	0.98	LEVERAGE-PLS	145	0.7	0.54	0.88
One Gauss, S(G0)	253				One Gauss, S(G0)	116				One Gauss, S(G0)	116			
Model: SVM-MZ					Model: SVM-MZ					Model: SVM-MZ				
LEVERAGE-PLS	195	0.5	0.51	0	LEVERAGE-PLS	90	0.58	0.55	0.17	STD-KNN-MZ	89	0.58	0.62	0.18
EUCLID-KNN-MZ	198	0.5	0.47	0	EUCLID-KNN-DR	91	0.58	0.54	0.18	EUCLID-KNN-DR	90	0.58	0.63	0.2
EUCLID-KNN-DR	205	0.5	0.47	0	EUCLID-KNN-MZ	92	0.58	0.52	0.26	EUCLID-KNN-MZ	93	0.58	0.6	0.38
STD-ASNN	207	0.5	0.54	0.01	LEVERAGE-OLS	92	0.58	0.53	0.24	TANIMOTO-MLR-FR	95	0.58	0.54	0.5
STD-KNN-DR	216	0.5	0.46	0.08	TANIMOTO-MLR-FR	94	0.58	0.54	0.38	LEVERAGE-PLS	95	0.58	0.56	0.5
EUCLID-MLR-FR	218	0.5	0.49	0.1	STD-CONS	95	0.58	0.77	0.55	PLSEU-PLS	98	0.58	0.49	0.79
TANIMOTO-KNN-FR	219	0.5	0.51	0.06	CORREL-ASNN	97	0.58	0.55	0.65	EUCLID-MLR-FR	98	0.58	0.53	0.73
TANIMOTO-MLR-FR	225	0.5	0.52	0.07	STD-KNN-MZ	98	0.58	0.56	0.67	STD-KNN-DR	100	0.58	0.59	0.74
CORREL-ASNN	225	0.5	0.5	0.02	EUCLID-KNN-FR	99	0.58	0.53	0.92	TANIMOTO-KNN-FR	101	0.58	0.56	0.83
LEVERAGE-OLS	228	0.5	0.51	0.04	TANIMOTO-KNN-FR	100	0.58	0.55	0.81	STD-CONS	105	0.58	0.57	0.77
EUCLID-KNN-FR	236	0.5	0.51	0.2	STD-KNN-DR	103	0.58	0.53	0.8	EUCLID-KNN-FR	109	0.58	0.53	0.98
PLSEU-PLS	238	0.5	0.51	0.4	PLSEU-PLS	116	0.58	0.5	0.98	LEVERAGE-OLS	112	0.58	0.53	0.89
STD-KNN-MZ	250	0.5	0.48	0.59	STD-ASNN	123	0.58	0.61	0.91	CORREL-ASNN	115	0.58	0.58	0.98
STD-CONS	341	0.5	0.65	0.88	EUCLID-MLR-FR	125	0.58	0.53	0.93	STD-ASNN	118	0.58	0.6	0.84
One Gauss, S(G0)	246				One Gauss, S(G0)	95				One Gauss, S(G0)	95			
Model: CONS					Model: CONS					Model: CONS				
STD-ASNN	146	0.4	0.46	0	STD-KNN-MZ	78	0.51	0.49	0.22	EUCLID-KNN-DR	78	0.51	0.45	0.19
TANIMOTO-KNN-FR	153	0.4	0.43	0.02	STD-KNN-DR	78	0.51	0.45	0.26	STD-KNN-MZ	79	0.51	0.47	0.24
EUCLID-KNN-MZ	159	0.4	0.4	0.03	TANIMOTO-MLR-FR	81	0.51	0.46	0.41	EUCLID-KNN-MZ	85	0.51	0.46	0.77
LEVERAGE-PLS	162	0.4	0.44	0.02	STD-CONS	82	0.51	0.61	0.55	STD-KNN-DR	85	0.51	0.45	0.74
EUCLID-KNN-DR	162	0.4	0.39	0.08	LEVERAGE-PLS	84	0.51	0.47	0.67	TANIMOTO-MLR-FR	86	0.51	0.43	0.74
TANIMOTO-MLR-FR	165	0.4	0.44	0.										

Figure A1. A screenshot of the browser of experimental measurements in the OCHEM system.

