DOI: 10.1070/RCR4614

Computer-aided prediction of xenobiotic metabolism in humans

Vladislav M. Bezhentsev,[a] Olga A. Tarasova,[a] Aleksandr V. Dmitriev,[a] Anastasiya V. Rudik,[a] Aleksey A. Lagunin,[a, b] Dmitriy A. Filimonov,[a] Vladimir V. Poroikov [a]

[a] *Institute of Biomedical Chemistry*
ul. Pogodinskaya 10/8, 119121 Moscow, Russian Federation
[b] *Pirogov Russian National Research Medical University*
ul. Ostrovityanova 1, 117997 Moscow, Russian Federation

The review describes the main databases containing information about the metabolism of xenobiotics, including data on drug metabolism, metabolic enzymes, schemes of biotransformation and the structures of some substrates and metabolites. Computational approaches used to predict the interaction of xenobiotics with metabolic enzymes, to identify the sites of metabolism in the molecule and to generate structures of potential metabolites for subsequent evaluation of their properties are considered. The advantages and limitations of particular computational methods for metabolism prediction are indicated and the prospects for their applications to improve the safety and efficacy of new drugs are discussed.
The bibliography includes 165 references.

Contents

I.	Introduction	854
II.	Computational methods of studying metabolism	857
III.	Databases containing information on the metabolism of xenobiotics	857
IV.	Prediction of interaction of xenobiotics with metabolic enzymes	861
V.	Methods based on analysis of enzyme structures	866
VI.	Prediction of the sites of biotransformation	868
VII.	Generation of the structures of probable metabolites	870
VIII.	Conclusion	875

I. Introduction

The number of unique chemical compounds included in the world largest information system, Chemical Abstracts Service (CAS), is above 100 million (www.cas.org). A major part of these compounds are xenobiotics, *i.e.*, chemical compounds that are foreign to living organisms, are not used by cells for the production of energy and do not serve as building materials for cells, cofactors or intracellular messengers.[1] Examples of xenobiotics are pesticides, radionuclides, nutritional supplements, dyes, preserving agents, substances used in household and in industry and pharmaceutical drugs.

Having got into the body, xenobiotics can disturb normal vital processes and cause intoxication [2] or allergic response;[3, 4] being accumulated in tissues, some of them can induce mutations.[5, 6] Many xenobiotics, including drugs, are metabolized (biotransformed) in the human body, that is, they undergo chemical transformations induced by human cell enzyme systems.[7, 8] In this case, the initial compound is called 'biotransformation substrate' and the product is the 'biotransformation metab-

V.M.Bezhentsev. Research Assistant at the IBMC, Student of the Higher School of Economy.
Telephone: + 7(499)255 – 3029, e-mail: vbezhentsev@gmail.com
O.A.Tarasova. Candidate of Biological Sciences, Researcher at the IBMC.
Telephone: + 7(499)255 – 3029,
e-mail: fioland@yandex.ru, olga.a.tarasova@gmail.com
A.V.Dmitriev. Candidate of Biological Sciences, Senior Researcher at the IBMC. Telephone: + 7(499)255 – 3029, e-mail: a.v.dmitriev@mail.ru
A.V.Rudik. Candidate of Biological Sciences, Senior Researcher at the IBMC. Telephone: + 7(499)255 – 3029, e-mail: rudik_anastassia@mail.ru
A.A.Lagunin. Doctor of Biological Sciences, Leading Researcher at the IBMC, Head of the Department for Bioinformatics at the RNRMU.
Telephone: + 7(499)255 – 3029, e-mail: alexey.lagunin@ibmc.msk.ru
D.A.Filimonov. Candidate of Physical and Mathematical Sciences, Leading Researcher at the IBMC.
Telephone: + 7(499)255 – 3029, e-mail: dmitry.filimonov@ibmc.msk.ru
V.V.Poroikov. Doctor of Biological Sciences, Ptofessor, Head of the Department for Bioinformatics at the IBMC.
Telephone: + 7(499)246 – 0920,
e-mail: vladimir.poroikov@ibmc.msk.ru; vvp1951@yandex.ru
Current research interests of the authors: molecular modelling, bioinformatics, chemoinformatics, medicinal chemistry, pharmacology, toxicology, systems biology, computer-aided drug design.

Received 19 November 2015
Translation: Z.P.Svitanko

© 2016 Russian Academy of Sciences and Turpion Ltd.

olite'. As a rule, a particular substrate is converted *via* multistep biotransformation reactions not to one but to several metabolites, which form a metabolic network.

Xenobiotics are metabolized in almost every tissue of the human body, but mainly this takes place in the liver. The whole set of chemical transformations that constitute the metabolism of xenobiotics and are catalyzed by diverse enzymes is conventionally subdivided into four types of metabolic reactions: oxidation, reduction, hydrolysis and conjugation, which are, in turn, historically considered as biotransformation phases I and II.[7,8] During phase I, the initial molecules are converted to more polar hydrophilic metabolites *via* oxidation, reduction or hydrolysis. Hydrophilic compounds have higher excretion rates from the body. The most widespread enzymes for the first phase of xenobiotic metabolism in the human body are five enzymes of the cytochrome P450 superfamily (isoforms CYP3A4, CYP2C9, CYP2C19, CYP2D6 and CYP1A2), which perform most of biotransformation reactions (Fig. 1).

The reactions that take place in phase II of biotransformation (so-called conjugation phase) are catalyzed by various transferases. The most abundant type of conjugation of exogenous compounds is the formation of glucuronides under the action of the UDP glucuronosyltransferase enzyme (UDP is uridine 5-diphosphate). Some researchers also distinguish the third phase of metabolism of xenobiotics, in which the metabolites react with carrier proteins.[10]

As a rule, biotransformations result in detoxification of xenobiotics and give rise to chemically stable metabolites, which possess neither pharmacological activity nor toxicity. Nevertheless, in some cases, the drug biotransformation products are either chemically reactive or pharmacologically active. Transformations of this type are called metabolic activation or bioactivation (Fig. 2).

Bioactivation can occur upon both the first and the second phases of drug metabolism. A pharmacologically active or toxic product can arise both in a single reaction and in a series of enzymatic reactions. Virtually all enzymes involved in drug metabolism can also perform bioactivation. However, the major role in this process belongs to cytochromes P450, which are most important not only for drug metabolism but also for metabolic activation.[9,11,12]

The resulting chemically reactive metabolites can interact with cells in different ways, *e.g.*, participate in the covalent binding to cell components, stimulate lipid peroxidation and destroy the membranes. In turn, pharmacologically active metabolites are usually chemically stable. These metabolites can exhibit (to a higher or lesser extent) either pharmacological properties of the starting compounds (on-target activity) or activity differ-

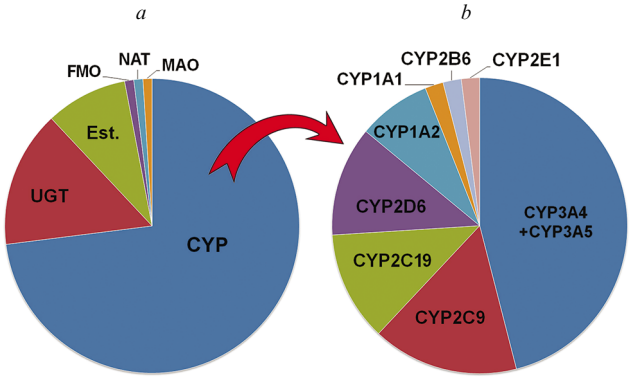


Figure 1. Role of various enzymes in drug metabolism.[9]
(*a*) The diagram shows the fractions of reactions of drug substances in the human body catalyzed by various human enzymes: cytochromes P450 (CYP), UDP glucuronyltransferase (UGT), esterase (Est), flavin-containing monooxygenases (FMO), N-acetyltransferases (NAT), monoamine oxidases (MAO); (*b*) analogous diagram for different cytochrome P450 isoforms.

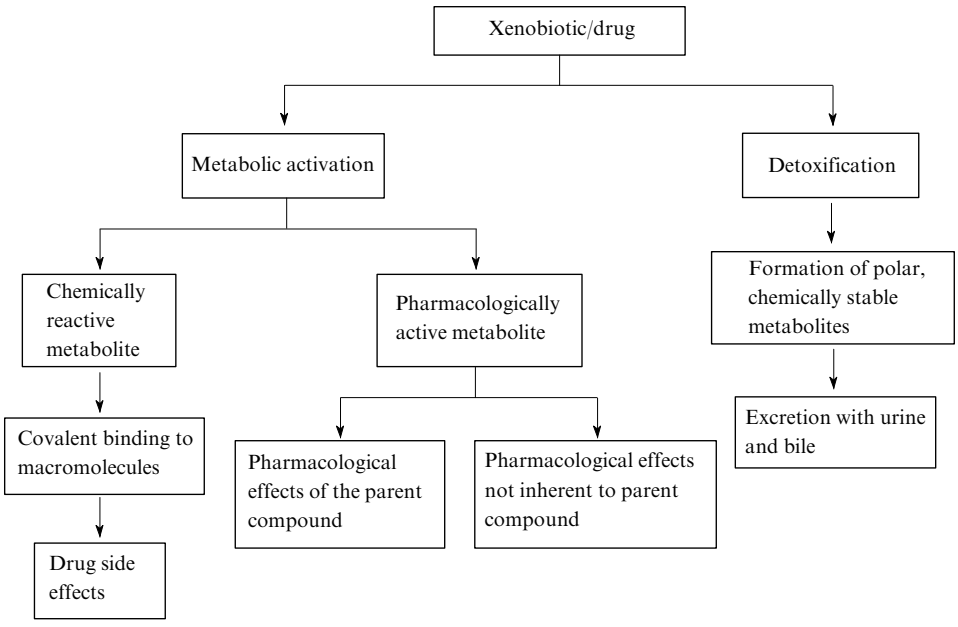


Figure 2. Drug bioactivation and detoxification.[11]

ing from that of the starting compounds (off-target activity); moreover, in some cases, the activity can be opposite to that of the starting compound. An example of this type of transformation is metabolism of the antidepressant drug nefazodone; one of metabolites of this drug, *m*-chlorophenylpiperazine, causes pronounced excitability and has been assigned to psychoactive agents.[11]

The starting drug substances that are converted to pharmacologically active metabolites can be subdivided into two groups. The first group comprises compounds that initially possess biological activity, while the second group includes prodrugs that do not possess potent biological activity before the metabolic activation.

As examples of drugs that initially possess pharmacological activity and are converted, upon biotransformation, to metabolites that also exhibit biological activity, note atorvastatin, tamoxifen, flutamide, bupropion, terfenadine, losartan, hydroxyzine, phenacetin, venlafaxine, sunitinib, fluoxetine and most of benzodiazepines and tricyclic antidepressant drugs used for therapeutic purposes.[11]

The prodrugs that are activated by cytochromes P450 include, for example, cyclophosphamide, ticlopidine, clopidogrel and prasugrel.[11] Other examples of prodrug bioactivation, including the bioactivation implemented *via* reduction rather than *via* oxidation typical of cytochromes P450, are presented in a review.[13]

A textbook case of drug toxicity arising upon metabolic activation induced by cytochrome P450 is acetaminophen, which is mainly converted under standard conditions to nontoxic conjugated metabolites, but its minor amount is transformed by cytochromes P450 into hepatotoxic *N*-acetyl-*p*-benzoquinoneimine (NAPQI) the metabolite, which is deactivated upon conjugation with glutathione. When acetaminophen is overdosed or when some other factor has led to depletion of glutathione reserves of the body, a considerable part of acetaminophen is converted to NAPQI.[14]

Covalent binding of metabolites to DNA or other macromolecules can also occur in some other cases, for example, for bioactivated sulfamethoxazole, dapsone, procainamide, chloramphenicol, nimesulide, dantrolene, nilutamide, phenytoin, thalidomide, furosemide and norethindrone.[11] One more well-studied example is hepatotoxic tetrachloromethane, which is metabolized to give DNA-alkylating compounds (alkylators) posessing genotoxicity (for other examples, see the paper of Montellano [13]).

