

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

QSAR approaches to predict human
cytochrome P450 inhibition

Sergii Novotarskyi

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. Rainer U. Meckenstock

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Hans-Werner Mewes
2. Univ.-Prof. Dr. Alexandre Varnek, Ph. D.
(Université de Strasbourg / Frankreich)

Die Dissertation wurde am 12.06.2013 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 09.09.2013 angenommen.

Abstract

In the recent decades, the quantitative structure-activity relationship (QSAR) approach to modeling chemical and biological properties of small molecules has gained considerable popularity. The benefits of the QSAR approach are low costs and high productivity levels of modeling of large chemical libraries and possibility to assess properties of non-existing and non-synthesized compounds. These benefits are in high demand in the area of drug design and discovery. According to numerous studies the main reason for failure is poor pharmacokinetical and toxicity properties. Therefore, it is vital for drug discovery process success to determine compounds with unacceptable ADME/T profiles as early as possible in the drug discovery pipeline.

The prediction of metabolism of molecules is of great interest for drug discovery. Cytochromes P450 (CYP) are a superfamily of enzymes, involved in metabolism of a large number of xenobiotic compounds. Approximately 75% of currently used drugs are cleared through metabolism and eight CYP forms in human liver carry out virtually the whole CYP-mediated metabolism. This makes CYP enzymes a primary target for early stage drug design screenings and introduces high demand on high-quality QSAR models for CYP inhibition. High promiscuity with regards to substrates, high flexibility and clinically significant genetic polymorphism of the CYP enzymes makes QSAR modeling of CYP inhibition a challenging task.

This thesis focuses on several aspects of QSAR modeling of human cytochrome P450 inhibition and suggests the methodology to increase the quality of CYP inhibition models. The validity of the methodology is demonstrated in comprehensive QSAR modeling of five most important CYP isoforms - CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4. It is shown that the addition of newly developed descriptors derived from docking simulations increases the predictive ability of the resulting models. It is also shown that using these descriptors in a modified QSAR modeling workflow allows to extrapolate modeling results across closely related cytochromes. This methodology allows to predict drug activity against mutated versions of genetically polymorphic cytochromes.

The studies were performed on the OCHEM platform (<http://ochem.eu>) and all the descriptors, datasets and models are publicly available to the scientific community.

Zusammenfassung

In den vergangenen Jahrzehnten hat der Ansatz der Quantitative Struktur-Wirkungs-Beziehung (QSAR: Quantitative Structure-Activity Relationship) zur computergestützten Vorhersage chemischer und biologischer Eigenschaften von kleinen Molekülen beträchtlich an Popularität gewonnen. Die Vorteile des QSAR Ansatzes sind zum einen die niedrigen Kosten und zum anderen die hohe Effektivität der Bearbeitung großer chemischer Bibliotheken. Hinzu kommt die Möglichkeit, die Eigenschaften von bisher nicht existierenden und nicht synthetisierten Verbindungen zu bewerten. Im Bereich der Wirkstoffforschung und -entwicklung besteht eine hohe Nachfrage nach ebendiesen Vorteilen. Wie in vielen Studien bestätigt, ist der Hauptgrund für das Scheitern eines Wirkstoffkandidaten seine schlechten pharmakokinetischen bzw. toxikologischen Eigenschaften. Daher ist es von entscheidender Bedeutung für eine erfolgreiche Wirkstoffentwicklung, Verbindungen mit inakzeptablem ADME/T-Profil so früh wie möglich in der Pipeline der Wirkstoffforschung herauszufiltern.

Die Vorhersage der Metabolisierung von Molekülen ist von großem Interesse für die Wirkstoffforschung. Eine wichtige Superfamilie der Enzyme sind die Cytochrome P450 (CYP), welche an der Verstoffwechslung einer Vielzahl von Xenobiotika beteiligt sind. Etwa 75% der derzeit verwendeten Medikamente werden durch Stoffwechselvorgänge abgebaut, wobei davon nahezu der gesamte CYP abhängige Metabolismus in der menschlichen Leber durch acht verschiedenen CYP-Formen erfolgt. Dies macht die CYP-Enzyme zu einem Primärziel des Frühphasen-Screenings in der Wirkstoffentwicklung und stellt somit einen hohen Anspruch an hochqualitative QSAR Modelle zur Vorhersage der CYP-Hemmung. Unterschiedlichste Substrate, hohe Flexibilität und klinisch signifikante, genetische Polymorphismen der CYP-Enzyme machen die Entwicklung von QSAR Modellen zur Vorhersage der CYP-Hemmung zu einer anspruchsvollen Aufgabe.

Der Fokus dieser Arbeit liegt auf den unterschiedlichen Aspekten der QSAR-Modellierung humaner Cytochrom-P450-Hemmung und schlägt eine neue Methodik vor, um die Qualität der Modelle zur Vorhersage der CYP-Hemmung zu verbessern. Die Methodik wird durch umfassende Modellierung der fünf wichtigsten CYP-Isoformen validiert, CYP1A2, CYP2C9, CYP2C19, CYP2D6 und CYP3A4. Des Weiteren wird gezeigt, dass die Hinzunahme von neu entwickelten, aus Docking-Berechnungen abgeleiteten, Deskriptoren die Vorhersagekraft der resultierenden Modelle erhöht. Ferner wird gezeigt, dass durch Verwendung dieser Deskriptoren in einem erweiterten QSAR Modellierungsansatz die Modellierungsergebnisse zwischen eng verwandten Cytochromen extrapoliert werden können. Durch diese Methode wird es möglich, eine potentielle Wirkung, auch gegenüber mutierten Versionen, genetisch polymorpher Cytochrome vorherzusagen.

Alle Studien wurden auf der Plattform OCHEM (<http://ochem.eu>) durchgeführt und alle Deskriptoren, Datensätze und Modelle sind für die wissenschaftliche Gemeinschaft öffentlich zugänglich.

Acknowledgments

I would like to start by expressing my thanks to all the members of the Institute of Bioinformatics and Systems Biology in Helmholtz Zentrum München for providing the great scientific environment and all the necessary infrastructure to make the research as comfortable as it possibly could be.

I would like to thank my doctoral supervisor Prof. Dr. Hans-Werner Mewes for his support and feedback, which made this PhD project possible. The advices of Prof. Mewes allowed me to see my work as a part of a bigger picture, inspired me to research the areas I would have otherwise missed, and laid the foundation for my further studies.

My particular thanks go to my thesis advisor and group leader Dr. Igor Tetko for helping me choose the topic for my doctoral work and enabling me to develop a strong understanding of the subject. Dr. Tetko provided me with important advice and critique, through which I have learned the best practices in QSAR research, and his help in facilitating my communications with some of the most outstanding scientists in the field is truly invaluable. I am grateful for the opportunity to work on my PhD project in his group.

I would like to express my great appreciation to my colleagues Iurii Sushko, Robert Körner and Anil Pandey for their daily help, suggestions, advices and feedback. I would also like to express my gratitude to present and former members of the Chemoinformatics group as well as numerous visiting students from other scientific groups for providing a great working environment and contributing their insights to the results presented in this work: Stefan Brandmaier, Ahmed Abdelaziz, Wolfram Teetz, Eva Schlosser, Matthias Rupp and many others.

In conclusion I want to give my warmest and deeply-felt thanks to my mom, my dad and my brother for their constant moral support. Thank you.

Sergii Novotarskyi

Table of Contents

1 Introduction.....	1
1.1 Quantitative structure-activity relationship.....	1
1.2 Cytochromes P450.....	2
1.3 CYP polymorphism.....	5
1.4 Motivation.....	6
2 General methodology.....	7
2.1 Molecule representation in QSAR.....	8
2.1.1 Small molecules.....	8
2.1.2 Protein structures and the Protein Data Bank.....	9
2.2 Molecule preprocessing, conformation sampling and optimization.....	11
2.2.1 Molecule preprocessing.....	11
2.2.2 Molecule conformation sampling and optimization.....	12
2.3 Molecular docking.....	14
2.3.1 Short classification of available methods.....	15
2.3.2 AutoDock Vina.....	17
2.4 Molecular descriptors.....	18
2.4.1 General purpose descriptors.....	20
2.4.2 Chemogenomics based descriptors.....	20
2.5 Machine learning methods.....	23
2.5.1 K-nearest Neighbors.....	24
2.5.2 Artificial Neural Networks.....	25
2.5.3 Support Vector Machines.....	26
2.5.4 Random Tree / Random Forest.....	26
2.5.5 C4.5 Decision Tree.....	27
2.5.6 Bootstrap aggregating (bagging).....	28
2.5.7 Local corrections and the LIBRARY approach.....	28
2.6 Model performance evaluation.....	29
2.6.1 Sensitivity and specificity.....	30
2.6.2 Accuracy.....	30
2.6.3 Balanced accuracy.....	30
2.6.4 Matthews correlation coefficient.....	31
2.7 Model validation.....	31
2.7.1 Test set validation.....	31
2.7.2 Cross-validation.....	32
2.7.3 Bagging validation.....	32
2.7.4 General considerations.....	32

2.8 Model comparison.....	33
2.9 Applicability domain methods.....	34
2.9.1 General concepts.....	34
2.9.2 Prediction-based DM measure for classification tasks.....	35
2.9.3 Analysis of model performance with applicability domain approach..	37
2.10 Summary.....	40
3 OCHEM – The database of experimental measurements and modeling environment.....	41
3.1 Motivation.....	41
3.2 Database of experimental properties.....	43
3.2.1 Structure overview.....	43
3.2.2 Data search and editing.....	44
3.2.3 Data introduction.....	46
3.2.4 Typical OCHEM usage scenario.....	47
3.3 Modeling framework.....	48
3.3.1 Overview.....	48
3.3.2 Dataset, machine learning method and validation method selection.	49
3.3.3 Data preprocessing.....	51
3.3.4 Molecular descriptors.....	51
3.3.5 Descriptors filtering.....	52
3.3.6 Machine learning method configuration.....	53
3.3.7 Model calculation start.....	54
3.3.8 Tasks management and load distribution.....	55
3.3.9 Model analysis.....	56
3.3.10 Additional model assessment tools.....	57
3.3.11 Applicability domain assessment.....	58
3.3.12 Model application.....	59
3.4 Implementation notes.....	60
3.5 Summary.....	60
4 QSAR studies of CYP inhibition.....	63
4.1 Datasets overview and analysis.....	63
4.1.1 Datasets description.....	63
4.1.2 Preliminary analysis of datasets.....	67
4.1.3 Fragment analysis.....	69
4.1.4 Summary.....	72
4.2 Benchmarking of QSAR models for CYP1A2 inhibitor classification.....	73
4.2.1 Materials and methods.....	73
4.2.2 Modeling results.....	74
4.2.3 PCA Plot model comparison.....	77
4.2.4 Applicability domain of models.....	79

4.2.5 External test set results.....	80
4.2.6 Summary.....	83
4.3 Using novel descriptors in QSAR modeling of CYP 1A2, 2C9, 2C19, 2D6 and 3A4...	84
4.3.1 Materials and methods.....	84
4.3.2 Modeling results.....	85
4.3.3 PCA plot model comparison.....	86
4.3.4 Applicability domain of models.....	87
4.3.5 Application of models to the external test sets.....	89
4.3.6 Summary.....	94
4.4 Novel descriptors in predicting CYP2C19 activity based on CYP2C9 dataset.....	95
4.4.1 Materials and methods.....	95
4.4.2 Modeling results.....	98
4.4.3 Applicability domain analysis.....	100
4.4.4 Fragment-based interpretation.....	108
4.4.5 Summary.....	111
5 Conclusions and outlook.....	113
References.....	117
Appendix.....	131
Publication record.....	139
Curriculum vitae.....	141

1 Introduction

In this chapter we introduce the QSAR field of knowledge and give basic definitions used in this thesis. We also introduce the cytochrome P450 superfamily and describe the challenges that lie in QSAR prediction of their inhibition activity. Finally, we provide the motivation behind this work and a brief overview of accomplishments achieved in this study.

1.1 Quantitative structure-activity relationship

The idea that the physiological effects of a substance depends on its chemical composition and structure was first formulated more than a hundred years ago [1]. However, quantitative estimates of such a relationship could be determined only at the beginning of the 20th century. It was first established that for a certain group of organic compounds, there is a connection between the sedative effect (narcotic/depressant action) and the oil / water partition coefficients of these compounds [2]. Later the mathematical proof was provided of the correlation of depressant action with the relative saturation of volatile compounds in the vehicle in which they were administered [3]. In biochemical studies the first work in this area was a work by Hansch and Muir, which studied the structure-activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity [4]. The results of these studies formed the basis of the mechanistic approach to quantitative structure-activity relationship (QSAR) model construction.

Today this approach is widely used in biochemical, pharmaceutical and other fields of science where predicting properties of chemical compounds is necessary. The popularity of this approach is based on the now obvious statement that the biological or physicochemical activity of the compound is a function of its structure, represented by a set of directly measurable or computable parameters [5]. However, establishing this functional relationship is a very time-consuming and non-trivial task, the successful outcome of which is dependent on the progress in the following fields:

- Preparation and availability of large data sets of experimental measurements of physicochemical and biological properties of chemical compounds. At the same time specialization and annotation of these data sets is important, since for each particular case study it is often required to consider only a subset of chemical space.
- Development of new tools and methods to describe a molecule by a set of measurable or computable parameters. New methods that would include additional information that was discarded in previous studies in the modeling process are important.
- Since no statistical model can be equally predictive on a whole chemical space, it is important to develop methods to estimate accuracy of predictions for a given model and given chemical compound.

- The creation of new computing technologies and computational tools that provide opportunities for QSAR analysis in a reasonable timeframe.
- Development of necessary methodology to ensure effective use of known mathematical methods to achieve the purposes of QSAR modeling. This includes development of new mathematical statistics and machine learning methods. For qualitative predictions of the presence of a particular type of activity the development of classification methods is important. If the quantitative prediction of properties of chemical compounds is necessary, the development and use of regression methods is important.

The benefits of QSAR modeling reflect the benefits of computational modeling in general and include:

- Low costs and high productivity levels of modeling of large chemical libraries
- Environmentally-friendly research, reduction in necessary chemical experiments and animal testing
- Possibility to assess properties of non-existing and non-synthesized compounds
- Requires minimal tools, staff and infrastructure

These benefits are in high demand in the area of drug design and discovery. A drug discovery is a process of narrowing down from millions of synthesizable compounds to a single drug. The average time to discover and get a drug to the market is 10 – 12 years. According to numerous reports [6–9] a large fraction of drug candidates fail on different steps of drug discovery pipeline. The most expensive fails are fails that are late in the pipeline: preclinical and clinical trials. According to studies [10–12] the main reason for failure is poor pharmacokinetic and toxicity properties – ADME/T (absorption, distribution, metabolism, excretion, toxicity). Therefore, it is vital for drug discovery process success to determine compounds with unacceptable ADME/T profiles as early as possible in the drug discovery pipeline. That's why QSAR, a methodology that doesn't require measurements or even synthesis of the tested compounds, is important in the early stages of the drug discovery process.

1.2 Cytochromes P450

The prediction of metabolism of molecules is of great interest for drug discovery. Cytochromes P450 (CYP) are a superfamily of enzymes, involved in metabolism of a large number of xenobiotic compounds [13]. CYP are involved in metabolism of a large amount of drugs currently present on the market [14]. Individual CYP enzymes in families 1, 2 and 3 metabolize xenobiotics, including the majority of small molecule drugs currently in use [15]. The distinctive feature of CYP enzymes is broad and overlapping substrate specificity [16]. Approximately 75% of currently used drugs are cleared through metabolism and eight CYP forms in human liver carry out virtually the whole CYP-mediated metabolism (Figure 1.1). It is worth noting that most drugs that are cleared by the CYP system are metabolized through several CYP forms. As a general rule, drugs that are metabolized by a single CYP form are more susceptible to drug interactions than drugs metabolized by multiple forms.

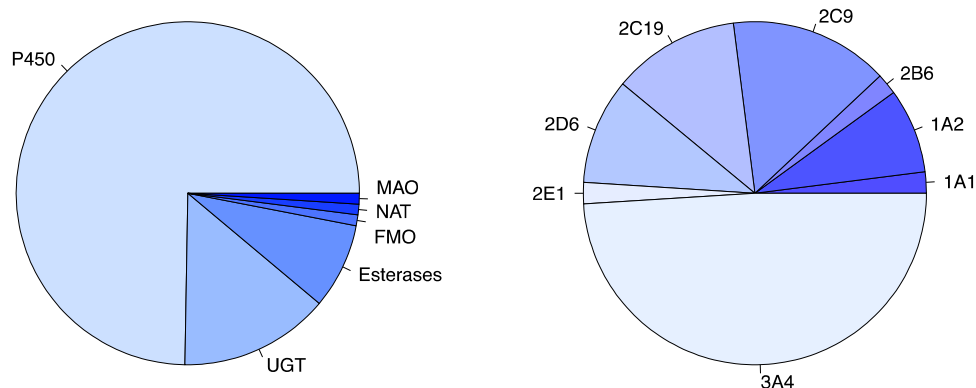


Figure 1.1. Percentage of currently marketed drugs metabolized by different human enzymes (left); percent of CYP-metabolized drugs by specific CYP isoforms (right) [17]

The promiscuity with respect to substrates makes the CYP prone to inhibition by a large amount of drugs, which may lead to clinically significant drug-drug interactions [18,19]. Similarly to a large number of other proteins, CYP enzymes are prone to both competitive and noncompetitive inhibition. In competitive inhibition, there is a competition between the substrate and inhibitor to bind to the same position on the active site of the enzyme. In the noncompetitive mode of inhibition, the active binding site of the substrate and inhibitor is different from each other. In the case of noncompetitive inhibition, the inhibitor binds to the enzyme–substrate complex, but not to the free enzyme entity. In practice, mixed-type inhibition displaying elements of both competitive and noncompetitive inhibition are frequently observed for CYP enzymes.

CYP inhibition can lead to decreased elimination of compounds dependent on metabolism for systemic clearance. If a drug is metabolized mainly via a single pathway, CYP inhibition may result in an increased steady-state concentration and accumulation ratio and non-linear kinetics as a consequence of the saturation of enzymatic processes. Especially with pro-drugs, inhibition may result in a decrease in the amount of the active drug form. Thus, inhibition of CYP may lead to toxicity or lack of efficacy of drugs [15]. Therefore, early prediction of CYP-related activity of compounds may help to avoid the pursuit of drug candidates with these undesirable effects. The metabolism of carcinogens, pro-carcinogens, and chemotherapeutics by CYP enzymes gives them an indisputable role in the cancer prevention and treatment strategies and a large number of studies research CYP inhibition for prevention and treatment of cancer [20–22]. This dictates a high interest in QSAR and computational chemistry methods of CYP inhibition prediction [23–25].

In this thesis the research is focused on five most involved isoforms: CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4.

CYP1A2 is a major enzyme in the metabolism of a number of important chemicals, which typically belong structurally to the group of planar polyaromatic amides and amines [26]. It accounts for 15% of total CYP contents in human liver and is responsible for the metabolism of approximately 10% of therapeutically used drugs [15,27,28]. Amitriptyline, ethoxyresorufin, caffeine, fluvoxamine, phenacetin, theophylline, clozapine, melatonin,

haloperidol, zolmitriptan and tizanidine are biotransformed predominantly by CYP1A2 [29]. CYP1A2 participation in xenobiotics metabolism and corresponding implications for drug development is an intensively studied topic in medicinal chemistry [30].

Multiple studies to a different degree of success performed QSAR modeling of CYP1A2 inhibition. Most of the studies focus on QSAR modeling of small classes of closely related compounds and pursue the goal of determining structural features of molecules responsible for inhibition [31–35]. Some studies also perform QSAR modeling on large sets of heterogeneous compounds [36–40].

CYP2C9 is expressed in the human liver to an extent of 15-20% of the total amount of CYP enzymes and is responsible for metabolism of around 15% of currently marketed drugs [17]. CYP2C9 is involved in the metabolism of drugs especially many of the commonly used polar acidic drugs in humans. CYP2C9 is competitively inhibited by non-steroidal anti-inflammatory drugs. Such drugs as diclofenac, ibuprofen, tolbutamide, glyburide, amitriptyline, tamoxifen and S-warfarin are predominantly metabolized by CYP2C9 [29]. Therefore, the need to evaluate drugs by QSAR for their ability to interact with CYP2C9 in their early stages of development is thought to be critical, since the chance of drug-drug interactions in a large fraction of patients is very high.

To date there are several QSAR studies on small groups of closely related compounds that successfully predict CYP inhibition concentration [41–46]. Available studies on large heterogeneous sets of compounds focus on inhibitor/non-inhibitor type classification tasks [40].

CYP2C19 is involved in metabolism of around 10% of the marketed drugs. Main known CYP2C19 substrates are some proton pump inhibitors (lansoprazole, omeprazole, pantoprazole), anti-epileptics (diazepam, phenytoin, phenobarbitone) and some other drugs (amitriptyline, clomipramine, primidone, R-warfarin). Among known CYP2C19 inhibitors are chloramphenicol, fluvoxamine, modafinil and topiramate [29]. Numerous QSAR studies on datasets of variable sizes were performed on CYP2C19 inhibition activity prediction [47–50].

CYP2D6 holds a 10% share of marketed drug metabolism and is involved in biotransformation of beta blockers (carvedilol, S-metoprolol, propafenone, timolol), antidepressants (amitriptyline, clomipramine, desipramine) and antipsychotics (haloperidol, perphenazine, thioridazine and zuclopenthixol) [29]. There are several successful QSAR studies dedicated to prediction of CYP2D6 inhibition and substrate activity [50–54].

CYP3A4 is involved in the largest fraction of drug metabolism (around 50% of marketed drugs are metabolized by this isoform) [17]. CYP3A4 substrates span across multiple chemical classes and include macrolide antibiotics (clarithromycin, erythromycin, telithromycin), anti-arrhythmics(quinidine), benzodiazepines (alprazolam, diazepam, triazolam), immune modulators (cyclosporine, tacrolimus), HIV Antivirals (indinavir, nelfinavir, ritonavir), antihistamines (astemizole, chlorpheniramine, terfenadine) and multiple other classes of drugs [29]. Several QSAR studies have been performed to predict CYP3A4 inhibition and substrate activity [40,55–59].

As several reviews [60,61] stress, another important CYP-related task is predicting the CYP isoform primarily responsible for the clearance of a particular small molecules. Several computational models are developed to address this task [62,63].

1.3 CYP polymorphism

The human CYP genes are highly polymorphic. Phenotypically, a specific population are composed of ultrarapid metabolizers (UMs), extensive metabolizers (EMs), intermediate metabolizers (IMs), and poor metabolizers (PMs). The distribution of the genetic variations and the phenotypes is ethnicity dependent [64]. The PM phenotype is due to the presence of two nonfunctional (null) alleles or deletion of entire gene, while the EM phenotype is due to one or two alleles with normal function. An IM phenotype is usually found in individuals carrying one null allele and another allele with reduced function, while UMs often carry more than one extra functional gene.

Genetic polymorphisms within CYPs mainly affect the metabolism of drugs that are substrates for those particular enzymes, probably leading to differences in drug response, in addition to an altered risk for adverse drug reactions [65,66]. Allelic variants resulting in altered protein expression or activity have significant effects on the disposition of drugs and may cause diseases as a phenotype. Genetic polymorphism is defined as a stable variation in a given locus of the genetic sequence, which is detected in 1% or more of a specific population. The most common genetic mutation in human CYP genes is single-nucleotide polymorphisms (SNPs), and nonsynonymous SNPs are functionally important SNPs, since they occur in a coding region and cause an amino-acid change in the corresponding CYP [64]. The functional CYP polymorphisms consist of gene deletions, gene duplications, and deleterious mutations creating inactive gene products, e.g., small insertions and deletions causing frame shift mutations, etc. Furthermore, amino acid changes might be introduced which, in some cases, can change the substrate specificity. An important aspect of the CYP polymorphism is the copy number variation where multiple functional gene copies of one allele can result in increased drug metabolism and absence of drug response at ordinary dosage. It was found that each human CYP gene contains a mean of 14.6 nonsynonymous SNPs and many of them are associated with altered drug metabolism or susceptibility to certain diseases [67].

In 1969 the first direct evidence from a twin study was provided that the metabolic clearance of nortriptyline was influenced by genetic factors [68]. It was later discovered that the metabolism of debrisoquine and sparteine, respectively, is polymorphic, and it was later shown that these drugs are metabolized by a common enzyme (i.e., CYP2D6 whose activity is determined by genetic trait)

The different alleles are summarized at the Human CYP allele nomenclature committee home page [69]. The page currently encompasses alleles for the CYP1A1, CYP1A2, CYP1B1, CYP2A6, CYP2A13, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, CYP2F1, CYP2J2, CYP2R1, CYP2S1, CYP3A4, CYP3A5, CYP3A7, CYP3A43, CYP4A11, CYP4A22, CYP4B1, CYP5A1, CYP8A1, CYP19A1, CYP21A2 and CYP26A1 genes. The database at present contains more than 350 functionally different CYP alleles, i.e., gene variants that affect the function and/or activity of the gene products.

The CYPs that are highly involved in drug metabolism have a large number of existing alleles. The most diverse is the CYP2D6 cytochrome with over one hundred registered alleles. CYP2C9 has over 30 registered variations, CYP2C9 – around 30 registered variations. CYP3A4 and CYP1A2 have around 20 registered variations [65].

It is estimated that the genetic variability of the CYP2C9, CYP2C19 and CYP2D6 genes can be estimated to significantly influence about 20–25% of drug treatment to such a large extent that they are of clinical importance for the outcome of drug therapy. The polymorphism of the different CYPs translates into interindividual variability to different extents depending on the enzyme in question and the impact of the allelic variant. Among the particularly important treatment regimens affected by these polymorphisms are therapies with several antidepressants, antipsychotics, antiulcer drugs, anti-HIV drugs, anticoagulants, antidiabetics and the anticancer drug tamoxifen [65].

Therefore, research in the field of QSAR modeling of activities of mutated CYP isoforms with minimal additional experimental data is important for early stage drug discovery purposes and personalized medicine approaches.

1.4 Motivation

This thesis focuses on several aspects of QSAR modeling of human cytochrome P450 inhibition.

While multiple studies were performed in the area of QSAR predictions of CYP inhibition, these studies were limited with respect to the number of applied machine learning methods and diversity of descriptors as well as the lack of a common approach to model evaluation and estimation of confidence of predictions on external test sets.

The goal of this work is to study how the accuracy of prediction of CYP inhibitors depends on the different machine learning methods and descriptor sets and to find the combination of descriptors and machine learning, ensemble and meta-learning methods that would yield the highest predictivity.

Another goal of this work is to assess expediency of introducing protein structure information (in the form of novel docking-derived descriptors) with respect to QSAR model quality and predictivity.

Lastly, this study focuses on the methodological aspects of predicting activities of mutated CYP isoforms. The combination of cytochrome structural information and a modification in traditional QSAR modeling workflow is studied with respect to extrapolating CYP inhibition prediction to structurally similar CYP isoforms.

All QSAR studies in this thesis are complemented with a fragment-based analysis to provide a mechanistic explanation of results. Applicability domain approaches are extensively used to analyze the practical usability of the obtained models both for CYP inhibition prediction and for experiment planning in the field of CYP inhibition measurements.

This study also provides publicly accessible models that could be easily used by chemoinformatics community to screen their compounds for CYP inhibition activity. While there were many publications in this area, in most cases the published models and data are not publicly available and can not be used by the community. Moreover, the use of these models will allow to better evaluate the usefulness of HTS screening techniques and *in silico* approaches for identification of CYP inhibitors.

2 General methodology

This chapter focuses on the methodological aspects of QSAR research in general and the studies presented in this thesis in particular. The QSAR study starts with the object of research - a physicochemical or biological property (or a set of properties) to be modeled. The scope of the study is often determined by the available dataset of experimentally obtained measurements.

The first step in QSAR modeling that defines all subsequent steps is selection, analysis and preparation of datasets. The datasets may be obtained from scientific literature, downloaded from specialized databases, or measured directly. Most of QSAR studies performed by pharmaceutical companies are based on in-house datasets of experimental measurements. Except for the experimentally measured value, the context and the conditions of the measurement are very important. A good dataset would include information on temperature, pressure, pH, concentration or other experimental conditions important for the measured property.

The next step consists of choosing the tools for analysis of available dataset.

Representing information about molecules in a computer-processable format is essential for QSAR analysis. Depending on the task requirements (storing 3D conformation information, storing charge information, etc) the molecule may be represented in one of the several common formats.

Molecule conformation sampling and molecule conformation optimization steps are essential in QSAR studies for properties dependent on 3D molecular information. The initial molecule structures are optimized to get a most probable bioactive conformation.

A key role in QSAR studies is the choice of descriptors. There are two basic approaches to the selection of descriptors to build QSAR. First, mechanistic, based on a priori choice of descriptors, based on known data about the property being studied and the most important structural features of the studied molecules. Second, statistical, based on the assumption that the choice of descriptors should not be made subjectively. In this case various chemometric methods are used to construct the models and select the most appropriate descriptor sets. The advantage of the statistical approach includes the absence of the subjective factor which may bias model performance.

The matrices of molecular descriptors and experimental values are processed by machine learning methods to produce a predictive model. The choice of a machine learning method is dictated by the dataset and problem specifics.

Proper model validation procedure is essential for correct estimation of model performance. Statistically valid tests should be performed during model comparison in order to make conclusions on advantages or disadvantages of each particular model.

The use of applicability domain approaches is essential for estimating individual accuracy of the model for each compound. Since model performance is non-uniform on the whole chemical space, separating molecules with confident predictions from the non-confident ones is beneficial.

2.1 Molecule representation in QSAR

2.1.1 Small molecules

The common way of representing a molecule in literature is a molecular name or a 2D sketch of the molecule. Both these ways have their own disadvantages and generally are not used for computations. All of the machine-readable molecule representations are based in one way or another on the representation of the molecule as a non-directed graph. The most common representations include SMILES, Molfile/SDF, MOL2 and InChI / InChIKey.

SMILES (Simplified molecular input line entry specification) is a way of unambiguously representing molecules with short, human-readable ASCII strings. While in general the same molecule may have several SMILES representations, the *canonical* SMILES, built following a specific rule, is unique for a molecule. The advantage of this format is its short and human-readable nature. The disadvantage is the inability to represent individual atomic coordinates in a molecule and thus to represent different molecular conformations.

In terms of graph theory, SMILES is obtained by printing symbols (atom and bond representations) encountered during a depth-first traversal of a chemical graph. Hydrogen atoms are often removed prior to traversal. All cycles (including aromatic rings) are broken in the graph and numbers are used to indicate connection points. Parentheses are used to indicate points of branching on the tree. Different extensions of the SMILES standard exist, such as *isomeric* SMILES.

InChI (IUPAC International Chemical Identifier) is also a textual one-string representation of a molecule designed to provide a human-readable and machine-processable standard for storing molecule information [70–72]. The standard is designed and maintained by IUPAC (International Union of Pure and Applied Chemistry). The InChI string contains “layers” and “sublayers” separated by slashes. Some layers and sublayers are optional. Each layer, except the first one, starts with a specific prefix. The layers include: main layer (chemical formula, atom connections, hydrogen atoms), charge layer (positive charges, negative charges), stereochemical layer (double bonds, tetrahedral stereochemistry, type of stereochemistry information), isotopic layer, fixed-H layer and reconnected layer. The advantage of separator-prefix format is the possibility to parse large amounts of molecules in InChI representation by wildcards or regular expressions to filter molecules with specific features.

InChIKey is a complementary format and is a hashed version of the full standard InChI using SHA-256 algorithm. It consists of three parts separated by hyphens: a 14 character hash of the connectivity information layers, 9 character hash of the rest of the layers plus one character identifying the version of the InChI calculation tool, and one character checksum information. The InChIKey has a fixed length, is unique for every molecule and therefore is extremely useful for searches in different database implementations.

Figure 2.1 displays some examples of SMILES, InChI and InChIKey representations of molecules.

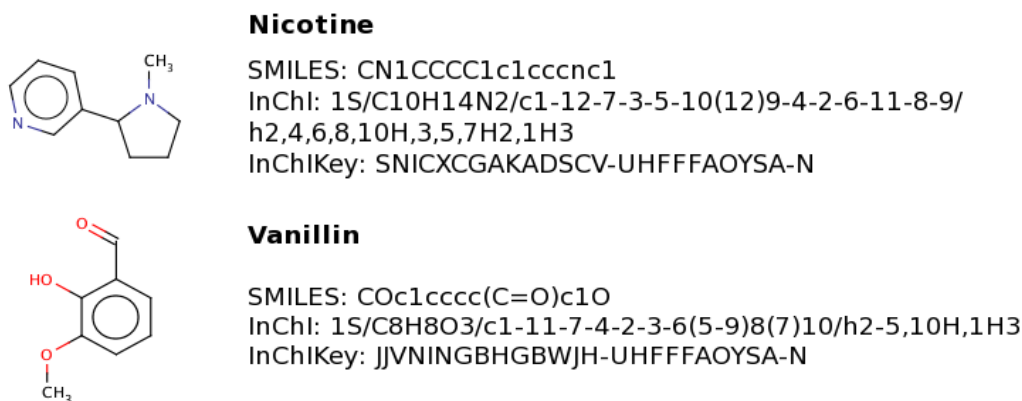


Figure 2.1. Sample SMILES, InChI and InChIKey representation of Nicotine and Vanillin

Molfile (or MDL) is a connection-table style molecule representation format. It consists of a header (three lines that can hold arbitrary information on the information origin, name, date, etc.), a summary line (containing information on the total number of atoms in a molecule, total number of bonds in a molecule, file format version, etc.), atom information section (each line of which holds atom type, three atomic coordinates, supplementary atom information) and bond information section (each line of which holds bond type, numbers of atoms a bond connects, supplementary bonds information). This file format is less human-readable, but can hold every particular details of the molecule conformation.

SDF (structure-data file) is an extension of the Molfile format. It allows to supplement the molecule structure information with a set of key-value pairs named “tags”. Tags can hold any additional information about a molecule a user would like to store - molecule names, molecular weight, some internal database identifier, etc. SDF format also allows to store several molecules in one file separated by \$\$\$\$ separator.

MOL2 is a molecule representation format similar to SDF. It allows storing information about atom and bond types, atom coordinates and additional information. This file format includes more atom types than SDF, since it discriminates, for example, aromatic and non-aromatic carbons. The distinctive feature of MOL2 file is the ability to store partial charges among other atomic information, which makes this format important for molecular descriptor calculation tools that incorporate information about charges.

2.1.2 Protein structures and the Protein Data Bank

The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules. The data contained in the archive include atomic coordinates, crystallographic structure factors and NMR experimental data. Aside from coordinates, each deposition also includes the names of molecules, primary and secondary structure information, sequence database references, where appropriate, and ligand and biological assembly information, details about data collection and structure solution, and bibliographic citations.

All the Protein Data Bank entries are stored in a special format - PDB. Every PDB file is presented in a number of lines. Each line in the PDB entry file consists of 80

columns and is self-identifying. The first six columns of every line contains a record name. Record names are listed and explained in great detail in the PDB format guide [73]. Another way to describe a PDB file is as a collection of record types. Each record type consists of one or more lines.

Records can be semantically grouped into sections.

- A *title* section contains records used to describe the experiment (title, authors, literature citations, etc) and the biological macromolecules present in the entry
- The *primary structure* section of a PDB formatted file contains the sequence of residues in each chain of the macromolecules. Embedded in these records are chain identifiers and sequence numbers that allow other records to link into the sequence
- The *heterogen* section contains the complete description of non-polymer chemical residues in the entry
- The *secondary structure* section describes helices, sheets, and turns found in protein and polypeptide structures
- Several additional information and annotation sections include the *connectivity annotation* section and *miscellaneous features* section
- A separate section describes the geometry of the crystallographic experiment and the coordinate system transformations.
- The *coordinate* section contains the collection of atomic coordinates
- The *connectivity* section provides information on atomic connectivity
- The *bookkeeping* section provides some final information about the file itself

The format, initially designed to describe biopolymers, can also serve as a structure format for small molecules. This way it only needs the title, coordinate, connectivity and bookkeeping section.

The AutoDock Vina [74] docking tool used in this study requires both protein and ligand be presented in PDBQT format. PDBQT [75] is a modification of the PDB format to store additional annotations of atoms in the protein and ligand structures.

2.2 Molecule preprocessing, conformation sampling and optimization

2.2.1 Molecule preprocessing

Prior to any further analysis by most computational methods molecules should be standardized, neutralized and subjected to salt- and counter-ion removal.

Standardization is the process of transforming a molecule according to a set of SMARTS templates. The templates used in the this study allow converting nitro mesomers. It is a required step to receive consistent molecule datasets. Due to limitations of molecule representation in QSAR, molecules with different nitro mesomer representations may be treated as different molecules. This is wrong from a chemical and biological point of view. Therefore, it is required to convert all analyzed molecules to a consistent representation.

Neutralization refers to neutralization of charged atoms in the molecules by attaching additional hydrogen atoms to them. Mesomers like nitro groups or quaternary nitrogens without hydrogens remain intact.

Remove salts is a procedure that allows removing salts, counter-ions, solvents and other molecule fragments from molecular structure. From all the detached fragments the biggest by mass is usually kept. It is an important step, since a large amount of molecule optimization or molecule descriptor calculation tools can not correctly process molecules containing salt or counter-ions. This procedure, however, results to loss of information on complete molecule structure and may lead to false duplicates in analyzed datasets.

Figure 2.2 shows the examples of molecule structure preprocessing.

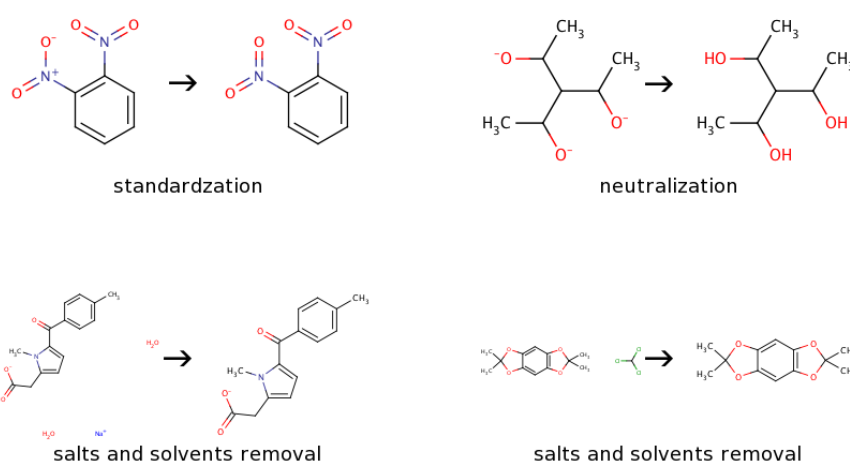


Figure 2.2. Examples of possible molecule preprocessing options

All preprocessing steps in this study were performed by Chemaxon software (Chemaxon Marvin and Chemaxon Standardizer libraries) [76,77].

2.2.2 Molecule conformation sampling and optimization

When intending to calculate 3D descriptors or perform studies related to protein-ligand interactions it is important to remember that most organic molecules of nontrivial size are not just three-dimensional, they are “four-dimensional”, because they exist as an ensemble of three-dimensional conformations interchanging over time (or equivalently, consist of a distribution of conformations at any time). Their properties and reactivities depend intimately on this ensemble. Knowing the “structure” of a molecule requires either knowing all the structures in this ensemble or knowing one structure of the ensemble that is favorable for a particular task. Respectively, the task of obtaining bioactive conformations of molecules is often divided to two major subtasks - conformation sampling (i.e., generation of multiple conformations) and conformation optimization.

Conformation generation algorithms fall into two broad categories: deterministic, which exhaustively enumerate all possible torsions at certain discrete intervals, and stochastic, which use a random element to explore the molecule’s conformational space. For flexible molecules, stochastic methods, such as molecular dynamics and Monte Carlo sampling can be considered preferable, since deterministic sampling of the torsions of all rotatable bonds would have exponential computational complexity with respect to the number of these bonds.

An alternative approach, known as distance geometry, is to generate conformations that satisfy a set of geometric constraints derived from the molecular connectivity table. There are two forms of constraints: distance constraints encoded in the form of upper (u_{ij}) and lower (l_{ij}) bounds for every interatomic distance d_{ij} (such that $l_{ij} < d_{ij} < u_{ij}$), and volume constraints that prevent the signed volume V_{ijkl} formed by four atoms i, j, k, l from exceeding certain limits. The latter are used to enforce planarity of conjugate systems and correct chirality of stereocenters. The advantage of distance geometry is that it generates chemically sensible conformations without any direct energy calculation.

Two main components of conformation optimizations are an optimization method and a target function. The target function for conformation optimization is generally energy as calculated by a specific force field. In the context of molecular modeling, a force field refers to the form and parameters of mathematical functions used to describe the potential energy of a system of atoms in a molecule. Force field functions and parameter sets are derived from both experimental work and high-level quantum mechanical calculations. One can differentiate between “all-atom” force fields that provide parameters for every type of atom in a system, including hydrogens, and “united-atom” force fields that treat the hydrogen and carbon atoms in methyl and methylene groups as a single interaction centers. For biopolymers even more crude representations are used in form of “coarse-grained” force fields.

The MMFF94 [78] force field is used to evaluate conformation quality by OpenBabel software package in the studies presented in this work. Other popular force fields (and their implementations in software packages) used for molecular dynamics are, for example, CHARMM, AMBER and GROMACS [79–82].

The total energy by *MMFF94* force field can be described as follows

$$E_{MMFF} = \sum EB_{ij} + \sum EA_{ijk} + \sum EBA_{ijk} + \sum EOO P_{ijk;l} \\ + \sum ET_{ijkl} + \sum E_{vdW}_{ij} + \sum EQ_{ij}$$

indices *i*, *j*, *k* and *l* indicate atoms, respectively.

The *EB* term represents bond stretching energy, *EA* - angle bending energy, *EBA* - stretch-bend interactions, *EOOP* - out-of-plane bending at tricoordinate centers, *ET* - torsion interactions. These terms represent bonded interactions within a molecule.

The *EvdW* term represents van der Waals interactions and *EQ* - electrostatic interactions. These terms represent the non-bonded interactions.

Each term is calculated by its own approximation function and is parametrized based on experimental data.

The optimization algorithm itself can be an implementation of any widely used general-purpose optimization algorithms from computer science studies. OpenBabel [83], for example, uses gradient descent for conformation optimization. Balloon software [84,85] uses gradient descent for initial conformation optimization and the specialized genetic algorithm for further conformation generation and optimization.

The software package used in this study for 3D optimization is Corina by Molecular Networks GmbH [86]. It uses a rule-based approach empirical optimizations.

By combining monocentric fragments with standard bond lengths and angles and by using appropriate dihedral angles a 3D model of a molecule is built. Bond lengths and angles are taken from a table. Since multiple solutions exist for torsion angles, Corina uses empirical approaches to tackle two separate problems: selection of bond torsions in a way that would ensure proper ring closure, and minimization of non-bond atom interactions (“atom crowding”).

For the ring closure problem rings of up to a size of nine atoms are processed by using a table of single ring conformations that implicitly ensure ring closure. In the case of fused or bridged systems, a backtracking search procedure finds a contradiction-free set of conformations for each single ring following some geometric and energy restrictions. The ring conformations are then translated into 3D coordinates and further refined using a simplified pseudo force field that contains only special geometric terms for the optimization of ring systems.

To minimize the non-bond interactions the principle of longest pathways has been implemented in Corina for acyclic fragments and molecules. The main chains are extended as much as possible by setting the torsion angles to *anti* or *trans* configurations, unless a *cis* double bond is specified. This method effectively minimizes non-bonding interactions.

After the combination of the three-dimensional fragments of the ring systems and of the acyclic parts, the complete 3D model is checked for overlap of atoms and for close contacts. If such situations are detected, Corina performs a reduced conformational analysis

in order to avoid these interactions. First, a strategic rotatable bond within the pathway connecting the two interacting atoms is determined, depending on topological features and double bond character. Secondly, the torsion angle of this bond is changed until the non-bonded interactions are eliminated. For appropriate torsion angles, Corina uses a set of rules and data obtained from a statistical analysis of the conformational preferences of open-chain portions in small molecule crystal structures.

Special extensions are made to handle big ring structures and organometallic complexes.

This makes Corina an extremely useful tool for conformation generation and optimization for the studies presented in this work.

2.3 Molecular docking

The number of algorithms available to assess and rationalize ligand protein interactions is large and ever increasing. Many algorithms share common methodologies with novel extensions, and the diversity in both their complexity and computational speed provides a plethora of techniques to tackle modern structure-based drug design problems [87]. Assuming the receptor structure is available, a primary challenge in lead discovery and optimization is to predict both ligand orientation and binding affinity; the former is often referred to as “molecular docking”.

Molecular docking is a computer simulation procedure to predict the conformation of a receptor-ligand complex, where the receptor is usually a protein or a nucleic acid molecule (DNA or RNA) and the ligand is either a small molecule or another protein. It can also be defined as a simulation process where a ligand position is estimated in a predicted or predefined binding site [88].

Molecular docking simulations may be used for reproducing experimental data through docking validations algorithms, where protein-ligand or protein-protein conformations are obtained *in silico* and compared to structures obtained from X-ray crystallography or nuclear magnetic resonance. Furthermore, docking is one of main tools for virtual screening procedures, where a library of several compounds is “docked” against one drug target and returns the best hit. The procedure of virtual screening through docking has become crucial when it is necessary to test a database of thousands (or even millions) of compounds against one or more targets in a short period of time. This search would be impossible to be reproduced experimentally at a so small economic and time cost. For this reason docking has been found to be a useful step in QSAR studies, where statistical analysis is applied to thousands of drug candidates.

In this thesis molecular docking is used as a step in calculation of novel Protein-Ligand Interaction-Based descriptors. The docking of the ligand is first performed against the target in question, and then a set of descriptors is calculated as functions of relative coordinates of atoms of the ligand and the target.

2.3.1 Short classification of available methods

A number of reviews on molecular docking algorithms, methods and software that provide a thorough classification and comparison of different approaches used in the area were published during the last years [87–91].

A general approach to classifying docking protocols is as a combination of two mostly independent components; a search algorithm and a scoring function. The search algorithm should generate an optimum number of configurations that include the experimentally determined binding mode. Generating a broad range of binding modes is ineffective without a model to rank each conformation that is both accurate and efficient. The scoring function should be able to distinguish the experimental binding modes from all other modes explored through the searching algorithm.

The quality of the docking methodology can be evaluated by performing the docking on the ligands, for which the crystallographic information is available. The quality score would then be root-mean-square deviation (RMSD) calculated on two sets of coordinates - obtained by crystallography and by docking simulations.

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N ((x_{ci} - x_{di})^2 + (y_{ci} - y_{di})^2 + (z_{ci} - z_{di})^2)}$$

where (x_{ci}, y_{ci}, z_{ci}) is a set of crystallography-obtained coordinates for the i -th atom of the ligand and (x_{di}, y_{di}, z_{di}) is a set of docking-obtained coordinates. The coordinates of the atoms of the protein should be the same for two compared cases.

All current docking approaches can be separated into three major groups.

Rigid ligand to rigid protein docking is a common approximation in early docking algorithms [92]. Both the ligand and target are treated as rigid bodies and only the six degrees of translational and rotational freedom are explored. This approach is extremely simple and not very computationally demanding. Although this method has been successful in certain cases, there is a limitation to the rigid body docking paradigm in that the ligand conformation must be close to the experimentally observed conformation when bound to the target [93]. Furthermore, numerous examples of conformational change of the target upon binding to a receptor limit the applicability of this type of methods [94].

Flexible ligand to rigid protein docking is the most popular approach used in modern docking studies. In this approach the ligand is considered flexible and the traversed search space includes the conformational space of the ligand itself in addition to the position of the ligand with respect to the protein. For this case the variety of search algorithms and conformation evaluation functions exist. The task pursued in flexible ligand docking is similar to the tasks of conformation generation and optimization described in section 2.2 , page 11.

Flexible ligand to flexible protein docking is the most computationally demanding type of docking and is a target of intensive research. The approaches used in this type of docking include docking to a fully rigid protein with some relaxed constraints on protein and ligand atoms overlapping [95,96], docking of a ligand to a protein with flexibility of only the predefined side-chains in the binding site [97–101], and docking of ligand to several rigid conformations of the protein (possibly generated by molecular dynamics methods) [102–

104].

Based on the used conformation search algorithm most of the docking methods can be classified into several large categories. In existing software the methods are often used complementarily.

Molecular dynamics methods involve the calculation of solutions to Newton's equations of motions. The goal of the MD simulations is generally finding the global minimum energy of a docked protein-ligand complex. Due to high computational demands and several methodological problems (due to its gradient nature, the method tends to get "stuck" in local minima, unable to step over an energy barrier to reach the favorable binding conformation) the method is often used on a local conformation optimization step, conformation being produced by some other algorithm [79,101,105].

Monte Carlo simulations is an established and popular approach in docking software [106–108]. A significant advantage of the MC technique compared with gradient based methods, such as MD, is that a simple energy function can be used which does not require derivative information. Furthermore, through a judicious choice of move type, energy barriers can simply be stepped over. Force fields are used to estimate favorability of each conformation. Most popular force fields used in docking software are different modifications of CHARMM and AMBER.

Since their inception, *genetic algorithms* have increased in popularity as an optimization tool. The genetic algorithm approach is also widely used in docking [103,109–111]. The essence of a GA is the evolution of a population of possible solutions via genetic operators (mutations, crossovers and migrations) to a final population, optimizing a predefined fitness function. The mutation operator randomly changes the value of a gene, crossover exchanges a set of genes from one parent chromosome to another, and migration moves individual genes from one sub-population to another. The fitness function of genetic algorithm based docking solutions is also an energy function as defined by a force field of choice.

The broad philosophy of *fragment based docking* methods can be described as dividing the ligand into separate portions or fragments, docking the fragments, followed by the linking of fragments [112–114]. These methods require subjective decisions on the importance of the various functional groups in the ligand, which can result in the omission of possible solutions, due to assumptions made about the potential energy landscape.