Revisiop.4:25 by midlighter checked in on 2010-11-22 11:14:50. Built from 146.107.60.52 on 2010-11-22 11:20:44

Firefox 3.6 on Mac - Supported

Welcome, Dear Mr.Sushkol (4) My Account Logout

A+ a-

Area of your interest: no tags selected [change]

Condition browser X Filtered records X

Compounds properties browser
Search for numerical compounds properties linked to scientific articles

1 - 5 of 2043

Records Tags

5 items on page 1 of 409 > >>

Papp(Caco-2)
 = 58.0 ± 5.0
 $10^{(-6)}$ cm/s
 Nordqvist, A.
 A General Model for Prediction of Caco-2 Cell Permeability...
 QSAR Comb. Sci. **2004**; 23 (5) 303 - 310
 Testosterone

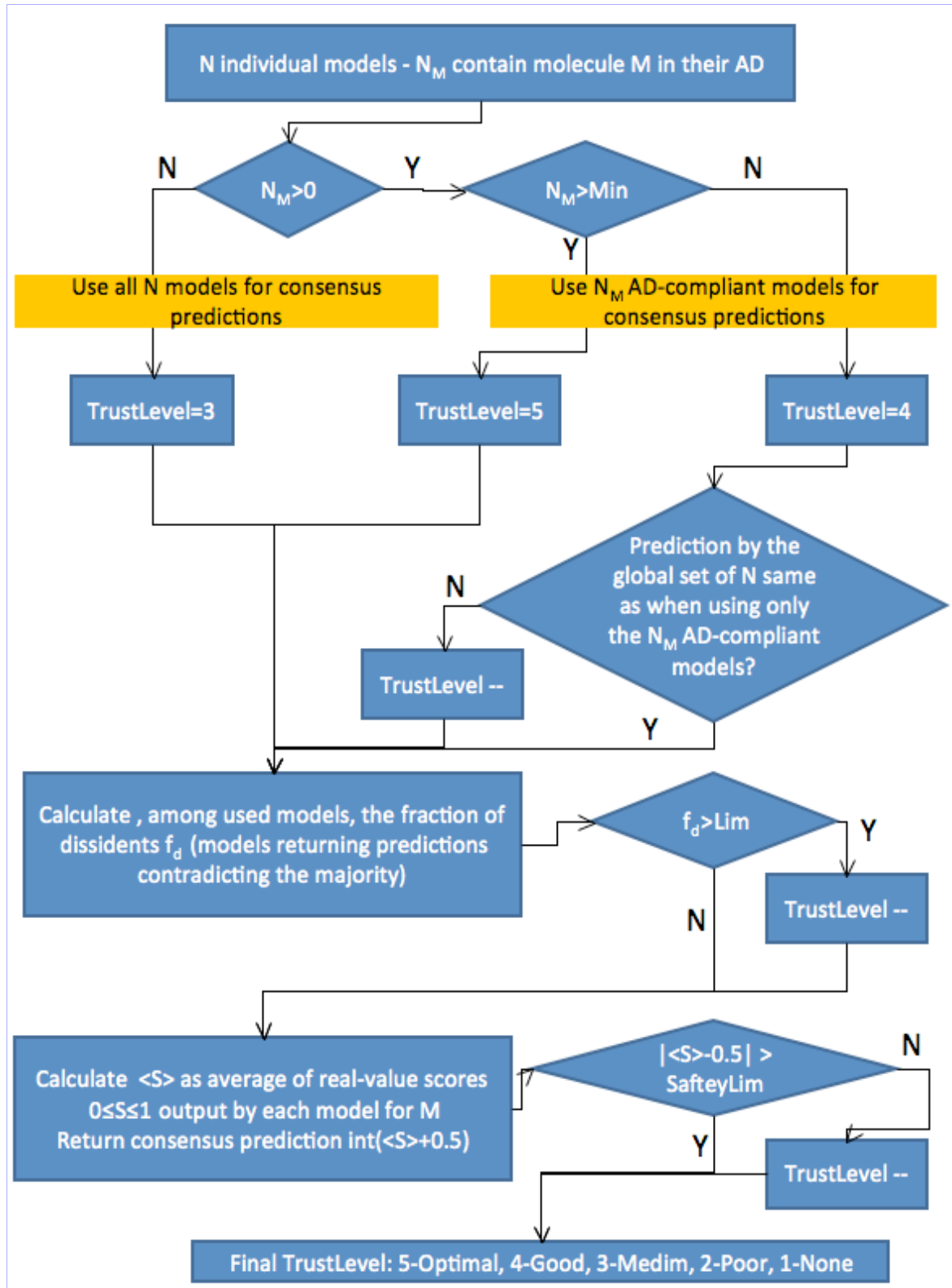
Papp(Caco-2)
 = 17.0 ± 4.0
 $10^{(-6)}$ cm/s
 Nordqvist, A.
 A General Model for Prediction of Caco-2 Cell Permeability...
 QSAR Comb. Sci. **2004**; 23 (5) 303 - 310
 Salicylic acid

