

VILNIUS UNIVERSITY

Justas Dapkūnas

COMPUTATIONAL MODELING OF CYTOCHROME P450-MEDIATED
DRUG METABOLISM

Doctoral dissertation
Physical sciences, biochemistry (04P)

Vilnius, 2011

The dissertation work was carried out at Vilnius University from 2007 to 2011 in cooperation with researchers at VŠĮ "Aukštieji algoritmai".

Scientific supervisor:

dr. Remigijus Didžiapetris (VŠĮ "Aukštieji algoritmai", physical sciences, biochemistry – 04P)

VILNIAUS UNIVERSITETAS

Justas Dapkūnas

CITOCROMŲ P450 KATALIZUOJAMO VAISTŲ METABOLIZMO
KOMPIUTERINIS MODELIAVIMAS

Daktaro disertacija
Fiziniai mokslai, biochemija (04P)

Vilnius, 2011

Disertacija rengta 2007 - 2011 metais Vilniaus universitete bendradarbiaujant su VŠĮ „Aukštieji algoritmai“.

Mokslinis vadovas:

dr. Remigijus Didžiapetris (VŠĮ „Aukštieji algoritmai“, fiziniai mokslai, biochemija – 04P)

Acknowledgements

I would like to thank my scientific advisor dr. Remigijus Didžiapetris and the director of VŠĮ “Aukštieji Algoritmai” dr. Pranas Japertas for their help, patience, and guidance through all the years of working there. I very appreciate the valuable discussions on development of the models and interpretation of the obtained results with my colleagues dr. Andrius Sazonovas and Kiril Lanevskij. Additionally, I acknowledge dr. Laura Steponėnienė and Liutauras Juška who helped me with the analysis of experimental regioselectivity data, and also Dainius Šimelevičius, dr. Rytis Kubilius, and Tomas Bukėnas for development of the software tools that were used in this work.

I would like to thank the people at the Department of Biochemistry and Biophysics of Vilnius University who allowed and encouraged me to do this work. Especially I am grateful to prof. dr. Vida Kirvelienė, prof. dr. Dobilas Kirvelis and dr. Saulius Gražulis for the complicated examinations during which I learned a lot.

Finally, special thanks go to my family. This thesis would have not been written without the help of my wife, parents, sister, brother, and, of course, daughters, who always supported me.

Table of Contents

Introduction	1
1 Overview of Literature	6
1.1 Cytochrome P450 and Drug Metabolism	6
1.1.1 Reactions Catalyzed by Cytochrome P450	6
1.1.2 Human Cytochrome P450 Enzymes	9
1.1.3 Inhibition of Cytochrome P450	13
1.2 Experimental Methods for Estimation of Drug Metabolism	15
1.2.1 Metabolite Identification	15
1.2.2 Cytochrome P450 Reaction Phenotyping	16
1.2.3 Cytochrome P450 Inhibition Assays	17
1.3 QSAR Models of CYP3A4 Inhibition	20
1.4 Prediction of Drug Metabolism Regioselectivity	23
1.4.1 Biotransformation Rules	24
1.4.2 Prediction of Regioselectivity by Quantum Chemistry Methods	25
1.4.3 Regioselectivity Models Using Structures of Enzymes	27
1.4.4 Data Mining Models for Prediction of Metabolism Sites	29
2 Data and Modeling Methods	33
2.1 CYP3A4 Inhibition Data	33
2.1.1 Literature Dataset	33
2.1.2 PubChem Dataset	34
2.1.3 Summary of CYP3A4 Inhibition Datasets	35
2.2 Regioselectivity Data	36
2.2.1 Modeling Dataset	36
2.2.2 External Validation Dataset	37
2.3 Descriptors	38
2.3.1 Fragmental Descriptors	38
2.3.2 Atom-centered Fragmental Descriptors	38
2.4 Statistical Methods	39
2.4.1 Global Model	41
2.4.2 Dynamic Similarity	42
2.4.3 Local Model	43

2.4.4	Estimation of Prediction Reliability	44
2.4.5	Training of the GALAS Model	45
2.5	Development and Validation of Models	46
2.5.1	CYP3A4 Inhibition Model	46
2.5.2	Regioselectivity Model	47
2.5.3	Software	49
3	Results and Discussion: CYP3A4 Inhibition Modeling	50
3.1	Global Model	51
3.2	Local Model	58
3.3	Training of the GALAS Model	61
3.3.1	Training with Data from a Similar Assay	62
3.3.2	Training with Data from an Assay with a Different Potency Threshold	64
3.3.3	Training with Compounds from a New Structural Class	65
4	Results and Discussion: Regioselectivity Modeling	68
4.1	Internal Validation of the Model	69
4.2	External Validation of the Model	73
4.3	Comparison to Other Models	79
4.4	Adaptation of the Model to Compounds of Novel Classes	81
4.5	Adaptation of the Model to Cytochrome P450 Phenotyping	83
	Conclusions	86
A	Results of External Validation of Regioselectivity Model	88
	References	103
	List of Publications	121
	Curriculum Vitae	123
	Summary in Lithuanian (Santrauka)	124

Abbreviations

ADME – Absorption, Distribution, Metabolism, Excretion;
ANN – Artificial Neural Networks;
BC – naive Bayesian Classifier;
BFC – 7-benzyloxy-4-trifluormethylcoumarin;
CYP1A2 – cytochrome P450 1A2;
CYP2C9 – cytochrome P450 2C9;
CYP2C19 – cytochrome P450 2C19;
CYP2D6 – cytochrome P450 2D6;
CYP3A4 – cytochrome P450 3A4;
DFT – Density Functional Theory;
DMCI – Data-Model Consistency Index;
GALAS – Global, Adjusted Locally According to Similarity;
HLM – Human Liver Microsomes;
HPLC – High Performance Liquid Chromatography;
HTS – High-Throughput Screening;
 IC_{50} – half maximal Inhibitory Concentration;
IGF-1R – Insulin-like Growth Factor-1 Receptor;
 K_i – inhibition constant;
 K_m – Michaelis constant;
kNN – k-Nearest Neighbors;
LC – Liquid Chromatography;
 LD_{50} – median lethal dose (the dose required to kill half of the members of a tested population);
 $LogP$ – 1-octanol/water partition coefficient;
LR – Logistic Regression;
MLR – Multiple Linear Regression;
MS – Mass Spectrometry;

NCBI – National Center for Biotechnology Information;
NMR – Nuclear Magnetic Resonance;
OECD – Organization for Economic Co-operation and Development;
 p – probability;
PLS – Partial Least Squares/Projection to Latent Structures;
QSAR – Quantitative Structure-Activity Relationship;
 RI – Reliability Index;
ROC – Receiving Operating Characteristic;
RT – Regression Trees;
 SI – Similarity Index;
 $SOME$ – Site of Metabolism Estimator;
 $SPORCalc$ – Substrate Product Occurrence Ratio Calculator;
SVM – Support Vector Machines;
 $SyGMa$ – Systematic Generation of Potential Metabolites.

Introduction

Hepatic metabolism by cytochrome P450 enzymes is the major clearance route for most of the drugs. The biotransformation reactions introduce hydrophilic functional groups into lipophilic molecules, usually resulting in more polar and water-soluble compounds that are readily excreted. Fast metabolism and rapid elimination of drug is undesired in the pharmaceutical industry therefore many efforts are focused on the optimization of metabolic stability of drug candidates [1]. Furthermore toxic metabolites are formed in some cases leading to unwanted adverse effects. Safety testing of metabolites has become a requirement according to the guidances for industry issued by regulatory institutions [2].

Consequently the metabolites of a drug candidate are desired to be known in the earliest stages of drug development. Knowing the regioselectivity of metabolism, i. e. major reaction sites, possible modifications of a compound to increase its metabolic stability can be suggested [1]. On the other hand, understanding the pharmacological and toxicological consequences of metabolism of a drug candidate is critical for pharmaceutical research, thus the biotransformation pathways are always identified in detail at the later stages of development.

Another crucial aspect related to drug metabolism is drug-drug interactions, predominantly caused by the inhibition of xenobiotic metabolizing enzymes. These are among the main problems in the modern drug discovery after several drugs have been withdrawn from market because of drug-drug interactions involving inhibition of CYP3A4. This is the most relevant enzyme, responsible for more than 50% of drug metabolism in human organism [3]. It is a broad specificity oxygenase which is able to metabolize compounds belonging to many diverse drug classes [4].

The early predictions of metabolism related properties, such as possible metabolites or cytochrome P450 inhibition, may alert researchers to safety risks associated both directly with the drug substance itself and with products of its biotransformation. All possible tools should be used to avoid failures, including *in silico* predictions. This approach is very attractive because *in silico* models may be applied in early stages of drug discovery at a very small cost. Predictions are very fast and can be obtained for virtual compounds prior to their synthesis.

Despite the complexity of biological systems, prediction of metabolism from molecular structure is possible [5]. Recently the predictive software has been successfully applied for optimization of metabolic stability of cyclooxygenase-2 inhibitors [6–8]. Experimental identification of metabolites by mass spectrometry is facilitated by using structures of metabolites estimated *in silico* [2, 9, 10]. The prediction of CYP3A4 inhibition allows rapid screening of virtual libraries for possible drug-drug interactions prior to their actual synthesis and enables compound prioritization before the experimental testing.

The (quantitative) structure-activity relationship is the most popular means for modeling ADME properties, including metabolism and cytochrome P450 inhibition [11]. A variety of regression and classification methods are used to express the mathematical relationship between activity and structure of compound, like Partial Least Squares (PLS) or Support Vector Machines (SVM). The chemical structures are encoded by 2D- or 3D-structural, topological, quantum chemical or physicochemical descriptors and the models are trained on the data for diverse compounds acquired from *in vitro* or *in vivo* experiments. One of the main limitations of all predictive structure-activity relationship models is that they are valid only in the part of chemical space closely related to the training set, called model applicability domain. According to OECD principles for (Q)SAR validation, any model that is proposed for regulatory use should be based on a defined experimental endpoint, tested on data that were not used for its development, and after all associated with a defined domain of applicability [12].

Most of the previously published metabolism and cytochrome P450 inhibition models are based on proprietary datasets, which automatically raises several issues regarding their potential application:

- The effective assessment of the applicability domain of such models is not possible as the actual training set structures are not available in public;
- In-house datasets usually consist of specific compounds that a particular institution is working with, rendering such models practically useless for anyone dealing with different compound classes;
- The modeling techniques used in model development do not allow estimation of prediction reliability.

The latter fact is especially important in case of metabolism site prediction as none of the previous regioselectivity models estimates its applicability domain.

A novel GALAS (Global, Adjusted Locally According to Similarity) modeling methodology was recently introduced, which allows forecasting the reliability of each prediction. The successful applications of this method in predicting continuous properties, such as $\text{Log}P$ and acute toxicity in terms of LD_{50} , have been described in detail [13, 14]. GALAS model is a combination of two approaches: a global model for the prediction of the property of interest, and a similarity based local correction model. This methodology not only allows the estimation of reliability of predictions, but also makes it possible to expand the applicability domain of a resulting model in a very straightforward manner, i. e. without time consuming full statistical re-parameterization of the model.

Objectives of the Study

The objectives of the study were to adapt the GALAS modeling method for prediction of properties related to drug metabolism, and to develop and validate a model predicting the regioselectivity of metabolism. The following tasks were set in order to achieve these objectives:

- Develop a probabilistic model for CYP3A4 inhibition as an instance of a drug metabolism related property;
- Develop a metabolism regioselectivity model predicting the probability to be oxidized in human liver microsomes for every atom in the molecule;

- Test the features of GALAS model (corrections according to similarity, estimation of prediction reliability, possibility to adapt the model to novel compounds) in the developed models;
- Validate the predictions of metabolism regioselectivity using new published experimental data;
- Compare the regioselectivity predictions with previously published ligand-based model *SMARTCyp*.

Scientific Novelty

The major aspects of the scientific novelty of the results described in the thesis are as follows.

- The GALAS modeling method was successfully applied for prediction of drug metabolism related properties.
- The developed model predicting sites of metabolism in human liver microsomes is the first QSAR model of regioselectivity that provides a definition of its applicability domain.
- The developed models predicting CYP3A4 inhibition and metabolism regioselectivity can be trained using experimental data for new compounds simply by adding them to the local part of the model without the need of full remodeling.

Practical Value

The reported GALAS models can be used in the pharmaceutical industry. *In silico* predictions may be applied in rapid screening of virtual libraries prior to actual synthesis of compounds. Estimation of CYP3A4 inhibition allows consideration of possible drug-drug interactions in the earliest drug discovery stages.

The good prediction results of regioselectivity model lead to its potential practical application in metabolic stability optimization or metabolite identification. Reliability estimation is provided for each prediction and enables compound prioritization before experimental testing. The trainability feature of a GALAS model enables adjusting it to the needs of any particular drug discovery project.

Statements Presented for Defense

- Structure-activity relationship models predicting CYP3A4 inhibition and regioselectivity of human liver microsomal metabolism were developed using GALAS modeling method.
- The developed models provide estimation of their applicability domain in the form of calculated prediction Reliability Index that effectively identifies correct predictions.
- The applicability domain of the developed GALAS models is easily expanded by adding new compounds to the local part of the model.
- The quality of regioselectivity predictions produced by the developed model is comparable to the results of recently published *SMARTCyp* software, based on quantum chemistry calculations.

Chapter 1

Overview of Literature

1.1 Cytochrome P450 and Drug Metabolism

The most significant part of biotransformation of drugs and other xenobiotics in human organism occurs in the liver. It is divided into two phases. Oxidation-reduction and hydrolysis reactions constitute Phase I. These reactions introduce hydrophilic functional groups into lipophilic xenobiotic molecules. In Phase II either the parent compounds or their oxidized metabolites are conjugated to glucuronic acid, sulfate, glutathione or amino acids. The conjugation reactions are usually very fast and thereby the Phase I is the rate limiting step of the drug metabolism. This section describes the enzymes of cytochrome P450 superfamily, which are the most important catalysts of the oxidative Phase I reactions, responsible for more than 2/3 of total drug metabolism [3].

1.1.1 Reactions Catalyzed by Cytochrome P450

The cytochrome P450 enzymes are hemoproteins. In mammalian organisms they are bound to the membrane of endoplasmic reticulum. The main reaction catalyzed by these enzymes is the incorporation of one oxygen atom from O₂

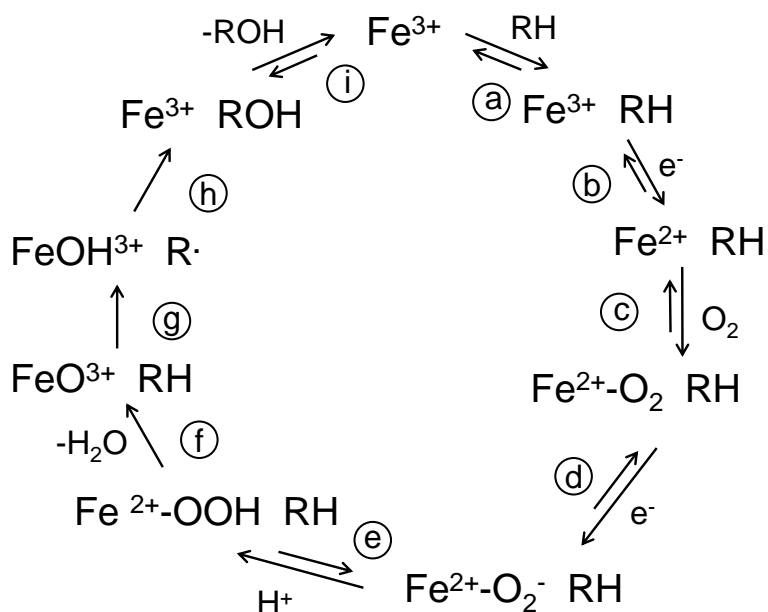


Figure 1.1: The reaction cycle of cytochrome P450 enzymes.

molecule into the molecule of substrate:



The chemical mechanism of a typical cytochrome P450 reaction is well known [15, 16] and is shown in Fig. 1.1. After binding of substrate (a), the iron in the active site of enzyme is reduced from ferric to ferrous state by flavoprotein NADPH-cytochrome P450 reductase (b). Then the enzyme binds an oxygen molecule (c), and a second electron from reductase (or in some cases from cytochrome b₅) further reduces the system (d). The O–O bond is cleaved after addition of a proton (e), generating water (H₂O) and an intermediate Compound I which is formally represented as FeO³⁺ or Fe^V=O (f). The exact electronic structure of this intermediate is not fully characterized because it cannot be trapped due to its very short lifetime [17].

The Compound I reacts with the substrate molecule. In case of aliphatic carbon hydroxylation, the hydrogen atom abstraction leads to formation of a radical that reacts with the FeOH³⁺ intermediate to form hydroxylated metabolite (Fig. 1.1, g and h; Fig. 1.2, a). If this oxidation occurs next to heteroatom, an unstable

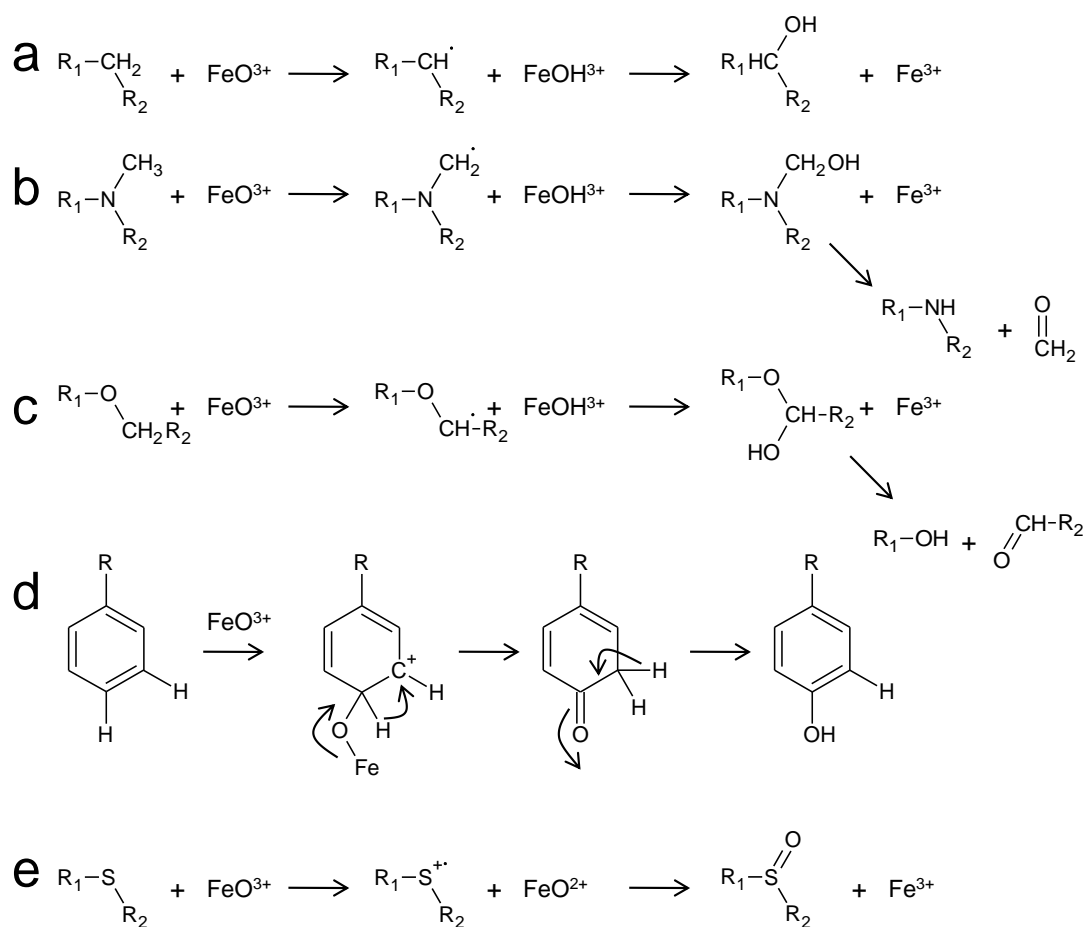


Figure 1.2: The mechanisms of reactions catalyzed by cytochrome P450 enzymes: a – aliphatic hydroxylation, b – N-dealkylation, c – O-dealkylation, d – aromatic hydroxylation, e – S-oxidation.

compound is formed and the carbon-heteroatom bond may split resulting in dealkylation (Fig. 1.2, b and c). Aromatic carbon is oxidized by a different mechanism (Fig. 1.2, d). In this case σ -complex with the substrate is formed, resulting in migration of hydrogen atom, called “NIH shift”. The heteroatoms (N, S) are oxidized by the electron abstraction mechanism (Fig. 1.2, e). After the formation of metabolite, the last step of the cytochrome P450 reaction cycle is the dissociation of the product from the active site of the enzyme (Fig. 1.1, i).

The above described reaction mechanism is a simplification. A number of alternative reactions can proceed, resulting in formation of reactive oxygen species, inhibition of enzyme because of covalent binding of substrate to the heme or protein, rearrangements in substrate molecule, etc. [16]. The outcome of

these processes is hardly predictable. Expansion or contraction of rings, various isomerisations and eliminations lead to production of diverse metabolites [18, 19]. Further investigation of these uncommon reactions should provide a deeper understanding of oxidation mechanisms and possible new insights into toxicity of chemicals.

1.1.2 Human Cytochrome P450 Enzymes

The major cytochrome P450 isoforms that mediate biotransformations of xenobiotics in human body are CYP3A4, CYP2D6, CYP2C9, CYP2C19, and CYP1A2 [3]. CYP3A4 is a broad specificity oxygenase responsible for a half for drug metabolism. The typical substrates of this enzyme are large hydrophobic compounds. They belong to multiple drug classes and have molar mass ranging from 151 Da (paracetamol) to > 1,200 Da (cyclosporins). CYP2D6 is a polymorphic oxygenase, its activity varies in different individuals up to > 1,000 times leading to adverse effects of drugs in poor metabolizers. Its substrates are basic aromatic compounds (opioids, anti-arrhythmics, antidepressants, β -blockers, drugs of abuse, etc.). On the contrary CYP2C9 oxidizes mostly acidic molecules like non-steroidal anti-inflammatory drugs. The specificity of CYP2C19 is similar to that of CYP2C9 except the fact that this enzyme prefers unionized drugs. Substrates of CYP1A2 are planar aromatic molecules. In addition to biotransformations of several drugs, this enzyme also activates promutagens and procarcinogens, such as polycyclic aromatic hydrocarbons and aromatic amines [4].

The determinants of CYP2D6 specificity are probably the best understood among human drug metabolizing enzymes. The substrates of this enzyme usually contain an aromatic ring in the distance of 5-10 Å from basic amino group [20]. The site-directed mutagenesis experiments show that the amino group from substrate molecule interacts in the active site with Asp301 and Glu216 residues [21]. Changing these amino acids into non-acidic residues altered the specificity of enzyme, and it oxidized non-basic compounds which are not substrates for wild-type CYP2D6. The X-ray crystallography experiments confirmed these data (Fig. 1.3) [22].

The aromatic amino acids that are known to be important for CYP2D6 specificity

are Phe120, Phe481, and Phe483. Phe120 is responsible for ligand orientation in the active site. After substituting it to non-aromatic residues the regioselectivity of metabolism changed for several substrates and the affinity for typical CYP2D6 inhibitor quinidine decreased [23, 24]. Phe481 and Phe483 are more distant from heme in the crystal structure (Fig. 1.3) but can be active in substrate recognition or binding in active site [22]. The mutagenesis of these residues changed kinetics of reactions, regioselectivity of metabolism, and even substrate specificity [25–27].

Despite of a relatively clear picture of substrate binding to CYP2D6, some recent findings show that the understanding of the specificity of this enzyme is not complete. A new compound pactimibe and its analogues were found to be clinically significant CYP2D6 substrates [28]. All these molecules contain an acidic group and are negatively charged at physiological pH. Docking experiments revealed that the carboxy group of pactimibe interacts with basic Arg221 residue in the active site of the enzyme [29]. Unfortunately these results are not confirmed by X-ray diffraction or mutagenesis experiments. The significance of acidic CYP2D6 ligands is unknown because only a few non-basic substrates of this enzyme are known up to date.

The knowledge on the determinants of ligand binding to other human drug metabolizing enzymes is even less exhaustive. Only a few amino acids that are responsible for substrate binding and catalysis are determined in CYP1A2. The role of Phe226 in interaction with planar aromatic ligands is clear [30, 31] and was later confirmed in crystal structure of CYP1A2 with inhibitor naphthoflavone [32]. The exact influence of other residues is still not understood.

Some reactions catalyzed by human cytochrome P450 enzymes do not obey classical Michaelis-Menten kinetics [33]. In fact, such “atypical kinetics” are characteristic for many substrates of CYP3A4 or CYP2C9. The analysis of metabolism of different chemicals by these enzymes are consistent with theoretical kinetic schemes involving binding of substrates into several binding sites [34]. A cytochrome P450 enzyme has only one active site that is in some cases large enough to bind two or even more ligand molecules. These interact with different amino acid residues of the protein, resulting in unusual kinetic curves representing activation, autoactivation, partial inhibition or substrate inhibition.

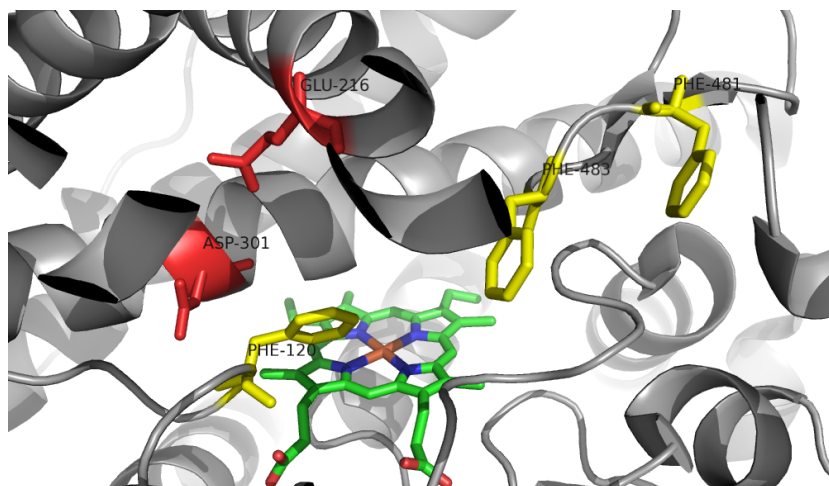


Figure 1.3: The structure of CYP2D6 active site (PDB ID 2F9Q). Acidic residues Glu216 and Asp301 are red, aromatic amino acids Phe120, Phe481 and Phe483 are yellow.

These results of kinetic studies are in agreement with later crystallographic and site-directed mutagenesis experiments. The structure of CYP2C9 was determined with substrates warfarin [35] and flurbiprofen [36]. Flurbiprofen forms ion bridge with Arg108 (Fig. 1.4, a). Warfarin binds in the active site by π - π interactions with Phe114 and Phe476 (Fig. 1.4, b). The same residues were found to be important by site-directed mutagenesis [37–39]. These experiments show that interactions in the active site of CYP2C9 depend on ligand.

The crystallographic analysis of CYP3A4 demonstrated that the active site of this enzyme is large enough to bind bulky substrates or several ligand molecules together. It contains hydrophobic mostly amino acids [40–42]. Fig. 1.5 shows two molecules of ketoconazole bound to CYP3A4. The hydrophobic cluster consisting of Phe213, Phe215, Phe219, Phe220, Phe241, and Phe304 breaks down after ligand binding, and the amino acid residues that lacked secondary structure in ligand-free protein now form a helix-like structure. Imidazole ring from ketoconazole molecule coordinates with heme, and hydrophobic interactions and multiple hydrogen bonds are also observed. However, the binding of erythromycin could not be explained by X-ray data [42]. Another CYP3A4 inhibitor ritonavir also forms complex with heme iron and induces large conformational changes in protein structure [43].

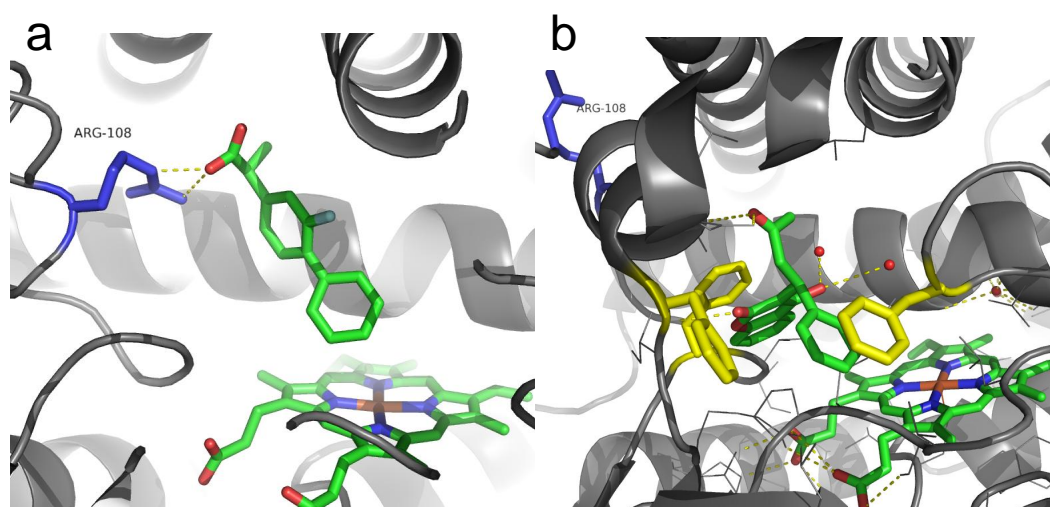


Figure 1.4: Different binding of substrates in the active site of CYP2C9: a – flurbiprofen forms salt bridge with Arg108 (PDB ID 1R9O); b – warfarin interacts with phenylalanine residues, and Arg108 is not in the active site of the enzyme (PDB ID 1OG5).

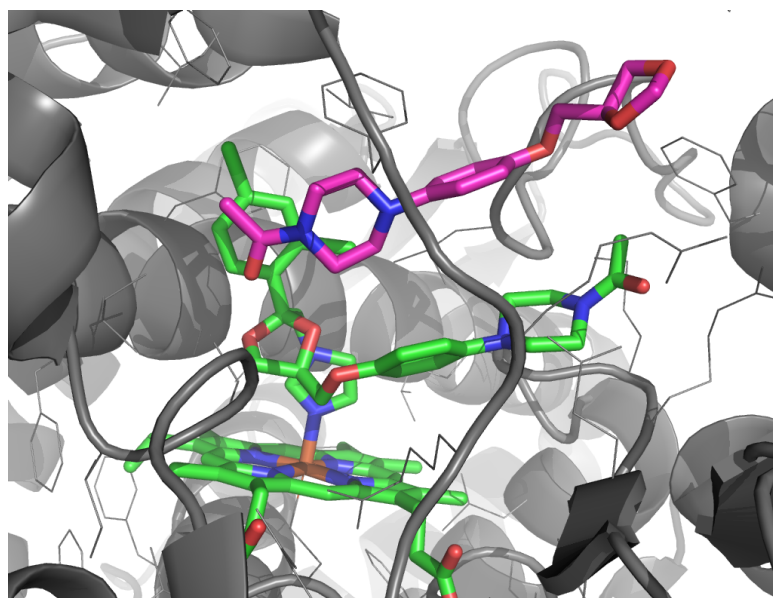


Figure 1.5: Two molecules of ketoconazole bound to CYP3A4 (PDB ID 2J0C).

Many amino acids that are responsible for ligand binding were found by site-directed mutagenesis of CYP3A4. For example, changing Ser119, Leu210, Leu211, Phe213, Asp214, Ile301, Phe304, Ala305, Ile369, Ala370 affect regioselectivity and kinetics of steroid metabolism [44–49]. Some of these residues have impact on midazolam oxidation (Ser119, Phe304) together with Ile120, Tyr307 and Thr309 [50]. Another benzodiazepine drug diazepam also interacts with these amino acids [51]. On the contrary, mutating Ser119 and Thr309 did not influence raloxifene biotransformation, but Phe215 appeared to be important for positioning this substrate in the active site of enzyme [52]. Taken together, the overall data lead to conclusion that a single mechanism of ligand binding to CYP3A4 does not exist, and different models are suitable for different ligands. These features explain the extreme chemical diversity of CYP3A4 substrates and inhibitors and support the general opinion on the significance of hydrophobic interactions for binding to this enzyme.

1.1.3 Inhibition of Cytochrome P450

The inhibitors of cytochrome P450 enzymes are widely used as therapeutic agents, insecticides or herbicides. On the other hand, inhibition of drug metabolizing cytochrome P450 enzymes (especially CYP3A4) can lead to undesired accumulation of their substrates in the organism potentially resulting in toxic side effects. Up to date a number of drugs (mibefradil, terfenadine, astemizole) has been withdrawn from the market because of drug-drug interactions [3]. As a result, testing novel compounds for human cytochrome P450 inhibition has become a common practice in pharmaceutical industry [1].