Docking ligands to the binding site of a receptor is often performed using *points of complementarity* between the protein and ligand. Many of the fragment based docking algorithms could also be included in this category, although an important distinction is generally made between algorithms that treat the ligand as a complete entity throughout the docking method, and those where the ligand is divided into fragments [115,116].

Tabu search algorithms are a family of docking algorithms which may be described as a stochastic evolution of the system using a tabu search with a generalized scoring function [117–119].

In this thesis an AutoDock Vina tool was used for flexible ligand to rigid protein docking.

2.3.2 AutoDock Vina

AutoDock Vina is a relatively new program for molecular docking and virtual screening [74]. It is a successor of an older docking suit, AutoDock 4. The developers claim that Vina achieves a two orders of magnitude speed-up compared to the previous version, while also significantly improving the accuracy of the binding mode predictions, judging by the tests on the training set used in AutoDock 4 development.

The optimization algorithm used in Vina is a variation of an “iterated local search” global optimizer [120,121]. Vina uses the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [122] method for the local optimization, which is considered to be an efficient quasi-Newton method.

The *scoring function* is a function developed especially for the Vina tool. The current implementation of Vina is designed to work with the scoring functions that have a general form of

$$c = \sum_{i < j} f_{t_i t_j}(r_{ij})$$

where the summation is over all of the pairs of atoms that can move relative to each other, normally excluding 1–4 interactions, i.e., atoms separated by three consecutive covalent bonds. Here, each atom i is assigned a type t_i , and a symmetric set of interaction functions $f_{t_i t_j}$ of the interatomic distance r_{ij} should be defined. This value is considered as a sum of intramolecular and intermolecular interactions.

$$c = c_{intra} + c_{inter}$$

The optimization algorithm attempts to find the global minimum of c and other low-scoring conformations, which it then ranks.

The predicted free energy of binding is calculated from the intermolecular part of the lowest-scoring conformation:

$$s_1 = g(c_{inter})$$

In the current implementation of Vina the atom typing scheme follows that of X-score [123]. The hydrogen atoms are not considered explicitly, other than for atom typing, and are omitted from the scoring function. The interaction functions $f_{t_i t_j}$ are defined relative to the surface distance $d_{ij} = r_{ij} - R_{t_i} - R_{t_j}$:

$$f_{t_i t_j}(r_{ij}) \equiv h_{t_i t_j}(d_{ij}) ,$$

where R_t is the van der Waals radius of the atom type t . The scoring function $h_{t_i t_j}$ is defined as a weighted sum of five terms:

$$h_{t_i t_j}(d_{ij}) = -0.0356 \cdot \text{gauss}_1(d_{ij}) - 0.00516 \cdot \text{gauss}_2(d_{ij}) + 0.840 \cdot \text{repulsion}(d_{ij}) \\ - 0.0351 \cdot \text{hydrophobic}(d_{ij}) - 0.587 \cdot \text{hydrogenbonding}(d_{ij})$$

The coefficients were tuned by the Vina authors using experimental data from PDBind. The individual terms are defined as following:

$$\text{gauss}_1(d) = e^{-(d/0.5\text{\AA})^2} \quad \text{gauss}_2(d) = e^{-(d-3\text{\AA})/2\text{\AA})^2} \\ \text{repulsion}(d) = \begin{cases} d^2, & \text{if } d < 0 \\ 0, & \text{if } d \geq 0 \end{cases}$$

$$\left. \begin{aligned} \text{hydrophobic}(d) &= \begin{cases} 1, & \text{if } d < 0.5 \text{ \AA} \\ 0, & \text{if } d > 1.5 \text{ \AA} \\ (-d + 1.5 \text{ \AA}) / 1 \text{ \AA}, & \text{if } 0.5 \text{ \AA} \leq d \leq 1.5 \text{ \AA} \end{cases} \\ \text{hydrogen bonding}(d) &= \begin{cases} 1, & \text{if } d < -0.7 \text{ \AA} \\ 0, & \text{if } d > 0 \text{ \AA} \\ -d / 0.7 \text{ \AA}, & \text{if } -0.7 \text{ \AA} \leq d \leq 0 \text{ \AA} \end{cases} \end{aligned} \right\}$$

The function g used for ranking the conformations and for predicting free binding energy is:

$$g(c_{inter}) = \frac{c_{inter}}{1 + 0.0585 \cdot N_{rot}}$$

where N_{rot} is the number of active rotatable bonds between heavy atoms in the ligand.

The implementation details are described in the original paper [74].

AutoDock Vina was chosen as a docking tool for the studies in this thesis for number of its distinctive features:

- high benchmarked accuracy of conformation predictions
- high speed of calculations
- runs on most Linux platforms and Mac OS X
- support for multiprocessor and multicore parallelization
- lightweight one-executable software, easy to use in a cluster environment
- available for free

2.4 Molecular descriptors

Machine learning methods are statistical and computer science methods that operate on numerical representation of entities. When applied to predicting chemical and biological properties of small molecules the important task arises to adequately represent a small molecule in the form of a numerical vector. The numbers of this vector that are used to represent different structural or functional aspects of the molecule are called molecular descriptors.

Since molecules and molecule interactions are complex entities, any numerical vector would only be an approximation of these entities with respect to some specific problem. Therefore it is important to choose a set of descriptors relevant to a particular problem.

There are numerous ways to classify descriptor sets and descriptor calculation software. A widely accepted classification approach is to rely on the type of structural information used by the method and split the whole variety of descriptors into five categories of 0D - 4D descriptors [124]:

- The 0D descriptors are the descriptors independent of any knowledge concerning

the molecular structure. The 1D descriptors are calculated over such one-dimensional representations of a molecule. Examples of 0D descriptors are total atom number, absolute or relative number of specific atom types, absolute or relative number of specific bond types, etc; the 1D descriptors include counts of fragments or functional groups of interest present in the molecule. The 0D/1D classification is often ambiguous.

- The 2D descriptors are derived from two-dimensional topological representation of the molecule and include topological information indices, molecular profiles and 2D autocorrelation descriptors.
- The 3D descriptors are based on a three-dimensional representation of the molecule and require a valid optimal three-dimensional conformation of a represented molecule. The examples of these descriptors include WHIM, GETAWAY and 3D-MoRSE descriptors from Dragon software package. The Protein-Ligand Interaction-Based descriptors are calculated based on the three-dimensional structure of a ligand obtained by docking simulations and, therefore, can be considered 3D descriptors.
- The 4D descriptors are descriptors calculated over an ensemble of molecular conformations. One approach to generating 4D descriptors would be to generate an ensemble of potentially active optimal conformations and calculate normal 3D descriptors over each of these conformations. Average values and standard deviations of these descriptors over a whole conformation ensemble would be classified as 4D descriptors.

Another way to classify the descriptors would be into two classes: general purpose descriptors and chemogenomics based descriptors.

- *General purpose descriptors* are based solely on the structure of the molecule itself. They may either be calculated on the molecule structure directly or be a result of simulations of interaction of a molecule with some default force field or probe atom / molecule. These descriptors are general and can be equally used for predicting both chemical and biological properties of molecules. There is a wide variety of successful general purpose descriptors used for QSAR studies.
- *Chemogenomics based descriptors* are calculated with the intention of describing not the small molecule itself, but rather a protein - small molecule interaction. These descriptors incorporate not only structural information of the small molecule, but of the protein as well. Depending on the approach the information may include atomic coordinates of the protein, specifics of the protein binding site, relative position of the protein and the small molecule in question, etc. These descriptors mostly make sense only in the context of biological properties related to protein-ligand interaction such as activation or inhibition of a specific protein, molecule binding affinity, etc. The presented Protein-Ligand Interaction-Based descriptors are chemogenomics based descriptors.

2.4.1 General purpose descriptors

Throughout this study a number of general purpose descriptors were used and their performance compared. These descriptors are described briefly below.

ISIDA SMF descriptors [125,126] were calculated using the fragmentation tool from the ISIDA suite. The *substructural molecular fragments* (SMF) method is based on the splitting of a molecule into fragments. The fragment type is then a descriptor, and the number of occurrences of this fragment in a molecule is the value for this descriptor. Two different types of fragments are considered: “sequences” and “augmented atoms”. For each type of fragment three subtypes can be defined **AB** (atom and bond types), **A** (atom types only), and **B** (bond types only). In the studies presented in this work the **AB** type descriptors were used.

Atom type E-state indices and molecular bond E-state indices are described in appropriate articles by Hall and Kier [127]. These descriptors combine electronic and topological properties of the described molecules. Each atom in the molecular graph is represented by an E-state variable, which encodes the intrinsic electronic state of the atom as perturbed by the electronic influence of all other atoms in the molecule within the context of the topological character of the molecule. The E-state index for an atom or bond consists of an intrinsic value for that atom/bond plus a term for its perturbation by all the other atoms in the molecule. For every atom type and bond type in the molecule the calculated indices are summed.

Dragon [124] is a software tool licensed by Talete inc. The Linux version of Dragon – dragonX 1.2.4, which calculates 1664 molecular descriptors, was used in this thesis. These descriptors cover 0D - 3D descriptors which are arranged into 20 blocks. Dragon descriptors are very popular and are often successfully used for QSAR modeling of various properties.

2.4.2 Chemogenomics based descriptors

Chemogenomics is an emerging interdisciplinary field described [128] as “the study of the genomic and/or proteomic response of an intact biological systems whether it be single cells or whole organisms to chemical compounds, or the study of the ability of isolated molecular targets to interact with such compounds”. This chapter focuses on several examples of chemogenomic methods used in QSAR studies. It also argues about the necessity of a new set of chemogenomics based descriptors. Finally, a novel set of descriptors is presented.

Chemical genomics based virtual screening approach applied in several studies demonstrated excellent results in prediction of biological activities of small molecules and in finding novel bioactive molecules [129–133].

An example study applying this approach [129] was aimed at predicting small molecule activity on a set of G-protein coupled receptors the initial dataset contained information on 5207 small molecule - protein interactions (a total of 317 unique GPCRs and 866 ligands). Descriptors were calculated separately for small molecules and proteins. Chemical descriptors for small molecules were calculated using traditional 2D molecular descriptors. Protein descriptors were calculated from the sequences alone based on a

mismatch-allowed spectrum kernel. The concatenated protein-ligand descriptors vector was used in the QSAR study of predicting small molecule activity against individual GPCRs.

This method has several important advantages. One of the advantages is the lack of necessity to generate bioactive conformations of small molecules and independence of the 3D structure of the protein in question. This allows fast and efficient screening for proteins for which the structure is unknown. Another advantage is the possibility to combine knowledge about interactions of small molecules with a wide variety of proteins in a systematic and sensible manner.

On the other hand, this method can not be applied to a dataset containing molecule activity against one protein or few closely related proteins, like CYP family. In this case the protein section of the descriptors would not be discriminative enough to increase the model predictivity compared to the traditional QSAR models build on traditional molecular descriptors. Since the crystal structures of many important CYPs have been reported, it would make sense to incorporate this information into the QSAR model.

COMBINE (COMparative BINding Energy) analysis [134,135] is a well established method to derive a system-specific expression to compute binding free energy using the three-dimensional structures of receptor-ligand complexes. The method uses empirical scoring functions that are quick to compute to estimate binding free energy using a single structure of a receptor-ligand complex. If some experimental binding data are available for a set of related complexes, then this information is used to derive a target-specific scoring functions. This is the approach taken in COMBINE analysis in which the binding free energy, inhibition constant or some related property is correlated with a subset of weighted interaction energy components determined from the structures of energy minimized receptor-ligand complexes. These energy terms can be considered as descriptors in terms of QSAR studies.

The COMBINE analysis method is based upon the assumption that the binding free energy (ΔG), measured in experiments as inhibition constants (K_i), can be correlated by Partial Least Squares (PLS) with selected weighted interaction energy terms. In the recent COMBINE studies Coulombic (ΔU_{elec}) and Lennard-Jones (ΔU_{vdw}) interaction energies as well as electrostatic solvation energy terms (ΔG_{solv}) were computed using the 3D coordinates of all-atom models of receptor-ligand complexes. The Coulombic and Lennard-Jones interaction energies were partitioned into interaction terms (Δu_{elec} and Δu_{vdw}) between each amino acid residue of the receptor and the ligand. The ligands were not subdivided because of the high diversity of their lead structures in the datasets studied. The electrostatic solvation energy terms were calculated for the ligand (ΔG_{solv}^L) and for the protein (ΔG_{solv}^R) by solving the Poisson-Boltzmann equation. The resulting model would have the form of the sum:

$$\Delta G = \sum_{i=1}^n w_i^{vdw} \Delta u_i^{vdw} + \sum_{i=1}^n w_i^{elec} \Delta u_i^{elec} + w_{solv}^R \Delta G_{solv}^R + w_{solv}^L \Delta G_{solv}^L + C$$

where (w_{vdw} , w_{elec} , w_{Rsolv} , w_{Lsol}) are determined by the appropriate terms contribution to the PLS model. The values (Δu_{vdw} , Δu_{elec} , ΔG_{Rsolv} , ΔG_{Lsol}) can then be considered as COMBINE descriptors in a PLS model.

To predict the binding affinities of new ligands, it is necessary to model the structures

of their protein-ligand complexes. When the ligands are similar, it may be possible to do this by analogy to experimentally determined protein-ligand complexes. However, in general, and particularly for the virtual screening of compound libraries, it is necessary to dock the ligands into the receptor binding sites *de novo*. For this reason, COMBINE analysis is coupled with a ligand-receptor docking step.

COMBINE analysis was originally developed to study the interactions of one target macromolecule with a set of related ligands. Since then, it has been shown that the method can be successfully applied to a wide variety of complexes including enzyme-substrate and inhibitor complexes, protein-protein/peptide complexes, and protein-DNA complexes [136–140].

While being successfully used in several studies, the COMBINE approach has several issues that require additional handling. First is that the resulting energy is considered to be a linear composition of its (energy-type) terms. Applying a non-linear machine learning method, while quite possible, would ruin the concept of a resulting energy being a sum of individual energy terms. Second issue lies in the calculation of the energy terms themselves. The software calculating the energy terms uses a particular force field, that is empirical by itself and is parametrized for some subset of molecular interactions. This conversion introduces additional noise to the resulting model. And third issue lies in the per-residue calculation of energy terms. This makes seamless combining of datasets for different proteins problematic. While this problem may be partially tackled for closely related proteins by aligning their binding sites and calculating descriptors for matching residues only, the efficiency of this method and the applicability of it to non-related proteins is an open question.

Protein-ligand atom pair descriptors is a set of descriptors to transparently describe protein-ligand interactions in a way suitable for (possibly non-linear) QSAR analysis. The descriptors presented and used in this thesis are a simplified modification of the Distance-Dependent Atom-Type Pair Descriptors [141] re-purposed for the use in QSAR studies and were developed with the following ideas in mind:

- easily calculable directly from the protein-ligand complex structural information, omitting intermediate empirical steps
- universal with respect to proteins and ligands
- descriptors of interactions of different proteins should be easily combinable to form a single dataset

The calculation of descriptors is performed on the protein-ligand complex obtained from the docking or molecular mechanics simulations. The protein-ligand complex should contain the coordinates of all protein and small molecule atoms.

All atom pairs between the protein and ligand are considered. Atom pairs are combined in groups. Atom pairs are combined to groups based on:

- Protein and ligand atom types
- Protein and ligand atom partial charge sign
- Binned distance between protein and ligand atoms (distances are binned into five

groups: 0–3Å, 3–4Å, 4–5Å, 5–6Å, 6–12Å). All atom pairs with distances larger than 12Å are ignored.

In each group the number of atom pairs and the sum of products of partial charges of atom pairs is calculated. This gives two descriptors for each atom pair group.

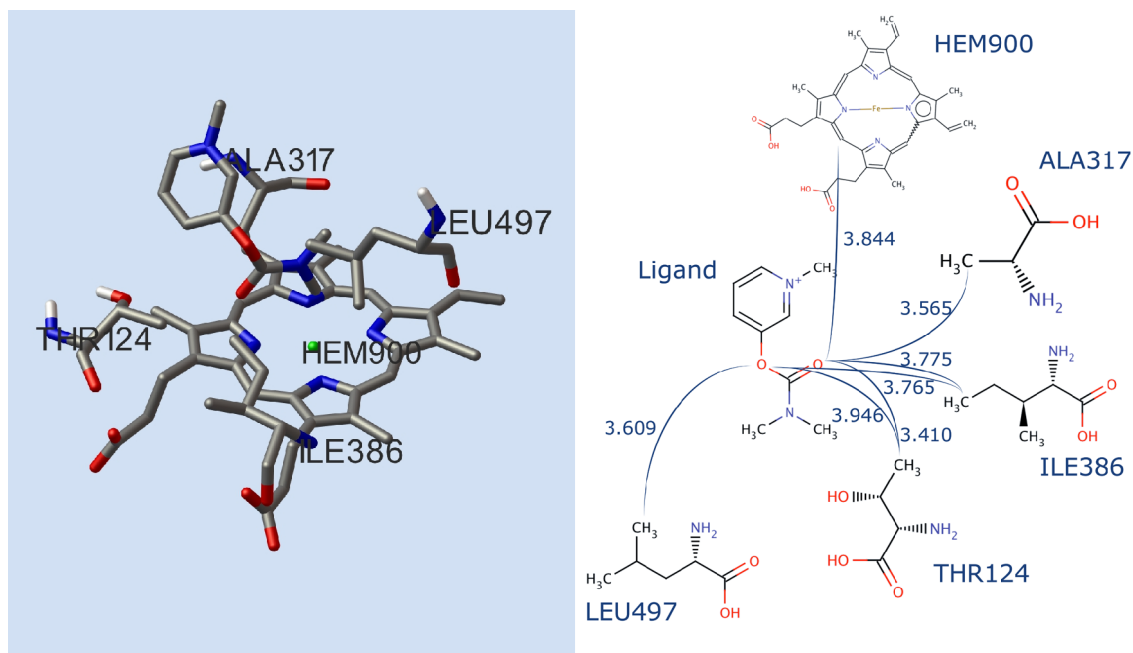


Figure 2.3. Docking conformation of a sample ligand in a fragment of CYP1A2 protein (left); sample oxygen-carbon atom pairs in the 3-4 Å distance bin (right)

Figure 2.3 (left) displays a sample ligand docked to the binding site of the CYP1A2 protein. Only amino acids with distances closer than 4Å to the ligand are displayed. Figure 2.3 (right) displays a schematic representation of protein-ligand atom pair calculation. Labeled arcs display seven atom pairs that fall to the “oxygen negatively charged, carbon positively charged, distance bin 3-4Å” group. Since for this protein-ligand atom pair the total number of atom pairs in this group is 7, the respective atom pair count descriptor value is 7.

2.5 Machine learning methods

Machine learning methods is a set of algorithms that aim to extract knowledge in some particular form from an amount of input data. The core objective of the machine learning method is to generalize the presented data and extract some specific information on the distribution of the data that would allow to either provide additional information on the existing data points or to predict specific properties of new data points.

In QSAR the most often used class of machine learning methods are so-called *supervised* machine learning methods. In supervised learning the input data consists of a number of *training examples*, which consist of the input object and the resulting signal (most commonly, vectors of numerical values or labels from a predefined set). The goal of the supervised machine learning method is to analyze a set of training examples and infer a functional dependency between the input and output objects in a dataset. Such a function is generally called a *classifier*, when the output object is a label from a discrete set, and a

regression function, when the output is a continuous numerical value. The problems in this case are called *classification* and *regression* problems respectively. Some machine learning methods can be adapted to both types of problems, while others are suited only to one type of problems.

In QSAR studies training examples most often include a vector of molecular descriptors for a particular small molecule as an input object and a (often experimentally measured) numerical value or a discrete label for a physicochemical or biological property for this molecule as an output object. The goal of the machine learning method is to build a regression function or a classifier that would allow to predict property values for new molecules, which were not yet measured or even synthesized.

The machine learning methods most often used in QSAR are multilinear regression (MLR), k-nearest neighbors (KNN), artificial neural networks (ANN), support vector machines (SVM), random trees / random forest (RT / RF), C4.5 decision trees (C4.5), etc. Linear methods additionally can be extended by kernel techniques. Kernel functions map the initial descriptor space of the problem to a higher dimensional space, thus allowing to solve non-linear problems by linear machine learning methods. Kernel-based methods are very popular in QSAR studies [142,143].

In addition, several methods used in machine learning can not be attributed to one specific machine learning method, but bear a common nature and can be applied in combination with a number of machine learning methods. These techniques can be generally called as meta-learning techniques.

This chapter gives a short summary on the machine learning methods and meta-learning methods used in this study. A physicochemical property or biological property that needs to be predicted will be referred to as *target property*. The capital letter J will represent a chemical compound, $x_i(J)$ – the i -th descriptor of a compound J , $y(J)$ and $\tilde{y}(J)$ – real and predicted values of the target property, M and N – the number of the used molecular descriptors and the number of molecules in the training set, respectively.

2.5.1 K-nearest Neighbors

KNN is a machine learning method that derives predictions on new instances based on the distance of this instance in descriptor space to k instances on the training set, where k is an optimizable parameter. Although the distance may be calculated using different metrics, an Euclidean distance is most often used. It's an example of instance-based machine learning method where the function is only approximated locally, no training process (except for, possibly, optimizing the k parameter) takes place, and all computation is deferred until the prediction phase.

For classification problems the prediction value is obtained by voting of the k nearest neighbors of the instance. For regression problems the value of the instance is generally calculated as the weighted average of values of the nearest neighbors of the instance. The weight can generally be $1/d$ where d is a distance to the neighbor. This scheme is a generalization of linear interpolation.

The KNN model is fully described by a matrix of descriptors of the training set x_{ij} .

2.5.2 Artificial Neural Networks

ANN is a family of mathematical models inspired by the functionality of a biological neuron. The neural networks most often used in QSAR studies are multilayered perceptrons [144]. A multilayered perceptron can be represented as a multilayered directed graph, where all nodes of some layer are connected to all the nodes of the previous layer. Mathematically the neural network predicting one property y for a compound J can be presented as following:

$$\tilde{y}(J) = f_{L-1,0}(J)$$

$$f_{i,j}(J) = \begin{cases} g\left(\sum_{k=1}^M w_{ijk} \cdot x_k(J)\right), & \text{for } i=0, j=0..Z_i-1 \\ g\left(\sum_{k=1}^{Z_{i-1}} w_{ijk} \cdot f_{i-1,k}(J)\right), & \text{for } i=1..L-1, j=0..Z_i-1 \end{cases}$$

where L is a total number of layers in a network, Z_i - a total number of neurons in layer i , w_{ijk} is a weight of input k of a neuron j in a layer i , and $f_{i,j}(J)$ is an output of neuron j in a layer i for a compound J . The function $g(x)$ is some nonlinear function, generally referred to as “neuron activation function”. Note that the resulting prediction is the output of a single neuron of the last layer of the network. An example of such function would be hyperbolic tangent, or some other sigmoid function. An ANN model, therefore, is completely defined by the set of neuron weights $W = \{w_{ijk}\}$, given that the configuration of the network (the number of layers, the number of neurons in each layer, the form of the activation function for every neuron) is fixed.

Neuron inputs (z_1, \dots, z_n) may either be a vector of descriptors (x_1, \dots, x_M) for the neurons of the first layer, or outputs of the neurons of the previous layer - otherwise.

The process of constructing of a predictive ANN (“training of the neural network”) lies in optimizing the input weights of all the neurons in the network, so that sum predefined cost function is minimized.

For the simplest case the cost function may be a sum of squared errors of predictions on the training set, similarly to the linear regression. When applying a gradient descent method to this minimization task, one gets a most straightforward yet very efficient algorithm for training neural networks - back propagation algorithm [145].

In this study the neural networks were trained by SuperSAB - an adaptive modification of the back propagation algorithm [146]. It's distinctive features compared to the classical back propagation algorithm is high conversion speed and insensitivity to the choice of parameter values.

Additionally, the neural networks used in this study are combined with a local correction LIBRARY approaches in a specific implementation called Associative Neural Networks (ASNN) [147,148].

2.5.3 Support Vector Machines

The original SVM is the non-probabilistic linear binary classifier and is initially suitable only for classification problems. More precisely, the algorithm tries to build a hyperplane in descriptor space that would separate the instances of two classes and that the distance to instances of each class (so-called “functional margin”) would be maximal. The algorithm is based on quadratic programming and was first introduced by Vapnik [149].

A generalization of SVM to multiple-class classification problem involves building multiple hyperplanes.

Since the construction of such a hyperplane that would separate all the instances is not always possible, an extension was suggested with a maximum margin idea that would allow misclassified instances [150]. If there exists no hyperplane that can split the examples of two classes, the *soft margin* method will choose a hyperplane that splits the examples as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. The additional penalty variables are introduced, which measure the degree of misclassification of each instance.

Although the original SVM is a linear method, an extension was proposed that allows to create nonlinear SVM classifiers by applying kernel modification to maximum-margin hyperplanes. The resulting algorithm is formally similar, except that every dot product is replaced by a nonlinear kernel function. As a result, SVM is performed not in the original space but in the *feature space* obtained via this nonlinear 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. Kernel-based SVM is very popular in QSAR studies due to the non-linear nature of QSAR problems [142,143]. The popular kernel functions are polynomial kernels, radial basis function, hyperbolic tangent, etc. Some kernels may have optimizable parameters that influence SVM performance for each particular task. These parameters, as well as a soft margin parameter of the SVM method itself, can be optimized via a grid search using cross-validation procedures.

2.5.4 Random Tree / Random Forest

Random Tree [151] is a simple example from a large family of *decision tree* algorithms. Decision trees (in general case they can be both classification and regression trees) map a number of observations (i.e, descriptors) about an item to the target property value. In a tree leaves represent the actual classification or regression values, and intermediate nodes - conjunctions of features that lead to those classifications.

Random Tree is most often used with a *bagging* meta-learning method to produce a Random Forest [152]. A single Random Tree is constructed as following (let the

number of training cases be N , and the number of variables in the classifier be M):

1. We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M .
2. A training set for this tree is formed by choosing N times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
3. For each node of the tree, randomly m variables are chosen on which to base the decision at that node. The best split based on these m variables in the training set is calculated.

For a Random Forest multiple trees are created and the result value is obtained by voting of individual trees.

2.5.5 C4.5 Decision Tree

C4.5 is a successor of an ID3 algorithm and is a decision tree algorithm based on the idea of entropy gain [153].

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists. For each list there are several base cases. If all the samples in the list belong to the same class, the algorithm creates a leaf node for the decision tree saying to choose that class. If none of the features provide any information gain, C4.5 creates a decision node higher up the tree using the expected value of the class.

Main distinctive features of C4.5 decision tree include:

- The ability to handle both continuous and discrete input attributes. In order to handle continuous attributes, C4.5 chooses a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Handling training data with missing attribute values. Missing attribute values are simply not used in gain and entropy calculations.
- Handling attributes with differing costs.
- Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

In this thesis a Java implementation of C4.5 algorithm from Weka machine learning package [154] - J48 - is used extensively for classification tasks.

2.5.6 Bootstrap aggregating (bagging)

Bootstrap aggregating (bagging) is a variation of machine-learning ensemble meta-algorithm that was proposed by Breiman [152] to increase prediction accuracy and stability and reduce the risks of over-fitting for random trees. Although it was first introduced for tree classifiers, the concept can be used for all machine learning methods, both for regression and classification problems. The method relies on building multiple classification or regression models and averaging the results (for regression tasks) or voting on the result (for classification tasks) to obtain the final prediction.

The training set for each individual model is obtained by resampling with replacement of the original training set. Given the uniform distribution of the selected training set instances, every resulting training set is likely to have 63.2% of unique instances of the original training set. This kind of training instance samples is called a bootstrap sample.

It is worth noting, that since the results of the individual regression models are averaged, the bagging approach does not increase the quality of predictions of linear models.

Bagging has been used intensively in the studies described in this thesis to successfully increase prediction accuracy of different machine learning methods, especially - random and C4.5 decision trees. Bagging approach also allows to estimate the accuracy of predictions for each individual compound, as described in the applicability domain section.

2.5.7 Local corrections and the LIBRARY approach

As discussed earlier, machine learning methods like ANN are memoryless approaches, as after the training is complete all information about the input patterns is stored in the neural network weights and the input data is no longer needed, i.e., there is no explicit storage of any presented example in the system. And contrary to that, such methods as the k -nearest-neighbors (KNN) represent the memory-based approaches, since their predictions are derived from local approximations derived directly from training data. A global model provides a good approximation of the global metric of the input data space. However, if the analyzed property is too complicated, there is no guarantee that all details of its fine structure will be represented. Thus, the global model can be inadequate because it does not describe equally well the entire state space with poor performance of the method being mainly due to a high bias of the global model in some particular regions of space.

The idea of local corrections [147] lies in combining global ANN models with local

corrections derived from KNN approach. For local corrections to make sense a special metrics of similarity of predicted instances in model output space is introduced. The most successful metrics is reported to be Spearman rank-order correlation coefficient between vectors of predictions for an instance by an ensemble of ANN networks. The KNN corrections are performed for k nearest neighbors in model output space.

Formally, let $\vec{y}(J)$ be a vector of property predictions of individual networks in an ensemble, and $\xi(\vec{z}^i, \vec{z}^j) = \left\| \frac{\vec{z}^i}{\|\vec{z}^i\|}, \frac{\vec{z}^j}{\|\vec{z}^j\|} \right\|$ - some similarity measure (e.g., Spearman rank-order correlation coefficient) between two vectors \vec{z}^i and \vec{z}^j .

Then a classical approach would be to average prediction values of the individual networks to obtain a final prediction

$$\bar{y}(J) = \frac{1}{M} \cdot \sum_{k=1}^M \tilde{y}_k(J) ,$$

where M is a number of networks in an ensemble. The local correction approach provides a corrected prediction $\bar{y}'(J)$

$$\bar{y}'(J) = \bar{y}(J) + \frac{\sum_{k=1}^K (y(J_k) - \bar{y}(J_k)) \cdot F(\xi(\vec{y}(J), \vec{y}(J_k)))}{\sum_{k=1}^K F(\xi(\vec{y}(J), \vec{y}(J_k)))}$$

where K - a number of nearest neighbors from the training set J_k to the predicted compound J (where ξ is the distance metrics), F - a weighting function for corrections introduced by individual neighbors from the distance to these neighbors. The simple case would be $F \equiv 1$ to account for corrections of individual neighbors equally.

The idea of LIBRARY approach lies in extending a ready model built with local corrections enable machine learning method with an additional set of experimental measurements (i.e., a library). This increases the overall accuracy of the model without the need of retraining it on the new data - the information from this new data will be included in the final predictions through the local corrections. The approach has been shown to increase the accuracy of LogP predictions on in-house datasets of several pharmaceutical companies [155–157].

2.6 Model performance evaluation

The general approach to model performance evaluation is to apply this model to a number of instances, for which the experimental values are known, and to calculate some performance metrics based on the real values and the prediction values obtained from the model. To estimate whether any given model has any predictive ability, a number of integral model performance measures is used. Regression and classification

models have different performance metrics.

Since the models in this thesis are only binary classification models, the performance metrics will be described in terms of binary classification. The metrics can be generalized to multiclass classification if necessary. In the definitions below numbers of true positives (instances that belong to “positive” class, which were correctly classified as “positives”) , true negatives (instances that belong to “negative” class, which were correctly classified as “negatives”), false positives (instances that belong to “negative” class, which were misclassified as “positives”), and false negatives (instances that belong to “positive” class, which were misclassified as “negatives”) are denoted as TP , TN , FP , FN , respectively. The metrics used in the studies are sensitivity, specificity, accuracy, balanced accuracy and Matthew's correlation coefficient.

2.6.1 Sensitivity and specificity

Sensitivity and specificity are measures of the ability of the model to correctly detect “positives” and “negatives” respectively.

$$SENS = \frac{TP}{TP + FN}, \quad SPEC = \frac{TN}{TN + FP}$$

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.

2.6.2 Accuracy

In classification tasks accuracy is determined as the ratio of correctly classified instances to a total amount of classified instances.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy can be an acceptable measure if the number of “positive” and “negative” instance in the set is approximately same (i.e., a set is balanced). For highly imbalanced sets it is more informative to use a balanced accuracy.

2.6.3 Balanced accuracy

Balanced accuracy is calculated as a weighted sum of accuracies within each class.

$$BACC = \frac{SENS + SPEC}{2} = \frac{1}{2} \cdot \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

2.6.4 Matthews correlation coefficient

Matthews correlation coefficient [158] takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between -1 and $+1$. A coefficient of $+1$ represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

Matthews correlation coefficient can be used to estimate classification model performance disregarding the differences in datasets.

2.7 Model validation

Model validation is a process of obtaining model performance metrics in a way to reflect the actual model prediction ability. Given a fixed limited set of experimentally measured points, a good validation strategy would try to simulate as close as possible the “real world” scenario, when the resulting model is presented with completely new instances, not used in training.

Estimating model performance on the same dataset that was used for model training could lead to over-optimistic results. Given a large enough number of descriptors, it is possible to construct a model that will “remember” the exact values for all instances in the training set. Such a model, however, will probably have a poor performance on any new data due to the lack of generalization abilities. These models are referred to as *over-fitted* models and should be avoided.

The validation strategies used in this thesis are test set validation, cross-validation and bagging validation.

2.7.1 Test set validation

The most straightforward validation strategy would be to divide the initial available dataset into two subsets - training set and validation set. The training set is then used for model training. The resulting model is then applied to the instances from the validation set and the performance metrics are calculated.

The training and test sets may either be of equal or different sizes, and may either be created by random splitting or by specialized procedures. Although specialized procedures (e.g., “optimal design”) can yield smaller training sets and better model performance results, it is important to use them with caution. Since in optimal design approaches the training set is constructed based on the information from the whole set, it can be regarded as introduction of a bias to the training set.

Test set validation is an easy and universal validation strategy. One of the variations of this strategy would be to construct a model on the data from one source, and validate on the data from another, unrelated source.

The negative side of test set validation is that only part of the data is used to build the model (hence, the inevitable loss of information), and that there's no estimation of model performance on the data that is used for model creation. To avoid these problems one can use cross-validation of bagging validation strategies.

2.7.2 Cross-validation

In the cross-validation strategy, the data is randomly divided into N folds. The modeling procedure is then repeated N times. During each modeling procedure one fold is used as a validation set, and the rest $N-1$ are combined to form a training set. Each of the N available folds is used once as a validation set and $N-1$ times as a part of a training set.

Performance measures are calculated based on the prediction values from validation folds. In cross-validation the model accuracy on the whole data is considered. In QSAR studies the N -fold cross-validation is the most popular validation strategy.

A special case of an N -fold cross-validation is the leave-one-out (LOO) strategy. In leave-one-out one item is excluded from the training set and forms a "validation set". The modeling procedure is repeated for each excluded item.

2.7.3 Bagging validation

Validation can also be performed using a bagging procedure, described in detail in section 2.5.6 (page 28). As described earlier, for each individual run a training set is formed as a bootstrap replica of the initial training set, i.e. the individual training set is formed by resampling with replacement. The individual training set then contains approximately 63% of unique instances from the initial training set. This means the rest 37% of instances can be used as a validation set for this run.

If we run the procedure multiple times (the bagging approach for models in this study used 100 individual models in a bag) the chances are high that each molecule in the initial dataset will appear in the validation set at least once. If the molecule appears in multiple validation sets, the prediction results are averaged.

2.7.4 General considerations

The described validation methods (external validation set, cross-validation and bagging validation) can be combined. Although cross-validation is considered to be the most popular validation strategy in QSAR studies, the bagging validation is favorable in some cases since it

- combines the ensemble model creation and validation; ensemble models have shown to have a higher average accuracy compared to single models.
- it provides multiple predictions for each compound; multiple predictions can be used to calculate the statistical information for estimating the applicability domain of the model (see section 2.9, page 34)

Bagging validation, however, has one disadvantage - it requires significantly more computational power. The bagging model with 100 individual models in a bag would, naturally, require 100 times more computational power to create and to apply to new compounds.

Regardless of the validation method it is important to avoid placing the same compound in the training and validation folds, as it will lead to over-optimistic model performance evaluation (i.e., over-fitting).

2.8 Model comparison

A correct comparison of two models' performance is a non-trivial task. The direct approach would be to just compare the selected performance measure values (RMSE, MAE, MCC, etc.) for two models on the same validation set. The model performance values, however, are obviously dependent on the contents of the validation set. That's why when declaring that one model has higher performance than the other, we also have to check whether this performance difference is not caused by a mere chance, but is significant in the statistical sense.

With this formulation the task of comparing the performance of two models using a given performance measure on a given validation set becomes a classical statistical hypothesis testing problem. The result (i.e., difference in model performance) would then be called statistically significant if it is unlikely to have occurred by chance alone, according to a predetermined threshold probability, the significance level.

A general approach to hypothesis testing is choosing one outcome of the test as a default, i.e. null hypothesis. For example, the null hypothesis in model comparison would be "two models have same performance". Then a p-value calculated based on the type of the statistical test performed. A p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. When the p-value is less than a predefined threshold (usually 0.05 or 0.01), the null hypothesis is rejected. When the null hypothesis is rejected, the result is said to be statistically significant. In our case models were considered statistically significantly different with significance level of 0.05

Two general types of tests are parametric and non-parametric. The parametric tests

(e.g., t-test) are usually based on some assumptions on the observed statistics distribution (most commonly the statistics is assumed to have a normal distribution). The p-value is then often calculated as a function from the statistics distribution parameters. The non-parametric tests (e.g., bootstrap test, Wilcoxon test) involve fewer assumptions and can be used universally.

In this work the non-parametric bootstrap test was used.

The bootstrap test involves generating N (in this work $N = 10000$) test set samples from the original test set by resampling with replacement. Then the tested models are applied to these generated test sets and their performance measure (e.g., RMSE for regression models or MCC for classification models) is compared in a pairwise manner. If one of the model has better performance than the other in 95% of all tested cases, we claim that this model has better performance with significance level $p = 0.05$.

2.9 Applicability domain methods

2.9.1 General concepts

The first general definition of the applicability domain was given as following [159]: “*The applicability domain of a QSAR model is the response and chemical structure space in which the model makes predictions with a given reliability*”.

Obviously, a QSAR model is created on a limited subset of chemical space and can not have the same level of accuracy and predictive ability for every molecule. Thus, it is very important to distinguish reliable and non-reliable predictions: the former predictions can be used in place of experimental measurements while the latter ones should be tested in experiments. The assessment of the applicability domain of the model generally involves a more general task - estimation of prediction accuracy for a given compound by a given model.

Typically, to assess the prediction accuracy, a QSAR model is validated using a validation strategy of choice and the average model performance metrics on this set is reported as the ultimate indicator of the model performance. This approach, however, does not reflect the actual model performance and is misleading for diverse datasets, since model accuracy within the validation set is inhomogeneous. Molecules similar to the ones present in the model's training set are likely to have a better-than-average prediction accuracies, while molecules holding some distinctive features different to the model's training set will probably be predicted with lower accuracy. In applicability domain studies the goal is to individually estimate the prediction accuracy for every predicted compound. The compound that have accuracy higher than some predefined level can then be considered to be “in the applicability domain of the model”.

The key concept used 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 [160]. Distances to models were used in several QSAR studies to estimate the AD of predictive models [160–163]. Distance to model is a general, abstract concept. It is considered that the compounds that have higher DM values are “far from the model” and thus have a lower average accuracy of predictions, while compounds with small DM values are “close to the model” and are more likely to be predicted accurately.

The traditional approaches to DM calculation include descriptor-based DMs [164–166]. The example of descriptor-based DM is leverage:

$$LEVERAGE(J) = \vec{x}(J) \cdot (X^T \cdot X)^{-1} \cdot \vec{x}(J)^T ,$$

where $\vec{x}(J)$ is a vector of molecular descriptors for the compound J , X is a matrix of descriptors for compounds from the training set. Higher leverage values indicated bigger distance of the predicted compound J from the model training set in descriptor space. Often a threshold is chosen, and compounds with leverage values bigger than the threshold are considered outside of the model's applicability domain.

One of the descriptor-based DMs is Tanimoto similarity index. It's distinction from the other descriptor-based methods is that it does not take into account the descriptors of the model itself, rather derives own fragment-based descriptors directly from molecule structure. It is calculated as following:

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

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 . 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.

In this work, however, prediction-based DM approaches are used. The prediction-based approaches were shown to have better results at discriminating accurate and inaccurate predictions [163].

2.9.2 Prediction-based DM measure for classification tasks

As in this work only binary classification problems were addressed, the DM measures were specific to the binary classification. For calculation purposes the “negative” or “non-active” class was assigned a numerical value “-1”, and the “positive” or “active” class - a numerical value “+1”. Given a sample of predictions for a particular compound by the

ensemble of models, we can calculate standard statistical measures for this sample - mean value, standard deviation, etc.

One possible measure based only on the mean value of model predictions is “rounding effect” or CLASS-LAG. Since the mean values of predictions is numeric, it will have to be rounded to the nearest label (-1 or +1) 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. The absolute value of difference between the mean prediction value and the nearest of the labels can be used as a DM. This measure is calculated as follows:

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

Another measure, similar to STD DM for regression tasks, would be concordance of ensemble predictions. The concordance of the ensemble can be defined as the biggest percent of models in the ensemble that give the same prediction. The opposite value (1-concordance) could serve as the DM. (The bigger concordance is, the smaller the distance of this compound to the model is, the bigger prediction reliability is). The DM measure may be defined, or example, as following:

$$d_{INCONCORDANCE}(J) = \min\left\{0.5 - \frac{\bar{y}(J)}{2}, 0.5 + \frac{\bar{y}(J)}{2}\right\}$$

Another DM called PROB-STD [163], combines the uncertainty related to rounding of predictions and the uncertainty the disagreement of different models. One can think of this measure as of a Gaussian probability (according to the observations) of the compound to be of a different, than assigned by a majority of votes, class. It is worth noting, that although our set of ensemble predictions is not necessarily distributed normally (and in case if every model in the ensemble returns the numerical values {-1,1} to represent each of the classes, the distribution will definitely not be normal), the probability is calculated based on expressions for normal distribution:

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

where $N(x, y(J), d_{STD}(J))$ is the normal distribution density function with a mean $y(J)$ and a standard deviation $d_{STD}(J)$.

Figure 2.4 displays four charts for four possible situations. Charts **a** and **b** represent the reliable and unreliable predictions for a {-1} class. Charts **c** and **d** represent the reliable and unreliable predictions for a {+1} class. Higher standard deviation values lead to larger areas representing the probability that the prediction is opposite compared to the one determined

by the majority of votes. This leads to higher PROB-STD DM values. The images on the left represent reliable predictions (small area); the images on the right – unreliable predictions (large area).

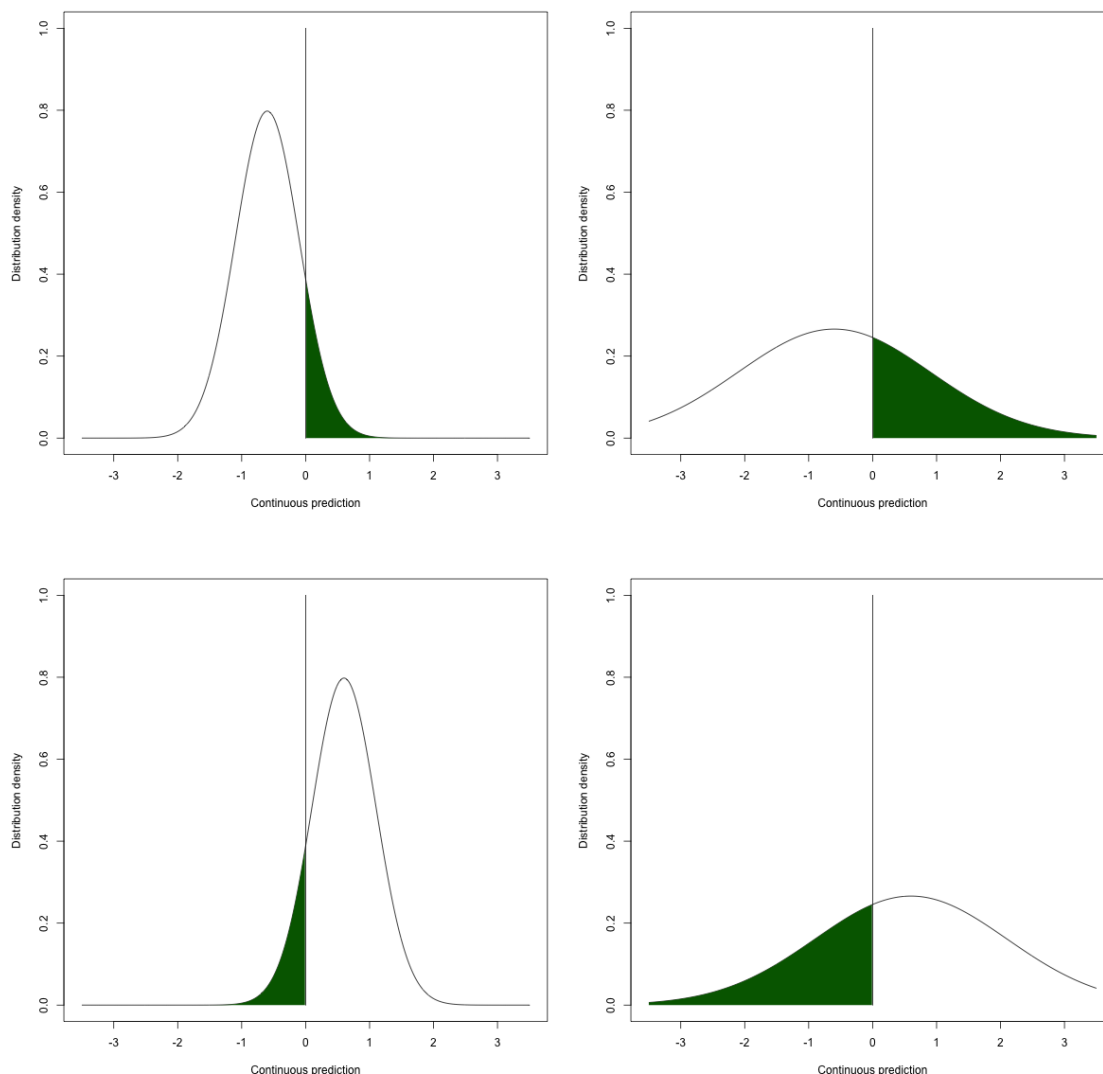


Figure 2.4. Charts a-d display reliable and unreliable predictions for class $\{-1\}$ and reliable and unreliable predictions for class $\{+1\}$ respectively. Green area represents PROB-STD DM measure (larger area for higher uncertainty, smaller area for lower uncertainty).

2.9.3 Analysis of model performance with applicability domain approach

The traditional approach in QSAR studies is to publish the average model performance (RMSE, MAE, classification accuracy or MCC) for the test set or for the cross-validated model. Sometimes the appropriate model performance measure is used to calculate confidence intervals for the whole model, and all model predictions are appended by these confidence intervals. This approach is not very illustrative and does not reflect the whole

information on model performance, since different classes of compounds (and even different compound within each class) have different prediction accuracies. For some compounds the estimated accuracy may be comparable to experimental accuracy, while for other compounds the prediction could have the reliability of a random guess. It is important to separate these compounds and to build confidence intervals based on this information about these compounds.

The DM measures described above give us some information on model prediction reliability. It is thus important to calculate the average estimated accuracy for a compound or a group of compounds based on this information.

The sliding window averaging (SWA) approach involves choosing a sliding window size N and averaging the accuracy on N adjacent compounds sorted by some particular DM. The resulting value gives a SWA estimate of prediction accuracy for a middle compound in the window. The window is then shifted by one compound and the averaging is repeated to get the accuracy estimate for the next compound. The SWA accuracy plot is a useful tool for assessing DM-based estimated accuracy for a compound and generally contains SWA accuracies (or errors) for compounds plotted against their DM values. For a successful DM measure the SWA accuracy plot should have an overall downward (for accuracies) or upward (for errors) trend. Figure 2.5 displays an example of SWA accuracy plot for a CYP classification model with BAGGING-STD DM.

A different approach to assessing prediction accuracy is cumulative averaging, the accuracy is averaged over all the compounds with DMs less than a particular (variable) threshold. The DM threshold is often given implicitly in the form of a percentage of the dataset that have DM values less than this threshold. The resulting values plotted against the percentage thresholds result into a cumulative averaging accuracy plot. This plot can easily display the average accuracy for top 10% best predicted compounds, for example. This cumulative averaging is easily interpretable and very stable against noise. However is strongly depends on the diversity of the set and its similarity to the training set of the model. Figure 2.6 displays an example of the cumulative averaging accuracy plot for the same three CYP classification models.

The provided examples display classification accuracy for a classification task. However, the concept stays the same for regression tasks. The only difference is that for regression tasks the error measure (like RMSE or MAE) is used instead of accuracy measure.

The functional dependency between the DM values and corresponding SWA estimated accuracy for the training set is the basis for estimating prediction accuracies for new predicted compounds. To obtain an accuracy estimate for a new compound, we calculate its DM value and determine the corresponding accuracy using SWA plot. 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.

The details about AD assessment approaches used in this study can be found in a comprehensive methodological work [167].