The toxicity of compounds that are bioactivated upon interaction with cytochromes P450 may be caused by reversible or irreversible deactivation of the cytochrome P450 itself. For example, drugs such as amphetamine, cimetidine and isoniazid, which are cytochrome substrates, can also inhibit cytochromes.[15]

It is noteworthy that metabolic enzyme inhibition- or induction-related effects are important factors for evaluation of drug–drug interactions. In the development of new drugs and the use of known drugs, considerable attention is paid to interaction of these drugs with other drugs in the human body. A change of metabolism of some drug substance under the action of another drug substance is one of the most significant consequences of drug–drug interactions. For example, inhibition of some metabolic enzyme may give rise to toxicity of a drug representing a substrate of this enzyme. Inhibition of the metabolism may increase the concentration of a competing substrate and thus enhance its action. If the increase in the concentration is substantial, this may bring about the toxicity. This is one of the most frequently occurring and important mechanisms of drug interactions in clinical practice. Since many drug substances are metabolized by the same cytochrome P450 isoforms, the competition between the substances is possible. For example, inhibitors of CYP1A2 (cytochrome P450 isoform) can increase the risk of toxicity of CYP1A2 substrates, clozapine and theophylline. The CYP2A9 inhibitors can cause toxicity of phenytoin, tolbutamide and oral anticoagulants such as warfarin. The CYP3A4 inhibitors (*e.g.*, phenytoin) increase the risk of toxicity for many drugs including carbamazepine, cyclosporine, lovastatin, HIV protease inhibitors, rifabutin, simvastatin and vinca alkaloids.[16]

The metabolism of drugs as xenobiotics that persistently and consistently enter a human body in biologically significant doses is an important factor affecting both their therapeutic effects and their toxicity.[17–20] Since the bioavailability, biological activity and toxicity of metabolites can differ substantially from those of the initial compounds, the integral evaluation of the behaviour of xenobiotics taking account of their metabolism in the human body is an essential task of medicinal chemistry. For such an evaluation, it is necessary to consider each stage of metabolism of xenobiotics, which implies, in the ideal case, the detailed construction of the metabolic network of enzymatic reactions with indication of qualitative and quantitative characteristics of all metabolites formed in each stage. The complexity of constructing such a network in an unambiguous and completed form is caused by the fact that this network cannot be unified for the whole diversity of genotypes and phenotypes encountered in the human population. In the ideal case, the construction of this network should make allowance for individual, including time-dependent, dynamic parameters affecting this network. For example, the following characteristics should be considered: the administered dose of the drug, other drugs taken simultaneously, patient's race, age and gender, physiological state of the body and diseases that can affect the expression of drug metabolism genes, food ration and so on.

In early stages of the development and investigation of the action of new drug substances, when the direct investigation of the drug metabolism in the human body is impossible for ethical reasons, while investigation on experimental animals or using test systems *in vitro* requires the synthesis and experimental testing of quite a number of chemical compounds and their metabolites, the only option is the application of computational methods.[21] The final goal is the *in silico* evaluation of the structure and properties of the metabolites formed upon biotransformation of pharmaceuticals. However, to achieve this goal, it is necessary to find out what particular enzymes interact with the test compound and to identify most probable sites of metabolism in the compound structure.

During the last decade, several reviews have been reported considering both experimental methods for

evaluation of xenobiotic metabolism [22, 23] and computational methods for predicting the metabolism.[2, 21 – 26] These reviews compare the existing metabolism databases and the prediction methods used to evaluate particular stages of the biotransformation of pharmaceutical substances and consider the options for combining experimental and computational approaches for predicting drug metabolism.[25]

This review is an attempt to consider more comprehensively the issues related to structural descriptors which are specially developed for predicting the interaction of xenobiotics with particular enzyme isoforms performing the biotransformation and for evaluating the sites of metabolism and the rules of selection of the most probable metabolites among the generated *in silico* candidates. Thus, this review covers the total integrated picture of evaluating metabolism of pharmaceuticals, including: (1) prediction of the interaction with particular biotransformation enzymes, (2) elucidation of the most probable sites of biotransformation, (3) identification of the most probable metabolites. The Conclusion gives a brief account of computer-aided estimation of the biological activity and toxicity of the parent compounds and the metabolites (these issues were considered in more detail by Zakharov and Lagunin [27]).

II. Computational methods of studying metabolism

The computational approaches for metabolism prediction can be subdivided into groups according to their application area: (1) prediction of the interaction of xenobiotics with metabolic enzymes, (2) prediction of the sites of metabolism (SOMs) in the chemical structure of a compound to be metabolized, (3) generation of the potential metabolite structures for the subsequent evaluation of their properties.

Figure 3 shows what results can be obtained by using various metabolism prediction methods as exemplified by the metabolism of codeine (an opioid analgesic, opiate receptor agonist). Drug-metabolizing enzymes convert codeine into more active metabolites: codeine 6-glucuronide (formed under the action of the UGT2B7 transferase), morphine (formed by CYP2B6) and norcodeine (formed by CYP3A4). The use of computer-aided methods provides answers to the questions of how (by which isoenzyme), where (sites of metabolism) and to what (metabolites) the compound of interest is metabolized.

Below we consider the computational approaches that are applied to solve various problems related to predicting metabolism of xenobiotics in the human body. They include approaches based on the ligand structure, such as the analysis of quantitative structure – property relationships (QSAR), construction and analysis of pharmacophore models, comparative analysis of the molecular fields and so on, and approaches based on the target protein structures, such as quantum chemistry, molecular docking, molecular dynamics and so on, or combinations of both types of approaches.

Using modern computational methods, it is possible to assess the binding energies of compounds to the enzymes that metabolize xenobiotics, identify the chemical structure moieties that are likely to be modified and determine the structures of metabolites formed upon the biotransformation. A popular approach to the prediction of metabolism is based on the application of expert systems.[23] As a rule, these methods use the knowledge bases containing the already known metabolic transformations of xenobiotics and the data on the structures of enzyme active sites involved in a particular biotransformation.[22]

An important result in the prediction of xenobiotic metabolism is the identification of the metabolites formed, which is necessary for the subsequent integral estimation of the properties of xenobiotics. Therefore, in this review, we consider more closely the computational systems that make it possible to generate metabolite structures. In most cases, these systems are based on analysis of training sets which include information on the substrate and metabolite structures and biotransformation reactions observed in *in vivo* and *in vitro* experiments and conditions for these reactions to proceed. The necessary information can be collected from available literature sources or retrieved from specialized free or commercially available databases (DBs). Below we give an account of databases containing relevant information.

III. Databases containing information on the metabolism of xenobiotics

The currently existing databases of biologically active compounds that can serve for predicting metabolism can be subdivided into three groups: (1) DBs containing information on the three-dimensional structures of protein – ligand complexes needed for implementation of methods for predicting the interaction between the substrate and metabolic enzymes and probable sites of metabolism; (2) DBs of biologically active compounds comprising the protein – ligand binding constants, which can be used for (quantitative) structure – property relationship analysis (prediction of the probability of interaction and estimation of the binding constants between substrates and metabolic enzymes); (3) DBs containing information on biotransformation reactions as well as structures of substrates and biotransformation products. The databases of the last-mentioned group are most important for metabolism prediction by methods based on the ligand structure.

Protein Data Bank (PDB) refers to the first group of DBs (those containing data on the three-dimentional structure of protein – ligand complexes).[28] This DB holds the coordinates of 3D structures for more than 35 000 proteins. However, most enzymes involved in xenobiotic metabolism refer, first of all, to the superfamily of cytochromes P450, membrane proteins for which experimental determination of the 3D structure is difficult. Therefore, for most eukaryotic cytochromes P450, three-dimensional structures have been determined after removal of the N-terminal transmembrane domain, which allowed the protein to be isolated and crystallized.[29] The PDB is accessed through a web interface with several software programs for 3D structure visualization and for amino acid sequence alignment. This allows the search for particular proteins by amino acid sequences, in particular, in the presence of signal-site amino acid substitutions.

The second group comprises the ChEMBL, PubChem and DrugBank databases containing quite a lot of information on xenobiotic metabolism.

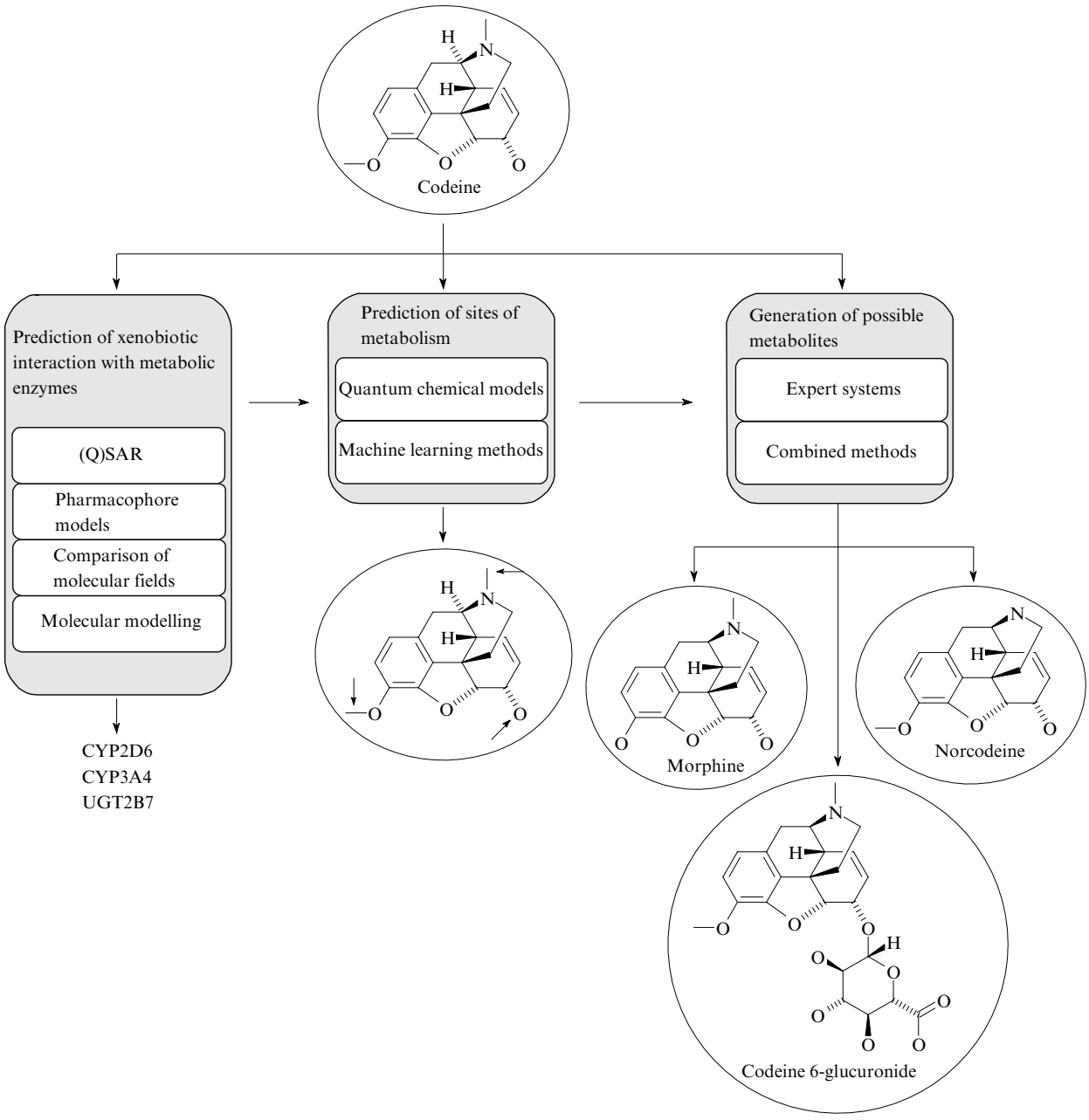


Figure 3. Key tasks and computer-aided methods used to evaluate the metabolism of xenobiotics exemplified by analysis of codeine metabolites (the hydrogen atoms at heteroatoms are not shown).