There are several notable aspects of cytochrome P450 inhibition mechanisms [53]. In case of reversible inhibition of the enzyme, two types of interaction between the inhibitor and enzyme are observed. Competitive inhibitors form hydrogen bonds with the side chains of active site amino acids or bind by hydrophobic interactions. Such compounds usually are tight binders but poor substrates. In addition, complex between the heme iron and the inhibitor can be formed. Pyridine and imidazole derivatives are commonly used cytochrome P450 inhibitors. The strongest interaction is observed for compounds that bind both to the side chains of amino acids and prosthetic heme iron (Fig.1.6).

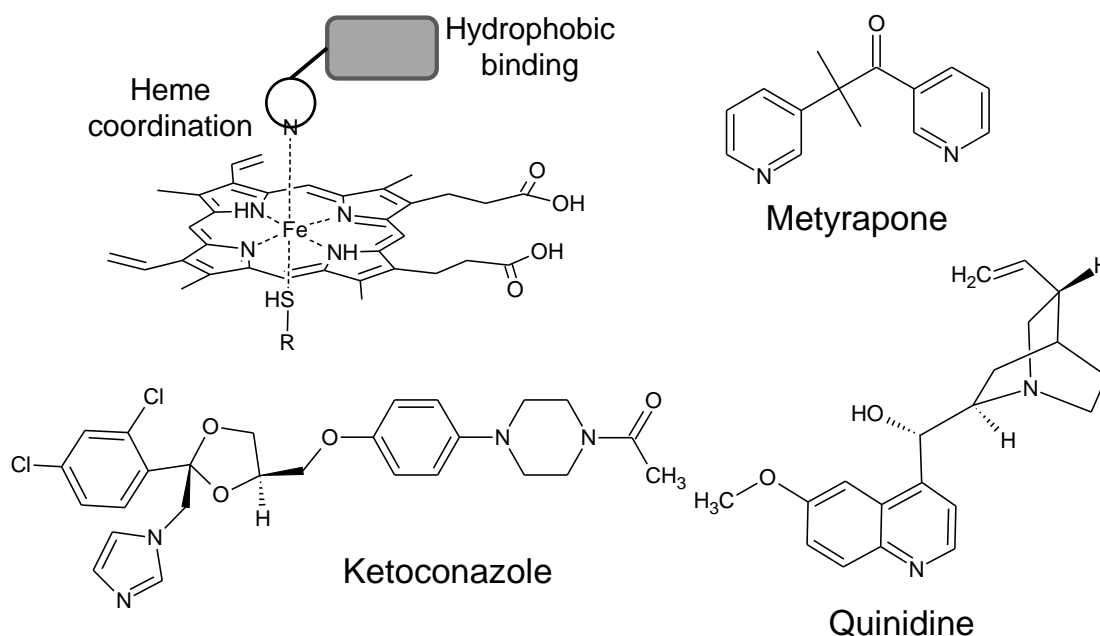


Figure 1.6: Reversible inhibition of cytochrome P450 enzymes. General mechanism of inhibition and examples of inhibitors: metyrapone (non-specific cytochrome P450 inhibitor), ketoconazole (fungal Lanosterol 14- α -demethylase and human CYP3A4 inhibitor), quinidine (human CYP2D6 inhibitor).

In addition to reversible inhibition, some inhibitors act as mechanism-based (or catalysis-dependent) cytochrome P450 inactivators. After binding to the active site, they undergo oxidation as substrates of these enzymes. An intermediate compound then reacts with the protein and can bind covalently to amino acids or heme, or form quasi-irreversible complex with heme iron. Mechanism-based cytochrome P450 inhibitors are nitrogen, sulfur and halogen containing compounds, terminal alkenes and alkynes, methylenedioxybenzenes [53]. Fig. 1.7 shows how methylenedioxybenzenes form quasi-irreversible complex with heme iron after oxidation.

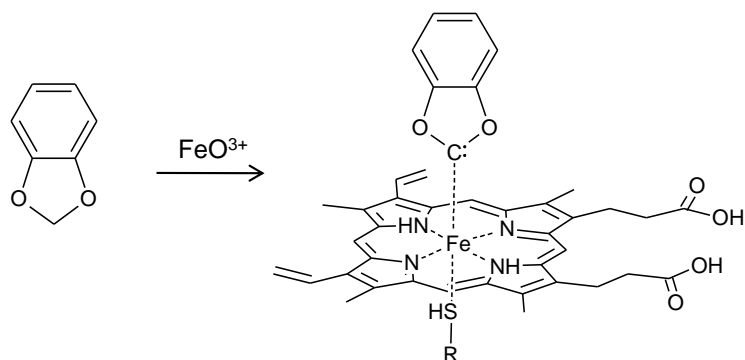


Figure 1.7: Mechanism of cytochrome P450 inhibition by compounds having methylenedioxybenzene substructure, according to [53].

1.2 Experimental Methods for Estimation of Drug Metabolism

The investigation of metabolism related properties of a drug candidate consists of the elucidation of the structures of its metabolites, identification of the enzymes that are responsible for formation of major biotransformation products, and determination of the possibility to inhibit main human xenobiotic metabolizing enzymes.

1.2.1 Metabolite Identification

The modern identification of metabolites is done using liquid chromatography coupled with mass spectrometry (LC/MS). The metabolites of a drug candidate are obtained by incubation *in vitro* with hepatocytes, microsomes or purified enzymes in early drug discovery. Later samples from blood, urine and feces are analyzed to investigate metabolism *in vivo* and identify circulating and excreted metabolites. A great variety of mass spectrometry methods and instruments have been developed and are utilized in elucidation of structures of metabolites [54, 55]. Mass difference between parent compound and metabolite is known for common biotransformation reactions (e. g., +16 for hydroxylation, -14 for demethylation). A number of software packages exist to ease analysis of spectra by incorporating the mass shifts most frequently observed in drug metabolism

[9, 54]. Still, simple MS methods cannot detect all metabolites. Additional data obtained by tandem mass spectrometry are then used to identify metabolism sites for less common biotransformations. The spectra of product ions are obtained in these multi-stage MS experiments and help in identifying unusual metabolites [54].

Common practice is to label drug candidate molecules by radioactive isotopes (^{14}C , ^3H) for metabolism studies. Besides alternative techniques have been developed for analysis of mass spectra containing peaks of impurities from complex biological samples. These methods are applied after data acquisition. Multiple reaction monitoring utilizes prior knowledge on structures of metabolites. For example, initially metabolites are characterized *in vitro* using liver microsomes, and the acquired data are employed in analysis of blood samples [56]. Mass defect filter uses accurate mass data to identify small mass defects that are within 50 mDa range for typical biotransformations [57, 58]. These techniques refine the high resolution MS data and automatically remove ions of molecules that interfere with drug metabolite ions.

Many software tools aid the interpretation of MS data in metabolite identification. In addition to programs provided by instrumentation vendors, the possibility to use *in silico* predicted metabolites has been proposed [9]. Recently the analysis of mass spectra has been successfully automated by combining it to site of metabolism estimations by *MetaSite* software [10]. This approach facilitates the work of biotransformation scientist and accelerates the process of metabolite identification.

If the exact elucidation of metabolite structure is not possible by MS only, more labor intensive methods are used. Purified metabolites are analyzed by NMR which can determine exact sites of hydroxylation. Chemical modifications are utilized in rare cases when unstable or hardly separable metabolites are produced [54].

1.2.2 Cytochrome P450 Reaction Phenotyping

Knowing the *in vitro* metabolic profile of a drug candidate in human liver microsomes, most projects later focus on phenotyping major cytochrome P450

enzymes. The enzymes that catalyze the formation of metabolites are determined using several methods [59]. The best approach for identification of the most significant cytochrome P450 isoforms is inhibition of microsomal metabolism using monoclonal antibodies. These inhibit individual enzymes and can therefore unambiguously confirm their contribution to the particular biotransformation reaction. Nevertheless the commercial availability of inhibitory antibodies is low.

Inhibition by specific chemical inhibitors is used as an alternative reaction phenotyping method. These compounds may inhibit several human cytochrome P450 isoforms, but the inhibition constants are different. As a result researchers obtain highly selective inhibition of enzymes by choosing an appropriate concentration of the inhibitor. Ketoconazole is such non-specific inhibitor. Its affinity for CYP3A4 is more than 10 times higher than for other human hepatic enzymes, therefore at low concentrations it selectively inhibits this isoform [60]. The names and structures of compounds frequently used for human cytochrome P450 phenotyping are listed in Table 1.1.

The specific inhibition experiments are complemented by incubation of compound of interest with recombinant cytochrome P450 enzymes. Interpretation of the results of such testing is straightforward only if a single enzyme catalyzes the biotransformation. If the metabolism by several enzymes is observed, the studies with inhibitory antibodies or selective chemical inhibitors are obligatory to estimate the contribution of each enzyme in microsomal incubations. Historically the correlation of the rate of drug biotransformation with cytochrome P450 marker reactions is also used. This method is the least reliable. It may produce numerous errors and should be utilized only for confirmation of the results of other experiments [59].

1.2.3 Cytochrome P450 Inhibition Assays

Many *in vitro* methods have been developed and are used today to screen large libraries of synthesized compounds for cytochrome P450 inhibition [62]. The conventional methods are based on inhibition of standard probe reactions measured in human liver microsomes or with recombinant cytochrome P450

Table 1.1: Selective inhibitors of human cytochrome P450 enzymes, according to [61].

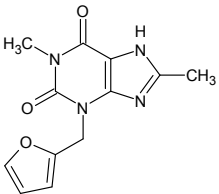
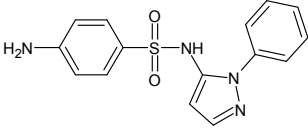
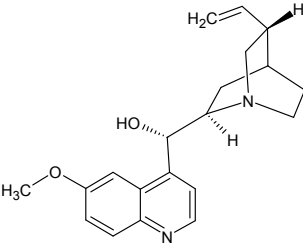
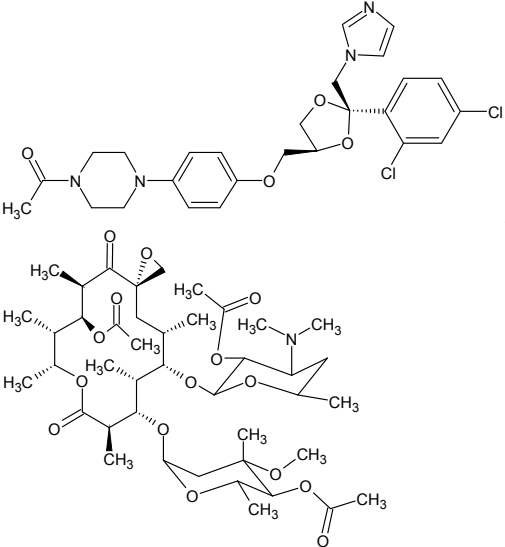
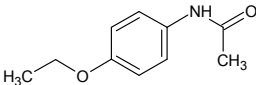
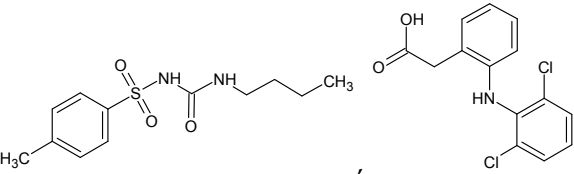
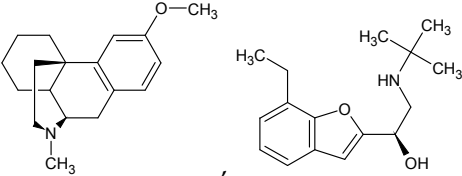
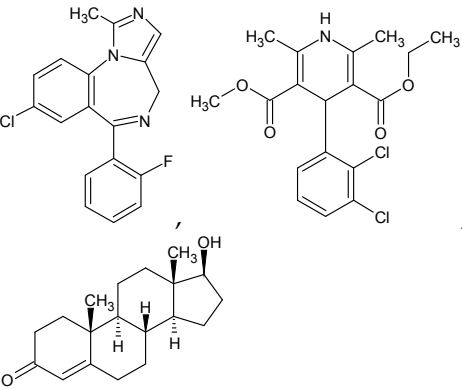
Enzyme	Inhibitor	Structure
CYP1A2	furafylline	
CYP2C9	sulfaphenazole	
CYP2D6	quinidine	
CYP3A4	ketoconazole, troleandomycin	

Table 1.2: Examples of standard substrates used in human cytochrome P450 inhibition studies [61, 63].

Enzyme	Inhibitor	Structure
CYP1A2	phenacetin	
CYP2C9	tolbutamide, diclofenac	
CYP2D6	dextromethorphan, bufuralol	
CYP3A4	midazolam, felodipine, testosterone	

enzymes [1, 63]. The specific substrates of human cytochrome P450 enzymes are shown in Table 1.2. Some of these compounds are metabolized by several microsomal enzymes, but different metabolites are formed. Diclofenac is hydroxylated in different positions by CYP2C9 and CYP3A4, dextromethorphan is O-dealkylated in different positions by CYP2C9 and CYP3A4, dextromethorphan is O-dealkylated by CYP2D6 and N-dealkylated by CYP3A4. Metabolites specific for the enzyme of interest are tracked in these cases, i. e. 4'-hydroxydiclofenac for CYP2C9 and dextropran (O-demethylated dextromethorphan) for CYP2D6. The reactions are performed with radiolabeled substrates or detection of metabolites by LC/MS is used.

Later high-throughput screening (HTS) assays based on fluorescent compounds

were developed and made the analysis faster and cheaper [64]. Substrates that produce fluorescent metabolites are incubated with recombinant cytochrome P450 enzymes, as these substrates are not enough selective for experimentation in human liver microsome [65]. Luminogenic assays were developed as an alternative for fluorescence experiments [66]. Analogues of luciferin are converted to luciferin by cytochrome P450 enzymes. The luciferase decarboxylates luciferin using ATP, generating an amount of light which is proportional to the amount of luciferine formed.

Cytochrome P450 inhibitors diminish the fluorescence in fluorometric experiments and light output in luminogenic assays. Both methods can be automated, and the enzymes and substrates for the HTS experiments are commercially available [62]. The main disadvantage of fluorometric assays is that they cannot be used for analysis of drug candidates that are also fluorescent. Luminogenic methods are more sensitive but may produce false positive results for compound that do not inhibit cytochrome P450 but interfere with luciferase activity.

Due to these facts, in later drug discovery stages the results of HTS have to be approved by an experiment with standard cytochrome P450 probes [1]. Cocktail assays are sometimes used to increase throughput of these methods. Specific substrates of several isoforms are incubated with human liver microsomes or a mixture of recombinant cytochrome P450 enzymes and inhibition potential is determined for all enzymes in one experiment. Finally, only detailed results from conventional experiments with determination of inhibition mechanism and kinetics can be reported to regulatory institutions [62].

1.3 QSAR Models of CYP3A4 Inhibition

The data acquired from *in vitro* studies of CYP3A4 inhibition have been used to develop *in silico* structure-activity relationship models [67–76], which can serve as virtual screening tools in evaluating the possibility for new compounds to cause drug-drug interactions. The results of these works are briefly summarized in Table 1.3. A great variety of descriptors and statistical methods were used for prediction of CYP3A4 inhibition representing the tools commonly used for QSAR modeling.

Table 1.3: Summary of classification structure-activity relationship models of CYP3A4 inhibition based on large data sets of diverse compounds

Model and data	Modeling details ^a	Results ^b
Zuegge et al., 2003 [67] Inhibitors: $IC_{50} < 1 \mu M$ (421, 36.5%) Non-inhibitors: $IC_{50} > 50 \mu M$ (160, 14.5%)	Methods: PLS, ANN Descriptors: fragmental, topological, physicochemical	Accuracy: 90% Sensitivity: 93% Specificity: 86%
Kriegel et al., 2005 [69, 70] ^c Strong inhibitors: $IC_{50} < 2 \mu M$ (243, 18%); Medium inhibitors: 2 $\mu M < IC_{50} < 20 \mu M$ (561, 42%) Non-inhibitors: $IC_{50} > 20 \mu M$ (541, 40%)	Methods: PLS, SVM Descriptors: physicochemical, topological, quantum chemical, 3D structural	Accuracy: 70% (three class model) Sensitivity: strong: 68% medium: 63%
Arimoto et al., 2005 [71] ^d Inhibitors: $IC_{50} < 3 \mu M$ (1578, 35%) Non-inhibitors: $IC_{50} > 3 \mu M$ (2892, 65%)	Methods: RT, kNN, BC, LR, SVM Descriptors: topological, fragmental	Accuracy: 83% Sensitivity: 82% Specificity: 81%
Jensen et al., 2007 [73] ^e Inhibitors: $IC_{50} < 20 \mu M$ (361, 26%) Non-inhibitors: $IC_{50} > 20 \mu M$ (1021, 74%)	Method: kNN Descriptors: fragmental	Accuracy: 88% Sensitivity: 65% Specificity: 94% Not classified: 14%
Gleeson et al., 2007 [74] ^d Inhibitors: $IC_{50} < 6.3 \mu M$ (145, 19.8%) Non-inhibitors: $IC_{50} > 15.8 \mu M$ (420, 57.2%)	Methods: PLS, RT Descriptors: fragmental, physicochemical	Accuracy: 89% Sensitivity: 67% Specificity: 96%
Choi et al., 2008 [75] Inhibitors: not defined (394, 42.5%) Non-inhibitors: not defined (533, 57.5%)	Method: RT Descriptors: topological, physicochemical, 3D structural	Accuracy: 73% Sensitivity: 83% Specificity: 54%

^a Abbreviations of statistical methods are given in Abbreviations list.

^b Test set classification results for the best model is reported in case of several models.

^c Inhibition of erythromycin metabolism by recombinant CYP3A4.

^d Inhibition of 7-benzyloxy-4-trifluoromethylcoumarin (BFC) metabolism by recombinant CYP3A4.

^e Inhibition of erythromycin metabolism in human liver microsomes.

All computational models have their limitations. For example, widely accepted 3D-QSAR models analyze spatial ligand-enzyme interactions assuming that binding mode for all compounds is the same. The predictive power of such CYP3A4 inhibition models is limited [77], because in reality a great variety of ligand binding modes to CYP3A4 exists as well as a large conformational degree of freedom in the active site is possible [42, 78–80]. Descriptors, based on 3D structure of chemicals have been rarely used for CYP3A4 specificity modeling (Table 1.3). Physicochemical, topological and 2D structural descriptors proved to be more suitable for this enzyme.

Despite the relative successes in CYP3A4 inhibition modeling a need of new computational approaches for identification of CYP3A4 inhibitors still exists. Most of published CYP3A4 inhibition models were developed in pharmaceutical companies and are based on in-house data. In such cases the molecules of the training set belong to specific chemical classes that are investigated in a particular company, and consequently wide application of the models is limited. A new QSAR model trained on publically available diverse data is needed.

Furthermore, not all previously developed models are associated to a measure of their applicability domain, which is a standard requirement for QSAR models that can be used for regulatory purposes [12]. The importance of this domain has been also shown for CYP3A4 inhibition model recently [76]. The methods for estimation of model applicability also should be improved. Similarity measures are widely utilized to evaluate the reliability of prediction. In addition to them the novel GALAS modeling method takes into account the consistency of experimental data with regard to the predictions of the global model. The necessity to include a measure for consistency of experimental data has been described in recent publication covering the topic of acute toxicity modeling [14]. In case of CYP3A4 inhibition, compound classes also exist where some representatives are potent inhibitors while others do not inhibit this enzyme at all despite being very similar. Such situation was observed for 1,4-dihydropyridine calcium channel antagonists [81].

1.4 Prediction of Drug Metabolism Regioselectivity

Attempts of drug metabolism regioselectivity prediction have been reviewed recently [5]. In this very informative article the models estimating potential metabolism sites are classified to substrate orientation-based, mechanism-based and empirical predictions. Predictions of substrate orientations either align three-dimensional structures of many substrates or analyze the interactions between substrate and enzyme. Mechanism-based predictions estimate the most reactive sites in the organic molecules by quantum chemistry methods. After analysis of large biotransformation databases models utilizing the empirical knowledge can be made, either rule-based expert systems to predict metabolites or statistical models predicting the most likely metabolism sites.

Another possible way for classification of drug metabolism regioselectivity predicting models is the using of ligand- or protein-based methods. Protein-based models investigate interactions between enzyme and its substrates while ligand-based ones use only the structures of ligands. Most of the predictions are made integrating various methods, e. g. docking of substrates to the structures of enzymes is often combined to reactivity predictions.

Historically the first attempts of metabolism prediction were ligand-based as the structures of drug metabolizing enzymes were not known. Analysis of databases containing information on drug metabolism reactions lead to derivation of biotransformation rules [82–86]. Another important ligand-based method is quantum chemistry calculations which predict the reactivity of compounds in cytochrome P450 catalyzed reactions [87–90]. Later the increasing amount of available experimental data lead to creation of several statistical methods for mining these data [91–97]. Protein-based modeling techniques for prediction of possible metabolism sites became popular when the structures of drug metabolizing enzymes were determined [98–102]. Further in this section we are going to review these and other significant models of drug metabolism regioselectivity in detail.

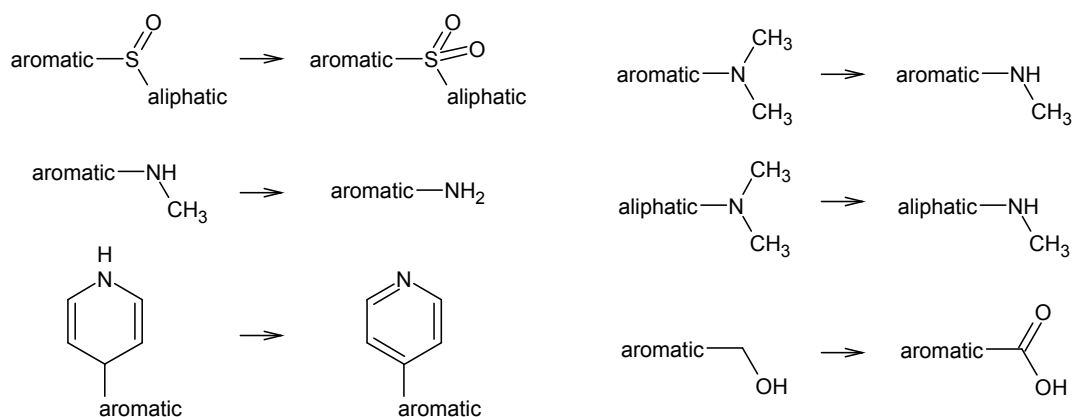


Figure 1.8: Examples of biotransformation rules used in the program *SyGMA* [95].

1.4.1 Biotransformation Rules

The earliest regioselectivity models predicted the metabolites using a set of biotransformation rules derived from experimental data. The substitutions of fragments of substrate molecule following these rules result in generation of structure of metabolite. Some examples of biotransformation rules are provided in Fig. 1.8. Different numbers of rules were derived by different authors, ranging from 70 in earliest models to > 300 in the latest [82–86].

The major drawback of rule-based regioselectivity models is that too many metabolites are generated. Most of models use some prioritization method to distinguish major metabolites from insignificant ones. In the oldest models expert ratings were used, which later evolved into more sophisticated reasoning systems [85, 86]. Unfortunately, some of these are too ambiguous: it is hardly possible for a user to rank the metabolism sites using categories “probable” and “plausible”, and many false positive predictions are provided after all [103]. The newest rule-based model learns priorities of different metabolism sites from a large database of experimental observations, and can be viewed also as data mining method [95].

1.4.2 Prediction of Regioselectivity by Quantum Chemistry Methods

It can be assumed that the rate of a biotransformation reaction is limited by the reactivity of the substrate and not by its binding to the enzyme, because many diverse compounds are oxidized by cytochrome P450 [4]. Under such circumstances it might be possible to predict the sites of metabolism by estimating the reactivity of compound in every possible reaction by quantum chemistry methods. The activation energy can be calculated for hydrogen atom abstraction in case of aliphatic hydroxylation and for formation of a bond between oxygen and carbon atom in case of aromatic hydroxylation.

Semiempirical AM1 calculations were used in the first quantum chemical models. The hydrogen atom abstraction energies were calculated for sets of small molecules, and their reactions with radicals mimicking the Compound I ($\text{Fe}^{\text{V}}=\text{O}$, FeO^{3+}) in the active site of cytochrome P450 were modeled [87]. The best predictions were obtained using *p*-nitrosoxyphenyl radical. Later the same authors extended their model to incorporate aromatic hydroxylation reactions using methoxy radical as model of Compound I [88]. Calculated energies were the calibrated according to experimental data to obtain combined model which predicted both aliphatic and aromatic hydroxylation, but this model was not validated against an external test set.

Later the reactivity of drug-like molecules in cytochrome P450 catalyzed reactions was predicted by *ab initio* calculations at the density functional theory (DFT) levels. These methods are the best of available theoretical tools to obtain reliable reactivity models [104]. $\text{Fe}(\text{porphine})(\text{SCH}_3)\text{O}$ was used as a more realistic model of Compound I. The results of calculations are briefly summarized in Table 1.4. For hydrogen atom abstraction, lowest activation energies were calculated for amines (28-32 kJ/mol), corresponding to N-dealkylation reaction. In case of ethers and atoms connected to sp^2 -hybridized carbon the energies were higher (45-55 kJ/mol and 48-56 kJ/mol, respectively), and substrates with only sp^3 -hybridized carbon atoms had the highest energy barrier (60-62 kJ/mol for secondary and tertiary atoms, and > 72 kJ/mol for methyl group) [89]. Activation energies for aromatic hydroxylation were 60-87 kJ/mol, indicating that aliphatic hydroxylation is more energetically favorable reaction.

Table 1.4: Activation energies for cytochrome P450 catalyzed reactions calculated by *ab initio* DFT methods [89, 90].

Atom type	Reaction	Calculated energy
Aliphatic carbon in amine	N-dealkylation	28-32 kJ/mol
Aliphatic carbon in ether	O-dealkylation	45-55 kJ/mol
Vinylic or benzylic carbon (next to sp ² -hybridized carbon)	Aliphatic hydroxylation	48-56 kJ/mol
Secondary or tertiary sp ³ -hybridized carbon	Aliphatic hydroxylation	60-62 kJ/mol
Primary sp ³ -hybridized carbon (methyl group)	Aliphatic hydroxylation	> 72 kJ/mol
Aromatic carbon	Aromatic hydroxylation	60-87 kJ/mol

In case of some substituents, preference for *ortho*- and *para*-oxidation was clearly shown [90, 105].

Despite high accuracy, the main disadvantage of the *ab initio* quantum chemistry methods is the intense computational power needed. A calculation for one typical drug-like compound can last for several weeks [89, 90]. Therefore these methods are more frequently viewed as tools to investigate the mechanisms of cytochrome P450 catalyzed reactions rather than means to predict the regioselectivity of metabolism for drug-like compounds [104, 105]. Semiempirical methods are usually chosen to reduce time required for calculation. Interestingly, the results of semiempirical calculations correlated well with the DFT results in case of aliphatic hydroxylation [89], but only low correlation was observed for aromatic hydroxylation [90]. Other studies later confirmed these results showing that semiempirical calculations accurately predict DFT energies only for alkanes [106]. When the activation energies were predicted using various descriptors, the results also were not enough accurate [89, 90, 105]. Therefore it was concluded that the simplified methods may give unreliable estimations.

As a consequence of the above described experiments, an interesting method of utilizing *ab initio* DFT methods to predict metabolism sites was developed in the program *SMARTCyp* [107, 108]. The activation energies are calculated for small molecular fragments representing most popular metabolism sites, like aliphatic carbon next to nitrogen or oxygen, vinylic or benzylic carbon, aromatic atoms in various rings, etc. When predicting for a new molecule, the software only looks up the energy value for every atom in the precalculated database. This solves the problem of long lasting computation, and the regioselectivity

predictions are even superior in comparison to semiempirical calculations. The reactivity predictions can be later corrected according to the accessibility of the atom, either by docking substrate into the structure of enzyme [107] or by simply calculating solvent accessible surface area [108]. Most metabolism sites of CYP1A2 [107] and CYP3A4 [108] catalyzed metabolism could be predicted by this regioselectivity model.

1.4.3 Regioselectivity Models Using Structures of Enzymes

Only the predictions of reactivity for every atom in the molecule is sometimes not sufficient for prediction of likely metabolism sites. Steric effects may make the reactive site not accessible to the heme-oxygen complex. Therefore most of the above mentioned reactivity based predictions are later combined with accessibility evaluation. Solvent-accessible surface area is calculated in ligand-based models. Such topological description is sufficient for broad specificity enzymes, like CYP3A4. More accurate methods are needed for more specific cytochrome P450 isoforms when the most likely metabolism site is determined by binding of the substrate to the active site of enzyme. As a result, models of reactivity are often combined to protein-based methods. In such cases docking a substrate molecule into a known structure of enzyme is a possible solution.

As it was already described above, the patterns of substrate interaction with CYP3A4 are not clear. The substrates can bind in several places of the active site, and the protein structure undergoes considerable changes upon ligand binding. Under these circumstances it is not surprising that the ability to predict of CYP3A4 metabolism sites by docking is lower than by other methods [99, 100]. A more successful prediction of regioselectivity of CYP3A4 catalyzed metabolism involved a specific docking methods with dominant reactivity component [101].

Many published protein-based models predict sites of metabolism by CYP2D6 [109–112], CYP2C9 [100, 102, 113] or CYP1A2 [114, 115]. These enzymes have a more defined specificity compared to CYP3A4 and are therefore more suitable for docking. In case of CYP2D6, docking studies mostly aim not to predict the sites of metabolism but to understand the determinants of the regioselectivity of the reactions catalyzed by this enzyme [29, 112].

An interesting method how to utilize docking for predictions of CYP2C9 metabolism sites was published recently [102]. First, the metabolites are generated using rule-based methods. Then these metabolites are docked into the active site of enzyme in order to remove false positives. Nevertheless regioselectivity of CYP2C9 metabolism is predicted less accurately by docking than by other methods [100], and this raises questions on the potential of wide application of such models.

One of possible reasons for failures of structural models is the nature of human drug metabolizing enzymes. The tertiary structures of the major human cytochrome P450 were determined using X-ray crystallography. The protein is usually modified before crystallization: the N-terminal membrane anchor is removed, and sometimes several amino acid residues are substituted to improve solubility [116]. These modifications cause oligomerization of engineered proteins which can be observed not only in the crystals but also in solution [117]. The significance of these facts is not sufficiently discussed. Moreover, both experimental and theoretical investigations showed that the structure of CYP2C9 and CYP3A4 changes upon ligand binding [35, 36, 42, 118], and water molecules can affect substrate binding [43]. Given these facts, the interpretation of data on tertiary structures of human cytochrome P450 enzymes is not straightforward.

Furthermore, docking algorithms have drawbacks which raise difficulties in their practical application [119]. The correct binding pose is not always provided with the highest score. Different programs show best results for particular proteins, and the performance is usually low having no prior knowledge on the protein, software and scoring functions. Under these circumstances, only a skilled computational chemist may produce accurate and reliable docking results.

A notable example of a use of protein-based model for prediction of metabolism sites is implemented in the software package *MetaSite* [98]. This program tries to solve aforementioned problems of docking methods by providing a fully automated analysis of substrate interaction with the active site of the selected enzyme and prediction of the most reactive atoms. These are predicted using a similar approach to *SMARTCyp*: hydrogen abstraction energy is calculated for small fragments by *ab initio* methods, and when predicting for a new compound,

the value is looked up in the database. When a certain fragment is not found among the precalculated ones, semiempirical AM1 method is used.

The protein structure is converted to a flexible Molecular Interaction Field using the GRID method [120]. The generated fingerprint of the active site is matched with the fingerprint of organic molecule. The interaction between substrate and enzyme is described by hydrophobic, hydrogen bond forming or charge capabilities. Such approach allows for the users of *MetaSite* to import their protein structures and apply the method for any cytochrome P450 enzyme.

The predictions of *MetaSite* has been validated several times in pharmaceutical companies. About 80% of metabolism sites were identified using this software in these evaluations [98–100, 121, 122]. The program was applied for metabolic stability optimization of cyclooxygenase-2 inhibitors [6–8] and for automation of MS data analysis in metabolite identification [10].

1.4.4 Data Mining Models for Prediction of Metabolism Sites

In the last decades metabolites have been determined experimentally for many compounds, leading to accumulation of large databases containing thousands of biotransformation reactions. As a result, various data mining techniques have been proposed for prediction of metabolism regioselectivity. Reaction sites have been predicted by fingerprints of known substrates [92, 93], statistical evaluation of biotransformation rules [95] or by QSAR models [91, 96, 97]. Such empirical approaches, briefly summarized in Table 1.5, provide the simplest way to utilize constantly increasing amounts of experimental cytochrome P450 regioselectivity data.