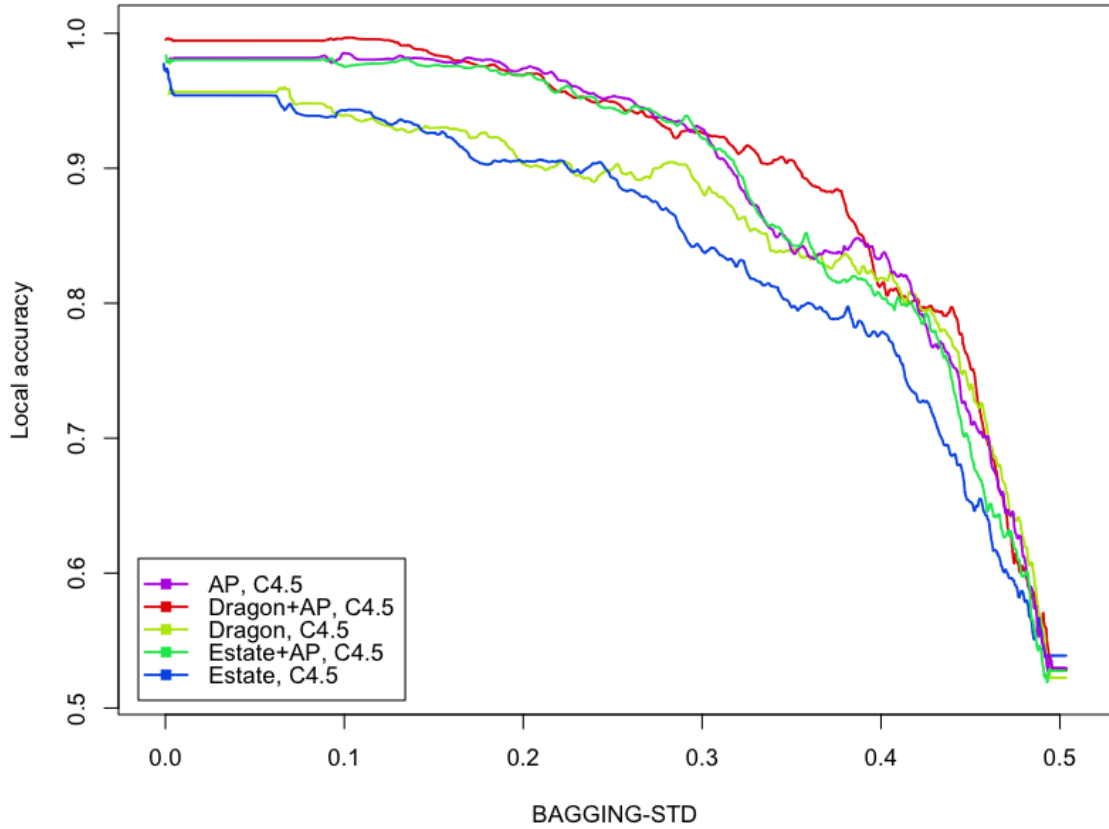


Figure 2.5. SWA accuracy graph for a classification model

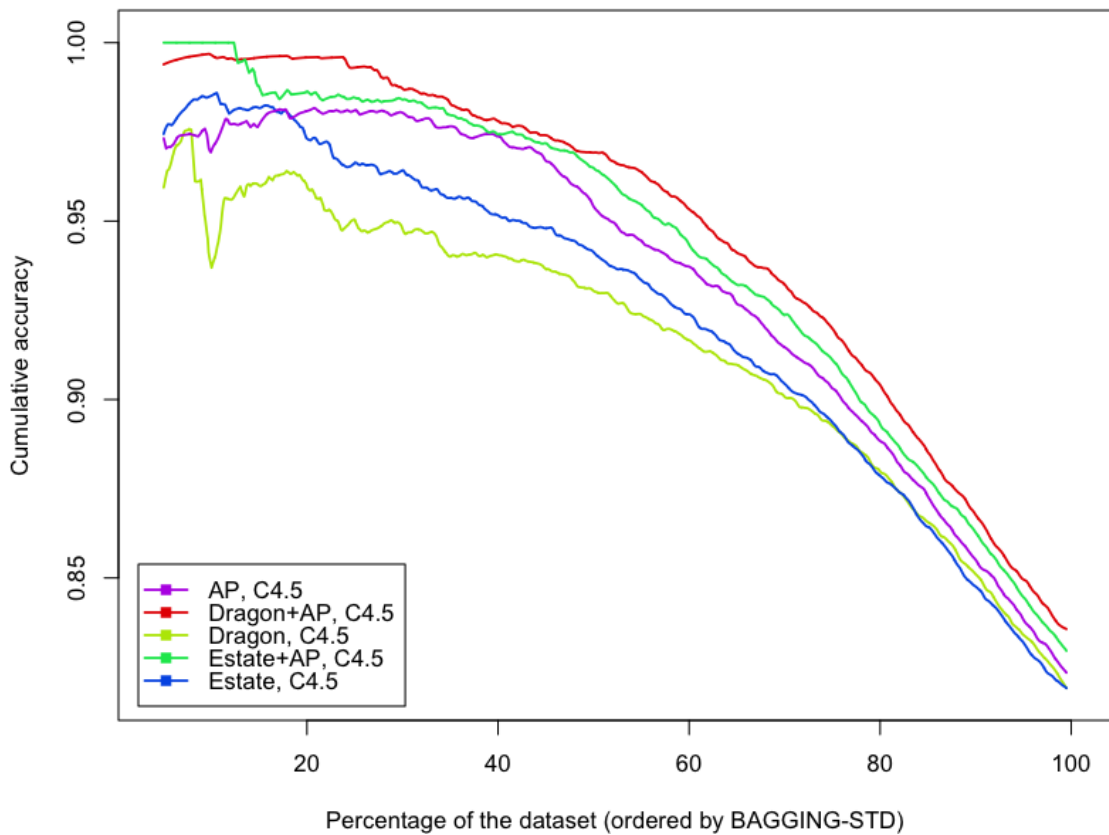


Figure 2.6. Cumulative accuracy averaging graph for a classification model

2.10 Summary

QSAR is a methodology that combines computational chemistry, statistics and machine learning methods to build predictive statistical models for chemical or biological properties of small molecules.

In QSAR studies small molecules are represented in one of the machine-readable formats: SMILES, SDF, MOL2 or InChI/InChIKey. All these formats are based on representation of a molecule as a non-directed graph. As a result of requirements of the studies in this thesis, SDF molecule format is used in all experiments described in the manuscript.

The atomic 3D coordinates most probable for the molecule bioactive confirmation should be determined before any 3D molecular information can be used in further studies. Numerous molecule conformation and sampling methods exist, among which deterministic and stochastic methods optimizing force field target functions, empirical or semi-empirical methods, etc. In studies in this thesis rule-based empirical optimization tool Corina is used.

Molecular docking is a field of computational chemistry that aims to predict the correct binding conformation of a small molecule in a binding site of a protein. Based on the search algorithm the methods are divided to molecular dynamics methods, Monte Carlo simulations, genetic algorithms, fragment-based docking, etc. In the studies in this section the docking is used to obtain protein-ligand conformations which are then used to calculate protein-ligand atom-based descriptors. The docking tool used in the studies is Autodock Vina.

The molecule conformations obtained from 3D optimization or docking are then used to calculate numerical features that represent some molecular properties – descriptors. The descriptors used in the studies described in this manuscript are ISIDA SMF descriptors, E-state indices, Dragon descriptors and novel protein-ligand atom pair descriptors.

Machine learning methods used in the studies are K-nearest neighbors, ASNN neural networks, Support Vector Machines, Random Forest and C4.5 decision trees. Bootstrap aggregating is used to increase model accuracy and for applicability domain purposes.

The validation of the models is performed via N-fold cross-validation, bagging validation and external set validation. Since this manuscript focuses on classification tasks, model quality is assessed by calculating sensitivity and specificity measures, accuracy and balanced accuracy, as well as Matthews correlation coefficient.

Applicability domain methods in model prediction space are used to estimate model prediction accuracy. The methods include calculation of BAGGING-STD distances to model.

3 OCHEM – The database of experimental measurements and modeling environment

This chapter describes the Online Chemical Modeling Environment project [168] (OCHEM, <http://ochem.eu>). It is a database of experimental measurements of physicochemical and biological properties of compounds integrated with the powerful QSAR modeling framework. OCHEM was used as a research and development tool for all the studies presented in this work.

The author of this work made an essential contribution to the OCHEM project development. The contribution involves design of the general concept and the modeling framework API, implementation of data integration tools, and development of a major part of the database system and data processing nodes of the modeling environment. All the described novel descriptor calculation methods were implemented and tested as parts of the OCHEM environment.

3.1 Motivation

A typical process of QSAR modeling involves several highly repetitive, time-consuming steps. Automatization of these steps is the key to shifting researcher's attention from routine steps of data preparation and management to model interpretation, applicability domain assessment and outliers' study.

One of the most time-consuming parts of building any QSAR model is data preparation. Most of the available experimental data can be found in published scientific articles. Most of the articles contain only one or several measurements on a small group of closely related compounds. This makes the process of collecting a chemically diverse dataset quite a tedious task. Additionally the measurements are often performed under different experimental conditions and the resulting values are reported in different measurement units. Careful comparison of the compatibility of measurement conditions and transformation of the values to the same measurement unit may be a challenging task for a QSAR researcher.

Employing a consistent and well-structured database of experimental measurements of molecular properties can minimize all these steps. The database may help find the experimental properties relevant to the topic, given they were previously uploaded by another scientist. Even if a researcher has collected a specific dataset of interest, any additional data may be used for validation of the developed model. The embedded tools may also automatically perform the routine tasks of unit conversion, duplicates control and filtering.

Another issue is model usability and reproducibility. Hundreds of QSAR models are published in scientific journals every year. More than 50 models published only for lipophilicity, logP, and water solubility in 2005 [169,170]. For most of these models the life cycle ends with a publication in a journal. Only a small share of the published models is maintained as standalone tools and is applied to new compounds. While being useful as proof-of-concept for the methods or descriptors used in the study, this type of models is useless for further study.

An attempt at reproducing a published model often meets several specific problems. One of the difficulties is reproducing the initial dataset. In a vast amount of cases, the dataset is not published with the model. Often the published dataset contains molecular names or structure depictions, both of which are ambiguous and prone to human errors. Many models involve a complicated procedure of selecting the training and test sets from the initial dataset, and unless the exact contents of these sets are explicitly published, the models become irreproducible.

The extreme variety of (often only commercially available) machine learning methods software and molecular descriptors calculation software make reproducing a published model yet more difficult. Even for the same tool the values are very dependent on the tool version, which makes it impossible to reproduce the older models, built with older versions of tools.

A central repository that would provide access to a wide variety of machine learning tools and descriptor calculation software and allow to store the data, the model and the model's protocol in the same place would allow to publish reusable and reproducible models.

There's a wide variety of online databases and modeling tools available, none of which resolve all of the mentioned problems. Some databases (PubChem, ChemSpider, DrugBank, ChemExper [171–177]) fulfill the role of storing the chemical data and providing tools for navigating this data. However, these solutions lack some important functionality for QSAR data preparation (unit conversion tools, consistent experimental conditions, etc.) The mentioned sites also lack any modeling capability and thus can only be used as a data source on the early stage of model development. A number of modeling tools are also available online [178,179] but are too generic and incapable of supporting a complete QSAR model development pipeline.

This chapter presents a unique online tool – OCHEM, the Online Chemical Modeling Environment that addresses all of the mentioned problems. The OCHEM consists of two essential parts – the database of experimental measurements, and the QSAR modeling framework. The database allows search, manipulation, upload and download of QSAR data. It contains tools to convert units, remove duplicates, form separate datasets based on experimental conditions, and perform other steps essential for preparing a good dataset. The modeling framework is integrated with the database and allows using the prepared datasets as input to the QSAR modeling workflow. The user has the possibility to choose from numerous machine learning methods and meta-learning techniques and dozens of descriptors to build the models even for the most demanding properties. The resulting models can be made available either for peer review or publicly, which encourages people to use or reproduce them, thus extending their life cycle.

3.2 Database of experimental properties

3.2.1 Structure overview

The central concept of the OCHEM database is *experimental property* or *record*. The experimental property represents a single measurement of a particular property for a particular molecule under defined conditions, and published in a defined article. This single record represents a single data point in a QSAR dataset. The simplified structure of the *experimental property* is shown on Figure 3.1.

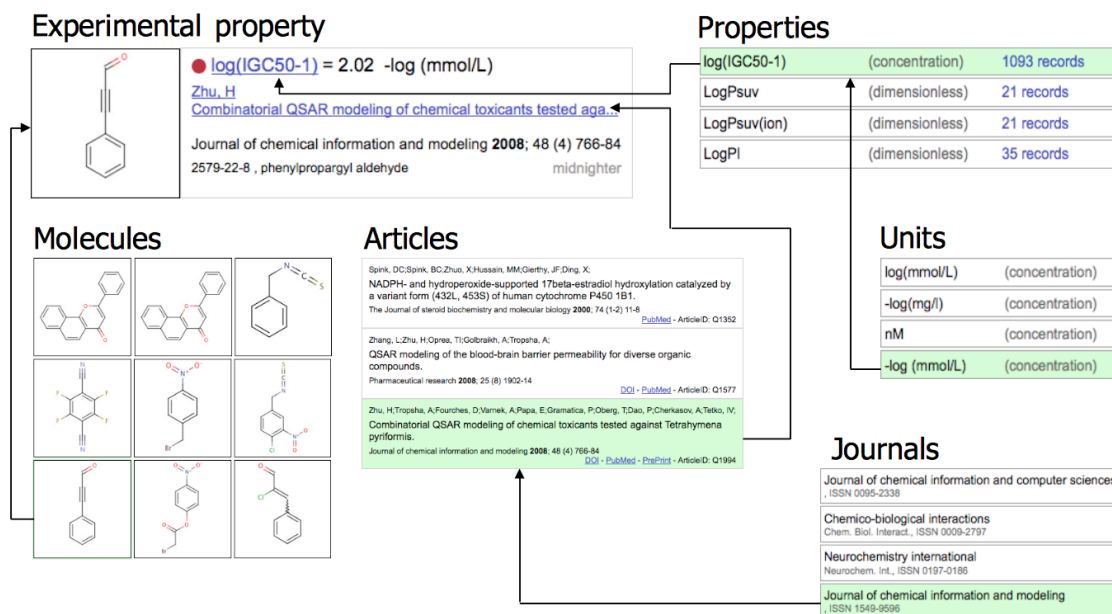


Figure 3.1. A schematic overview of the *experimental property* structure

The *experimental property* itself is the central entity of the database and represents a measurement value. The value may be numeric or represent a qualitative observation.

The *property* represents a physical, chemical or biological property being measured. The property may be quantitative – then an exact numerical value of the measurement is provided, or qualitative.

For qualitative properties one must provide a list of available options. For example, for qualitative “DMSO Solubility” property only two options are available “soluble” and “insoluble”. This approach to solubility is useful when the studied fact of interest is solubility of a particular compound above some defined threshold.

A *measurement unit* must always complement a numeric value for a quantitative property. The units are grouped by categories. For example, the units *mol/l*, *mg/l*, *g/cm³* all belong to “concentration” category. Main categories available in OCHEM database involve temperature, concentration, speed, time, dose, pressure, density, etc. The database policy is to store the unit exactly the way it was published in the referenced article. However, within one category units can be automatically converted for modeling purposes. For quantitative properties a unit category is associated with the property. For example, the property “Melting point” has a “temperature” unit category associated with it. The user may provide

data points in degrees Celsius, Kelvin or Fahrenheit.

Since QSAR data reflects measurements for particular molecules, information about molecule structure is an essential part of this data. Each *experimental property* contains information about the *molecule*, for which the measurement was performed. The OCHEM supports rich facilities for describing a molecule, which involve common molecule formats (SDF, MOL2, SMILES), molecule names, and possibilities to draw a structure in a molecular editor.

Both *property* and *molecule* can be marked by a set of *tags*. Tags provide a non-hierarchical way to classify data according, for example, to tasks or areas of interest. Several thousand molecules in the OCHEM, for example, are marked with “ChemBridge” tag. This represents, that these molecules are available in form of DMSO solution from ChemBridge provider (as mentioned in tag description). This may be useful information for someone performing virtual screening before buying compounds for actual laboratory experiments.

Another important part of the *experimental property* entity is the article (or more generally – the source), where the value was published. The OCHEM database policy requires a user to provide the reference to the article, from which the value was obtained. This allows for a better data quality control. In case of any suspicion (for example, if a data point represented by a particular experimental property is a distinctive outlier in a model) the user of the data has the possibility to check the original article for the correct value.

One of the features that distinguish the OCHEM from other chemical databases is the possibility to store *conditions of experimental measurements*. It is clear, that values of the “Boiling point” property are dependent on the pressure, under which they were measured. Similarly, the “Solubility” values depend on the temperature of the solution. Thus it is very important to store this kind of information with the measurement to be able to form a consistent dataset.

3.2.2 Data search and editing

Browsers are an important concept in the OCHEM database. It is a special dialog that is associated with almost every meaningful entity in the database. Currently there are specific browsers created for experimental properties, molecules, properties, conditions, units, articles, journals, molecule sets, tags and models.

Every browser generally consists of the *viewing area* that displays a page with a list of items and the navigation bar that allows changing pages, and a set of *filters* that allows narrowing the set of items displayed in the viewing area. By default, the filters are empty and the viewing area displays all available items (with user-specific restrictions applied). Every item in the list holds some functional elements (buttons, icons or clickable links) that can be used for editing or deleting this item, obtaining some additional information about the item, or navigating to some other browser in the database that is relevant to this item. Figure 3.2 displays a browser for properties with a name filter applied (the items area only display properties that have “con” in their names). Every property in the list has the “edit” and “look-up in wiki” icons and the clickable link that will open an experimental property browser with this property’s records.

The screenshot shows the 'Properties browser' interface. At the top, there is a search bar with the text 'con' entered. Below the search bar, there are navigation options: '1 - 10 of 11', '10 items on page', and '1 of 2 > >>'. The main content area displays a list of properties, each with a checkbox, a name, a description, the number of records, a brief description, and a list of contributors. The 'Lethal Concentrations' property is highlighted in yellow.

Property Name	Description	Records	Contributors
Atmospheric NO3 Rate Constant	(Rate reaction constant 2nd order)	2 records	henri / mojca
Atmospheric O3 Rate Constant	(Rate reaction constant 2nd order)	2 records	henri / itetko
Atmospheric OH Rate Constant	(Rate reaction constant 2nd order)	781 records	itetko / mojca
Bioconcentration factor	(Concentration)	66 records	boris / mojca
BioConcentrationFactors (Dimensionless)		7 records	Linus
Critical micelle concentration (CMC)	(Concentration)	26 records	wolfram / simona
Fishtine Constant KF	(Dimensionless)	22 records	wolfram
Lethal Concentrations	(Concentration)	1590 records	wolfram
Log Henry's law constant [Pa*m ³ /mol]	(Henry's Law Constant)	58 records	wolfram / mojca
Medium of Concern	(qualitative)	0 records	i.tetko

Figure 3.2. Properties browser

The main browser of the OCHEM database is the **experimental properties browser**. It is the biggest, most function-rich and the most complicated browser in the system. It allows searching and editing of experimental measurements and grouping them into sets (that can be later used as training or test sets in the modeling environment). The available filters allow narrowing the scope of the displayed measurements:

- *Property filters* make most sense, since a QSAR researcher is generally interested in a single property or a group of closely related properties
- *Molecule filters* allow finding records for one specific molecule (by filtering by a molecule name or InChI key) or for a group of molecules (by filtering by a subfragment or a range of molecular weights) and are useful for researches interested in specific families of compounds (e.g., triazoles).
- *Article filters* allow specifying the desired source article for the records.
- *Filter by experimental conditions* is very important to form a consistent dataset for QSAR modeling
- *Additional filters* allow selecting records from a specific set, displaying error records or records with wrong names only, selecting only records originally measured in the articles, etc. These filters are created to help with data quality control and data set creation and manipulation.

As it is very important to keep the data consistent, the OCHEM implements a set of rules to identify duplicate data. Two types of duplicate records are considered in the system: *strong duplicates* and *weak duplicates*. Two experimental measurement records are

considered strong supplicates if they describe measurements that are performed for the same property and for the same molecule under same conditions, published in the same article and have the same values (with precision to 3 significant digits, given the values are converted to the same measurement unit). Strong duplicates are not allowed to hold a “valid” status; one of the duplicates should be explicitly marked as “error”. Weak duplicates are records that share only part of the data (e.g., property and molecule). These records are allowed, but the experimental properties browser has the tools to search for such records.

A separate topic of consideration is duplicate control among molecules. The OCHEM implements the industry standard method of molecule duplicate control – control by InChI key. InChI key is an IUPAC-developed fingerprint hash [70–72]. Two molecules are considered same if their InChI keys are same.

3.2.3 Data introduction

Since the OCHEM is a user-contributed platform, it is essential to provide the user with simple and efficient tools to introduce data into the system.

Experimental property browser provides the single record edit tool. It allows modifying every aspect of the record and is useful for data correction or single record introduction, but becomes impossible to use when the introduction of hundreds or thousands of records is required.

For efficient and fast introduction of large amounts of data, the OCHEM includes the “mass data introduction tool” or the “batch upload” tool. The input data for the tool is a specially prepared Excel workbook, CSV or SDF file.

The preferred file format for the batch upload tool is Excel file. The example Excel file with all possible columns and explanations can be downloaded directly at the first page of the batch upload tool. As described earlier in section 3.3.1 (page 48), the essential information, contained in the record, is a value of a biological or chemical property for a specific molecule, published in a specific article. Although the Excel file format allows providing all the detailed information about the record (number of a page in an article, where a particular value was published, accuracy of measurements, textual comments to the record, record evidence, etc.), the minimal valid file should contain information on property value, molecule structure and article for every uploaded record. In case if some information is not provided (i.e. unit of measurement), the default values are taken.

The number of features makes record uploading easier. For example, information about the molecule structure can be provided in form of SMILES, SDF or MOL. If the structure of the molecule is not available, it is possible to provide a molecule name or CASRN [180] – the tool will make an attempt to fetch the structure from PubChem automatically. The article can be provided either in form of internal OCHEM article identifier or PubMed [181] identifier. The sheet can also contain information about the measurement conditions. For proper work of the data upload, the information about the property itself and all the required conditions and units should be already present in the database. For numeric properties the user can provide predicates, such as >, <, ≥, ≤, ~, >>, <<, ≈.

After the file has been created, the user can use the batch upload tool to introduce data to OCHEM. The tool is created in the form of a wizard with a step-by-step approach to the upload process. The wizard will highlight all the potential data problems (like unrecognized property or unit name, unknown molecule, duplicate data, etc.) and provide previews and reports on the data upload process. Figure 3.3 displays a screenshot of a preview page of the data upload wizard.

Batch upload browser
Please review your data and confirm it, either modify it in your file and upload it again.

Information

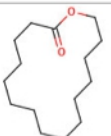
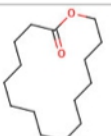
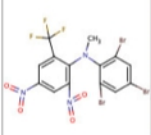
Column (C) "NUM" was not recognized and therefore it has been ignored

Column (A) CASRN recognized correctly
Property at (B) LD50 recognized correctly
Column (D) PAGE recognized correctly
Column (E) ARTICLEID recognized correctly
Column (F) REFERENCE recognized correctly
Condition at (G) Administration route recognized correctly
Condition at (H) Organism recognized correctly

Number of errors: 1
Number of empty structures: 0
Number of structures attempted to get from PubChem: 0
Number of structures from QSPR: 28

SHOW ALL SHOW ONLY VALID SHOW ONLY ERRORS SHOW EXTERNAL DUPLICATES SHOW INTERNAL DUPLICATES

1 - 5 of 28 5 items on page 1 of 6 > >>

	<p>● LD50 = 5.0 g/kg/once = 5000 mg/kg/day</p> <p>Hexyl salicylate... P: 787 Food and Cosmetics Toxicology 1975; 13 (6) 807 - 808 106-02-5 Row 2</p>	<p>Organism = rabbit Administration route = skin</p> <p><input type="radio"/> Don't save <input checked="" type="radio"/> Save as new</p>
	<p>● LD50 = 5.0 g/kg/once = 5000 mg/kg/day</p> <p>Hexyl salicylate... P: 787 Food and Cosmetics Toxicology 1975; 13 (6) 807 - 808 106-02-5 Row 3: This item is a duplicate of another item in the uploaded dataset (Row 2)!</p>	<p>Organism = rabbit Administration route = skin</p> <p><input checked="" type="radio"/> Don't save</p>
	<p>● LD50 = 1.0 g/kg/once = 1000 mg/kg/day</p> <p>Worthing, C. The Pesticide manual: a world compendium... P: 96 1991; 63333-35-7 Row 4: This record is a duplicate of another PUBLIC record, already existent in database! [show]</p>	<p>Organism = rabbit Administration route = skin</p> <p><input type="radio"/> Don't save <input type="radio"/> Save as duplicate <input type="radio"/> Overwrite</p>
Row 5: Some obligatory conditions for property LD50 have not been specified: [Organism]		
Row 6: Invalid value for property "LD50": abc		

1 - 5 of 28 5 items on page 1 of 6 > >>

DON'T SAVE ALL SAVE AS NEW ALL OVERWRITE ALL

CONFIRM AND SAVE TO DATABASE RETURN TO SHEET PREVIEW RETURN TO FILE UPLOAD [DOWNLOAD XLS] Save Batch File

Figure 3.3. A preview page of the data upload wizard. Erroneous records are highlighted.

3.2.4 Typical OCHEM usage scenario

If a QSAR researcher has some data of his own (collected from literature or obtained from some other database, for example), he is most likely to start with data introduction. With the help of appropriate browsers he will introduce a source article for his data, property and unit (in case they are missing). After that he will form an Excel sheet with his data and

use the batch upload tool to introduce his data to the OCHEM. He may mark his data “private” while doing so.

Once the data is in the database, he is likely to create a training set with the uploaded records, by using the appropriate experimental properties browser tools.

Additionally the researcher would most likely search for any additional relevant data to form an external validation set. That can be achieved in the experimental properties browser by filtering by relevant property, measurement conditions and (possibly) molecule fragments. After reviewing any appropriate data and using the duplicate and error control tools to avoid redundancy in it, the researcher can create a data set, which can be used as an external validation set.

Once the training and validation sets are ready, they can be used as input parameters in the modeling part of the OCHEM system.

3.3 Modeling framework

3.3.1 Overview

The main goal of the OCHEM environment is to reduce the amount of work a QSAR researcher should perform to obtain a predictive model.

The preparation and management of the data, filtering, grouping as well as storing and reusing of the data can be performed by the database part, described earlier. With the help of appropriate tools a researcher can prepare training and validation sets for further use in modeling.

Modeling framework is an essential part of the OCHEM. It functions in a tight integration with the database and provides rich tools for model creation and analysis. It is built in a form of a step-by-step wizard, each step of which allows modification of different settings of the model-building framework.

The OCHEM software is built to handle wide range of available data. To suit this purpose, typical time-consuming steps (like conversion of all available data to the same measurement unit) are taken care of. The OCHEM supports both regression problems (prediction of the numerical values of properties, e.g., solubility in mg/l) and classification problems (prediction of qualitative properties, e.g. inhibitor/non-inhibitor, mutagenic/non-mutagenic, etc.) The OCHEM also supports simultaneous prediction of several properties, i.e. multilearning [182].

A rich selection of descriptors calculation software allows choosing the descriptors best suitable for a specific task. In addition to popular and industry-standard software, the OCHEM provides access to a wide variety of descriptor calculation tools developed by scientific groups all over the world. Some tools (like protein-ligand interaction based descriptors, which are an implementation of the approaches described in the General methodology section (page 7) of this work, or atomic-based descriptors) are a result of research in the OCHEM group.

A special parallel computation back-end suitable for coarse-grain parallel tasks typical for QSAR modeling was developed. The central MetaServer node governs distribution of computational tasks and collecting results for a computational cluster of over 300 cores.

Finally, the OCHEM provides tools for analysis and modification of the models. These tools allow AD assessment and prediction accuracy estimation, outlier removal, fragment-based model interpretation, etc.

3.3.2 Dataset, machine learning method and validation method selection

The modeling framework was built with a standard QSAR model building workflow in mind. The wizard takes the user through typical steps of dataset selection, selection and configuration of the appropriate machine learning method, selection and configuration of the descriptor calculation software, configuration of descriptor selection block, selection of the appropriate validation (or meta-learning) method.

The wizard starts with the first step necessary for any QSAR modeling – the selection of the training and the (optional) validation sets. The sets may contain one or several related properties, both quantitative and qualitative. The content of the sets defines the applicable machine learning methods. For datasets containing quantitative properties the OCHEM supports automatic unit conversion. The modeling wizard page suggests the user to select a default unit for each modeled quantitative property from the appropriate category, and all values from the dataset for this property will be converted to that unit. This allows seamless combining of heterogeneous measurements in one dataset.

Based on the contents of the dataset the researcher is presented with the choice of machine learning methods suitable for the data points in the dataset. Several methods (e.g., Weka implementation of C4.5 decision tree – WEKA-J48) can only handle classification data, and the only method capable of multilearning so far is Associative Neural Networks (ASNN).

The machine learning methods currently available are ASNN (ASsociative Neural Networks), FSMLR (Fast Stagewise Multiple Linear Regression), KNN (K-Nearest Neighbors), KPLS (Mathematica implementation of Kernel Partial Least Squares), KRR (Kernel Ridge Regression), LIBRARY (a model-based local correction method), MLR (Multiple Linear Regression), LibSVM (Support Vector Machines implementation) and WEKA-J48 (Weka-based implementation of C4.5 decision tree). These methods are reviewed in detail in section 2.5, page 23.

Model validation is an important part of any model-building process. A correct validation approach can ensure that the model is not overfitted and is not prone to some bias (e.g. descriptor selection bias). A missing or incorrect validation procedure may result to misleading over-optimistic results [183,184]. The OCHEM modeling framework supports two types of validation: N-fold cross-validation and bagging validation (described in detail in the section 2.7, page 31).

If the N-fold cross-validation is selected, the training set is randomly split into N folds (the default and most used value for N is 5). Special care is taken, that records with same molecules (disregarding stereochemistry differences) are placed necessarily placed in the

same fold. The whole modeling process is then performed N times, one different fold being used as a validation set, and other N-1 combined – as a training set. This way we receive model performance estimation on the whole set, while avoiding predictions of the training data.

For the bagging validation the modeling process is performed several (by default - 100) times. For each time the training set is formed randomly from the original training set by resampling with replacement. The samples not selected for the training set (on the average 33% of the original training set) form the test set. The predictions of all validation sets of all models are then combined (duplicate results are averaged) to form a model performance evaluation for the initial training set.

The bagging in OCHEM has a modification called stratified bagging. It is useful for highly imbalanced classification dataset. The stratified bagging tries to form balanced training sets from the original training set by undersampling the occurrence of the overrepresented class.

Figure 3.4 displays the screenshot of the OCHEM modeling wizard with the dataset, machine learning method and validation method selection dialogs.

Create a model
Select the training and validation sets, the machine learning method and the validation protocol

Select the training and validation sets:
Training set (*required*): [Thesis_CYP1A2](#) [details]
[Add a validation set](#)

The model will predict this property:
CYP450 modulation using unit:

Choose the learning method:
Suggested modeling methods:

- ASNN (ASsociative Neural Networks) [W](#)
- FSMLR (Fast Stagewise Multiple Linear Regression) [W](#)
- KNN (K-Nearest Neighbors) [W](#)
- Library model (A model based on another ASNN model enriched with new compounds data) [W](#)
- LibSVM wrapper with grid-search parameter optimisation [W](#)
- MLR (Multiple Linear Regression) [W](#)
- PLS (Partial Least Square) [W](#)
- WEKA-J48 (Weka-based implementation of C4.5 decision tree) [W](#)
- WEKA-RF (Weka-based implementation of Random Forest) [W](#)

Models under development. (Do not use unless you are sure how to use):

- ANNC (Molecule-centric, experimental!) [W](#)
- BLASSO (Bayesian regression) [W](#)
- Consensus model (experimental) [W](#)
- KRR (Kernel Ridge Regression) [W](#)

Model validation
Validation method:
Number of folds:
 Stratified cross-validation

You can create a model from template: [import an XML model template](#) or [use another model as a template](#)

Figure 3.4. First page of the OCHEM modeling wizard

3.3.3 Data preprocessing

The user has the possibility to choose some data preprocessing and data handling options before processing to the descriptor selection page.

Data preprocessing options refer to molecular structure preprocessing. The available options are “standardization”, “neutralization” and “remove salts”. The details about the effects of these options can be found in section 2.2.1, page 11.

Data handling options refer to handling of non-typical numerical values. The OCHEM allows storing “greater than” and “less than” values (e.g., melting point > 100 °C), interval values (e.g., solubility = 7 – 9 mg / l), and values with accuracies (e.g., boiling point = 25 ± 3 °C). Since most of the machine learning methods can't use these kinds of values explicitly (the only exception being ASNN method), the user has the choice either to exclude data with these values from his dataset, or to convert them to “equals” values. When converting, boundary values are taken for “greater than” and “less than” data and average values – for interval data and data with accuracies.

3.3.4 Molecular descriptors

The available descriptors are grouped by the software that contributes them: ADRIANA.Code [185], CDK descriptors [186], Chirality codes [187–189], Dragon descriptors [124], E-State indices [127], ETM descriptors [190], GSfrag molecular fragments [191], Inductive descriptors [192], ISIDA molecular fragments [125], Quantum chemical MOPAC 7.1 descriptors [193], MERA and MERSY 3D descriptors [194–196], MolPrint 2D descriptors [197], ShapeSignatures [198] and logP and aqueous solubility calculated with AlogPS program [157]. The descriptors used in this work are described in greater detail in section 2.4 , page 18.

Many of the descriptors also include additional options, e.g., the minimal and maximal fragment length for ISIDA, the individual descriptor block selection for Dragon, Inductive and CDK descriptors, etc. The list also includes experimental Protein-Ligand Interaction-Based descriptors, described in this work (page 22). The descriptor screen is shown in Figure 3.5.

If the property has some important conditions (temperature condition for solubility, pressure condition for boiling point, pH condition for a variety of chemical and biological properties), it is possible to include these condition values in the dataset as descriptors. By using this approach it is possible to build a single consistent model for a dataset of measurements performed under different conditions.

For example, by including pH condition as a descriptor it is possible to combine the whole amount of LogP data measured under different pH into one model. This model would then be able to predict LogP values for different pH values, which widens the applicability area of the model.

Depending on whether there were any 3D-dependent descriptors selected on this step, the user may be asked for molecule 3D structure optimization options on the next step of the wizard. Current options allow the user to skip optimization and stay with the 2D-structure, use Corina [86] tool, or use MOPAC [193] tool for optimization.

Select descriptor blocks

Please select the MOLECULAR descriptors:

- E-state [W](#)
- OESState [W](#)
- ALogPS [W](#)
- AMBIT Descriptors [W](#)
- MolPrint [W](#)
- GSFragment [W](#)
- Dragon (5.4) [W](#)
- Dragon (6.0) [W](#)
- ISIDA fragments [W](#)

Aromatize structures: Chemaxon Basic [W](#)

Fragments from 2 to 5

Type of fragments: Sequences of atoms and bonds [W](#)

- MOPAC descriptors (3D) [W](#)
- ADRIANA.Code (3D) [W](#)
- CDK molecular descriptors [W](#)

Aromatize structures: Chemaxon Basic [W](#)

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

- constitutional descriptors
- geometric descriptor
- hybrid descriptor (3D)
- topological descriptors
- electronic descriptor (3D)

- ShapeSignatures (3D) [W](#)
- 'Inductive' descriptors (3D) [W](#)
- MERA descriptors (3D) [W](#)
- MERSY descriptors (3D) [W](#)
- Protein-Ligand Interaction-Base descriptors (alpha version)(3D) [W](#)
- Chemaxon descriptors [W](#)
- Chiral Descriptors [W](#)
- ETM descriptors [W](#)

Figure 3.5. Descriptor screen of the OCHEM modeling wizard

3.3.5 Descriptors filtering

The OCHEM implements several descriptors filtering tools. The filtering tools are unsupervised, that is, the target property values are not used in the selection process.

The available filtering methods are filter-by-value, filter-by-variance, pairwise correlation based grouping, unsupervised forward selection [199] and principal component analysis.

For the filter-by-value tool the user specifies the minimum number of different values for a descriptor column to be selected. For example, the filter-by-value tool with the value “2” will select all non-constant descriptor columns (i.e., columns with at least two different values). This tool is most often used as a preprocessing tool for the machine learning method implementations that do not filter out constant-value columns themselves (e.g., ASNN tool).

Pairwise correlation based tool removes columns that correlate with coefficient more than some predefined value. The user can define the correlation coefficient threshold on the descriptor selection settings page. If the amount of descriptors makes it computationally infeasible to calculate a full correlation matrix, the descriptors are split into sets and filtering is performed within each set. This leads to a bigger amount of resulting descriptors, but reduces the necessary calculation time. The correlation matrices are calculated for the descriptor sets, and the descriptors are selected based on their correlation values with other descriptors in the same set. The goal of filtering in each set is to obtain the smallest subset of

non-correlating descriptors.

Unsupervised forward selection (UFS) aims to generate a subset of descriptors from any given data set in which the resultant variables are relevant, redundancy is eliminated, and multicollinearity is reduced [199].

PCA is an orthogonal transformation that allows substituting the original matrix of (possibly) correlated descriptors with the matrix of uncorrelated values – principal components. A user can define a variance threshold (all principal components with smaller variance will be removed) and a threshold value for the total number of principal components.

The user can also upload a text file with a list of descriptors he would like to keep. This is useful for a number of scenarios – reproducing of a published model, for example. Figure 3.6 displays a screenshot of the descriptor selection settings.

Select filters of descriptors ?

Eliminate descriptors with less than unique values

Delete descriptors that have absolute values larger than

Delete descriptors that have variance smaller than

Group descriptors, that have pair-wise correlations Pearson's correlation coefficient R larger than

Use Unsupervised Forward Selection to delete variables using the above value of multiple correlation coefficient R

Perform principal component analysis

After filtering, I want to select necessary descriptors myself (advanced)

Normalisation parameters

Descriptors normalization

Values normalization

<<Back Next>>

Figure 3.6. Descriptor selection settings

3.3.6 Machine learning method configuration

The final screen of the modeling wizard allows the user to configure the selected machine learning method. Machine learning methods and their parameters are described in detail in the methods section. This section gives a short overview of the modifiable parameters.

- For **ASNN** the user can modify the training algorithm, number of neurons in the hidden layer of a network, the number of networks in an ensemble, and the maximum number of training iterations. The parameters allow reaching a balance in the “calculation time” vs. “model complexity” vs. “model generalization ability”.
- For **FSMLR** the user can specify shrinkage (influences the generalization ability of the model) and the relative size of an internal validation set (influences model

complexity).

- For **KNN** a user can specify the number of neighbors to use and the distance metrics. Both parameters have a data-specific effect on the resulting model.
- In **KRR** the user can either specify the *Lambda* regularization parameter explicitly or set the parameters for the cross-validation loop that will determine it automatically. The user can also select the kernel type.
- The **LIBRARY** method has no user-customizable settings.
- For **MLR** the user can specify the *Alfa* parameter that regulates internal descriptor selection procedure.
- The **SVM** implementation allows wide customization, involving SVM type, kernel type, and the boundaries for the parameter-optimization grid-search.
- The **WEKA-J48** wrapper inherits all the customizable settings from the Weka implementation.

3.3.7 Model calculation start

Once all the aspects of the modeling workflow are configured, the model can be sent for calculation. The final screen, where the user has the possibility to give a name to his calculation task and (in case of privileged users) specify its priority is displayed in Figure 3.7.

Start calculation of the model

Now we are ready to start calculation.
Please provide the name for your model:

Default priority
 Use privileged servers
 Debug priority

Figure 3.7. Last screen of the modeling wizard

The data is then preprocessed according to user setting and fed to the modeling workflow. The workflow itself is schematically presented on Figure 3.8. This workflow is valid for most of the OCHEM models, except for some special cases (e.g., LogP-LIBRARY method does not require descriptors).

While the workflow itself is consecutive, there are parts of it that can be calculated in parallel. For example, different descriptor calculation tools or different instances of a machine learning method nodes in a validation procedure are calculated simultaneously, if there is enough free calculation power.

Since for big datasets and for specific settings of machine learning methods the model calculation may take hours or even days, it would be inconvenient for the user to wait for the modeling process to finish. That's why every model once sent to the calculation workflow is placed in a *pending tasks registry*. The user can access the registry page to view the status of his model. It is important to note, that a user can schedule several models this way, and if there's enough computational power the models will be calculated simultaneously. The screenshot of the pending task registry is displayed on Figure 3.9.

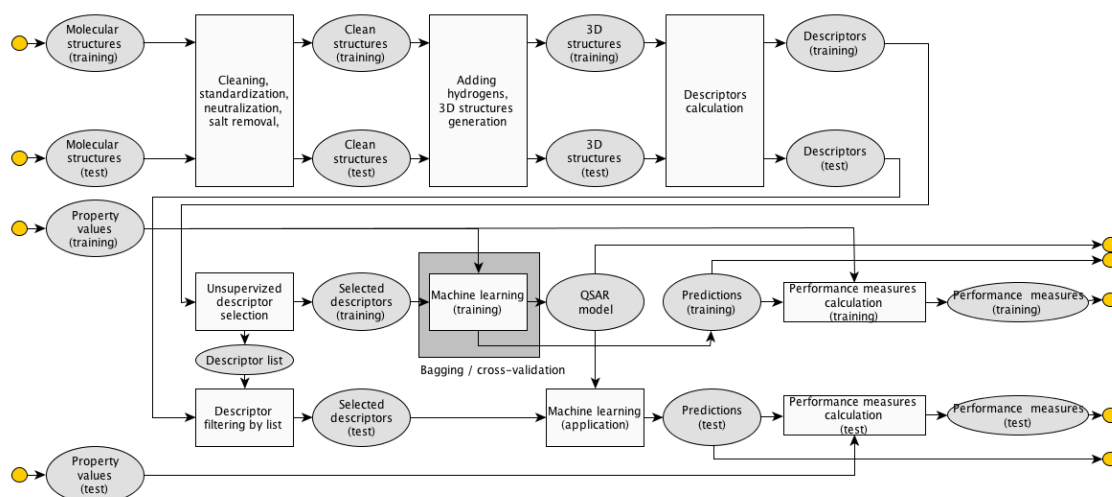


Figure 3.8. Simplified structure of the modeling workflow

Pending tasks						
The overview of all running tasks and all completed tasks awaiting your action						
All tasks [Refresh]						
Time started	Task type	Model	Property	Method	Status	Details
2011-05-04 16:30:47	Model training	CYP2D6 Reference Model stratified	CYP450 modulation	WEKA-J48	assigned	Processing task Selection - De [more>>] terminate
2011-05-02 16:24:55	Model training	CYP450 modulation, 8385	CYP450 modulation	WEKA-J48	ready	- recalculate
2011-04-15 15:03:40	Prediction	CYP450 modulation all unbiased	CYP450 modulation	WEKA-J48	ready	- recalculate

Figure 3.9. Pending task registry page

3.3.8 Tasks management and load distribution

Building a predictive QSAR model is often a calculation-intensive task. Memory, computation power and time requirements may be very high for specific cases.

Descriptor calculation tasks may take up to several hours for an average dataset. If the dataset is large or if flexible docking or molecular dynamics simulations must be performed prior to descriptor calculations, this stage of QSAR modeling may take days on a single machine.

Descriptor selection and filtering tasks involve operations on large matrices and may take hours for a relatively large dataset.

Depending on the machine learning method itself and the settings of this machine learning method, the model training stage may yet be the most computationally intensive task in the whole QSAR modeling process. And when ensemble modeling is involved, the amount of required calculations increases. And ensembles have been shown to produce better modelings results and are also required for ensemble-based applicability domain methods, described in section 2.9 , page 34. A good example would be an average ASNN

model with Bagging validation. One instance of ASNN model is an ensemble of 64 networks. The bag size of 100 instance would lead to training of 6400 individual neural networks.

However, most of the QSAR modeling subtasks can be calculated in a parallel manner - calculating descriptors for individual molecules and training individual instances of models in an ensemble or individual folds in validation procedure can be performed independently. That's why a powerful parallel calculation subsystem is important for a modeling framework.

The OCHEM has an implementation of a parallel calculation system. The calculations may be distributed to over 500 CPUs of the Helmholtz Center's Institute of Bioinformatics LSF cluster and around 20 CPUs of the desktop computers of the members of the group. The central Metaserver node is responsible for scheduling and distributing tasks, collecting and storing results from individual computational nodes. Individual computational nodes may perform descriptors calculation tasks, machine learning method computations (neural networks, kNN, SVM, WEKA-J48, etc.), and general QSAR modeling workflow management.

3.3.9 Model analysis

An important part of every QSAR modeling based research is the model analysis. The researcher should not only build the model, but estimate the model performance with regards to its generalization and prediction abilities, identify the outliers and study the reasons why the outliers occurred, access the model's applicability domain, etc. The OCHEM provides a set of convenient tools to perform these tasks.

The established metrics of the **regression model** performance are root mean squared error (RMSE), mean absolute error (MAE) and the squared correlation coefficient (r^2). The OCHEM provides these values for both training and validation sets of the model on the model summary page (Figure 3.10).

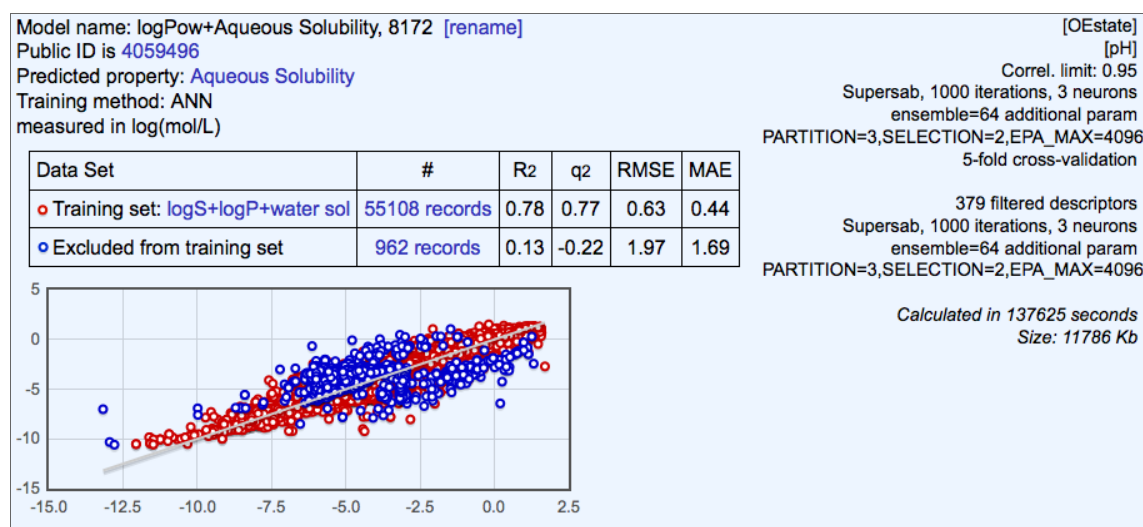


Figure 3.10. Model summary for a aqueous solubility model. Displays basic model statistics, a detailed summary of model parameters and a real-vs-predicted plot.

Another extremely illustrative tool for model analysis is the real-vs-predicted plot. It is a plot where every compound is displayed as a dot, and x-axis shows the real observed property value for this compound, and y-axis represents a value obtained for this compound via model prediction. This plot allows to estimate model performance in a single glance, as well as identify outliers - compounds, for which the real and predicted values are significantly different.

The OCHEM real-vs-predicted plot allows clicking on every individual dot and seeing the compound represented by this dot, as well as descriptors for this compound. This is a unique feature and is only possible due to integration of the experimental properties database and the modeling environment. By clicking on a very distinctive outlier a user may find out, that the record in question has an error in its value, has wrong or inconsistent measurement conditions, etc. Since all the references for the measurements are stored in the database, the user may track the experimental value to the original publication. If there are suspicions on the quality of particular datapoint, the user can exclude it from the training set by a single click.

The **classification** models are best characterized by the classification accuracy - overall percentage of correctly classified instances, and accuracies within each class. In the OCHEM the confusion matrix is displayed for classification models. All correctly classified or misclassified records can be accessed by one click on the appropriate confusion matrix cell. The example of the classification model summary is displayed on Figure 3.11.

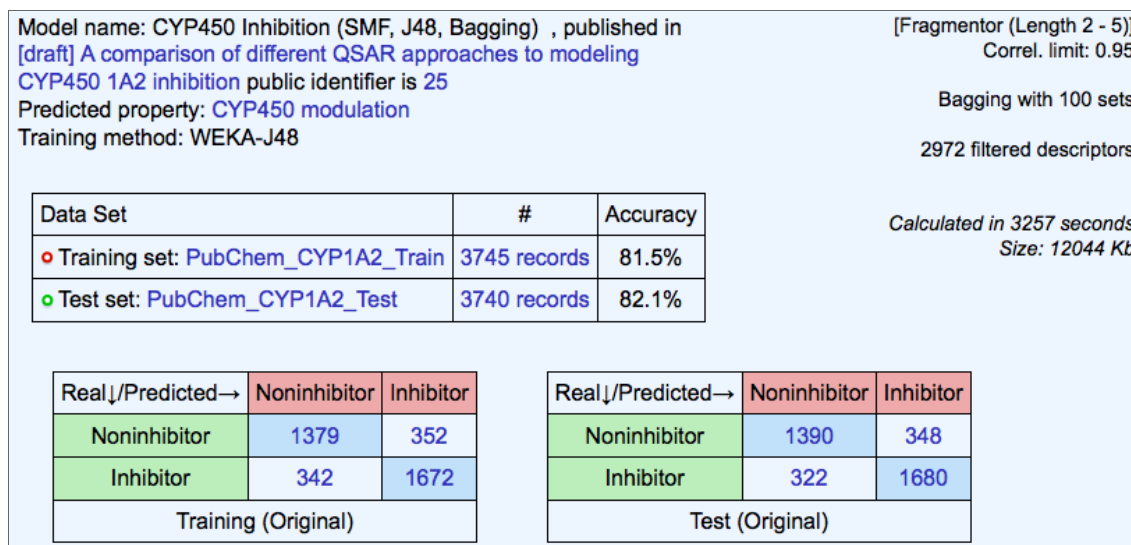


Figure 3.11. Summary of the classification model predicting CYP1A2 inhibitors. It displays the classification accuracy values, confusion matrices for both training and test sets, and a summary of model parameters.