ChEMBL[30] is a DB created and maintained by the European Bioinformatics Institute (EBI, UK). The ChEMBL database (version 20) holds information about the biological activity of more than 1400 thousand compounds. Among more than 13.5 million records on the biological activity of these compounds present in ChEMBL, there are data on inhibition of cytochromes P450 involved in metabolism of xenobiotics (CYP17A1, CYP19A1, CYP1A2, CYP2A6, CYP2C19, CYP2C9, CYP2D6, CYP2E1, CYP3A4 and CYP5A1 isoforms). Altogether ChEMBL comprises data on more than 150 thousand compounds inhibiting cytochrome P450 family, including binding constants of enzymes to xenobiotics [half maximal inhibitory concentration (IC_{50}) for almost 25 000 compounds and inhibition constants (K_i) for 7500 compounds]. There is information on species or strains for which the DB contains inhibitors of the corresponding cytochrome isoform. The ChEMBL database is accessed through a web interface which provides the keyword search and structural formula search for small organic compounds. In addition, there is the possibility to download the database in the SDF format (SDF stands for structure-data file, Biovia, http://accelrys.com) or in the SQL format for subsequent import into widely used database management systems (DBMS) (Oracle, MySQL, Postgre SQL); hence, the automatic processing of the data can be performed locally. The developers of this DB implemented quite a number of

mechanisms of automatic access to the DB by means of various web services, including hypertext transfer protocol requests.[31] In particular, the ADME SARfari web service[32] provides the possiblity to retrieve ADME (Absorption, Distribution, Metabolism, Excretion) data, including data for compounds inhibiting particular targets (upon a query). For example, using ADME SARfari, it is possible to retrieve data about more than 800 assays described in various publications and used to test the inhibition of the CYP1A2 human cytochrome and to estimate the number of compounds investigated by means of particular assays.

PubChem[33, 34] is a free-access database maintained by the National Center for Biotechnology Information (NCBI, USA) and holding data on the interaction of chemical compounds with target proteins, including the interaction of xenobiotics with biotransformation enzymes (data on more than 26 000 compounds inhibiting various cytochrome P450 isoforms and on more than 400 metabolites of existing pharmaceutical drugs) in the human body. This DB is accessed through a web interface. Furthermore, a hypertext transfer protocol access[35] and data download in the XML, SDF and ASN.1 formats are envisaged.

DrugBank[36] is a free-access DB created and maintained by the Canadian Institutes of Health Research (CIHR) and the the Metabolomics Innovation Centre (TMIC). Initially, it was developed as a DB containing information on drugs, drug targets and mechanisms of action. The DrugBank 4.2 version holds ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) data for drugs, including 8000 records of metabolism. For about 2500 drug substances, there are also data on biotransformations, including the enzyme name, type of reactions and the structures of metabolites (more than 2200 structural formulae). The queries to the DrugBank database can be made through a web interface. A drug substance can be searched by the structural formula or the target protein sequence. The DrugBank database can be downloaded in the XML format to perform the subsequent automatic processing locally.

BindingDB[37] is a free-access DB with a web interface (https://www.bindingdb.org/). The database comprises 1 207 821 binding constants for protein – ligand complexes, which have been determined experimentally for 6235 target proteins and 529 618 ligands (as per May 1, 2016). The data were taken from scientific publications, the emphasis being placed on known or potential drug target proteins the structures of which are present in PDB. The BindingDB web site allows one to make various types of queries, in particular, to search by chemical structure and substructure, amino acid sequences, names of ligands and proteins, ranges of binding constant and the molecular mass. The answer to a query can be exported in the SDF format for subsequent analysis. The BindingDB records are linked to the structural data of PDB and literature sources from PubMed by means of PDB ID and PubMed ID identifiers, respectively.

The third group databases include specialized free-access DBs and commercial DBs containing xenobiotic metabolism data. Since exactly these DBs are needed to design prediction systems able to generate metabolite structures and, over a longer term, to provide integrated evaluation of xenobiotic metabolism, we will consider databases of this type in more detail.

Metabolite (**BIOVIA**) is a database developed by the MDL Information Systems and later purchased by the Accelrys (currently BIOVIA) company, which is now maintaining this DB (http://accelrys.com/products/databases/bioactivity/metabolite.html). This DB contains more than 116 thousands of records of biotransformation reactions of organic compounds (mainly drug and drug-like substances) collected from various sources, including the results of *in vitro* and *in vivo* experiments. For each reaction, the database has a schematic record including structural formulae of the substrate and the product. For some substrates and reaction products, the DB contains data about the biological activity, toxicity, CAS registry number and so on. For biotransformation reactions, there are references to publications, names of enzymes and isoenzymes catalyzing the reaction, data on experimental conditions, way of drug administration in the experiment, the type of biotransformation and so on. The Metabolite database is available through paid subscription.

The database included in the **MetaDrug** software has been created and is maintained by the Thomson Reuters company; the authors claim it to be a platform implementing a systemic pharmacological approach to the analysis of the biological activity of small organic compounds taking account of their ADME properties and toxicological characteristics (http://thomsonreuters.com/en/products-services/pharma-life-sciences/pharmaceutical-research/metadrug.html). This DB can also be used to analyze the properties of known and new drug substances at the development stage. The scope of analysis includes the search by structure of a related compound, calculated ADME properties and construction of QSAR models for particular enzymes. This analysis provides data on the targets being affected by biologically active compounds and evaluations of potentially toxic and biologically active metabolites and protein – protein interactions. Also, the possibility of predicting the transformation pathways of a compound in the body and the compound toxicity is envisaged. The database of MetaDrug contains information on the structures of more than 700 thousand small organic compounds, their biotransformations and the binding constants of xenobiotics to metabolic enzymes. The information has been taken from scientific publications and verified by experts. The information on reactions included in this DB is also accessible by means of another Thomson Reuters product — DB MetaBase (http://thomsonreuters.com/en/products-services/pharma-life-sciences/pharmaceutical-research/metabase.html). MetaBase is a knowledge base, a component part of the platform comprising MetaCore and MetaDrug. MetaBase is a database managed by Oracle and containing more than 200 tables. Thomson Reuters provides access to MetaBase through SQL queries. MetaBase contains information on more than 1 million interactions of proteins with other proteins, DNA, RNA and xenobiotics, more than 9000 metabolic reactions and about 700 000 compounds with indication of the targets and biological activity. These software products and DBs are available through paid subscription.

Human Metabolome Database (HMDB) (http://www.hmdb.ca/) is a free-access DB, which has been

created and maintained by the Canadian Institutes of Health Research and the Metabolomics Innovation Centre.[38] This database includes information on the metabolites of small organic compounds (xenobiotics and endogenous compounds) formed in the human body. The HMDB database, version 3.6, contains data on almost 42 thousand metabolites and more than 5.5 thousand amino acid sequences of proteins — metabolic enzymes and transport proteins. The DB is accessed through a web interface. Apart from the keyword search, it is possible to search by small organic compound structures and amino acid sequences of proteins. The HMDB data can be exported in the XML, SDF (metabolite structures) and FASTA (amino acid sequences) formats.

Fujitsu ADME Database has been created and is maintained by Fujitsu Kyushu Systems (https://life-science.kyushu.fujitsu.com/admedb/). The DB contains information about the interactions of more than 5200 chemical compounds with drug-metabolizing enzymes such as esterases, transferases, monooxidases including cytochrome P450 isoforms (CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP2E1, CYP3A4, CYP3A5, including their mutant forms and so on) and with transport proteins (P-glycoprotein and so on) with indication of the corresponding animal species or microbial strains. About 30% of biotransformation reactions included in the DB are accompanied by the structural formulae of metabolites. The information is collected from more than 5 thousand scientific publications. The Fujitsu ADME Database is available through paid subscription. The search through the DB can be performed using a standard form permitting the user to construct complex Boolean queries by choosing the category of drug, reflecting its pharmacological behaviour, chemical class, name of compound, type of interaction with metabolic enzymes (substrate, inhibitor, inducer, *etc.*), name of the protein, experimental testing results, type of metabolite (major or minor). The user can submit more complex queries and perform search by the structural formula of the drug. The developers do not provide an application programming interface (API) for operating the Fujitsu ADME database; therefore, automatic access to the DB is impossible.

Integrity has been created by the Prous Science company, which was later purchased by Thomson Reuters; currently, the latter maintains this DB and provides the Internet access to this DB through paid subscription (https://integrity.thomson-pharma.com/integrity/xmlxsl/). In particular, the database contains data on the pharmacokinetics of xenobiotics, including the maximum concentration of drug metabolites in blood plasma. A distinctive feature of the Integrity DB is that, apart from the drug metabolite concentration, it is also possible to retrieve data on the experimental method used to measure the concentration (taking account of the animal species, gender and age) and the type of biological sample (*e.g.*, whole blood, plasma and so on). Owing to these detailed data, one can compare the metabolite compositions and concentrations under different experimental conditions. The database is accessed through a web interface, which envisages the possibility to submit complex queries including various parameters such as pharmacokinetic data, dosage form, the model used in the experiment, drug and metabolite names and many other. However, automatic downloading of data from the DB Integrity is impossible, because according to the paid subscription terms, information can be retrieved only for a specified number of compounds by forming the queries manually.

KEGG[39] is an extensive resource combining genomic, chemical and functional information. As regards the metabolic studies, of considerable interest is the KEGG PATHWAY database, which provides a set of KEGG maps. A KEGG map is a diagram reflecting molecular interactions, metabolic networks, processes that occur in the cell and many other.

EAWAG-BBD[40] is a part of the EAWAG-BBD Pathway Prediction System.[40] This DB covers information on more than 1300 compounds, 900 enzymes, 1500 reactions and 500 microorganisms; also, it contains about 250 rules describing biotransformation reactions. The EAWAG-BBD database has a web interface (http://eawag-bbd.ethz.ch/), which allows the search for compounds by the trivial or systematic name, CAS registry number, structural formula or a substructure incorporated in the molecule.

Table 1 presents comparative characteristics of the considered DBs, which contain information on the xenobiotic metabolism.

All of the described DBs provide the user with advanced search options, including the substructure search and the similarity search, which forms a potent tool for generation of hypotheses (actually prediction) for metabolism of xenobiotics in the human body, in particular, for virtual, not yet synthesized molecules. However, only some of these DBs (Fujitsu ADME, Metabolite) contain biotransformation pathways and products and metabolites formed in these reactions, which restricts the possibility of direct utilization of most existing databases for construction of predictive systems.

Other resources. Apart from the DBs considered above, which hold information about the inhibitors and substrates of metabolic enzymes and about the biotransformation reactions, there are several more dedicated resources also relevant to drug metabolism.

SuperCYP[41] database holds records about 57 cytochromes P450. For more than 1000 single nucleotide polymorphisms and 1200 mutations in the proteins, the effects on expression and activity are indicated. The database also contains data on 1170 drugs and 3800 drug–drug interactions and drug interactions with biotransformation enzymes. The access to SuperCYP is through a web site (http://bioinformatics.charite.de/supercyp/) with a convenient user interface. This web site offers a wide range of functions for dealing with the SuperCYP database, including cytochrome 3D structure visualization and amino acid sequence alignment.