Papp(Caco-2)
 = $0.28 \cdot 10^{(-6)}$
 cm/s
 Nordqvist, A.
 A General Model for Prediction of Caco-2 Cell Permeability...
 QSAR Comb. Sci. **2004**; 23 (5) 303 - 310
 Mannitol

Name / QID / InchiKey
 [search by fragment] [caddaster substructure search]
 Molecular mass between and
 MISCELLANEOUS
 Current set [?]:
 Show all
 All users
 Records by introducers:
 Original records
 Primary records
 Not validated
 Error records
 Error in chies
 Mismatching names
 Include stereochem.
 Empty molecules
 Sort by:
 Creation time Asc

Basket
 1 - 5 of 2043

Figure A2. The algorithm for calculation of the “trust score” used by the ULP group in the Ames mutagenicity challenge



Curriculum vitae

Iurii Sushko

Personal data

Date of birth	07/09/84
Nationality	Ukrainian
Marital status	Single

Education

2007 – pres.	PhD student <i>Institution:</i> Helmholtz-Zentrum, Munich, Germany <i>Topic:</i> QSAR/QSPR research, development of the methodology for assessment of applicability domain of QSAR models <i>Supervisor:</i> Prof. Mewes <i>Advisor:</i> Dr. Tetko
2001-2007	Master of Science with distinction (average mark 5.0 / 5.0) <i>Institution:</i> Institute of Applied System Analysis National technical University of Ukraine <i>Topic:</i> Applications of the superparamagnetic clustering technique <i>Supervisor:</i> Prof. Makarenko
1998-2001	High school Kyiv Polytechnic Lyceum, Faculty of Physics and Mathematics
1991-1998	Elementary school

Scientific interests

Chemoinformatics, QSAR, Computer Science, Mathematics, Physics

Computer skills

Programming	Programming languages: Java, C++, Delphi Database management: SQL, Java Hibernate, MySQL Operating systems: Linux, MacOSX, Windows
Analytical tools	R
Web development	Javascript, Ajax, HTML+CSS, jQuery

Languages

Russian, Ukrainian	native
English	fluent
German	solid basics

Publication record

Journal articles

Sushko I, Novotarskyi S, Körner R, Pandey AK, Kovalishyn VV, Prokopenko VV, Tetko IV. Applicability domain for in silico models to achieve accuracy of experimental measurements. *Journal of Chemometrics*. 2010;24(3-4):202-208.

Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A, Li J, Gramatica P, Hansen K, Schroeter T, Müller K, Xi L, Liu H, Yao X, Öberg T, Hormozdiari F, Dao P, Sahinalp C, Todeschini R, Polishchuk P, Artemenko A, Kuz'min V, Martin TM, Young DM, Fourches D, Muratov E, Tropsha A, Baskin I, Horvath D, Marcou G, Varnek A, Prokopenko VV, Tetko IV. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *Journal of Chemical Information and Modeling* [Internet]. In press. Available from: <http://pubs.acs.org/doi/abs/10.1021/ci100253r>

Tetko IV, **Sushko I**, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model*. 2008 Sep;48(9):1733-1746.

Sushko I, Pandey AK, Novotarskyi S, Körner R, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko V, Tanchuk V, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin I, Palyulin V, Radchenko E, Welsh W, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *Submitted (in revision)*

Conference talks

Sushko I, Novotarskyi S, Pandey AK, Körner R, Tetko IV. Online chemical modeling environment: models. The 238th ACS National Meeting, Washington, DC, August 16-20, 2009

Tutoring

Sushko I, Novotarskyi S. Online Chemical Modeling Environment. Environmental Chemoinformatics course. Achievements and applications of contemporary informatics, mathematics and physics (AACIMP), Kiev, 2009.

Sushko I, Novotarskyi S, Körner R. Introduction to the QSAR research using the novel chemical modeling framework. 1st Autumn School of Environmental ChemOinformatics (ECO), Munich, 18-22 October

Declaration / Erklärung

Ich erkläre an Eides statt, dass ich die der Fakultät für Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Promotionsprüfung vorgelegte Arbeit mit dem Titel:

Applicability Domain of QSAR models

im Institut für Bioinformatik und System Biologie des Helmholtz Zentrums München

unter der Anleitung und Betreuung durch

Prof. Dr. Hans-Werner Mewes

ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß § 6 Abs. 5 angegebenen Hilfsmittel benutzt habe.

Ich habe die Dissertation in dieser oder ähnlicher Form in keinem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt.

Ich habe den angestrebten Doktorgrad noch nicht erworben und bin nicht in einem früheren Promotionsverfahren für den angestrebten Doktorgrad endgültig gescheitert.

Die Promotionsordnung der Technischen Universität München ist mir bekannt.

München, den

.....

Iurii Sushko