3.3.10 Additional model assessment tools

The OCHEM provides several additional tools for model evaluation and comparison.

For multilearning models (models predicting several properties simultaneously) the user can have a look at the summary page with short statistics for each of the predicted properties (Figure 3.12).

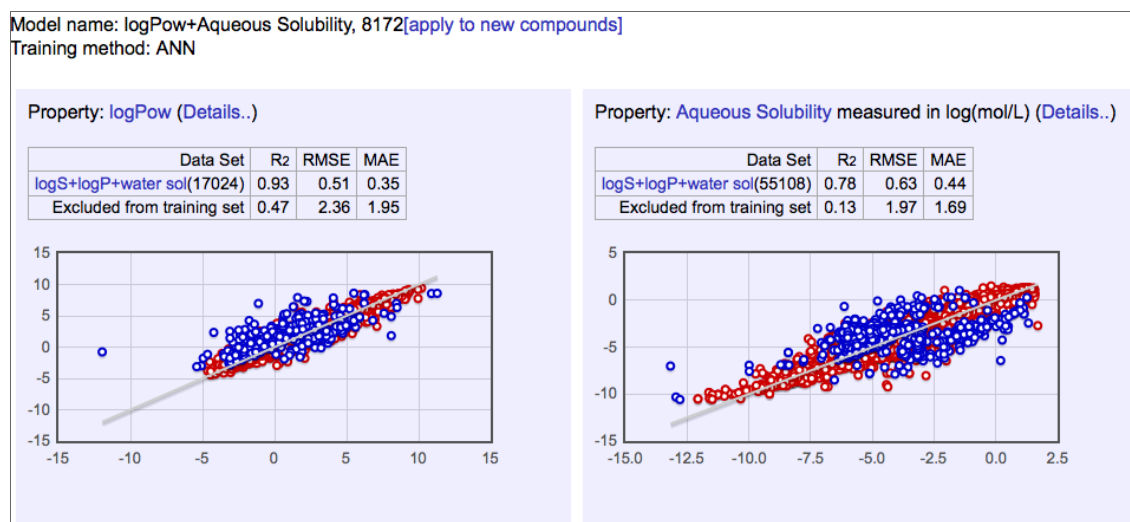


Figure 3.12. A summary page for logP+Solubility multilearning model-building.

If there are several models built on the same training set, the user can have a comparison summary of the performances of all the models for this set. The sample multi-model summary is depicted on Figure 3.13.

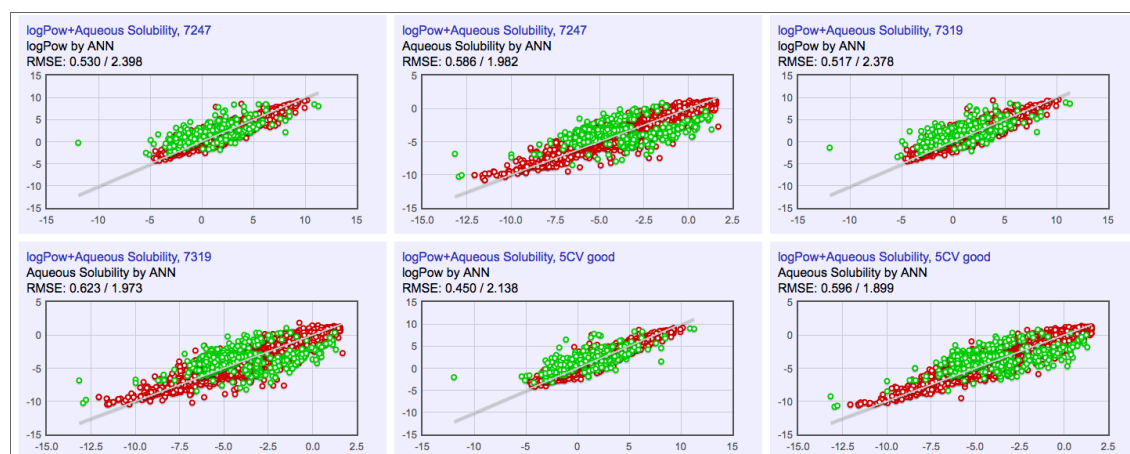


Figure 3.13. Multi-model comparison for a logP+Solubility training set

3.3.11 Applicability domain assessment

The OCHEM has powerful applicability domain tools that implement methods described in section 2.9, page 34.

The DMs implemented in the OCHEM modeling framework are: LEVERAGE, ASNN-STD, BAGGING-STD, CORREL, CLASS-LAG and STD-PROB. Every model is complemented with a applicability domain summary, that is appropriate for the modeling task type (classification or regression), machine learning method and validation method used in the model creation. Figure 3.14 displays the sample applicability domain chart for a classification problem with bagging validation.

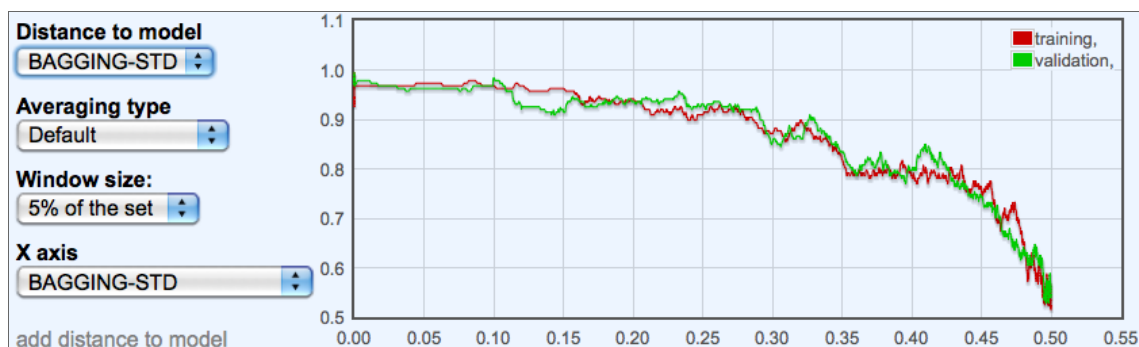


Figure 3.14. Applicability domain for classification CYP1A2 model with bagging validation

The controls left to the applicability domain chart allow dynamic changing of the DM metrics, averaging type (sliding window, bin-based, cumulative), size of the window (where applicable), and the X axis meaning (DM value or percentage of compounds). The Y axis displays accuracy metrics (for classification problems - total ratio of correctly classified instances).

3.3.12 Model application

Since the actual purpose of the QSAR model is prediction of new compounds, the OCHEM platform includes a model applier facility. The user has a possibility to choose one or several models from a model registry (Figure 3.15) and apply them to a set of new compounds. The set may either be a molecule set in the OCHEM database (and then the user has the possibility to compare model predictions with the actual property values) or a set of molecule structures uploaded as an SDF file during the application process.

Step 1. Select a model from the list

Filter by model name: **CYP** and property name: or by article id: Models visibility: **Public** [refresh]

1 - 5 of 5

CYP450 Inhibition (Full Set, ASNN, Bagging) , published by novserj	predicts CYP450 modulation using PubChem_CYP1A2_Train (3745) validated by PubChem_CYP1A2_Test (3740)	ANN	2011-02-14	
CYP450 Inhibition (SMF, J48, Bagging) , published by novserj	predicts CYP450 modulation using PubChem_CYP1A2_Train (3745) validated by PubChem_CYP1A2_Test (3740)	WEKA-J48	2011-02-10	
CYP450 Inhibition (Dragon, J48, Bagging) , published by novserj	predicts CYP450 modulation using PubChem_CYP1A2_Train (3745) validated by PubChem_CYP1A2_Test (3740)	WEKA-J48	2011-02-10	
CYP450 Inhibition (Full Set, J48, Bagging) , published by novserj	predicts CYP450 modulation using PubChem_CYP1A2_Train (3745) validated by PubChem_CYP1A2_Test (3740)	WEKA-J48	2011-02-10	
CYP450 modulation e-state , published by novserj	predicts CYP450 modulation using PubChem_CYP1A2_Train (3745) validated by PubChem_CYP1A2_Test (3740)	ANN	2011-01-25	

1 - 5 of 5

Figure 3.15. Model registry displaying published models filtered by “CYP” keyword. Three first models are selected for model applier.

The result of model application is the list of predictions (one prediction from each model for each molecule) accompanied by accuracy values, deduced from the AD methods. The results of model application can be exported in Excel or CSV format. Figure 3.16 displays predictions for one molecule by three models from the previous example.

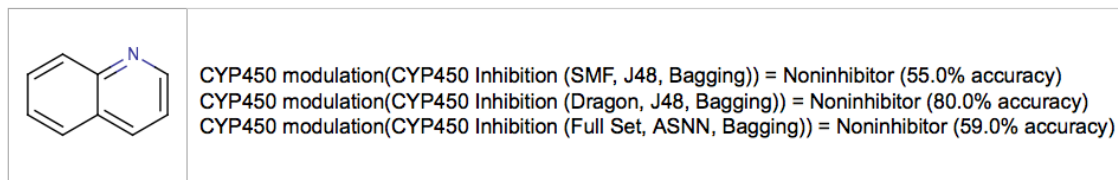


Figure 3.16. Predictions for CYP1A2 inhibition activity for a molecule by three models

3.4 Implementation notes

The OCHEM server-side code is developed with Java 6 programming language (compiles with OpenJDK6) and runs on the Tomcat 6 servlet container. The database is stored in the MySQL 5.1 DMBS, table engines are InnoDB. The Apache 2.2 HTTP server is used for seamless integration of several OCHEM services into single URL space.

The Java Hibernate 3.6 is used for object-relationship mapping abstraction layer. The Spring Framework 3 is used to support the MVC programming paradigm (separation of code, data and representation). The XML generation is performed by the JAXB Reference Implementation libraries. The XSLT technology is used to convert XML data representation to HTML web pages.

On the client side Javascript is the main programming language. The jQuery and YUI libraries are used. The AJAX technology is used to create an interactive user experience and JavascriptMVC framework is used to convert JSON data and javascript templates to HTML code snippets to be displayed to the user.

All chemistry-related processing (i.e., molecule format conversion, molecule depiction generation, molecule standardization, neutralization and salt removal) is performed by Chemaxon library set.

Individual calculation server implementations include tools mostly written in Java and C++.

The OCHEM comprises about 100,000 lines of Java code. Several of its components were inspired by the Virtual Computational Chemistry Laboratory (VCCLAB, <http://www.vcclab.org>).

3.5 Summary

The Online Chemical Modeling Environment (OCHEM) is a set of tools to facilitate QSAR research. One of its two main components is the user-contributed database of experimental measurements. The user-friendly interface allows searching for specific data by property, publication, molecular properties, etc. Special tools allow introducing of large amounts of user data for further analysis. The database allows storing various supplementary information (e.g., conditions of experimental

measurements), which allows creation of refined, consistent datasets.

The modeling framework is integrated with the database and can use the data from the database in the modeling process. The framework contains calculation nodes for all the typical steps of the QSAR modeling. The molecule preprocessing nodes allow performing such essential steps as molecule normalization and neutralization. Molecule structure optimization nodes give the user the choice between several options of molecule 3D optimization for three-dimensional descriptors. A large amount of available descriptor calculation nodes include both such industry standard descriptors as Dragon and such experimental descriptors as Protein-Ligand Interaction-Based descriptors and allow modeling of a wide range of physicochemical and biological properties. An important feature of the OCHEM is using experimental property conditions as descriptors to allow accurate modeling of condition-dependent properties. The machine learning method nodes include implementations of several most widespread methods and support both regression and classification tasks, as well as multilearning. The resulting model profile displays most integral model performance metrics and also allows to track model performance to an individual compound in the model's training set.

A cluster of over 500 cores employed by OCHEM makes even rather computationally intensive tasks feasible.

The OCHEM is available at <http://ochem.eu> and contains over 750,000 data points for around 500 different properties. Tools are being developed to integrate OCHEM with other databases such as ChemExper (<http://www.chemexper.com>) for physical properties such as boiling point, melting point and density or ChemSpider (<http://www.chemspider.com>).

The OCHEM platform was used as a main framework for all studies presented in this work. The Protein-Ligand Interaction-Based descriptors presented in this work are available for a limited number of proteins. The top performing models for human CYP inhibition presented in this thesis are available on the OCHEM online.

4 QSAR studies of CYP inhibition

This chapter is dedicated to describing the three QSAR studies of human CYP inhibition modeling. The methods used in the studies are briefly mentioned in appropriate “Methodology” sections of the studies and are described in great detail in chapter 2 (page 7). The studies were performed using the OCHEM platform (chapter 3, page 41).

In section 4.1 dataset overview and analysis is performed. The datasets are described, their qualitative and quantitative properties are determined. The cross-dataset relationships are calculated. Finally, the fragment-based analysis is performed and structural features of the datasets are described.

Section 4.2 presents benchmarking of traditional QSAR approaches, that involve well-established descriptors and machine learning techniques. The CYP1A2 dataset was chosen for this benchmarking study. The most successful descriptors and machine learning methods are determined in this section.

Section 4.3 demonstrates modeling results obtained by the most successful model configurations for CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 isoforms. The protein-ligand atom pair descriptors introduced in subsection 2.4.2 are used and the results are benchmarked against traditional approaches. Applicability domain and fragment-based analyses are performed.

In section 4.4 an attempt is made to predict activities of small molecules against a protein based on experimental data measured for another protein. The motivation is to determine whether the atom pair descriptors allow extrapolation of QSAR modeling results to clinically significant mutations of cytochromes. Due to lack of consistent dataset for mutated cytochrome activity, the CYP2C19 activity was chosen as the target property and CYP2C9 activity as the measured property.

4.1 Datasets overview and analysis

4.1.1 Datasets description

Throughout the studies presented in this chapter datasets from different sources were used. All molecules in all datasets were processed as described in section 2.2.1 , page 11.

PubChem AID410 dataset. PubChem [172,200] is a project hosted by a database by National Center for Biotechnology Information of National Library of Medicine of National Institute of Health funded by the U.S. government. The PubChem database is a rich storage of information about the properties of small molecules.

PubChem BioAssay – a section of the database that stores results of bioassays with brief descriptions. BioAssays in this database have unique numerical assay identifiers

(AIDs).

AID410 [201] holds the results of high throughput screening (HTS) measurements of human cytochrome 1A2 inhibition activity of small molecules, deposited in October 2007. The results include the structural information about the molecules, inhibition activities of small molecules at different concentrations, the AC50 values obtained from fitted concentration-response curves, and the resulting “inhibitor” / “non-inhibitor” labels. Only structure information and “inhibitor” / “non-inhibitor” labels were used to form the PubChem AID410 dataset in this study.

The description of the AID410 bioassay experiment shows that the demethylation of luciferin 6' methyl ether (the Luciferin-ME P450 Glo-Buffer provided by Promega-Glo) to luciferin was used as a target reaction for human CYP1A2 for this dataset. The luciferin was then measured by luminescence after the addition of a luciferase detection reagent. The dataset obtained from this bioassay contained 8,348 compounds, out of which 4,175 were determined as active, 3,673 – inactive, 713 – inconclusive. The protocol summary of the assay is available from the assay page on PubChem. The detailed protocol description is available in the Promega-Glo technical bulletin [202].

The dataset was subject to preprocessing (as described in the “Data preprocessing” section). After the preprocessing, if the same molecule was in both “active” and “inactive” sets or if a molecule was found in an “inconclusive” set, as specified by PubChem, it was removed from all sets. This was the case for 241 molecules. The number of non-conflicting inconclusive compounds was 543. There were also 66 molecules, that were duplicates within “inactive” or “active” lists, respectively. As a result of this preprocessing a non-redundant set of 4,016 active and 3,470 inactive molecules (a total of 7,486 molecules) was formed.

PubChem AID883 dataset. AID883 is a qHTS assay for inhibitors of cytochrome P450 2C9, deposited in December 2007 [203]. Similarly to the previously described assay, it contains molecular structures, inhibition activities of tested molecules at different concentrations, as well as the resulting “inhibitor” / “non-inhibitor” labels. The PubChem AID883 dataset for this study includes the structural information and “inhibitor” / “non-inhibitor” labels.

The assay used human CYP2C9 to measure the hydroxylation of deoxy luciferin (the Luciferin-H P450 Glo-Buffer provided by Promega-Glo) to luciferin. The luciferin was then measured by luminescence after the addition of a luciferase detection reagent. Luciferin-H concentration in the assay was equal to its Michaelis constant for CYP2C9. It contains a total of 9,567 molecules out of which 1,273 were determined as active, 6,937 - inactive, and 1,357 - inconclusive. The technical details about the assay can be found on AID883 assay page or in the Promega-Glo technical bulletin [202].

During the preprocessing stage 74 molecules were removed from the dataset as duplicates within “active” or “inactive” classes, and 208 molecules were removed as duplicates across “active”, “inactive” or “inconclusive” classes. The resulting PubChem AID883 dataset holds a total of 7,879 molecules, among them 1,167 are inhibitors and 6,712 - non-inhibitors.

PubChem AID899 dataset. AID899 is a qHTS assay for inhibitors of cytochrome P450 2C19, deposited to PubChem in December 2007 [204].

This assay used human CYP2C19 to measure the hydroxylation of ethylene glycol ester of 6' deoxyluciferin (Luciferin-H EGE by Promega-Glo) to luciferin. In this assay activities of 9,621 molecules were measured, 1,901 were reported active, 6,441 - inactive, 1,279 - inconclusive.

Based on this assay a dataset was formed. There were 67 duplicates within "active" or "inactive" molecule classes and 267 duplicates between "active" and "inactive" classes. The resulting PubChem AID899 dataset contained 7,922 molecules (1,756 active molecules and 6,166 inactive ones).

PubChem AID891 dataset. AID891 is a qHTS assay for inhibitors of cytochrome P450 2D6, deposited to PubChem in December 2007 [205]. Structural information and "inhibitor" / "non-inhibitor" labels from the assay were used to form this dataset.

This assay used human CYP2D6 to measure the demethylation of ethylene glycol ester of luciferin 6' methyl ether (Luciferin-ME EGE from Promega-Glo) to luciferin. The luciferin is then measured by luminescence after the addition of a luciferase detection reagent. Luciferin-ME EGE concentration in the assay was equal to its Michaelis constant for CYP2D6. The assay reports 9,598 molecules: 1,623 - active, 6,335 - inactive, 1,640 - inconclusive. The assay details and comments on molecule scoring and "active"/"inactive" labeling can be found on the assay page.

On the preprocessing stage 73 molecules were removed as duplicates within activity classes and 240 - as duplicates across activity classes. As a result, PubChem AID891 dataset contains 7,574 molecules (1,468 of them are active, and 6,106 - inactive).

PubChem AID884 dataset. AID884 is a qHTS assay for inhibitors of cytochrome P450 3A4, deposited to PubChem in December 2007 [206]. Structural information and "inhibitor" / "non-inhibitor" labels from the assay were used to form this dataset.

This assay used human CYP3A4 to measure the dealkylation of luciferin-6' phenylpiperazinyl (Luciferin-PPXE; luciferin detection buffer) to luciferin. The luciferin is then measured by luminescence after the addition of a luciferase detection reagent. Luciferin-PPXE concentration in the assay was equal to its Michaelis constant for CYP3A4. In total 13,312 molecules were measured. Among them 3,438 were reported as active, 7,066 - inactive, and 2,808 - inconclusive.

After removal of duplicates within activity classes (210 molecules) and between activity classes (311 molecules) the resulting dataset contains 9,979 molecules (3,303 - active, and 6,676 - inactive).

PubChem AID1851 datasets. AID1851 is a cytochrome panel assay with activity outcomes [207,208]. The study determined potency values for 17,143 compounds against five CYP isozymes (1A2, 2C9, 2C19, 2D6 and 3A4) using an in vitro bioluminescent assay. The compounds included libraries of US FDA-approved drugs and screening libraries. Among these molecules 8,019 were the compounds from the Molecular Libraries Small Molecule Repository, including compounds chosen for diversity and rule-of-five compliance, synthetic tractability and availability; 6,144 compounds were from biofocused libraries, which included 1,114 FDA-approved drugs; and the rest 2,980 compounds were from combinatorial chemistry libraries, containing privileged structures targeted at G protein-coupled receptors and kinases and containing

purified natural products or related structures.

This assay used various human CYP P450 isozymes to measure the dealkylation of various pro-luciferin substrates to luciferin. The luciferin is then measured by luminescence after the addition of a luciferase detection reagent. Pro-luciferin substrate concentration in the assay was equal to its Michaelis constant for its CYP P450 isozyme. Inhibitors and some substrates limit the production of luciferin, and decrease measured luminescence.

To address potential artifacts due to the assay format, particularly important for pan-active compounds, we used a database of potency values determined for the variant of the firefly luciferase used in the assay to remove any compounds that interfered with luciferase detection (only 0.7% were found to be interfering in the compound collection used for the assay).

Since this thesis focuses only on CYP inhibitors, all the activators and inconclusive molecules were discarded.

The AID1851 dataset was deposited in July 2009 and contains larger amount of molecules and from more diverse sources than AID410, AID883, AID899, AID891 and AID884 assays. This makes data from AID1851 a perfect validation set for unbiased evaluation of predictive abilities of QSAR models for CYP inhibition.

Only confident results with $\log(AC_{50}) < -5$ were taken as “inhibitors”. The molecules with low confidence results and molecules with $\log(AC_{50}) > -5$ were considered inconclusive and removed from the set.

Only molecules that were explicitly marked as “inactive” were taken as “non-inhibitors” in the set.

AID1851 CYP1A2 Full dataset contains 12,157 molecules (5,430 inhibitors and 6,727 non-inhibitors). To have an unbiased estimation of QSAR models built on **AID410** dataset, the **AID1851 CYP1A2 Filtered** set was prepared. This set contained only molecules not present in **AID410** dataset and therefore not used in model development. The set contained 6,636 molecules (3,016 inhibitors and 3,620 non-inhibitors).

Similarly, **AID1851 CYP2C9 Full** dataset contains 12,034 compounds (3,983 inhibitors and 8,051 non-inhibitors). The **AID1851 CYP2C9 Filtered** (against **AID883**) dataset contains 5,728 molecules (3,306 inhibitors and 2,422 non-inhibitors).

AID1851 CYP2C19 Full dataset contains 11,717 compounds (4,893 inhibitors and 6,824 non-inhibitors). The **AID1851 CYP2C19 Filtered** (against **AID899**) dataset contains 5,569 molecules (3,790 inhibitors and 1,779 non-inhibitors).

AID1851 CYP2D6 Full dataset contains 12,914 compounds (2,372 inhibitors and 10,542 non-inhibitors). The **AID1851 CYP2D6 Filtered** (against **AID891**) dataset contains 6,817 molecules (1,415 inhibitors and 5,402 non-inhibitors).

AID1851 CYP3A4 Full dataset contains 11,412 compounds (4,211 inhibitors and 7,201 non-inhibitors). The **AID1851 CYP3A4 Filtered** (against **AID884**) dataset contains 6,029 molecules (2,664 inhibitors and 3,365 non-inhibitors).

Table 4.1 summarizes quantitative contents of the studied datasets.

		CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP3A4	
Training sets	Active	4016	1167	1756	1468	3303	
	Inactive	3470	6712	6166	6106	6676	
	CSI	0.31	0.31	0.32	0.31	0.34	
	Assay	AID410	AID883	AID899	AID891	AID884	
Test sets full	Active	5430	3983	4893	2372	4211	
	Inactive	6727	8051	6824	10542	7201	
	CSI	0.28	0.27	0.28	0.29	0.27	
	Assay	AID1851	AID1851	AID1851	AID1851	AID1851	
Test sets filtered	Active	3016	3306	3790	1415	2664	
	Inactive	3620	2422	1779	5402	3365	
	CSI	0.21	0.21	0.2	0.21	0.21	
	Assay		AID1851	AID1851	AID1851	AID1851	AID1851
			excluding AID410	excluding AID883	excluding AID899	excluding AID891	excluding AID884

Table 4.1. Distribution of inhibitors and non-inhibitors in the studied datasets . CSI (chemical similarity index) – an averaged pairwise Tanimoto similarity index calculated on 512 bit daylight structural fingerprint for all molecules in a set; it represents a basic chemical diversity measure for a set (lower values – higher diversity).

Experimental accuracy of the PubChem data sets. To have a basis for comparison of model accuracy to the experimental accuracy, we considered the inconclusive compounds in the dataset as experimental errors. Since not all inconclusive compounds should be treated as experimental errors, this value is an overestimation. However, it provides a lower boundary for accuracy estimation.

For PubChem AID410 Dataset the error rate is $713 / 8,348 = 0.085 \sim 9\%$. For PubChem AID883 error rate is $1,357 / 9,567 = 0.141 \sim 14\%$. For PubChem AID899 dataset the error rate is $1,279 / 9,621 = 0.133 \sim 13\%$. For PubChem AID891 the error rate is $1,640 / 9,598 = 0.171 \sim 17\%$. For PubChem AID884 the error rate is $2,808 / 13,312 = 0.210 \sim 21\%$.

We can see that for the training sets our estimated accuracies lie in the ranges of 80 - 91%.

For the AID1851 the experimenters have performed confirmatory measurements on 91 compound randomly selected from the 17,143 molecules and have reported the reproducibility of measurements in the ranges of 84-90% [206]. This number agrees with our estimation.

4.1.2 Preliminary analysis of datasets

A short preliminary analysis of the datasets was performed. For this we have calculated daylight fingerprints (as implemented in Chemaxon software) on the molecular structures of the datasets. The length of each fingerprint is 512 bit.

We then calculated chemical similarity indices (CSI), which are averaged Tanimoto indices on these fingerprints within each dataset to evaluate the structural diversity of the sets. Table 4.1 displays the ratios of active / inactive compounds in each set along with the chemical similarity index for each set.

We can see that the AID1851 filtered datasets are more diverse than the datasets used

for model training (average chemical similarity index of 0.21 for test sets, average chemical similarity index of 0.31 for training sets). This means that the test sets represent a larger fraction of chemical space and contain more diverse classes of compounds. We can also observe that the ratios of active and inactive compounds in the AID1851 datasets are different from the ratios in the training set. This gives additional reasons to evaluate models using measures that account for accuracies within each class equally.

The preliminary conclusion is that the models built on the training sets can not be applicable to all the compounds of the test set and the average prediction accuracies are expected to be lower than cross-validated accuracies on the training set. Applicability domain methods should be used to separate reliable model predictions from unreliable ones.

To assess the similarity of activity classes of different studied CYP isoforms ratios of molecules with the same inhibitory activity were calculated, separately for training sets, full test sets and filtered test sets.

For each group of datasets a set of molecules was selected, that is present in all datasets in a group (a total of 3,878 molecules for training sets, 5,270 for full test sets and 1,651 for filtered test sets). After that for each pair of datasets activity classes for these molecules were compared. A similarity measure was calculated as a ratio of molecules with the same activity class to the overall amount of molecules.

Table 4.2 displays these pairwise similarity measures for three groups of datasets.

	1A2	2C9	2C19	2D6	3A4
1A2	1,000	0,656	0,708	0,675	0,722
2C9	0,656	1,000	0,898	0,790	0,797
2C19	0,708	0,898	1,000	0,835	0,832
2D6	0,675	0,790	0,835	1,000	0,773
3A4	0,722	0,797	0,832	0,773	1,000

	1A2	2C9	2C19	2D6	3A4
1A2	1,000	0,692	0,716	0,667	0,657
2C9	0,692	1,000	0,849	0,702	0,749
2C19	0,716	0,849	1,000	0,722	0,755
2D6	0,667	0,702	0,722	1,000	0,707
3A4	0,657	0,749	0,755	0,707	1,000

	1A2	2C9	2C19	2D6	3A4
1A2	1,000	0,669	0,697	0,631	0,598
2C9	0,669	1,000	0,793	0,581	0,696
2C19	0,697	0,793	1,000	0,563	0,686
2D6	0,631	0,581	0,563	1,000	0,656
3A4	0,598	0,696	0,686	0,656	1,000

Table 4.2. Similarity measures for activities of isoforms for training sets, full test sets and filtered test sets.

We can see that the highest similarity in all datasets is consistently among the CYP2C9 and CYP2C19 isoforms. This can be attributed to the fact that these isoforms belong to the same 2C subfamily and share a large amount of structural similarity [209,210].

CYP3A4 is closer to CYP2D6, CYP1A2, CYP2C9 than these isoforms among themselves. CYP3A4 is especially close to CYP2C19. This fact reflects high promiscuity of CYP3A4 isoform. CYP3A4 is involved in around 50% of all CYP-mediated metabolism [17] (see section 1.2, page 2).

4.1.3 Fragment analysis

Figures 4.1 - 4.3 display the structurally diverse fragments that were found to influence CYP inhibition activity. The bar plots show the negative decimal logarithm of the p-value for abundance of active (p-value active) or inactive (p-value inactive) compounds containing the fragment, relative to the whole dataset (calculated as a binomial probability of such abundance when randomly sampling a subset of such size from the original set). Larger bars indicate stronger correlation between the presence of a specific fragment in a molecule and the CYP-related activity of this molecule. Only four top fragments were selected for analysis from each fragment group.

Simple fragments

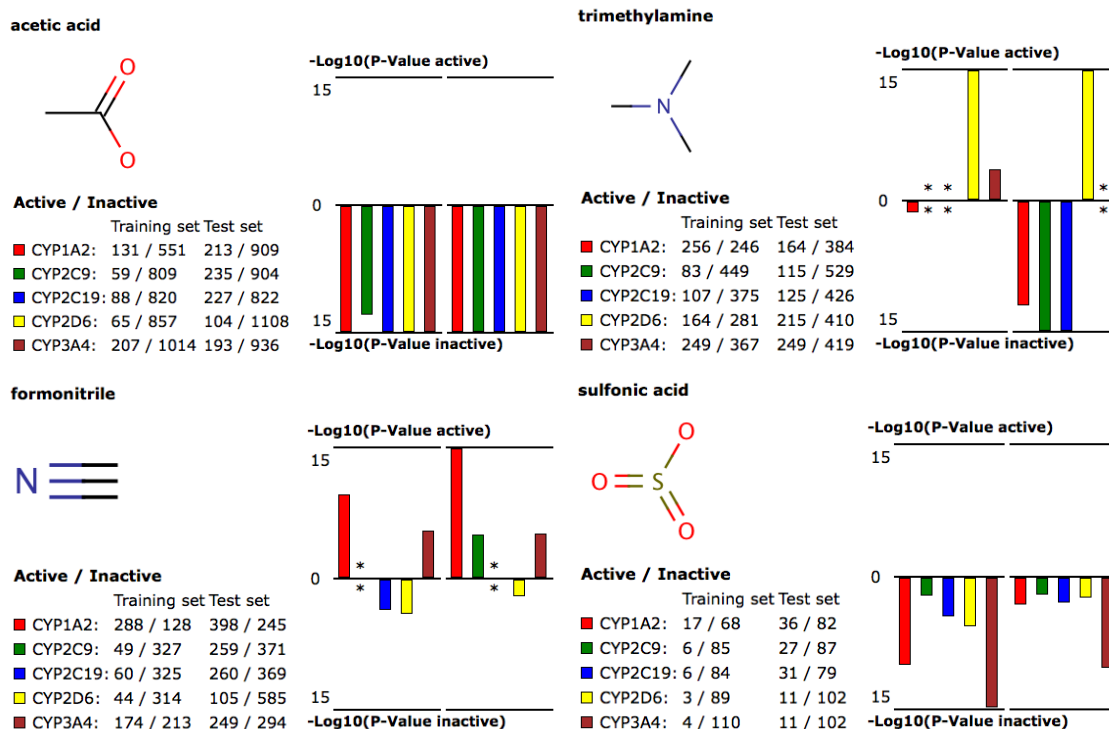


Figure 4.1. Simple fragments with disproportionally distributed activity classes relative to the full molecule sets.

Among determined simple fragments we can observe two different activity patterns.

Acetic acid and *sulfonic acid* fragments demonstrate a strong correlation with CYP non-inhibitor molecules. Based on the analyzed dataset the presence of either of the fragments statistically significantly makes a molecule CYP non-inhibitor for all of the studied CYP isoforms. The effect is most apparent for the *acetic acid* fragment.

The *trimethylamine* fragment demonstrates a case of isoform selectivity. Based on both

training and test datasets, the presence of this fragment strongly indicates CYP2D6 inhibition activity. The analysis for the rest of the isoforms is inconclusive. However, it suggests a non-inhibitor correlation for CYP1A2, CYP2C9 and CYP2C19 and a weak inhibitor correlation for CYP3A4.

Formonitrile fragment displays strong correlation with CYP1A2 and CYP3A4 inhibition activity and weak correlation with CYP2D6 non-inhibition. The results for the other isoforms are inconclusive and can't be determined statistically based on the given datasets.

Simple aromatic-containing fragments

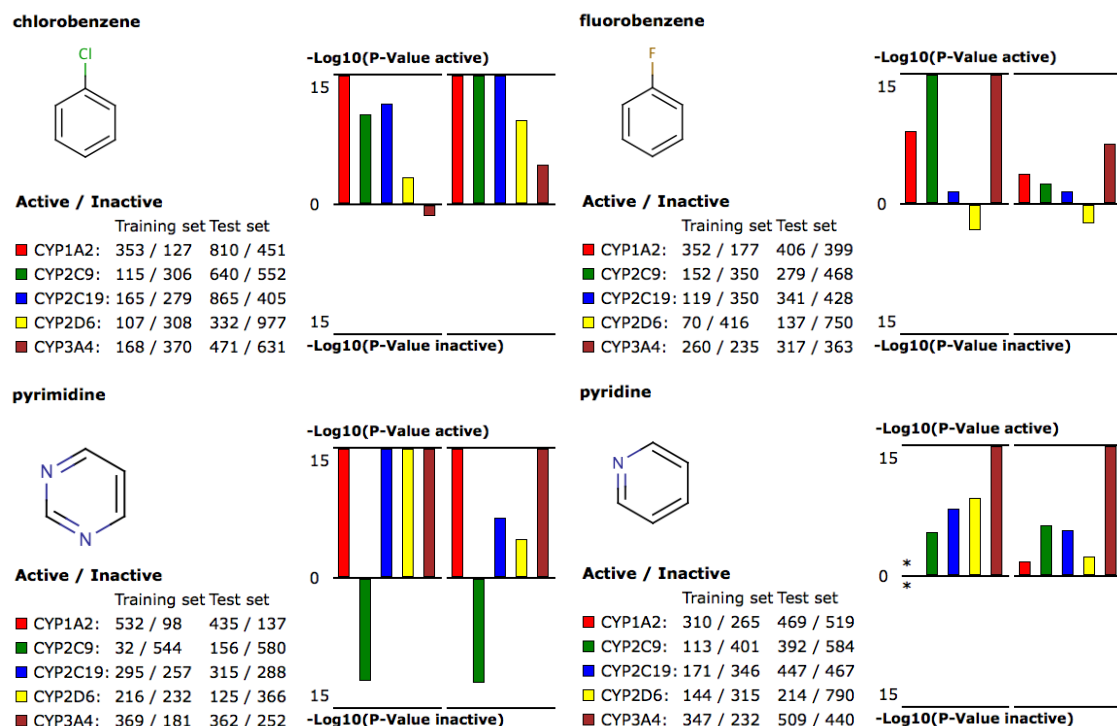


Figure 4.2. Simple aromatic-containing fragments with disproportionally distributed activity classes relative to the full molecule sets.

Chlorobenzene fragment displays a correlation with CYP inhibition activity for 1A2, 2C9, 2C19 and 2D6 isoforms and has inconclusive results for 3A4 isoform. The correlation between the presence of this fragment and 1A2, 2C9 and 2C19 inhibition activity is very strong with $-\text{Log}_{10}(\text{P-Value}) > 9$ for both studied datasets.

Fluorobenzene fragment can be associated with CYP1A2 and CYP3A4 inhibition activity. There is also weak correlation between presence of the fragment in a molecule and a non-inhibition activity of this molecule for CYP2D6 isoform. Activity against other isoforms is inconclusive for this fragment.

Despite the structural similarity the pyrimidine and pyridine fragments have different activity patterns.

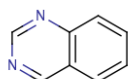
The pyrimidine displays a strong case of isoform selectivity. There is strong correlation between the presence of this fragment in a molecule and inhibition activity of this molecule against CYP1A2, CYP2C19, CYP2D6 and CYP3A4. However, apparently presence of this

fragment in a molecule makes it a non-inhibitor for CYP2C9. This is one of the cases of a strong difference in CYP2C9 and CYP2C19 activities.

The pyridine displays inconclusive results for CYP1A2 and somewhat less apparent (compared to pyrimidine) correlation with CYP2C19 and CYP2D6 inhibition activities. It's noteworthy that pyridine-containing molecules demonstrate weak correlation with CYP2C9 inhibition activity.

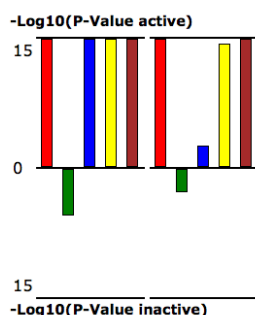
Complex heterocycles

quinazoline

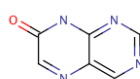


Active / Inactive

	Training set	Test set
■ CYP1A2:	472 / 13	405 / 19
■ CYP2C9:	32 / 386	138 / 366
■ CYP2C19:	215 / 179	170 / 178
■ CYP2D6:	202 / 131	102 / 158
■ CYP3A4:	351 / 52	253 / 61

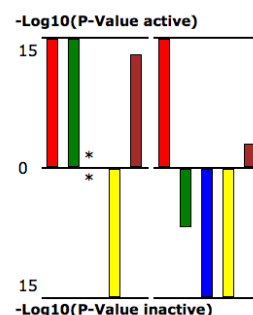


8H-pteridin-7-one

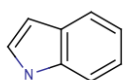


Active / Inactive

	Training set	Test set
■ CYP1A2:	907 / 74	622 / 72
■ CYP2C9:	219 / 556	162 / 510
■ CYP2C19:	170 / 653	106 / 622
■ CYP2D6:	28 / 819	60 / 813
■ CYP3A4:	324 / 354	239 / 321

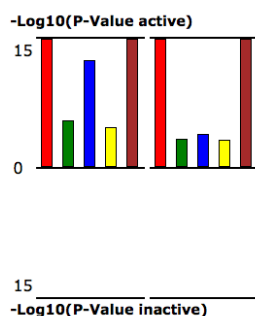


1H-indole

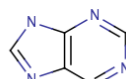


Active / Inactive

	Training set	Test set
■ CYP1A2:	182 / 40	175 / 69
■ CYP2C9:	50 / 128	112 / 150
■ CYP2C19:	92 / 110	139 / 124
■ CYP2D6:	58 / 121	68 / 190
■ CYP3A4:	150 / 75	163 / 95



9H-purine



Active / Inactive

	Training set	Test set
■ CYP1A2:	39 / 89	33 / 81
■ CYP2C9:	5 / 119	10 / 110
■ CYP2C19:	14 / 117	15 / 99
■ CYP2D6:	7 / 115	11 / 105
■ CYP3A4:	69 / 174	19 / 93

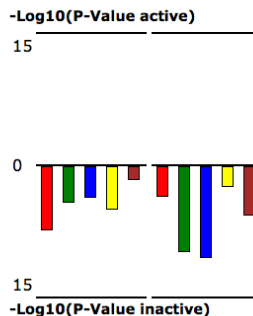


Figure 4.3. Complex heterocyclic fragments with disproportionally distributed activity classes relative to the full molecule sets.

Quinazoline and *8H-pteridin-7-one* were earlier determined as a fragment that correlates with CYP1A2 inhibition activity [39].

In this analysis we can see that *quinazoline* has strong correlation with 1A2, 2C19, 2D6 and 3A4 inhibition activity. The fragment also demonstrates isoform selectivity, since, based on the studied datasets, *quinazoline*-containing molecules tend to be CYP2C9 non-inhibitors with $-\text{Log}_{10}(\text{P-Value}) \sim 5$.

8H-pteridin-7-one can be statistically associated with CYP1A2 and CYP3A4 inhibition activity and CYP2D6 non-inhibition activity. The results for other isoforms are inconclusive. The analysis of the training dataset displays a strong correlation between the fragment presence and CYP2C9 inhibition activity. However, the test set analysis demonstrates opposite results, i.e. the fragment is mostly associated with non-inhibition activity.

1H-indole fragment displays a consistent correlation with CYP inhibition activity against all studied isoforms.

9H-purine fragment displays a stable non-inhibition pattern.

4.1.4 Summary

For the experiments presented in this thesis datasets from PubChem BioAssay database were obtained.

Both training and test sets are results of high throughput screening experiments against five major cytochrome P450 isoforms. The training sets are obtained from five different assays – AID410, AID883, AID899, AID891 and AID884. The experimental measurements in these assays were performed using same methodology and on similar sets of compounds.

The test sets were based on AID1851 assay. This assay was performed using the same measurement methodology but on a wider and more chemically diverse set of compounds. The test sets were obtained by removing all molecules present in the training sets from the AID1851 assay dataset.

All datasets were processed in the same way by removing salts, standardizing and neutralizing the molecules. All duplicates, conflictual or inconclusive results were removed.

Based on inconclusive molecule analysis, the experimental accuracy of the datasets are 80-91%. This corresponds with the similar estimates of dataset authors, who reported 84-90% accuracy.

A correlation of inhibition activity across datasets for different isoforms was calculated. The CYP2C9 and CYP2C19 showed the highest degree of similarity with 85-89% of similar activity molecules. CYP3A4 was found to be the closest isoform to CYP1A2, CYP2C9, CYP2C19 and CYP2D6. This corresponds to the known fact that CYP3A4 has the largest fraction of metabolized exogenous compounds and often serves as a secondary metabolism pathway in case the activity of other isoforms is lowered.

The fragment-based analysis of the datasets was performed and activity profiles were built. Some fragments demonstrated a statistically significant correlation with inhibition (*chlorobenzene*, *pyridine*, *1H-indole*) or non-inhibition (*Acetic acid*, *sulfonic acid*, *9H-purine*) activity. Other fragments displayed statistically significant isoform selectivity (*trimethylamine* only inhibitor for CYP2D6, *pyrimidine* and *quinazoline* are inhibitors for all isoforms except CYP2C9, *8H-pteridin-7-one* - inhibitor for CYP1A2 and CYP3A4 and non-inhibitor for CYP2D6).

The datasets were prepared for further use in all QSAR studies in this thesis.

4.2 Benchmarking of QSAR models for CYP1A2 inhibitor classification

4.2.1 Materials and methods

Datasets. For this study the PubChem AID410 dataset was used as a training set for all the models. All cross-validation and bagging results are reported for this dataset. The PubChem AID1851 CYP1A2 Filtered dataset was used as an external test set for this study.

Descriptors. One of the goals of the study was to determine the influence of different representation of molecules on the quality of models for the prediction of CYP1A2 inhibitor. One of the specific questions researched in this study is whether the 3D descriptors are necessary to achieve high prediction accuracy for this property. Three descriptors sets were used: fragments-base descriptors (ISIDA SMF)[125], 2D topological descriptors (E-state)[127] and a diverse set of 0D – 3D descriptors (Dragon) [124]. The descriptors were described in the methodology section (section 2.4 , page 18). The models were built using the following descriptor combinations and configurations:

- *Estate*: E-State indices (atom type indices and bond type indices)
- *ISIDA*: ISIDA descriptors (fragment length from 2 to 5 atoms)
- *Dragon2D*: 0D-2D descriptors from Dragon package
- *Dragon*: 0D-3D descriptors from Dragon package
- *All*: the full set containing all descriptors from the aforementioned sets

In these sets only Dragon and All configurations include 3D descriptors. The detailed description of all used tools can be found in section 2.4 on page 18.

The models in this study were created both with full set of descriptors and with the use of descriptor selection procedure. The descriptor selection procedure is based on a custom-implemented unsupervised correlation-based method. The descriptors are split into subsets and in each subset the correlation coefficients among descriptors are calculated. The descriptors having a correlation coefficient of more than some particular threshold are filtered out. The descriptors that correlate the least amount of other descriptors in a subset are kept. For the purpose of this study a correlation threshold of 0.7 was adopted.

Machine learning methods. Several popular machine-learning methods that were found efficient for QSAR modeling were used in this study. When applied to the same datasets, these methods provide a basis for comparison of efficiency of each method to predict CYP1A2 inhibitors. The analyzed methods were Associative Neural Networks (ASNN) [147,148], k Nearest Neighbors (kNN), Random Forest (RF) [151,152], C4.5 Tree (J48) [153], and Support Vector Machines (SVM)[149,150] as implemented in LibSVM [211]. The detailed description of the methods is provided in section 2.5 , page 23. Two validation strategies were used: the 5-fold cross-validation, and bagging. The number of bagging instances was 64.

Applicability domain. For bagging models the applicability domain approaches were studied. The DM measure was BAGGING-STD [163]. The ability of this measure to separate accurate and inaccurate predictions was studied. The details of applicability domain methods and BAGGING-STD DM are described in section 2.9 , page 34.

4.2.2 Modeling results

The benchmarking performed in this study included combination of four parameters - descriptor set, machine learning method, validation strategy and descriptor selection approach. The total amount of possible models was $5 \times 5 \times 2 \times 2 = 100$. Only 94 models were built, however. The SVM bagging models for non-decorrelated descriptors and for decorrelated full set were too large and could not be calculated.

Table 4.3 displays 30 top performing models out of these 94. The selection was performed based on overall model balanced accuracy (BACC). The table contains the details of each model (descriptor set, machine learning method, ensemble approach) as well as additional accuracy measures - accuracy (ACC), sensitivity (SENS), specificity (SPEC) and Matthew's correlation coefficient (MCC) (section 2.6, page 29). The table is divided into three groups. Within each group the models are statistically non-significantly different to the top model in the group with the significance level of 0.05.

Descriptors	Method	Validation	Selection	ACC	BACC	SENS	SPEC	MCC
All	Ann	Bagging	None	0,828	0,829	0,826	0,832	0,656
All	J48	Bagging	None	0,827	0,827	0,828	0,825	0,653
All	Ann	Bagging	Decor	0,826	0,826	0,823	0,829	0,651
All	J48	Bagging	Decor	0,826	0,826	0,827	0,825	0,650
Dragon	J48	Bagging	None	0,823	0,823	0,822	0,824	0,644
Dragon	J48	Bagging	Decor	0,822	0,821	0,833	0,808	0,641
Estate	J48	Bagging	None	0,821	0,819	0,848	0,791	0,640
All	RF	Bagging	Decor	0,823	0,819	0,866	0,772	0,643
Dragon	RF	Bagging	Decor	0,822	0,819	0,863	0,774	0,642
Dragon	Ann	Cv	None	0,817	0,817	0,822	0,812	0,633
Estate	RF	Bagging	None	0,819	0,817	0,851	0,782	0,636
Dragon2D	RF	Bagging	Decor	0,820	0,817	0,858	0,775	0,637
All	Ann	Cv	None	0,817	0,816	0,821	0,812	0,632
Estate	J48	Bagging	Decor	0,818	0,816	0,849	0,783	0,634
All	Ann	Cv	Decor	0,816	0,816	0,828	0,803	0,631
Dragon2D	J48	Bagging	Decor	0,818	0,815	0,857	0,774	0,634
Dragon	Svm	Bagging	Decor	0,815	0,815	0,813	0,818	0,629
Estate	RF	Bagging	Decor	0,818	0,815	0,852	0,778	0,633
All	RF	Bagging	None	0,818	0,815	0,862	0,767	0,634
Dragon	RF	Bagging	None	0,817	0,814	0,861	0,766	0,632
Dragon	Ann	Bagging	Decor	0,813	0,813	0,809	0,817	0,625
All	Svm	Cv	Decor	0,811	0,813	0,783	0,842	0,624
Dragon	Ann	Cv	Decor	0,813	0,813	0,815	0,810	0,625
ISIDA	Svm	Bagging	Decor	0,809	0,811	0,784	0,839	0,621
Dragon2D	J48	Bagging	None	0,813	0,811	0,847	0,775	0,624
ISIDA	J48	Bagging	None	0,812	0,809	0,850	0,768	0,621
ISIDA	J48	Bagging	Decor	0,808	0,807	0,814	0,801	0,614
Dragon2D	RF	Bagging	None	0,810	0,807	0,849	0,765	0,617
Dragon	Ann	Bagging	None	0,807	0,807	0,810	0,803	0,613
ISIDA	RF	Bagging	None	0,810	0,806	0,853	0,760	0,617

Table 4.3. The performance of best 30 models for the set of CYP1A2 inhibitors and non-inhibitors from PubChem BioAssay database. ANN – Associative Neural Networks [147,148], RF and J48 – random trees [152] and C4.5 pruned trees [153] as implemented in WEKA [154], SVM - support vector machines [149] as implemented in LibSVM [211]. Dragon, Dragon2D - 3D and 2D descriptors by software by Talete inc. [124], ISIDA - substructural molecular fragments as implemented in ISIDA [125], Estate - electrotopological state indices [127].

To investigate the influence of the studied parameters of models on their test set accuracy, the cumulative plots were built. Firstly, the models were sorted according to BACC in descending order and n top-performing models were selected. Secondly, among the list of n -top ranked models the percentage of models of each particular machine learning method was calculated.

Figure 4.4a shows the calculated percentage of models, built with the use of a particular descriptor set (y axis) among the n top-performing models (x axis). Number n changes from 2 to 94 with a step of 2 models. Methods with higher areas in the left part of the plot had higher performance.

Figure 4.4b uses the same concept to illustrate the difference in machine learning method performance, while Figure 4.4c shows the performance of bagging versus single (cross-validated) method. Figure 4.4d uses the same concept to illustrate the difference between models build on full sets of descriptors versus decorrelated sets of descriptors.