An interesting resource is the **Transformer** database,[42] providing information about the interactions of 3000 drugs and more than 350 food ingredients with more than 400 proteins such as the metabolic enzymes for the first and second phases of drug metabolism and with proteins of the so-called third phase, *i.e.*, transport proteins. Furthermore, the Transformer DB offers con-

Table 1. Comparative characteristics of the databases on xenobiotic metabolism.[a]

DB name	Total number of chemical structures and compounds interacting with metabolic enzymes	Total number of biotransformation product structures	Biotransformation reaction schemes
Protein Data Bank	79 complexes of cytochrome P450 isoforms with ligands	–	–
ChEMBL	> 150 000	–	–
PubChem	> 26 000	> 400	–
DrugBank	> 8000	> 2200	–
Metabolite [b]	> 25 000	> 50 000 [c]	> 110 000
Metadrug [b]	> 29 000	nd [d]	> 9000
HMDB	> 22 000	> 40 000	–
FujitsuADME Database [b]	> 5200	> 1500	–
Integrity [b]	nd,[d] present	nd,[d] present	–

[a] The dash means that no data are available, [b] DB is accessible through paid subscription, [c] total number of data without taking account of the number of unique compounds, [d] nd means cannot be determined because of the limited access.

venient tools for the search for drug – drug interactions described in the literature.

Pharmacokinetics Knowledge Base (PKKB)[43] is a free-access DB designed for accumulating information on the ADMET properties of compounds. The PKKB database contains records of 1685 drugs including the LogP and LogS coefficients,† intestinal absorption, Caco-2 permeability, human bioavailability, half-life excretion and other. The PKKB is available online at http://cadd.suda.edu.cn/admet.

The information available from the considered DBs can be used for different stages of metabolism prediction, ranging from the interaction with metabolic enzymes to generation of potential metabolite structures.

The quantity of data on the interaction of proteins with small organic ligands is constantly increasing in both scientific publications and databases of biologically active compounds. According to investigations,[44 – 46] data on the affinity of one and the same small organic compound contained in DBs of biologically active compounds are substantially scattered, which is expectable if these data have been obtained under different experimental conditions, but the discrepancies are still observed when experimental protocols are similar. It is quite likely that the reason is insufficiently detailed descriptions of experimental conditions in scientific publications, which precludes development of an unambiguous classification of experimental methods to be included in relevant DBs. Evidently, this brings about the problem of choosing well-substantiated data for numerical experiments; furthermore, ambiguity arises when data obtained in incomparable experiments are included in the test sets.

The key reasons for the errors in the data used to predict the ADMET properties are disclosed by Waldman *et al.*[47] In particular, the authors point out the errors associated with inappropriately indicated units of measurements in the DBs (*e.g.*, mM instead of μM) upon retrieving data from a scientific publication; faulty correlation of the synonyms of chemical names and trade names of the products; and errors arising due to tautomerism. The study cited describes also cases of ambiguous classification of compounds into cytochrome P450 substrates and non-substrates, in particular, caused by the influence of some enzymes [for example, nicotinamide adenine dinucleotide phosphate-H oxidase (NADPH oxidase)] on the structure of the initial compound, which is erroneously attributed to the effect of cytochrome on this compound.

The identification of data variability in the existing information resources and analysis of its sources[44 – 47] will be helpful for the development of approaches for eliminating this problem. This will enable designing more accurate models of quantitative structure – activity relationships with higher predictive power.

IV. Prediction of interaction of xenobiotics with metabolic enzymes

IV.1. Prediction methods based on ligand structure

Most of the methods of analysis of structure – activity relationships imply representing a small organic molecule as a set of descriptors, most often, as a vector with fixed positions of each type of descriptors.

IV.1.a. Descriptors of the structure of organic compounds

In most cases, organic compounds are described by one-, two- and three-dimensional (1D, 2D, 3D) descriptors.[48 – 50] One-dimensional descriptors are calculated from the molecular formula and can describe parameters like molecular mass. Two-dimensional descriptors are constructed from structural formula and can reflect the number of particular groups present in the molecule, the numbers of hydrogen bond donors and acceptors and so on. Three-dimensional descriptors are constructed from the spatial structure of a small organic compound and can describe parameters such as the charge distribution (dipole and quadrupole moments). So-called four-dimensional and higher-dimensional descriptors can take into account, for example, the relative positions of the protein and the ligand in space or molecular dynamic parameters over a specified period of time.

† LogP and LogS are descriptors corresponding to the $\log P$ and $\log S$, where P is the n-octanol – water partition coefficient characterizing the lipophilicity; S is the water solubility, mol litre^{-1}.

Note that the knowledge of the structural formula is necessary and sufficient for calculating the descriptors, as any descriptor ranging from the molecular mass to quantum chemical values is finally calculated on the basis of the structural formula.[51] Prediction of a property of an organic compound is usually based on a set of descriptors rather than a single descriptor. Therefore, a key task of the prediction [52] is to select the minimal set of descriptors that is sufficient for characterizing the property to be predicted.

While predicting the metabolism, one should take into account not only characteristics of one molecule (*e.g.*, xenobiotic) but also the possible changes in the xenobiotic structure upon the biotransformation. We developed structure descriptors for metabolism prediction, which can be used, in particular, to describe the site of metabolism and biotransformation reactions. Below we consider these descriptors in more detail.

VI.1.b. The neighbourhood descriptors MNA/LMNA

The Multilevel Neighbourhoods of Atoms (MNA) descriptors [53] are based on a structure representation that explicitly includes all hydrogen atoms according to the atom valences and charges but does not specify the types of bonds. Such a structural formula is fully unambiguous and does not depend, for example, on alternative ways of depicting aromatic systems.

The MNA descriptors for each atom of the molecule are constructed iteratively on the basis of this structure representation in the following way:

the zero-level MNA descriptor is the symbol of the atom A;

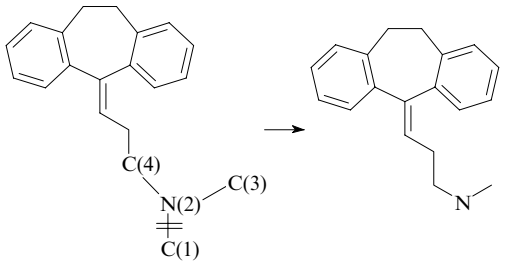
the MNA descriptor of each successive level is the identification code of the neighbourhood $A(D_1D_2 \ldots D_i \ldots)$, where D_i is the MNA descriptor of the previous level for the ith neighbour of the given atom designated by A. The neighbour descriptors $D_1D_2 \ldots D_i \ldots$ are written in some unambiguous order, for example, lexicographical one.

This iteration procedure can be continued up to any level. It is important to emphasize that the atom designation is not necessarily the chemical symbol according to the Periodic Table but can also include some additional information, for example, whether the atom is incorporated in a chain or in some ring or whether this atom is a site of biotransformation. This approach has been used in the development of LMNA descriptors (LMNA stands for Labelled Multilevel Neighbourhoods of Atoms),[54] which take account of what particular atom is considered as the hypothetical site of metabolism (SOM). An examples of LMNA descriptors for characterizing the SOM for N-dealkylation of amitriptyline (Scheme 1) is presented in Table 2 for four atoms that form the biotransformation site.

Table 2. Example of LMNA descriptors for the biotransformation site of N-dealkylation of amitriptyline.

Level	Descriptors
Zero	LMNA0(1) = -*C
	LMNA0(2) = -N
	LMNA0(3) = -C
	LMNA0(4) = -C
First	LMNA1(1) = -*C(-N-H)
	LMNA1(2) = -N(-*C-C)
	LMNA1(3) = -C(-N-H)
	LMNA1(4) = -C(-N-C-H)
Second	LMNA2(1) = -*C(-N-H)
	LMNA2(2) = -N(-*C(-N-H)-C(-N-H)-C(-N-C-H))
	LMNA2(3) = -C(-N(-C*-C)-H(-C))
	LMNA2(4) = -C(-N(-*C-C)-C(-C-H)-H(-C))

Scheme 1



The zero-level LMNA descriptor for the C(1) carbon atom (see Scheme 1) is additionally labelled by '*' as a hypothetical SOM [C(1) is the actual SOM for N-dealkylation of amitriptyline]. For other atoms, the zero-level LMNA descriptors do not have this label. Thus, the LMNA descriptors describe the site of metabolism and can be used to implement the computer-aided prediction algorithms.

As regards presenting chemical information to be applied to metabolism-related problems, the problem of descriptor-based representation of chemical reactions comes to the forefront. Therefore, descriptors of chemical structures, in which data on chemical reactions are stored in the compact form and the required operations can be effectively performed, are most important for metabolism prediction. Examples of descriptors that represent chemical reactions are RMNA descriptors (RMNA stands for the Reaction Multilevel Neighbourhoods of Atoms) [55] created on the basis of MNA descriptors.

IV.1.c. RMNA descriptors of biotransformation reactions

The principal difference between the RMNA [55] and MNA descriptors is the additional labelling of atoms and simultaneous consideration of two structures, namely, the substrate and product structures:

$$RNMA_1(A) = ([-]A[T]A_1[B]A_2[B] \ldots A_i[B] \ldots)$$

where A is the symbol of an atom; [-] is the label that implies that the atom is not incorporated in a ring; [T] is the label for an atom that changes during the biotransformation (acquires the value '<' for the incoming atom and '>' for the outgoing atom, respectively); $A_1, \ldots, A_i \ldots$ are the types of the closest neighbours of the atoms listed in the lexicographical order; [B] is the label for the bond that changes during the biotransformation (acquires the value '/' if the bond between the atom A and some neighbouring atom is cleaved, '\' if the bond is formed, ']' if the type of bond in the substrate changes and '[' if the type of bond in the product changes).

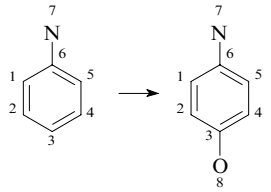
Table 3. Example of RMNA descriptors for the 'aromatic hydroxylation' reaction of aniline molecule.[55]

Atom	$RMNA_0$	$RMNA_1$	$RMNA_2$
1	C	C-HCC	(C-HCC)(C-HCC)(C-NCC)(-HC)
2	C	C-HCC	(C-HCC)(C-HCC)(-HC)(C-O < CC)
3	C	C-O < CC	(C-O < CC\)(C-HCC)(C-HCC)(-O < -H < C\\)
4	C	C-HCC	(C-HCC)(C-HCC)(-HC)(C-O < CC\)
5	C	C-HCC	(C-HCC)(C-HCC)(C-NCC)(-HC)
6	C	C-NCC	(C-NCC)(C-HCC)(C-HCC)(-N-H-HC)
7	N	-N-H-HC	(-N-H-HC)(C-NCC)(-H-N)(-H-N)
8	O	-O < -H < C	(-O < -H < C\\)(C-O < CC\)(-H < -O < \)

The RMNA descriptors of successive levels are concatenations of the previous-level descriptors of the given atom and the closest neighbours.

An example of generation of different-level descriptors (up to the second level inclusively) for biotransformation (aromatic hydroxylation) of the aniline molecule is shown in Scheme 2 (the hydrogen atoms are omitted) and in Table 3.

Scheme 2



Other molecular structure[56, 57] and quantum chemical[58, 59] descriptors are also used to predict the sites of biotransformation. Biotransformations can also be encoded as a system of special rules that are considered in more detail below.