SPORCalc (Substrate Product Occurrence Ratio Calculator) and *SyGMA* (Systematic Generation of Potential Metabolites) use experimental databases to calculate the frequency of observed metabolites in case of metabolism sites or biotransformation rules. *SPORCalc* uses atom-centered structural fingerprints [92, 93]. Every atom and its neighborhood up to 6 atoms is described by fingerprints based on 33 atom types. Then similar atoms are retrieved from database and occurrence ratio of metabolism sites among them is calculated. *SPORCalc* has been applied in case of both Phase I and Phase II metabolism, and more than

Table 1.5: Summary of data mining models for prediction of drug metabolism regioselectivity.

Model and data	Modeling details	Test set prediction results
<p><i>SPORCalc</i> [92, 93] Training set: 20,029 compounds (64,650 reaction)^{a,b} Test set: 30 new compounds</p>	<p>Calculation of occurrence ratios for metabolism sites among similar ones in database, described by atom-centered structural fingerprints</p>	<p>87% of experimental sites within 3 highest ranked positions</p>
<p><i>SyGMA</i> [95] Training set: 1,848 compounds (6,187 reactions)^b Test set: update of database^{a,b} P450 test set: 106 compounds (127 reactions)^c</p>	<p>Probabilities for 144 biotransformation rules are calculated according to fraction of experimentally observed metabolites</p>	<p>68% of total metabolites were reproduced. In the compounds representing cytochrome P450 reactions, 84% of metabolites were predicted (107/127).</p>
<p>Sheridan et al. [91] Training set: 316, 124, and 92 compounds for CYP3A4, CYP2D6 and CYP2C9 Test set: 25 compounds</p>	<p>Random Forest classification using atom-centered fragmental and accessibility descriptors</p>	<p>Metabolism site was among the two top ranked atoms for 84% (16/19), 70% (7/10) and 67% (6/9) of molecules in case of CYP3A4, CYP2D6 and CYP2C9, respectively.</p>
<p><i>CypScore</i> [96] Training set: 844 compounds (2,336 reactions)^c Test set: 345 compounds (612 reactions)</p>	<p>Six binary models for 7 reactions using MLR and quantum chemical and surface descriptors</p>	<p>61% of sites obtained high score (373/612); for 74% (254/345) of compounds experimental site was among 3 top ranked atoms.</p>
<p><i>SOME</i> [97] Training set: 1,819 metabolism sites^{a,c} Test set: Sheridan et al. data; 24 compounds (45 metabolism sites)</p>	<p>Classification models for 6 reactions using SVM and quantum chemical descriptors</p>	<p>Accuracy: 79% (295/373 atoms) Sensitivity 80% (36/45 metabolism sites)</p>

^a Data from MDL Metabolite Database.

^b Phase I and Phase II biotransformations.

^c Cytochrome P450 catalyzed Phase I metabolism reactions.

87% of metabolism sites were predicted among three top ranked sites for 30 test set compounds. Later this method was implemented in the software program *MetaPrint2D* [94]. The user of this program can choose similarity thresholds for customized searching of metabolism sites and select databases to obtain predictions for human, rat or dog metabolism.

SyGMA is based on prioritization of biotransformation rules [95]. Each of 144 rules covering Phase I and Phase II metabolism are applied to the whole database. The number of metabolites that match experimental ones is divided by the number of all metabolites generated according to every rule. This ratio is the probability score for the rule. If possible, the rules were further refined to make them more specific and increase the number of correctly predicted metabolites obtaining higher probability values. For example, aromatic hydroxylation could be split according to presence of substituents in *ortho*-, *meta*- or *para*- positions, and aliphatic metabolism sites were attached to aromatic, heteroaromatic or aliphatic cores. This model predicted 71% of all metabolites in training set and 68% of metabolites in test set. In case of oxidative Phase I reactions the results were even better with 84% of metabolites predicted. The number of false positive predictions is shown to be smaller than in case of other rule-based biotransformation predictors.

The third approach that is used to develop drug metabolism regioselectivity models using available experimental data is structure-activity relationship based on statistical classification and regression methods. First such model was developed by Sheridan et al. for prediction of metabolism by three major human cytochrome P450 enzymes, CYP3A4, CYP2D6 and CYP2C9 [91]. The majority of experimental metabolism sites were within atoms ranked first or second by models for all three enzymes. Later two structure-activity relationships were created using quantum chemical descriptors, *CypScore* [96] and *SOME* (Site of Metabolism Estimator) [97]. Position ranked top by *CypScore* was metabolism site for 46% of compounds, and 61% of experimental metabolism sites obtained rank higher than 38, considered as threshold for classification. *SOME* correctly predicted about 80% of metabolism sites.

All previously developed structure-activity relationships of metabolism regioselectivity have the same drawback: the applicability domain of the model is

not defined, as the modeling methods do not allow estimation of prediction reliability. Therefore it was decided to create a new model, predicting the regioselectivity of metabolism using the GALAS modeling method. The preliminary results already established the effectiveness of the methodology and the usefulness of calculated Reliability Index in identifying the correct predictions, i. e. the evaluation of a model applicability domain [123].

Chapter 2

Data and Modeling Methods

2.1 CYP3A4 Inhibition Data

2.1.1 Literature Dataset

Two datasets have been used in the development and validation of CYP3A4 inhibition models. CYP3A4 inhibition data in the first set (“Literature dataset”) were collected from various literature sources (scientific publications, drug prescribing information). Inhibition of metabolism of probe CYP3A4 substrates (midazolam, testosterone, erythromycin, 7-benzyloxy-4-trifluormethylcoumarin (BFC), nifedipine, and others) has been considered. Classification of compounds was performed only after critical analysis of original literature in order to identify any cases of contradictions in experimental CYP3A4 inhibition data, such as substrate dependency [79, 124] or inconsistency between data obtained using human liver microsomal and recombinant enzyme [125]. Only IC_{50} values determined at substrate concentration close to K_m value were used for classification. Compounds having $IC_{50} < 40 \mu M$ were classified as CYP3A4 inhibitors, having $IC_{50} > 60 \mu M$ were classified as non-inhibitors, compounds with intermediate IC_{50} values (40-60 μM) or discrepant results in different assays were marked as inconclusive. In cases when detailed analysis of CYP3A4 inhibition was available and inhibition constant K_i has been reported, compounds having $K_i < 20 \mu M$ were classified as CYP3A4 inhibitors.

2.1.2 PubChem Dataset

The second set (“PubChem dataset”) was downloaded from National Center for Biotechnology Information (NCBI) PubChem database (assay ID 884) on 10 June 2008 [126]. PubChem database contained 14127 entries concerning CYP3A4 inhibition, determined using luminogenic CYP3A4 inhibition screening method [66]. PubChem data were preprocessed prior to further analysis: entries containing inorganic compounds, non-covalent complexes, and mixtures were excluded; salts were converted to corresponding acids or bases; water molecules were removed from hydrates. All compounds that were marked as CYP3A4 activators in PubChem database were excluded from further analysis. The resulting entries were classified as CYP3A4 inhibitors, non-inhibitors or inconclusive. To identify inhibitors in general the same rules were used as in PubChem: entry was classified as inhibitor if observed assay score was ≥ 40 and non-inhibitor if the score was 0, with inconclusive range lying between scores of 0 and 40. PubChem activity score is assigned from fitted IC_{50} value, with respect to completeness of dose-response curve and efficacy of inhibition (maximum inhibition response). For compounds having PubChem activity score > 40 the IC_{50} is less than $40 \mu M$, therefore the classifications of PubChem and literature datasets in this case are consistent with each other. No specific effort was made to review any supporting experimental information and PubChem scores were used as provided.

Following the classification of the entire PubChem database, the attention was switched to the compounds provided with several experimental results, i.e. represented by multiple entries in the database (3797 entries for 1546 compounds with stereoisomers treated as duplicates). Only 55 compounds (4%) had contradicting classification as active (activity score > 40) in one experiment while inactive (activity score 0) in another. Such compounds were marked as inconclusive. Additional 398 compounds (26%) had inconclusive result in one of the experiments and therefore were also marked as inconclusive. For the remaining 1093 compounds (71%) that had consistent classification one entry per compound was left in database.

Another classification scheme was applied to distinguish effective inhibitors in PubChem dataset: compounds with $IC_{50} < 5 \mu M$ and maximal efficacy $> 70\%$

Table 2.1: Distribution of compounds in data sets with regard to CYP3A4 inhibition.

Data set	No. of compounds	Inhibitors	Non-inhibitors	Inconclusive compounds
Literature dataset ^a	907	335 (36.9%)	497 (54.8%)	75 (8.3%)
PubChem dataset (general inhibition) ^b	11,060	303 2 (27.4%)	5,496 (49.7%)	2,532 (22.8%)
PubChem dataset (effective inhibition) ^c	11,060	1,238 (11.2%)	6,401 (57.9%)	3,421 (30.9%)

^a Inhibitors: $IC_{50} < 40 \mu M$; non-inhibitors: $IC_{50} > 60 \mu M$; inconclusive compounds: IC_{50} 40-60 μM or contradicting results in different experiments.

^b Inhibitors: PubChem Activity score > 40 ; non-inhibitors: PubChem Activity score 0; inconclusive compounds: other (including contradictory results in repeating experiments).

^c Inhibitors: $IC_{50} < 5 \mu M$ and efficacy $> 70\%$; non-inhibitors: $IC_{50} > 30 \mu M$; inconclusive compounds: other.

were classified as effective inhibitors, compounds with activity score of 0 or $IC_{50} > 30 \mu M$ were treated as inactive, remaining compounds were classified as inconclusive.

2.1.3 Summary of CYP3A4 Inhibition Datasets

Number of compounds and their distribution with regard to CYP3A4 inhibition classes for both datasets are summarized in Table 2.1. Inconclusive data from all datasets were not used in modeling. Compounds that were present in both literature and PubChem databases were excluded from the latter one, which was used as a validation set.

Consistency of experimental data for 297 compounds present in both Literature and correspondingly classified PubChem datasets has been checked before modeling. 119 compounds have inconclusive classification in one of the sets. Detailed analysis of experimental values for remaining compounds is given in Table 2.2. For 156 out of 178 compounds (88%) consistent classification of inhibition was observed in both datasets.

Table 2.2: Consistency of experimental data present in both literature and PubChem datasets.

		Literature	
		Inhibitor	Non-inhibitor
PubChem	Inhibitor	36	13
	Non-inhibitor	9	120
		Agreement: 87.6%	

2.2 Regioselectivity Data

2.2.1 Modeling Dataset

Experimental data on metabolism in human liver microsomes for 873 compounds were collected from scientific publications dealing with analytical identification of the metabolites observed after the incubation of compound with human liver microsomes or recombinant human cytochrome P450 enzymes.

Every carbon atom having at least one hydrogen atom attached was marked as a site of metabolism, if hydroxylation or oxidation at the atom was observed, or site of no metabolism otherwise. For dealkylation reactions, carbon atoms of the leaving groups were marked in the same manner. Sulfur atoms having < 4 neighbors were marked as metabolism sites if they were oxidized. Some atoms were marked as "inconclusive" and consequently not used in the modeling. The overall dataset contained 8,608 atoms, 1,326 of them were marked as metabolism sites.

The complete dataset was then divided into N-dealkylation, O-dealkylation, aliphatic hydroxylation, aromatic hydroxylation, and S-oxidation subsets according to the reaction types. The description and composition of the obtained subsets in case of human liver microsomal metabolism is outlined in Table 2.3.

Additionally, for compounds having experimental data on CYP2D6 catalyzed metabolism, the atoms of the modeling dataset were marked if they are sites of metabolism by this enzyme. Table 2.4 shows the composition of this dataset.

Table 2.3: Datasets of atoms used for regioselectivity modeling.

Subset	Description	No. of compounds	No. of metabolism sites	Total No. of marked atoms
N-dealkylation	Aliphatic CH attached to N	511	333	1,173
O-dealkylation	Aliphatic CH attached to O	488	260	1,033
Aliphatic hydroxylation	Aliphatic CH (other)	723	318	29,04
Aromatic hydroxylation	Aromatic CH	739	358	3,341
S-oxidation	S with less than 4 neighbors	135	57	157
Total		873	1326	8,608

Table 2.4: CYP2D6 regioselectivity dataset.

Subset	No. of compounds	No. of metabolism sites	Total No. of marked atoms
N-dealkylation	440	104	1,020
O-dealkylation	372	84	777
Aliphatic hydroxylation	569	58	2,227
Aromatic hydroxylation	577	108	2,632
S-oxidation	83	15	100
Total	673	369	6,756

2.2.2 External Validation Dataset

After the model has been developed, experimental data for 42 compounds were collected from scientific literature, mostly from newly published articles. This dataset served as external validation set, and was also used for comparison of the predictions with previously published regioselectivity models. All atoms of the compounds in this dataset were marked if they are sites of metabolism using the same criteria as in modeling dataset. The list of validation dataset compounds with references is presented in Chapter 4, Section 4.2.

2.3 Descriptors

2.3.1 Fragmental Descriptors

Fragmental descriptors were chosen for the modeling of CYP3A4 inhibition. Molecules were fragmented using a set of 379 predefined fragments. The major part of this set is intended for the description of general chemical compound constitution and was comprised of conventional fragmental descriptors, such as atoms, functional groups, molecular shape fragments, and others. These descriptors have been already used and proved effective in previous projects, involving modeling various biological activities and chemical properties [13, 14]. Additionally several typical fragments describing CYP3A4 specificity were added (e.g., nitrogen containing heterocycles, methylenedioxybenzene, fragments representing possible cytochrome P450 metabolism sites, etc.).

2.3.2 Atom-centered Fragmental Descriptors

The regioselectivity of metabolism is not a whole-molecule property: the possibility of metabolism has to be predicted for every single atom in the molecule. Therefore modifications of the standard fragmental descriptors were necessary in order to generate different values of descriptors for every atom in the molecule. The atom-centered fragmentation was developed. It provides information about the atom types present at equidistant positions from the selected atom, called levels. Fig. 2.1 shows how these levels are computed for different atoms in different molecules.

In order to describe both the metabolism site and the whole molecule, atom-centered fragmentation up to level 15 was performed. The nature of the selected atom itself and its neighborhood have the largest influence in the reactivity of the metabolism site. They were described using fragmental descriptors in which atoms were differentiated according to element (C, N, O, S, P, halogens, etc.), hybridization, number of attached hydrogens, cyclization and aromaticity. This detail description was used up to level 2 for aliphatic carbon atoms. In case of aromatic atoms it was extended up to level 4 with the intention to cover

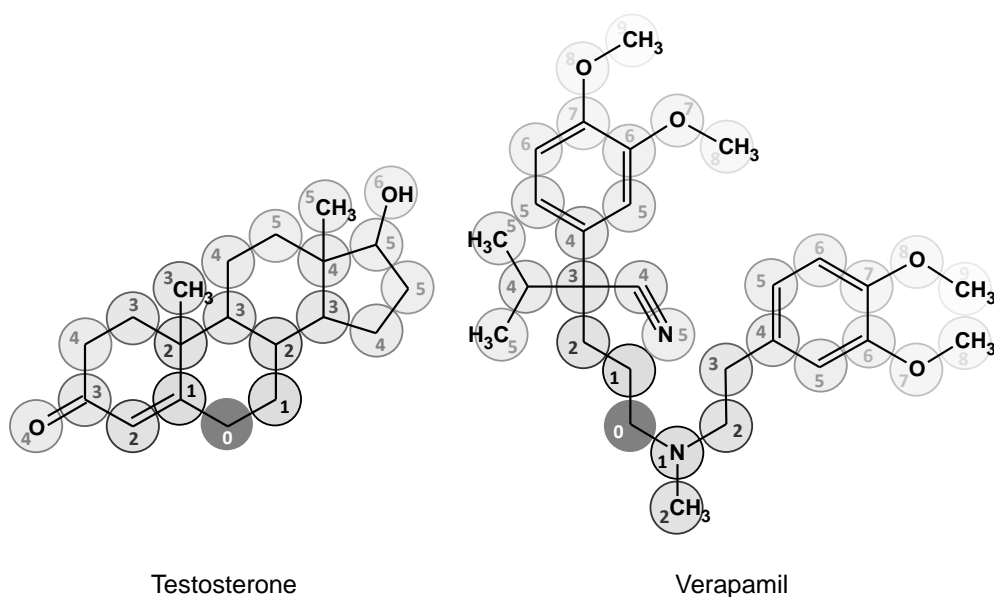


Figure 2.1: Examples of levels in atom-centered fragmentation. Reprinted with permission from [127].

the first level of all possible substituents in a six-membered aromatic ring. In addition, slightly different sets of predefined fragments were used for aliphatic and aromatic carbon as well as sulfur atoms.

The further levels have less influence to the reactivity of metabolism site and can be described more roughly. Merging of them was performed as it is shown in Fig 2.2 to reduce the number of descriptors and account for the flexibility of the molecule. Another merging scheme served as description of the presence of functional groups having strong electronic effects ($-\text{NO}_2$, halogens, $-\text{CF}_3$, $-\text{CN}$, esters, charged groups etc.). The significance of presence of such groups depending on the distance from the selected atom was assigned in analogy with the known extent of the electronic interaction propagation over the distance in the molecule, and then all levels were summed up to one descriptor.

2.4 Statistical Methods

GALAS modeling methodology has been developed utilizing extensive experience from numerous QSAR modeling projects. Recently the successful

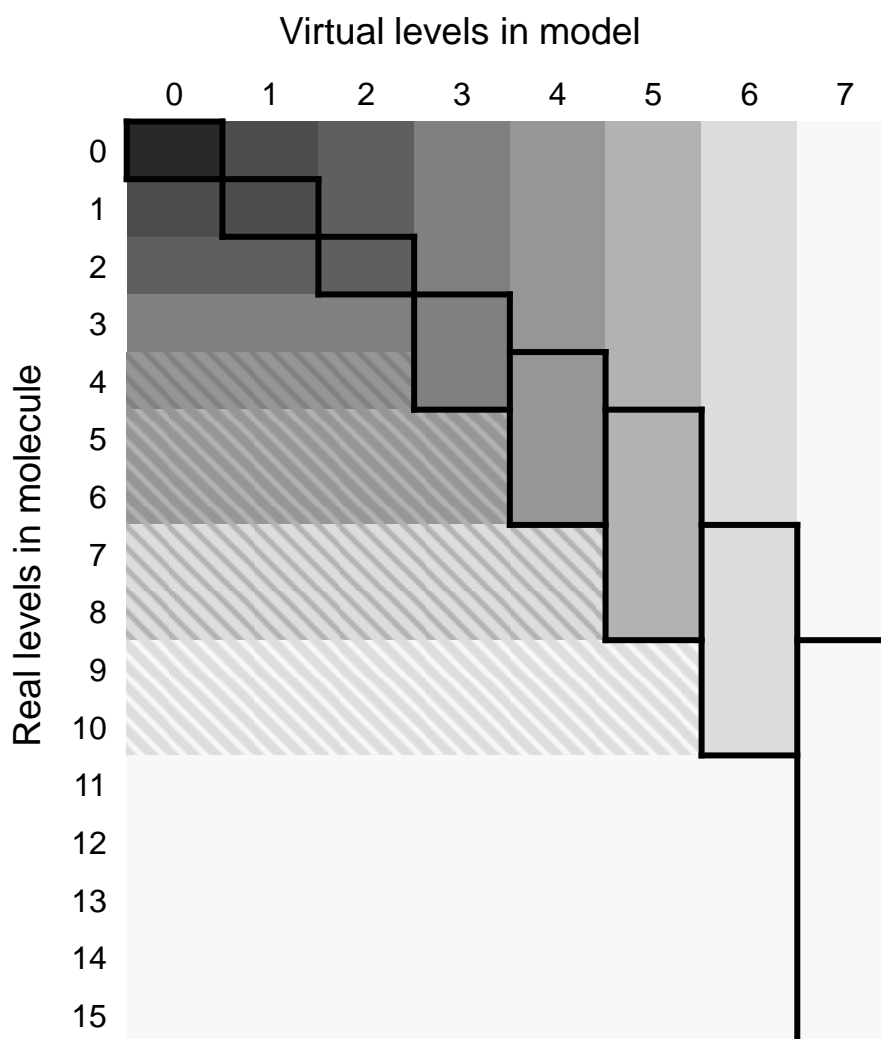


Figure 2.2: Merging of levels in atom-centered fragmentation. Reprinted with permission from [127].

application of this methodology in predicting various properties of continuous nature (e.g., $\text{Log}P$, LD_{50}) has been reported [13, 14]. Two main parts constituting the basis of the method are a global QSAR model providing baseline predictions for the property of interest and local corrections calculated according to the experimental data for the most similar compounds from the training set. The global model for CYP3A4 inhibition is intended to capture the general trends in enzyme specificity and identify structural fragments tending to affect inhibition. The global models of regioselectivity describe the reactivity of the atom in cytochrome P450 catalyzed reactions. The purpose of the second step is identifying

any systematic errors made by the global model in the local chemical space of the test compound and compensating for it. In case of regioselectivity model it can be also viewed as the corrections to reactivity emerging from the binding to active sites of the enzymes.

In further sections general description of the GALAS modeling method is provided. The same statistical method was applied for both modeling of CYP3A4 inhibition and prediction of metabolism sites. The only differences between the models are that the regioselectivity models use data for atoms instead of whole compounds, and the atom-centered method for generation of fragmental descriptors.

2.4.1 Global Model

Baseline QSAR is a linear model built using logistic PLS regression. This method is a variation of ordinary PLS, possessing all its useful features in combination with the ability to analyze binary data. The predicted value in this case is the logit transformation of probability for a compound to exhibit activity (p) which is calculated as the sum of all the fragmental contributions:

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right) = \sum_i a_i f_i + c \quad (2.1)$$

here p is the probability for a compound to be active; f_i is the occurrence sum of a particular fragment in a molecule; a_i – statistical coefficient of the fragment, determined using logistic PLS; c – intercept.

Such model predicting CYP3A4 inhibition was compared to those developed using linear Support Vector Machine (SVM) and nonlinear Random Forest methods. The statistical results were similar for all three models (see Chapter 3, Section 3.1). In fact, the implementation of GALAS modeling method described in this article is also possible using other global models as baseline. Logistic PLS was chosen for further works due to its easier interpretability and our experience in development of PLS-based QSAR models.

2.4.2 Dynamic Similarity

Another important role of the baseline model is determining the “dynamic” similarity key used to compare compounds in prediction space, i.e. in terms of a particular analyzed property. Igor Tetko and his colleagues were among the first to realize the advantages of such approach over the use of typical methods that define the similarity of two molecules in descriptor space and *a priori* [128, 129]. The same predefined similarity key used in the models for different properties disregards the possibility that what is similar in case of property A can be significantly less relevant for property B.

Global model is developed combining logistic PLS with the bootstrapping technique. The latter method implies random compound sampling from the initial training set [130], i.e. generation of new training sub-sets and derivation of independent model for each sub-set. Performing this procedure 100 times provides each compound with a vector of 100 predictions, each based on a slightly different part of the initial training set. This ensemble of models contains information about the influence and stability of every independent descriptor. Broad variability of the particular descriptor coefficient value across the models is an indication that the global QSAR cannot correctly estimate its influence. As a result, molecules containing the same highly variable descriptor are the best candidates to correct the potentially unreliable baseline prediction. The more variation is observed in the values of a particular descriptor coefficient, the more significant contribution the corresponding atom or functional group should have in the similarity assessment.

On the other hand, the most stable descriptors are those most widely encountered (-CH₂-, -CH₃, aromatic carbon and similar in case of fragmental descriptors). Their minimal contribution in the similarity assessment (i.e., the compounds that differ only by one -CH₂- or -CH₃ group will be treated as nearly identical by the model) conforms to the general chemical logic that compounds in homologous series are the most similar compounds among themselves.

The quantitative measure of the individual similarity between any two compounds (Similarity Index, SI_i) in the GALAS model is the square of correlation coefficient (r^2) between the prediction vectors. When comparing these vectors,

the same trends in variability of 100 logistic PLS predictions indicate that the two compounds possess a very similar pattern of structural features found by the model to be of great influence in case of the analyzed property. This leads to a conclusion that two such compounds are indeed similar. Every difference in the set of significant fragments will inevitably reduce the correlation between the prediction vectors decreasing the SI_i .

2.4.3 Local Model

Determining the similarity between any two compounds is a key process in the second layer of the GALAS modeling methodology. Here the predictions of the baseline model are compared to experimental values of the 5 training set compounds most similar to the query molecule. The final probability estimation is a combination of global prediction and local correction:

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right) = \sum_i a_i f_i + \Delta + c \quad (2.2)$$

here Δ is the correction calculated according to the experimental data for the most similar compounds.

The Δ value itself is calculated as a weighted average from the differences between global QSAR predictions and experimental data for the 5 most similar compounds in the training set:

$$\Delta = \frac{\sum_{i=1}^5 w^{i-1} \cdot SI_i \cdot \Delta_i}{\sum_{i=1}^5 w^{i-1}} \quad (2.3)$$

here Δ is a correction that should be applied for the given prediction from the global model; w is a weighting constant (simple average is calculated if this constant is 1); SI_i is an individual Similarity Index between given compound and the i -th most similar compound in the training set; Δ_i is the difference between logit value of the experimental result and the value predicted by global

model for the i -th most similar compound prior to its transformation into the baseline probability.

Since function logit approaches negative and positive infinity when its argument approaches 0 and 1 respectively (the numbers used as experimental result indications in our case), 5% cut-offs are used at both ends of the probability value range to avoid unreasonable values while calculating Δ_i in eq. 2.3, i.e., $\text{logit}(> 0.95) = \text{logit}(0.95) = 2.94$ and $\text{logit}(< 0.05) = \text{logit}(0.05) = -2.94$. For example, Δ_i for experimental positive result and predicted baseline probability 0.7 is calculated as the difference between $\text{logit}(0.95)$ and $\text{logit}(0.7)$, and in case of baseline probability > 0.95 Δ_i is zero.

The application of the Δ correction prior to the transformation using the logistic function (as shown in eq. 2.2) and subsequent specifics of the Δ_i calculation (as described in eq. 2.3) are prompted by necessity to confine the final corrected probability (p in eq. 2.2) into the interval between 0 and 1. Straightforward application of the correction to the baseline probability would leave the possibility of the final probability exceeding 1 or falling below 0 in certain cases. Conversely, the above described workflow assures the different influence of the correction on the final probability depending on the initial baseline prediction. A proposed large positive correction will have a great impact if the baseline probability is low, but will not influence the prediction if the baseline probability is already close to 1.

2.4.4 Estimation of Prediction Reliability

GALAS modeling methodology also allows estimating the quality of prediction. This feature is even more significant than the accuracy improvement, compared to the global statistical methods. Knowing the prediction reliability is very important given the fact that any QSAR model is characterized by its applicability domain, outside of which, the performance of the model is usually poor [12, 14, 76]. No prediction can be considered reliable if there are no similar compounds in the training set. In the situations when such compounds do exist, but experimental data for them are inconsistent with regard to the global model, predictions based on such data cannot be confident as well. These two

assumptions provide the basis for the calculation of Reliability Index (RI) which has been made directly dependent on these two factors.

The presence or absence of similar compounds in the training library is indicated by the compound Similarity Index (SI) to the entire data set. This index is calculated by weighted averaging of all the individual Similarity Indices (SI_i) for the test molecule and each of the 5 most similar compounds from the training library. Data-Model Consistency Index ($DMCI$) is used to quantitatively evaluate the consistency of experimental data of similar compounds with the global baseline model. $DMCI$ value compares the individual differences between experimental and predicted baseline property values (Δ_i) for the same most similar compounds from the training library with the overall local correction for the compound of interest calculated by the eq. 2.3. The more individual differences are scattered around the calculated average (Δ), the more inconsistent are the data for the similar compounds with regards to the global baseline model and vice versa.

The final prediction Reliability Index characterizing the applicability domain of the model is calculated in the following manner:

$$RI = SI \cdot DMCI \quad (2.4)$$

It is a value set to vary between 0 and 1 with larger RI values indicating more reliable predictions. A more in-depth consideration of some of the mathematical background of the GALAS modeling methodology is available in a recent publication [14].

2.4.5 Training of the GALAS Model

Another important feature of GALAS modeling methodology is the possibility for an easy and straightforward expansion of the applicability domain of resulting models. Training of the GALAS models is performed by simply adding new compounds with experimental data to the similarity correction (local) part of the modeling, as shown in Fig. 2.3. If the baseline model is unable to predict

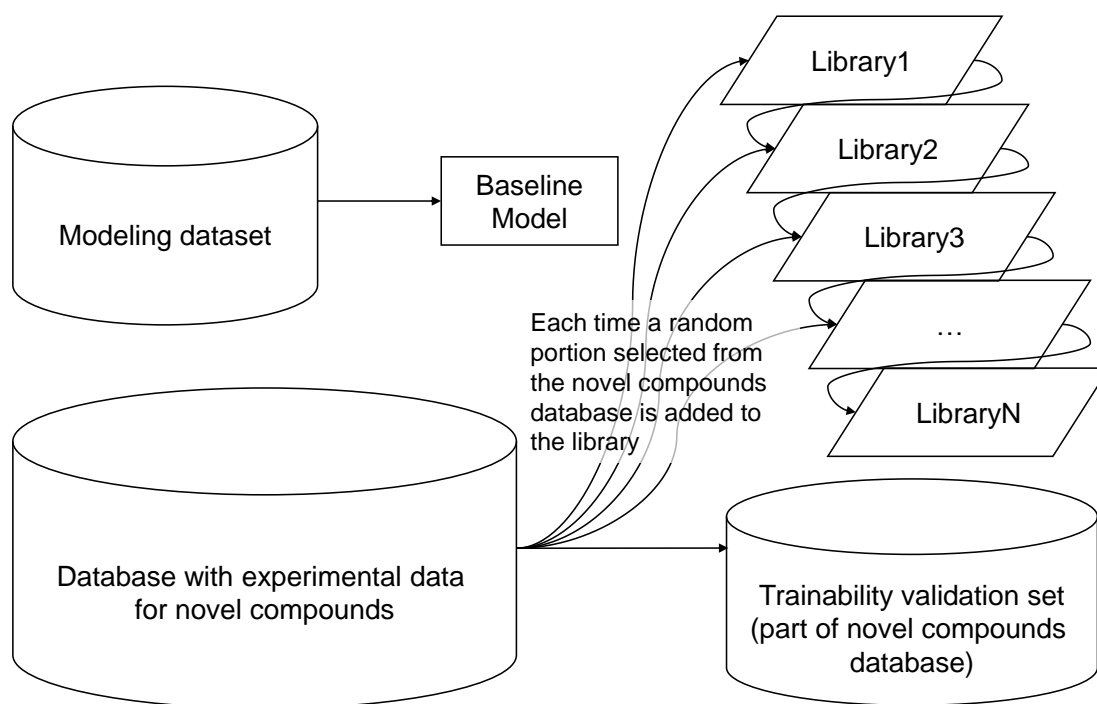


Figure 2.3: The workflow of the GALAS model trainability testing.

accurately for a certain class of compounds, the local similarity correction can compensate this inaccuracy by calculating the appropriate Δ value. Introduction of the new data results in the adaptation of the local model to a new part of the chemical space, represented by the newly imported compounds, while the same global model is used. The latter fact allows performing model training “on-the-fly” without time consuming full statistical re-parameterization of the model.

2.5 Development and Validation of Models

2.5.1 CYP3A4 Inhibition Model

The model predicting CYP3A4 inhibition is based on the Literature dataset and predicts probability for a compound to be a CYP3A4 inhibitor with $IC_{50} < 40 \mu M$. Literature dataset was randomly split into the training and test set (190 compounds in the test set, ca. 20%). Initially the model was tested on this internal

test set, followed by a more rigorous validation using the PubChem dataset as an external test set, which was published after the model has been developed. As it was already mentioned, compounds that were present in both databases were excluded from the external test set. Compounds with predicted probability > 0.5 were considered as inhibitors, whereas probability of < 0.5 indicated non-inhibitors. Additionally, the performance of global and local models was adjusted using receiver operating characteristic (ROC) curves [131].

Following performance test the possibility to train GALAS model has been investigated. In order to verify whether the model can be trained using data from a different assay two different trainability experiments have been conducted on PubChem data classified as CYP3A4 inhibitors using different thresholds. In the first scenario the model was trained using PubChem dataset, classified according to PubChem activity score, i.e. data from similar experimental assay. The second trainability example used PubChem dataset with only effective CYP3A4 inhibitors classified as positive compounds, mimicking a different experimental assay with a different potency threshold to identify inhibitors. In both cases the PubChem dataset was divided into several parts: one half of it was reserved as the validation set, while the remaining parts were added one by one to the training library (Fig. 2.3). The similarity correction libraries contained ca. 2.5, 5, 10, 25, and 50 percent of the PubChem dataset.