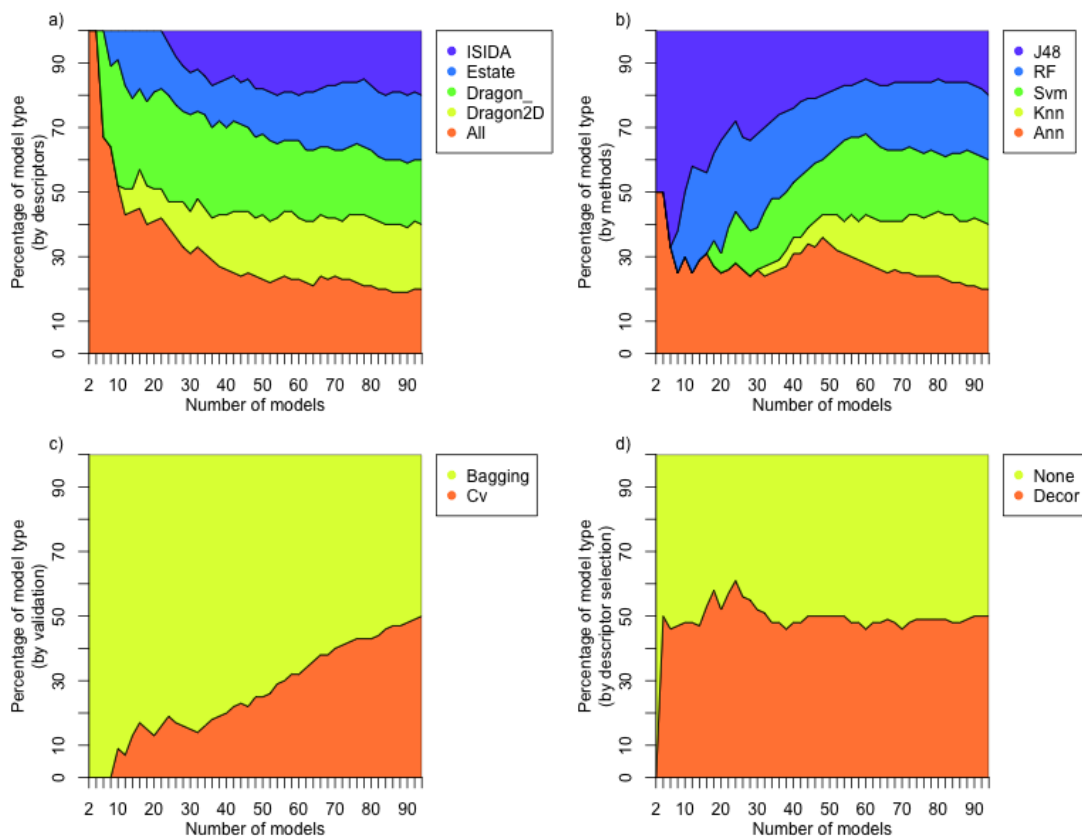


Figure 4.4. Cumulative charts of share of models of each type among the top-performing models. The horizontal axis displays the amount of top performing models taken into account; the vertical axis displays a share of each type of machine learning methods, descriptors or ensemble approaches among these models. Larger areas (J48 and ANN, Dragon and All, Bagging) demonstrate more successful approaches.

Comparison of machine learning methods. Among the used methods, best performances were achieved with ANN and J48 methods. The model with the highest balanced accuracy - 82.9% correctly classified instances - is nonsignificantly better than the following five models with balanced accuracies of 82.1% - 82.7%, with the significance level of 0.05 (hypothesis testing was performed as described in “Bootstrap testing” in the methodology section).

Overall most of the J48 and Ann models performed significantly better than models based on other machine learning methods. The KNN was the only machine learning method that didn't produce a model within top 30 most accurate models. Thus, this method had a lower performance than other machine learning methods analyzed in our work.

Comparison of descriptors. The models based on descriptors with 3D information (Dragon and All) significantly outperformed models based on 2D descriptors only. When used separately, Dragon descriptors demonstrated the highest performance. Among the 2D descriptors E-State indices performed best, followed by Dragon2D and ISIDA with approximately equal performance. These results show, that a combination of the descriptor sets calculated with different approaches brought new information to the model and increased its performance.

It is important to note that 3D descriptors increased model performance, which demonstrated the importance of 3D information for modeling CYP1A2 inhibition activity. On the other hand, the generation of 3D structures can be a limiting step and can significantly increase the time required for application of models using these sets of descriptors.

Bagging/ensembles provided better results compared to the single models. The charts at Figure 1c demonstrate that bagging/ensemble methods performed better than single-models. Table 1 confirms this result and also indicates that bagging and ensemble approaches significantly improved the performance of ASNN, SVM, RT and J48 models. However, these approaches had less or no influence on the KNN models. The KNN methods is more stable and is less influenced by distortions of the training set due to bagging. The former four methods, however, have a stronger intrinsic variability and models calculated with such methods using different bagging replica have larger variations.

The standard deviations of predictions for bagging-validated molecules were 0.31, 0.27, 0.26 and 0.19 for J48, RF, SVM and ASNN methods, respectively, while they were only 0.08 for KNN. This result indicates that methods with higher variation of predictions (SVM, RT, ASNN and J48) had a higher gain from using bagging approach, as it is clear from Table 4.3. This result is in agreement of previous conclusions of Breiman [152], who reported similar results by considering bias and variance of models. He assumed that methods with higher variation of results may have lower bias and their low performance could be mainly due to higher variation of their predictions. The average of predictions of such methods decreases their variance and improves their accuracy. The performances of more stable methods (e.g., KNN) are to

a larger degree dependent on their biases. Therefore, the use of ensemble approach does not improve their accuracy.

The increase of model accuracy came at a price of increasing usage of computational resources both for training and application of a model. In the presented study the bag consisted of 64 model instances. This led to a 64 times increase of computational time required to create and apply these models and the corresponding increase in the size of the models.

Unsupervised descriptor selection does not influence modeling results. In this study the models trained on the preselected set of descriptors performed similar to those trained on the full set of descriptors. This illustrates that unsupervised correlation-based descriptor selection method did not reduce the amount of information contained in the descriptor set. It is therefore beneficial to use the descriptor selection procedure to reduce the computational complexity of the models, decrease their calculation time and overall size.

Figure 4.5 shows the PCA plots of the descriptor sets before and after decorrelation procedure. The decorrelation procedure keeps the variability of the descriptor set while removing the possible bias caused by correlating descriptors. The effect is most obvious for ISIDA and Dragon descriptors.

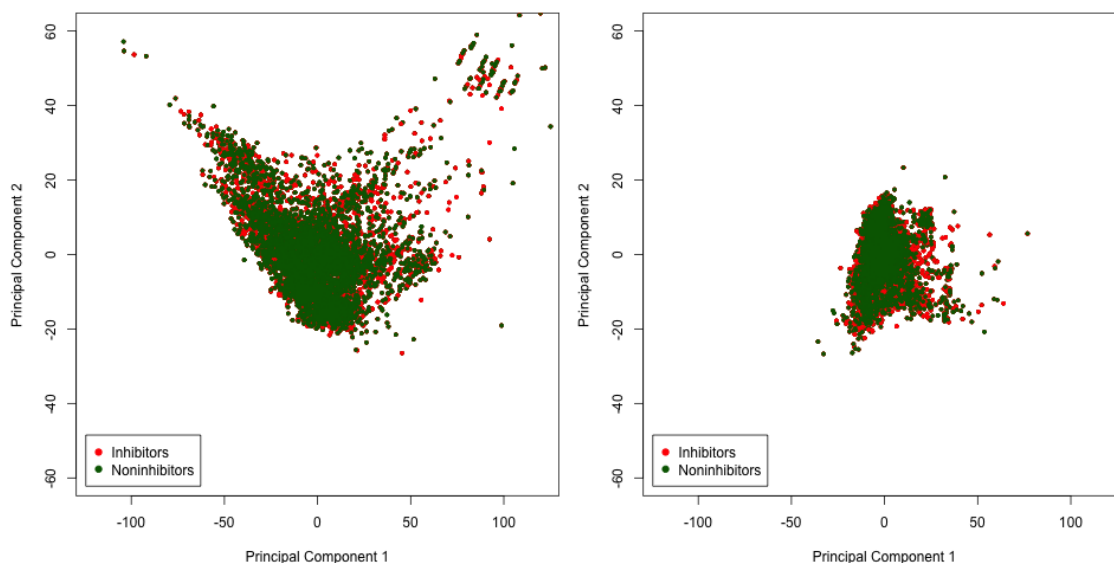


Figure 4.5. PCA plots of descriptors before (left panel) and after (right panel) decorrelation procedure.

4.2.3 PCA Plot model comparison

Figure 4.6 displays a PCA plot of the researched set of models in the space of predictions. The colors indicate the used machine learning method, a total of 5 different colors. The size of the points indicate a validation approach (bigger points - bagging, smaller points - cross-validation). The point shape denotes the descriptor set, a total of 5 shapes.

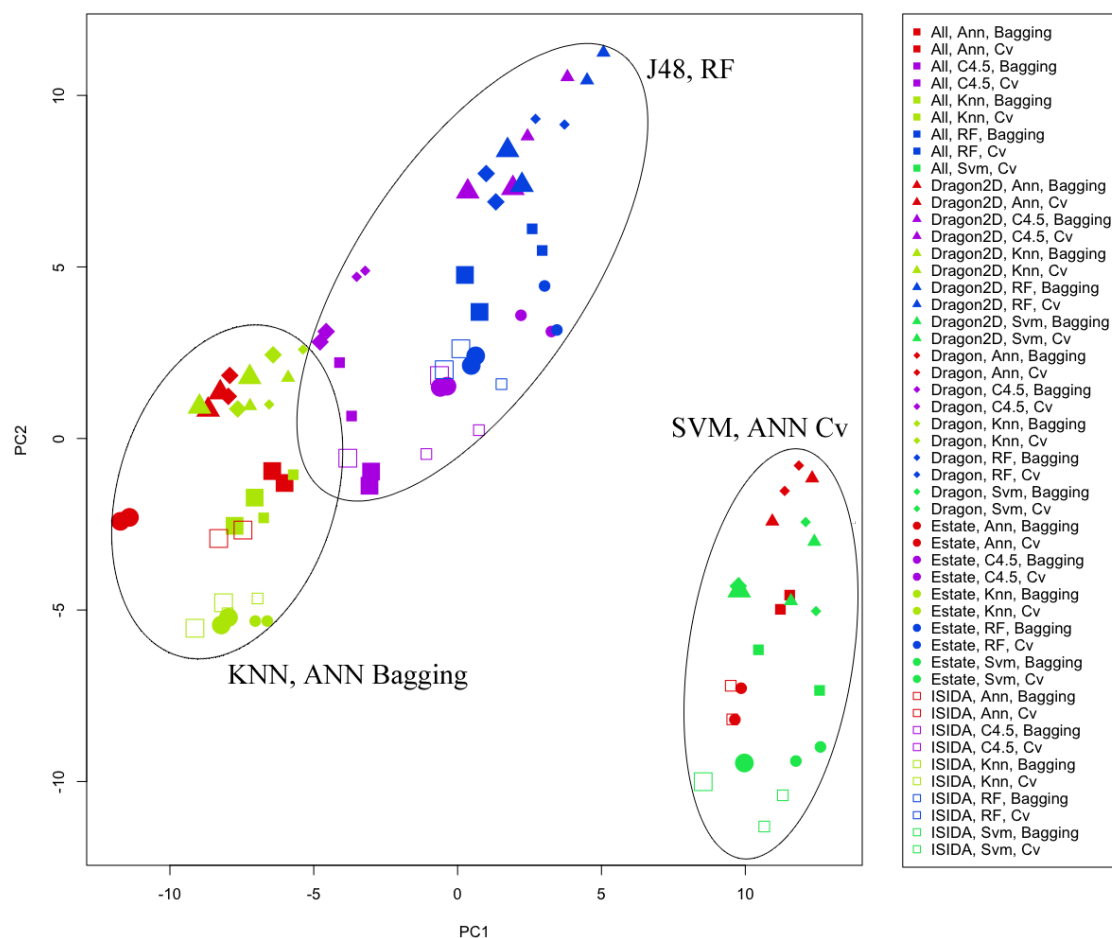


Figure 4.6. PCA plot of models in prediction space

It is visible that models are grouped into several distinctive clusters. The clusters are machine learning method based, that is, the distinctions between model predictions can be attributed to the difference in machine learning methods rather than descriptor sets. The three main clusters of models that can be observed are KNN and ANN-Bagging models, decision tree models, and SVM and ANN-Cv models.

Several conclusion can be made based on the plot:

- KNN is the only method for which the cross-validated and bagging models form a single cluster, that is, bagging has little influence on the KNN prediction results
- RF cross-validated and bagging results form separate clusters, but the clusters are nearby. This can be explained by the aggregating nature of the RF method itself. It internally uses a bagging procedure for individual trees to form a final results
- The bagging clusters are smaller and more dense, that is, the results of the bagging models are less dependent on the descriptor set and descriptor selection procedure used
- J48 models have the biggest variance of predictions among the studied methods
- In a majority of cases the decorrelation of descriptors had little impact on prediction results

4.2.4 Applicability domain of models

Figures 4.7 and 4.8 show two different chart types that illustrate the ability to differentiate accurate and inaccurate predictions for CYP1A2 models using the BAGGING-STD DM. Both figures 4 and 5 display the charts for all descriptors, no decorrelation, bagging models. Since bagging model could not be produced for full set of descriptors for SVM models, the SVM chart is missing.

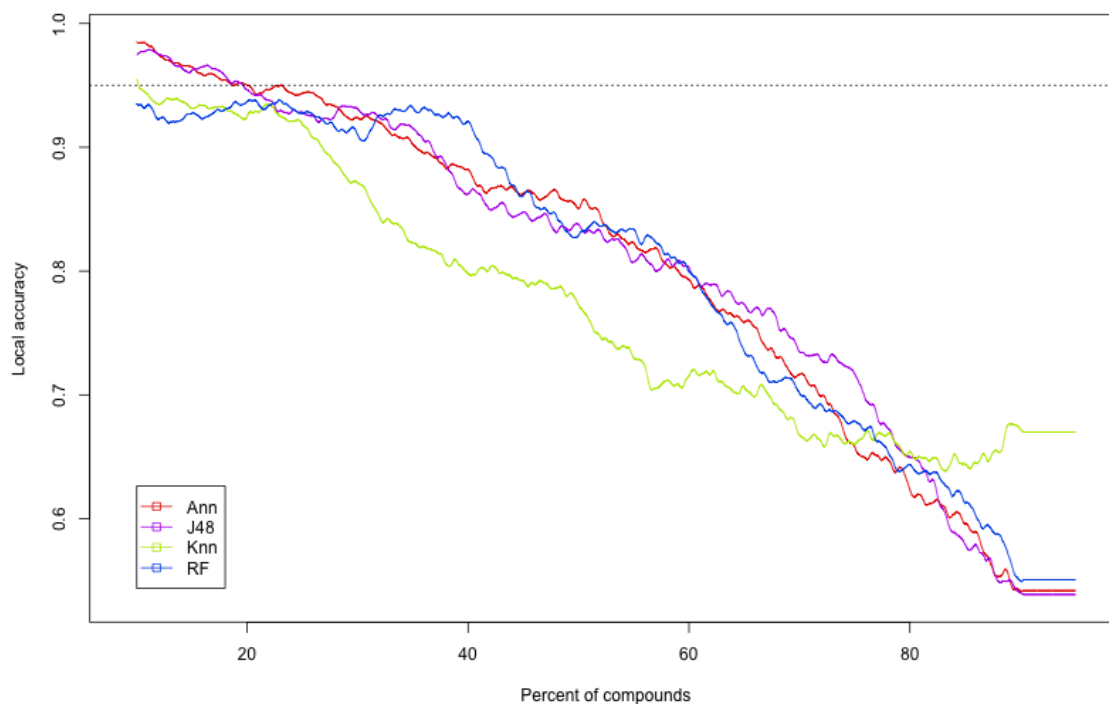


Figure 4.7. Local balanced accuracy of model predictions, when ordered by BAGGING-STD DM. The charts are shown for all descriptors, no decorrelation, bagging models only.

The charts are plotted for the internal test set compounds sorted by BAGGING-STD. Figure 4.7 displays the accuracy of predictions calculated as simple moving average over a window of 10% compounds. The plot shows the percentage of correct predictions in a window for each particular value of BAGGING-STD measure. For unification reasons the X axis of the plot does not display the BAGGING-STD value itself, but rather the percentage of compounds that have the BAGGING-STD value lower than a particular threshold. The plot has a general downward trend that shows a strong correlation of the prediction accuracy and a DM.

The molecules with higher balanced accuracies can be considered confidently predicted. The molecules with balanced accuracies around 50-60% can be treated as “randomly guessed” by the model and should be experimentally measured.

Figure 4.8 represents cumulative accuracy-coverage plots. This chart displays balanced prediction accuracy (y axis) for a group of compounds, having DM less than some threshold against percentage of this group of compounds in the whole set (x axis). The plot start from high accuracy values (for compounds with low BAGGING-STD measures) and drop to the level of approximately 83% - the average accuracy for the whole set.

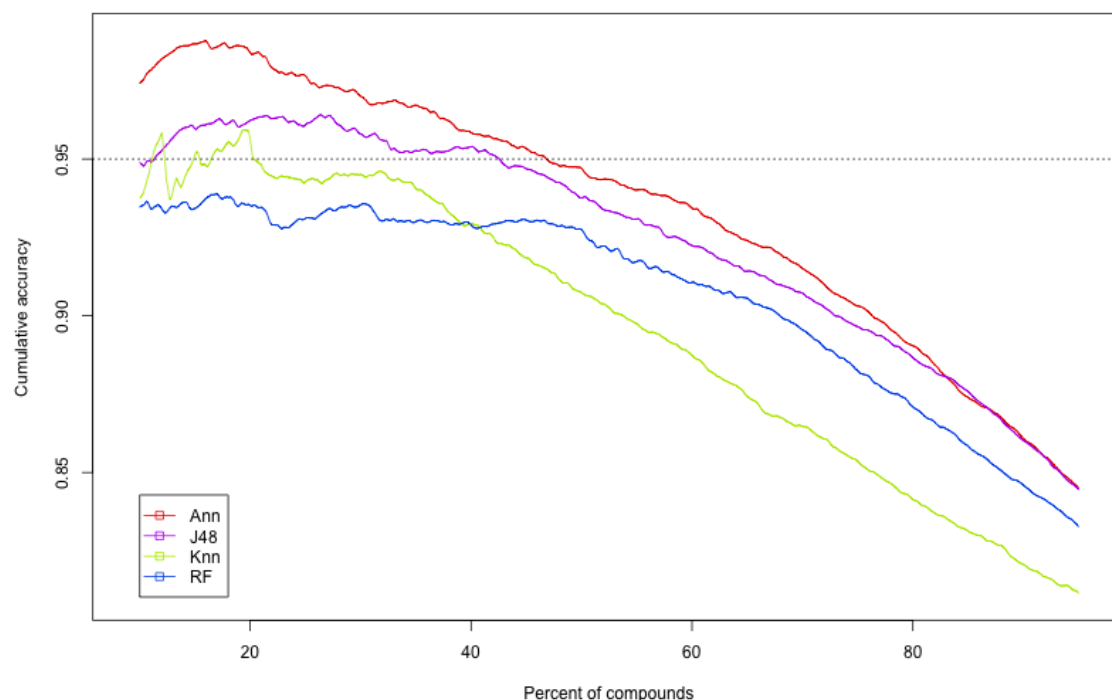


Figure 4.8. Overall balanced accuracy of model predictions for a subset of compounds as a function of the size of this subset, when ordered by BAGGING-STD DM. The charts are shown for ISIDA, decorrelated, bagging models.

We can see that BAGGING-STD measure is less successful for the KNN and RF methods. On the subset of 20% most confident predictions, however, the cumulative accuracy for KNN method allowed to reach the balanced accuracy of 95%. The BAGGING-STD measure was most successful for ANN method, allowing to achieve the balanced accuracy of 95% of correctly classified instances on a subset of approximately 43% of most confident predictions. This result is not significantly lower for the J48 method (balanced accuracy of 95% on a subset of 41% most confident predictions).

As we can see, the behavior of the plots is similar for different models. This means the BAGGING-STD DM worked universally, and was successfully used with all sets of descriptors and machine learning methods, as long as bagging approach was used.

4.2.5 External test set results

Table 4.4 shows the external set model performance measures of the 30 top models in this study. It also displays the balanced accuracy on a top 20% most confidently predicted molecules.

The overall balanced accuracy of predictions for this external dataset was 71% - 83% of correctly classified instances. The balanced accuracy of the top 20% most confidently predicted molecules is 83-96%. The high accuracy values on the subsets of most confidently predicted molecules, the size of the models, the computational intensity of the model training and application process, and the better balance between sensitivity and specificity make the ANN and J48 decision trees, developed using bagging, the most successful approaches in this study.

Descriptors	Method	Validation	Selection	ACC	BACC	SENS	SPEC	MCC	BACC(AD)
All	Ann	Bagging	None	0,79	0,77	0,91	0,63	0,58	0,92
All	J48	Bagging	None	0,81	0,8	0,87	0,74	0,61	0,95
All	Ann	Bagging	Decorr.	0,84	0,83	0,91	0,75	0,68	0,9
All	J48	Bagging	Decorr.	0,8	0,79	0,87	0,71	0,59	0,96
Dragon	J48	Bagging	None	0,8	0,79	0,86	0,72	0,59	0,94
Dragon	J48	Bagging	Decorr.	0,8	0,79	0,88	0,7	0,59	0,94
Estate	J48	Bagging	None	0,78	0,76	0,91	0,6	0,55	0,92
All	RF	Bagging	Decorr.	0,78	0,76	0,93	0,59	0,56	0,86
Dragon	RF	Bagging	Decorr.	0,78	0,75	0,92	0,58	0,55	0,85
Dragon	Ann	Cv	None	0,78	0,76	0,89	0,64	0,55	-
Estate	RF	Bagging	None	0,77	0,75	0,91	0,59	0,54	0,85
Dragon2D	RF	Bagging	Decorr.	0,78	0,75	0,92	0,59	0,55	0,87
All	Ann	Cv	None	0,78	0,76	0,91	0,62	0,56	-
Estate	J48	Bagging	Decorr.	0,78	0,76	0,91	0,61	0,55	0,92
All	Ann	Cv	Decorr.	0,72	0,71	0,78	0,64	0,43	-
Dragon2D	J48	Bagging	Decorr.	0,78	0,75	0,92	0,59	0,55	0,91
Dragon	Svm	Bagging	Decorr.	0,43	0,5	0	1	0	0
Estate	RF	Bagging	Decorr.	0,78	0,76	0,91	0,6	0,55	0,88
All	RF	Bagging	None	0,78	0,76	0,92	0,59	0,56	0,86
Dragon	RF	Bagging	None	0,78	0,75	0,92	0,58	0,55	0,83
Dragon	Ann	Bagging	Decorr.	0,79	0,77	0,9	0,65	0,57	0,94
All	Svm	Cv	Decorr.	0,82	0,82	0,8	0,84	0,64	-
Dragon	Ann	Cv	Decorr.	0,73	0,71	0,86	0,57	0,45	-
ISIDA	RF	Cv	None	0,75	0,74	0,86	0,61	0,49	-
Dragon2D	J48	Bagging	None	0,79	0,77	0,9	0,64	0,57	0,94
ISIDA	J48	Bagging	Decorr.	0,78	0,77	0,83	0,7	0,54	0,91
ISIDA	Ann	Cv	None	0,7	0,69	0,75	0,63	0,38	-
Dragon2D	RF	Bagging	None	0,78	0,76	0,9	0,62	0,56	0,87
Dragon	Ann	Bagging	None	0,8	0,78	0,91	0,66	0,59	0,93
ISIDA	RF	Bagging	Decorr.	0,79	0,77	0,91	0,62	0,57	0,9

Table 4.4. The performance of best 30 models for the external validation set of CYP1A2 inhibitors and non-inhibitors from PubChem BioAssay database. ANN – Associative Neural Networks[147,148], RF and J48 – random trees [152] and C4.5 pruned trees [153] as implemented in WEKA [154], SVM - support vector machines [149] as implemented in LibSVM [211]. Dragon, Dragon2D - 3D and 2D descriptors by software by Talete inc. [124], ISIDA - substructural molecular fragments as implemented in ISIDA [125], Estate - electrotopological state indices [127].

Figure 4.9 represents cumulative accuracy-coverage plot of expected and observed accuracies of the top model in the list (Ann, full descriptor set, bagging). The expected accuracies are plotted based on the DM to local accuracy relationship derived from the training set of the same model. We can see in Figure 4.9 that the behavior of the accuracy-coverage plot of the external test set is similar to the estimated behavior, however the actual accuracy is up to five percent higher than estimated accuracy.

The difference between the real and expected accuracies can be attributed both to the differences in the relationship between the DM measure and the accuracy of prediction for a specific compound in the training and the external test set, and to the different distribution of compounds with specific DM values in the training and the test set.

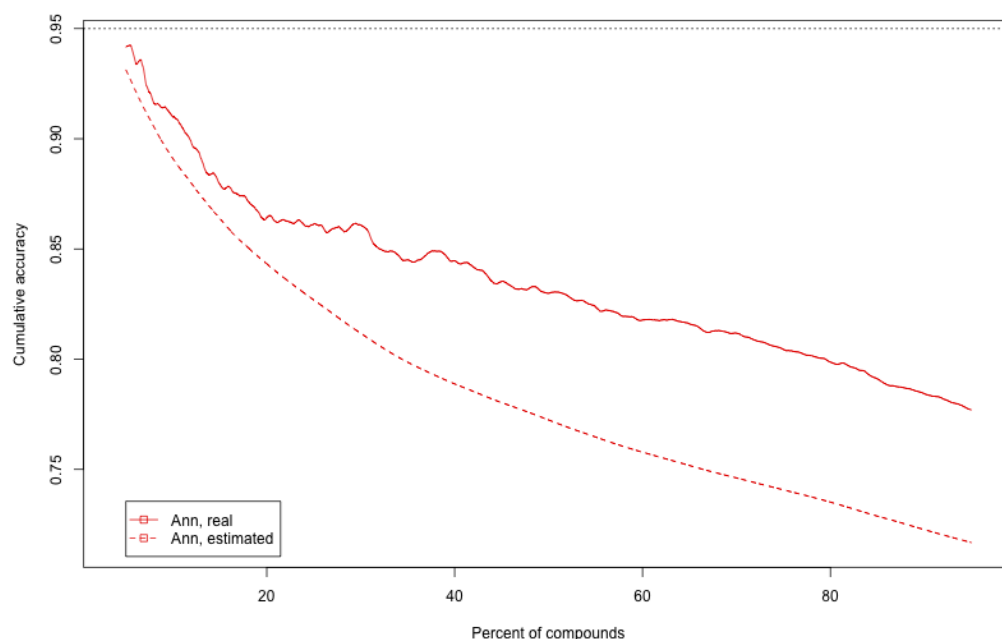


Figure 4.9 Cumulative balanced accuracy of model predictions, when ordered by BAGGING-STD DM for the training and external test sets. The charts are shown for full set of descriptors, ANN bagging model.

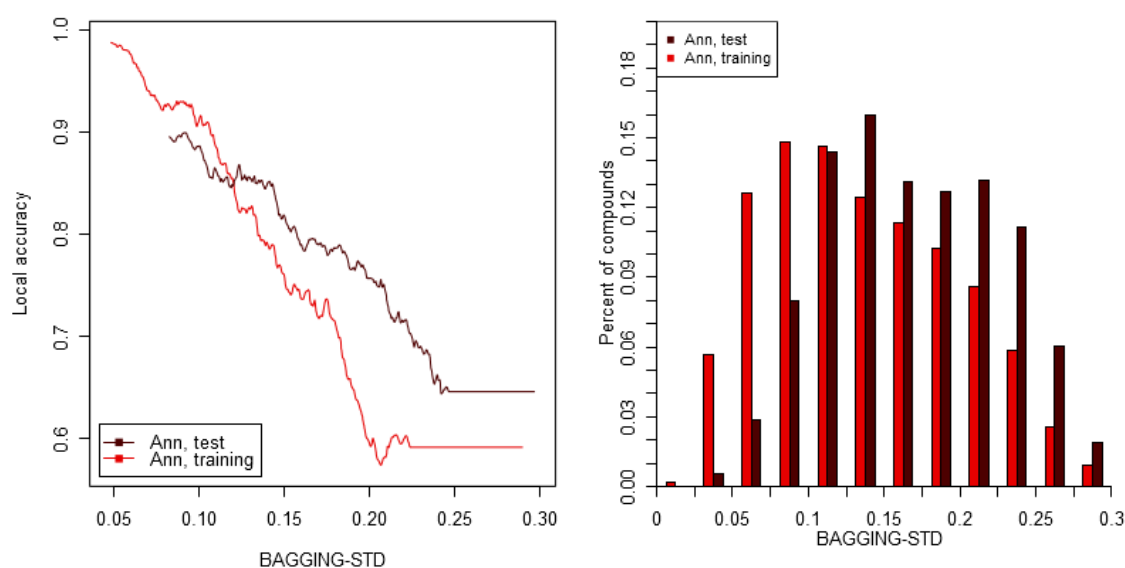


Figure 4.10. The relationship between the BAGGING-STD DM and local balanced accuracy for the training and the external test sets (left panel) – the chart shows that for the same values of BAGGING-STD DM the compounds from the training set generally have lower accuracy than those from the test set; the distribution of molecules with a particular BAGGING-STD DM value in the training and test sets (right panel) – the training set contains a significantly higher amount of compounds with lower BAGGING-STD DM values than the test set. The charts are shown for full set of descriptors, ANN bagging model.

Figure 4.10a shows the relationship between the BAGGING-STD DM and the local balanced accuracy, as described earlier. The molecules with the same BAGGING-STD DM value tend to have higher accuracy of predictions for the external test set. This means that the DM measure is somewhat pessimistic in estimating the prediction accuracy.

Figure 4.10b shows the distribution of molecules in the training and external sets by BAGGING-STD values. It can serve as a visual representation of differences between training and test set in prediction space.

The higher accuracy values for DM values of 0.15 - 0.3 and higher percentage (over 60%) of test set compounds having these values constitute to a higher than estimated prediction accuracy of the model on the test set.

4.2.6 Summary

In this part of the study, different QSAR approaches for the prediction of CYP1A2 inhibition were compared. Dragon full set, Dragon 2D set only, E-State and ISIDA SMF descriptors were used. The kNN, SVM, ASNN, RF and J48 methods were studied. Models built on the PubChem BioAssay A410 dataset were tested by cross-validation on the same set, and applied to predict an external test set from another assay - PubChem BioAssay A1851.

SVM and J48 models displayed highest accuracy among the used methods. The top performing model is SVM on a Dragon descriptor set with balanced accuracy of 83%. Several other models, including SVM models on ISIDA descriptors and J48 models on full descriptor set displayed the balanced accuracies over 82% and were non-significantly different from the top-performing model. On average, the 3D descriptors (Dragon set and full set) outperformed the 2D descriptors.

The decorrelation of descriptors had no influence on model accuracy, but greatly increased the speed of model creation and application.

For all methods, except for KNN, the bagging approach allowed a statistically significant increase of performance.

Based on PCA in model prediction space, three groups of models were determined: KNN and ANN-Bagging models, J48 and RF models, SVM and ANN-Cv models. Predictions of these groups of models form distinctive clusters on the PCA plot.

The external test accuracies for the models are 71% - 82% correctly classified instances. Using the BAGGING-STD measure allowed us to increase the accuracies to 83%-96% on about 10% of external set compounds. The top performing bagging model (ANN model on full set of descriptors) displayed 96% accuracy on 10% of most confident predictions.

As we have shown in section 4.1.1 (page 67), the prediction accuracy of the models on the most confidently predicted compounds is close to experimental accuracy of measurements for CYP inhibition. This proves that the models can be used to decrease the number of experiments on a subset of studied compounds.

4.3 Using novel descriptors in QSAR modeling of CYP 1A2, 2C9, 2C19, 2D6 and 3A4

4.3.1 Materials and methods

Datasets. For this study the PubChem AID410, AID883, AID899, AID891 and AID884 datasets were used as training sets for the models. The PubChem AID1851 Filtered datasets were used as test sets for the models.

Descriptors. The goal of this study is to determine whether descriptors derived from protein-ligand complex obtained by docking procedure can increase the predictive capabilities of QSAR models for CYP inhibition. Therefore the base descriptors were chosen based on the results of the previous study.

- *Estate*: E-State indices (atom type indices and bond type indices)
- *Dragon*: 0D-3D descriptors from Dragon package
- AP: docking-derived protein-ligand atom pair descriptors
- Estate+AP: A combination of two sets of descriptors
- Dragon+AP: A combination of two sets of descriptors

The Estate indices were chosen since this is the set of 2D descriptors that showed the best performance in the previous study. The Dragon descriptors was chosen as the set of 3D descriptors that showed performance statistically similar to the best performing model in the previous study.

The models built on combined sets will allow to evaluate whether docking-derived descriptors bring new information to the model and increase its performance.

All the models in this study are built with the use of descriptor selection procedure. The previous study showed that unsupervised correlation-based descriptor selection does not decrease model performance.

Machine learning methods. Based on the results of the previous study, the machine learning methods used were ASNN neural networks and J48 decision trees. These methods showed the best performance in predicting CYP1A2 inhibition. Both methods were chosen for their complementary nature - neural networks is a regression-based machine learning method, while J48 is a classification decision tree.

All models were built with bagging meta-learning method. The number of model instance in each bag was 32 for ANN models and 512 for J48 models. Higher number of instances for J48 models further increased model performance, as well as provided higher

resolution for applicability domain measurements.

Applicability domain. In the previous study the BAGGING-STD DM was shown to be successful in CYP modeling task, therefore the same DM was used in the current study.

4.3.2 Modeling results

For each isoform in this study a total of 10 models were built (two different machine learning methods, five different descriptor sets). To keep the results consistent, balanced accuracy (BACC) was used to assess models predictive abilities. This metrics is especially useful in case of highly imbalanced datasets. For isoforms 2C9 and 2D6 the ratio of non-inhibitors to inhibitors is as high 5 to 1. Using average accuracy for these datasets might produce misleading results.

Table 4.5 shows the balanced accuracies for all models built in this study.

Descriptors	Method	BACC				
		CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP3A4
Dragon+AP	J48	0.835	0.833	0.827	0.849	0.87
Estate+AP	J48	0.833	0.807	0.803	0.843	0.861
AP	J48	0.823	0.773	0.752	0.779	0.821
Dragon	J48	0.817	0.811	0.807	0.832	0.854
Estate	J48	0.817	0.803	0.799	0.838	0.854
Dragon+AP	Ann	0.824	0.807	0.816	0.833	0.864
Estate+AP	Ann	0.822	0.788	0.783	0.817	0.846
AP	Ann	0.795	0.742	0.704	0.766	0.803
Dragon	Ann	0.808	0.8	0.809	0.825	0.856
Estate	Ann	0.777	0.77	0.781	0.798	0.809

Table 4.5. The performance of the models for CYP inhibitors and non-inhibitors. ANN – Associative Neural Networks [147,148], J48 – C4.5 pruned trees [153] as implemented in WEKA [154]. Dragon - 3D descriptors by software by Talete inc. [124], Estate - electrotopological state indices [127], AP - docking-derived protein-ligand atom pair descriptors (section 2.4.2, page 22).

In this and all following prediction accuracy tables yellow background cells designate the best models and models statistically similar to the best models. Green background cells represent the second group of models, statistically similar between themselves, but statistically worse than the best model. Blue background cells represent the rest of the models with lower performance.

We can see that for this study the decision tree machine learning method was more successful than the neural networks. For each of the isoforms the best performing model is the J48 decision tree model, and CYP3A4 is the only isoform for which the neural networks managed to produce the model statistically similar to the best performing model. On average model balanced accuracy is 2 - 3% higher for the decision tree model and the neural network model built on the same descriptors for the same CYP isoform.

The models where atom pair descriptors were the only descriptor set used performed among the worst in the study (balanced accuracy of 77% - 82% depending on the isoform in question). This can be explained by the design of these descriptors. The

atom pair descriptors only describe interaction between the atoms of the small molecule and the protein and contain no information about structural arrangement of atoms within a molecule or any additional molecular properties. This way only the atom pair descriptors do not contain enough information to produce predictive models and are meant to be used in combination with traditional molecular descriptors.

When comparing models with and without atom pair descriptors we can see that in all models regardless of the molecular descriptors and machine learning method used and regardless of the CYP isoform being modeled, addition of atom pair descriptors to the descriptor set increased model performance. In 12 out of 16 cases studied (4 isoforms, 2 base descriptor types, 2 machine learning methods) the increase of model performance was statistically significant with the significance level of 0.05.

The best performing model for all isoforms is built on a combination of Dragon and atom pair descriptors and has the balanced accuracy of 83% - 87%.

4.3.3 PCA plot model comparison

Figure 4.11 displays a PCA plot of the researched set of models in the space of predictions. Each point on the plot represents a single model, a total of 50 studied models. Point colors designate descriptors, point sizes - machine learning methods (smaller points - neural networks, bigger point - decision trees). Point shape represent the CYP isoform for which the model is built.

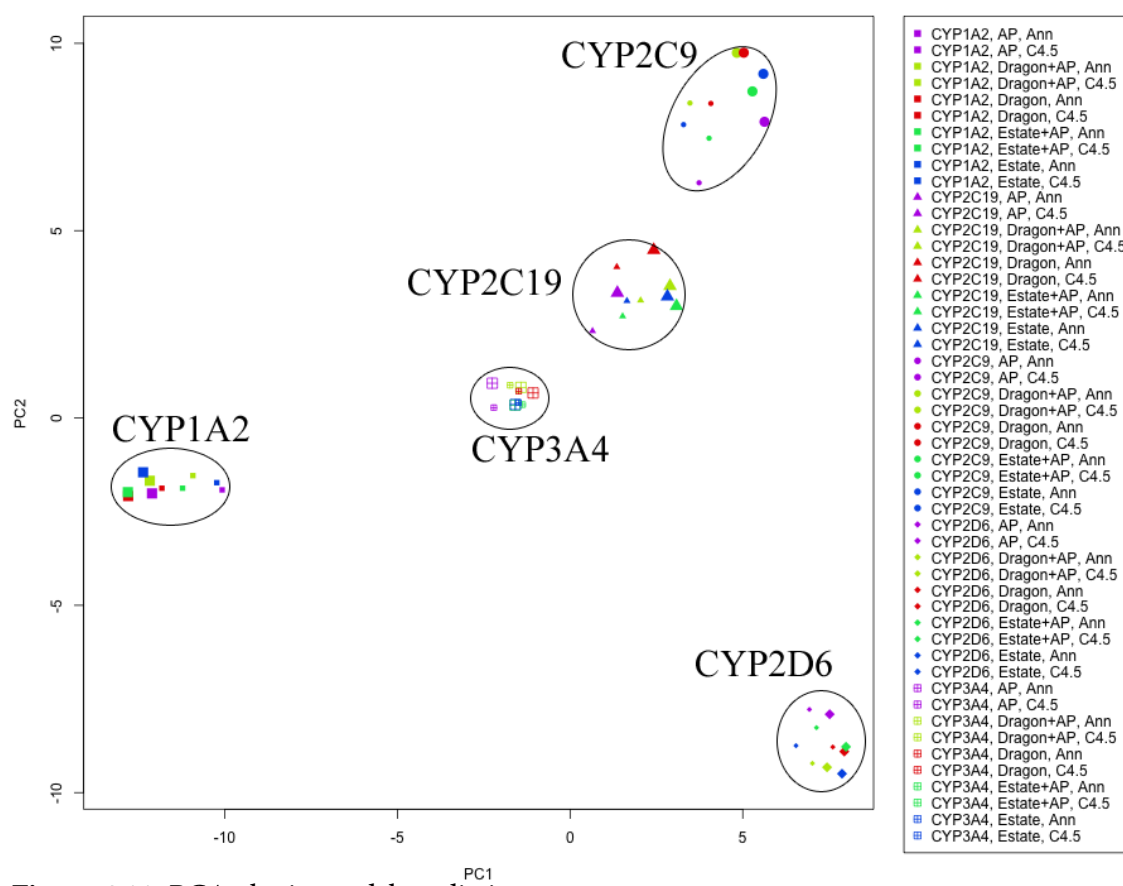


Figure 4.11. PCA plot in model prediction space

Several conclusion can be made based on the plot:

- The main characteristic defining the clusters on the plot is the CYP isoform - all five isoforms form distinctive clusters.
- While CYP1A2 and CYP2D6 clusters are isolated and far apart, CYP2C9, CYP2C19 and CYP3A4 are closer together.
- The placement of the clusters reflects the conclusions of the dataset analysis from the previous section. CYP2C9 and CYP2C19 clusters are close to each other. CYP3A4 cluster is closer to CYP2D6, CYP1A2, CYP2C9 than these clusters among themselves. CYP3A4 cluster is especially close to CYP2C19 cluster.
- Inside each individual CYP isoform cluster sub-clusters by machine learning method can be seen. This confirms the conclusions from the previous study.

4.3.4 Applicability domain of models

Figure 4.12 shows cumulative and local balanced accuracy applicability domain charts. These types of charts were introduced in the previous study. The charts are presented for CYP1A2 isoform only. The charts for other isoforms can be found in the appendix section, Figure A1.

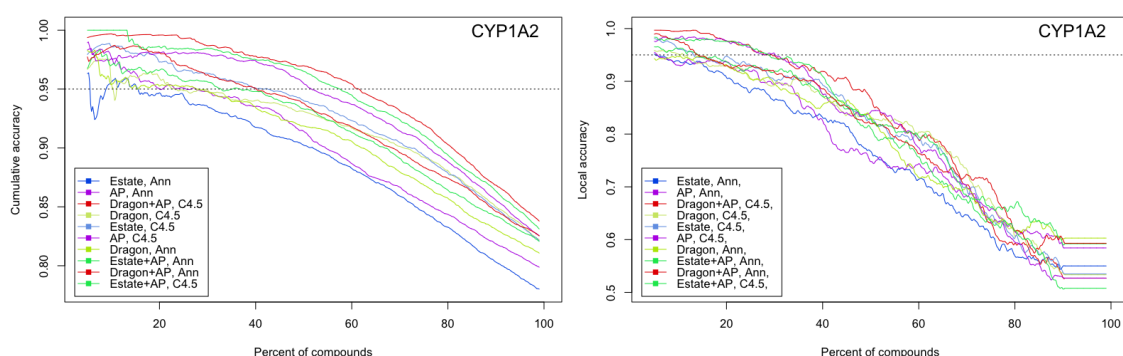


Figure 4.12. Cumulative (left) and local (right) balanced accuracies of model predictions, ordered by BAGGING-STD DM.

The cumulative charts (Figure 4.12, left) display balanced prediction accuracy (y axis) for a group of compounds, having DM less than some threshold against percentage of this group of compounds in the whole set (x axis). The plot start from high accuracy values (for compounds with low BAGGING-STD measures) and drop to the the average balanced accuracy for the whole set. These average values are also reported in Table 4.5.

Local accuracy charts display the accuracy of predictions calculated as simple moving average over a window of 10% compounds. The plot shows the percentage of correct predictions in a window for each particular value of BAGGING-STD measure. These charts reflect the estimated accuracy of predictions of each individual compound with a specific BAGGING-STD value, which makes it different from the cumulative charts, that display average balanced accuracies of groups of compounds that have BAGGING-STD values less than a certain threshold.

Table 4.6 displays training set validated accuracies for top 20% most confident predictions.

Descriptors	Method	BACC for top 20% most confident predictions				
		CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP3A4
Dragon+AP	J48	0,995	0,97	0,987	0,988	0,994
Estate+AP	J48	0,985	0,981	0,97	0,99	0,992
AP	J48	0,98	0,971	0,902	0,972	0,983
Dragon	J48	0,956	0,936	0,951	0,976	0,981
Estate	J48	0,973	0,955	0,943	0,98	0,983
Dragon+AP	Ann	0,982	0,951	0,97	0,97	0,987
Estate+AP	Ann	0,961	0,954	0,947	0,965	0,984
AP	Ann	0,95	0,918	0,88	0,919	0,965
Dragon	Ann	0,954	0,93	0,938	0,974	0,978
Estate	Ann	0,948	0,917	0,924	0,964	0,966

Table 4.6. The performance of the models for CYP inhibitors and non-inhibitors for top 20% most confident predictions of the validated training sets. ANN – Associative Neural Networks [147,148], J48 – C4.5 pruned trees [153] as implemented in WEKA [154]. Dragon - 3D descriptors by software by Talete inc. [124], Estate - electrotopological state indices [127], AP - docking-derived protein-ligand atom pair descriptors (section 2.4.2, page 22).

We can see that BAGGING-STD works well for differentiating between confident and unconfident predictions for all five studied isoforms. Large fractions of the datasets could be predicted with balanced accuracy of 0.95. Table 4.7 summarizes model performance.

Isoform	Best performing models	Fraction of the set predicted by the best model with given BACC	
		BACC=0.95	BACC=0.90
CYP1A2	C4.5 AP+Dragon	60%	83%
CYP2C9	C4.5 AP+Estate	43%	62%
	C4.5 AP+Dragon		73%
CYP2C19	C4.5 AP+Dragon	30%	65%
	C4.5 AP+Estate		60%
	Ann AP+Dragon		62%
CYP2D6	C4.5 AP+Dragon	63%	82%
	C4.5 AP+Estate		
CYP3A4	C4.5 AP+Dragon	71%	90%

Table 4.7. Fractions of the datasets predicted with a given accuracy

The most successful modeling technique that allowed to achieve the highest balanced accuracy results was the decision trees with a combination of atom pair and Dragon descriptors.

Models built on combined descriptors (AP+Estate and AP+Dragon models) outperformed models with no protein-ligand atom pair information both for BACC=0.95 and BACC=0.90 thresholds. For BACC=0.95 adding atom pair descriptors increased the fraction of the dataset predicted with this accuracy by 7% - 10% depending on the isoform.

We can also see that decision trees in general were more successful regardless of the descriptor set.

Applicability domain methodology also worked for models built entirely on atom pair descriptors - there is an obvious correlation between the actual prediction accuracy of the compound and its BAGGING-STD measure.

4.3.5 Application of models to the external test sets

This section describes the results of applying the models from the study to external test sets. As described in the dataset section, the training and test sets consist of data measured using the same methodology. The test sets, however, include molecules from a larger number of libraries and, therefore, can be used to model a real life scenario where models are used to replace measurements in novel sections of chemical space. Table 4.8 shows the balanced accuracy values of applying the models from the previous section to the AID1851 datasets for cytochromes 1A2, 2C9, 2C19, 2D6 and 3A4.

Descriptors	Method	BACC				
		CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP3A4
Dragon+AP	J48	0.791	0.775	0.748	0.763	0.784
Estate+AP	J48	0.777	0.713	0.729	0.736	0.744
AP	J48	0.762	0.723	0.702	0.68	0.717
Dragon	J48	0.777	0.736	0.722	0.74	0.754
Estate	J48	0.758	0.687	0.711	0.731	0.756
Dragon+AP	Ann	0.79	0.757	0.746	0.756	0.781
Estate+AP	Ann	0.78	0.727	0.718	0.744	0.767
AP	Ann	0.75	0.698	0.677	0.701	0.723
Dragon	Ann	0.77	0.732	0.734	0.738	0.761
Estate	Ann	0.749	0.693	0.701	0.705	0.745

Table 4.8. The performance of the models for CYP inhibitors and non-inhibitors for the external test sets. ANN – Associative Neural Networks [147,148], J48 – C4.5 pruned trees [153] as implemented in WEKA [154]. Dragon - 3D descriptors by software by Talete inc. [124], Estate - electrotopological state indices [127], AP - docking-derived protein-ligand atom pair descriptors (section 2.4.2, page 22).

We can see that the accuracy of the results of application of the models to the external test sets is significantly lower compared to the bagging-validated accuracies derived from the training sets. The drop in prediction accuracy ranges from about 4% (for the top performing model for CYP1A2 inhibition) to as much as 9% (for the top performing models for CYP2D6 and CYP3A4 inhibition). This drop in performance was expected and was due to the higher chemical diversity of the test sets and the presence of additional chemical libraries (that represent different chemical classes) in the test sets. (see “Dataset analysis and interpretation” section).

Applicability domain methods can help separate the reliable model predictions (predictions for the compounds similar enough to the compounds in the training sets) from unreliable ones (predictions of novel chemical scaffolds, unfamiliar to the

model). The threshold of 20% most confident predictions was chosen and cumulative accuracy on this fraction of the test sets were calculated. Table 4.9 shows the results of the models, when only 20% of most confident predictions (according to the applicability domain measure) are considered.

Descriptors	Method	BACC for top 20% most confident predictions				
		CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP3A4
Dragon+AP	J48	0,964	0,898	0,871	0,908	0,941
Estate+AP	J48	0,922	0,834	0,837	0,886	0,914
AP	J48	0,918	0,811	0,829	0,829	0,902
Dragon	J48	0,924	0,832	0,865	0,861	0,916
Estate	J48	0,891	0,77	0,821	0,862	0,894
Dragon+AP	Ann	0,897	0,865	0,863	0,88	0,921
Estate+AP	Ann	0,887	0,807	0,827	0,849	0,919
AP	Ann	0,886	0,788	0,808	0,798	0,88
Dragon	Ann	0,869	0,84	0,829	0,839	0,905
Estate	Ann	0,848	0,793	0,807	0,783	0,887

Table 4.9. The performance of the models for CYP inhibitors and non-inhibitors for top 20% most confident predictions of the external test sets. ANN – Associative Neural Networks [147,148], J48 – C4.5 pruned trees [153] as implemented in WEKA [154]. Dragon - 3D descriptors by software by Talete inc. [124], Estate - electrotopological state indices [127], AP - docking-derived protein-ligand atom pair descriptors (section 2.4.2, page 22).

Based on the relationship between the BAGGING-STD DM and actual prediction accuracy derived from the training set, we evaluated the expected model accuracy based on DM values for predicted compounds. This estimation was based on the assumption that the relationship between the DM values and prediction accuracy was exactly same for the training and the test sets. Figure 5 features three plots for each CYP isoform studied: the cumulative applicability domain plot with the actual and estimated accuracy curves, the local accuracy plot for the training and test sets, and the applicability domain bar plot for the training and test sets. All the plots are built for the top performing model for each isoform (the C4.5 model based on Dragon + Docking descriptors).