IV.1.d. Line notation rules for chemical structures

A compact notation for transformations that is applicable for predicting particular biotransformation reactions and generation of probable metabolite structures makes use of the SMIRKS system,[60] which is, in turn, an extended version of SMARTS rules (http://www.daylight.com/dayhtml/doc/theory/theory. smarts.html) based on representation of the chemical structure as a linear SMILES string.[61] Although strictly speaking, these rules are not the same as compound descriptors, they still make it possible to compare molecular structures and perform substructure search through databases of organic compounds. The SMIRKS rules are used to describe a set of biotransformation reactions; therefore, it is necessary to consider the most popular rules of line notation for chemical compounds.

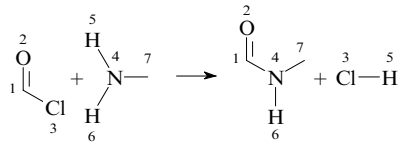
The **SMILES** (Simplified Molecular Input Line Entry System) rules[61] encode a structural formula of a molecule as a string of symbols. In most cases, this notation is much more compact than other chemical notations. For example, the SMILES string for ethane is CC. The construction of the SMILES string is conveniently described in terms of the graph theory. The SMILES string is formed by symbols of atoms present in the molecule listed in the order of depth-first traversal of the molecular graph. Thus, it is implied that the atoms adjoining each other in the SMILES string are chemically bonded. With this approach, the information on some edges belonging to rings can be lost. In order to avoid this, the symbols of atoms connected by such a 'lost' bond are provided with equal numeric suffixes. The branching points in the depth-first traversal tree are marked by brackets. The advantage of the SMILES system for the description of organic structures is the relative simplicity of implementation of the depth-first traversal algorithm, whereas the shortcomings are due to ambiguity of the choice of the spanning tree and the order of traversal, which results in the possibility of generation of numerous SMILES strings for one and the same structural formula. An attempt to overcome this ambiguity was made in the development of the generally accessible InChI format (international chemical identifier) similar to SMILES (http://www.iupac.org/inchi/).

The **SMARTS** (SMILES ARrbitrary Target Specification) system of rules (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html) is designed for the description of molecular substructures by means of rules that are straightforward extensions of SMILES. The use of SMARTS facilitates the search for specified substructures in molecules, which is especially desirable for predicting the structures of metabolites. The SMARTS language is almost the same as that of SMILES. However, SMARTS provides special labels for atoms and bonds, which allow one to describe a group of structures rather than a single structure. For example, the SMARTS atomic symbol [C,N] means that this atom is either aliphatic carbon or aliphatic nitrogen, and the bond symbol '~' (tilde) matches any bond. Furthermore, SMARTS notation has logical operators, so that logical expressions can be used to designate atoms and bonds. For example, if some atom is labelled as !C, then this SMARTS string matches any structure in which this atom is not C.

SMIRKS is a system of rules for describing general reactions; a SMIRKS description can match a group of reactions rather than a single reaction.[60] These descriptions are useful in those cases where it is necessary to apply a set of transformations to a group of molecules, which is especially often required for the generation of metabolite structures. Any reaction can be represented as a list of changes undergone by edges and atoms of a molecule during the transformation. If the reaction rules are designated as 'edge elimination', 'decrease in the bond order by 1,' 'decrease in the atomic charge by 1,' particular atom and bond labels play no part. This simple idea forms the basis of SMIRKS. The description of a generalized reaction in this format consists of SMARTS encodings of the substrate and the product,

with mapping being specified between the substrate and product atoms. Thus, appropriate reaction can be applied to any molecule matching the description of the substrate.[60] An example of a reaction written in the SMIRKS notation is presented in Scheme 3.

Scheme 3



SMIRKS reaction

[C:1](=[O:2])[Cl:3].[H:5][N:4]([H:6])[C:7] > > [C:1](=[O:2])[N:4](H:6)[C:7].[Cl:3][H:5]

Chemical structure descriptors can be used for both predicting the interaction of xenobiotics with metabolic enzymes and predicting the sites of metabolism and metabolite structures. The selection of the set of descriptors that is sufficient for a particular method of prediction of a property of organic compounds (*e.g.*, metabolism) is determined by the researcher and influences substantially the prediction outcome.[48, 52] The application of descriptors for prediction of metabolism is discussed in more detail below in the consideration of particular prediction algorithms.

IV.2. Models for (quantitative) structure – property relationship analysis

The Quantitative Structure – Activity/Property Relationships (QSAR/QSPR) are used to find correlations between the structure of molecules and the properties they exhibit. QSAR models have a wide range of application and are often used, for example, in the development of antibacterial[62] and antitumour[63] agents. The constructed models of organic molecules are used to predict their properties, for example, interaction with cytochromes P450 and sites of metabolism (biotransformation) of small organic compounds.[64] In the construction of QSAR models, the compound property is considered as the target variable and the descriptors of compounds serve as independent variables. The set of compounds with known experimental properties, composed using information taken from databases, is subdivided into the training set, which serves for constructing the model, and the test set applied to evaluate the accuracy of the model. The QSAR models are constructed using machine learning methods, in particular, Multiple Linear Regression (MLR), Projection to Latent Structures (PLS), Artificial Neural Networks (ANN), Support Vector Machine (SVM) and Bayes assessment-based approach. The QSAR models can be used to predict substrate binding in the active site[65] and inhibition[66] and induction[67] of cytochromes P450.[68] The substrates of different cytochromes are classified using SVM, *k* nearest neighbours algorithm, decision tree and so on.[69]

One of the first attempts of constructing QSAR models for metabolism prediction was made by White and McCarthy[70] in 1986 for ten *para*-substituted toluenes. The catalytic rate constant for hydroxylation (k_{cat}), the apparent dissociation constant of the enzyme – substrate complex (K_d) and the Gibbs energy of interaction (ΔG_{int}) were used as the target variables characterizing the interaction of these compounds with purified cytochrome P450 (LM2 isoform currently designated as CYP2B4) from the microsomal fraction of rabbit liver. They revealed linear correlations between these constants and physicochemical characteristics of substrates. Linear dependences were found between the rate constant for hydroxylation and the Hansch hydrophobic constant (π) and the Hammett electronic constant (σ); between the apparent dissociation constant and the molecular volume and the constant π; and between the interaction energy and the Hansch hydrophobic constant and the reciprocal of the dielectric constant. White and McCarthy arrived at the conclusion that the empirical dependences they found would be useful for both prediction of the properties of new substrates and the understanding of the chemical structure and mechanistic details of the action of the enzyme active site. Later, the QSAR models for the rate constants for hydroxylation and the binding energies constructed for this set of *para*-substituted toluenes were improved by Lewis *et al.*,[71] who used parameters of the substrates calculated by quantum chemistry methods.

Lewis[72] analyzed a small set of substrates of different cytochrome P450 isoforms and data on the substrate binding to the enzymes using multiple stepwise regression and physicochemical descriptors. This analysis demonstrated that the affinity of binding of substrates in the active site depends on their lipophilicity, acidity and the number of hydrogen bond donors. However, the authors of that study indicate only that the accuracy of the models was assessed using the set of 14 compounds and do not specify whether it was the test or training set, and the accuracy was calculated by leave-one-out cross-validation.

Mishra *et al.*[73] used the SVM method to find what cytochrome P450 isoform (CYP3A4, CYP2D6, CYP1A2, CYP2C9 or CYP2C19) interacts with xenobiotics being metabolized. The models were constructed and validated on a set of 266 drugs authorized for use in medicine. Using a genetic algorithm, the best molecular descriptors (41 out of 2392) were selected and then the models were constructed by the SVM method. The *k*-fold cross-validation technique estimated the prediction accuracy as ~86% and validation by an independent set of 146 drugs gave a value of ~71%. The resulting models were implemented as a web server for predicting the cytochrome P450 isoform that would metabolize particular organic compounds (http://crdd.osdd.net/raghava/metapred).

One approach used for determination whether or not a compound is a substrate of the particular cytochrome P450 isoform is based on artificial neural networks and is widely used to solve poorly defined problems.[67] Lewis[68] used a set comprising 485 substrates of cytochrome P450 and 523 biotransformation products. Seven physicochemical descriptors — $\log D_{7.4}$ ($\log P$ at pH 7.4 with ionization corrections), HOMO (Highest Occupied Molecular Orbital) energy, Jurs-TPSA (Total Polar Surface Area calculated by the Jurs method),[74] Jurs-PPSA (Partial Positive Surface Area, *i.e.*, the surface area of the positively charged molecular parts), molecular mass, the number of hydrogen bond acceptors, the number of hydrogen bond donors — were selected to be used as the

input parameters for artificial neural networks. Two networks were constructed.

The first one was a Kohonen self-organizing network. Two-dimensional coordinates of the Kohonen map, which reflected discrimination between substrates and products, were the output parameters. The network was built and trained using a program included in the ChemoSoft system.[69] The model correctly classified 76.7% of substrates and 62.7% of products.

The second one was the neural network with error back propagation algorithm using the supervised learning principle. The dataset (1008 compounds) was randomly divided into the training (60%), validation (20%) and test (20%) sets. The validation set was used to prevent system overtraining (overfitting). As the output parameter, the researchers used the difference between the scores characterizing the substrates and the products. The neural network was constructed and trained using the NeuroSolution 4.0 software.[75] The discrimination accuracy for this neural network was, on average (for 3 experiments), 60% for substrates and 65% for products.

The possibility of classification of a compound as a substrate, inhibitor or an inducer of various cytochromes P450 is implemented on the PASSOnline web resource (http://www.way2drug.com/PASSonline/). The average leave-one-out prediction accuracy for these classes of compounds on the training set was >90%. The PASS program[76] is the web resource core providing high-accuracy prediction of the biological activity spectra of small organic compounds on the basis of their structural formulae. The molecular structures of organic compounds are written by MNA descriptors. The prediction result is presented as a ranked list of names of the corresponding classes and the probabilities that a compound belongs (*Pa*) or does not belong (*Pi*) to a particular class.

Using the PASSOnline technology with a special training set, a web service was created (SMP: Prediction of substrate/metabolite specificity, http://www.way2drug.com/SMP) for the prediction of interaction with 18 key isoenzymes of the cytochrome P450 (CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4) and UDP glucuronosyltransferase families (UGT1A10, UGT1A1, UGT2B7, UGT1A7, UGT2B15, UGT1A8, UGT1A4, UGT2B17, UGT2B10, UGT1A3, UGT1A9, UGT1A6, UGT2B4).

Apart from the studies and web services considered above, there are also commercial software products for the prediction of interaction with drug-metabolizing enzymes: Discovery Studio (BIOVIA), ACD/Percepta (ACD/Labs), ADME QSAR module for StarDrop (Optibrium), PASS (interaction with drug-metabolizing enzymes), ADMET Predictor (Simulation Plus Inc.) and so on. Comparison of these software products is given in a review by Peach *et al.*[2]

IV.3. Advantages and shortcomings of QSAR methods

As noted above, the determination of the spatial structure of cytochromes P450 by X-ray crystallography is difficult as they are membrane proteins. However, quite a few experimental data on the substrates of cytochromes P450 have been accumulated to date; hence, methods based on the ligand structure alone can be used for metabolism prediction. The quantitative structure–property relationships have a number of advantages over pharmacophore models, because they do not require three-dimensional alignment of ligand structures, which can be ambiguous.[77,78]

Although QSAR models are able to predict the interaction of xenobiotics with various cytochrome P450 isoforms, the use of this approach is insufficient for predicting the metabolic pathways and constructing the metabolic networks. The key limitation for the construction of QSAR models with the up-to-date level of accumulated data on the structure–activity relationships is the substantial scatter of data contained in the databases of biologically active compounds[46] (this problem was described in greater detail in Section III).