The third trainability example involved compounds having a completely new structural scaffold. CYP3A4 inhibition data for 10 insulin-like growth factor-1 receptor (IGF-1R) inhibitors have been recently published [132]. All these compounds were classified as CYP3A4 inhibitors having $IC_{50} < 40 \mu M$. Five compounds were randomly selected and added one by one to the similarity correction library which in case of original model consisted of the literature dataset. After the addition of each training molecule the probability to inhibit CYP3A4 and corresponding RI values were calculated for five remaining compounds.

2.5.2 Regioselectivity Model

Individual models were built for all types of reactions listed in Table 2.3 using the GALAS modeling method and atom-centered fragmental descriptors. Each

subset of atoms was randomly divided into the training (70%) and test (30%) sets. Initially the model has been validated using these internal test sets. Atoms with calculated probability > 0.5 were considered metabolism sites. The influence of local modeling was also visualized by ROC curves.

The predictions of regioselectivity model have been further evaluated using newly published data (external validation set). The predictions were divided into four classes according to their quality. The definitions of classes are as follows:

- “Excellent” predictions were those where the model produced probabilities of > 0.5 for all experimentally determined metabolism sites and the atom ranked 1st was experimentally determined as a metabolism site.
- In cases where most metabolism sites were predicted with probability > 0.5 , the prediction was marked as “good”, though for some compounds the atom ranked by the model as most probable metabolism site was experimentally not found to be metabolize.
- When less than a half of experimentally determined metabolism sites obtained probabilities > 0.5 , the prediction was labeled “satisfactory”. If the only experimentally determined metabolism site was ranked as one of three most probable sites, but the probability was < 0.5 , the prediction was also labeled “satisfactory”.
- If the model failed to identify metabolism sites both by score and rank, the prediction was marked as “unsatisfactory”.

For comparison with previously published regioselectivity models, the metabolism sites in external validation set compounds were predicted using *SMART-Cyp* software [108, 133]. Similarly, these predictions were divided into the same quality classes according to analogous criteria with some modifications due to different output of the program:

- “Excellent” predictions were those where the atom ranked 1st was experimentally determined as a metabolism site.

- In cases where most metabolism sites were among top 3 atoms, the prediction was marked as “good”.
- The prediction was labeled “satisfactory” if experimentally determined metabolism sites was ranked as one of 6 most probable sites.
- If the software failed to identify most of metabolism sites among 6 top ranked atoms, the prediction was marked as “unsatisfactory”.

After the development and validation of regioselectivity model the trainability feature was tested using data for fluorinated propranolol derivatives [134, 135]. In this case all analogues and metabolites of propranolol (10 compounds) were removed from the modeling dataset, and a special model was built without them. Then three randomly chosen compounds (trifluoropropranolol, propranolol derivative TFE and propranolol *tert*-butyl derivative) were added one-by-one to the similarity correction library which in case of initial model consisted of the atoms of the modified modeling set (Fig. 2.3). The probability of metabolism was predicted for all remaining propranolol analogues after each addition.

In order to test the possibility to adapt the GALAS model to the specificity of a particular enzyme, we marked the modeling dataset according to available CYP2D6 metabolism data (Table 2.4). CYP2D6 was chosen as the most specific of major human drug metabolizing enzymes. The same baseline models were used, and the similarity correction library consisted of the part of CYP2D6 dataset that intersects with the training set. This model was then tested on the rest of CYP2D6 regioselectivity dataset.

2.5.3 Software

Molecule fragmentation and all subsequent statistical analysis were performed using *Algorithm Builder 1.8* software [136, 137], except SVM and Random Forest models for CYP3A4 inhibition which were developed using *R 2.6.2* [138]. ROC curves were generated by a web-based calculator [139]. The trainability of GALAS models was tested on *ACD/ADME Suite 4.95* software application [140]. *SMARTCyp* predictions of metabolism sites were obtained using a web-based version of this software [133].

Chapter 3

Results and Discussion: CYP3A4 Inhibition Modeling

Classification scheme, distinguishing compounds between inhibitors and non-inhibitors according to their experimental data is the first aspect of any attempt to develop statistical classification models of enzyme inhibition. Inhibition potency thresholds selected for classification in earlier structure-activity relationship studies of CYP3A4 inhibition cannot be directly compared between each other, because experimental estimation of CYP3A4 inhibition depends on the methods used [63, 79, 80]. The percent of active compounds identified in different screening programs also confirms that; and the distribution of compounds according to CYP3A4 inhibition potency is not the same in available databases. PubChem screening program identified 27% of compounds with IC_{50} values less than $40 \mu M$ (see Table 2.1). The study performed on *Novo Nordisk* in-house database, utilizing the assay of erythromycin metabolism inhibition in human liver microsomes, identified a similar fraction of 26% of compounds as active, yet a lower threshold was used ($IC_{50} = 20 \mu M$, instead of $IC_{50} = 40 \mu M$) [73]. Even greater proportions of active compounds were reported in other studies using IC_{50} thresholds as low as $< 10 \mu M$ for compound classification (see Table 1.3) [67, 69, 71, 74].

The aforementioned facts suggest that universal and objective thresholds for classification of compounds into CYP3A4 inhibitors and non-inhibitors cannot be established [1]. In this work two classification thresholds for diversifying

the compounds according to their experimental CYP3A4 inhibition data were chosen. The first one is used to identify inhibitors in general ($IC_{50} < 40 \mu M$), while the second one distinguishes only effective inhibitors ($IC_{50} < 5 \mu M$) from the rest. The structure-activity relationship model described in this article predicts general CYP3A4 inhibitory properties ($IC_{50} < 40 \mu M$) and was built using data from literature sources (Literature dataset). This unusually high threshold to classify inhibitors is chosen for consistency of classification of literature and PubChem data. It allows identification of the most general properties related to CYP3A4 inhibition.

Further it will be demonstrated how this model, built using GALAS modeling methodology, is able to adapt itself even to the binary data obtained using IC_{50} thresholds other than in the construction of the training set, i. e. when only effective inhibitors with $IC_{50} < 5 \mu M$ were classified as active compounds. Consequently, proposed classification model based on any selected threshold can later serve as baseline model in training with CYP3A4 inhibition data from any available inhibition assay: either based on fluorescent substrate metabolism by recombinant CYP3A4, or inhibition studies with human liver microsomes, or other.

3.1 Global Model

The first step of the presented GALAS model is a logistic PLS with predefined set of fragments as independent variables (see Section 2.4) – a baseline model of CYP3A4 inhibition. This is a linear and additive model. It produces relevant predictions for both internal and external test sets with accuracy of about 80%. The overall results are comparable to other standard machine learning methods like SVM and Random Forest (Table 3.1). The fact that an additive model accurately describes probability for a compound to inhibit CYP3A4 is in agreement with the very broad specificity of this enzyme [4]. The active site cavity of CYP3A4 is considerably larger than that of any other cytochrome P450 isoform. It also has the potential to expand considerably on ligand binding [41, 42] and is even able to accept several molecules simultaneously [34, 42]. The dependence of CYP3A4 inhibition potency on molecular weight has also been

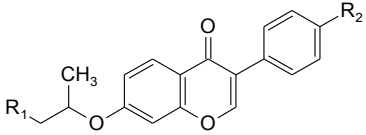
shown to be markedly different from that of the other cytochrome P450 isoforms with no drop in the mean potency for large compounds with molecular weight of $> 750 Da$ [141]. Therefore, additivity of the model still persists for compounds with high molecular weight.


The most attractive feature of *in silico* screening tools is the ability to use them in the assessment of the properties for virtual compound libraries. Large databases of structures can be screened even prior to the actual synthesis of the substances. A part of this study has been devoted to the analysis of the baseline model predictions for some virtual compounds in order to understand how changes in chemical structure affect the probability of drug-like compound to inhibit CYP3A4. Table 3.2 presents predicted probabilities to inhibit CYP3A4 for virtual ipriflavone analogues. Ipriflavone itself is not a CYP3A4 inhibitor, but its metabolites are weak inhibitors [142, 143]. Similar flavones that inhibit this enzyme are present in PubChem database. Baseline model predicts ipriflavone as non inhibitor of CYP3A4 – the probability for this compound to have $IC_{50} < 40 \mu M$ is 0.3 ($\text{logit}(p) = -0.85$).

Two studies have been published recently that are dedicated to the analysis of the influence of physicochemical properties and most popular substituents of organic compounds on ADME and toxicity parameters, including CYP3A4 inhibition [141, 144]. The predictions of the baseline model obtained for virtual compounds in Table 3.2 are within a good agreement with results of these studies. The predicted probabilities were converted to corresponding $\text{logit}(p)$ values in order to maintain linear scale before comparison of the published changes in average pIC_{50} [144] with the predicted values of our models. Then the difference between the $\text{logit}(p)$ of a virtual compound and the $\text{logit}(p)$ of ipriflavone was calculated ($\Delta \text{logit}(p)$). This value shows the influence of a particular substituent in our baseline model. The overall correlation of published changes in pIC_{50} and $\Delta \text{logit}(p)$ for considered virtual ipriflavone analogues is high ($r^2 = 0.70$, Fig. 3.1). Some notable substituents are analyzed further in the text.

Introduction of an acidic group makes the inhibition of CYP3A4 enzyme almost impossible (predicted probability for virtual analogue $p = 0.05$, $\Delta \text{logit}(p) = -2.1$). The negative impact of any acidic group on IC_{50} for CYP3A4 inhibition value has been also shown using proprietary data ($\Delta pIC_{50} = -0.55$) [141, 144].








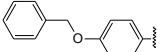

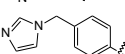

Table 3.2: Effects of structure modifications of virtual ipriflavone analogues on predicted baseline probability for a compound to inhibit CYP3A4.



R ₁	R ₂	ΔpIC_{50} ^a	<i>p</i>	logit(<i>p</i>)	Δ logit(<i>p</i>)
Ipriflavone:					
-H	-H	0	0.3	-0.85	0.00
Ionogenic substituents:					
-COOH	-H	-0.55	0.05	-2.94	-2.10
-H	-COOH	-0.55	0.04	-3.18	-2.33
-NH ₂	-H	0.07	0.23	-1.21	-0.36
-H	-NH ₂	0.07	0.26	-1.05	-0.20
-N(CH ₃) ₂	-H	0.09	0.15	-1.73	-0.89
-H	-N(CH ₃) ₂	0.36	0.49	-0.04	0.81
	-H	0.00	0.19	-1.45	-0.60
Hydrophilic substituents:					
-OH	-H	-0.23	0.25	-1.10	-0.25
-H	-OH	-0.23	0.29	-0.90	-0.05
-CH ₂ OH	-H	-0.05	0.33	-0.71	0.14
-CONH ₂	-H	-0.01	0.21	-1.32	-0.48
-SO ₂ CH ₃	-H	0.12	0.28	-0.94	-0.10
-H	-SO ₂ CH ₃	0.12	0.25	-1.10	-0.25
-H	-NHSO ₂ CH ₃	0.28	0.23	-1.21	-0.36
Ethers, esters and amides:					
-COOCH ₃	-H	0.26	0.48	-0.08	0.77
-H	-COOCH ₃	0.26	0.34	-0.66	0.18
-H	-OCH ₃	0.11	0.37	-0.53	0.32
-H	-OCF ₃	0.44	0.53	0.12	0.97
-H	-OCH ₂ CH ₃	0.22	0.42	-0.32	0.52
-SCH ₃	-H	0.24	0.35	-0.62	0.23
-H	-SCH ₃	0.24	0.33	-0.71	0.14
-CH ₂ OCH ₃	-H	0.19	0.45	-0.20	0.65
-OCOCH ₃	-H	0.16	0.48	-0.08	0.77
-NHCOCH ₃	-H	0.23	0.39	-0.45	0.40

^a ΔpIC_{50} values are taken from a previous publication [144].

Table 3.2: Effects of structure modifications of virtual ipri-flavone analogues on predicted baseline probability for a compound to inhibit CYP3A4. (continued)

R ₁	R ₂	ΔpIC_{50} ^a	<i>p</i>	logit(<i>p</i>)	Δ logit(<i>p</i>)
Aliphatic substituents:					
-CH ₃	-H	0.11	0.36	-0.58	0.27
-CH ₂ CH ₃	-H	0.32	0.39	-0.45	0.40
-CH ₂ CH ₂ CH ₃	-H	0.28	0.43	-0.28	0.57
-CH(CH ₃) ₂	-H	0.29	0.39	-0.45	0.40
-CH ₂ CH ₂ CH ₂ CH ₃	-H	0.20	0.48	-0.08	0.77
-CH(CH ₃)(CH ₂ CH ₃)	-H	0.50	0.41	-0.36	0.48
-C(CH ₃) ₃	-H	0.36	0.57	0.28	1.13
-H		0.51	0.53	0.12	0.97
-H		0.42	0.63	0.53	1.38
Aromatic substituents:					
	-H	0.46	0.77	1.21	2.06
	-H	0.65	0.88	1.99	2.84
	-H	0.54	0.57	0.28	1.13
	-H	0.73	0.54	0.16	1.01
	-H	0.86	0.68	0.75	1.60
	-H	^b	0.97	3.48	4.32
	-H	^b	0.86	1.82	2.66
	-H	^b	0.91	2.31	3.16
	-H	^b	0.9	2.20	3.04
Halogen containing substituents:					
-H	-F	0.07	0.27	-0.99	-0.15
-H	-Cl	0.21	0.37	-0.53	0.32
-H	-Br	0.27	0.35	-0.62	0.23
-H	-CF ₃	0.24	0.48	-0.08	0.77
-CH ₂ F	-H	0.52	0.39	-0.45	0.40
-CN	-H	0.08	0.36	-0.58	0.27
-H	-CN	0.08	0.23	-1.21	-0.36

^b The ΔpIC_{50} for these substituents was not reported in article [144].

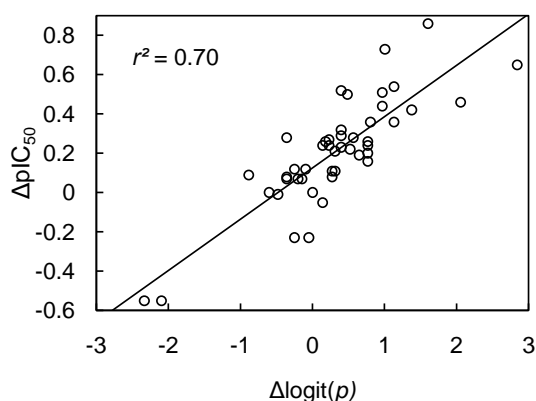


Figure 3.1: The correlation between published changes in average pIC_{50} (ΔpIC_{50}) [144] and changes in values predicted by the baseline model ($\Delta \text{logit}(p)$).

Data on HMG-CoA reductase inhibitors also show that, while in acidic form, these compounds show no activity towards CYP3A4, whereas their lactone form inhibits this enzyme [145]. Similar effect is observed in a novel series of arginine vasopressin V2 receptor agonists. The compound with carboxy group does not inhibit CYP3A4 while others do [146]. Introduction of a methyl ester shows positive impact on CYP3A4 inhibition both analyzing experimental data and predictions of our model (p in the range from 0.34 to 0.48 for esters, corresponding $\Delta \text{logit}(p)$ in the range between 0.18 and 0.77). Therefore it may be concluded that the negative charge of an acidic group is the most significant property reducing the interaction between inhibitor and CYP3A4.

Introduction of a strong basic group (e. g., aliphatic amine) slightly decreases the probability for a compound to inhibit CYP3A4, but the effect is nowhere comparable to that of an acidic group (p varies between 0.15 and 0.23, corresponding $\Delta \text{logit}(p)$ in the range of -0.89 to -0.36). Similarly, average inhibition potency for basic compounds has been found to decrease slightly compared with neutral compounds in large datasets of CYP3A4 inhibitors analyzed in other studies [141].

Negative impact of various hydrophilic groups (hydroxyl or amide, p ranges between 0.21 and 0.33, $\Delta \text{logit}(p)$ varies within -0.48 and -0.05 with an exception for hydroxymethyl substituent having $\Delta \text{logit}(p) = 0.14$) is possibly related to the decrease of lipophilicity of the compound. On the contrary, increasing the

size of the molecule with hydrophobic aliphatic or aromatic residue results in the rise of the probability for a compound to inhibit CYP3A4 (p in the range 0.36 to 0.97, $\Delta \logit(p)$ varies between 0.27 and 4.32). The majority of previously reported CYP3A4 specificity models highlighted the importance of multiple hydrophobic groups for effective interaction with this enzyme [74]. This characteristic dependency of CYP3A4 inhibition potency on molecular weight and lipophilicity was as well observed in a recent analysis of a large amount of CYP3A4 inhibition data [141]. The significance of hydrophobic interactions for effective CYP3A4 inhibition might be expected from the known crystal structure of this enzyme, in which a phenylalanine cluster plays an important role in defining the CYP3A4 active site [41]. This particular fact explains why aromatic substituents inflict a larger influence on predicted probabilities compared to an aliphatic chain (probabilities 0.36 to 0.57 for aliphatic chains and 0.57 to 0.88 for aromatic rings). The introduction of nitrogen containing aromatic rings increases the probability to inhibit enzyme even more (predicted probability in the range from 0.68 to 0.97). Possession of the nitrogen-containing heterocyclic moieties (such as imidazole, quinoline, pyrimidine) gives the compound a possibility to form complexes with the heme iron inside cytochrome P450 enzymes [53].

Methoxy and ethoxy groups ($R-OCH_3$, $R-OC_2H_5$) connected to the aromatic ring increases probability to inhibit CYP3A4 (p varies from 0.37 to 0.42, $\Delta \logit(p)$ ranges from 0.32 to 0.52). Methoxy group connected to aromatic ring is a possible site of CYP3A4 mediated metabolism [15]. Substituents of this type increase the probability for a compound to become a mechanism-based CYP3A4 inhibitor. One of the biggest impacts on predicted CYP3A4 inhibition probability was observed following the addition of a methylenedioxybenzene substituent ($p = 0.9$, $\Delta \logit(p) = 3.04$). This group, well known for its relevance in mechanism-based inhibition, is frequent among cytochrome P450 inhibitors [15, 53].

Although predictive models described here are not able to distinguish mechanism-based inhibitors from competitive ones, this shortcoming is not likely to be important in practical applications. Many of the known clinically relevant inhibitors like azole type drugs interact with CYP3A4 as mixed type inhibitors – both competitive and mechanism-based [147]. High-throughput screening experiments used in determination of cytochrome P450 inhibitors cannot identify mechanism-based inhibition as well.

3.2 Local Model

The global model described above successfully reflects the general trends of CYP3A4 inhibition. However, the sensitivity of the model is lower than specificity (75% in the internal test set and 61% in the external test set, see Table 3.1). This can be attributed to the features of some CYP3A4 inhibitors which could not be described using linear model. Under these circumstances a method is needed that is able to account for possible nonlinear effects. Local correction of the baseline model predictions according to experimental data for the most similar compounds (a second layer of the GALAS model) has been developed to deal with this kind of problems. Moreover, this routine allows estimation of the prediction reliability in the form of calculated Reliability Index (RI), as well as training of the model using new experimental data.

Detailed results of the validation of the GALAS model for CYP3A4 inhibition are presented in Table 3.3. Predictions having $RI < 0.3$ fall outside the applicability domain of the model [14] and are not considered here. This affects only a small fraction of the internal test set. The majority (73%) of its compounds obtain predictions of acceptable reliability. The statistical parameters of the GALAS model for such compounds are superior compared to those of the baseline model (overall accuracy 89%, sensitivity 83%, specificity 93%). This is actually greater than classification consistency between the datasets compiled from literature sources and PubChem database (see Table 2.2). When considering only predictions with high reliability ($RI > 0.5$), the accuracy increases to 95%, with both sensitivity and specificity being higher than 90%. These values approach the accuracy of experimental measurements that were observed in the analysis of compounds having multiple experimental values reported in the PubChem database. Such results serve as the first confirmation that the employed Reliability Index calculation methodology effectively assesses the quality of prediction and defines the applicability domain of the model.

The benefits of correcting baseline predictions according to experimental data for similar compounds can be seen while analyzing receiver operating characteristic (ROC) curves for the internal test set (Fig. 3.2). These graphs show the dependence between false positive rate (1 - specificity) and true positive rate (sensitivity) as the discrimination threshold of the model is varied [131]. The

Table 3.3: The final results of the GALAS model of CYP3A4 inhibition for the internal and external test set compounds falling within the model applicability domain ($RI > 0.3$) and obtaining high reliability predictions ($RI > 0.5$).

		External test set (8528 compounds)					
		Internal test set (190 compounds)			External test set (8528 compounds)		
		Pred. True	Pred. False		Pred. True	Pred. False	
$RI > 0.3$	Obs. True	43	9	Sensitivity: 82.7%	Obs. True	214	Sensitivity: 69.0%
	Obs. False	6	80	Specificity: 93.0%	Obs. False	2,638	Specificity: 93.3%
	% of compounds within RI range: 72.6%				% of compounds within RI range: 41.3%		Accuracy: 88.5%
$RI > 0.5$	Obs. True	30	3	Sensitivity: 90.9%	Obs. True	59	Sensitivity: 65.1%
	Obs. False	1	41	Specificity: 97.6%	Obs. False	1,012	Specificity: 98.2%
	% of compounds within RI range: 39.1%				% of compounds within RI range: 14.1%		Accuracy: 93.5%

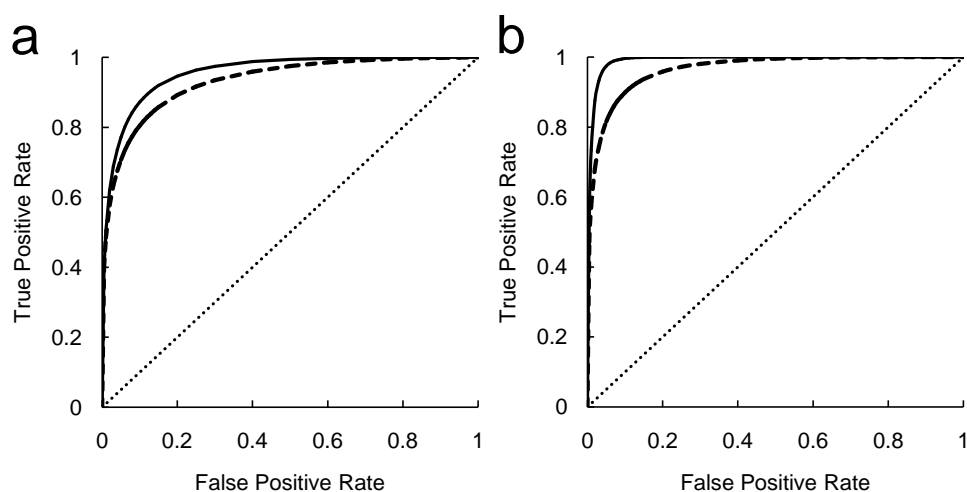


Figure 3.2: ROC curves comparing the performance of global and similarity corrected models on the internal test set: a – compounds within the applicability domain of the model ($RI > 0.3$); b – compounds obtaining high reliability predictions ($RI > 0.5$). Dotted line indicates random classifier, dashed line – baseline model, solid line – similarity corrected model.

point (0;1) corresponds to perfect classification. The closer to this point gets the curve, the better is the model. As it can be seen from the ROC curves for compounds in the applicability domain of the model (Fig. 3.2, a), local modeling shifts the baseline predictions towards better classification. The influence of similarity correction is bigger for compounds obtaining high reliability predictions (Fig. 3.2, b). This can be explained by the fact that significant Δ values are calculated only in the cases when highly similar compounds with consistent experimental data are present in the training library.

Comparison of the GALAS model results for internal test set compounds within different RI ranges reveal the improvement of all the statistical parameters with the increase of RI values (see Fig. 3.3). Similar trends are observed when testing the model on the external set – PubChem dataset. For compounds that belong to the applicability domain of the model ($RI > 0.3$) the statistical results are also better than those of the baseline model (Table 3.3 vs. Table 3.1). The accuracy of predictions is 89% with acceptable sensitivity (69%) and very good specificity (93%). Overall prediction accuracy is even better (93%) for compounds having high Reliability Index ($RI > 0.5$). However, only less than half (41%) of the

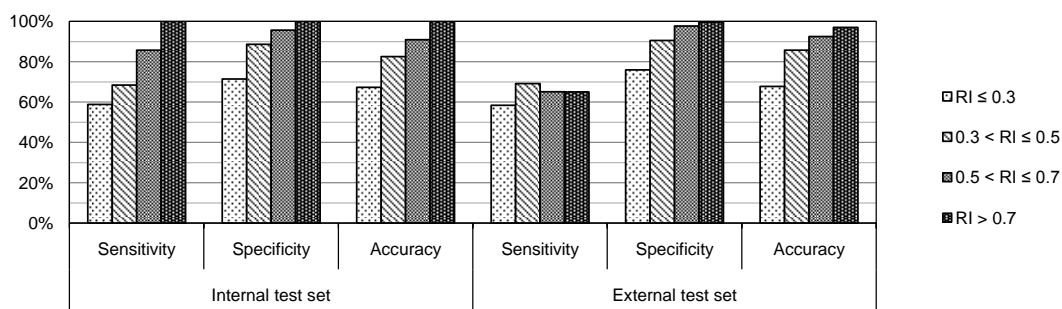


Figure 3.3: The statistical parameters of GALAS models for internal and external test set compounds classified according to the Reliability Index values.

PubChem dataset actually belongs to the applicability domain of the model, and only a small fraction of compounds (14%) obtain high reliability predictions. In situations like this, the possibility to expand the applicability domain of the model would be of the utmost importance.

3.3 Training of the GALAS Model

The similarity correction procedure described in the methodological part plays the major role in the ability to train the model with external data. In addition to the original training library, initially containing the same training set compounds used in the development of a baseline model, new compounds can be added allowing the model to cover a novel, previously unknown, part of the chemical space.

The above described procedure does not introduce large changes if the baseline predictions are already correct. Yet if a query molecule representing a part of the chemical space just added to the model is mispredicted by the global model which remains unchanged, the similar compounds just added to the similarity correction library effectively correct this baseline prediction based on the newly available experimental data.

3.3.1 Training with Data from a Similar Assay

PubChem dataset was used to demonstrate how training of the model is able to improve the predictions for new compounds. A randomly selected part of this database was used as a validation set for trainability testing, while remaining compounds were added in several portions to form a new training library of the model, as described in Section 2.5. The results for models containing different fractions of the PubChem dataset in a training library are presented in Fig. 3.4.

The baseline model predicts CYP3A4 inhibition specificity for the trainability validation part of the PubChem dataset with overall accuracy of 76%, sensitivity of 59%, and specificity of 85% (see Fig. 3.4, a). These parameters are comparable to the baseline model performance on the whole PubChem dataset. Addition of just a small part of the remaining dataset (ca. 200 compounds) as the training library markedly improves the statistical performance (accuracy 79%, sensitivity 65%, specificity 87%), whereas further expansion of this library with PubChem compounds keeps increasing the values of statistical parameters of the model (see Fig. 3.4, a), reaching a maximum accuracy of 86%, sensitivity of 77%, and specificity of 90% when > 4,000 compounds are added to the library.

Notably, in all cases the specificity of the model was significantly better than sensitivity. This means that a compound dissimilar to the training set is always more likely to be classified as a non-inhibitor. The small number of correctly identified inhibitors can be named as a major drawback of the baseline model. This number noticeably increases with the addition of new compounds to the training library. Just 200 compounds added result in 35% of inhibitors being identified with $RI > 0.3$, and this percent increases to 46% when 400 compounds are added (see Fig. 3.4, b). The fraction of inhibitors identified with high reliability ($RI > 0.5$) increases approximately twofold comparing the results for 200 and 400 compounds added to the library (13% and 25% respectively). Even more inhibitors can be identified if larger training libraries are used because of expanding of the model applicability domain to cover more compound structural classes. Following the addition of all available (> 4,000) PubChem compounds to the training library 50% of CYP3A4 inhibitors are found with high reliability of prediction.

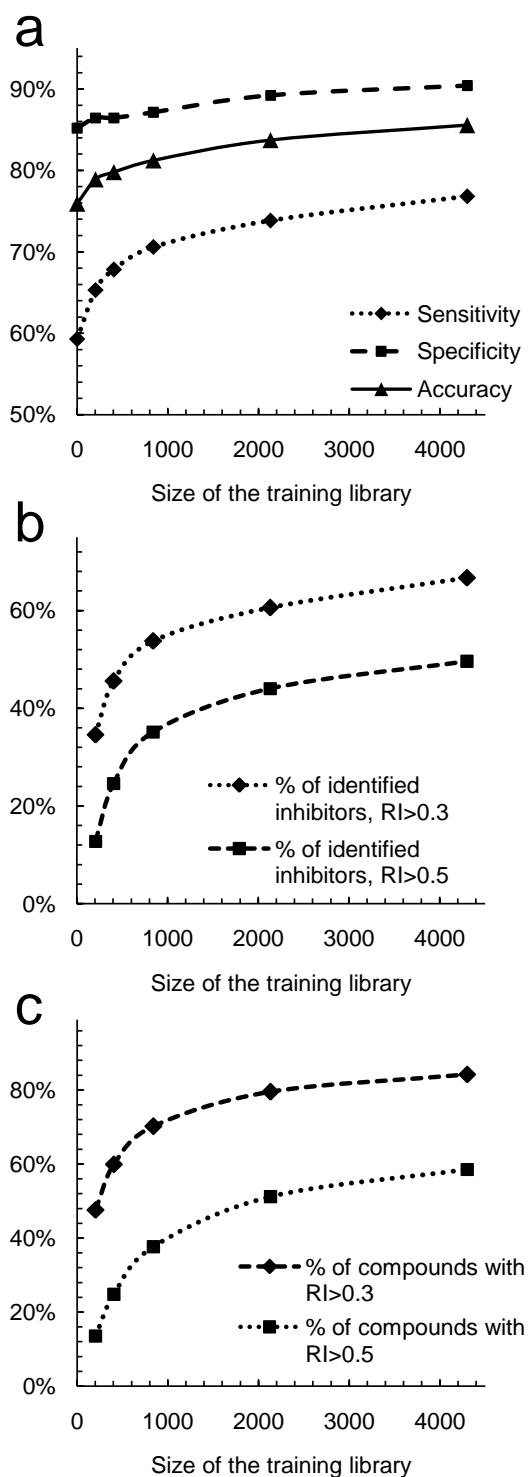


Figure 3.4: Training of the GALAS model using data from a similar assay: a – statistical parameters of the model; b – percentage of identified inhibitors; c – percentage of compounds predicted with $RI > 0.3$ and $RI > 0.5$.

In order to directly illustrate the actual expansion of the GALAS model applicability domain as a result of new compounds added to the training library; let us consider the following results. Only 48% of the trainability validation set belongs to the applicability domain ($RI > 0.3$) of the model trained with just 200 compounds. The number of acceptable reliability predictions increases to 60% after adding another 200 compounds (400 in total) (see Fig. 3.4, c). Further enlargement of the training library eventually results in applicability domain of the model covering 89% of the test set at > 4,000 compounds added (ca. twofold increase in total). The relative growth in the number of high reliability predictions ($RI > 0.5$) is even bigger, starting from 14% at 200 compounds added, and reaching 59% at > 4,000 compounds added (ca. four times more). These results definitely show the adaptation of the model, based on the literature data, to the part of the chemical space occupied by the PubChem dataset compounds.

3.3.2 Training with Data from an Assay with a Different Potency Threshold

Following the successful training of CYP3A4 inhibition model with data from a similar experimental assay classified using the same IC_{50} threshold, a natural question arises: can the trainable model described above be adapted for the discrimination between CYP3A4 inhibitors and non-inhibitors identified using a different scale? In other words, do the binary data (inhibitor/non-inhibitor) used in model training necessarily have to be obtained using exact same criteria as for the training set, or data produced by any screening programs can be used to train the model, no matter what experimental protocol or inhibition potency threshold was used to classify compounds in terms of CYP3A4 inhibition? To test this possibility, the initial model was trained with the PubChem dataset classified using significantly different inhibition potency threshold ($IC_{50} = 5 \mu M$ versus $IC_{50} = 40 \mu M$ used to classify the training set data). The results for the trained models obtained during this experiment are presented in Fig. 3.5.