We can see in Figure 4.13 that the behavior of the accuracy-coverage plots of the external test sets is similar to the estimated behavior. The plots are given for CYP1A2 isoform only. Plots for other studied isoforms are available in the appendix, Figure A2.

There's one general pattern for all the accuracy-coverage plots in Figure 4.13. The real accuracies are lower than the estimated ones in the area of high-confidence predictions and higher than the estimated ones in area of all predictions. That is, using AD estimated accuracy values would be over-optimistic for a fraction of most confident predictions and over-pessimistic for the whole dataset.

One reason for this is the difference in the relationship between the DM measure and the accuracy of prediction for a specific compound in the training and the external test set. As we can see in Figure 4.13 local accuracy plots, the actual accuracy for the molecules with low BAGGING-STD DM values is lower for the test set than for the training set. This makes the accuracy over-optimistic for the most confident predictions.

4.3 Using novel descriptors in QSAR modeling of CYP 1A2, 2C9, 2C19, 2D6 and 3A4

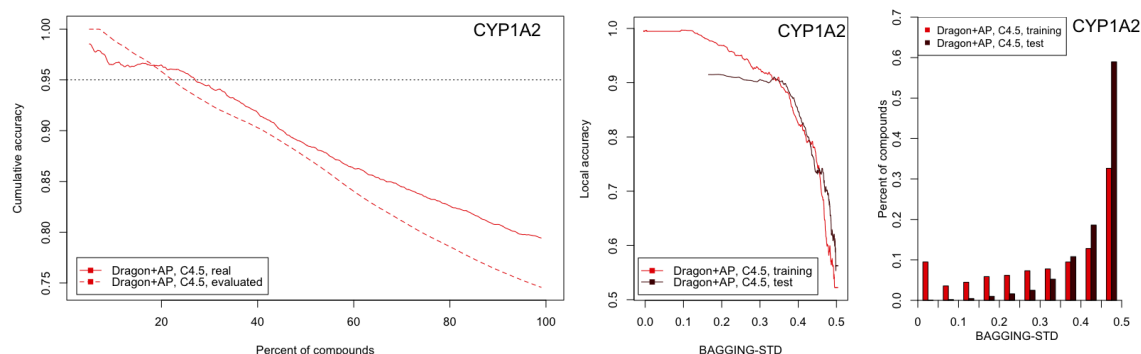


Figure 4.13. Actual and estimated cumulative applicability domain accuracy plots. Displayed model is C4.5 decision tree model built on Dragon + atom pair descriptors.

The distribution of molecules among different DM values is another reason. The large fraction of the test sets (50-60%) have fallen to the DM area of 0.4-0.5, which indicates low confidence of the model in the accuracy of predictions.

The DM distribution bar chart (which has some disadvantages, since it relies on a particular DM measure and the distribution of its values in the training and test sets) was replaced for further analysis by a percent-based DM distribution bar chart (Figure 4.14). In Figure 4.14 the boundaries for plot bins are given in “percentage of the training set” scale rather than DM values scale. In the “percentage of the training set” scale the actual DM values for the boundaries are chosen in a way that would make the DM distribution bar chart universally distributed. This way the deviation of the test set bar chart from the universal distribution would serve as model prediction based dataset similarity measure.

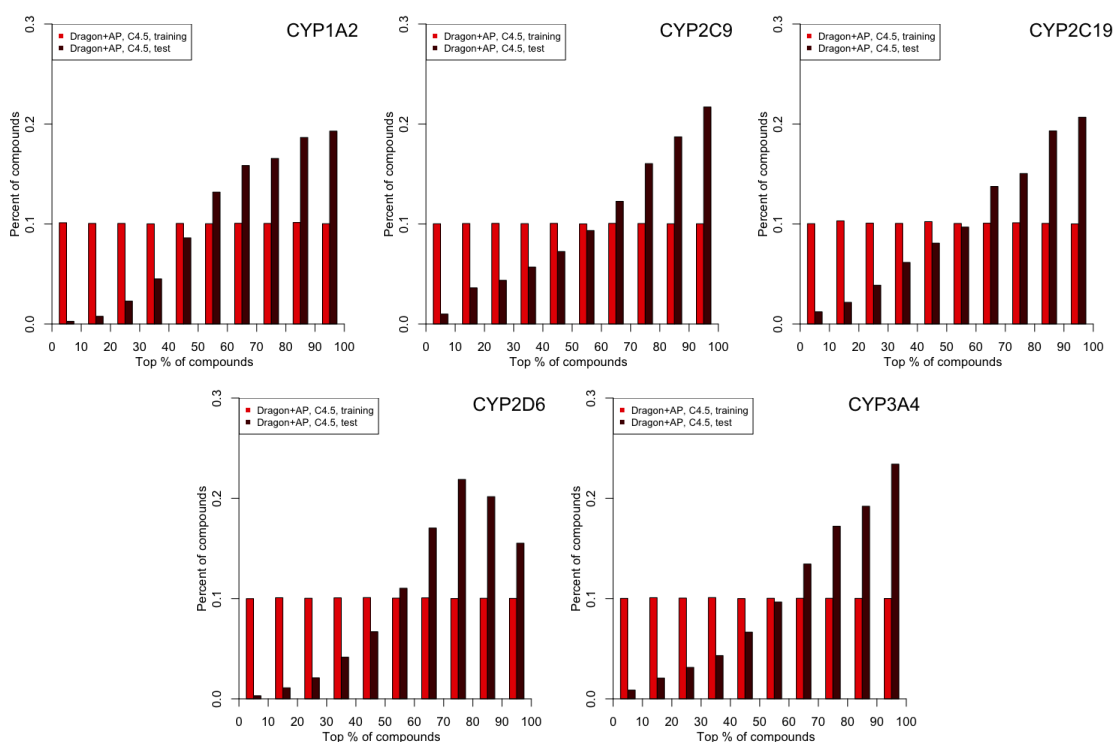


Figure 4.14. Balanced DM distribution plot for training and test sets (C4.5 models built on Dragon + AP descriptors).

Note that these charts do not require experimentally measured values for the test set and therefore can be built even for a virtual dataset of compounds to estimate its similarity to the model's training set.

As we can see on Figure 4.14, the test datasets in this study are dissimilar to the training sets; therefore, the drop in model performance on the test set compared to the cross-validated training set is to be expected.

Since both prediction accuracy and prediction confidence are non-uniformly distributed among different chemical classes, applicability domain based fragment analysis can be helpful to identify substructures reliably or unreliably predicted by the model. Figure 4.15 and Figure 4.16 display three most confidently predicted and three most unconfidently predicted fragments for CYP1A2 isoform. Similar diagrams for other cytochrome isoforms can be found in the appendix, Figure A3 - Figure A12.

Fragments in Figure 4.15 and Figure 4.16 were selected in such a way so that at least a 100 molecules would contain each of them. If two fragment-containing molecule groups contain the same amount of molecules and one of the fragments is the exact subfragment of the other, the bigger fragment was selected. For each fragment-containing molecule group average BAGGING-STD DM was calculated. The top three and bottom three fragments were selected as “most confidently predicted fragments” and “least confidently predicted fragments”, respectively.

High confidence of predictions for certain fragment-containing molecules indicate that the training set contained a big amount of diverse compounds containing this fragment and adding more compounds with this substructure will not significantly improve the overall model performance.

Low confidence of predictions for certain fragment-containing molecules indicate absence or low number of compounds of this chemical class in the model training set. When using the prediction models for decision support in experiment planning, it is beneficial to select these kinds of molecules for experimental testing. Introducing measured values for these molecules would have the highest benefit for the overall model performance. This also may indicate that the fragment itself possesses properties that may prevent reliable activity measurement or prediction (i.e. - interaction with the solvent).

We can see that the most confidently predicted fragments are linear and branched fragments containing 4 - 6 atoms, among which - carbon, nitrogen and oxygen. Of 15 presented fragments only two contain aromatic rings and one - non-aromatic circular structure.

Least confidently predicted fragments are mostly aromatic (12 out of 15 presented fragments contain aromatic rings). On average they contain more atoms and have a higher molecular weight. One particular fragment (trifluoromethylbenzene) was marked as least confidently predicted for three isoforms out of five (CYP2C9, CYP2C19 and CYP3A4).

4.3 Using novel descriptors in QSAR modeling of CYP 1A2, 2C9, 2C19, 2D6 and 3A4

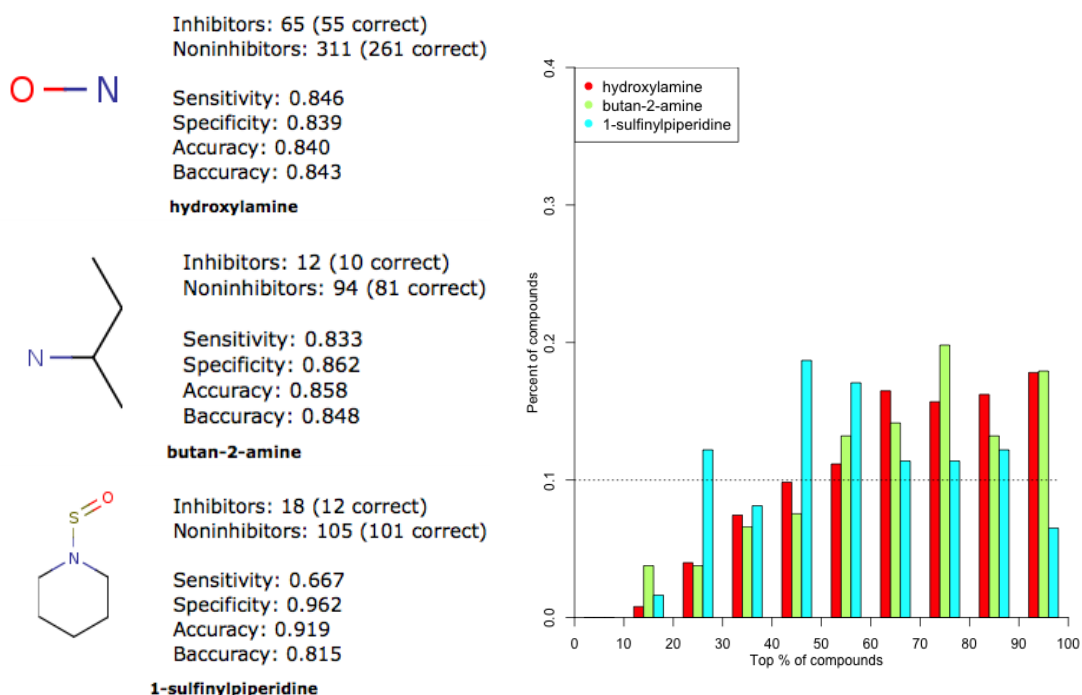


Figure 4.15. Diagram of best-predicted fragments for CYP1A2 isoform

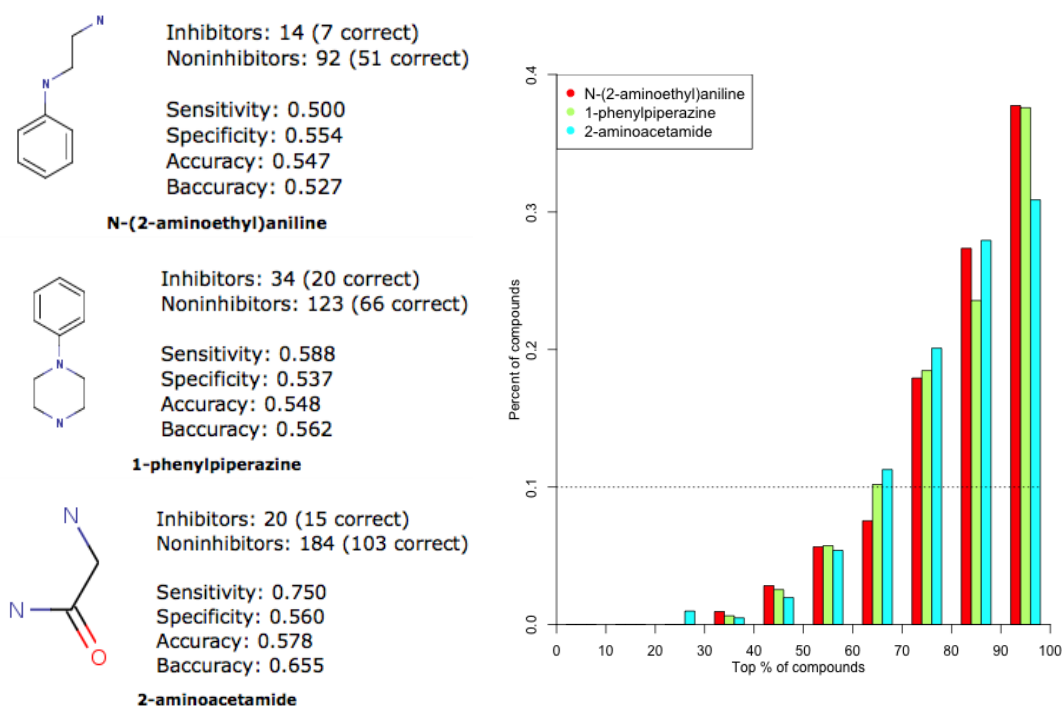


Figure 4.16. Diagram of worst-predicted fragments for CYP1A2 isoform

4.3.6 Summary

In this part of the study the most successful descriptors (Dragon descriptors and Estate indices) and machine learning methods (ASNN and C4.5 decision tree) determined in section 4.2 were used to model CYP inhibition activity for five different isoforms. Additionally, protein-ligand atom pair descriptors were used in the study in combination with traditional descriptors to benchmark their predictive abilities.

Confirming the results from the study in section 4.2, the C4.5 decision tree models were consistently more predictive, with an average increase of 2-3% of correctly classified instances.

Models containing Dragon descriptors were on average more predictive than Estate indices. The atom pair descriptors provided lower prediction accuracy, with the average balanced accuracy of 77% - 82% depending on the isoform.

However, the best performing models (with the significance value of 0.05) for all isoforms were built on a combination of Dragon and atom pair descriptors and had the balanced accuracy of 83% - 87%.

The PCA analysis in model prediction space confirmed the dataset similarity analysis results from section 4.1.2 summarized in Table 4.2, page 68. CYP2C9 and CYP2C19 isoform activity results formed the closest clusters. The CYP3A4 cluster was the closest cluster to all other isoforms.

The external test accuracies for the models are 74% - 79% correctly classified instances for the best performing model (C4.5 decision tree, Dragon and atom pair descriptors). A drop in prediction accuracy can be explained by a significant structural difference between training and test sets.

Using the BAGGING-STD measure allowed us to increase the accuracies to 87%-96% on about 20% of external set compounds for the top performing model. The detailed applicability domain analysis showed that the used applicability domain approach is somewhat optimistic in estimating model accuracy.

Fragment-based applicability domain analysis determined the fragments predicted with more than average and less than average confidence. The molecules that were predicted most confidently contained linear and branched fragments with the size of 4 - 6 atoms, among which - carbon, nitrogen and oxygen. Molecules, which were predicted with the lowest confidence contained fragments that were mostly aromatic and on average contained more atoms and had a higher molecular weight. One particular fragment (trifluoromethylbenzene) was marked as least confidently predicted for three out of five isoforms (CYP2C9, CYP2C19 and CYP3A4).

The fragments overrepresented in least confidently predicted molecules can be used as structural hints for additional experimental measurements.

4.4 Novel descriptors in predicting CYP2C19 activity based on CYP2C9 dataset

4.4.1 Materials and methods

In this part of the study we research the possible methods to extrapolate activity prediction across closely related cytochrome targets, measure the accuracies of possible extrapolation approaches and evaluate the practical applications of these methods. For this study we have chosen the cytochromes from CYP2C subfamily: the models built on CYP2C9 dataset are applied to predict CYP2C19 activity.

The datasets involved in this study are PubChem AID883 for CYP2C9 data and PubChem AID899 and AID1851 for CYP2C19 data.

The idea of the study is based on the fact that docking-derived atom pair descriptors are based on the protein-ligand complex structure rather than small molecule structure. The hypothesis is that for closely related protein structures same atom pair descriptors retain (qualitatively and quantitatively) the same relation to the modeled activity (in our case - CYP inhibition activity). Therefore it would be possible to reuse the models built on CYP2C9 data to predict inhibition activity of small molecules for CYP2C19.

We compare the performance of QSAR model in three different scenarios. All the three scenarios focus on prediction CYP2C19 inhibition activity for new molecules, but differ in the amount of information that is used in the model creation process.

QSAR for CYP2C19 data. The most straightforward scenario (that is labeled “CYP2C19 to CYP2C19”) is traditional QSAR. We assume that we have some amount of experimental data for the target itself we are interested in - CYP2C19 inhibition. We then build the QSAR model on this data and use it to predict novel compounds. For this scenario we mirror all the conditions from the previous study.

We use PubChem AID899 dataset as the training set and build ten different QSAR models. These models were built using two different machine learning approaches (ASNN - neural networks, and J48 - decision trees) for five different descriptor sets (Estate indices only, Dragon descriptors only, atom pair descriptors only, Estate indices with atom pair descriptors, Dragon descriptors with atom pair descriptors). Similarly to the previous study, BAGGING-STD DM was used to evaluate prediction accuracy on the subset of most confident predictions.

The models are then used to predict both the bagging-validated results for the same PubChem AID899 set and the external test set represented by PubChem AID1851 dataset.

QSAR for CYP2C9 data, naive extrapolation to CYP2C19 activity. The “naive” scenario (labeled “CYP2C9 to CYP2C19 (naive)”) can be used when there's no available experimental data for the target itself (CYP2C19 in our case). We then create traditional QSAR models for the closely related target with available experimental data (CYP2C9 in our case).

That is, we use PubChem AID883 dataset as the training set and build ten different QSAR models for CYP2C9 inhibition activity.

Given no additional data, we just assume that CYP2C9 and CYP2C19 activities are completely same and under this assumption try to predict CYP2C19 activity by CYP2C9 models.

As we have shown in Table 4.2 (page 68), CYP2C9 and CYP2C19 datasets are strongly correlated and the number of compounds sharing the same activity values for both cytochromes reaches about 80-90% of a significantly diverse set of compounds. Due to this fact our current scenario may produce a reasonable amount of accurate CYP2C19 inhibition predictions.

QSAR for CYP2C9 data, novel extrapolation to CYP2C19 activity. This novel scenario (labeled “CYP2C9 to CYP2C19 (novel)”) is also used when no experimental data for CYP2C19 is available and uses CYP2C9 model to predict CYP2C19 activity.

The model creation stage is exactly same as in the previous scenario: we use PubChem AID883 dataset as the training set and build ten different QSAR models for CYP2C9 inhibition activity. The important part that makes this scenario possible is including the protein-ligand complex descriptors to the modeling process.

The difference from previous scenario is in the model application process. The descriptors for the predicted structures are calculated based on CYP2C19 protein-ligand complexes rather than CYP2C9 complexes. This introduces new information to the modeling process and enhances the prediction results.

The process of applying a model to a set of compounds in this scenario therefore differs from the traditional QSAR approach in the phase of descriptor calculation. The molecule-based descriptors (Estate indices, Dragon descriptors) are calculated as usual. The protein-ligand atom pair descriptors are calculated on CYP2C19 protein structure rather than CYP2C9 structure. The combined sets of descriptors are then used in CYP2C9 model to predict CYP2C19 inhibition activity. The diagram of the process is displayed on Figure 4.17.

Same as for the previous scenario, the models were applied to the two available CYP2C19 datasets to make the models comparable to other scenarios.

Applying the suggested approaches to both PubChem AID899 and PubChem AID1851 CYP2C19 datasets models the real life scenarios of virtual screening of compounds to predict activities of new structures on a target based on experimental data for a closely related target. The difference is that PubChem AID899 experiments reflect the situation when the screened compounds and the compounds in the training set come from the same molecular library and therefore contain a high amount of structurally similar entities. The PubChem AID1851 CYP2C19 experiments represent screening of more structurally diverse molecules.

Both situations may arise in early stage drug discovery when the available amount of measurements for the specific cytochrome is used to screen potential drug candidates for inhibition activity not only for this cytochrome but also for its most probable variations.

Figure 4.18 illustrates which datasets were used as training and test sets in each experiment outlined in this chapter.

4.4 Novel descriptors in predicting CYP2C19 activity based on CYP2C9 dataset

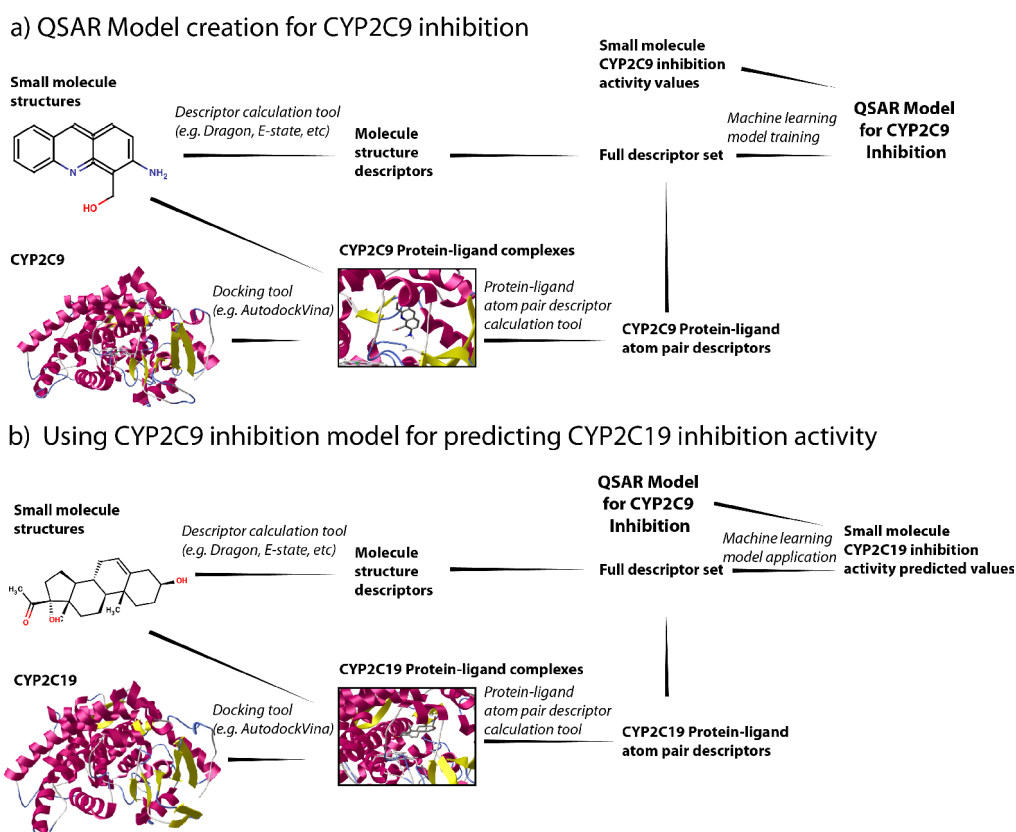


Figure 4.17. QSAR model creation and application processes for CYP2C9 to CYP2C19 novel extrapolation scenario

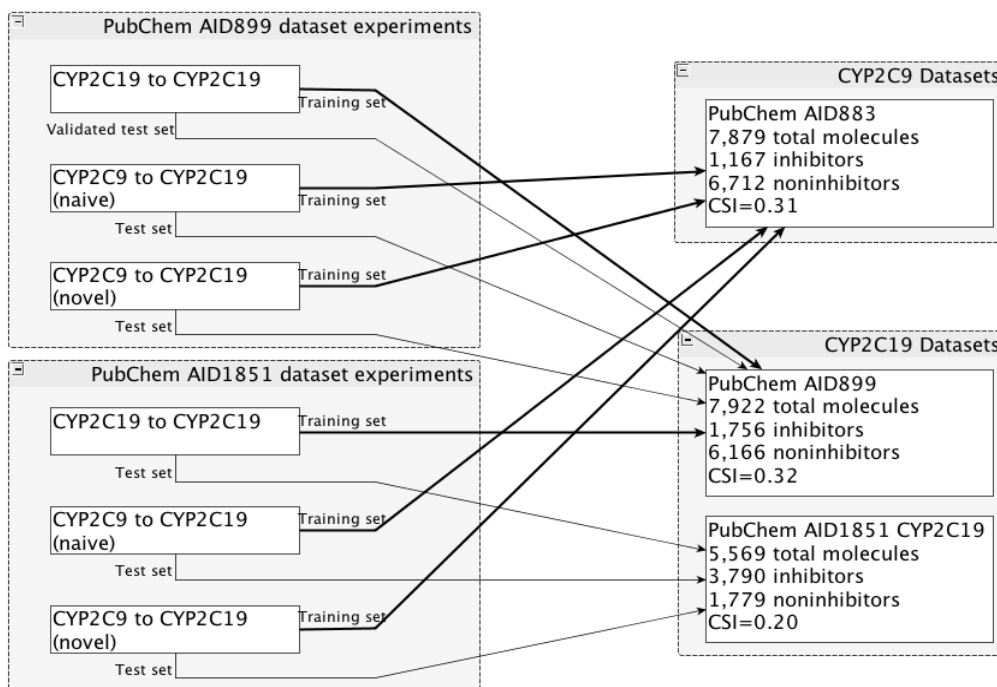


Figure 4.18. Relationships between datasets and modeling experiments described in this section

4.4.2 Modeling results

Table 4.10 and Table 4.11 summarize the model application results.

Descriptors	Method	BACC		
		CYP2C9 to CYP2C19 (novel)	CYP2C9 to CYP2C19 (naive)	CYP2C19 to CYP2C19
Dragon+AP	J48	0,811	0,781	0,827
Estate+AP	J48	0,799	0,779	0,803
AP	J48	0,757	0,751	0,752
Dragon	J48	0,79		0,807
Estate	J48	0,788		0,799
Dragon+AP	Ann	0,793	0,762	0,816
Estate+AP	Ann	0,759	0,733	0,783
AP	Ann	0,692	0,681	0,704
Dragon	Ann	0,779		0,809
Estate	Ann	0,752		0,781

Table 4.10. Results of three different approaches to CYP2C19 inhibition modeling, PubChem AID899 dataset. ANN – Associative Neural Networks [147,148], J48 – C4.5 pruned trees [153] as implemented in WEKA [154]. Dragon - 3D descriptors by software by Talete inc. [124], Estate - electrotopological state indices [127], AP - docking-derived protein-ligand atom pair descriptors (section 2.4.2, page 22).

Descriptors	Method	BACC		
		CYP2C9 to CYP2C19 (novel)	CYP2C9 to CYP2C19 (naive)	CYP2C19 to CYP2C19
Dragon+AP	J48	0,751	0,728	0,748
Estate+AP	J48	0,731	0,714	0,729
AP	J48	0,695	0,652	0,702
Dragon	J48	0,728		0,722
Estate	J48	0,68		0,711
Dragon+AP	Ann	0,732	0,722	0,746
Estate+AP	Ann	0,71	0,699	0,718
AP	Ann	0,695	0,659	0,677
Dragon	Ann	0,719		0,734
Estate	Ann	0,696		0,701

Table 4.11. Results of three different approaches to CYP2C19 inhibition modeling, PubChem AID1851 CYP2C19 dataset. ANN – Associative Neural Networks [147,148], J48 – C4.5 pruned trees [153] as implemented in WEKA [154]. Dragon - 3D descriptors by software by Talete inc. [124], Estate - electrotopological state indices [127], AP - docking-derived protein-ligand atom pair descriptors (section 2.4.2, page 22).

The “CYP2C9 to CYP2C19 (novel)” column shows the results of applying CYP2C9 models to predict CYP2C19 activity through extrapolation as described in the methods section. Evaluation of quality and applicability of this method is the main goal of this study. The “CYP2C9 to CYP2C19 (naive)” displays the model accuracies when CYP2C9 models are normally applied to data to get CYP2C9 activity predictions and then checking if these predictions are accurate for CYP2C19 cytochrome as well. This approach is well known and requires neither additional

experimental data nor additional computational resources to use. Therefore, it will serve as a baseline for comparison. The “CYP2C19 to CYP2C19” column represents the results of bagging-validated CYP2C19 modeling from the previous section. It shows the situation when the experimentally measured data for the mutated cytochrome is available and, therefore, no special techniques are required to model this protein's activity.

Note, that for models built only on Dragon or Estate descriptors the results for “naive” and “novel” columns are same, since they represent exactly the same models; the difference between the two approaches stems from the difference in methodology of calculation of docking descriptors. For the models where docking descriptors are not used the models and prediction results are identical.

The results presented in Table 4.10 and Table 4.11 confirm the general conclusions about QSAR modeling of CYP inhibition from the previous chapter.

Decision tree methods in general have displayed slightly higher performance for this classification task than neural networks: average performance of J48 models is around 3% higher than ANN models. The best performing J48 model is on average 1.5 - 2% more accurate than the best performing ANN model. The difference is statistically significant (with significance value of 0.05) according to the bootstrap test performed on 10000 replicas.

Three dimensional descriptors (represented in these experiments by Dragon) have been proven to be essential for CYP inhibition modeling. In the majority of the performed experiments models containing Dragon descriptors are significantly better than those not containing them. Adding atom pair descriptors significantly increased the performance of both Estate and Dragon models. For most of the experiments therefore Dragon+AP descriptor set yielded the models with highest prediction accuracy.

The traditional QSAR models built on CYP2C19 training data (“CYP2C19 to CYP2C19” models) have shown the best performance among the three studied approaches, in most cases significantly outperforming the other approaches in the study. Since this is the only approach among three that used CYP2C19 experimental data on the stage of model training, the higher accuracy results are explainable and once more confirm the importance of relevant experimental data to produce predictable QSAR models. This approach allowed the 0.827 accuracy for the bagging-validated training set. Applying the models to the structurally diverse external validation set resulted into a significant drop of model performance (the balanced accuracy of 0.748). This confirms the importance of good structural diversity of the training set and importance of applicability domain analysis for QSAR models to avoid applying them to data that is too different from the model training set.

The naive approach showed the least accurate results among the three studied approaches for all the models built in the study. This shows that for proteins from one subfamily that share up to 90% of activity values just assuming same activity results does not yield models with acceptable prediction accuracy. The best

performing models for this approach reached the *0.781* balanced accuracy value for the PubChem AID899 dataset and *0.728* balanced accuracy for the PubChem AID1851 CYP2C19 dataset. The speciality of this set of models is that additional four out of ten models have achieved accuracy values statistically similar to the top performing model.

In most performed modeling experiments the use of novel approach allowed to significantly increase the balanced accuracy (for up to 3%) as compared to the naive approach. For Dragon+AP and Estate+AP for both decision tree and neural networks machine learning methods the difference in prediction quality of this extrapolation approach from the “CYP2C9 to CYP2C19” approach was not statistically significant with significance value of 0.05 (as measured by bootstrap test with 10000 bootstrap replicas). The Dragon+AP and Estate+AP models built using decision tree approach displayed best results in predicting CYP2C19 activity based on CYP2C9 data. For the PubChem AID899 dataset the top performing model was the Dragon+AP J48 model with the balanced accuracy of *0.812*, which was not significantly lower than the performance of the same model built on CYP2C19 data (balanced accuracy of *0.827*). The models built using this approach experience the drop in prediction accuracy common for all other QSAR models when presented with structurally diverse external validation set data. For the PubChem AID1851 CYP2C19 dataset the balanced accuracies of predictions by novel model and CYP2C19 model are *0.751* and *0.748*, respectively. The difference is statistically insignificant, which means that the extrapolation approach in this experiment managed to achieve the accuracy comparable to the model built on relevant experimental data.

We can conclude that applying QSAR models to predict even closely related targets in most cases leads to insufficiently accurate results (as we can see from the “CYP2C9 to CYP2C19 naive” experiments). Introducing new information to the modeling process, however, leads to significant increase in modeling accuracy. The best kind of information is the experimental data for the specific target (the highest results are achieved by “CYP2C19 to CYP2C19” traditional QSAR approach”). In some cases similar results can be achieved by introducing only some additional information about the target (in form of atom pair descriptors calculated on the crystal structure of the target) and using modified methodology of model building and application.

4.4.3 Applicability domain analysis

Applicability domain analysis methods can be applied to the models in this study in a similar way to traditional QSAR. Since all models were built and applied through bootstrap aggregation, BAGGING-STD DM can be calculated for each predicted molecule.

Validated training set analysis

Figure 4.19 displays cumulative and local balanced accuracies of model predictions ordered by BAGGING-STD DM for all three methods of CYP2C19 activity prediction for PubChem AID899 dataset.

4.4 Novel descriptors in predicting CYP2C19 activity based on CYP2C9 dataset

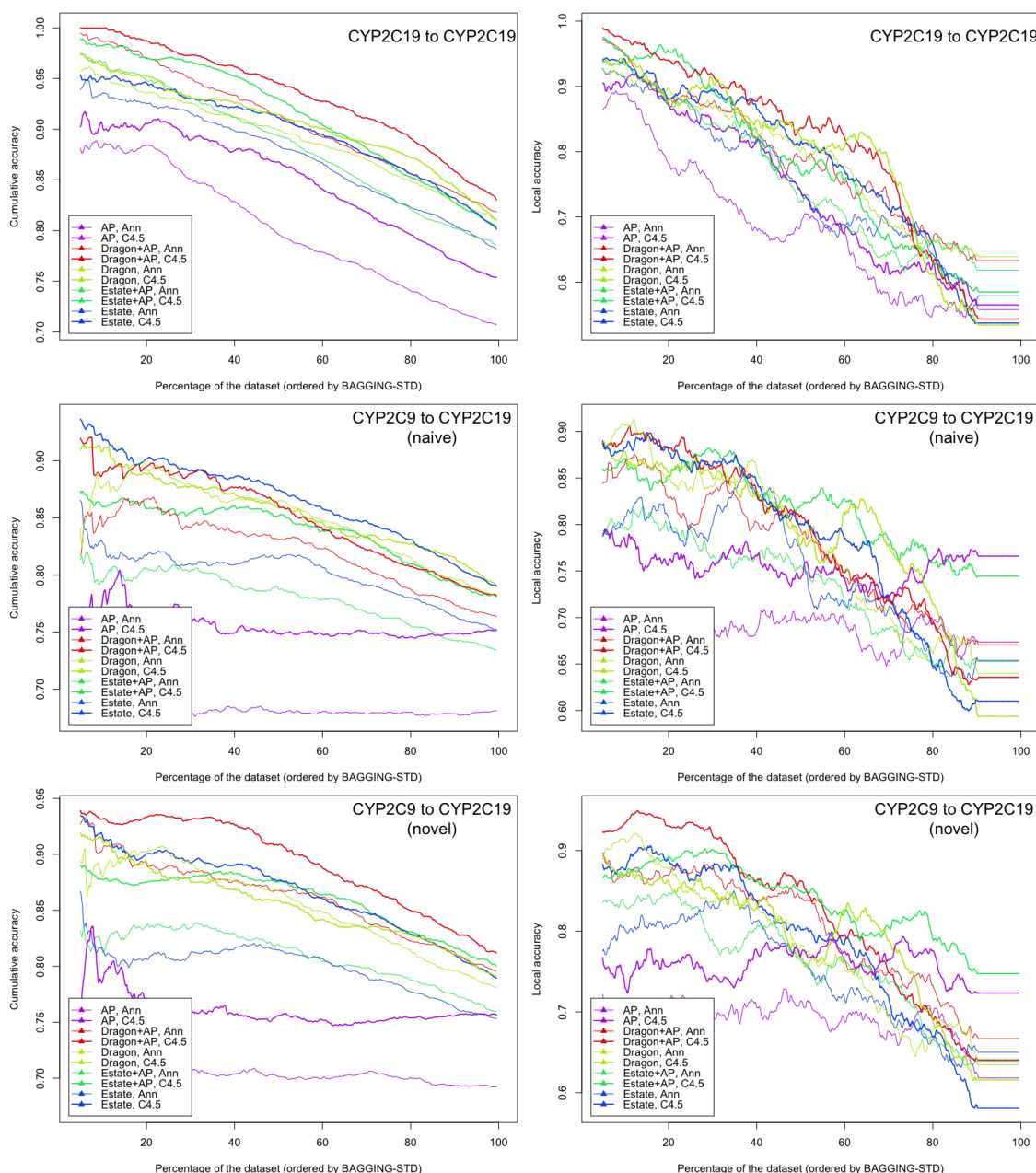


Figure 4.19. Cumulative and local balanced accuracies (ordered by BAGGING-STD DM) for PubChem AID899 dataset.

As we can see on Figure 4.19, applicability domain charts for extrapolation approaches (both novel and naive) exhibit the same behavior as for the traditional QSAR. On the cumulative charts the “CYP2C19 to CYP2C19” models demonstrate the balanced accuracy of up to 100% on a fraction of 10% of most confident predictions. The cumulative accuracy curves then gradually fall to the average model accuracy values (82.7% for the best performing Dragon+AP J48 model). The local accuracies for individual molecules are as high as 100% for the most confidently predicted molecules and as low as 60% for the least confidently predicted ones. This shows that the chosen applicability domain approach works well for the studied models and allows separation of confidently and unconfidently predicted molecules. The approach works best for Dragon+AP J48 model and worst - for the AP-only models.

The “CYP2C9 to CYP2C19 naive” cumulative accuracy charts fall from the maximum of 92% balanced accuracy for a subset of 5% most confident predictions to the average accuracy of 79% on the whole set. Local accuracy charts show that the highest individual balanced accuracy is around 92% for the most confident predictions and is around 65% for the least confident ones.

We can see that for both ANN Dragon descriptor containing models (Dragon ANN and Dragon+AP ANN) the charts do not follow the usual pattern and start at around 82% for the top 5% most confident predictions and rise, reaching a peak on the mark of around 20% most confident predictions, and then fall similarly to the rest of the models. This displays that for these models high confidence of predictions does not correspond to the real prediction accuracy. This may be caused by a subset of CYP2C9-specific descriptors having a high weight in the resulting model. That is, a similarity measure defined in the CYP2C9 inhibition property space fails for the CYP2C19 property space.

The chosen AD approach fails as well for the AP-only models in this modeling scenario. The local accuracy charts for these models show no correlation between the BAGGING-STD DM and the actual prediction accuracy for the particular molecule. As a consequence, the cumulative accuracy charts for these models display the average accuracy for any subset of most confidently predicted molecules. This highlights a drawback of AP descriptors and demonstrates, that AP information alone is insufficient to define similarity in extrapolated CYP2C19 property space.

The “CYP2C9 to CYP2C19 novel” models do not have some of the problems described for the “CYP2C9 to CYP2C19 naive” models. The best performing model is Dragon+AP J48 and it achieves the accuracy of around 93% on the top 40% of the dataset. The local accuracy chart displays good correlation between the BAGGING-STD DM and the real prediction accuracy for individual molecules. The most confidently predicted molecules have the local accuracy of 93%, the least confidently predicted molecules - 65%.

The AD analysis fails for the AP-only models in this modeling scenario as well. This highlights the necessity of traditional molecule-centered descriptors in models to reliably define the “distance to model” in property space.

Two different approaches to assessing the success of applying the applicability domain analysis to a set of models include: determining the average accuracy on a subset of fixed size that would contain only most confident predictions; and determining a size of the subset of most confident predictions on which a fixed predetermined accuracy could be expected.

From the cumulative balanced accuracy graph we can fill Table 4.12 with balanced accuracy results when only 20% most confident predictions are taken into account (as identified by BAGGING-STD DM).

Descriptors Method	BACC for top 20% most confident predictions		
	CYP2C9 to CYP2C19 (novel)	CYP2C9 to CYP2C19 (naive)	CYP2C19 to CYP2C19
Dragon+AP J48	0.933	0.895	0.987
Estate+AP J48	0.876	0.859	0.97
AP J48	0.755	0.752	0.901
Dragon J48	0.892		0.952
Estate J48	0.902		0.944
Dragon+AP Ann	0.893	0.858	0.971
Estate+AP Ann	0.839	0.801	0.949
AP Ann	0.693	0.702	0.884
Dragon Ann	0.895		0.937
Estate Ann	0.802		0.925

Table 4.12. Results of three different approaches to CYP2C19 inhibition modeling, PubChem AID899 dataset; only top 20% most confident predictions are considered. ANN – Associative Neural Networks [147,148], J48 – C4.5 pruned trees [153] as implemented in WEKA [154]. Dragon - 3D descriptors by software by Talete inc. [124], Estate - electrotopological state indices [127], AP - docking-derived protein-ligand atom pair descriptors (section 2.4.2, page 22).

As we can see, the most successful models displayed the balanced accuracies of 98%, 89% and 93% for the “CYP2C19 to CYP2C19”, “CYP2C9 to CYP2C19 naive” and “CYP2C9 to CYP2C19 novel” approaches, respectively. The extrapolation approaches are significantly worse than the approach based on relevant experimental data. However the novel approach is good enough to be practically use in virtual screening studies.

Table 4.13 reflects a second view on the applicability domain accuracy charts.

Descriptors Method	Fraction of the set predicted by the best model with given BACC					
	CYP2C9 to CYP2C19 (novel)		CYP2C9 to CYP2C19 (naive)		CYP2C19 to CYP2C19	
	BACC =0.90	BACC =0.85	BACC =0.90	BACC =0.85	BACC =0.90	BACC =0.85
Dragon+AP J48	54%	80%	9%	55%	77%	93%
Estate+AP J48	3%	71%	-	50%	62%	81%
AP J48	-	-	-	-	27%	57%
Dragon J48	12%	56%	12%	56%	60%	83%
Estate J48	26%	64%	26%	64%	56%	88%
Dragon+AP Ann	15%	67%	2%	27%	55%	83%
Estate+AP Ann	1%	7%	-	1%	46%	69%
AP Ann	-	-	-	-	4%	31%
Dragon Ann	22%	59%	22%	59%	49%	80%
Estate Ann	3%	6%	3%	6%	39%	66%

Table 4.13. Fractions of the PubChem AID899 dataset predicted by a specific approach with given balanced accuracy

This table can be interpreted as a summary of success of using a BAGGING-STD DM applicability domain approach with each of the models in the study. We fix a particular balanced accuracy requirements and evaluate whether a model is capable of fulfilling it.

The experimental data based approach is the most successful, with all the models allowing to some extent the 90% balanced accuracy (the most unsuccessful model is the AP Ann model with only 4% of the dataset, and the most successful - the Dragon+AP J48 model with 77% of the dataset).

The best approach for the “CYP2C9 to CYP2C19 naive” from the point of view of this table was Estate J48 model. It achieved the 90% accuracy on 26% of most confidently predicted compounds.

The most successful model of the “CYP2C9 to CYP2C19 novel” approach was Dragon+AP J48 as well. The balanced accuracy of 90% could be achieved on around 54% of most confident predictions.

External test set analysis

The analysis can be also performed for the external validation set - PubChem AID1851 CYP2C19 dataset. Figure 4.20 displays the local and cumulative balanced accuracies obtained by applying the studied models to the external validation set.

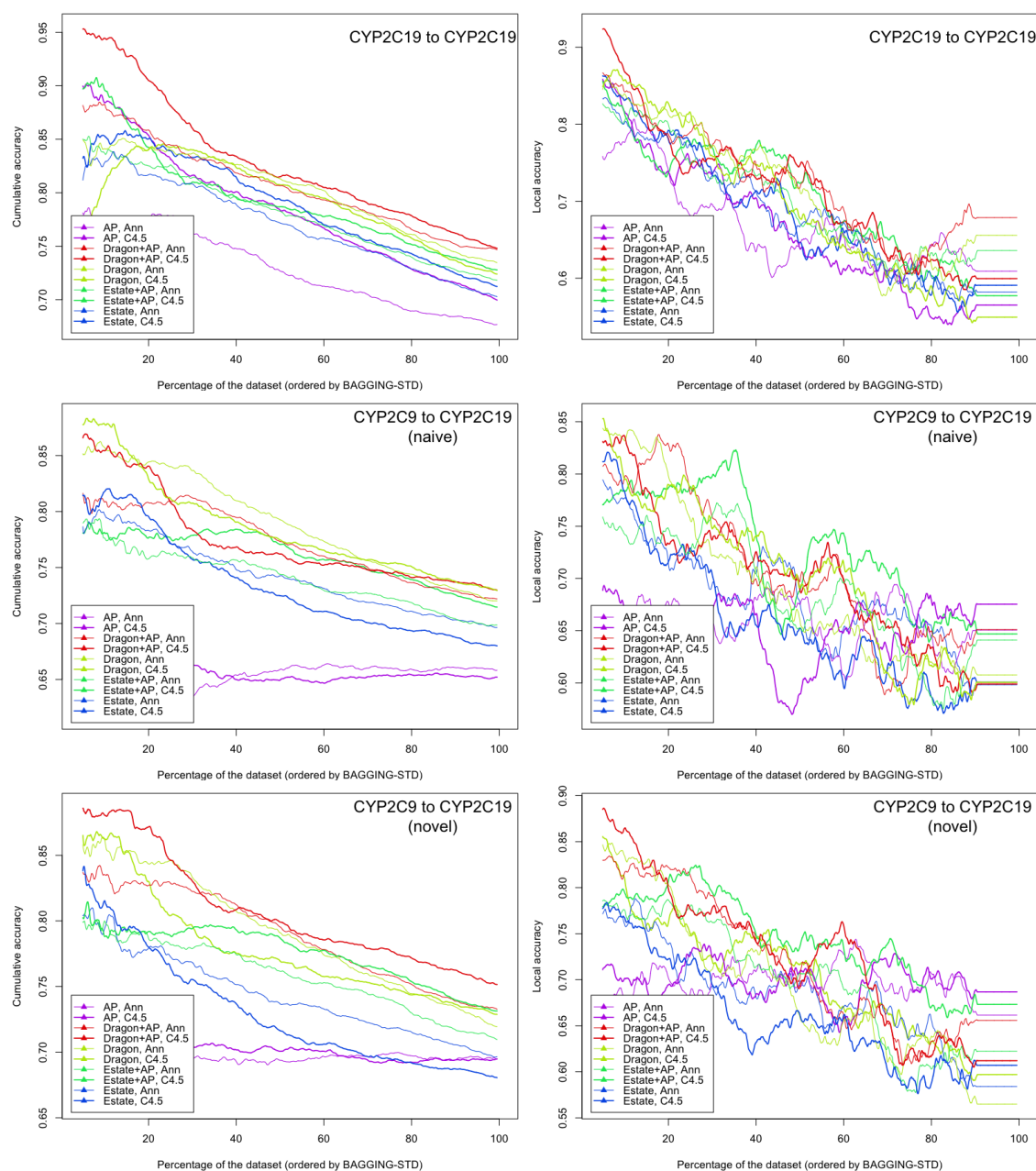


Figure 4.20. Cumulative and local balanced accuracies (ordered by BAGGING-STD DM) for PubChem AID1851 CYP2C19 dataset.

The external validation set results are similar to those of the PubChem AID899 dataset analysis.

The best balanced accuracy results are achieved by “CYP2C19 to CYP2C19” models. The highest accuracy model was Dragon+AP J48 and achieved over 90% balanced accuracy on the top 20% most confident predictions. The local accuracies for this dataset range from around 95% for the most confidently predicted compound to 60% for the least confidently predicted compounds. The AP J48 model displayed reasonable correlation between BAGGING-STD DM and prediction accuracy. The most confidently predicted molecules have the local accuracy of around 90%, and the least confidently predicted molecules - around 58%.

The most accurate model in the “CYP2C9 to CYP2C19 novel” scenario demonstrates the 90% local accuracy for the most confidently predicted molecules. As a result, around 20% of the most confidently predicted molecules have the balanced accuracy of 87%.

Table 4.14 contains balanced accuracy results for top 20% most confident predictions from the PubChem AID1851 CYP2C19 dataset.

Descriptors	Method	BACC for top 20% most confident predictions		
		CYP2C9 to CYP2C19 (novel)	CYP2C9 to CYP2C19 (naive)	CYP2C19 to CYP2C19
Dragon+AP	J48	0,87	0,84	0,905
Estate+AP	J48	0,79	0,776	0,842
AP	J48	0,711	0,674	0,852
Dragon	J48	0,825		0,841
Estate	J48	0,785		0,848
Dragon+AP	Ann	0,828	0,805	0,855
Estate+AP	Ann	0,784	0,76	0,826
AP	Ann	0,672	0,679	0,774
Dragon	Ann	0,844		0,84
Estate	Ann	0,776		0,817

Table 4.14. Results of three different approaches to CYP2C19 inhibition modeling, PubChem AID1851 CYP2C19 dataset; only top 20% most confident predictions are considered. ANN – Associative Neural Networks [147,148], J48 – C4.5 pruned trees [153] as implemented in WEKA [154]. Dragon - 3D descriptors by software by Talete inc. [124], Estate - electrotopological state indices [127], AP - docking-derived protein-ligand atom pair descriptors (section 2.4.2, page 22).

Since none of the models in the extrapolation scenarios achieved the 90% balanced accuracy on any fraction of the dataset, Table 4.15 only presents results for BACC=0.85 threshold.

As we can see, under given accuracy threshold the use of novel extrapolation approach allows to extend the fraction of compounds predicted with this accuracy from 14% to 24% of the chemically diverse external validation set.

Descriptors Method	Fraction of the set predicted by the best model with given BACC		
	CYP2C19 (novel)	CYP2C19 (naive)	CYP2C19 to CYP2C19
	BACC =0.85	BACC =0.85	BACC =0.85
Dragon+AP J48	24%	14%	32%
Estate+AP J48	1%	-	19%
AP J48	-	-	20%
Dragon J48	16%	16%	22%
Estate J48	1%	1%	21%
Dragon+AP Ann	-	-	21%
Estate+AP Ann	-	2%	9%
AP Ann	-	-	-
Dragon Ann	17%	17%	21%
Estate Ann	-	-	-

Table 4.15. Fractions of the PubChem AID1851 CYP2C19 dataset predicted by a specific approach with given balanced accuracy.