IV.4. Construction of pharmacophore models

Pharmacophore modelling is also used for predicting metabolism of xenobiotics.[79–81] By pharmacophore[82] is meant the ensemble of structural and physicochemical features of the molecule (hydrogen bond donors and acceptors, positively and negatively charged atoms, hydrophobic areas, aromatic rings and so on — so-called pharmacophore sites) that are significant for a particular sort of biological activity (in the context of this review, metabolism). As a rule, these features are derived from comparison of the structures of known substrates of particular enzymes using structural formulae or, most often, by three-dimensional alignment of structures of compounds with known biological activity and identification of pharmacophore sites in this way. The application of pharmacophore models for predicting xenobiotic metabolism is considered in reviews.[83,84]

A pharmacophore model allows one to discriminate between the structures that belong to the substrate class and that belong to the opposite ('non-substrate') class and to propose hypotheses about the mechanism of substrate to enzyme binding. This can be used to assign compounds to the class of substrates of particular enzymes.

It is noteworthy that pharmacophore models based on the ligand structure do not always adequately reflect the reality. For example, it was found using the pharmacophore model constructed for the CYP2D6 isoform that substrates for this isoform are molecules, which contain a basic nitrogen atom and/or positive charge, and that CYP2D6 oxidizes atoms located at a 5–7 Å distance from a nitrogen atom.[85,86] However, some substrate structures are beyond the scope of this model and have much longer distances from nitrogen to the oxidation site; an example is tamoxifen (>10 Å distance). This is attributable to the fact that models of pharmacophores based on cytochrome P450 substrates[79,81,87] have a limited applicability domain determined by the training set composition and reflect the static picture of the substrate and catalyzed reaction recognition without taking account of the shape of the cytochrome active site and its change during the reaction. Pharmacophore models also fail to predict the strength of ligand binding in the enzyme active site.

IV.5. Comparison of molecular fields

Yet another method based on the ligand structure is Comparative Molecular Field Analysis (CoMFA). It also relies on superimposition of the 3D molecular structures

of known ligands of a particular target protein, possibly, going through their different conformations one by one. After superposition of the training set molecules, the molecular fields of electrostatic, steric and other interactions are calculated. This is followed, most often, by calculation of the values of the fields at three-dimensional lattice points or the values averaged over the lattice cell volumes. The set of variables obtained in this way is used to construct a model by some machine learning method, for example, by the PLS method. The superposition of training set molecules is performed under the assumption that all molecules bind to the enzyme active site in the same way; however, in reality, different molecules have different orientations; it is even possible that several orientations of a molecule relative to the active site are inherent to the same compound.[88]

An example of application of the CoMFA algorithm is the construction of the active site model for the CYP1A2 enzyme.[89] This model was constructed using molecular docking and molecular dynamics for determination of the most probable orientation of substrates in the enzyme active site corresponding to experimental data. Then the obtained superimpositions of the productive conformations were projected onto the residues incorporated in the active site, which thus gave its simplified 'trace'. The model was tested in computational and real experiments; it was shown to possess fairly high predictive performance.

One more method of ligand-based analysis is the Comparative Molecular Similarity Indices Analysis (CoMSIA). In this case, training set molecules are also superimposed on a grid but apart from the steric and electrostatic fields, the hydrogen bond donor and acceptor fields as well as hydrophobicity fields are calculated. For each point of the grid, these field characteristics are converted to similarity indices and then correlation analysis is performed. As an example of xenobiotic metabolism prediction by the CoMSIA method, consider the paper of Locuson *et al.*,[90] who analyzed the specificity determinants for CYP2C9 using a library of highly active benzbromarone inhibitors. Three-dimensional contour maps and quantitative interaction models were constructed; this provided an explanation for the fact that anionic substrates with bulky hydrophobic substituents have higher affinity for CYP2C9.

The Molecular Field Topology Analysis (MFTA) is a topological analogue of CoMFA.[91] This method is based on automatic construction of the molecular supergraph. Each structure included in the set can be represented as a subgraph of this supergraph and the structure of a compound being predicted can be superimposed on the supergraph. Superimpositions of the structures of particular compounds on the supergraph form the descriptor vectors, which can serve for modelling by machine learning procedures. The authors demonstrated that the MFTA combined with latent variable regression gives robust models even if the number of descriptors is comparable with or exceeds the number of objects.

In most cases, methods based on comparison of molecular fields produce a generalized view of the enzyme active site, and these data can serve for the search for inhibitors and substrates characterized by considerable chemical novelty.

Despite the substantial progress of CoMFA and CoMSIA methods in recent years, they still have limitations. It is necessary to calculate correct and mutually correlated conformations of many molecules, this calculation being complicated by difficulty of correct account of the effect of reaction centre on substrate conformations; superimposition of many molecules is not always unambiguous and does not take account of the substrate arrangement in the reaction centre either. Often, it is necessary to repeatedly vary the model parameters to achieve an acceptable accuracy, which brings about a considerable ambiguity of the results, and not all of the obtained parameters are cited in publications. This leads to non-reproducible results of one particular CoMFA and CoMSIA study and can markedly hamper the practical application of these methods for predicting the properties of organic compounds.[92]

V. Methods based on analysis of enzyme structures

The methods considered below require the data on 3D protein structure. Usually, proteins consist of several thousand atoms; therefore, prediction methods based on the enzyme structure are associated with considerable computational time, which restricts the application of these methods for screening. Nevertheless, despite these shortcomings, these methods are indispensable when researchers need to understand the nature of ligand interaction with the enzyme active site.

V.1. Determination of three-dimensional protein structure

The 3D protein structures are determined, most often, by X-ray diffraction crystallography and nuclear magnetic resonance spectroscopy. When no experimental data on the spatial structure are available, researchers use models constructed by homology [*e.g.*, for CYP2C19 (see Ref. 93)] on the basis of amino acid sequences of similar proteins with known 3D structures.[64] These models are actively used to assess the interaction of metabolic enzymes with small organic ligands. For example, de Groot *et al.*[87] docked 51 substrates into the model cytochrome CYP2D6 active site constructed by homology. This study provided interpretation for 72 biotransformation reactions catalyzed by this cytochrome. The homological modelling of 3D structures can be used for the subsequent construction of 3D QSAR models. Belkina *et al.*[94] presented a model of human CYP1A2 cytochrome P450 constructed by homology with four bacterial cytochromes. The obtained 24 complexes of the CYP1A2 isoform of cytochrome P450 with its known ligands served to construct a model with high predictive performance, which made it possible to predict 680 new potential CYP1A2 ligands.

Currently, 3D structures are known for some isoforms of cytochrome P450 involved in the first phase of xenobiotic metabolism: CYP3A4,[95] CYP2C9,[96] CYP2D6 [97] and CYP1A2.[98] These structures are included in the PDB database considered above. Studies of the obtained 3D structures of cytochromes P450 are continued [99] in order to specify the active site shape and the protein – ligand interaction mechanism. In the experimental determination of 3D structures of cytochromes, these proteins are subjected to various modifications to increase their solubility, which is necessary to obtain

crystals. These modifications introduce some uncertainties into the resulting 3D structure of the protein; hence, the structures of the same cytochrome P450 isoform obtained under different conditions can differ.[100]

The 3D structures of the isoforms of UDP glucuronyl transferase (UGT enzymes), which accomplish most of conjugation reactions of the metabolism second stage, are less studied. The point is that while determining the 3D structures of UGT, researchers face even more challenging problems than while dealing with cytochromes P450. The problems are associated with the possible overexpression of this enzyme and the isolation and crystallization difficulties.[101,102] The studies of these enzymes are still limited to the elucidation of the 3D structures of particular domains.[103]

V.2. Molecular modelling methods

The basic methods of molecular modelling[104] used to study the substrate–enzyme interaction are molecular mechanics, molecular dynamics and docking.

Molecular mechanics methods use classical mechanics and represent molecules as point masses and charges corresponding to atoms and elastic forces that act between them and correspond to deviations from the 'natural' bond lengths and bond angles, van der Waals forces and electrostatic interactions. This approach is rough and the calculation results depend on approximations and parameters used to describe the potential functions called force field, which are selected depending on the type of problem being solved, for example, the solvent molecules are included either directly or indirectly.[105]

Molecular dynamics is based on the calculation of the trajectories of motion of the macromolecule in the phase space of the coordinates and impulses of its atoms where the molecule is considered as a system of interacting classical particles.[106] Molecule dynamics serves to follow the time variation of a system of the interacting ligand and enzyme molecules. The method helps to consider various orientations and conformations of the ligand in the enzyme active site and the role of solvent molecules if they are present.[107] Using special calculation methods, it is possible to estimate the thermodynamic characteristics of molecular complexes without determination of the formally accurate trajectories of motion of atoms forming the complexes.[104]

Molecular docking consists in the computer-aided search for the optimal geometry of the possible substrate–enzyme molecular complex and assessment of the formation energy of the complex, which is calculated as scoring functions.[108]

Molecular docking can be used at the stage of evaluation of the ligand interaction with a protein, *viz.*, a metabolic enzyme. The molecular docking analysis allows predicting the binding mechanism between the protein and the ligand. Molecular docking makes it possible to detect new compounds, which can be illustrated, for example, by the discovery of new antitumour agents.[109]

A number of papers published in the recent years report on assessment of the spatial arrangement of ligands in the active site of some cytochrome P450 isoforms by molecular docking and calculation of the binding energies for the protein–ligand complexes formed upon docking. Vasanthanathan *et al.*[110] used molecular docking to evaluate the interaction of 7469 compounds taken from the PubChem database with CYP1A2. As a result, active conformations of small organic ligands in the cytochrome active site (so-called binding poses) were determined.[110] However, the balanced accuracy of compound classification into cytochrome 'ligands' and 'non-ligands' was 65.5%, which implies a relatively low prediction accuracy. Subsequently, a combined approach was used: docking of eight structures of known pharmaceuticals into the cytochrome CYP1A2 active site was followed by application of molecular dynamics and calculation of the binding affinity, which was defined as a linear combination of energies of several types of intermolecular interactions.[111] By validation on an independent test set, it was shown that the constructed model has a reasonable predictive performance (the root-mean-square error for the binding affinity was 2.1 kJ mol^{-1} for the test set). As a drawback of this study, note that the training set comprised only eight compounds and the test set included only five compounds. Similarly, ligand affinity was estimated for a number of other cytochromes: CYP2C9,[112] CYP2D6[113] and CYP3A4.[114] In these studies, the numbers of involved compounds did not exceed 15 either, which precluded extrapolation of the results to the whole diversity of structures of possible ligands. As a benefit of these methods, note that, while using them, researchers found that there are simultaneously several energetically favourable (active) ligand conformations in the cytochrome active site. This can indirectly attest to a conformational flexibility of the active site of CYP3A4,[114] CYP2C9[112] and CYP2D6.[113]

Molecular docking techniques as well as combinations of molecular docking and molecular dynamics are used, in particular, for predicting the SOMs (see below), because upon application of molecular docking, it is possible to choose the active conformations of the ligand in the complex with the protein, which is significant for assessment of spatial accessibility of the potential ligand site of metabolism. The key criterion that is used to consider some atom as a potential site of metabolism is the distance between this atom and the iron atom in the enzyme active site in the active conformation of the ligand; on average, this distances is taken to be $\leqslant 6$ Å.[115,116] The protein–ligand interaction models with implemented attempts to predict the SOMs have now been constructed for CYP3A4,[117] CYP1A2[98] and CYP2D6.[115] A combined molecular docking/molecular dynamics approach to determination of the sites of metabolism induced by CYP2D6 has been reported by Hritz *et al.*[118]

The key limitation for the use of molecular docking algorithms in the prediction of SOMs is caused by the fact that the standard docking algorithms neglect the flexibility of some fragments of the enzyme active site and the role of water molecules in the ligand interaction with the enzyme. When the molecular docking procedure tries to take account of water molecules in an explicit form, the results — selected energetically favourable ligand conformations, predicted interaction of water molecules with the substrate or enzyme atoms or the

predicted presence/absence of a water molecule in the active site — can differ,[110] and this causes difficulties in their interpretation. In a study devoted to prediction of the sites of CYP2D6-induced metabolism,[113] it was shown that taking account of water molecules in the substrate docking increases the accuracy of the estimate to a minor extent (57% and 52% with and without water molecules, respectively).