The first look at the results reveals that initially many false positive predictions are observed. Using the baseline model for the classification of the PubChem validation set, the positive predictive value is only 46%, i. e. only about a half of compounds predicted as active are indeed experimentally determined as

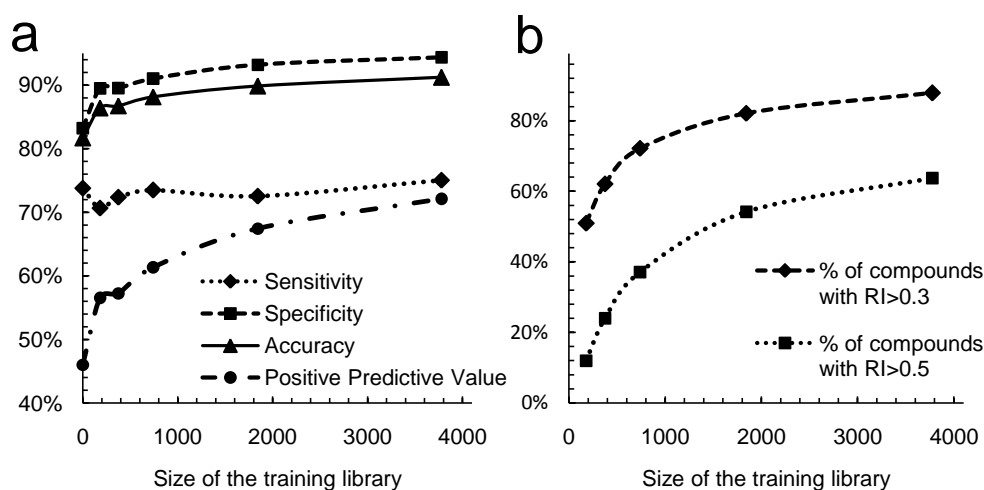


Figure 3.5: Training of the GALAS model using data from an assay with different potency threshold: a – statistical parameters of the model; b – percentage of compounds predicted with $RI > 0.3$ and $RI > 0.5$.

effective CYP3A4 inhibitors. This is an inevitable consequence of the differences in classification thresholds used. Percent of false positive predictions decreases and positive predictive value of the model increases following the addition of the new compounds from the PubChem set to the training library (see Fig. 3.5, a). Improvements in other statistical parameters are observed as well (Fig. 3.5, a). The fraction of reliable and high reliability predictions depends on the number of new compounds added to the library analogously as observed in the previous training example (Fig. 3.5, b). These results demonstrate the ability of the GALAS modeling methodology, employed in this work, to handle data from different types of experimental CYP3A4 inhibition studies.

3.3.3 Training with Compounds from a New Structural Class

The above examples show the adaptation of the model to experimental data obtained using a different method or even a different classification threshold. Obviously, having such a large library with experimental data for new compounds (exceeding the original training set several times) is a reasonable argument for the revision of the model. Still a simple statistical re-parameterization of the existing model not necessarily produces better results than the described

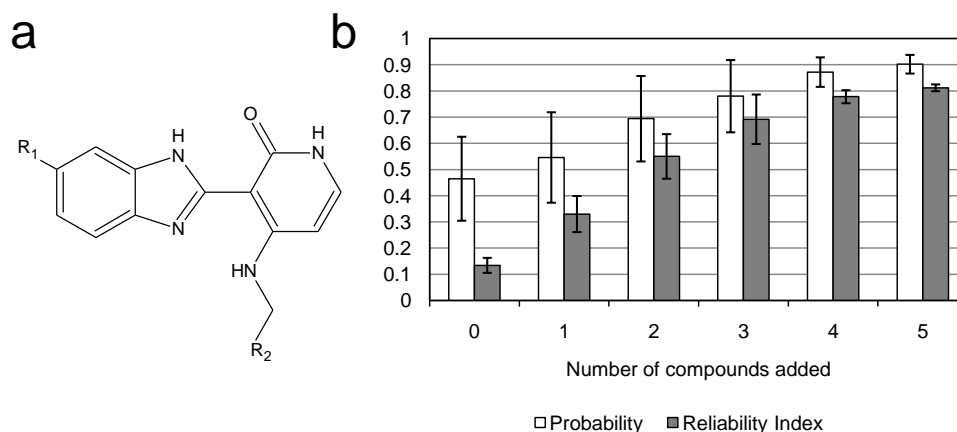


Figure 3.6: Training of the GALAS model with compounds belonging to a novel chemical class: a – general scaffold of compounds (R_1 – aliphatic heterocycle, R_2 – aromatic substituent); b – increase in average predicted probability and *RI* values (error bars indicate standard deviation).

training procedure. In contrast, full remodeling is a long process, involving a more in-depth analysis of data, calculation of new descriptors, etc. When new experimental data are obtained for only a few compounds belonging to a novel chemical class, it is usually desirable to avoid these time consuming procedures.

We have decided to test the trainability feature with recently published inhibitors of insulin-like growth factor-1 receptor (IGF-1R) as a model of such a novel drug class [132]. The common scaffold of these compounds is not present among the molecules in the Literature and PubChem datasets and is shown in Fig. 3.6, a. According to the classification rules used in this article, all 10 published IGF-1R inhibitors are CYP3A4 inhibitors having $IC_{50} < 40 \mu M$. Five randomly selected compounds were added one by one to the similarity correction library of the model initially containing literature dataset. The changes of average predicted probability and average Reliability Index values for the remaining five compounds are shown in Fig. 3.6, b.

The original model (0 compounds added) predicts three compounds as non-inhibitors with probabilities ranging from 0.24 to 0.39, and two inconclusively with probabilities around 0.5. Average probability is 0.46 for all five compounds. The *RI*s are low for all predictions with average of 0.13 (values range from 0.09 to 0.16), indicating that novel IGF-1R inhibitors fall outside the applicability

domain of original CYP3A4 inhibition model. As expected, after adding similar compounds to the training library both predicted probabilities and *RIs* increase, indicating successful model training. In this case adding three compounds is enough to adapt the model to the novel class of IGF-1R inhibitors as calculated probabilities to inhibit CYP3A4 and corresponding *RI* values become higher than 0.5 for all test compounds.

Recently, the importance of the applicability domain evaluation in QSAR modeling was illustrated in the study involving models for the prediction of plasma protein binding and CYP3A4 inhibition generated using *GlaxoSmithKline* in-house data [76]. Using internal test set the average error of prediction of CYP3A4 inhibition potency (expressed as IC_{50}) was close to experimental variability. Switching to the external validation set significantly increased the error which kept steadily growing following each subsequent expansion of this set with newly available data measured after the model was built. The bigger was the time gap between model development and new CYP3A4 inhibition measurements, the larger prediction errors were observed [76]. Such time dependence is a direct consequence of structural diversity changes in the in-house databases of experimentally tested compounds. Training the model using above described methodology would be useful in keeping the models up to date and adjusting them to the parts of chemical space of researcher's interest. While full revisions of QSAR models are inevitable from time to time having a substantial amount of new experimental data, the proposed training procedure of the GALAS model can be used on a daily basis.

Chapter 4

Results and Discussion: Regioselectivity Modeling

Regioselectivity of metabolism is not a typical property for structure-activity relationship analysis. A QSAR model usually predicts whole molecule properties like *LogP*, toxicity or enzyme inhibition. The model of metabolism regioselectivity predicts the possibility of reaction for every atom in a molecule. Moreover, different reactions – hydroxylation of aliphatic and aromatic carbon as well as heteroatom oxidation – have distinct reaction mechanisms. Therefore building a QSAR model for prediction of regioselectivity is a complicated task. First of all the atoms of modeling dataset compounds were marked whether they are metabolism sites or not. In order to handle the difficulties with reaction mechanisms the whole database of atoms has been split into subsets which have the same types of reactions [15]. Then 5 separate models were developed for N-dealkylation, O-dealkylation, aromatic and aliphatic hydroxylation, and S-oxidation. Atom-centered fragmentation was used to describe atoms and their neighborhood.

The developed regioselectivity model was initially validated using an internal test set, which contained 30% of atoms of the modeling dataset. Then predictions were obtained for 42 compounds that were not used in modeling on purpose to see what is the model performance in real-life application on drug candidates. After this thorough validation, the predictions were compared to those of *SMARTCyp* software. This program was chosen because it is also a

Table 4.1: The performance of the regioselectivity model on the initial test sets constituting 30% of modeling dataset atoms.

		Pred. True	Pred. False	
Baseline model	Obs. True	273	128	Sensitivity: 68.1% Specificity: 80.7% Accuracy: 78.8%
	Obs. False	420	1759	
After local modeling	Obs. True	260	141	Sensitivity: 64.8% Specificity: 93.9% Accuracy: 89.4%
	Obs. False	133	2046	
After local modeling, $RI > 0.3$	Obs. True	203	97	Sensitivity: 67.7% Specificity: 96.3% Accuracy: 92.5%
	Obs. False	71	1864	
After local modeling, $RI > 0.5$	Obs. True	119	34	Sensitivity: 77.8% Specificity: 98.4% Accuracy: 96.2%
	Obs. False	22	1310	

ligand-based model but uses a completely different computational method – quantum chemistry calculations. Finally, the trainability features of GALAS model were tested.

4.1 Internal Validation of the Model

The results of initial model validation are presented in Table 4.1. The test sets of all reactions are joined into one dataset in this table. The performance of baseline models is already good, 68% of metabolism sites predicted with probability > 0.5 , and 79% of total atoms classified correctly.

Table 4.2 shows the descriptors having largest coefficients in the aliphatic and aromatic hydroxylation models. The largest positive influence for aliphatic hydroxylation comes from the sp^2 -hybridized aliphatic or aromatic system next to the metabolism site. Oxygen and nitrogen atoms in the second level have largest negative contribution, probably because the reaction is more likely to occur right next to these heteroatoms, resulting in O- or N-dealkylation. The large positive increments for methyl groups show that secondary carbon atom in $\omega-1$ position and tertiary atom in isopropyl group are more susceptible to

Table 4.2: The descriptors having largest coefficients in aliphatic and aromatic hydroxylation models.

Fragment	Level	Coefficient	% of metabolism sites ^a	Examples of metabolism sites ^b
<i>Aliphatic hydroxylation:</i>				
CH ₃	1	0.45	20% (38/189)	
Csp ² (aliphatic or aromatic)	1	0.38	24% (156/644)	
Csp ² (aliphatic)	1	0.23	21% (80/385)	
Csp ² or sp (aliphatic)	1	0.22	21% (81/391)	
CH ₂ (cyclic)	0	0.22	12% (67/552)	
C, connected to Csp ²	0	0.20	24% (155/638)	
CH ₃	2	0.20	13% (65/358)	
N	2	-0.23	5% (29/543)	
O	2	-0.26	5% (22/441)	
<i>Aromatic hydroxylation:</i>				
OH	2	0.49	30% (24/81)	
<i>meta</i> -substituted ring	1	0.43	11% (231/2040)	
<i>ortho</i> -substituted ring	2	0.42	11% (233/2061)	
<i>para</i> -substituted ring	0	0.33	20% (177/902)	
CH (aromatic)	1	0.31	12% (233/1947)	
N (aromatic)	3	-0.32	6% (17/263)	
Csp ² or sp (aliphatic)	2	-0.33	1% (2/258)	
C (aliphatic)	4	-0.37	6% (77/1308)	
O	3	-0.52	3% (15/491)	

^a Number of metabolism sites having the fragment in corresponding level divided by number of atoms having this fragment.

^b The metabolism site (marked atom) for which the fragment is found is circled, R means any structure.

oxidation than primary carbon. These facts are consistent with the results of *ab initio* quantum chemistry calculations [89]. In addition to vinylic, benzylic and ω -1 metabolism sites, positive influence of methyl groups in level 2 leads to prediction of oxidation in tertiary butyl groups, which are typical cytochrome P450 metabolism sites [15].

In case of aromatic hydroxylation carbon atom having a hydrogen next to the oxidation site has a positive influence. This fact is in agreement with the reaction mechanism [15, 16]. As it can be seen from other descriptors having largest positive increments, the metabolism of monosubstituted benzene is predicted in *para*-position. Quantum chemistry models estimate oxidation in *ortho*- and *para*-positions [90, 105], but our baseline model tends to predict *ortho*-hydroxylation only in case of phenols, as it can be seen from large coefficient for OH group in level 2. This empirical observation and most of negative influences for aromatic hydroxylations are harder to explain using published quantum chemistry data. Thus the baseline models can be viewed as description of the reactivity with some empirical corrections.

Further advancement of the model is needed because of the high number of false positives among baseline predictions (about 16% of all atoms). This situation noticeably changes after local similarity corrections (see Table 4.1). Slight decrease of sensitivity to 65% is compensated by large increase in specificity to 94%, indicating the diminishing of false positives. The overall accuracy increases to 89%. The influence of local modeling can be nicely visualized using the ROC curves (Fig. 4.1). Improvement of classification after similarity based corrections is observed for all reactions.

Another important feature of the local similarity correction in the GALAS model is the possibility to calculate the Reliability Index. This value takes into account the presence and consistence of experimental data for similar atoms. When considering only predictions with higher values of Reliability Index, all the statistical parameters improve (Table 4.1). Superior classification among predictions of higher reliability can be also seen from the shift of ROC curves towards the point of perfect classification (Fig. 4.1). These results demonstrate that *RI* effectively identifies correct predictions in a similar manner that was observed in CYP3A4 inhibition modeling (Chapter 3).

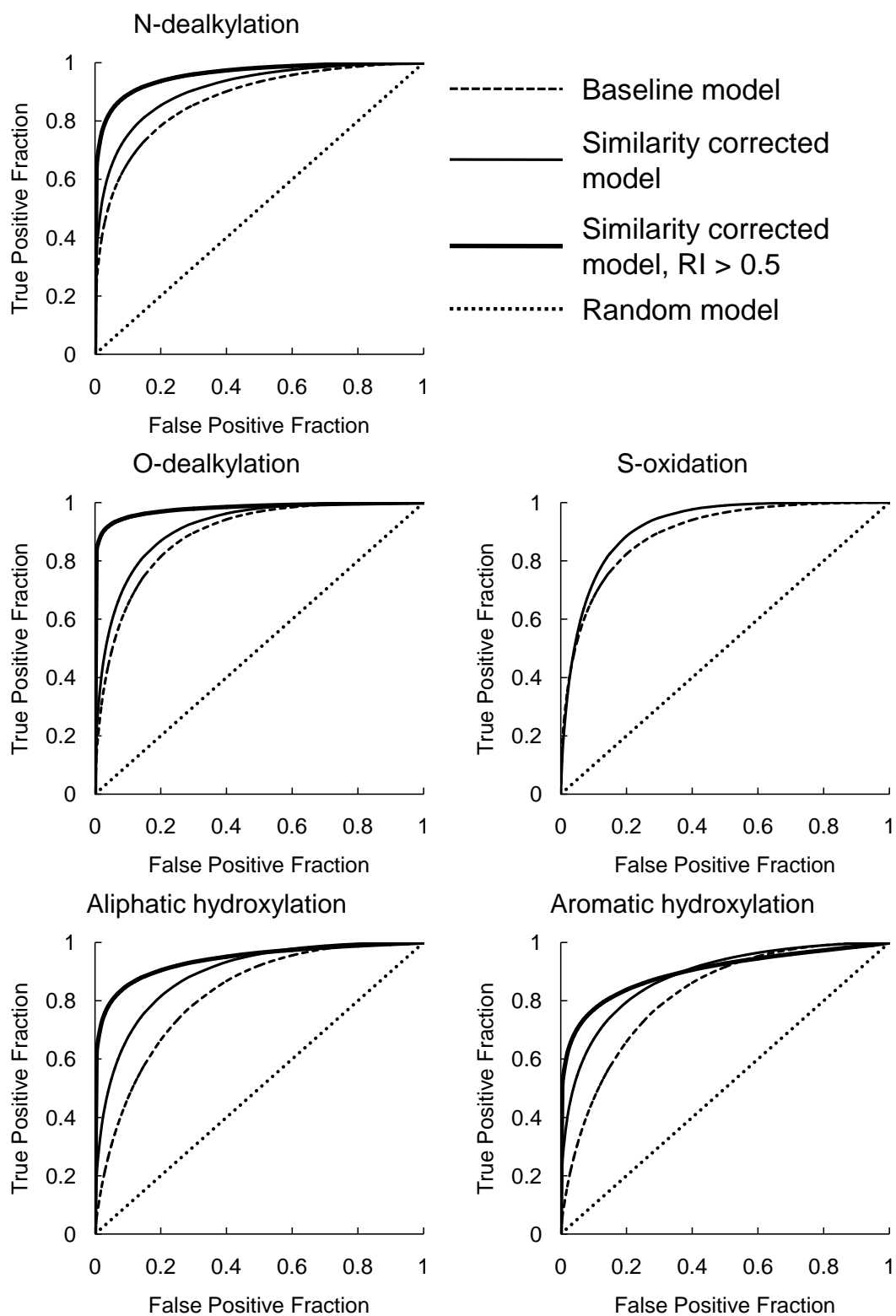


Figure 4.1: ROC curves comparing the performance of baseline and similarity corrected regioselectivity models for different reaction types.

4.2 External Validation of the Model

After obtaining acceptable results for the internal test sets, the model was evaluated on an external validation set of 42 compounds that were not used in modeling. The results show good agreement between predictions and experimental data. The experimentally observed metabolism site obtained highest probability for 30 compounds (71%), and 35 compounds (83%) had at least one experimental site among three top ranked atoms. 43 of 73 metabolism sites are predicted by the model with probability > 0.5 (sensitivity 59%).

Most of the previously published regioselectivity prediction evaluation studies conducted either by model developers or by independent researchers [91, 97, 99, 100, 122] focus only on a few top ranked metabolism sites. However, this approach is limited. Top ranked atom is obviously the most probable place for oxidation in the molecule, but its calculated score (probability of reaction, activation energy, etc.) can still be negligible, indicating that actually the model suggests this metabolism site as nonsignificant and the whole compound as not metabolized. Therefore a new approach for analyzing the predictions is needed which examines not only ranking of atoms according to calculated possibility of metabolism, but also the values of calculated metabolism scores.

Such method was used in evaluation of the developed model. The probability of being a metabolism site was calculated for every atom of 42 external test set compounds. Analyzing the results for individual compounds, the predictions were divided according to their quality into four classes: "excellent", "good", "satisfactory", and "unsatisfactory". Brief descriptions for these classes are as follows:

- Excellent: all metabolism sites predicted;
- Good: most metabolism sites predicted;
- Satisfactory: some metabolism sites predicted;
- Unsatisfactory: no metabolism sites predicted.

The detailed definitions of the quality classes are given in the Chapter 2, Section 2.5.2. The whole list of compounds with description of the prediction quality

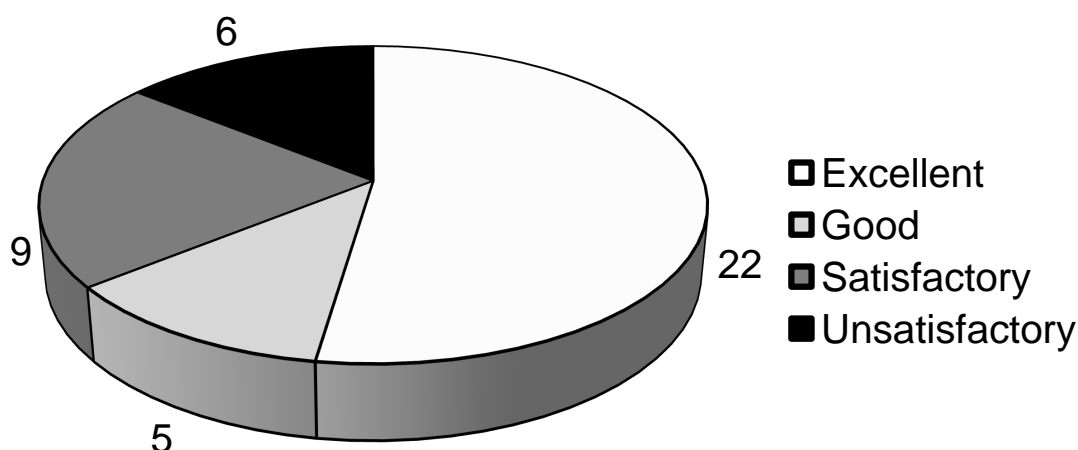


Figure 4.2: The distribution of regioselectivity predictions for external test set according to their quality.

is in Table 4.3, and the detailed predictions can be found in Appendix A. Fig. 4.2 shows the distribution of the external test set compounds according to the quality of regioselectivity predictions.

Most metabolism sites were predicted for more than 60% of the compounds in the validation set (22 “excellent” and 5 “good” predictions). As it can be seen in the Table 4.3, all types of reactions are represented among these predictions. Fig. 4.3 shows compounds predicted “excellent” that are from well-known drug classes, such as steroids (medroxyprogesterone acetate), opioids (dimemorfan), PDE5 inhibitors (vardenafil), which are well represented in the training sets. The majority of atoms in these compounds obtained high reliability predictions.

Fig. 4.4 shows examples of “excellent” and “good” predictions for compounds belonging to novel drug classes. Some of them have only one experimentally determined metabolism site which is ranked as one of three most probable atoms (NSC 639829, harmine, bepridil, tadalafil). The false positive predictions might be minor metabolites that were not identified experimentally. The opposite is observed for TSU-68, carbosulfan, and voreloxin: one of the experimental sites is not predicted. The two major metabolites are predicted for TSU-68. Interestingly, in case of voreloxin, the O-dealkylation site is ranked second having probability 0.37. The S-oxidation site of carbosulfan is ranked 4th with probability of 0.23.

Table 4.3: List of external validation set compounds with the evaluation of prediction.

Name	Main reactions	No. of exp. sites	No. of pred. sites	Prediction evaluation
BDB [148]	O-dealkylation	1	1	Excellent
Bepiridil [122]	Aromatic hydroxylation	1	2	Excellent
<i>p</i> -Cresol [149]	Hydroxylation	2	2	Excellent
Dimemorfan [150]	N-dealkylation, aliphatic hydroxylation	2	2	Excellent
Gemfibrozil glucuronide [151]	Hydroxylation	3	3	Excellent
Harmaline [152]	O-dealkylation	1	1	Excellent
Medroxyprogesterone acetate [153]	Aliphatic hydroxylation	3	3	Excellent
5-Methoxy-N,N-dimethyltryptamine [154]	O-dealkylation	1	2	Excellent
Methyleugenol [155, 156]	Aliphatic hydroxylation	1	1	Excellent
2-Nitroanisole [157]	O-dealkylation	1	1	Excellent
Osthol [158]	O-dealkylation, aliphatic hydroxylation	2	2	Excellent
Pefloxacin [122]	N-dealkylation	1	1	Excellent
Pinoline [159]	O-dealkylation	1	1	Excellent
R-125528 (pactimibe metabolite) [160]	Aliphatic hydroxylation	1	1	Excellent
Sibutramine [161]	N-dealkylation	1	1	Excellent
Sibutramine metabolite M1 [161]	N-dealkylation	1	1	Excellent
SKF 525A [162]	N-dealkylation	1	1	Excellent
SKF8742 [162]	N-dealkylation	1	1	Excellent
Tadalafil [122]	N-dealkylation	1	4	Excellent
Trabectedin [163]	Dealkylation, hydroxylation	5	4	Excellent
Trichostatin A [122]	N-dealkylation	1	1	Excellent
Vardenafil [164]	N-dealkylation	1	1	Excellent

Table 4.3: List of external validation set compounds with the evaluation of prediction. (continued)

Name	Main reactions	No. of exp. sites	No. of pred. sites	Prediction evaluation
Carbosulfan [165, 166]	Aliphatic hydroxylation, S-oxidation	2	1	Good
Harmine [152]	O-dealkylation	1	3	Good
3-Hydroxycarbamazepine [167]	Aromatic hydroxylation	1	3	Good
NSC 639829 [168]	N-dealkylation	1	3	Good
TSU-68 [169]	Aromatic hydroxylation	3	2	Good
Voreloxin [170]	Dealkylation	2	1	Good
CP-533,536 [171]	N-dealkylation, hydroxylation	2	2	Satisfactory
Dasatinib [172]	Hydroxylation	3	2	Satisfactory
Fluticasone [173]	S-oxidation	1	1	Satisfactory
2-Hydroxycarbamazepine [167]	Aromatic hydroxylation	1	2	Satisfactory
Pactimibe [160]	Hydroxylation	2	2	Satisfactory
Piperacillin [174]	N-dealkylation	1	0	Satisfactory
Tanespimycin [175]	Dealkylation, aliphatic hydroxylation	4	1	Satisfactory
TZB-30878 [176]	Hydroxylation, N-dealkylation	5	3	Satisfactory
Chenodeoxycholic acid [177]	Aliphatic hydroxylation	2	0	Unsatisfactory
Cholic acid [177]	Aliphatic hydroxylation	1	0	Unsatisfactory
Flu-1 [178, 179]	Aromatic hydroxylation	1	0	Unsatisfactory
Sanguinarine [180]	Aromatic hydroxylation	2	0	Unsatisfactory
Zearalenol [181]	Aromatic hydroxylation	2	0	Unsatisfactory
Zearalenone [182]	Aromatic hydroxylation	2	0	Unsatisfactory

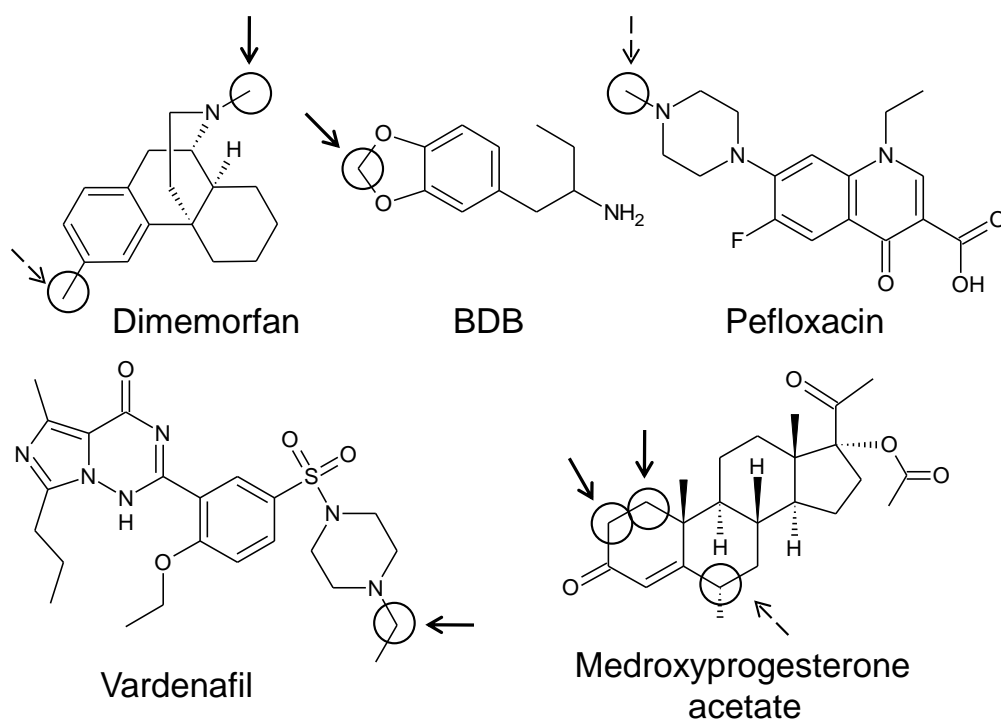


Figure 4.3: “Excellent” predictions for drugs of well-known classes. Experimental metabolism sites are circled, solid arrows indicate confidently predicted metabolism sites (probability > 0.5 and *RI* > 0.5), dashed arrows show predicted sites with *RI* < 0.5.

Fig. 4.5 shows two examples of these compounds that obtained best of “satisfactory” predictions. In case of pactimibe, the major metabolism site – pyrrolidine ring oxidation – is not predicted, but site of ω -1 aliphatic hydroxylation obtained probability of 0.52. N-dealkylation that is estimated by the model is in fact a minor hardly detectable pathway in human liver microsomes [160]. The sulfur atom of fluticasone is ranked second and obtained probability of 0.43, close to the threshold value. Similar situation is observed for piperacillin, where N-dealkylation site is ranked highest with probability of 0.44. Many false positive metabolites are notable among other “satisfactory” predictions. It is important to note that most of them obtained low Reliability Indices. Still, some of the experimentally determined metabolism sites are identified correctly by the model.

No metabolism sites were predicted for only 6 compounds (“unsatisfactory” results). Five of them are natural compounds, such as bile acids or secondary

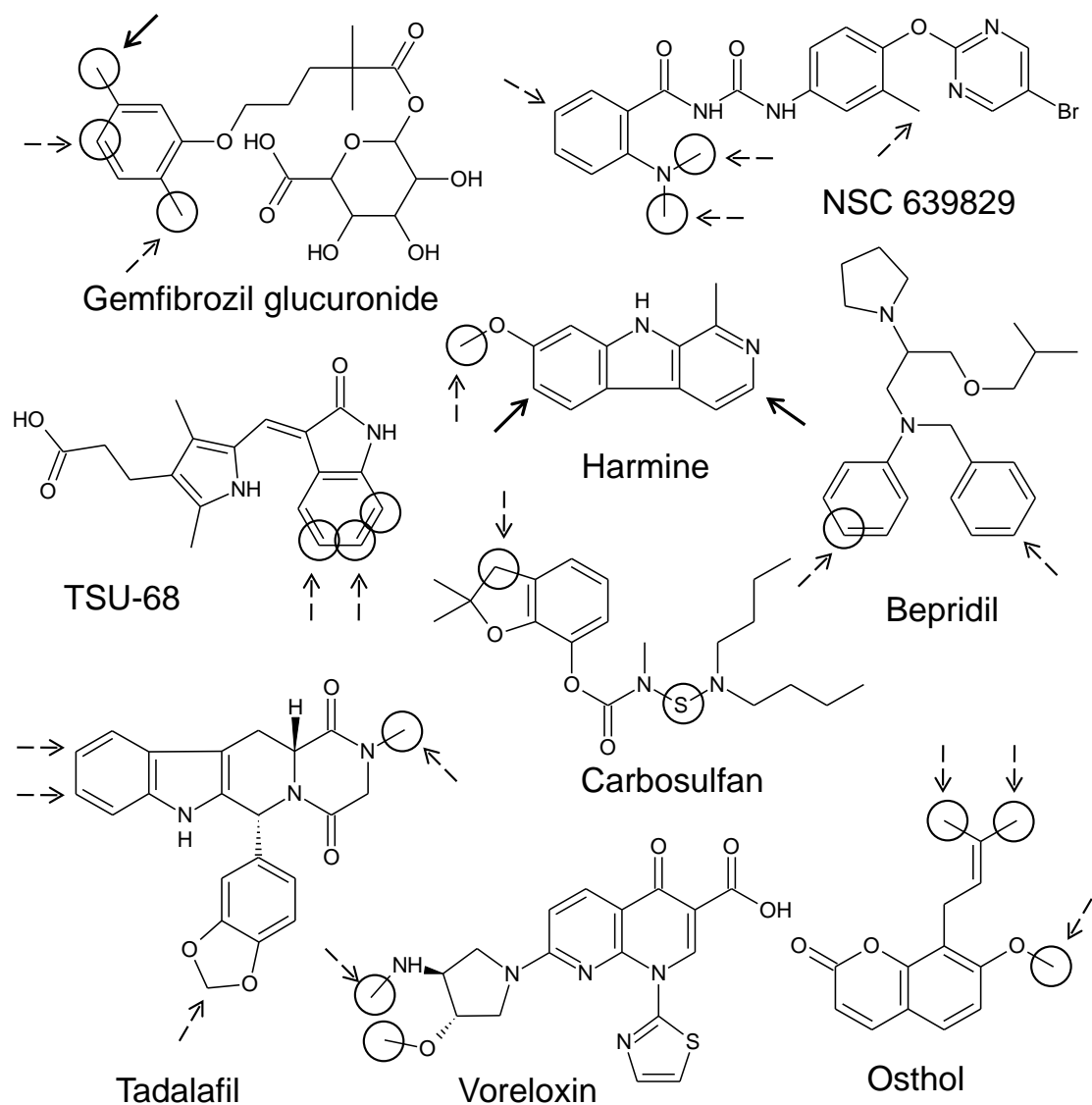


Figure 4.4: Examples of “excellent” and “good” predictions. Experimental metabolism sites are circled, arrows indicate atoms predicted with probability > 0.5 , solid arrows show atoms with $RI > 0.5$, dashed arrows show atoms with $RI < 0.5$.