Since the most useful task of applicability domain analysis is prediction of model accuracy for a specific compound, we performed a detailed estimated prediction accuracy vs. real prediction accuracy for a top performing model for each of the three scenarios in the study. Dragon+AP, J48 model is the top performing model both for PubChem AID899 and PubChem AID1851 CYP2C19 datasets, and the behavior of applicability domain plots is qualitatively and quantitatively similar for both sets. Therefore to avoid redundancy we will perform the analysis for the “worst case” - PubChem AID1851 CYP2C19 experiments only.

Figure 4.21 displays the training and test set DM distribution bar plots, training and test set local accuracy plots and real and estimated applicability domain cumulative balanced accuracy plots.

DM distribution plots display the fractions of the training and test sets that have DM values within specific ranges. We can build this plots for any external validation sets with no prior knowledge about the actual activity values for the predicted molecules, since DM values are produced by the model for each individual molecule.

We can see that the distribution plots for the three scenarios are similar. The squared mean values of BAGGING-STD values are 0.391, 0.392 and 0.392 for training sets for the three scenarios, and 0.449, 0.442 and 0.437 for the test sets, respectively. This means that the cause of the difference between real and estimated accuracies for the studied models stem from different relationships between DM values and prediction accuracies for training and test sets.

4.4 Novel descriptors in predicting CYP2C19 activity based on CYP2C9 dataset

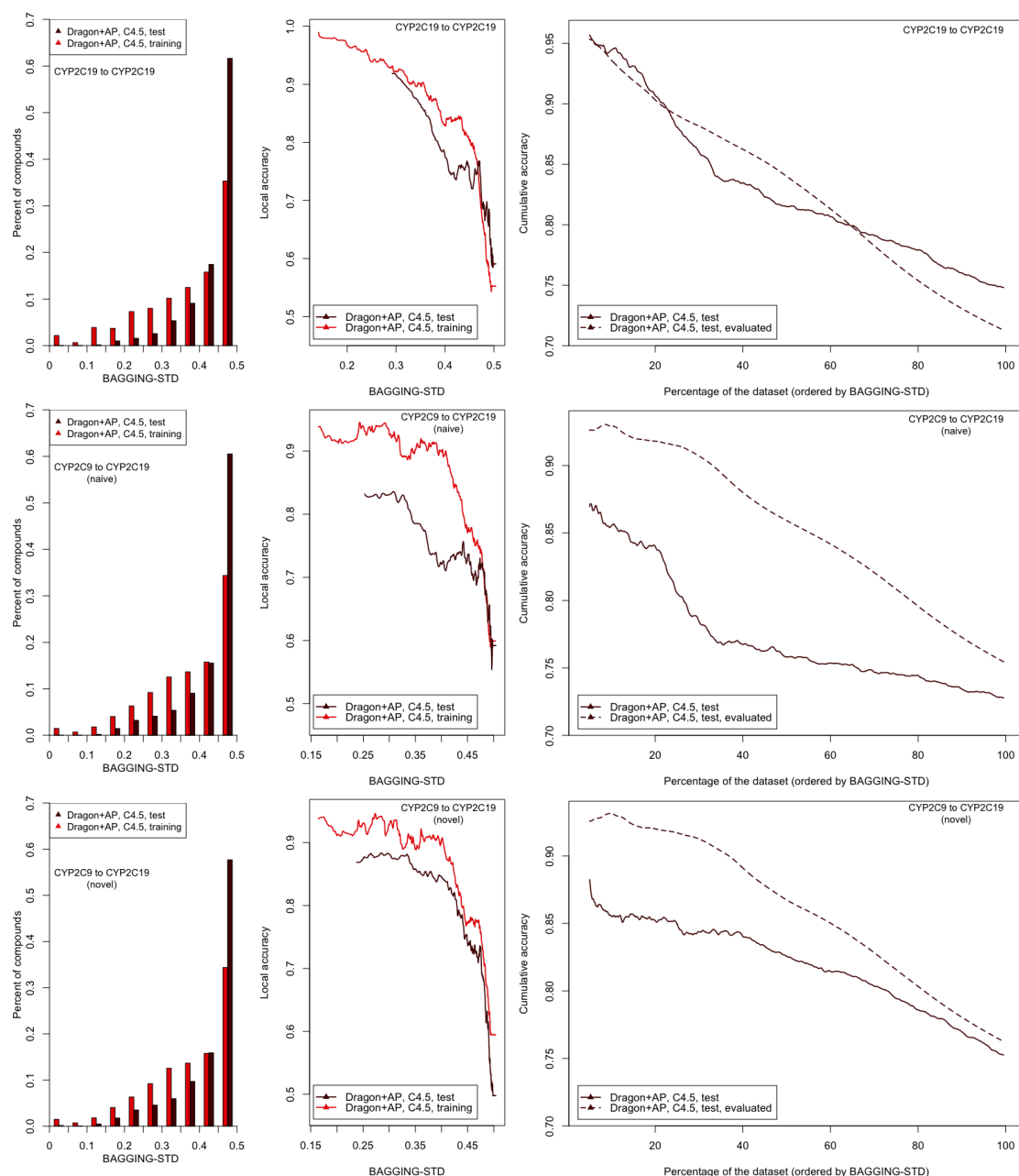


Figure 4.21. DM value distribution diagrams, local accuracy plots, and actual and estimated cumulative applicability domain accuracy plots for three studied methods. Displayed model is C4.5 decision tree model built on Dragon + Docking descriptors, PubChem AID1851 CYP2C19 dataset.

Local accuracy plots display the relationship between BAGGING-STD DM and the accuracy of prediction of every particular compound (as determined by averaging prediction accuracy by sliding window over DM-sorted compounds). We can not build this plot if we don't know actual activity values for the predicted set. Therefore, when estimating accuracy of prediction of each particular molecule we assume that the relationship between DM values and prediction accuracy is same for training and test sets. We then proceed to evaluate the accuracy of the test set based on the DM-accuracy dependency for the training set. Figure 4 shows that this assumption is not totally correct. The squared mean of differences between training and test accuracies for the same BAGGING-STD for the

CYP2C19 to CYP2C19 scenario is 0.045, for CYP2C9 to CYP2C19 naive is 0.107 and for CYP2C9 to CYP2C19 novel is 0.071. The lower value for the novel approach (as compared to naive approach) identifies the higher quality of expected accuracy estimations.

The “real” cumulative accuracy plot is exactly the same plot for Dragon+AP J48 model from the previous chapter. It can only be built if the test set experimental values are known. The “estimated” cumulative accuracy plot is the same cumulative accuracy plot, but built for the test set of the model based on the hypothesis that the relationship between DM and compound prediction accuracy is quantitatively same for training and test sets of the model. The estimated plot can be built for any test set and does not require experimental values. The difference between the real and estimated plot reflects the quality of prediction accuracy estimation using the chosen AD approach.

Therefore using the novel extrapolation technique we increase the accuracy of model prediction (as demonstrated in previous chapter), and also achieve higher quality accuracy estimations by applicability domain techniques.

4.4.4 Fragment-based interpretation

In this section the applicability domain measures are studied on fragment-based subsets of the PubChem AID1851 CYP2C19 dataset. Fragments are generated in a way similar to the previous studies. Only fragments that are part of at least 100 molecules in a set were considered. If two fragment-containing molecule groups contain the same amount of molecules and one of the fragments is the exact subfragment of the other, the bigger fragment was selected. The BAGGING-STD DM measures were calculated based on the predictions for a group of molecules containing a fragment only. Within each fragment-containing subset of molecules balanced accuracy was calculated.

The goal of this section is fragment-based explanation of higher accuracy values of novel approach as compared to naive approach.

Among all generated fragment-containing subsets four subsets demonstrated an increase in prediction accuracy that could not be explained by general increase in model prediction accuracy: molecules containing *acetophenone*, *3-nitrotoluene*, *N-phenylthiourea* and *cyclohexane* fragments have higher increase in prediction accuracy than the whole dataset with significance value of 0.05.

Figure 4.22 shows statistics for the determined fragments and DM values distribution plot for the C4.5 decision tree model built on Dragon + Docking descriptors, PubChem AID1851 CYP2C19 dataset, naive and novel approaches.

Note that the size of molecule subsets slightly differ due to the fact that not all molecules could be successfully docked to both CYP2C9 and CYP2C19 isoforms. Balanced distance to model value distribution plots built for these fragment-containing subsets show a shift from “mostly unconfident predictions” to “mostly confident predictions” for the naive vs. novel approach.

4.4 Novel descriptors in predicting CYP2C19 activity based on CYP2C9 dataset

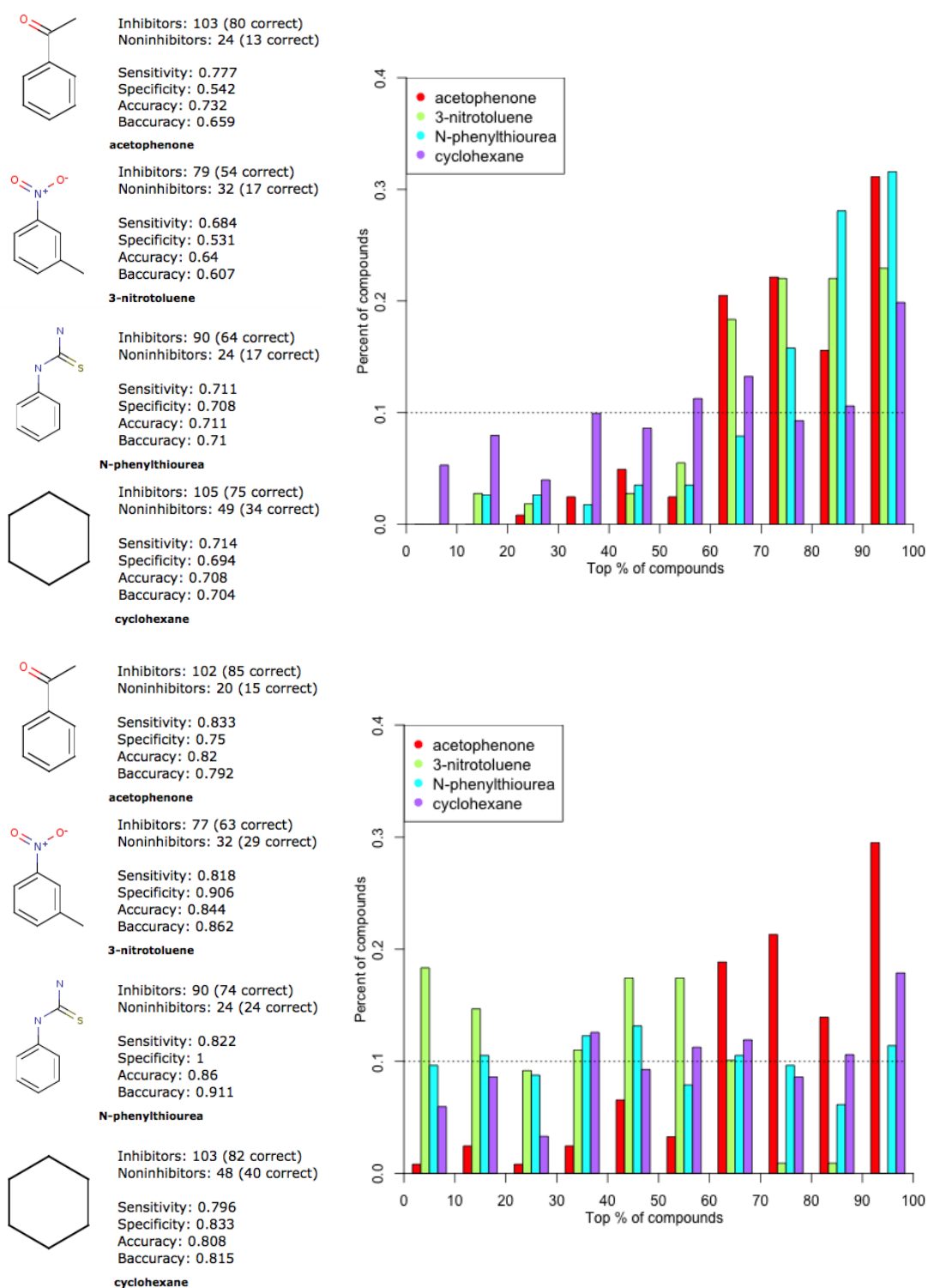


Figure 4.22. Fragments statistics and DM value distribution plot for C4.5 decision tree model built on Dragon + Docking descriptors, naive (top) and novel (bottom) approach.

Table 4.16 demonstrates a comparison of real and estimated balanced accuracies for the studied models.

	CYP2C9 to CYP2C19 (naive)		CYP2C9 to CYP2C19 (novel)		CYP2C9/CYP2C19 correlation
	BACC (Estimated)	BACC	BACC (Estimated)	BACC	
acetophenone	0,708	0,659	0,749	0,792	0,702
3-nitrotoluene	0,712	0,607	0,881	0,862	0,672
N-phenylthiourea	0,694	0,71	0,892	0,911	0,595
cyclohexane	0,785	0,704	0,794	0,815	0,642
FULL SET	0,77	0,728	0,77	0,751	0,793

Table 4.16. Comparison of model balanced accuracies (real and estimated) for determined fragments, naive and novel approach.

From Table 4.16 we can see that for the determined fragments the percentage of molecules with the same activity values for CYP2C9 and CYP2C19 is significantly lower than for the whole external test set (~0.6-0.7 values for the fragment-containing subsets and ~0.8 for the whole dataset). This can be explained by the possibility that for these molecules the CYP2C9 activity and CYP2C19 activity is significantly different (consequence of substrate selectivity of the studied CYP isoforms). As a result we can see that in the case of naive approach the balanced accuracies are comparatively low (compared with the average model accuracy) and the estimated accuracy values obtained via applicability domain approach are over optimistic.

The novel approach model, however, takes into account the target protein structure and compensates for some of isoform selectivity differences. We can see that the novel approach estimated accuracies are close to the real model accuracies. The accuracy values of the fragment-containing subsets are significantly higher than of the full external test set (~0.79-0.91 for the subsets, 0.751 for the whole set).

To explain the increase in model accuracy we analyze the docking conformations of the selected subset of molecules. Several sources [209,210,212] indicate the important role of ILE99 (CYP2C9) / HIS99 (CYP2C19) amino-acids in CYP substrate selectivity for a wide range of compounds.

The sample docking conformations of cyclohexane- and 3-nitrotoluene- containing compounds for CYP2C9 and CYP2C19 proteins are displayed on Figure 4.23 and figure Figure 4.24, respectively. The heme is displayed in green color. The protein is displayed in secondary structure representation. The ILE99/HIS99 amino-acids are explicitly displayed to the left of the sample compounds.

For the selected fragment subsets the fraction of molecules docked within 4Å of the ILE99/HIS99 amino-acid is 87%, which is significantly higher than for the whole set (only 72%).

Therefore the naive vs. novel approach fragment-based comparison has provided a testable hypothesis of importance of ILE99/HIS99 amino-acid for the CYP2C9/CYP2C19 substrate selectivity for the determined subset of molecules.

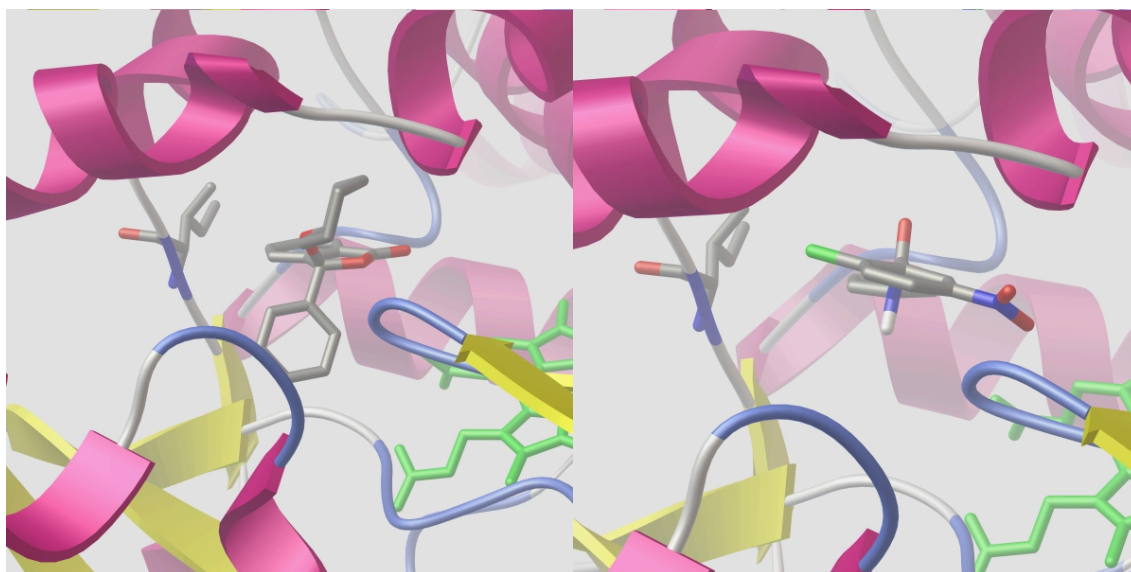


Figure 4.23. Docking conformations of sample cyclohexane (left) and 3-nitrotoluene (right) containing compounds in the CYP2C9 pocket near the ILE99 amino acid.

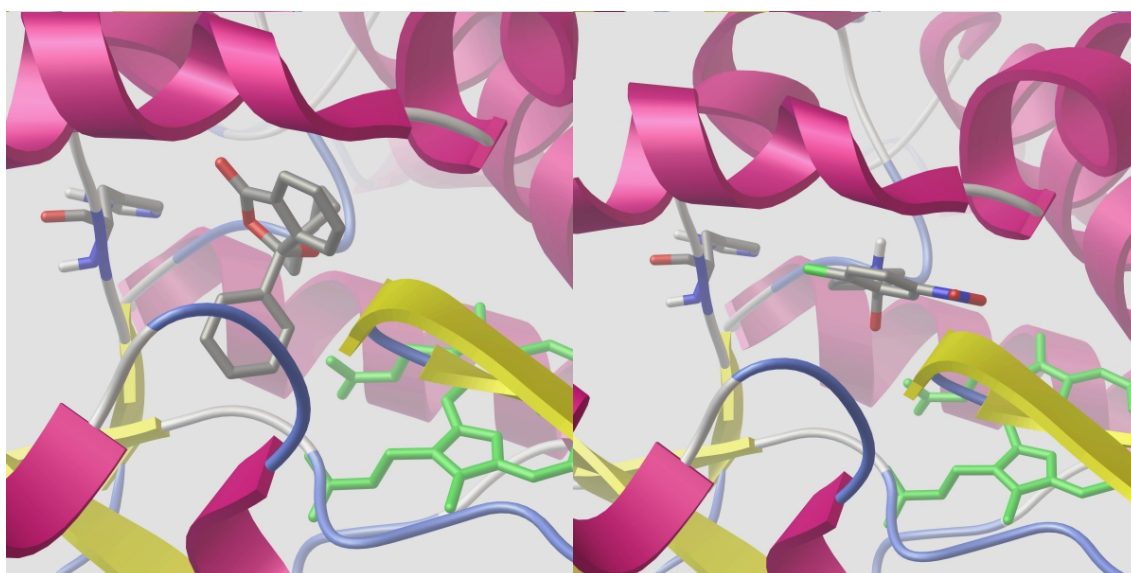


Figure 4.24. Docking conformations of sample cyclohexane (left) and 3-nitrotoluene (right) containing compounds in the CYP2C19 pocket near the HIS99 amino acid.

4.4.5 Summary

In this study we predicted activities of small molecules against a protein based on experimental data measured for another protein. The motivation was to determine whether the atom pair descriptors allow extrapolation of QSAR modeling results to clinically significant mutations of cytochromes. Due to lack of consistent dataset for mutated cytochrome activity, the CYP2C19 activity was chosen as the target property and CYP2C9 activity as the measured property.

The QSAR models were built using the same descriptors (Dragon and Estate indices, protein-ligand atom pair descriptors) and machine learning methods (C4.5 decision trees and ASNN neural networks) as in the previous study.

Three kinds of experiments were performed. The CYP2C9 to CYP2C19 experiments used CYP2C9 experimental data to predict the CYP2C19 activity and was used as the upper threshold for expected experiment accuracy. The CYP2C9 to CYP2C19 naive experiments just assumed that CYP2C9 and CYP2C19 activities are exactly same, and the prediction accuracies were calculated from predictions based on this hypothesis. The CYP2C9 to CYP2C19 novel experiments used atom pair descriptors and modified methodology to predict CYP2C19 activity based on CYP2C9 experimental data.

The novel approach allowed to increase the accuracy of predictions by up to 4% (as compared to the naive approach) and achieve the balanced accuracy of 81% for AID899 dataset and 79% for the AID1851 dataset. For the best performing model for the AID1851 dataset the prediction accuracy was statistically similar to the CYP2C9 to CYP2C19 approach.

The applicability domain approach used in combination with the CYP2C9 to CYP2C19 novel approach achieved the balanced accuracy of 87% on the top 20% most confident predictions for the AID1851 dataset. It was also determined that the use of novel approach increased the quality of accuracy estimation: the squared error of real vs. estimated accuracy was 7% (as compared to 11% for the naive approach).

The fragment-based analysis of model accuracy increase for “naive” vs. “novel” approach has determined four specific fragments. Prediction accuracy for molecules containing these fragments has increased significantly more (as compared to average accuracy increase). The *acetophenone*, *3-nitrotoluene*, *N-phenylthiourea* and *cyclohexane* fragments demonstrate a significantly higher prediction accuracy and applicability-domain based accuracy estimation compared to the whole set. The DM value distribution plots also display that the molecules containing these fragments are distributed in the mostly confident predictions area for the “novel” approach. This is a direct result of the approach, which included protein-specific information into the model and therefore captured substrate selectivity behavior for these classes of compounds.

This signals that the fragments hold structural features that can be especially accurately predicted by using atom pair descriptors in combination with the correct target protein structure. The hypothesis is that the molecules containing these fragments are mostly located near the ILE99 (CYP2C9) / HIS99 (CYP2C19) amino acid of the binding site. It has been shown that this amino acid substitution plays an important role in CYP2C9/CYP2C19 substrate selectivity.

5 Conclusions and outlook

In this work the challenges and possibilities of QSAR approaches to prediction of human cytochrome P450 inhibition were investigated. The CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 isoforms were selected for detailed analysis based on their reported importance in drug metabolism and potential drug-drug interaction studies.

Datasets of sufficient chemical diversity coming from high-throughput screening experiments were chosen for analysis in this study. The training sets (the ones presented to the machine learning methods) and the test sets (used to evaluate the quality of the obtained models) were chosen in a way to prevent any chance of falsely optimistic results and to present a realistic estimation of performance of the methodology introduced in this study. The experimental accuracy of the datasets was estimated to be 80-91%.

Preliminary analysis of the datasets confirmed a high degree of similarity between CYP2C9 and CYP2C19 isoforms, as well as a relatively high degree of similarity between CYP3A4 and the other analyzed isoforms.

The molecule-fragment-based analysis methodology was presented and fragment-based analysis of the datasets was performed. It was found that some molecular fragments demonstrated a statistically significant correlation with cytochrome inhibition (*chlorobenzene*, *pyridine*, *1H-indole*) or non-inhibition (*acetic acid*, *sulfonic acid*, *9H-purine*) activity. Other fragments displayed statistically significant isoform selectivity (*trimethylamine* only inhibitor for CYP2D6, *pyrimidine* and *quinazoline* are inhibitors for all isoforms except CYP2C9, *8H-pteridin-7-one* - inhibitor for CYP1A2 and CYP3A4 and non-inhibitor for CYP2D6).

A comprehensive QSAR study was performed on the CYP1A2 datasets using well-established general molecular descriptors and machine learning methods. Several models showed the bagging-validated prediction accuracy of 82% - 83% of correctly classified instances of the initial training set. These results are similar to the performance of models published in several related papers. The external test accuracies for the models are 71% - 82% correctly classified instances. Using the distance to model based applicability domain approaches allowed us to increase the accuracies to 83% - 96% on about 10% of external set compounds. These prediction accuracy values are close to estimated experimental accuracy values for this set. This proves that QSAR models can be used to decrease the number of experimental measurements on a subset of studied compounds in early stage drug discovery scenarios.

In an effort to incorporate more problem-specific information into the QSAR models, the novel set of chemogenomics based descriptors was developed. The descriptors are calculated on a protein-ligand complex (instead of the ligand structure only). The protein-ligand complex for the descriptors can be obtained by molecular docking experiments.

The docking experiments on the studied datasets were performed and the database of protein-ligand complexes for five studied CYP isoforms was created. A comprehensive

QSAR study on all studied CYP isoform datasets was performed. The novel chemogenomics based descriptors were used in the study. It was shown that incorporation of additional information (via novel descriptors) into the model results into statistically significant increase in model performance. For all studied isoforms the top-performing model included novel descriptors, with an average increase of 2-3% of correctly classified instances. The resulting models had the balanced accuracy of 83% - 87% on the validated training sets and 74% - 83% on the test sets. Using the distance to model based applicability domain approaches allowed us to increase the accuracies to 87% - 96% on about 20% of external set compounds for the top performing model.

A fragment-based applicability domain analysis methodology was presented and the analysis determined groups of molecular fragments that are predicted with more than average confidence and less than average confidence. The molecules that were predicted most confidently contained linear and branched fragments with the size of 4 - 6 atoms, among which - carbon, nitrogen and oxygen. Molecules, which were predicted with the lowest confidence contained fragments that were mostly aromatic and on average contained more atoms and had a higher molecular weight. One particular fragment (trifluoromethylbenzene) was marked as least confidently predicted for three out of five isoforms (CYP2C9, CYP2C19 and CYP3A4). This indicates that the datasets contain inconclusive data for molecules containing this fragment and additional experimental measurements are required to reliably predict CYP inhibition activity for trifluoromethylbenzene containing molecules.

The clinically significant genetic polymorphism of the studied CYP isoforms makes it challenging to use models built for non-mutated proteins in early stage drug design. In this study the possibility to use the novel descriptors to extrapolate modeling results for one protein to a family of closely related proteins is explored.

The methodology was presented to build QSAR using novel descriptors in a way that would allow to extrapolate their predictions to a family of closely related proteins. As a proof of concept, a series of QSAR studies was performed where modeling results of CYP2C9 were extrapolated to predict CYP2C19 activity. The presented CYP2C9 to CYP2C19 extrapolation approach demonstrated the increase of up to 4% of correctly classified instances (as compared to the naive approach) and achieves the prediction accuracy that was statistically similar to QSAR modeling of CYP2C19 activity on experimental CYP2C19 data. The applicability domain approach used in combination with the CYP2C9 to CYP2C19 novel approach achieved the balanced accuracy of 87% on the top 20% most confident predictions for the external test dataset.

The fragment-based analysis of model accuracy increase for “naive” vs. “novel” approach has determined four specific fragments, for which the increase in prediction accuracy was significantly higher than for the dataset in general: *acetophenone*, *3-nitrotoluene*, *N-phenylthiourea* and *cyclohexane*. The hypothesis is that the molecules containing these fragments are mostly located near the ILE99 (CYP2C9) / HIS99 (CYP2C19) amino acid of the binding site. It has been shown that this amino acid substitution plays an important role in CYP2C9/CYP2C19 substrate selectivity. Since novel descriptor include protein-ligand interaction information, their presence in the model explain the specific increase of prediction accuracy for these fragments.

All the models built in this thesis are available online on the OCHEM platform. They can be used by researchers to validate the introduced methodology, benchmark their own models and screen molecular databases for cytochrome P450 activity.

The goal of QSAR models is not only to help screen molecular libraries and prioritize compounds for experimental measurements, but also to guide the drug design process. This should be the main focus of the future work that follows this study. To be able to generate useful conclusions from QSAR models the researcher should not only know that a particular molecule is active, but also *why* it's active and what can be done to make it inactive. Therefore, the model interpretation facilities and methodologies are of crucial importance in QSAR modeling in general and QSAR approaches to the prediction of human cytochrome P450 inhibition in particular.

The fragment-based analysis methodology can be extended to take into account fragment hierarchies to account for synergistic effects of different fragments present in one molecule. QSAR and 3D-QSAR studies on subsets of molecules only containing specific fragments can be performed to derive the mechanistic explanation of importance of each determined fragment. Combination of applicability domain and fragment analysis can be employed to determine classes of compounds with higher prediction errors.

Further development of the suggested descriptor set is required. The simple atom types used in this study can be extended to include information on nearest neighboring atoms and bond types, as well as information on functional group into which the atom is incorporated. This will make the descriptors more interpretable. Well established methods of linear modeling and PCA in descriptor space can then be performed to obtain information on particular protein-ligand interaction importance.

Modeling results extrapolation across closely related proteins is a prospective topic, which can be researched further by including further protein or protein-ligand specific information into the modeling process. The assumption that all descriptors bear the same relation to inhibition activity (both qualitatively and quantitatively) for all closely related proteins can not be true in a general case. Therefore further study is required to validate this assumption.

The author hopes that this work will contribute to the applicability of QSAR approaches in early stage drug discovery and drug-drug interaction studies, as well as toxicity and risk assessment fields. The models of the study were made available on OCHEM platform to promote standards of model reproducibility and collaboration in QSAR community working on the cytochrome P450 inhibition prediction.

References

1. Brown AC, Fraser TR. On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *J Anat Physiol.* 1868;2(2):224–42.
2. Overton E. *Studien uber die Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie.* G. Fischer, Jena; 1901.
3. Ferguson J. The Use of Chemical Potentials as Indices of Toxicity. *Proc. R. Soc. Lond. B.* 1939 Jul 4;127(848):387–404.
4. Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. , Published online: 14 April 1962; | doi:10.1038/194178b0. 1962 Apr 14;194(4824):178–80.
5. Esposito EX, Hopfinger AJ, Madura JD. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.* 2004;275:131–214.
6. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov.* 2004 print;3(8):711–6.
7. Kubinyi H. Drug research: myths, hype and reality. *Nature Reviews Drug Discovery.* 2003 Aug 1;2(8):665–8.
8. Kerns EH, Di L. Pharmaceutical profiling in drug discovery. *Drug Discovery Today.* 2003 Apr 1;8(7):316–23.
9. Khanna I. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discovery Today.* 2012 Oct;17(19–20):1088–102.
10. Dimasi JA. Risks in new drug development: approval success rates for investigational drugs. *Clin. Pharmacol. Ther.* 2001 May;69(5):297–307.
11. DiMasi JA, Feldman L, Seckler A, Wilson A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin. Pharmacol. Ther.* 2010 Mar;87(3):272–7.
12. Prueksaritanont T, Tang C. ADME of Biologics– What Have We Learned from Small Molecules? *AAPS J.* 2012 Sep 1;14(3):410–9.
13. Rendic S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* 2002;34(1-2):83–448.
14. Masimirembwa CM, Thompson R, Andersson TB. In vitro high throughput screening of compounds for favorable metabolic properties in drug discovery. *Comb. Chem. High T. Scr.* 2001 May;4(3):245–63.
15. Pelkonen O, Turpeinen M, Hakkola J, Honkakoski P, Hukkanen J, Raunio H. Inhibition and induction of human cytochrome P450 enzymes: current status. *Arch. Toxicol.* 2008 Oct;82(10):667–715.
16. Guengerich FP, Wu Z-L, Bartleson CJ. Function of human cytochrome P450s: Characterization of the orphans. *Biochem. Bioph. Res. Co.* 2005 Dec 9;338(1):465–9.
17. Williams JA, Hyland R, Jones BC, Smith DA, Hurst S, Goosen TC, Peterkin V, Koup JR, Ball SE. Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUC_i/AUC) ratios. *Drug Metab. Dispos.* 2004 Nov;32(11):1201–8.

18. Pirmohamed M, Park BK. Cytochrome P450 enzyme polymorphisms and adverse drug reactions. *Toxicology*. 2003 Oct 1;192(1):23–32.
19. Michalets EL. Update: clinically significant cytochrome P-450 drug interactions. *Pharmacotherapy*. 1998 Feb;18(1):84–112.
20. Oyama T, Kagawa N, Kunugita N, Kitagawa K, Ogawa M, Yamaguchi T, Suzuki R, Kinaga T, Yashima Y, Ozaki S, Isse T, Kim Y-D, Kim H, Kawamoto T. Expression of cytochrome P450 in tumor tissues and its association with cancer development. *Front. Biosci.* 2004 May 1;9:1967–76.
21. McFadyen MCE, Melvin WT, Murray GI. Cytochrome P450 enzymes: novel options for cancer therapeutics. *Mol. Cancer Ther.* 2004 Mar;3(3):363–71.
22. Löhr M, McFadyen MCE, Murray GI, Melvin WT. Cytochrome P450 enzymes and tumor therapy. *Mol. Cancer Ther.* 2004 Nov;3(11):1503; author reply 1503–1504.
23. Lewis DFV. Quantitative structure–activity relationships (QSARs) within the cytochrome P450 system: QSARs describing substrate binding, inhibition and induction of P450s. *Inflammopharmacology*. 2003 Feb 1;11(1):43–73.
24. Chohan KK, Paine SW, Waters NJ. Quantitative Structure Activity Relationships in Drug Metabolism. *Current Topics in Medicinal Chemistry*. 2006;6(15):1569–78.
25. Sridhar J, Liu J, Foroozesh M, Stevens CLK. Insights on Cytochrome P450 Enzymes and Inhibitors Obtained Through QSAR Studies. *Molecules*. 2012 Aug 3;17(8):9283–305.
26. Lewis DF. 57 varieties: the human cytochromes P450. *Pharmacogenomics*. 2004 Apr;5(3):305–18.
27. Wolf CR, Smith G. Pharmacogenetics. *Brit. Med. Bull.* 1999;55(2):366–86.
28. Pelkonen O, Mäenpää J, Taavitsainen P, Rautio A, Raunio H. Inhibition and induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica*. 1998 Dec;28(12):1203–53.
29. Flockhart D.A. Drug Interactions: Cytochrome P450 Drug Interaction Table. Indiana University School of Medicine [Internet]. 2007 [cited 2010 Nov 2]. Available from: <http://medicine.iupui.edu/clinpharm/ddis/table.asp>
30. Wang B, Zhou S-F. Synthetic and natural compounds that interact with human cytochrome P450 1A2 and implications in drug development. *Curr. Med. Chem.* 2009;16(31):4066–218.
31. Iori F, da Fonseca R, Ramos MJ, Menziani MC. Theoretical quantitative structure-activity relationships of flavone ligands interacting with cytochrome P450 1A1 and 1A2 isozymes. *Bioorg. Med. Chem.* 2005 Jul 15;13(14):4366–74.
32. Korhonen LE, Rahnasto M, Mähönen NJ, Wittekindt C, Poso A, Juvonen RO, Raunio H. Predictive three-dimensional quantitative structure-activity relationship of cytochrome P450 1A2 inhibitors. *J. Med. Chem.* 2005 Jun 2;48(11):3808–15.
33. Roy K, Roy PP. Comparative QSAR studies of CYP1A2 inhibitor flavonoids using 2D and 3D descriptors. *Chem Biol Drug Des.* 2008 Nov;72(5):370–82.
34. Sridhar J, Jin P, Liu J, Foroozesh M, Stevens CLK. In silico studies of polyaromatic hydrocarbon inhibitors of cytochrome P450 enzymes 1A1, 1A2, 2A6, and 2B1. *Chem. Res. Toxicol.* 2010 Mar 15;23(3):600–7.
35. Sridhar J, Foroozesh M, Stevens CLK. QSAR models of cytochrome P450 enzyme 1A2 inhibitors using CoMFA, CoMSIA and HQSAR. *SAR QSAR Environ Res.*

- 2011 Oct;22(7-8):681–97.
36. Chohan KK, Paine SW, Mistry J, Barton P, Davis AM. A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *J. Med. Chem.* 2005 Aug 11;48(16):5154–61.
 37. Burton J, Ijjaali I, Barberan O, Petitot F, Vercauteren DP, Michel A. Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset. *J. Med. Chem.* 2006 Oct 19;49(21):6231–40.
 38. Vasanathanathan P, Taboureau O, Oostenbrink C, Vermeulen NPE, Olsen L, Jørgensen FS. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* 2009 Mar;37(3):658–64.
 39. Novotarskyi S, Sushko I, Körner R, Pandey AK, Tetko IV. A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. *J. Chem. Inf. Model.* 2011 Jun 27;51(6):1271–80.
 40. Cheng F, Yu Y, Shen J, Yang L, Li W, Liu G, Lee PW, Tang Y. Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers. *J. Chem. Inf. Model.* 2011 May 23;51(5):996–1011.
 41. Mancy A, Broto P, Dijols S, Dansette PM, Mansuy D. The substrate binding site of human liver cytochrome P450 2C9: an approach using designed tienilic acid derivatives and molecular modeling. *Biochemistry.* 1995 Aug 22;34(33):10365–75.
 42. Jones JP, He M, Trager WF, Rettie AE. Three-dimensional quantitative structure-activity relationship for inhibitors of cytochrome P4502C9. *Drug Metab. Dispos.* 1996 Jan;24(1):1–6.
 43. Rao S, Aoyama R, Schrag M, Trager WF, Rettie A, Jones JP. A refined 3-dimensional QSAR of cytochrome P450 2C9: computational predictions of drug interactions. *J. Med. Chem.* 2000 Jul 27;43(15):2789–96.
 44. Ekins S, Bravi G, Binkley S, Gillespie JS, Ring BJ, Wikel JH, Wrighton SA. Three- and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. *Drug Metab. Dispos.* 2000 Aug;28(8):994–1002.
 45. Lardy MA, LeBrun L, Bullard D, Kissinger C, Gobbi A. Building a Three-Dimensional Model of CYP2C9 Inhibition Using the Autocorrelator: An Autonomous Model Generator. *J. Chem. Inf. Model.* 2012 May 25;52(5):1328–36.
 46. Yasuo K, Yamaotsu N, Gouda H, Tsujishita H, Hirono S. Structure-Based CoMFA As a Predictive Model - CYP2C9 Inhibitors As a Test Case. *J. Chem. Inf. Model.* 2009 Apr 27;49(4):853–64.
 47. Suzuki H, Kneller MB, Haining RL, Trager WF, Rettie AE. (+)-N-3-Benzyl-nirvanol and (-)-N-3-benzyl-phenobarbital: new potent and selective in vitro inhibitors of CYP2C19. *Drug Metab. Dispos.* 2002 Mar;30(3):235–9.
 48. Locuson CW, Wahlstrom JL. Three-dimensional quantitative structure-activity relationship analysis of cytochromes p450: effect of incorporating higher-affinity ligands and potential new applications. *Drug Metab. Dispos.* 2005 Jul;33(7):873–8.
 49. Lewis DFV, Lake BG, Ito Y, Dickins M. Lipophilicity relationships in inhibitors of CYP2C9 and CYP2C19 enzymes. *J Enzyme Inhib Med Chem.* 2006 Aug;21(4):385–9.
 50. Jónsdóttir SÓ, Ringsted T, Nikolov NG, Dybdahl M, Wedebye EB, Niemelä JR. Identification of cytochrome P450 2D6 and 2C9 substrates and inhibitors by QSAR analysis. *Bioorg. Med. Chem.* 2012 Mar 15;20(6):2042–53.
 51. Lewis DFV, Modi S, Dickins M. Structure-activity relationship for human

- cytochrome P450 substrates and inhibitors. *Drug Metab. Rev.* 2002 May;34(1-2):69–82.
52. Vaz RJ, Nayeem A, Santone K, Chandrasena G, Gavai AV. A 3D-QSAR model for CYP2D6 inhibition in the aryloxypropanolamine series. *Bioorg. Med. Chem. Lett.* 2005 Sep 1;15(17):3816–20.
 53. Ai C, Li Y, Wang Y, Chen Y, Yang L. Insight into the effects of chiral isomers quinidine and quinine on CYP2D6 inhibition. *Bioorg. Med. Chem. Lett.* 2009 Feb 1;19(3):803–6.
 54. Hammann F, Gutmann H, Baumann U, Helma C, Drewe J. Classification of cytochrome p(450) activities using machine learning methods. *Mol. Pharm.* 2009 Dec;6(6):1920–6.
 55. Mao B, Gozalbes R, Barbosa F, Migeon J, Merrick S, Kamm K, Wong E, Costales C, Shi W, Wu C, Froloff N. QSAR modeling of in vitro inhibition of cytochrome P450 3A4. *J Chem Inf Model.* 2006 Oct;46(5):2125–34.
 56. Roy K, Pratim Roy P. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. *Eur J Med Chem.* 2009 Jul;44(7):2913–22.
 57. Didziapetris R, Dapkunas J, Sazonovas A, Japertas P. Trainable structure-activity relationship model for virtual screening of CYP3A4 inhibition. *J. Comput. Aided Mol. Des.* 2010 Nov;24(11):891–906.
 58. Gleeson MP, Davis AM, Chohan KK, Paine SW, Boyer S, Gavaghan CL, Arnby CH, Kankkonen C, Albertson N. Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models. *J. Comput. Aided Mol. Des.* 2007 Nov;21(10-11):559–73.
 59. Dagliyan O, Kavakli IH, Turkay M. Classification of cytochrome P450 inhibitors with respect to binding free energy and pIC50 using common molecular descriptors. *J. Chem. Inf. Model.* 2009 Oct;49(10):2403–11.
 60. Crivori P, Poggesi I. Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. *Eur J Med Chem.* 2006 Jul;41(7):795–808.
 61. Li H, Sun J, Fan X, Sui X, Zhang L, Wang Y, He Z. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J. Comput. Aided Mol. Des.* 2008 Nov;22(11):843–55.
 62. Terfloth L, Bienfait B, Gasteiger J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J Chem Inf Model.* 2007 Aug;47(4):1688–701.
 63. Michielan L, Terfloth L, Gasteiger J, Moro S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *J. Chem. Inf. Model.* 2009 Nov;49(11):2588–605.
 64. Chowbay B, Zhou S, Lee EJD. An interethnic comparison of polymorphisms of the genes encoding drug-metabolizing enzymes and drug transporters: experience in Singapore. *Drug Metab. Rev.* 2005;37(2):327–78.
 65. Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C. Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol. Ther.* 2007 Dec;116(3):496–526.
 66. Kirchheiner J, Seeringer A. Clinical implications of pharmacogenetics of cytochrome P450 drug metabolizing enzymes. *Biochim. Biophys. Acta.* 2007

- Mar;1770(3):489–94.
67. Wang L-L, Li Y, Zhou S-F. A Bioinformatics Approach for the Phenotype Prediction of Nonsynonymous Single Nucleotide Polymorphisms in Human Cytochromes P450. *Drug Metab Dispos.* 2009 May 1;37(5):977–91.
 68. Alexanderson B, Evans DA, Sjöqvist F. Steady-state plasma levels of nortriptyline in twins: influence of genetic factors and drug therapy. *Br Med J.* 1969 Dec 27;4(5686):764–8.
 69. Human Cytochrome P450 (CYP) Allele Nomenclature Database [Internet]. [cited 2012 Nov 26]. Available from: <http://www.cypalleles.ki.se/>
 70. 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.
 71. 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):1832–4.
 72. McNaught A. The IUPAC international chemical identifier : InChI - A new standard for molecular informatics. *Chemistry International.* 2006;28(6):12–5.
 73. PDB File Format - Contents Guide Version 3.30 [Internet]. [cited 2012 Nov 26]. Available from: <http://www.wwpdb.org/documentation/format33/v3.3.html>
 74. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry.* 2010;31(2):455–61.
 75. What is the format of a PDBQT file? – AutoDock [Internet]. [cited 2012 Nov 26]. Available from: <http://autodock.scripps.edu/faqs-help/faq/what-is-the-format-of-a-pdbqt-file>
 76. Chemaxon. Marvin Beans, JChem 5.4 [Internet]. [cited 2011 Feb 16]. Available from: <http://www.chemaxon.com>
 77. Chemaxon. Standardizer, JChem 5.4 [Internet]. [cited 2011 Feb 16]. Available from: <http://www.chemaxon.com>
 78. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry.* 1996;17(5-6):490–519.
 79. Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry.* 1983;4(2):187–217.
 80. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* 1995 May 1;117(19):5179–97.
 81. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry.* 2005;26(16):1701–18.
 82. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* 2008 Mar 1;4(3):435–47.
 83. O’Boyle N, Banck M, James C, Morley C, Vandermeersch T, Hutchison G. Open Babel: An open chemical toolbox. *Journal of Cheminformatics.* 2011 Oct 7;3(1):33.

84. Vainio MJ, Johnson MS. Generating conformer ensembles using a multiobjective genetic algorithm. *J Chem Inf Model*. 2007 Dec;47(6):2462–74.
85. Puranen JS, Vainio MJ, Johnson MS. Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. *J Comput Chem*. 2010 Jun;31(8):1722–32.
86. Molecular Networks GmbH: Erlangen, Germany. CORINA [Internet]. [cited 2010 Nov 2]. Available from: <http://www.molecular-networks.com/>
87. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des*. 2002 Mar;16(3):151–66.
88. Dias R, de Azevedo WF Jr. Molecular docking algorithms. *Curr Drug Targets*. 2008 Dec;9(12):1040–7.
89. Huang S-Y, Zou X. Advances and challenges in protein-ligand docking. *Int J Mol Sci*. 2010;11(8):3016–34.
90. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*. 2004 Nov 1;3(11):935–49.
91. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*. 2002 Jun 1;47(4):409–43.
92. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol*. 1982 Oct 25;161(2):269–88.
93. Shoichet BK, Kuntz ID. Predicting the structure of protein complexes: a step in the right direction. *Chem. Biol*. 1996 Mar;3(3):151–6.
94. Wüthrich K, von Freyberg B, Weber C, Wider G, Traber R, Widmer H, Braun W. Receptor-induced conformation change of the immunosuppressant cyclosporin A. *Science*. 1991 Nov 15;254(5034):953–4.
95. Jiang F, Kim SH. “Soft docking”: matching of molecular surface cubes. *J. Mol. Biol*. 1991 May 5;219(1):79–102.
96. Ferrari AM, Wei BQ, Costantino L, Shoichet BK. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem*. 2004 Oct 7;47(21):5076–84.
97. Leach AR. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol*. 1994 Jan 7;235(1):345–56.
98. Frimurer TM, Peters GH, Iversen LF, Andersen HS, Møller NPH, Olsen OH. Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities. *Biophys. J*. 2003 Apr;84(4):2273–81.
99. Meiler J, Baker D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins*. 2006 Nov 15;65(3):538–48.
100. Nabuurs SB, Wagener M, de Vlieg J. A flexible approach to induced fit docking. *J. Med. Chem*. 2007 Dec 27;50(26):6507–18.
101. Antes I. DynaDock: A new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. *Proteins*. 2010 Apr;78(5):1084–104.
102. Knegtel RM, Kuntz ID, Oshiro CM. Molecular docking to ensembles of protein structures. *J. Mol. Biol*. 1997 Feb 21;266(2):424–40.
103. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*.