Thus, it can be concluded that modern molecular docking software cannot be applied for SOM recognition with high accuracy, even when special algorithms are used to take account of water molecules in the cytochrome P450 active site.

Some researchers make attempts to use corrections for the reactivity of particular functional groups of ligands. The reactivity of a functional group can be taken into account, for example, by correcting the scoring function after application of the standard molecular docking algorithm. In this case, reactivity corrections are made for functional groups located in the close neighbourhood of the iron atom or the oxygen atom of the enzyme active site.[119–121] One more way to take account of the reactivity is calculation of the activation energy for a particular functional group[121] (for the description of application of quantum chemical models, see Section VI.1 below).

An approach that would provide an exhaustive description of the substrate and enzyme interaction and predict biotransformations *in silico* should consider the ability of an enzyme to interact with a small organic compound, the enzyme active site structure, the position and conformation of the ligand in the active site and the possibility for the substrate to bind in the active site in different ways (binding may be mediated by water molecules). Furthermore, it is known that the substrate regioselectivity,[121] which affects the substrate–enzyme interaction, should also be considered. (By regioselectivity is meant the situation where one chemical bond cleavage or formation pathway predominates over the other pathways.) The substrate ability to interact with the enzyme and, hence, to participate in a particular biotransformation reaction can change depending on the conformation assumed by the substrate molecule. Finally, such an approach should include consideration of the affinity of the product released from the enzyme active site. None of the currently existing methods covers the whole set of the above requirements.[122]

The score functions used to calculate the binding energies in molecular docking are not sufficiently accurate, which considerably hampers the formation of versatile criteria for sorting out the results.

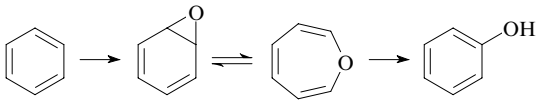
VI. Prediction of the sites of biotransformation

The majority of the developed algorithms for prediction of the sites of biotransformation use ligand-based computational methods. Initially, quantum chemical models were employed most often for this purpose. This, apparently, stems from the fact that the chemical bond breakage and formation are, in principle, quantum chemical phenomena and prediction of biotransformations using computational methods without applying quantum chemistry seemed improper.

VI.1. Quantum chemical models

Oxenoid model. The approach was proposed by Korzekwa *et al.*[123] for predicting the oxidation sites of substituted benzenes. Aromatic compounds are typical substrates of cytochrome P450. Many aromatic compounds, in particular, pharmaceuticals, can be considered as benzene derivatives with simple or complex substituents. Short-lived but exceptionally reactive products formed upon activation of benzene and benzene derivatives in the body are arene oxides (Scheme 4).[123]

Scheme 4



According to the model of Korzekwa and co-workers, the role of the enzyme is to cleave the O_2 molecule to give atomic or nearly atomic reactive oxygen ('oxene'), which binds to the substrate. The oxenoid mechanism of oxidation of various substrates was studied experimentally and using quantum chemical calculations.[124] In the case of benzene derivatives, the formation of an intermediate with a tetrahedral carbon atom is the key step triggering the metabolism process.

Presumably,[125] the rate of hydroxylation is determined by the stability of the 'tetrahedral' intermediate with respect to the parent compound. The formation energies of the parent compounds and intermediates with a tetrahedral carbon atom were calculated by the MNDO method.[126] In the calculation of intermediates, the geometry of the metabolized ring was taken to be the same as in the benzene intermediate and the substituent geometry was taken to be the same as in the parent substrate molecule. The substrate geometry was found by optimization from the condition of the total energy minimum. All possible sites for aromatic hydroxylation in the molecule were considered; if there were two or more benzene rings in the molecule, the energies were calculated for intermediates formed upon successive addition of oxygen to all unsubstituted positions of the benzene rings. According to the oxenoid model, the smaller is the difference between the formation enthalpies (ΔH) of the intermediate and the starting substrate, the more easily should be hydroxylation of the given position. The enthalpy difference between the intermediate and unsubstituted benzene as the starting molecule was taken to be zero. Positive ΔH values correspond to a less easy hydroxylation compared with unsubstituted benzene. If the ΔH value is positive and large, the reaction will not proceed.

Methoxy radical model. Instead of the reactive oxene involved in the oxenoid model,[123] the model of Jones *et al.*[127] uses the methoxy radical mimicking the oxidizing species of cytochrome P450. The structures of both the substrate and the metabolite are optimized using the semiempirical AM1 method.[128] The heats of formation of tetrahedral intermediates resulting from addition of the methoxy radical to carbon atoms of the aromatic ring were calculated for all possible sites of aromatic hydroxylation. The enthalpy formation difference (ΔH) between the tetrahedral intermediate and the substrate approximately describes the activation energy for oxida-

tion and, according to the model, the lower values are inherent in aromatic hydroxylation sites observed in the experiment.

As compared with the oxenoid model, the methoxy radical model is suitable for predicting the aromatic hydroxylation not only for substituted benzenes but also for heterocyclic and polycyclic aromatic compounds.[129]

The practical application of quantum chemical models of cytochrome P450-induced oxidation requires considerable computing time; furthermore, these models do not take account of the substrate orientation in the cytochrome active site. Hence, the applicability of these models is limited; however, in combination with molecular modelling methods and with introduction of statistical parameters, they can be successfully employed for prediction of the sites of metabolism.[59]

Currently, several approaches to SOM prediction based on the use of quantum chemical descriptors and on the construction of classification models by statistical methods have been developed.

Zheng *et al.*[58] used a set of 23 quantum chemical descriptors reflecting charge distribution in atoms, atom valences and energies of potential interaction between atoms and constructed the classifier for discrimination between SOM and non-SOM by the SVM method. Altogether, the SOMs for six biotransformation reactions were considered (aliphatic hydroxylation, aromatic hydroxylation, N-dealkylation, O-dealkylation, S-oxidation and N-oxidation). Training sets contained 74 to 300 compounds depending on the reaction. The test set included 24 compounds. The authors indicate that the balanced accuracy exceeded 0.9 for the training set and was more than 0.8 for the test set. Similarly, Hennemann *et al.*[130] applied quantum chemical and physicochemical descriptors and the multiple linear regression for model construction. However, in the cited examples, the authors did not evaluate the developed methods against several independent datasets, for which earlier prediction results are available.

In the SMARTCyp method,[131] the score function S by which the atoms are ranked according to the probability of being the CYP2C sites of metabolism has the following form:

$$S = E - 8A$$

where E is the reactivity descriptor, A is the accessibility descriptor. The descriptor E is an estimate of the energy required for a cytochrome to react with a particular atom in the molecule; SMARTCyp uses a database of structure fragments (the fragments are described according to the SMARTS rules), the activation energy for the reaction of the given fragment with the cytochrome being calculated beforehand by density functional theory (DFT).[132] For each atom, the search for fragments containing this atom is performed. If there are several such fragments, the one characterized by the lowest energy is chosen. The descriptor A is defined as the longest path distance to the given atom divided by the path distance in the whole molecule. This descriptor is similar to the SPAN descriptor used by Sheridan *et al.*[133]

The topological (structure) descriptors in combination with quantum chemical ones are applicable in the implementation of the RS-Predictor algorithm.[57] The SOM for the CYP3A4 isoform of cytochrome P450 was predicted by a modified SVM method. The authors used cross-validation with 10-fold splitting of the dataset into five partitions, one partition being the internal test set and one more being the set used to select parameters in the modelling. As a result of this procedure, each compound of the test set had 10 putative sites of metabolism, which were ranked by the algorithm developed by the authors. The average prediction accuracy varied from 56% to 87% depending on the set of descriptors. A comparison with some other SOM prediction methods was made. The basic limitation of this method is that the sites of metabolism were predicted only for CYP3A4, which is the most important but not the only one isoenzyme of the P450 family that catalyzes the metabolism of xenobiotics. In addition, considerable scatter of the accuracy of classification models depending on the set of descriptors precludes unambiguous definition of such set for application of these models to predict SOMs in other compounds, for example, considerably differing in the chemical structure from the compounds chosen for the test set.

Tyzack *et al.*[56] considered the application of molecular structure descriptors of organic compounds and various classifiers (naive Bayes classifier and original classifiers developed by the authors) to predict the possible sites of biotransformation catalyzed by cytochromes P450 (CYP2C9, CYP2D6 and CYP3A4 isoforms). The authors presented also the dependence of the training results on the number of descriptors selected for the model construction. The prediction methods were tested on an external test set and compared with other methods; however, as the parameter for comparison, the authors used the value of correct SOM predictions out of two most probable SOMs rather than the accuracy of SOM prediction (for the interaction with a particular cytochrome P450 isoform). The choice of this parameter for comparison is ambiguous, which is a drawback of the reported approach.

VI.2. Models based on LMNA descriptors

We developed[54] a method for predicting the sites of biotransformation based on LMNA descriptors (the definition of these descriptors is given above) and a PASS software algorithm.[76] Five major isoforms of cytochrome P450 (CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4) were chosen. An independent test was performed on a dataset of 57 drugs acting on the cardiovascular system. The average invariant accuracy of prediction was ~83%. The accuracy of prediction for some cytochrome P450 isoforms was found to be higher than that achieved by other SOM prediction methods, in particular, RS-Predictor.[57]

The developed method was implemented as a free-access SOMP web site (Site Of Metabolism Prediction, http://www.way2drug.com/SOMP/), which is capable of predicting the sites of metabolism for CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 and UDP glucuronyltransferase family.[134]

We will consider the results of metabolism prediction for codeine (**1**) using the PASSOnline, SOMP, SMARTCYP and RS-Predictor web services (Table 4).

It can be seen from Table 4 that the prediction results were generally correct. According to the PAS-

Structure 1

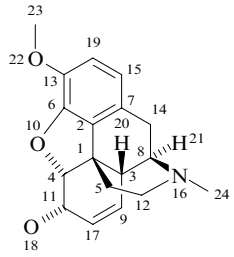


Table 4. Results of prediction of the codeine interaction with known metabolic enzymes and codeine sites of metabolism.

Web service	Prediction results		
	Pa	*Pi*	Type of activity
PASSOnline	0.984	0.002	CYP2D6 substrate*
	0.889	0.006	CYP3A4 substrate*
	0.699	0.002	UGT2B7 substrate*
	Atom	Rank	Score ($Pa-Pi$)
SOMP	CYP3A4		
	24	1	0.748*
	23	2	0.431
	CYP2D6		
	24	1	0.755
	23	2	0.752*
	UGT		
	18	1	0.417*
	1	2	−0.883
SMARTCyp	CYP2D6		
	23	1	0.755*
	24	2	0.752
RS-Predictor	CYP3A4		
	24	1*	
	23	2	
	CYP2D6		
	23	1*	
	16	2	

Note. The asterisks mark the correct predictions.

SOnline, codeine is likely to be a substrate of CYP2D6, CYP3A4 and UGT2B7. According to the SOMP, the known sites of metabolism for CYP3A4 and UGT were ranked first, and those for CYP2D6 were ranked second. The SMARTCYP and RS-Predictor also ranked first the known codeine sites of metabolism for those enzymes for which they were able to make predictions; however, SMARTCYP does not predict the sites of metabolism for CYP3A4 or UGT, while the RS-Predictor is inapplicable to UGT. This clearly demonstrates that higher prediction accuracy can be achieved if different software programs are used in combination.