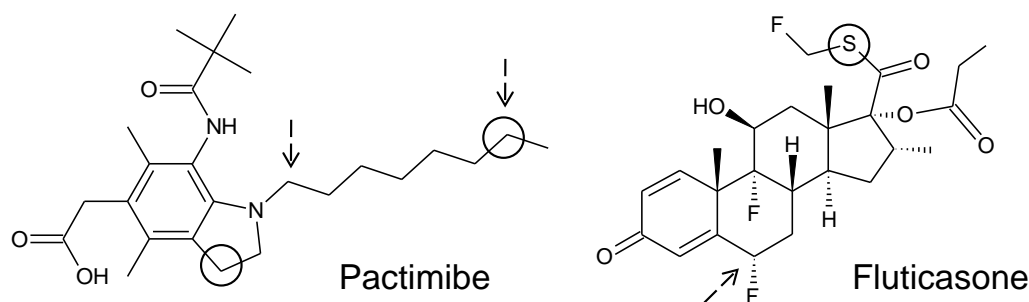


Figure 4.5: Examples of “satisfactory” predictions. Experimental metabolism sites are circled, dashed arrows show atoms predicted with $p > 0.5$ and $RI < 0.5$.

metabolites. There were no similar metabolism sites in the training set for these structures. Natural compounds often contain atypical sites of metabolism which are difficult to predict, thus the predictions are still in agreement with the general trends of cytochrome P450 reactivity of drug-like chemicals.

4.3 Comparison to Other Models

After development and validation of a new model it is useful to compare it to the existing ones. It is worth noting that direct comparison of different regioselectivity models is hardly possible because different methods produce different scales of predicted metabolism scores, and the output of programs also varies. This is the likely reason why only few such studies exist [99, 100].

The predictions of the developed regioselectivity model were compared to those of *SMARTCyp* [108]. This software is built using quantum chemical calculations – a completely different method for obtaining ligand-based models. The *SMARTCyp* predictions for 42 external validation set compounds were also categorized into the same four quality categories (“excellent”, “good”, “satisfactory”, and “unsatisfactory”) according to similar criteria as were used for evaluation of our model.

As we can see in Table 4.4, the overall performance of models is similar. Both models tend to correctly identify metabolism sites in most compounds. Like our model, *SMARTCyp* also predicts 28 compounds as “excellent” or “good”. The higher number of “unsatisfactory” predictions produced by our model (6 vs. 3)

Table 4.4: Comparison of our model predictions with the ones of *SMARTCyp*.

	<i>SMARTCyp</i> :			
	Excellent	Good	Satisfactory	Unsatisfactory
Our model:				
Excellent	11	4	5	2
Good	2	3	1	0
Satisfactory	0	4	4	0
Unsatisfactory	1	3	1	1

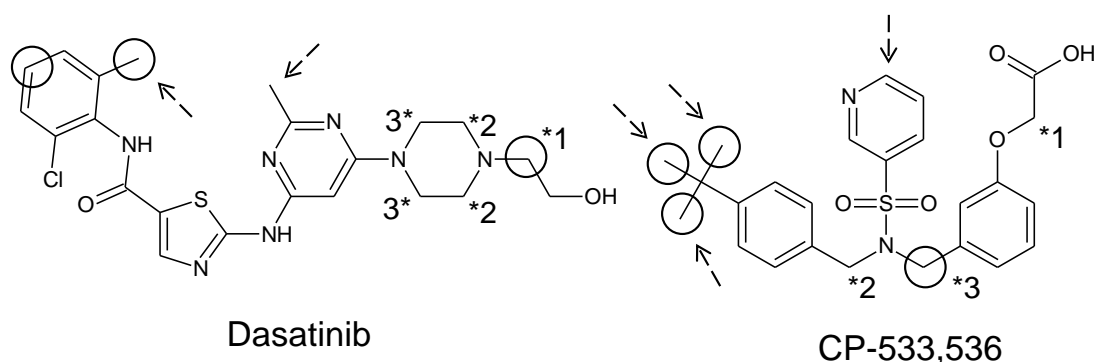


Figure 4.6: Combination of our model and *SMARTCyp* to predict sites of metabolism. Experimental metabolism sites are circled, arrows indicate our predictions, asterisks and numbers show sites predicted by *SMARTCyp* and their rank.

can be explained by the fact that according to the less strict evaluation criteria *SMARTCyp* predicts at least one metabolism site in every molecule.

An important outcome of comparing two regioselectivity models is that combination of different *in silico* approaches helps to predict more metabolism sites. Fig. 4.6 shows two example compounds for which the predictions of both models are satisfactory: only one metabolism site is found using individual approaches. Interestingly, different metabolism sites are predicted for dasatinib and CP-533,536. *SMARTCyp* forecasts N-dealkylation for both compounds, and our model suggests sites for aliphatic hydroxylation. Furthermore, the N-dealkylation site of dasatinib is ranked 3rd by our model with borderline probability of 0.46, and aromatic hydroxylation obtained 4th rank with probability 0.17. As a result, more possible metabolism sites predicted by the sum of two models could better guide medicinal chemists about possible modifications to improve metabolic stability or ease the experimental identification of metabolites by facilitating analysis of experimental mass spectral data.

It is also important to note that further comparison studies are necessary for regioselectivity models. The site of metabolism predictors have to be evaluated by independent researchers using experimental data obtained for novel compounds as the developers of one model cannot objectively compare it to the others. On the other hand, the regioselectivity models based on GALAS modeling methodology have some more useful features in addition to the good predictive power.

4.4 Adaptation of the Model to Compounds of Novel Classes

Although most of the metabolism sites of external validation set compounds have been predicted correctly, the calculated Reliability Indices were low for many of them, indicating that these metabolism sites are not in the model applicability domain. We have already shown how to expand this domain in case of the GALAS model of CYP3A4 inhibition, adapting the model to recognize novel compounds. New CYP3A4 inhibitors could be predicted after application of a straightforward training procedure which consists of adding of new compounds to the similarity correction part of the GALAS model. An analogous experiment was conducted for the developed regioselectivity models. Propranolol derivatives which are metabolized by CYP2D6 and CYP1A2 were chosen as an example of compounds of a novel chemical class. These compounds have two aromatic hydroxylation sites and one N-dealkylation site [134, 135].

In order to test the possibility to adapt the regioselectivity model to this novel compound class, baseline models were trained on the modified modeling dataset with all propranolol analogues and metabolites removed, and the same dataset was used as similarity correction library. Then three randomly chosen propranolol analogues were added one by one to this library.

The predictions after model training were similar for all propranolol analogues. Fig. 4.7 shows the changes in predicted metabolism sites for fluoropropranolol. The initial model with 0 analogues added predicts only one metabolism site of three. Such prediction would be classified as “satisfactory” according to the

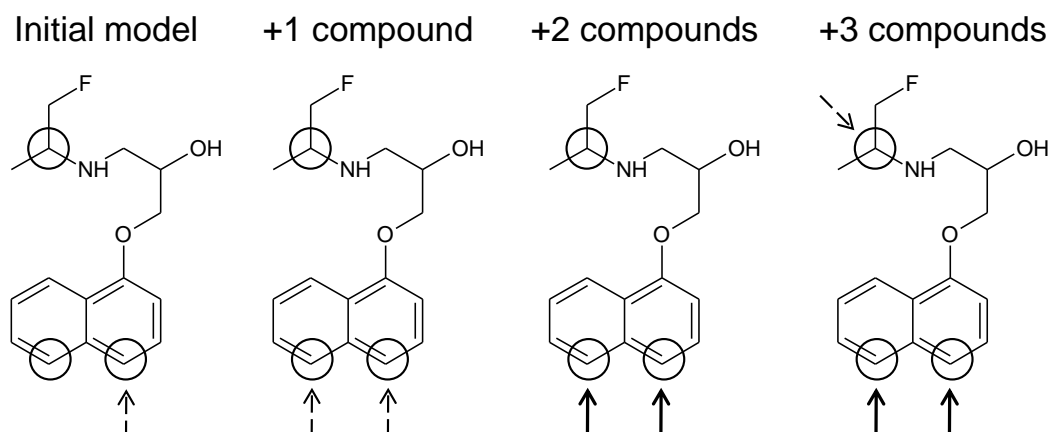


Figure 4.7: Changes in prediction of metabolism sites for fluoropropranolol. Experimental metabolism sites are circled, solid arrows indicate confidently predicted metabolism sites (probability > 0.5 and $RI > 0.5$), dashed arrows show sites predicted with probability > 0.5 and $RI < 0.5$.

model evaluation criteria (described in detail in Section 2.5). After adding the first similar compound both aromatic hydroxylation sites are predicted, and after adding another one, these sites are already estimated with high reliability. N-dealkylation site is found after addition of the third propranolol analogue, resulting in “excellent” prediction.

Fig. 4.8 shows the changes in the predicted probability and the Reliability Index. The steady increase of the calculated probability is observed for all metabolism sites of fluoropropranolol after adding its analogues to the training library of the model. The Reliability Indices also increase for all atoms, indicating that the compounds of novel class are now in the model applicability domain.

This example demonstrates the potential for practical applications of such trainable regioselectivity model, especially given the fact that the described improvements in predictions following the addition of similar compounds were instant and required no rebuilding of the baseline models. The regioselectivity model is more complex than any usual QSAR models and rebuilding it requires considerably more resources. In this situation the possibility to adapt the model to new chemical classes and new types of metabolism sites is even more significant in this case than it is for CYP3A4 inhibition. Furthermore, this training procedure can be automated by connecting it to the database with the metabolites of in-house compounds.

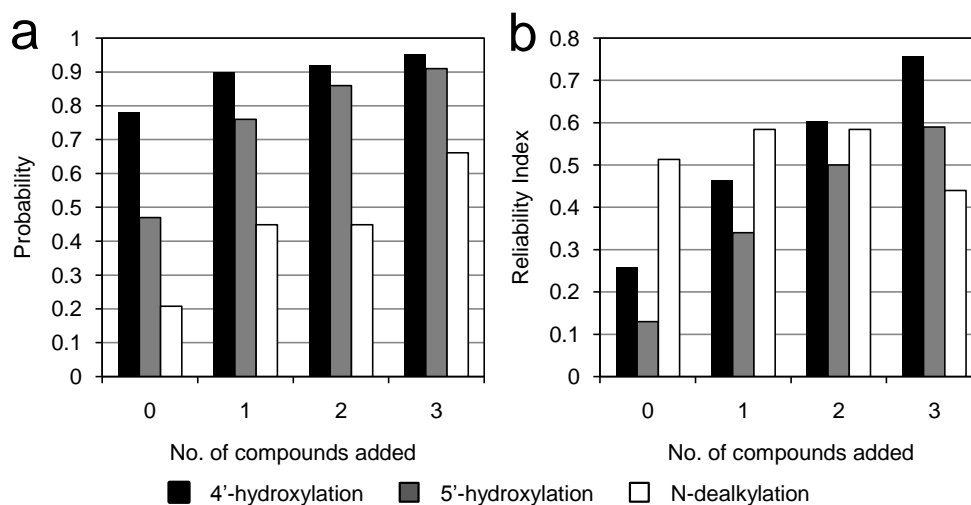


Figure 4.8: Changes in the predicted probability and *RI* values for metabolism sites of fluoroproranolol after training the model.

4.5 Adaptation of the Model to Cytochrome P450 Phenotyping

Determination of metabolites formed after incubation with human liver microsomes is only the first step of analysis of metabolism for a new chemical molecule. When the major metabolites are already known, the studies are focused to the identification of enzymes which catalyze their formation – reaction phenotyping. In this case *in silico* regioselectivity predictions for individual enzymes would be useful.

The GALAS models can be a valuable tool in this case. As it was described above, the global model can be viewed as a description of general reactivity trends. Local corrections then correspond to binding to the microsomal enzymes. The model predicts metabolism sites of all cytochrome P450 enzymes if we use a similarity correction database with the data on regioselectivity of human liver microsomal metabolism. Alternatively, having a database with only atoms that are oxidized by a particular enzyme marked as metabolism sites we can adapt the GALAS model to the specificity of this enzyme.

CYP2D6 was chosen for the experiment of model adaptation to cytochrome

Table 4.5: The performance of the regioselectivity model on the CYP2D6 test sets after training the model with CYP2D6 data.

		Pred. True	Pred. False	
Baseline model	Obs. True	93	24	Sensitivity: 79.5% Specificity: 74.0% Accuracy: 74.4%
	Obs. False	508	1449	
After local modeling	Obs. True	79	38	Sensitivity: 67.5% Specificity: 93.3% Accuracy: 91.9%
	Obs. False	131	1826	
After local modeling, $RI > 0.3$	Obs. True	61	26	Sensitivity: 79.1% Specificity: 96.6% Accuracy: 95.3%
	Obs. False	58	1646	
After local modeling, $RI > 0.5$	Obs. True	28	11	Sensitivity: 71.8% Specificity: 98.4% Accuracy: 97.5%
	Obs. False	17	1046	

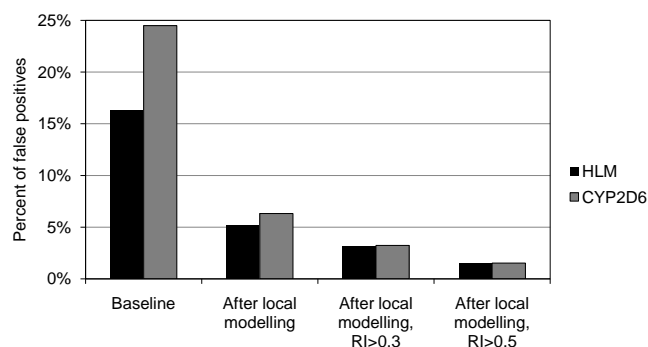


Figure 4.9: The percent of false positive predictions in the human liver microsomes model and the model, retrained using CYP2D6 data.

P450 reaction phenotyping. This enzyme has a more defined specificity comparing to other human drug metabolizing enzymes. Its substrates are aromatic compounds, having a basic center in the distance of 5-10 Å from an aromatic ring [20, 22]. The major reactions of this enzyme are aromatic hydroxylation and O-dealkylation at the aromatic ring and N-dealkylation in the basic amino group.

The results of prediction of regioselectivity of CYP2D6 metabolism after retraining can be found in Table 4.5. As it was expected, the baseline models

applied on CYP2D6 data produce large numbers of false positive predictions. These constitute about 24% of all predictions, compared to 16% by model, predicting sites of human liver microsomal metabolism (Fig. 4.9). This results from the fact that HLM models take into account metabolites produced by all enzymes, yet one enzyme can be only responsible for some of them. In other words, if a particular atom is metabolized by CYP2D6, it is metabolized in human liver microsomes, whereas the reverse is not necessarily true. After the similarity corrections the percent of false positive predictions decreases and is comparable to the microsomal model (Fig. 4.9). The trends of increasing quality of predictions after filtering them according to the reliability are similar to those observed in case of all microsomal enzymes. Thus the adaptation of the model to predict the CYP2D6 metabolism sites is obvious, and we believe that adjusting the model to the specificity of other microsomal enzymes should be also possible.

Conclusions

1. A structure-activity relationship model of CYP3A4 inhibition was developed using the novel GALAS modeling method. It exhibits good agreement between experimental and predicted values: the accuracy of predictions for compounds within the model applicability domain is 89%.
2. A model that predicts regioselectivity of metabolism in human liver microsomes was developed. It calculates probability of oxidation for every atom in the molecule and provides the reliability of prediction, defining the model applicability domain. Among metabolism sites that were observed experimentally and predicted with high reliability, 78% were identified correctly.
3. Local similarity-based corrections improved the accuracy of classification in case of both CYP3A4 inhibition and metabolism regioselectivity models, and calculated Reliability Index values clearly correlated with the quality of predictions.
4. The possibility to train the GALAS model was confirmed by adding experimental data for new chemical compounds into the local part of the CYP3A4 inhibition model. Its applicability domain expanded and new CYP3A4 inhibitors were predicted following this training procedure.
5. The trainability of regioselectivity model was also demonstrated. The model easily adapted to recognize additional metabolism sites in compounds with a new structural scaffold. Furthermore, the model was re-trained to recognize the specificity of a particular enzyme, CYP2D6, showing the possibility to use it in the cytochrome P450 reaction phenotyping.

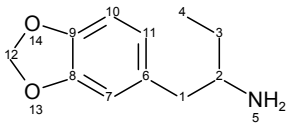
6. The regioselectivity model was validated on an external set of 42 compounds. Major metabolism sites were predicted for 27 of them, and at least one site was predicted for 36 compounds. These results are comparable to the predictivity of *SMARTCyp*, a quantum chemistry based software.

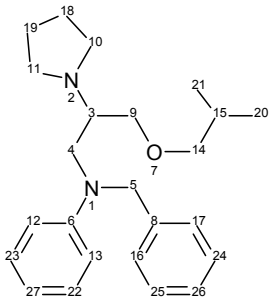
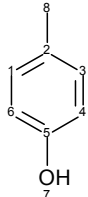
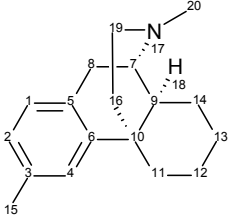
Appendix A

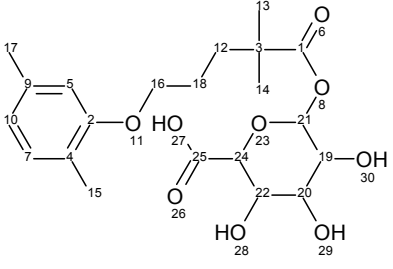
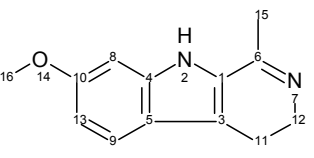
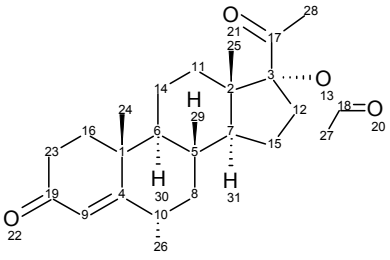
Results of External Validation of Regioselectivity Model

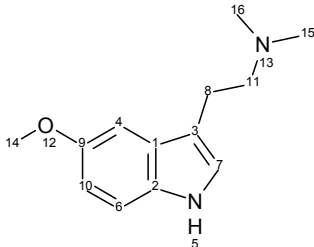
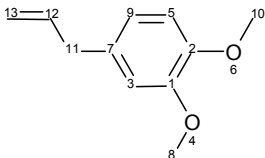
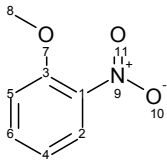
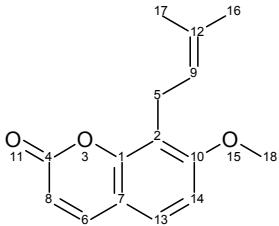
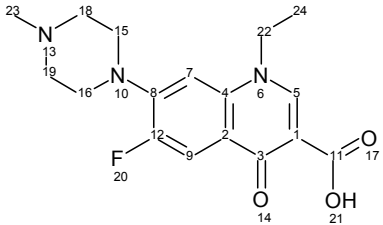
The tables below show all compounds of the external validation set with predictions for each atom. Only atoms that were marked (i.e., C–H and S atoms) are listed. Only one of equivalent atoms is included. The experimental sites of metabolism are bold. AtomNo is the number of atom, $p(\text{baseline})$ – baseline probability, p – final probability, RI – Reliability Index value.

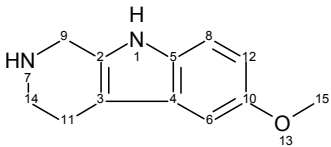
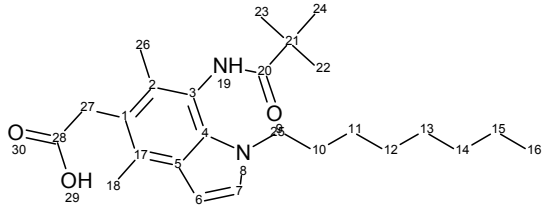
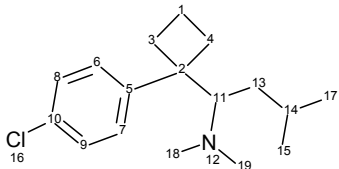
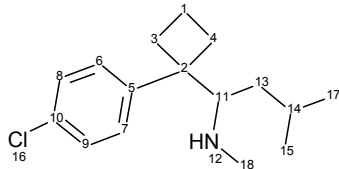
Excellent Predictions

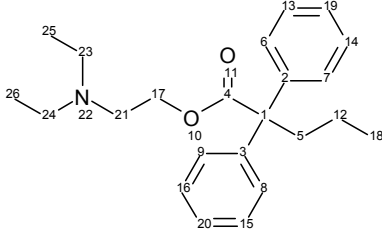
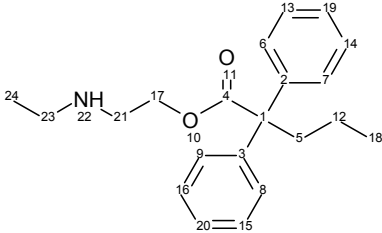
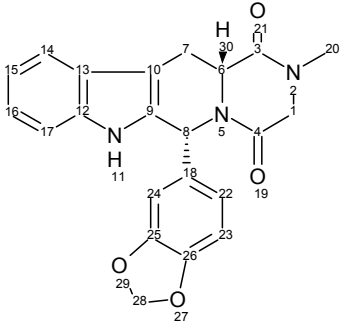
Compound	AtomNo	$p(\text{baseline})$	p	RI
BDB [148]				
	1	0.492	0.069	0.764
	2	0.117	0.046	0.708
	3	0.182	0.098	0.552
	4	0.192	0.120	0.485
	7	0.256	0.056	0.917
	10	0.124	0.053	0.929
	11	0.452	0.060	0.914
	12	0.856	0.944	0.921

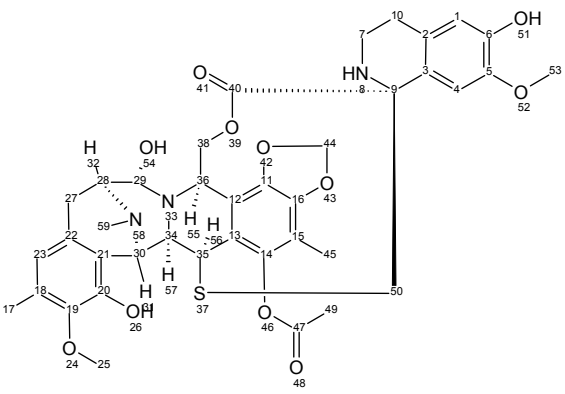
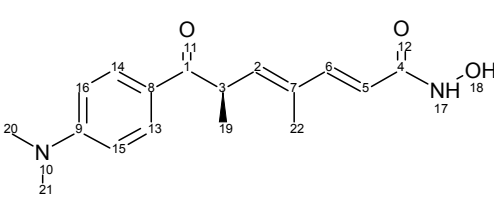
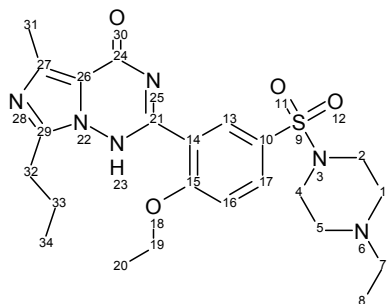
Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
Bepridil [122]				
	3	0.100	0.088	0.251
	4	0.029	0.018	0.469
	5	0.400	0.308	0.254
	9	0.020	0.016	0.512
	10	0.452	0.366	0.365
	13	0.072	0.043	0.561
	14	0.012	0.036	0.176
	15	0.225	0.210	0.314
	17	0.212	0.123	0.250
	19	0.110	0.042	0.735
	21	0.333	0.247	0.295
	23	0.180	0.116	0.501
	25	0.566	0.291	0.530
	26	0.943	0.607	0.373
27	0.957	0.824	0.363	
p-Cresol [149]				
	1	0.515	0.151	0.741
	6	0.816	0.765	0.203
	8	0.852	0.741	0.387
Dimemorfan [150]				
	1	0.063	0.063	0.651
	2	0.304	0.100	0.592
	4	0.027	0.027	0.829
	7	0.005	0.005	0.932
	8	0.379	0.079	0.895
	9	0.210	0.070	0.907
	11	0.315	0.096	0.828
	12	0.343	0.100	0.805
	13	0.283	0.089	0.768
	14	0.194	0.075	0.797
	15	0.902	0.877	0.479
	16	0.132	0.070	0.877
	19	0.007	0.007	0.928
	20	0.964	0.970	0.800

Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
Gemfibrozil glucuronide [151]				
	5	0.187	0.116	0.208
	7	0.057	0.065	0.251
	10	0.575	0.506	0.283
	12	0.211	0.131	0.436
	13	0.098	0.098	0.643
	15	0.969	0.920	0.195
	16	0.074	0.074	0.419
	17	0.963	0.921	0.172
	18	0.046	0.046	0.421
	19	0.002	0.005	0.434
	20	0.002	0.002	0.492
	21	0.010	0.010	0.585
	22	0.006	0.006	0.425
	24	0.013	0.021	0.185
Harmaline [152]				
	8	0.673	0.133	0.471
	9	0.212	0.090	0.576
	11	0.698	0.484	0.365
	12	0.564	0.175	0.574
	13	0.673	0.244	0.285
	15	0.600	0.320	0.331
	16	0.837	0.841	0.366
Medroxyprogesterone acetate [153]				
	5	0.033	0.031	0.869
	6	0.083	0.038	0.852
	7	0.035	0.023	0.747
	8	0.048	0.044	0.716
	10	0.577	0.790	0.357
	11	0.057	0.029	0.722
	12	0.271	0.109	0.614
	14	0.029	0.029	0.860
	15	0.083	0.137	0.692
	16	0.316	0.617	0.601
	23	0.620	0.762	0.578
	24	0.106	0.046	0.779
	25	0.110	0.045	0.747
	26	0.134	0.080	0.507
27	0.323	0.399	0.416	
28	0.230	0.183	0.316	

Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
5-methoxy-N,N-dimethyltryptamine [154]				
	4	0.445	0.108	0.818
	6	0.325	0.074	0.738
	7	0.631	0.077	0.793
	8	0.442	0.094	0.611
	10	0.550	0.201	0.304
	11	0.332	0.200	0.364
	14	0.875	0.918	0.553
	15	0.778	0.735	0.158
Methyleugenol [155, 156]				
	3	0.296	0.099	0.732
	5	0.217	0.109	0.773
	8	0.724	0.384	0.458
	9	0.575	0.088	0.788
	10	0.798	0.397	0.565
	11	0.949	0.904	0.699
2-Nitroanisoole [157]				
	2	0.322	0.120	0.768
	4	0.941	0.813	0.343
	5	0.654	0.094	0.767
	6	0.372	0.226	0.538
	8	0.846	0.847	0.411
Osthol [158]				
	5	0.652	0.299	0.488
	6	0.350	0.332	0.127
	8	0.323	0.293	0.189
	13	0.147	0.116	0.697
	14	0.531	0.203	0.553
	17	0.872	0.928	0.470
	18	0.698	0.752	0.363
Pefloxacin [122]				
	7	0.009	0.009	0.557
	9	0.013	0.013	0.545
	16	0.194	0.223	0.356
	19	0.145	0.137	0.400
	22	0.119	0.216	0.346
	23	0.901	0.862	0.129
	24	0.128	0.094	0.447

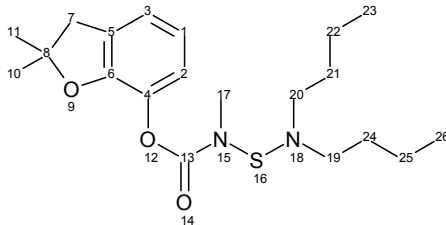
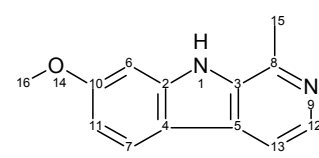
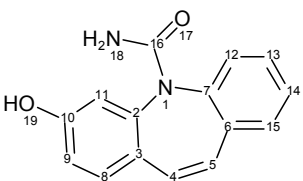
Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
Pinoline [159]				
	6	0.493	0.143	0.563
	8	0.278	0.118	0.530
	9	0.318	0.201	0.433
	11	0.589	0.431	0.619
	12	0.593	0.348	0.248
	14	0.106	0.134	0.530
	15	0.890	0.862	0.260
R-125528 (pactimibe metabolite) [160]				
	6	0.501	0.264	0.327
	7	0.450	0.281	0.255
	9	0.423	0.458	0.263
	10	0.248	0.149	0.352
	11	0.136	0.073	0.535
	12	0.163	0.109	0.393
	13	0.240	0.162	0.415
	14	0.406	0.235	0.424
	15	0.625	0.633	0.152
	16	0.681	0.471	0.317
	18	0.407	0.367	0.325
	22	0.690	0.184	0.544
	26	0.357	0.082	0.543
27	0.759	0.249	0.459	
Sibutramine [161]				
	1	0.198	0.083	0.614
	3	0.088	0.082	0.501
	6	0.064	0.059	0.599
	8	0.074	0.062	0.544
	11	0.137	0.089	0.362
	13	0.085	0.075	0.463
	14	0.104	0.156	0.333
	15	0.309	0.129	0.551
	19	0.685	0.866	0.546
	Sibutramine metabolite M1 [161]			
	1	0.185	0.076	0.627
	3	0.146	0.107	0.426
	6	0.079	0.071	0.587
	8	0.091	0.074	0.526
	11	0.168	0.100	0.344
	13	0.107	0.085	0.464

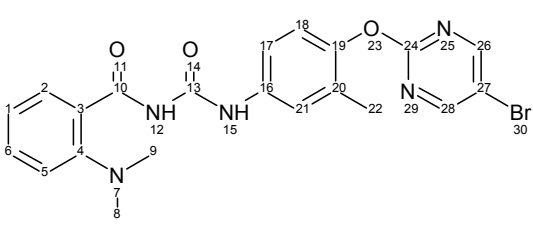
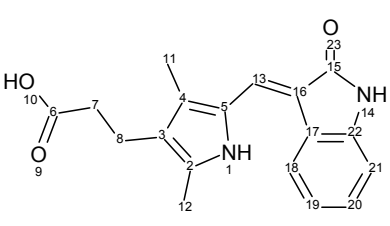
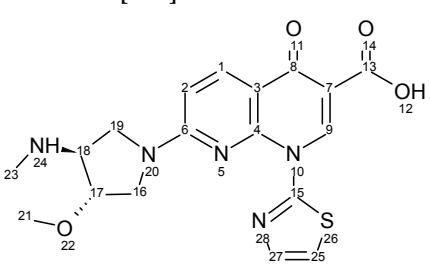
Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
	14	0.164	0.152	0.250
	15	0.295	0.131	0.570
	18	0.712	0.761	0.393
SKF 525A [162]				
	5	0.200	0.145	0.474
	6	0.043	0.036	0.754
	12	0.095	0.110	0.339
	13	0.145	0.102	0.525
	17	0.032	0.030	0.610
	18	0.112	0.067	0.575
	19	0.798	0.251	0.615
	21	0.381	0.301	0.292
	24	0.711	0.826	0.282
	25	0.043	0.040	0.644
SKF8742 [162]				
	5	0.206	0.148	0.504
	6	0.038	0.033	0.777
	12	0.106	0.121	0.369
	13	0.115	0.089	0.649
	17	0.039	0.037	0.661
	18	0.142	0.085	0.636
	19	0.790	0.213	0.654
	21	0.484	0.369	0.288
	23	0.783	0.795	0.225
	24	0.086	0.057	0.646
Tadalafil [122]				
	1	0.028	0.035	0.151
	6	0.002	0.002	0.467
	7	0.140	0.123	0.502
	8	0.017	0.020	0.423
	14	0.314	0.132	0.469
	15	0.813	0.573	0.128
	16	0.735	0.698	0.432
	17	0.417	0.178	0.339
	20	0.819	0.804	0.245
	22	0.095	0.091	0.440
	23	0.032	0.019	0.549
	24	0.046	0.036	0.396
	28	0.653	0.683	0.144

Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
Trabectedin [163]				
	1	0.096	0.100	0.321
	4	0.070	0.024	0.168
	7	0.006	0.005	0.311
	10	0.422	0.156	0.418
	17	0.924	0.880	0.192
	23	0.017	0.017	0.362
	25	0.478	0.598	0.129
	27	0.129	0.094	0.344
	28	0.000	0.000	0.466
	29	0.002	0.002	0.363
	30	0.001	0.001	0.326
	34	0.000	0.000	0.424
	36	0.010	0.009	0.332
	37	0.065	0.085	0.163
	38	0.001	0.001	0.256
	44	0.406	0.455	0.159
	45	0.290	0.230	0.396
49	0.616	0.350	0.241	
53	0.595	0.783	0.317	
59	0.611	0.608	0.310	
Trichostatin A [122]				
	3	0.776	0.494	0.234
	13	0.038	0.036	0.418
	15	0.049	0.042	0.338
	19	0.164	0.145	0.428
	21	0.677	0.742	0.362
	22	0.419	0.258	0.454
Vardenafil [164]				
	1	0.007	0.007	0.859
	2	0.004	0.006	0.844
	7	0.165	0.847	0.754
	8	0.299	0.079	0.614
	13	0.002	0.002	0.902
	16	0.013	0.013	0.895
	17	0.012	0.012	0.893
	19	0.142	0.065	0.803
	20	0.417	0.074	0.797
	31	0.545	0.266	0.301
	32	0.544	0.149	0.637