- 1998;19(14):1639–62.
104. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins*. 2002 Jan 1;46(1):34–40.
 105. Pak Y, Wang S. Application of a Molecular Dynamics Simulation Method with a Generalized Effective Potential to the Flexible Molecular Docking Problems. *J. Phys. Chem. B*. 2000 Jan 1;104(2):354–9.
 106. Hart TN, Read RJ. A multiple-start Monte Carlo docking method. *Proteins*. 1992 Jul;13(3):206–22.
 107. Morris GM, Goodsell DS, Huey R, Olson AJ. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* 1996 Aug;10(4):293–304.
 108. Trosset J-Y, Scheraga HA. Prodock: Software package for protein modeling and docking. *Journal of Computational Chemistry*. 1999;20(4):412–27.
 109. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 1997 Apr 4;267(3):727–48.
 110. Taylor JS, Burnett RM. DARWIN: a program for docking flexible molecules. *Proteins*. 2000 Nov 1;41(2):173–91.
 111. Grosdidier A, Zoete V, Michielin O. EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins*. 2007 Jun 1;67(4):1010–25.
 112. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 1996 Aug 23;261(3):470–89.
 113. Rarey M, Kramer B, Lengauer T. The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins*. 1999 Jan 1;34(1):17–28.
 114. Claussen H, Buning C, Rarey M, Lengauer T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* 2001 Apr 27;308(2):377–95.
 115. Burkhard P, Taylor P, Walkinshaw MD. An example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 Å X-ray structure of a thrombin-ligand complex. *J. Mol. Biol.* 1998 Mar 27;277(2):449–66.
 116. Wu M-Y, Dai D-Q, Yan H. PRL-Dock: protein-ligand docking based on hydrogen bond matching and probabilistic relaxation labeling. *Proteins*. 2012 Aug;80(9):2137–53.
 117. Tang H-X, Ye Y-Z, Ding D-F. Flexible Docking of Proteins and “Drug-like” Ligands. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao*. 1998;30(6):623–30.
 118. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins*. 1998 Nov 15;33(3):367–82.
 119. Pei J, Wang Q, Liu Z, Li Q, Yang K, Lai L. PSI-DOCK: towards highly efficient and accurate flexible ligand docking. *Proteins*. 2006 Mar 1;62(4):934–46.
 120. Baxter J. Local Optima Avoidance in Depot Location. *The Journal of the Operational Research Society*. 1981 Sep;32(9):815.
 121. Blum C, Roli A, Sampels M, editors. *Hybrid Metaheuristics: An Emerging Approach to Optimization*. 1st ed. Springer; 2008.
 122. Numerical Optimization [Internet]. [cited 2012 Nov 26]. Available from:

<http://www.springer.com/mathematics/book/978-0-387-30303-1>

123. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.* 2002 Jan;16(1):11–26.
124. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References.* 2nd ed. Wiley-VCH; 2009.
125. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solovev V, Hoonakker F, Tetko IV, Marcou G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aid. Drug.* 2008 Sep;4:191–198(8).
126. Solov'ev, Varnek, Wipff. Modeling of ion complexation and extraction using substructural molecular fragments. *J. Chem. Inf. Comput. Sci.* 2000 May;40(3):847–58.
127. Kier LB, Hall LH. *Molecular Structure Description: The Electrotopological State.* Academic Press; 1999.
128. Bredel M, Jacoby E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* 2004 Apr;5(4):262–75.
129. Yabuuchi H, Niijima S, Takematsu H, Ida T, Hirokawa T, Hara T, Ogawa T, Minowa Y, Tsujimoto G, Okuno Y. Analysis of multiple compound–protein interactions reveals novel bioactive molecules. *Mol Syst Biol.* 2011 Mar 1;7:472.
130. Mahasenani KV, Li C. Novel Inhibitor Discovery through Virtual Screening against Multiple Protein Conformations Generated via Ligand-Directed Modeling: A Maternal Embryonic Leucine Zipper Kinase Example. *J. Chem. Inf. Model.* 2012 May 25;52(5):1345–55.
131. Wendt B, Uhrig U, Bös F. Capturing Structure–Activity Relationships from Chemogenomic Spaces. *J. Chem. Inf. Model.* 2011 Apr 25;51(4):843–51.
132. Meslamani J, Rognan D. Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel. *J. Chem. Inf. Model.* 2011 Jul 25;51(7):1593–603.
133. Weill N, Rognan D. Development and Validation of a Novel Protein–Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands. *J. Chem. Inf. Model.* 2009 Apr 27;49(4):1049–62.
134. Ortiz AR, Pisabarro MT, Gago F, Wade RC. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* 1995 Jul 7;38(14):2681–91.
135. Henrich S, Feierberg I, Wang T, Blomberg N, Wade RC. Comparative binding energy analysis for binding affinity and target selectivity prediction. *Proteins.* 2010 Jan;78(1):135–53.
136. Murcia M, Ortiz AR. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. *J. Med. Chem.* 2004 Feb 12;47(4):805–20.
137. Coderch C, Klett J, Morreale A, Fernando Díaz J, Gago F. Comparative binding energy (COMBINE) analysis supports a proposal for the binding mode of eptophilones to β -tubulin. *ChemMedChem.* 2012 May;7(5):836–43.
138. Nakamura S, Nakanishi I, Kitaura K. Binding affinity prediction of non-peptide inhibitors of HIV-1 protease using COMBINE model introduced from peptide inhibitors. *Bioorg. Med. Chem. Lett.* 2006 Dec 15;16(24):6334–7.
139. Tomić S, Bertosa B, Wang T, Wade RC. COMBINE analysis of the specificity of

- binding of Ras proteins to their effectors. *Proteins*. 2007 May 1;67(2):435–47.
140. Tomic S, Nilsson L, Wade RC. Nuclear receptor-DNA binding specificity: A COMBINE and Free-Wilson QSAR analysis. *J. Med. Chem.* 2000 May 4;43(9):1780–92.
 141. Kramer C, Gedeck P. Global free energy scoring functions based on distance-dependent atom-type pair descriptors. *J Chem Inf Model*. 2011 Mar 28;51(3):707–20.
 142. Müller K-R, Mika S, Rätsch G, Tsuda K, Schölkopf B. An introduction to kernel-based learning algorithms. *IEEE TRANSACTIONS ON NEURAL NETWORKS*. 2001;12(2):181–201.
 143. Cristianini N, Shawe-Taylor J. An introduction to support Vector Machines: and other kernel-based learning methods. New York, NY, USA: Cambridge University Press; 2000.
 144. Rosenblatt F. The Perceptron - a perceiving and recognizing automaton. Cornell Aeronautical Laboratory; 1957. Report No.: 85-460-1.
 145. Rumelhart DE, Hinton GE, Williams RJ. Learning Internal Representations by Error Propagation. 1985 Sep.
 146. Tollenaere T. SuperSAB: fast adaptive back propagation with good scaling properties. *Neural Netw*. 1990 Oct;3(5):561–73.
 147. Tetko IV. Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* 2002 Jun;42(3):717–28.
 148. Tetko IV. Associative neural network. *Meth. Mol. Biol.* 2008;458:185–202.
 149. Vapnik VN. *Statistical Learning Theory*. Wiley-Interscience; 1998.
 150. Cortes C, Vapnik V. Support-Vector Networks. *Mach. Learn.* 1995 Sep;20(3):273–97.
 151. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Chapman & Hall, New York, NY; 1984.
 152. Breiman L. Bagging predictors. *Mach. Learn.* 1996 Aug;24:123–40.
 153. Quinlan R, Quinlan JR. *C4.5: Programs for Machine Learning*. Revised, Update. Morgan Kaufman Publ Inc; 1992.
 154. University of Waikato: Waikato, New Zeland. Weka: Waikato Environment for Knowledge Analysis [Internet]. [cited 2010 Nov 2]. Available from: <http://www.cs.waikato.ac.nz/ml/weka/>
 155. 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):1407–21.
 156. 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):3103–10.
 157. 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):5601–4.
 158. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* 1975 Oct 20;405(2):442–51.
 159. 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):155–73.
160. 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. *J. Chemometr.* 2010 Apr;24(3-4):202–8.
161. 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):1733–46.
162. Tetko IV, Poda GI, Ostermann C, Mannhold R. Accurate In Silico log P Predictions: One Can't Embrace the Unembraceable. *QSAR & Combinatorial Science.* 2009;28(8):845–9.
163. Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A, Li J, Gramatica P, Hansen K, Schroeter T, Müller K-R, Xi L, Liu H, Yao X, Oberg 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, Muller C, Varnek A, Prokopenko VV, Tetko IV. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* 2010 Dec 27;50(12):2094–111.
164. Oberg T. A QSAR for baseline toxicity: validation, domain of application, and prediction. *Chem. Res. Toxicol.* 2004 Dec;17(12):1630–7.
165. Luilo GB, Cabaniss SE. Quantitative structure-property relationship for predicting chlorine demand by organic molecules. *Environ. Sci. Technol.* 2010 Apr 1;44(7):2503–8.
166. Papa E, Gramatica P. Externally validated QSPR modelling of VOC tropospheric oxidation by NO₃ radicals. *SAR QSAR Environ Res.* 2008;19(7-8):655–68.
167. Sushko I. Applicability domain of QSAR models. PhD Thesis. TUM, Lehrstuhl für Genomorientierte Bioinformatik; 2011.
168. Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang Q-Y, 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. *J. Comput. Aided Mol. Des.* 2011 Jun;25(6):533–54.
169. 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):223–41.
170. Tetko IV, Livingstone DJ. 5.27 - Rule-Based Systems to Predict Lipophilicity. In: Editors-in-Chief: John B. Taylor, David J. Triggle, editors. *Comprehensive Medicinal Chemistry II* [Internet]. Oxford: Elsevier; 2007 [cited 2012 Nov 27]. p. 649–68. Available from: <http://www.sciencedirect.com/science/article/pii/B008045044X001449>
171. Kaiser J. Science resources. Chemists want NIH to curtail database. *Science.* 2005 May 6;308(5723):774.

172. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*. 2009 Jun 4;37(Web Server):W623–W633.
173. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH. PubChem's BioAssay Database. *Nucleic Acids Res*. 2012 Jan;40(Database issue):D400–412.
174. ChemSpider | The free chemical database [Internet]. [cited 2012 Nov 27]. Available from: <http://www.chemspider.com/>
175. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. DrugBank 3.0: a comprehensive resource for “omics” research on drugs. *Nucleic Acids Res*. 2011 Jan;39(Database issue):D1035–1041.
176. ChemExper - catalog of chemicals suppliers, physical characteristics and search engine [Internet]. [cited 2012 Nov 27]. Available from: <http://www.chemexper.com/>
177. Williams AJ. Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today*. 2008 Jun;13(11-12):502–6.
178. 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):453–63.
179. Hardy B, Douglas N, Helma C, Rautenberg M, Jeliaskova N, Jeliaskov V, Nikolova I, Benigni R, Tcheremenskaia O, Kramer S, Girschick T, Buchwald F, Wicker J, Karwath A, Gütlein M, Maunz A, Sarimveis H, Melagraki G, Afantitis A, Sopasakis P, Gallagher D, Poroikov V, Filimonov D, Zakharov A, Lagunin A, Glorizova T, Novikov S, Skvortsova N, Druzhilovsky D, Chawla S, Ghosh I, Ray S, Patel H, Escher S. Collaborative development of predictive toxicology applications. *J Cheminform*. 2010;2(1):7.
180. Chemical Substances - CAS REGISTRY [Internet]. [cited 2012 Nov 27]. Available from: <http://www.cas.org/content/chemical-substances>
181. PubMed - The world's largest collection of biomedical literature citations [Internet]. [cited 2012 Nov 27]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/>
182. 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):133–44.
183. Livingstone DJ, Manallack DT, Tetko IV. Data modelling with neural networks: advantages and limitations. *J. Comput. Aided Mol. Des*. 1997 Mar;11(2):135–42.
184. Tropsha A, Gramatica P, Gombar VK. 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.
185. ADRIANA.Code - Calculation of Molecular Descriptors [Internet]. [cited 2012 Nov 27]. Available from: <http://www.molecular-networks.com/products/adrianacode>
186. 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. *Curr. Pharm. Des*. 2006;12(17):2111–20.
187. Aires-de-Sousa J, Gasteiger J. New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective

- reactions. *J Chem Inf Comput Sci*. 2001 Apr;41(2):369–75.
188. 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):373–88.
 189. Aires-de-Sousa J, Gasteiger J. Prediction of enantiomeric excess in a combinatorial library of catalytic enantioselective reactions. *J Comb Chem*. 2005 Apr;7(2):298–301.
 190. Dimoglo A s., Shvets N m., Tetko I v., Livingstone D j. Electronic-Topological Investigation of the Structure – Acetylcholinesterase Inhibitor Activity Relationship in the Series of N-Benzylpiperidine Derivatives. Quantitative Structure-Activity Relationships. 2001;20(1):31–45.
 191. Skvortsova MI, Baskin II, Skvortsov LA, Palyulin VA, Zefirov NS, Stankevich IV. Chemical graphs and their basis invariants. *Journal of Molecular Structure*. 1999;466(1):211–7.
 192. 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):2047–56.
 193. Stewart JJP. Optimization of parameters for semiempirical methods I. *Method. Journal of Computational Chemistry*. 1989;10(2):209–20.
 194. 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):1040–4.
 195. Potemkin VA, Grishina MA. A new paradigm for pattern recognition of drugs. *J. Comput. Aided Mol. Des*. 2008 Jul;22(6-7):489–505.
 196. Potemkin VA, Pogrebnoy AA, Grishina MA. Technique for energy decomposition in the study of “receptor-ligand” complexes. *J Chem Inf Model*. 2009 Jun;49(6):1389–406.
 197. Bender A, Mussa HY, Glen RC, Reiling S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J Chem Inf Comput Sci*. 2004 Oct;44(5):1708–18.
 198. 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):5674–90.
 199. Whitley DC, Ford MG, Livingstone DJ. Unsupervised forward selection: a method for eliminating redundant variables. *J Chem Inf Comput Sci*. 2000 Oct;40(5):1160–8.
 200. National Library of Medicine, National Institute of Health. The PubChem Project [Internet]. [cited 2010 Nov 2]. Available from: <http://pubchem.ncbi.nlm.nih.gov/>
 201. National Library of Medicine, National Institute of Health. PubChem BioAssay AID-410 [Internet]. [cited 2012 Nov 27]. Available from: <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=410>
 202. Promega. P450-Glo(TM) Assays [Internet]. [cited 2010 Nov 2]. Available from: <http://www.promega.com/tbs/tb325/tb325.html>
 203. National Library of Medicine, National Institute of Health. PubChem BioAssay AID-883 [Internet]. [cited 2012 Nov 27]. Available from: <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=883>

204. National Library of Medicine, National Institute of Health. PubChem BioAssay AID-899 [Internet]. [cited 2012 Nov 27]. Available from:
<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=899>
205. National Library of Medicine, National Institute of Health. PubChem BioAssay AID-891 [Internet]. [cited 2012 Nov 27]. Available from:
<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=891>
206. National Library of Medicine, National Institute of Health. PubChem BioAssay AID-884 [Internet]. [cited 2012 Nov 27]. Available from:
<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=884>
207. Veith H, Southall N, Huang R, James T, Fayne D, Artemenko N, Shen M, Inglese J, Austin CP, Lloyd DG, Auld DS. Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nature Biotechnology*. 2009;27(11):1050–5.
208. National Library of Medicine, National Institute of Health. PubChem BioAssay AID-1851 [Internet]. [cited 2012 Nov 27]. Available from:
<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1851>
209. Ibeanu GC, Ghanayem BI, Linko P, Li L, Pedersen LG, Goldstein JA. Identification of Residues 99, 220, and 221 of Human Cytochrome P450 2C19 as Key Determinants of Omeprazole Hydroxylase Activity. *J. Biol. Chem.* 1996 May 24;271(21):12496–501.
210. Reynald RL, Sansen S, Stout CD, Johnson EF. Structural characterization of human cytochrome P450 2C19: active site differences between P450's 2C8, 2C9 and 2C19. *J. Biol. Chem.* 2012 Nov 1;
211. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. 2001.
212. Wada Y, Mitsuda M, Ishihara Y, Watanabe M, Iwasaki M, Asahi S. Important amino acid residues that confer CYP2C19 selective activity to CYP2C9. *J. Biochem.* 2008 Sep;144(3):323–33.

Appendix

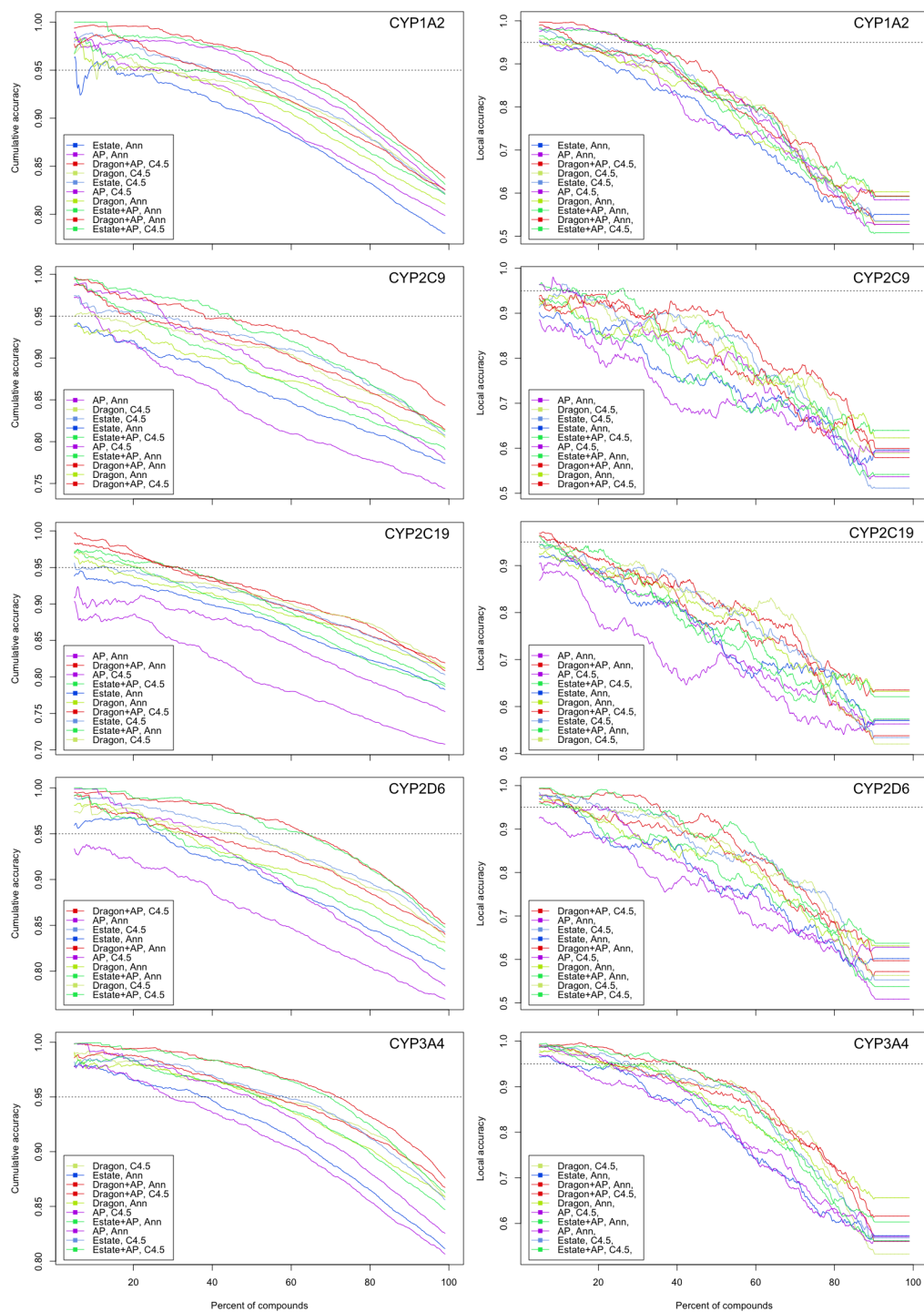


Figure A1. Cumulative and local balanced accuracies of model predictions, ordered by BAGGING-STD DM.

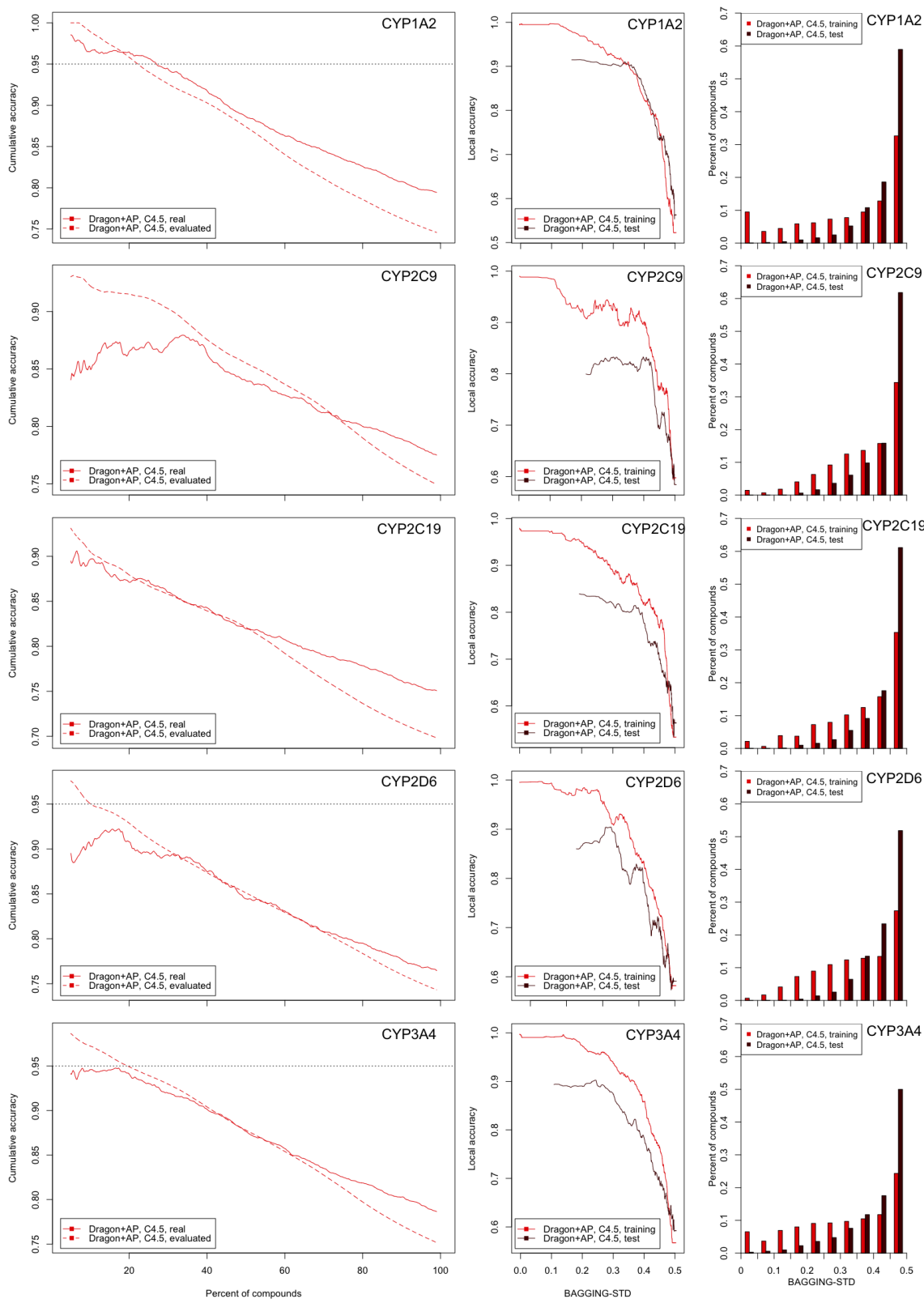


Figure A2. Actual and estimated cumulative applicability domain accuracy plots. Displayed model is C4.5 decision tree model built on Dragon + atom pair descriptors.

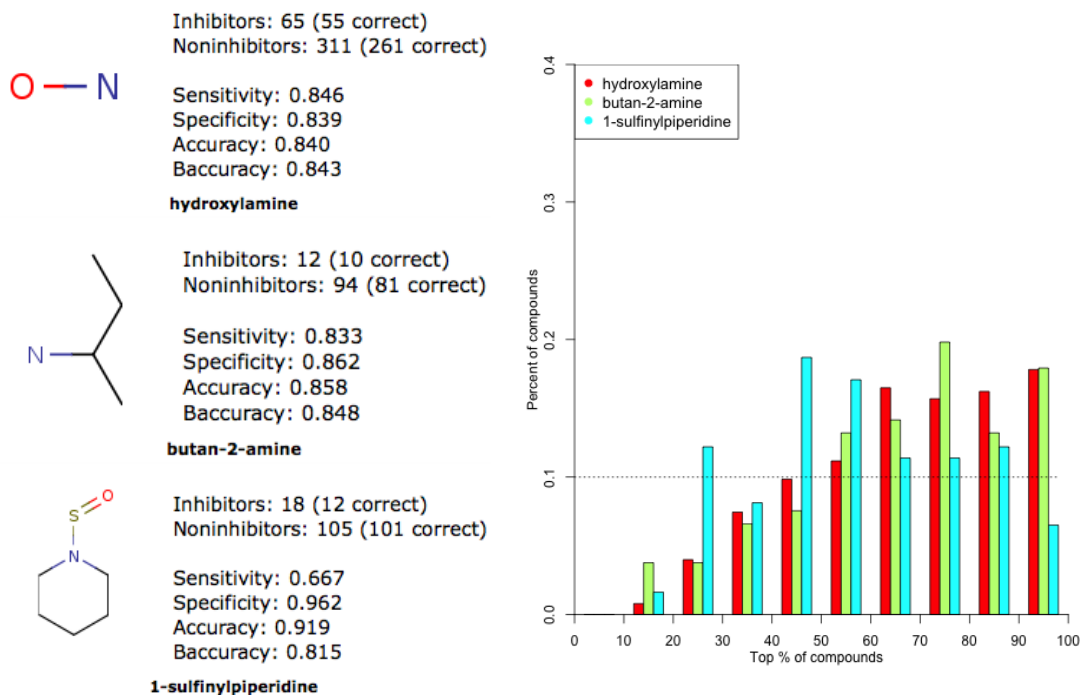


Figure A3. Diagram of best-predicted fragments for CYP1A2 isoform

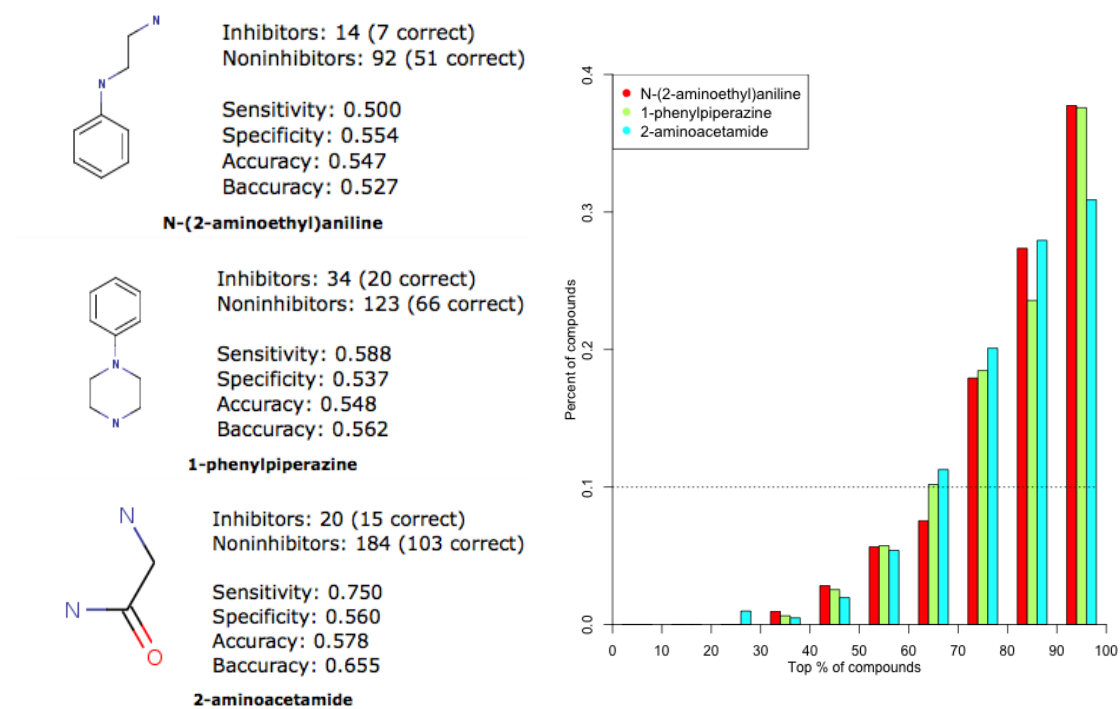


Figure A4. Diagram of worst-predicted fragments for CYP1A2 isoform

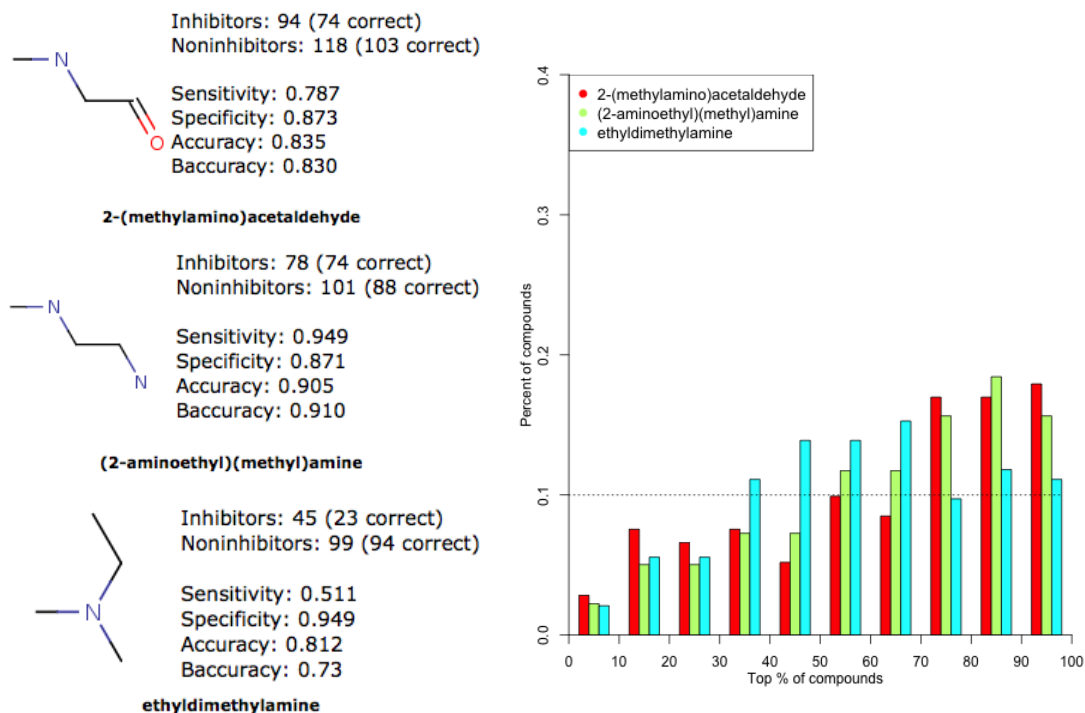


Figure A5. Diagram of best-predicted fragments for CYP2C9 isoform

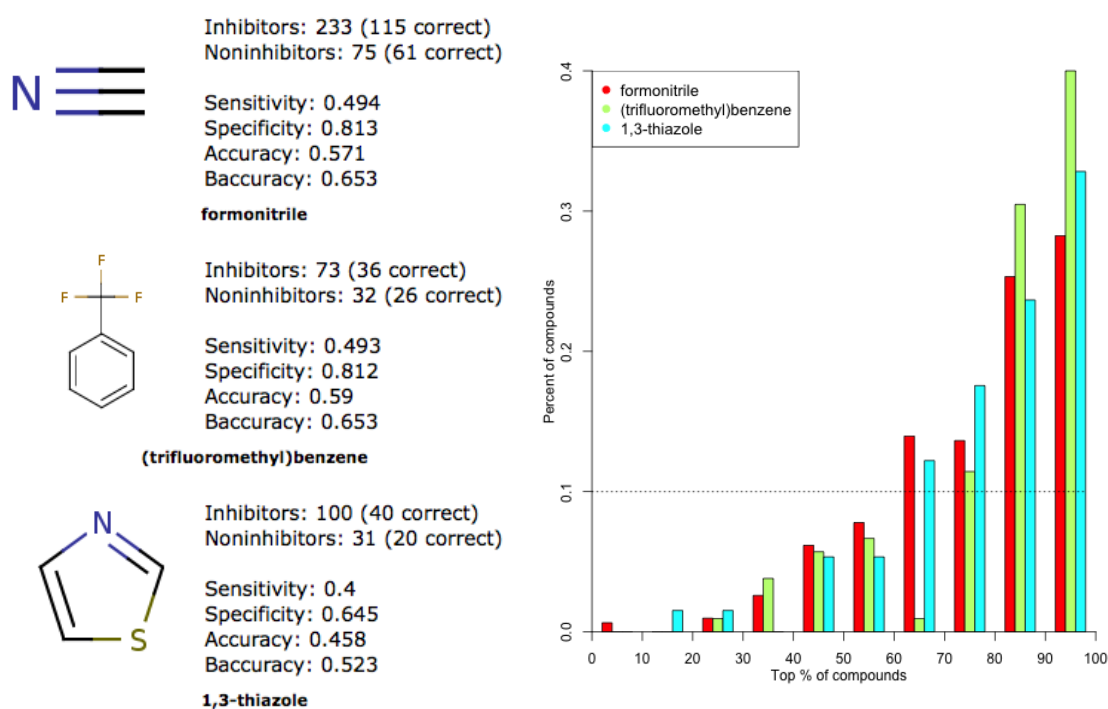


Figure A6. Diagram of worst-predicted fragments for CYP2C9 isoform

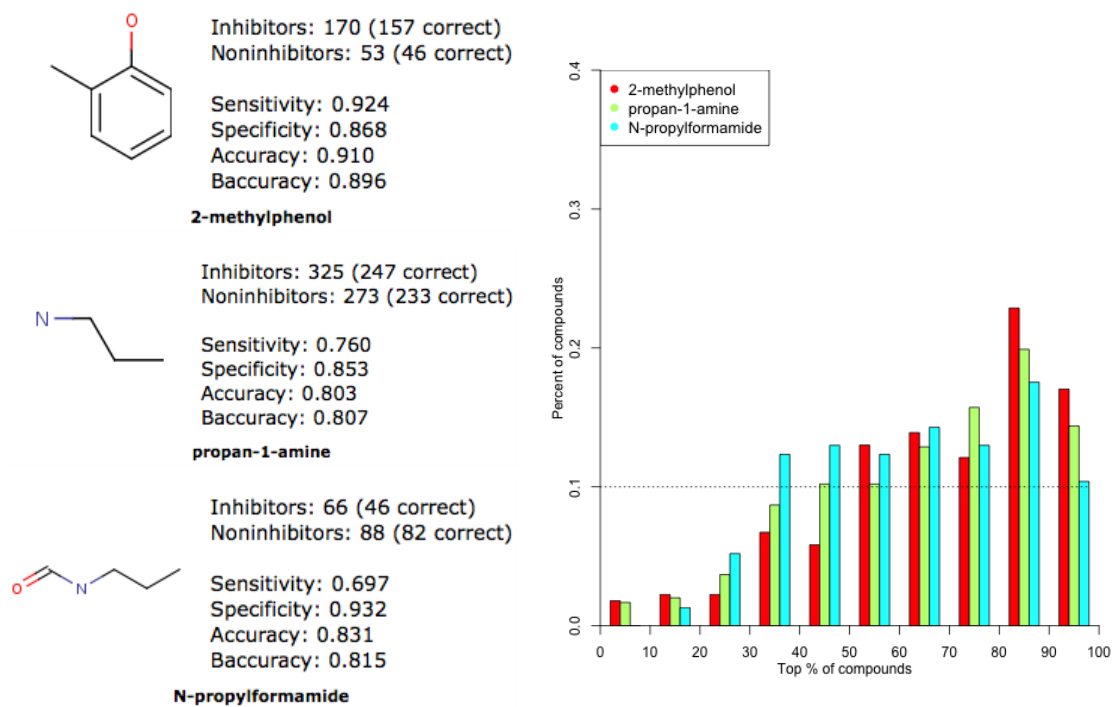


Figure A7. Diagram of best-predicted fragments for CYP2C19 isoform

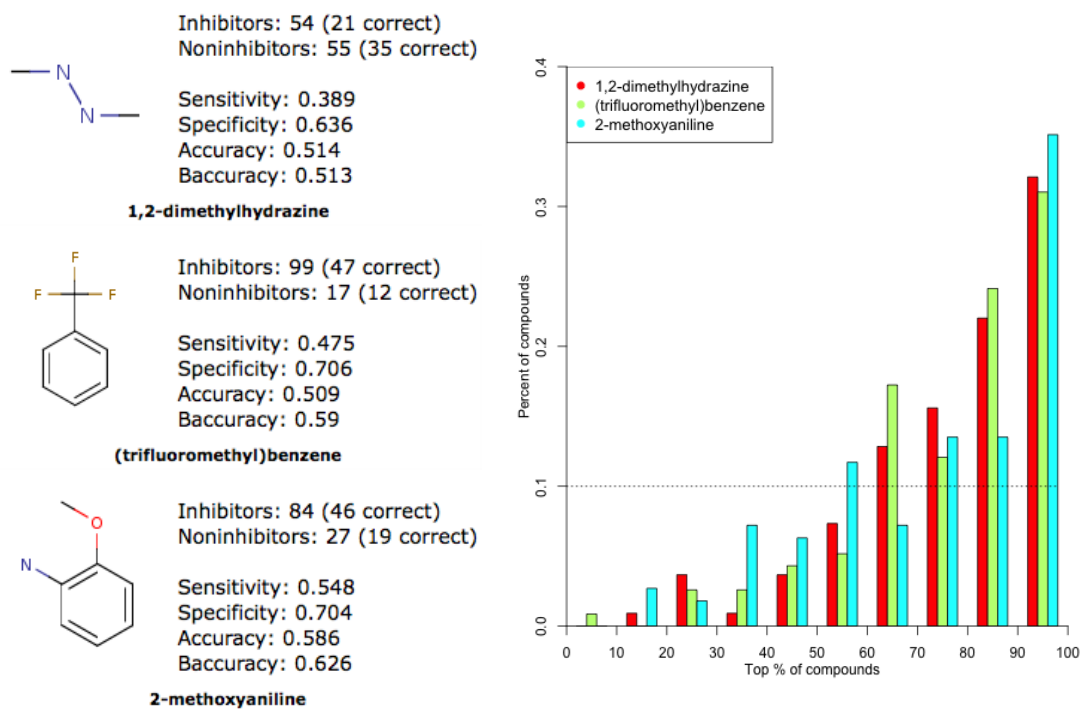


Figure A8. Diagram of worst-predicted fragments for CYP2C19 isoform

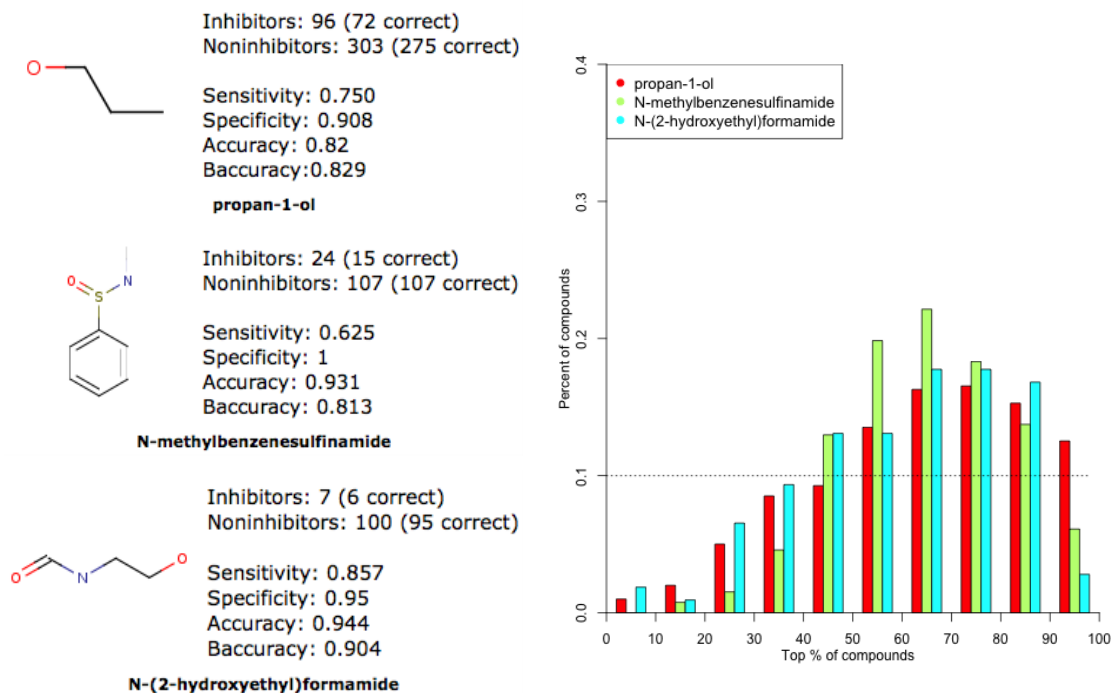


Figure A9. Diagram of best-predicted fragments for CYP2D6 isoform

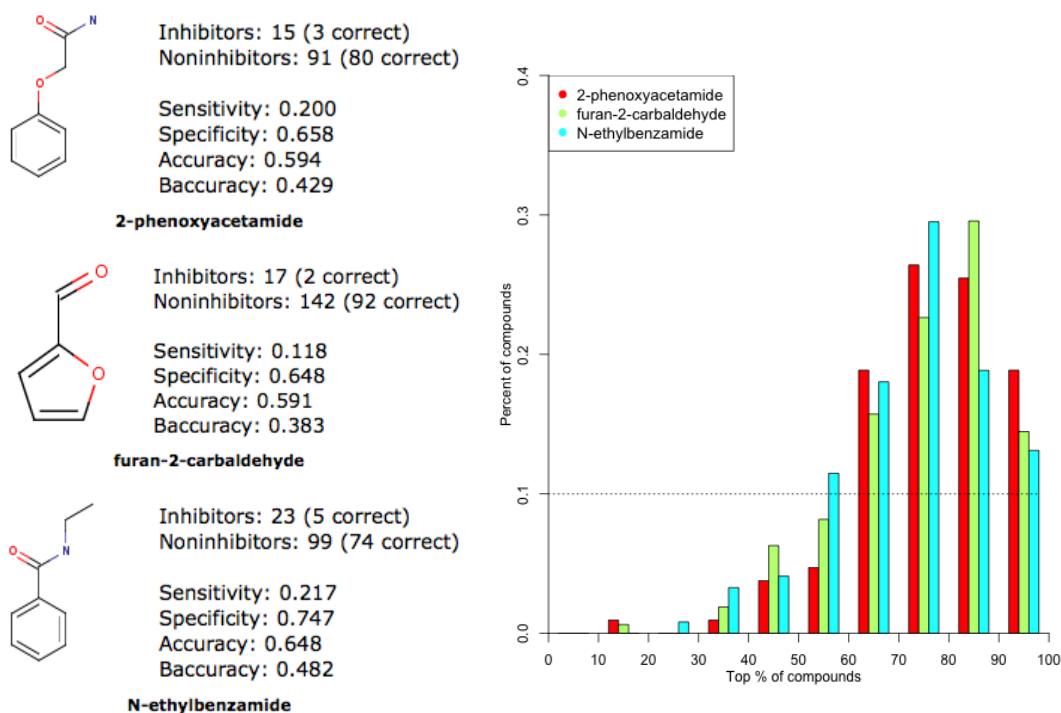


Figure A10. Diagram of worst-predicted fragments for CYP2D6 isoform

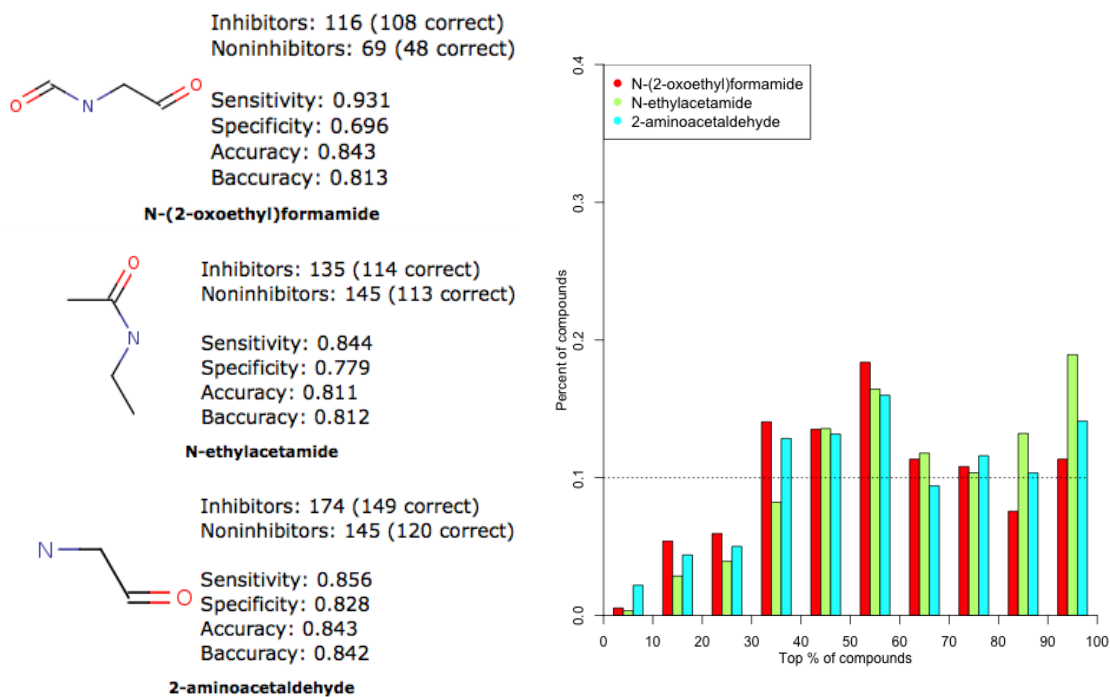


Figure A11. Diagram of best-predicted fragments for CYP3A4 isoform

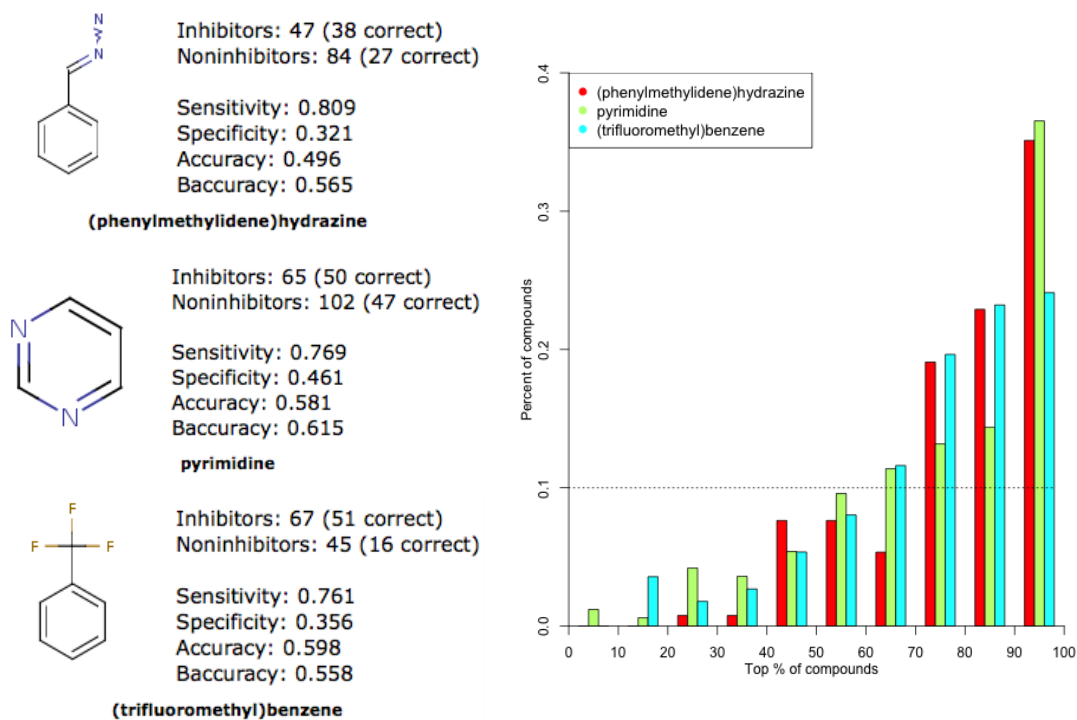


Figure A12. Diagram of worst-predicted fragments for CYP3A4 isoform

Publication record

Journal articles

Sushko I, **Novotarskyi S**, Körner R, Pandey AK, Kovalishyn VV, Prokopenko VV, et al. Applicability domain for in silico models to achieve accuracy of experimental measurements. *J. Chemometr.* 2010 Apr;24(3-4):202–8.

Sushko I, **Novotarskyi S**, Körner R, Pandey AK, Cherkasov A, Li J, et al. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* 2010 Dec 27;50(12):2094–111.

Novotarskyi S, Sushko I, Körner R, Pandey AK, Tetko IV. A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. *J. Chem. Inf. Model.* 2011 Jun 27;51(6):1271–80.

Sushko I*, **Novotarskyi S***, Körner R*, Pandey AK*, Rupp M, Teetz W, et al. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* 2011 Jun;25(6):533–54. *equal contribution

Oprisiu I, **Novotarskyi S**, Tetko IV. Modeling of Non-Additive Mixture Properties using an Online Chemical Database and Modeling Environment. *J. Cheminform.* 2013 Jan 15;5(1):4

Brandmaier S, **Novotarskyi S**, Sushko I, Tetko IV. From descriptors to predicted properties: Experimental design using the applicability domain estimation. *ATLA* 2013 Mar; 41(1):33-47. pp.33-47

Tetko I, Sopasakis P, Kunwar P, Brandmaier S, **Novotarskyi S**, Charochkina L, Prokopenko V, Peijnenburg W. Prioritisation of Polybrominated Diphenyl Ethers (PBDEs) by Using the QSPR-THESAURUS Web Tool. *ATLA* 2013 Mar; 41:127-135. pp.127-135

Conference talks

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

Novotarskyi S, Sushko I, Körner R, Tetko IV. An applicability domain approach in QSAR modeling of human cytochrome P450 inhibition. XIII Chemometrics in Analytical Chemistry, Budapest, Hungary, 25-29 June 2012

Tutoring

Novotarskyi S, Sushko I. Online Chemical Modeling Environment - an introductory lecture. Achievements and applications of contemporary informatics, mathematics and physics (AACIMP-08), Kiev, 11-24 August 2008

Novotarskyi S, Sushko I. Online Chemical Modeling Environment. Environmental Chemoinformatics course. Achievements and applications of contemporary informatics, mathematics and physics (AACIMP-09), Kiev, 5-16 August 2009

Novotarskyi S, Sushko I, Körner R. Introduction to the QSAR research using the novel chemical modeling framework. 1st Autumn School of Environmental ChemOinformatics (ECO), Munich, 18-22 October 2010

Novotarskyi S, Introduction to molecular docking. Winter School of Environmental ChemOinformatics (ECO), Kalmar, 25-28 February 2013

Curriculum vitae

Sergii Novotarskyi

Personal data

Date of birth 02 September 1984
Nationality Ukrainian
Marital status Single

Education

2008 – 2011 PhD student
Helmholtz-Zentrum, Munich, Germany
Topic: QSAR approaches to predict human cytochrome P450 inhibition
Supervisor: Prof. H.-W. Mewes
Advisor: Dr. I. Tetko

2001-2007 Master of Science
with distinction (average mark 5.0 / 5.0)
Chair of Computing Technics
Faculty of Informatics and Computing Technics
National Technical University of Ukraine
Major: Computer Systems and Networks
Topic: Locally-asynchronous models of computational structures oriented on solving boundary value problems
Supervisor: Prof. V. Shyrochyn

1998-2001 High school
Kyiv Polytechnic Lyceum, Faculty of Physics and Mathematics

1991-1998 Elementary school

Scientific interests

chemoinformatics, QSAR/QSPR, CYP inhibition prediction, early stage drug design, applicability domain, molecular docking, molecular descriptors

Computer skills

Programming Programming languages: Java, PHP, Python, Object Pascal
Database management: MySQL / Percona, MongoDB
Operating systems: Linux, Mac OS X, Windows

Analytical tools R

Web development XSLT, Javascript, Ajax, HTML+CSS, jQuery

Languages

Ukrainian native
Russian native
English fluent
German solid basics