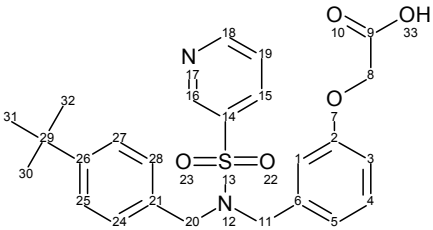
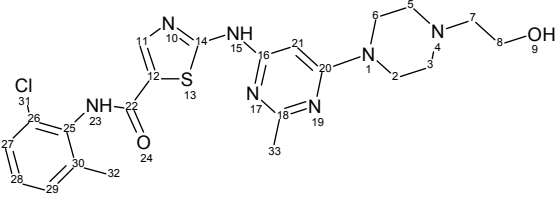
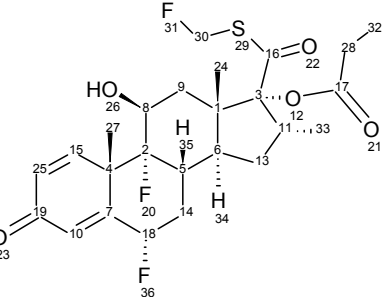
Compound	AtomNo	$p(\text{baseline})$	p	RI
	33	0.409	0.058	0.809
	34	0.659	0.145	0.695

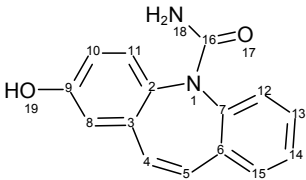
Good Predictions

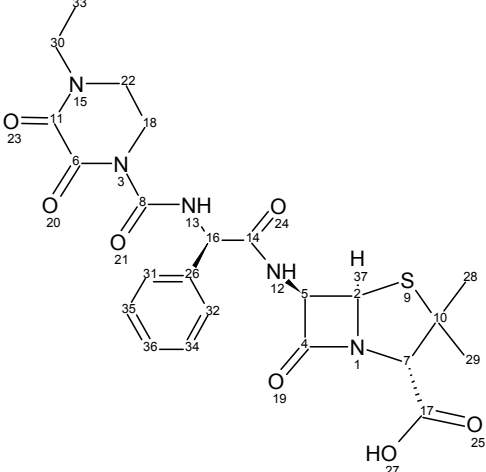
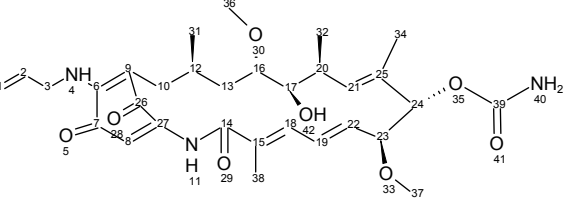
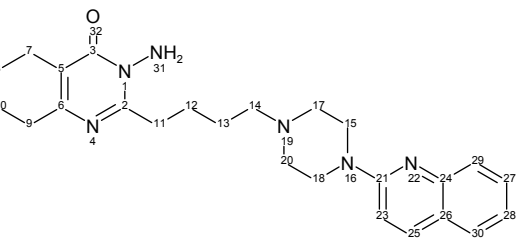
Compound	AtomNo	$p(\text{baseline})$	p	RI
Carbosulfan [165, 166]				
	1	0.334	0.135	0.294
	2	0.314	0.138	0.342
	3	0.137	0.079	0.314
	7	0.784	0.715	0.395
	10	0.373	0.157	0.478
	16	0.220	0.232	0.145
	17	0.754	0.302	0.465
	19	0.146	0.158	0.286
	21	0.101	0.060	0.486
	22	0.322	0.169	0.284
23	0.545	0.291	0.434	
Harmine [152]				
	6	0.574	0.145	0.613
	7	0.287	0.064	0.550
	11	0.546	0.697	0.557
	12	0.834	0.869	0.559
	13	0.746	0.155	0.665
	15	0.721	0.291	0.385
	16	0.728	0.717	0.173
3-Hydroxycarbamazepine [167]				
	8	0.035	0.046	0.481
	9	0.873	0.720	0.157
	11	0.377	0.439	0.539
	12	0.192	0.099	0.698
	13	0.717	0.717	0.383
	14	0.901	0.825	0.511
	15	0.106	0.078	0.628

Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
NSC 639829 [168]				
	1	0.872	0.869	0.255
	2	0.096	0.085	0.296
	5	0.266	0.161	0.369
	6	0.421	0.279	0.319
	9	0.753	0.755	0.204
	17	0.101	0.120	0.489
	18	0.073	0.065	0.342
	21	0.040	0.054	0.367
	22	0.822	0.724	0.210
26	0.021	0.024	0.321	
TSU-68 [169]				
	7	0.371	0.173	0.339
	8	0.601	0.447	0.374
	11	0.499	0.182	0.508
	12	0.488	0.234	0.373
	18	0.086	0.071	0.386
	19	0.839	0.846	0.357
	20	0.767	0.656	0.177
21	0.200	0.084	0.509	
Voreloxin [170]				
	1	0.004	0.004	0.510
	2	0.042	0.037	0.457
	9	0.033	0.029	0.206
	16	0.091	0.082	0.337
	17	0.039	0.037	0.261
	18	0.012	0.012	0.331
	19	0.096	0.087	0.329
	21	0.407	0.373	0.119
	23	0.707	0.665	0.208
	25	0.308	0.245	0.322
	26	0.046	0.044	0.115
27	0.272	0.246	0.217	

Satisfactory Predictions

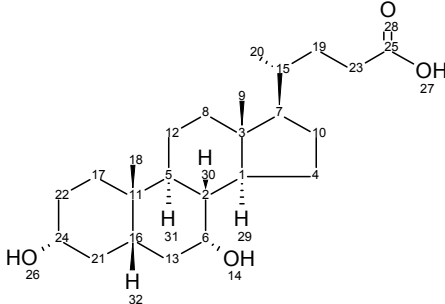
Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
CP-533,536 [171]				
	1	0.086	0.058	0.339
	3	0.516	0.431	0.107
	4	0.040	0.039	0.487
	5	0.338	0.150	0.385
	8	0.395	0.255	0.259
	11	0.134	0.057	0.280
	15	0.260	0.136	0.383
	16	0.358	0.187	0.428
	18	0.944	0.592	0.427
	19	0.568	0.284	0.436
	20	0.253	0.215	0.216
	24	0.014	0.013	0.502
	25	0.153	0.092	0.234
32	0.777	0.719	0.220	
Dasatinib [172]				
	2	0.029	0.040	0.185
	3	0.044	0.059	0.130
	7	0.433	0.457	0.220
	8	0.110	0.095	0.511
	11	0.006	0.006	0.386
	13	0.027	0.024	0.377
	21	0.036	0.030	0.453
	27	0.066	0.066	0.431
	28	0.276	0.171	0.246
	29	0.069	0.056	0.310
	32	0.891	0.875	0.328
33	0.875	0.730	0.082	
Fluticasone [173]				
	5	0.028	0.027	0.767
	6	0.019	0.016	0.662
	8	0.001	0.001	0.790
	9	0.016	0.016	0.670
	11	0.228	0.105	0.502
	13	0.060	0.071	0.427
	14	0.019	0.018	0.645
18	0.467	0.799	0.332	

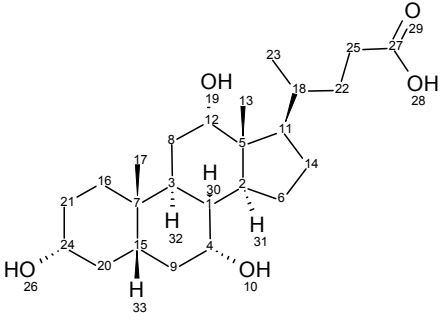
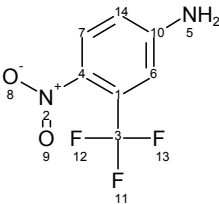
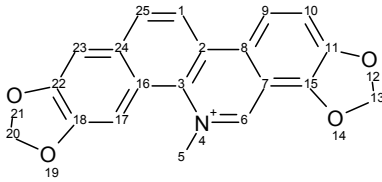
Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
	24	0.071	0.051	0.598
	27	0.083	0.059	0.625
	28	0.462	0.185	0.181
	29	0.262	0.428	0.377
	32	0.411	0.198	0.372
	33	0.135	0.099	0.512
2-Hydroxycarbamazepine [167]				
	8	0.198	0.146	0.473
	10	0.702	0.481	0.230
	11	0.077	0.052	0.599
	12	0.229	0.113	0.667
	13	0.784	0.696	0.513
	14	0.901	0.825	0.511
	15	0.106	0.078	0.628
				
Pactimibe [160]				
	8	0.091	0.093	0.261
	9	0.253	0.176	0.357
	10	0.585	0.557	0.180
	11	0.084	0.055	0.460
	12	0.090	0.061	0.394
	13	0.127	0.096	0.342
	14	0.195	0.122	0.423
	15	0.323	0.180	0.504
	16	0.562	0.517	0.173
	17	0.633	0.419	0.303
	18	0.211	0.100	0.382
	19	0.672	0.210	0.385
	23	0.294	0.151	0.440
	28	0.608	0.298	0.306

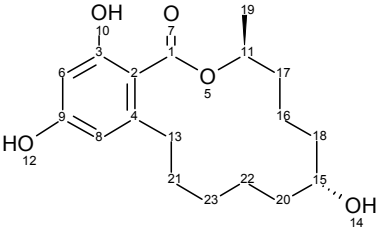
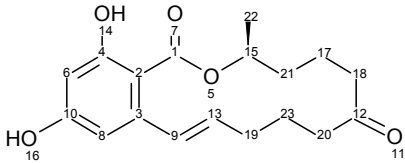
Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
Piperacillin [174]				
	2	0.001	0.001	0.468
	5	0.000	0.000	0.505
	7	0.007	0.007	0.289
	9	0.022	0.024	0.301
	16	0.013	0.013	0.458
	18	0.002	0.002	0.351
	22	0.002	0.002	0.369
	28	0.112	0.084	0.540
	30	0.312	0.442	0.099
	31	0.019	0.018	0.388
	33	0.201	0.171	0.300
	34	0.021	0.020	0.446
	36	0.084	0.082	0.427
Tanespimycin [175]				
	3	0.036	0.036	0.342
	10	0.101	0.061	0.358
	12	0.010	0.010	0.337
	13	0.023	0.020	0.358
	16	0.014	0.014	0.366
	17	0.059	0.059	0.384
	20	0.122	0.052	0.340
	23	0.015	0.015	0.458
	24	0.069	0.149	0.238
	31	0.058	0.028	0.442
	32	0.048	0.028	0.368
	34	0.080	0.065	0.340
	36	0.474	0.640	0.169
	37	0.182	0.438	0.135
	38	0.519	0.195	0.429
TZB-30878 [176]				
	7	0.825	0.723	0.182
	8	0.533	0.336	0.224
	9	0.769	0.699	0.159
	10	0.582	0.369	0.334
	11	0.238	0.207	0.319
	12	0.041	0.035	0.389
	13	0.021	0.020	0.528
	14	0.886	0.864	0.205
	15	0.006	0.006	0.488

Compound	AtomNo	$p(\text{baseline})$	p	RI
	17	0.013	0.012	0.494
	23	0.024	0.024	0.445
	25	0.077	0.073	0.369
	27	0.398	0.134	0.295
	28	0.557	0.366	0.281
	29	0.371	0.287	0.472
	30	0.476	0.242	0.263

Unsatisfactory Predictions

Compound	AtomNo	$p(\text{baseline})$	p	RI
Chenodeoxycholic acid [177]				
	1	0.067	0.047	0.675
	2	0.032	0.028	0.681
	4	0.050	0.038	0.602
	5	0.148	0.110	0.618
	6	0.012	0.012	0.682
	7	0.226	0.083	0.611
	8	0.058	0.038	0.570
	9	0.094	0.053	0.704
	10	0.161	0.120	0.551
	12	0.026	0.023	0.728
	13	0.052	0.051	0.539
	15	0.070	0.055	0.626
	16	0.381	0.185	0.464
	17	0.327	0.166	0.470
	18	0.112	0.080	0.639
	19	0.119	0.105	0.623
	20	0.041	0.036	0.706
	21	0.111	0.091	0.534
	22	0.152	0.088	0.534
	23	0.340	0.089	0.509
	24	0.026	0.026	0.563

Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>	
Cholic acid [177]					
	1	0.018	0.015	0.697	
	2	0.057	0.041	0.625	
	3	0.126	0.088	0.627	
	4	0.011	0.011	0.676	
	6	0.028	0.024	0.621	
	8	0.012	0.011	0.716	
	9	0.038	0.037	0.608	
	11	0.188	0.093	0.583	
	12	0.010	0.010	0.592	
	13	0.080	0.045	0.722	
	14	0.094	0.066	0.593	
	15	0.312	0.166	0.507	
	16	0.260	0.147	0.571	
	17	0.084	0.059	0.643	
	18	0.041	0.035	0.548	
	20	0.084	0.070	0.598	
	21	0.120	0.086	0.550	
	22	0.091	0.083	0.645	
	23	0.030	0.027	0.675	
	24	0.023	0.023	0.581	
	25	0.284	0.135	0.224	
	Flu-1 [178, 179]				
		6	0.516	0.303	0.544
		7	0.237	0.163	0.716
		14	0.636	0.258	0.715
Sanguinarine [180]					
	1	0.101	0.087	0.391	
	5	0.236	0.202	0.350	
	6	0.210	0.198	0.237	
	9	0.364	0.203	0.506	
	10	0.392	0.236	0.391	
	13	0.557	0.323	0.414	
	17	0.181	0.181	0.478	
	20	0.547	0.396	0.361	
	23	0.797	0.451	0.313	
	25	0.424	0.174	0.285	

Compound	AtomNo	<i>p</i> (baseline)	<i>p</i>	<i>RI</i>
Zearalenol [181]				
	6	0.680	0.054	0.635
	8	0.295	0.070	0.402
	11	0.045	0.045	0.379
	13	0.168	0.083	0.312
	15	0.014	0.052	0.460
	16	0.096	0.047	0.525
	17	0.051	0.036	0.553
	18	0.045	0.035	0.445
	19	0.068	0.063	0.524
	20	0.069	0.032	0.614
	21	0.055	0.031	0.618
	22	0.086	0.039	0.622
	23	0.048	0.023	0.729
Zearalenone [182]				
	6	0.649	0.047	0.618
	8	0.203	0.095	0.209
	15	0.031	0.031	0.362
	17	0.161	0.080	0.543
	18	0.176	0.152	0.468
	19	0.375	0.246	0.503
	20	0.317	0.199	0.490
	21	0.084	0.049	0.608
	22	0.094	0.075	0.459
	23	0.215	0.127	0.604

References

- [1] Kerns, E., and Di, L. *Drug-like Properties: Concepts, Structure Design and Methods: from ADME to Toxicity Optimization*, 1st ed.; Academic Press: San Diego, 2008.
- [2] Nedderman, A. N. R. (2009) Metabolites in safety testing: metabolite identification strategies in discovery and development. *Biopharm Drug Dispos* 30, 153–162.
- [3] Wienkers, L. C., and Heath, T. G. (2005) Predicting in vivo drug interactions from in vitro drug discovery data. *Nat Rev Drug Discov* 4, 825–833.
- [4] Rendic, S., and Di Carlo, F. J. (1997) Human cytochrome P450 enzymes: a status report summarizing their reactions, substrates, inducers, and inhibitors. *Drug Metab Rev* 29, 413–580.
- [5] Tarcsay, A., and Keseru, G. M. (2011) In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin Drug Metab Toxicol* 7, 299–312.
- [6] Ahlström, M. M., Ridderström, M., Zamora, I., and Luthman, K. (2007) CYP2C9 structure-metabolism relationships: optimizing the metabolic stability of COX-2 inhibitors. *J Med Chem* 50, 4444–4452.
- [7] Ahlström, M. M., Ridderström, M., and Zamora, I. (2007) CYP2C9 structure-metabolism relationships: substrates, inhibitors, and metabolites. *J Med Chem* 50, 5382–5391.
- [8] Boyer, D., Bauman, J. N., Walker, D. P., Kapinos, B., Karki, K., and Kalgutkar, A. S. (2009) Utility of MetaSite in improving metabolic stability of the neutral indomethacin amide derivative and selective cyclooxygenase-2 inhibitor 2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)-N-phenethyl-acetamide. *Drug Metab Dispos* 37, 999–1008.
- [9] Anari, M. R., and Baillie, T. A. (2005) Bridging cheminformatic metabolite prediction and tandem mass spectrometry. *Drug Discov Today* 10, 711–717.

- [10] Bonn, B., Leandersson, C., Fontaine, F., and Zamora, I. (2010) Enhanced metabolite identification with MS(E) and a semi-automated software for structural elucidation. *Rapid Commun Mass Spectrom* 24, 3127–3138.
- [11] Gleeson, M. P., Hersey, A., and Hannongbua, S. (2011) In-silico ADME models: a general assessment of their utility in drug discovery applications. *Curr Top Med Chem* 11, 358–381.
- [12] Worth, A. P., Hartung, T., and Van Leeuwen, C. J. (2004) The role of the European centre for the validation of alternative methods (ECVAM) in the validation of (Q)SARs. *SAR QSAR Environ Res* 15, 345–358.
- [13] Japertas, P. Fragmentinių ir statistinių metodų naudojimas biologiškai aktyvių junginių struktūros-aktyvumo sąryšio tyrimuose. Ph.D. thesis, Vilniaus universitetas, Vilnius, 2007.
- [14] Sazonovas, A., Japertas, P., and Didziapetris, R. (2010) Estimation of reliability of predictions and model applicability domain evaluation in the analysis of acute toxicity (LD50). *SAR QSAR Environ Res* 21, 127–148.
- [15] Testa, B. *The Metabolism of Drugs and Other Xenobiotics: Biochemistry of Redox Reactions*; Academic Press: San Diego, 1995.
- [16] Guengerich, F. P. (2001) Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem Res Toxicol* 14, 611–650.
- [17] Jung, C., de Vries, S., and Schünemann, V. (2011) Spectroscopic characterization of cytochrome P450 Compound I. *Arch Biochem Biophys* 507, 44–55.
- [18] Isin, E. M., and Guengerich, F. P. (2007) Complex reactions catalyzed by cytochrome P450 enzymes. *Biochim Biophys Acta* 1770, 314–329.
- [19] Ortiz de Montellano, P. R., and Nelson, S. D. (2011) Rearrangement reactions catalyzed by cytochrome P450s. *Arch Biochem Biophys* 507, 95–110.
- [20] Chohan, K. K., Paine, S. W., and Waters, N. J. (2006) Quantitative structure activity relationships in drug metabolism. *Curr Top Med Chem* 6, 1569–1578.
- [21] Paine, M. J. I., McLaughlin, L. A., Flanagan, J. U., Kemp, C. A., Sutcliffe, M. J., Roberts, G. C. K., and Wolf, C. R. (2003) Residues glutamate 216 and aspartate 301 are key determinants of substrate specificity and product regioselectivity in cytochrome P450 2D6. *J Biol Chem* 278, 4021–4027.
- [22] Rowland, P., Blaney, F. E., Smyth, M. G., Jones, J. J., Leydon, V. R.,

- Oxbrow, A. K., Lewis, C. J., Tennant, M. G., Modi, S., Eggleston, D. S., Chenery, R. J., and Bridges, A. M. (2006) Crystal structure of human cytochrome P450 2D6. *J Biol Chem* 281, 7614–7622.
- [23] Flanagan, J. U., Maréchal, J., Ward, R., Kemp, C. A., McLaughlin, L. A., Sutcliffe, M. J., Roberts, G. C. K., Paine, M. J. I., and Wolf, C. R. (2004) Phe120 contributes to the regiospecificity of cytochrome P450 2D6: mutation leads to the formation of a novel dextromethorphan metabolite. *Biochem J* 380, 353–360.
- [24] Keizers, P. H. J., Lussenburg, B. M. A., de Graaf, C., Mentink, L. M., Vermeulen, N. P. E., and Commandeur, J. N. M. (2004) Influence of phenylalanine 120 on cytochrome P450 2D6 catalytic selectivity and regiospecificity: crucial role in 7-methoxy-4-(aminomethyl)-coumarin metabolism. *Biochem Pharmacol* 68, 2263–2271.
- [25] Hayhurst, G. P., Harlow, J., Chowdry, J., Gross, E., Hilton, E., Lennard, M. S., Tucker, G. T., and Ellis, S. W. (2001) Influence of phenylalanine-481 substitutions on the catalytic activity of cytochrome P450 2D6. *Biochem J* 355, 373–379.
- [26] Lussenburg, B. M. A., Keizers, P. H. J., de Graaf, C., Hidestrand, M., Ingelman-Sundberg, M., Vermeulen, N. P. E., and Commandeur, J. N. M. (2005) The role of phenylalanine 483 in cytochrome P450 2D6 is strongly substrate dependent. *Biochem Pharmacol* 70, 1253–1261.
- [27] McLaughlin, L. A., Paine, M. J. I., Kemp, C. A., Maréchal, J., Flanagan, J. U., Ward, C. J., Sutcliffe, M. J., Roberts, G. C. K., and Wolf, C. R. (2005) Why is quinidine an inhibitor of cytochrome P450 2D6? The role of key active-site residues in quinidine binding. *J Biol Chem* 280, 38617–38624.
- [28] Kotsuma, M., Tokui, T., Freudenthaler, S., and Nishimura, K. (2008) Effects of ketoconazole and quinidine on pharmacokinetics of pactimibe and its plasma metabolite, R-125528, in humans. *Drug Metab Dispos* 36, 1505–1511.
- [29] Kotsuma, M., Hanzawa, H., Iwata, Y., Takahashi, K., and Tokui, T. (2008) Novel binding mode of the acidic CYP2D6 substrates pactimibe and its metabolite R-125528. *Drug Metab Dispos* 36, 1938–1943.
- [30] Parikh, A., Josephy, P. D., and Guengerich, F. P. (1999) Selection and characterization of human cytochrome P450 1A2 mutants with altered catalytic properties. *Biochemistry* 38, 5283–5289.
- [31] Yun, C. H., Miller, G. P., and Guengerich, F. P. (2000) Rate-determining

- steps in phenacetin oxidations by human cytochrome P450 1A2 and selected mutants. *Biochemistry* 39, 11319–11329.
- [32] Sansen, S., Yano, J. K., Reynald, R. L., Schoch, G. A., Griffin, K. J., Stout, C. D., and Johnson, E. F. (2007) Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. *J Biol Chem* 282, 14348–14355.
- [33] Hutzler, J. M., and Tracy, T. S. (2002) Atypical kinetic profiles in drug metabolism reactions. *Drug Metab Dispos* 30, 355–362.
- [34] Korzekwa, K. R., Krishnamachary, N., Shou, M., Ogai, A., Parise, R. A., Rettie, A. E., Gonzalez, F. J., and Tracy, T. S. (1998) Evaluation of atypical cytochrome P450 kinetics with two-substrate models: evidence that multiple substrates can simultaneously bind to cytochrome P450 active sites. *Biochemistry* 37, 4137–4147.
- [35] Williams, P. A., Cosme, J., Ward, A., Angove, H. C., Matak Vinković, D., and Jhoti, H. (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature* 424, 464–468.
- [36] Wester, M. R., Yano, J. K., Schoch, G. A., Yang, C., Griffin, K. J., Stout, C. D., and Johnson, E. F. (2004) The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0-Å resolution. *J Biol Chem* 279, 35630–35637.
- [37] Dickmann, L. J., Locuson, C. W., Jones, J. P., and Rettie, A. E. (2004) Differential roles of Arg97, Asp293, and Arg108 in enzyme stability and substrate specificity of CYP2C9. *Mol Pharmacol* 65, 842–850.
- [38] Melet, A., Assrir, N., Jean, P., Pilar Lopez-Garcia, M., Marques-Soares, C., Jaouen, M., Dansette, P. M., Sari, M. A., and Mansuy, D. (2003) Substrate selectivity of human cytochrome P450 2C9: importance of residues 476, 365, and 114 in recognition of diclofenac and sulfaphenazole and in mechanism-based inactivation by tienilic acid. *Arch Biochem Biophys* 409, 80–91.
- [39] Mosher, C. M., Hummel, M. A., Tracy, T. S., and Rettie, A. E. (2008) Functional analysis of phenylalanine residues in the active site of cytochrome P450 2C9. *Biochemistry* 47, 11725–11734.
- [40] Williams, P. A., Cosme, J., Vinkovic, D. M., Ward, A., Angove, H. C., Day, P. J., Vornrhein, C., Tickle, I. J., and Jhoti, H. (2004) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone.

- Science* 305, 683–686.
- [41] Yano, J. K., Wester, M. R., Schoch, G. A., Griffin, K. J., Stout, C. D., and Johnson, E. F. (2004) The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-Å resolution. *J Biol Chem* 279, 38091–38094.
- [42] Ekroos, M., and Sjögren, T. (2006) Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc Natl Acad Sci USA* 103, 13682–13687.
- [43] Sevrioukova, I. F., and Poulos, T. L. (2010) Structure and mechanism of the complex between cytochrome P4503A4 and ritonavir. *Proc Natl Acad Sci USA* 107, 18422–18427.
- [44] He, Y. A., He, Y. Q., Szklarz, G. D., and Halpert, J. R. (1997) Identification of three key residues in substrate recognition site 5 of human cytochrome P450 3A4 by cassette and site-directed mutagenesis. *Biochemistry* 36, 8831–8839.
- [45] Harlow, G. R., and Halpert, J. R. (1998) Analysis of human cytochrome P450 3A4 cooperativity: construction and characterization of a site-directed mutant that displays hyperbolic steroid hydroxylation kinetics. *Proc Natl Acad Sci USA* 95, 6636–6641.
- [46] Stevens, J. C., Domanski, T. L., Harlow, G. R., White, R. B., Orton, E., and Halpert, J. R. (1999) Use of the steroid derivative RPR 106541 in combination with site-directed mutagenesis for enhanced cytochrome P-450 3A4 structure/function analysis. *J Pharmacol Exp Ther* 290, 594–602.
- [47] Roussel, F., Khan, K. K., and Halpert, J. R. (2000) The importance of SRS-1 residues in catalytic specificity of human cytochrome P450 3A4. *Arch Biochem Biophys* 374, 269–278.
- [48] Domanski, T. L., He, Y. A., Khan, K. K., Roussel, F., Wang, Q., and Halpert, J. R. (2001) Phenylalanine and tryptophan scanning mutagenesis of CYP3A4 substrate recognition site residues and effect on substrate oxidation and cooperativity. *Biochemistry* 40, 10150–10160.
- [49] Fowler, S. M., Taylor, J. M., Friedberg, T., Wolf, C. R., and Riley, R. J. (2002) CYP3A4 active site volume modification by mutagenesis of leucine 211. *Drug Metab Dispos* 30, 452–456.
- [50] Khan, K. K., He, Y. Q., Domanski, T. L., and Halpert, J. R. (2002) Midazolam oxidation by cytochrome P450 3A4 and active-site mutants: an evaluation of multiple binding sites and of the metabolic pathway that leads to

- enzyme inactivation. *Mol Pharmacol* 61, 495–506.
- [51] He, Y. A., Roussel, F., and Halpert, J. R. (2003) Analysis of homotropic and heterotropic cooperativity of diazepam oxidation by CYP3A4 using site-directed mutagenesis and kinetic modeling. *Arch Biochem Biophys* 409, 92–101.
- [52] Moore, C. D., Shahrokh, K., Sontum, S. F., Cheatham, r., Thomas E, and Yost, G. S. (2010) Improved cytochrome P450 3A4 molecular models accurately predict the Phe215 requirement for raloxifene dehydrogenation selectivity. *Biochemistry* 49, 9011–9019.
- [53] Ortiz de Montellano, P. R. *Cytochrome P450: Structure, Mechanism and Biochemistry*, 2nd ed.; Plenum Press: New York, 1995; pp 305–364.
- [54] Prakash, C., Shaffer, C. L., and Nedderman, A. (2007) Analytical strategies for identifying drug metabolites. *Mass Spectrom Rev* 26, 340–369.
- [55] Tolonen, A., Turpeinen, M., and Pelkonen, O. (2009) Liquid chromatography-mass spectrometry in in vitro drug metabolite screening. *Drug Discov Today* 14, 120–133.
- [56] Mauriala, T., Chauret, N., Oballa, R., Nicoll-Griffith, D. A., and Bateman, K. P. (2005) A strategy for identification of drug metabolites from dried blood spots using triple-quadrupole/linear ion trap hybrid mass spectrometry. *Rapid Commun Mass Spectrom* 19, 1984–1992.
- [57] Zhang, H., Zhang, D., and Ray, K. (2003) A software filter to remove interference ions from drug metabolites in accurate mass liquid chromatography/mass spectrometric analyses. *J Mass Spectrom* 38, 1110–1112.
- [58] Zhang, H., Zhang, D., Ray, K., and Zhu, M. (2009) Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry. *J Mass Spectrom* 44, 999–1016.
- [59] Lu, A. Y. H., Wang, R. W., and Lin, J. H. (2003) Cytochrome P450 in vitro reaction phenotyping: a re-evaluation of approaches used for P450 isoform identification. *Drug Metab Dispos* 31, 345–350.
- [60] Eagling, V. A., Tjia, J. F., and Back, D. J. (1998) Differential selectivity of cytochrome P450 inhibitors against probe substrates in human and rat liver microsomes. *Br J Clin Pharmacol* 45, 107–114.
- [61] Smith, D. A., Abel, S. M., Hyland, R., and Jones, B. C. (1998) Human cytochrome P450s: selectivity and measurement in vivo. *Xenobiotica* 28, 1095–1128.

- [62] Zlokarnik, G., Grootenhuis, P. D. J., and Watson, J. B. (2005) High throughput P450 inhibition screens in early drug discovery. *Drug Discov Today* 10, 1443–1450.
- [63] Walsky, R. L., and Obach, R. S. (2004) Validated assays for human cytochrome P450 activities. *Drug Metab Dispos* 32, 647–660.
- [64] Miller, V. P., Stresser, D. M., Blanchard, A. P., Turner, S., and Crespi, C. L. (2000) Fluorometric high-throughput screening for inhibitors of cytochrome P450. *Ann NY Acad Sci* 919, 26–32.
- [65] Stresser, D. M., Turner, S. D., Blanchard, A. P., Miller, V. P., and Crespi, C. L. (2002) Cytochrome P450 fluorometric substrates: identification of isoform-selective probes for rat CYP2D2 and human CYP3A4. *Drug Metab Dispos* 30, 845–852.
- [66] Cali, J. J., Ma, D., Sobol, M., Simpson, D. J., Frackman, S., Good, T. D., Daily, W. J., and Liu, D. (2006) Luminogenic cytochrome P450 assays. *Expert Opin Drug Metab Toxicol* 2, 629–645.
- [67] Zuegge, J., Fechner, U., Roche, O., Parrott, N. J., Engkvist, O., and Schneider, G. (2002) A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quant Struc Act Relat* 21, 249–256.
- [68] Ekins, S., Berbaum, J., and Harrison, R. K. (2003) Generation and validation of rapid computational filters for CYP2D6 and CYP3A4. *Drug Metab Dispos* 31, 1077–1080.
- [69] Kriegel, J. M., Arnhold, T., Beck, B., and Fox, T. (2005) A support vector machine approach to classify human cytochrome P450 3A4 inhibitors. *J Comput Aid Mol Des* 19, 189–201.
- [70] Kriegel, J. M., Eriksson, L., Arnhold, T., Beck, B., Johansson, E., and Fox, T. (2005) Multivariate modeling of cytochrome P450 3A4 inhibition. *Eur J Pharm Sci* 24, 451–463.
- [71] Arimoto, R., Prasad, M., and Gifford, E. M. (2005) Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors. *J Biomol Screen* 10, 197–205.
- [72] Mao, B., Gozalbes, R., Barbosa, F., Migeon, J., Merrick, S., Kamm, K., Wong, E., Costales, C., Shi, W., Wu, C., and Froloff, N. (2006) QSAR modeling of in vitro inhibition of cytochrome P450 3A4. *J Chem Inf Model* 46, 2125–2134.

- [73] Jensen, B. F., Vind, C., Padkjaer, S. B., Brockhoff, P. B., and Refsgaard, H. H. F. (2007) In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J Med Chem* 50, 501–511.
- [74] Gleeson, M. P., Davis, A. M., Chohan, K. K., Paine, S. W., Boyer, S., Gavaghan, C. L., Arnby, C. H., Kankkonen, C., and Albertson, N. (2007) Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models. *J Comput Aid Mol Des* 21, 559–573.
- [75] Choi, I., Kim, S. Y., Kim, H., Kang, N. S., Bae, M. A., Yoo, S., Jung, J., and No, K. T. (2009) Classification models for CYP450 3A4 inhibitors and non-inhibitors. *Eur J Med Chem* 44, 2354–2360.
- [76] Weaver, S., and Gleeson, M. P. (2008) The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model* 26, 1315–1326.
- [77] Ekins, S., Bravi, G., Binkley, S., Gillespie, J. S., Ring, B. J., Wikel, J. H., and Wrighton, S. A. (1999) Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors. *J Pharmacol Exp Ther* 290, 429–438.
- [78] Galetin, A., Clarke, S. E., and Houston, J. B. (2003) Multisite kinetic analysis of interactions between prototypical CYP3A4 subgroup substrates: midazolam, testosterone, and nifedipine. *Drug Metab Dispos* 31, 1108–1116.
- [79] Wang, R. W., Newton, D. J., Liu, N., Atkins, W. M., and Lu, A. Y. (2000) Human cytochrome P-450 3A4: in vitro drug-drug interaction patterns are substrate-dependent. *Drug Metab Dispos* 28, 360–366.
- [80] Lu, P., Lin, Y., Rodrigues, A. D., Rushmore, T. H., Baillie, T. A., and Shou, M. (2001) Testosterone, 7-benzyloxyquinoline, and 7-benzyloxy-4-trifluoromethyl-coumarin bind to different domains within the active site of cytochrome P450 3A4. *Drug Metab Dispos* 29, 1473–1479.
- [81] Katoh, M., Nakajima, M., Shimada, N., Yamazaki, H., and Yokoi, T. (2000) Inhibition of human cytochrome P450 enzymes by 1,4-dihydropyridine calcium antagonists: prediction of in vivo drug-drug interactions. *Eur J Clin Pharmacol* 55, 843–852.
- [82] Darvas, F. (1988) Predicting metabolic pathways by logic programming. *J Mol Graph* 6, 80–86.
- [83] Klopman, G., Dimayuga, M., and Talafous, J. (1994) META. 1. A program

- for the evaluation of metabolic transformation of chemicals. *J Chem Inf Comput Sci* 34, 1320–1325.
- [84] Talafous, J., Sayre, L. M., Mieyal, J. J., and Klopman, G. (1994) META. 2. A dictionary model of mammalian xenobiotic metabolism. *J Chem Inf Comput Sci* 34, 1326–1333.
- [85] Klopman, G., Tu, M., and Talafous, J. (1997) META. 3. A genetic algorithm for metabolic transform priorities optimization. *J Chem Inf Comput Sci* 37, 329–334.
- [86] Button, W. G., Judson, P. N., Long, A., and Vessey, J. D. (2003) Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *J Chem Inf Comput Sci* 43, 1371–1377.
- [87] Korzekwa, K. R., Jones, J. P., and Gillette, J. R. (1990) Theoretical studies on cytochrome P-450 mediated hydroxylation: a predictive model for hydrogen atom abstractions. *J Am Chem Soc* 112, 7042–7046.
- [88] Jones, J. P., Mysinger, M., and Korzekwa, K. R. (2002) Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metab Dispos* 30, 7–12.
- [89] Olsen, L., Rydberg, P., Rod, T. H., and Ryde, U. (2006) Prediction of activation energies for hydrogen abstraction by cytochrome P450. *J Med Chem* 49, 6489–6499.
- [90] Rydberg, P., Ryde, U., and Olsen, L. (2008) Prediction of activation energies for aromatic oxidation by cytochrome P450. *J Phys Chem A* 112, 13058–13065.
- [91] Sheridan, R. P., Korzekwa, K. R., Torres, R. A., and Walker, M. J. (2007) Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J Med Chem* 50, 3173–3184.
- [92] Boyer, S., Arnby, C. H., Carlsson, L., Smith, J., Stein, V., and Glen, R. C. (2007) Reaction site mapping of xenobiotic biotransformations. *J Chem Inf Model* 47, 583–590.
- [93] Smith, J., and Stein, V. (2009) SPORCalc: A development of a database analysis that provides putative metabolic enzyme reactions for ligand-based drug design. *Comput Biol Chem* 33, 149–159.
- [94] Carlsson, L., Spjuth, O., Adams, S., Glen, R. C., and Boyer, S. (2010) Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinformatics* 11,

- 362.
- [95] Ridder, L., and Wagener, M. (2008) SyGMA: combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* 3, 821–832.
- [96] Hennemann, M., Friedl, A., Lobell, M., Keldenich, J., Hillisch, A., Clark, T., and Göller, A. H. (2009) CypScore: quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory. *ChemMedChem* 4, 657–669.
- [97] Zheng, M., Luo, X., Shen, Q., Wang, Y., Du, Y., Zhu, W., and Jiang, H. (2009) Site of metabolism prediction for six biotransformations mediated by cytochromes P450. *Bioinformatics* 25, 1251–1258.
- [98] Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., and Vianello, R. (2005) MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J Med Chem* 48, 6970–6979.
- [99] Zhou, D., Afzelius, L., Grimm, S. W., Andersson, T. B., Zauhar, R. J., and Zamora, I. (2006) Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metab Dispos* 34, 976–983.
- [100] Afzelius, L., Arnby, C. H., Broo, A., Carlsson, L., Isaksson, C., Jurva, U., Kjellander, B., Kolmodin, K., Nilsson, K., Raubacher, F., and Weidolf, L. (2007) State-of-the-art tools for computational site of metabolism predictions: comparative analysis, mechanistical insights, and future applications. *Drug Metab Rev* 39, 61–86.
- [101] Oh, W. S., Kim, D. N., Jung, J., Cho, K., and No, K. T. (2008) New combined model for the prediction of regioselectivity in cytochrome P450/3A4 mediated metabolism. *J Chem Inf Model* 48, 591–601.
- [102] Tarcsay, A., Kiss, R., and Keseru, G. M. (2010) Site of metabolism prediction on cytochrome P450 2C9: a knowledge-based docking approach. *J Comput Aid Mol Des* 24, 399–408.
- [103] Testa, B., Balmat, A., Long, A., and Judson, P. (2005) Predicting drug metabolism—an evaluation of the expert system METEOR. *Chem Biodivers* 2, 872–885.
- [104] de Visser, S. P., Kumar, D., Cohen, S., Shacham, R., and Shaik, S. (2004) A Predictive Pattern of Computed Barriers for C-H Hydroxylation by

- Compound I of Cytochrome P450. *J Am Chem Soc* 126, 8362–8363.
- [105] Bathelt, C. M., Ridder, L., Mulholland, A. J., and Harvey, J. N. (2004) Mechanism and structure-reactivity relationships for aromatic hydroxylation by cytochrome P450. *Org Biomol Chem* 2, 2998–3005.
- [106] Mayeno, A. N., Robinson, J. L., Yang, R. S. H., and Reisfeld, B. (2009) Predicting activation enthalpies of cytochrome-P450-mediated hydrogen abstractions. 2. Comparison of semiempirical PM3, SAM1, and AM1 with a density functional theory method. *J Chem Inf Model* 49, 1692–1703.
- [107] Rydberg, P., Vasanthanathan, P., Oostenbrink, C., and Olsen, L. (2009) Fast prediction of cytochrome P450 mediated drug metabolism. *ChemMedChem* 4, 2070–2079.
- [108] Rydberg, P., Gloriam, D. E., Zaretski, J., Breneman, C., and Olsen, L. (2010) SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med Chem Lett* 1, 96–100.
- [109] de Graaf, C., Oostenbrink, C., Keizers, P. H. J., van der Wijst, T., Jongejan, A., and Vermeulen, N. P. E. (2006) Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J Med Chem* 49, 2417–2430.
- [110] Hritz, J., de Ruiter, A., and Oostenbrink, C. (2008) Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J Med Chem* 51, 7469–7477.
- [111] Santos, R., Hritz, J., and Oostenbrink, C. (2010) Role of water in molecular docking simulations of cytochrome P450 2D6. *J Chem Inf Model* 50, 146–154.
- [112] Oláh, J., Mulholland, A. J., and Harvey, J. N. (2011) Understanding the determinants of selectivity in drug metabolism through modeling of dextromethorphan oxidation by cytochrome P450. *Proc Natl Acad Sci USA* 108, 6050–6055.
- [113] Sykes, M. J., McKinnon, R. A., and Miners, J. O. (2008) Prediction of metabolism by cytochrome P450 2C9: alignment and docking studies of a validated database of substrates. *J Med Chem* 51, 780–791.
- [114] Jung, J., Kim, N. D., Kim, S. Y., Choi, I., Cho, K., Oh, W. S., Kim, D. N., and No, K. T. (2008) Regioselectivity prediction of CYP1A2-mediated phase I metabolism. *J Chem Inf Model* 48, 1074–1080.
- [115] Vasanthanathan, P., Hritz, J., Taboureau, O., Olsen, L., Jorgensen, F. S.,

- Vermeulen, N. P. E., and Oostenbrink, C. (2009) Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. *J Chem Inf Model* 49, 43–52.
- [116] Li, H., and Poulos, T. L. (2004) Crystallization of cytochromes P450 and substrate-enzyme interactions. *Curr Top Med Chem* 4, 1789–1802.
- [117] Schoch, G. A., Yano, J. K., Wester, M. R., Griffin, K. J., Stout, C. D., and Johnson, E. F. (2004) Structure of human microsomal cytochrome P450 2C8. Evidence for a peripheral fatty acid binding site. *J Biol Chem* 279, 9497–9503.
- [118] Fishelovitch, D., Hazan, C., Shaik, S., Wolfson, H. J., and Nussinov, R. (2007) Structural dynamics of the cooperative binding of organic molecules in the human cytochrome P450 3A4. *J Am Chem Soc* 129, 1602–1611.
- [119] Warren, G. L., Andrews, C. W., Capelli, A., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., and Head, M. S. (2006) A critical assessment of docking programs and scoring functions. *J Med Chem* 49, 5912–5931.
- [120] Goodford, P. J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28, 849–857.
- [121] Caron, G., Ermondi, G., and Testa, B. (2007) Predicting the oxidative metabolism of statins: an application of the MetaSite algorithm. *Pharm Res* 24, 480–501.
- [122] Trunzer, M., Faller, B., and Zimmerlin, A. (2009) Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC-MS/MS validation. *J Med Chem* 52, 329–335.
- [123] Dapkunas, J., Sazonovas, A., and Japertas, P. (2009) Probabilistic prediction of the human CYP3A4 and CYP2D6 metabolism sites. *Chem Biodivers* 6, 2101–2106.
- [124] Stresser, D. M., Blanchard, A. P., Turner, S. D., Erve, J. C., Dandeneau, A. A., Miller, V. P., and Crespi, C. L. (2000) Substrate-dependent modulation of CYP3A4 catalytic activity: analysis of 27 test compounds with four fluorometric substrates. *Drug Metab Dispos* 28, 1440–1448.
- [125] Nomeir, A. A., Ruegg, C., Shoemaker, M., Favreau, L. V., Palamanda, J. R., Silber, P., and Lin, C. C. (2001) Inhibition of CYP3A4 in a rapid microtiter

- plate assay using recombinant enzyme and in human liver microsomes using conventional substrates. *Drug Metab Dispos* 29, 748–753.
- [126] The PubChem Project. <http://pubchem.ncbi.nlm.nih.gov/>.
- [127] Sazonovas, A. Organinių medžiagų ūmaus toksiškumo ir metabolizmo vietos molekulėje prognozavimas taikant GALAS metodą. Ph.D. thesis, Vilniaus universitetas, Vilnius, 2010.
- [128] Tetko, I. V. (2002) Associative Neural Network. *Neur Proc Lett* 16, 187–199.
- [129] Tetko, I. V. (2002) Neural network studies. 4. Introduction to Associative Neural Networks. *J Chem Inf Comput Sci* 42, 717–728.
- [130] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *Ann Statist* 7, 1–26.
- [131] Fawcett, T. (2006) An introduction to ROC analysis. *Patt Recog Lett* 27, 861–874.
- [132] Velaparthi, U. et al. (2008) Discovery and evaluation of 4-(2-(4-chloro-1H-pyrazol-1-yl)ethylamino)-3-(6-(1-(3-fluoropropyl)piperidin-4-yl)-4-methyl-1H-benzo[d]imidazol-2-yl)pyridin-2(1H)-one (BMS-695735), an orally efficacious inhibitor of insulin-like growth factor-1 receptor kinase with broad spectrum in vivo antitumor activity. *J Med Chem* 51, 5897–5900.
- [133] Rydberg, P., Gloriam, D. E., and Olsen, L. (2010) The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* 26, 2988–2989.
- [134] Uphagrove, A. L., and Nelson, W. L. (2001) Importance of amine pKa and distribution coefficient in the metabolism of fluorinated propranolol derivatives. Preparation, identification of metabolite regioisomers, and metabolism by CYP2D6. *Drug Metab Dispos* 29, 1377–1388.
- [135] Uphagrove, A. L., and Nelson, W. L. (2001) Importance of amine pKa and distribution coefficient in the metabolism of fluorinated propranolol analogs: metabolism by CYP1A2. *Drug Metab Dispos* 29, 1389–1395.
- [136] Japertas, P., Didziapetris, R., and Petrauskas, A. (2002) Fragmental methods in the design of new compounds. Applications of the Advanced Algorithm Builder. *Quant Struc Act Relat* 21, 23–37.
- [137] Algorithm Builder 1.8. ACD/Labs Inc., Toronto, ON, Canada, <http://www.acdlabs.com>.
- [138] R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.r-project.org/>.

- [139] Eng, J. ROC analysis: web-based calculator for ROC curves. 2006; <http://www.jrocfite.org>.
- [140] ADME Suite 4.95. ACD/Labs Inc., Toronto, ON, Canada, <http://www.acdlabs.com>.
- [141] Gleeson, M. P. (2008) Generation of a set of simple, interpretable ADMET rules of thumb. *J Med Chem* 51, 817–834.
- [142] Monostory, K., Vereczkey, L., Lévai, F., and Szatmári, I. (1998) Ipriflavone as an inhibitor of human cytochrome P450 enzymes. *Br J Pharmacol* 123, 605–610.
- [143] Moon, Y., Kim, S. Y., Ji, H. Y., Kim, Y. K., Chae, H. J., Chae, S. W., and Lee, H. S. (2007) Characterization of cytochrome P450s mediating ipriflavone metabolism in human liver microsomes. *Xenobiotica* 37, 246–259.
- [144] Gleeson, P., Bravi, G., Modi, S., and Lowe, D. (2009) ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg Med Chem* 17, 5906–5919.
- [145] Ishigami, M., Honda, T., Takasaki, W., Ikeda, T., Komai, T., Ito, K., and Sugiyama, Y. (2001) A comparison of the effects of 3-hydroxy-3-methylglutaryl-coenzyme a (HMG-CoA) reductase inhibitors on the CYP3A4-dependent oxidation of mexazolam in vitro. *Drug Metab Dispos* 29, 282–288.
- [146] Tsukamoto, I., Koshio, H., Kuramochi, T., Saitoh, C., Yanai-Inamura, H., Kitada-Nozawa, C., Yamamoto, E., Yatsu, T., Shimada, Y., Sakamoto, S., and Tsukamoto, S.-i. (2009) Synthesis and structure-activity relationships of amide derivatives of (4,4-difluoro-1,2,3,4-tetrahydro-5H-1-benzazepin-5-ylidene)acetic acid as selective arginine vasopressin V2 receptor agonists. *Bioorg Med Chem* 17, 3130–3141.
- [147] Zhou, S., Yung Chan, S., Cher Goh, B., Chan, E., Duan, W., Huang, M., and McLeod, H. L. (2005) Mechanism-based inhibition of cytochrome P450 3A4 by therapeutic drugs. *Clin Pharmacokinet* 44, 279–304.
- [148] Meyer, M. R., Peters, F. T., and Maurer, H. H. (2009) Investigations on the human hepatic cytochrome P450 isozymes involved in the metabolism of 3,4-methylenedioxy-amphetamine (MDA) and benzodioxolyl-butanamine (BDB) enantiomers. *Toxicol Lett* 190, 54–60.
- [149] Yan, Z., Zhong, H. M., Maher, N., Torres, R., Leo, G. C., Caldwell, G. W., and Huebert, N. (2005) Bioactivation of 4-methylphenol (p-cresol) via cyto-

- chrome P450-mediated aromatic oxidation in human liver microsomes. *Drug Metab Dispos* 33, 1867–1876.
- [150] Chou, Y., Chung, Y., Liu, T., Wang, S., Chau, G., Chi, C., Soucek, P., Krausz, K. W., Gelboin, H. V., Lee, C., and Ueng, Y. (2010) The oxidative metabolism of dimemorfan by human cytochrome P450 enzymes. *J Pharm Sci* 99, 1063–1077.
- [151] Baer, B. R., DeLisle, R. K., and Allen, A. (2009) Benzylic oxidation of gemfibrozil-1-O-beta-glucuronide by P450 2C8 leads to heme alkylation and irreversible inhibition. *Chem Res Toxicol* 22, 1298–1309.
- [152] Yu, A., Idle, J. R., Krausz, K. W., Küpfer, A., and Gonzalez, F. J. (2003) Contribution of individual cytochrome P450 isozymes to the O-demethylation of the psychotropic beta-carboline alkaloids harmaline and harmine. *J Pharmacol Exp Ther* 305, 315–322.
- [153] Zhang, J., Liu, Y., Zhao, J., Wang, L., Ge, G., Gao, Y., Li, W., Liu, H., Liu, H., Zhang, Y., Sun, J., and Yang, L. (2008) Metabolic profiling and cytochrome P450 reaction phenotyping of medroxyprogesterone acetate. *Drug Metab Dispos* 36, 2292–2298.
- [154] Shen, H., Wu, C., Jiang, X., and Yu, A. (2010) Effects of monoamine oxidase inhibitor and cytochrome P450 2D6 status on 5-methoxy-N,N-dimethyltryptamine metabolism and pharmacokinetics. *Biochem Pharmacol* 80, 122–128.
- [155] Gardner, I., Wakazono, H., Bergin, P., de Waziers, I., Beaune, P., Kenna, J. G., and Caldwell, J. (1997) Cytochrome P450 mediated bioactivation of methyleugenol to 1'-hydroxymethyleugenol in Fischer 344 rat and human liver microsomes. *Carcinogenesis* 18, 1775–1783.
- [156] Jeurissen, S. M. F., Bogaards, J. J. P., Boersma, M. G., ter Horst, J. P. F., Awad, H. M., Fiamegos, Y. C., van Beek, T. A., Alink, G. M., Sudhölter, E. J. R., Cnubben, N. H. P., and Rietjens, I. M. C. M. (2006) Human cytochrome p450 enzymes of importance for the bioactivation of methyleugenol to the proximate carcinogen 1'-hydroxymethyleugenol. *Chem Res Toxicol* 19, 111–116.
- [157] Miksanová, M., Sulc, M., Rýdlová, H., Schmeiser, H. H., Frei, E., and Stiborová, M. (2004) Enzymes involved in the metabolism of the carcinogen 2-nitroanisole: evidence for its oxidative detoxication by human cytochromes P450. *Chem Res Toxicol* 17, 663–671.

- [158] Yuan, Z., Xu, H., Wang, K., Zhao, Z., and Hu, M. (2009) Determination of osthol and its metabolites in a phase I reaction system and the Caco-2 cell model by HPLC-UV and LC-MS/MS. *J Pharm Biomed Anal* 49, 1226–1232.
- [159] Jiang, X., Shen, H., and Yu, A. (2009) Pinoline may be used as a probe for CYP2D6 activity. *Drug Metab Dispos* 37, 443–446.
- [160] Kotsuma, M., Tokui, T., Ishizuka-Ozeki, T., Honda, T., Iwabuchi, H., Murai, T., Ikeda, T., and Saji, H. (2008) CYP2D6-Mediated metabolism of a novel acyl coenzyme A:cholesterol acyltransferase inhibitor, pactimibe, and its unique plasma metabolite, R-125528. *Drug Metab Dispos* 36, 529–534.
- [161] Bae, S. K., Cao, S., Seo, K., Kim, H., Kim, M., Shon, J., Liu, K., Zhou, H., and Shin, J. (2008) Cytochrome P450 2B6 catalyzes the formation of pharmacologically active sibutramine (N-1-[1-(4-chlorophenyl)cyclobutyl]-3-methylbutyl-N,N-dimethylamine) metabolites in human liver microsomes. *Drug Metab Dispos* 36, 1679–1688.
- [162] Franklin, M. R., and Hathaway, L. B. (2008) 2-Diethylaminoethyl-2,2-diphenylvalerate-HCl (SKF525A) revisited: comparative cytochrome P450 inhibition in human liver microsomes by SKF525A, its metabolites, and SKF-acid and SKF-alcohol. *Drug Metab Dispos* 36, 2539–2546.
- [163] Vermeir, M., Hemeryck, A., Cuyckens, F., Francesch, A., Bockx, M., Van Houdt, J., Steemans, K., Mannens, G., Avilés, P., and De Coster, R. (2009) In vitro studies on the metabolism of trabectedin (YONDELIS) in monkey and man, including human CYP reaction phenotyping. *Biochem Pharmacol* 77, 1642–1654.
- [164] Ku, H., Ahn, H., Seo, K., Kim, H., Oh, M., Bae, S. K., Shin, J., Shon, J., and Liu, K. (2008) The contributions of cytochromes P450 3A4 and 3A5 to the metabolism of the phosphodiesterase type 5 inhibitors sildenafil, udenafil, and vardenafil. *Drug Metab Dispos* 36, 986–990.
- [165] Abass, K., Reponen, P., Mattila, S., and Pelkonen, O. (2009) Metabolism of carbosulfan. I. Species differences in the in vitro biotransformation by mammalian hepatic microsomes including human. *Chem Biol Interact* 181, 210–219.
- [166] Abass, K., Reponen, P., Mattila, S., and Pelkonen, O. (2010) Metabolism of carbosulfan II. Human interindividual variability in its in vitro hepatic biotransformation and the identification of the cytochrome P450 isoforms

- involved. *Chem Biol Interact* 185, 163–173.
- [167] Pearce, R. E., Lu, W., Wang, Y., Uetrecht, J. P., Correia, M. A., and Leeder, J. S. (2008) Pathways of carbamazepine bioactivation in vitro. III. The role of human cytochrome P450 enzymes in the formation of 2,3-dihydroxycarbamazepine. *Drug Metab Dispos* 36, 1637–1649.
- [168] Rudek, M. A., Zhao, M., Smith, N. F., Robey, R. W., He, P., Hallur, G., Khan, S., Hidalgo, M., Jimeno, A., Colevas, A. D., Messersmith, W. A., Wolff, A. C., and Baker, S. D. (2005) In vitro and in vivo clinical pharmacology of dimethyl benzoylphenylurea, a novel oral tubulin-interactive agent. *Clin Cancer Res* 11, 8503–8511.
- [169] Kitamura, R., Asanoma, H., Nagayama, S., and Otagiri, M. (2008) Identification of human liver cytochrome P450 isoforms involved in autoinduced metabolism of the antiangiogenic agent (Z)-5-[(1,2-dihydro-2-oxo-3H-indol-3-ylidene)methyl]-2,4-dimethyl-1H-pyrrole-3-propanoic acid (TSU-68). *Drug Metab Dispos* 36, 1003–1009.
- [170] Evanchik, M. J., Allen, D., Yoburn, J. C., Silverman, J. A., and Hoch, U. (2009) Metabolism of (+)-1,4-dihydro-7-(trans-3-methoxy-4-methylamino-1-pyrrolidinyl)-4-oxo-1-(2-thiazolyl)-1,8-naphthyridine-3-carboxylic acid (voreloxin; formerly SNS-595), a novel replication-dependent DNA-damaging agent. *Drug Metab Dispos* 37, 594–601.
- [171] Prakash, C., Wang, W., O'Connell, T., and Johnson, K. A. (2008) CYP2C8- and CYP3A-mediated C-demethylation of (3-[(4-tert-butylbenzyl)-(pyridine-3-sulfonyl)-amino]-methyl-phenoxy)-acetic acid (CP-533,536), an EP2 receptor-selective prostaglandin E2 agonist: characterization of metabolites by high-resolution liquid chromatography-tandem mass spectrometry and liquid chromatography/mass spectrometry-nuclear magnetic resonance. *Drug Metab Dispos* 36, 2093–2103.
- [172] Wang, L., Christopher, L. J., Cui, D., Li, W., Iyer, R., Humphreys, W. G., and Zhang, D. (2008) Identification of the human enzymes involved in the oxidative metabolism of dasatinib: an effective approach for determining metabolite formation kinetics. *Drug Metab Dispos* 36, 1828–1839.
- [173] Pearce, R. E., Leeder, J. S., and Kearns, G. L. (2006) Biotransformation of fluticasone: in vitro characterization. *Drug Metab Dispos* 34, 1035–1040.
- [174] Ghibellini, G., Bridges, A. S., Generaux, C. N., and Brouwer, K. L. R. (2007) In vitro and in vivo determination of piperacillin metabolism in humans.

- Drug Metab Dispos* 35, 345–349.
- [175] Lang, W., Caldwell, G. W., Li, J., Leo, G. C., Jones, W. J., and Masucci, J. A. (2007) Biotransformation of geldanamycin and 17-allylamino-17-demethoxygeldanamycin by human liver microsomes: reductive versus oxidative metabolism and implications. *Drug Metab Dispos* 35, 21–29.
- [176] Minato, K., Suzuki, R., Asagarasu, A., Matsui, T., and Sato, M. (2008) Biotransformation of 3-amino-5,6,7,8-tetrahydro-2-4-[4-(quinolin-2-yl)piperazin-1-yl]butylquinazolin-4(3H)-one (TZB-30878), a novel 5-hydroxytryptamine (5-HT)_{1A} agonist/5-HT₃ antagonist, in human hepatic cytochrome P450 enzymes. *Drug Metab Dispos* 36, 831–840.
- [177] Deo, A. K., and Bandiera, S. M. (2008) Identification of human hepatic cytochrome p450 enzymes involved in the biotransformation of cholic and chenodeoxycholic acid. *Drug Metab Dispos* 36, 1983–1991.
- [178] Goda, R., Nagai, D., Akiyama, Y., Nishikawa, K., Ikemoto, I., Aizawa, Y., Nagata, K., and Yamazoe, Y. (2006) Detection of a new N-oxidized metabolite of flutamide, N-[4-nitro-3-(trifluoromethyl)phenyl]hydroxylamine, in human liver microsomes and urine of prostate cancer patients. *Drug Metab Dispos* 34, 828–835.
- [179] Kang, P., Dalvie, D., Smith, E., Zhou, S., Deese, A., and Nieman, J. A. (2008) Bioactivation of flutamide metabolites by human liver microsomes. *Drug Metab Dispos* 36, 1425–1437.
- [180] Deroussent, A., Ré, M., Hoellinger, H., and Cresteil, T. (2010) Metabolism of sanguinarine in human and in rat: characterization of oxidative metabolites produced by human CYP1A1 and CYP1A2 and rat liver microsomes using liquid chromatography-tandem mass spectrometry. *J Pharm Biomed Anal* 52, 391–397.
- [181] Hildebrand, A., Pfeiffer, E., and Metzler, M. (2010) Aromatic hydroxylation and catechol formation: a novel metabolic pathway of the growth promotor zeranol. *Toxicol Lett* 192, 379–386.
- [182] Pfeiffer, E., Hildebrand, A., Damm, G., Rapp, A., Cramer, B., Humpf, H., and Metzler, M. (2009) Aromatic hydroxylation is a major metabolic pathway of the mycotoxin zearalenone in vitro. *Mol Nutr Food Res* 53, 1123–1133.

List of Publications

Journal articles

1. Dapkunas, J., Sazonovas, A., and Japertas, P. (2009) Probabilistic prediction of the human CYP3A4 and CYP2D6 metabolism sites. *Chem Biodivers* 6, 2101-2106.
2. Didziapetris, R., Dapkunas, J., Sazonovas, A., and Japertas, P. (2010) Trainable structure-activity relationship model for virtual screening of CYP3A4 inhibition. *J Comput Aided Mol Des* 24, 891-906.

Oral presentations

1. Dapkunas, J., Sazonovas, A., and Japertas, P. Probabilistic prediction of the human CYP3A4 and CYP2D6 metabolism sites. LogP2009, Zürich, Switzerland, February 8-11, 2009.
2. Dapkunas, J., Sazonovas, A., and Japertas, P. Probabilistic model of regioselectivity of metabolism in human liver microsomes. ADMET Europe, Munich, Germany, March 28-29, 2011.

Posters

1. Dapkunas, J., Sazonovas, A., and Japertas, P. Probabilistic prediction of the human CYP3A4 metabolism sites in a molecule. 235th ACS National

- Meeting, New Orleans, LA, USA, April 6-10, 2008.
2. Dapkunas, J., Sazonovas, A., and Japertas, P. Probabilistic model of regioselectivity of metabolism in human liver microsomes. 16th North American ISSX Meeting, Baltimore, MD, USA, October 18-22, 2009.
 3. Sazonovas, A., Didziapetris, R., Dapkunas, J., Juska, L., and Japertas, P. GALAS modeling methodology applications in the prediction of the drug safety related properties. 16th North American ISSX Meeting, Baltimore, MD, USA, October 18-22, 2009.
 4. Dapkunas, J., Didziapetris, R., Sazonovas, A., and Japertas, P. Evaluation of ACD/ADME Suite 5.0 regioselectivity predictions. 9th International ISSX meeting, Istanbul, Turkey, September 4-8, 2010.
 5. Didziapetris, R., Dapkunas, J., Sazonovas, A., and Japertas, P. Probabilistic GALAS models for the prediction of the human cytochrome P450 inhibition. 9th International ISSX meeting, Istanbul, Turkey, September 4-8, 2010.
 6. Dapkunas, J., Sazonovas, A., Didziapetris, R., and Japertas, P. In silico identification of metabolic soft spots: case study using ACD/ADME Suite software. 241st ACS National Meeting & Exposition, Anaheim, CA, USA, March 27-31, 2011.
 7. Dapkunas, J., Sazonovas, A., and Japertas, P. QSAR model of regioselectivity of metabolism in human liver microsomes: development, validation, comparison and adaptation to novel compounds. 241st ACS National Meeting & Exposition, Anaheim, CA, USA, March 27-31, 2011.

Curriculum Vitae

Justas Dapkūnas

Date of birth: December 25, 1983.

Place of birth: Vilnius, Lithuania.

E-mail: justas.dapkunas@gmail.com

Education:

From October 2007 – PhD student of biochemistry at Vilnius University.

2007 – Vilnius University, MSc, Biochemistry.

Master thesis: “Identification of physicochemical and structural properties of human cytochrome P450 substrates”.

2005 – Vilnius University, BSc, Biochemistry.

Bachelor thesis: “*In silico* investigation of cytochrome P450 3A4 specificity”.

2001 – Vilnius Užupis Gymnasium, secondary education.

Working experience:

From June 2003 – ADME/Tox researcher at *Aukštieji Algoritmai* (part of *Pharma Algorithms* until 2009, and part of *ACD/Labs* from 2009).

Scientific interests:

Modeling of drug metabolism-related properties: prediction of regioselectivity and possible metabolites of hepatic metabolism, QSAR analysis of human cytochrome P450 inhibition.

Modeling of other ADME properties and toxicity: binding to Estrogen receptor α , carrier-mediated transport in intestine, genotoxicity.

Summary in Lithuanian (Santrauka)

Pagrindinis šio darbo tikslas buvo kiekybinio struktūros ir aktyvumo ryšio modelių, prognozuojančių su vaistų metabolizmu susijusias savybes, kūrimas. Modeliai, prognozuojantys CYP3A4 slopinimą ir žmogaus kepenų mikrosomų katalizuojamo metabolizmo regioselektyvumą, buvo sukurti naudojant GALAS (angl. Global, Adjusted Locally According to Similarity; Globalus, lokaliai pakoreguotas pagal panašumą) modeliavimo metodą, kuris geba įvertinti prognozės patikimumą, taip apibrėždamas modelio pritaikymo sritį. Sukurtų modelių prognozės buvo tikrinamos naudojant eksperimentinius naujų cheminių junginių duomenis. Visų globalių modelių prognozės gerėjo po korekcijų pagal panašumą, o neteisingų spėjimų skaičius buvo ženkliai mažesnis tarp aukšto patikimumo prognozių. Visgi daugiau nei pusė išorinių duomenų nepatenka į šių modelių pritaikymo sritį. GALAS modeliai gali būti gana paprastai apmokomi, pridėdant naujus duomenis į lokalią modelio dalį ir apskaičiuojant reikiamą korekciją. Po tokios apmokymo procedūros CYP3A4 slopinimo modelis prisitaikė prie PubChem duomenų bazės cheminių junginių ir taip pat prie vaistų, turinčių naują cheminį karkasą. Pridėjus naujų junginių ir apmokius regioselektyvumo modelį, jis pradėjo prognozuoti naujas metabolizmo vietas. Pastarasis modelis taip pat buvo pritaikytas atskirų fermentų katalizuojamo metabolizmo prognozavimui.