VII. Generation of the structures of probable metabolites

VII.1. Expert systems

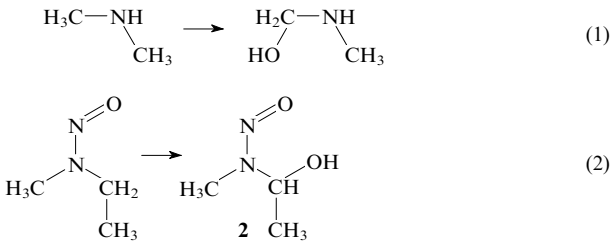
Most of pharmaceutical drugs are subjected to several biotransformation reactions in the body. For example, Fig. 4 shows the quetiapine molecule in which atoms that are involved in the major metabolic reactions in the human body are marked.[135–138]

Each biotransformation of the substrate molecule is accompanied by particular changes (transformations). The databases of transformations composed of experimental data form the basis of expert systems and are used to generate the structures of probable metabolites. Most often, the transformations are described in the form: IF 'structure fragment X is present', THEN 'replace it by fragment Y'. This condition can be extended to include several structure fragments, or can contain a list of fragments that hamper the reaction. A set of two fragments (X and Y) is called a transformation.

Examples of fragment description in the META program[139] are presented in Table 5. In the second column, the upper line is an example of 'structure fragment X', while the lower line is 'structure fragment Y'.

For example, the application of a rule from the biotransformation dictionary of the META program [Scheme 5, reaction (1)] to methylethyl-*N*-nitrosamine results in its hydroxylated derivative **2** [see Scheme 5, reaction (2)].

Scheme 5



Thus, one iteration of the generation of potential metabolites for a specified compound consists of application of rules from the biotransformation dictionary to the substrate, resulting in a set of predicted metabolites. If the prediction has to be continued, metabolites generated in the first iteration become the substrates for the second generation iteration. The key drawback of this approach is the exponential growth of the number of

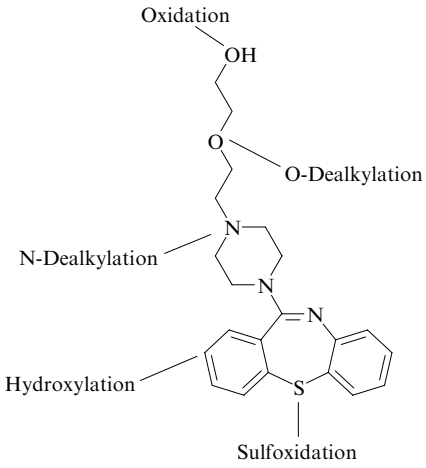


Figure 4. Quetiapine structure and key sites involved in biotransformation reactions.

Table 5. Examples of transformations (phase I) in the META program.[140]

Type of reaction	Example of META transformations	Enzyme example
Aromatic hydroxylation	CH=CH- C=CH- < OH >	P450
Aliphatic hydroxylation	CH3-CH2- CH3-CH- < 2-OH >	P450
O-Dealkylation	CH''-O-CH2- CH''-O-CH- < 3-OH >	P450
N-Oxidation	CH2-N-CH3- < 2-C > CH2-NO-CH3- < 2-C >	P450 and flavin monooxygenase

Note. The following special symbols are used: < > indicates that the substituent is attached to this site; '' is an sp^2-hybridized carbon atom.

predicted metabolites. Therefore, one of the tasks is prioritization of the results and choice of a termination criterion for construction of the metabolic tree (undirected acyclic graph with particular metabolites being the graph nodes). A possible termination criterion is, for example, the appearance of a metabolite with low LogP value (which is sufficiently water-soluble to be excreted from the body). One more limitation of expert systems is that they contain, as a rule, information about the biotransformations that take place in different biological organisms. However, the metabolic pathways in different organisms can be considerably different; therefore, for operation of an expert system, it is important to select the information related only to a particular group of organisms.[23]

The most significant expert systems developed for metabolism prediction are surveyed in Table 6.

The first attempts at the design of metabolite prediction systems resulted, in 1985, in the release of the system **MetabolExpert**.[141] Initially, its set of rules was formed by the metabolic pathways described in the book of Testa and Jenner.[142] The list of metabolic pathways was arranged in such a way that biotransformations associated with modifications of identical functional groups were located in the vicinity of one another. This data arrangement allowed, for the first time, predicting biotransformations for compounds that have not yet been studied. Subsequently, the set of rules was extended by addition of data from other literature sources. The transformation rules in MetabolExpert are similar to the logical expressions in the Prolog programming language. The operation of MetabolExpert involves the construction of the tree of predicted metabolites by successive application of the rules to the input molecule. The construction of the tree stops as soon as a second phase metabolite has been generated. The prediction results are represented as a dendrogram (tree). According to the prediction, the structural formula, the hydrophobicity and some other characteristics are presented for each metabolite. An important feature of MetabolExpert is the possibility to choose only those rules that refer to the type of metabolism that interests the researcher, *e,g*., animal or plant metabolism or degradation of organic compounds in soil. However, a considerable drawback of MetabolExpert is that the predicted metabolites are not ranked in any way, which hampers the isolation of most probable metabolites.

Apart from MetabolExpert, **META** was one of the first software programs designed for computer-aided prediction of metabolites.[139] The META system has a number of considerable advantages over MetabolExpert. First, META uses much more extended biotransformation dictionaries: there are several dictionaries, each presenting a particular metabolic model including the description of, for example, the animal species or organs. In particular, there are dictionaries that cover mammalian metabolism, aerobic and anaerobic degradation of organic compounds by bacteria and photodegradation. This separation allows the user to concentrate only on those types of metabolism he is interested in. While composing the dictionaries, the developers gave particular attention to mammalian metabolism. They set themselves the goal to include as many biotransformations of this type as possible, but using only the most reliable data from literature sources.

The biotransformations in META are prioritized. First, the priorities were established by expert chemists, and after that, a genetic algorithm was incorporated to increase the accuracy. Prioritization makes it possible to arrange the predicted metabolites in terms of the probability of their formation, which is important for the control of combinatorial explosion. During the META prediction, the structure of a molecule is split into fragments and then a biotransformation dictionary is applied to this structure to generate metabolites. Each generated metabolite is checked for stability. META has a dictionary of 'spontaneous reactions' reflecting the intermediate unstable steps of metabolism. Each unstable metabolite is subjected to transformations from this dictionary. This process is continued until all metabolites become stable. The operation of META generates a tree of predicted metabolites, which is terminated when compounds that are sufficiently soluble to be excreted are obtained.

Metabolizer (http://www.chemaxon.com/products/metabolizer/) is a commercially available module included in the JChem program package developed by ChemAxon in 2007 and designed for metabolism prediction. The Metabolizer program lists all possible metabolites of a specified compound and is able to predict metabolic pathways, major metabolites and stability of metabolites. The input data for Metabolizer are the biotransformation dictionary and the list of substrates for which metabolite prediction is to be performed. Currently, for the use of this program, one has to create own biotransformation dictionaries as the embedded library of biotransformation dictionaries for Metabolizer is still under development. Apart from the human metabolism, the program supports the prediction for many other species such as rats, bacteria and plants.

The **TIMES** platform[143] was developed in 2004 at the Laboratory of Mathematical Chemistry of the Burgas Free University (Bulgaria). The key designation of this program is to predict the toxicity of metabolites. The scenario of metabolite prioritization depending on their toxicity consists of four stages. The first stage is the determination whether the compound in question is organic. The subsequent stages are carried out only if the compound is organic. The second stage is the

Table 6. Characterization of expert systems developed for metabolism prediction.[a]

Name of the method	Type of the result	Termination criteria	Prioritization of metabolites and reactions	Web resource for prediction of reaction products (program downloading)
MetabolExpert	Tree of metabolites	Phase II metabolite has been generated	None	http://www.compudrug.com/metabolexpert
META	The same	LogP is sufficiently low for the compound to be excreted	Relative prioritization with a genetic algorithm	None
Meteor	"	The same	Relative prioritization of rules	Only paid subscription
EAWAG-PPS (UM-PPS)	"	A terminal compound has been reached, or a compound with less than 3 carbon atoms has been reached, or a layer of the tree comprising more than 10 compounds has been reached	The same	http://eawag-bbd.ethz.ch/predict/
SyGMa	"	The user specifies the maximum depth of the tree of metabolites	Metabolites are assigned probability scores based on probabilities of particular reactions	None
TIMES	"	The estimated probability of metabolite formation or the LogP value has reached a specified threshold	Metabolite prioritization according to toxicity based on QSAR models	None
JChem Metabolizer module	"	Termination criteria are specified manually	Reactions are prioritized according to their rates	https://www.chemaxon.com/products/online-tryouts/metabolizer/
Metaprint2D-React	List of possible sites of metabolism and biotransformations	None	Sites of metabolism and biotransformations are prioritized in terms of probability	http://www-metaprint2d.ch.cam.ac.uk/metaprint2d-react
MetaSite	List of direct metabolites	None	Present	http://www.moldiscovery.com/software/metasite/
MetaDrug	List of potential metabolites	–	–	Only paid subscription

[a] The dash stands for 'no data', *i.e.*, we did not find relevant data.

calculation of the physicochemical properties of the molecule. The third stage is the assessment of the probability of the compound to be metabolized in one of the enzyme systems of the body (*e.g.*, liver, skin or other tissues). If the compound is found to be biotransformable in the human tissues, metabolites are generated on the basis of biotransformation dictionary. Each biotransformation is assigned a priority reflecting its prevalence. TIMES is based on the heuristic algorithm; it generates a metabolic network, resulting from application of biotransformation rules to organic molecules which serve as the input data. Generation is stopped as soon as metabolites with a specified threshold probability of formation or a specified LogP value are achieved. The resulting set of metabolites can be sorted according to the toxicity predicted by means of QSAR models. Furthermore, TIMES can assess the confidence level of the generated metabolic pathways and metabolites.

The **MetaDrug** software,[144] as has already been noted, was developed by the GeneGo company, which was later purchased by Thomson Reuters. It comprises a number of QSAR models for substrates and inhibitors of cytochrome P450 enzymes; the models are applicable both to the parent compounds (substrates) and to predicted metabolites. The prediction is based on binary classifiers for CYP1A2, CYP2B6, CYP2D6 and CYP3A4 substrates and for CYP1A2, CYP2C19, CYP2C9, CYP2D6 and CYP3A4 inhibitors (for $IC_{50} < 10$ μmol litre^{-1}). For substrate prediction, the models were trained on published data retrieved from the MetaBase. The inhibitor prediction models were based on the activity data for cytochrome inhibitors from the PubChem DB. Training was accomplished by the recursive partitioning. The substrates and inhibitors were described by means of physicochemical descriptors. Each MetaBase model comprises a description that

includes information about the training set, the numbers of molecules in the training and test sets and training parameters such as sensitivity, specificity and accuracy. The applicability domain of each model is determined using the Tanimoto coefficient, which characterizes the similarity of the molecule being predicted and the training set molecules. MetaDrug is accessed through a web interface. Compounds that are to be subjected to evaluation can be uploaded in the SDF format and the prediction results can be either viewed online or exported as an Excel table. The MetaDrug software is not designed for analysis of large data arrays — predictions are possible only for files containing not more than 500 compounds.

The **Meteor** software[145] was developed by Lhasa in 1997 as an outgrowth of the Derek system designed for toxicity prediction. The idea was that some non-toxic compounds can be converted to toxic metabolites; therefore, for exhaustive evaluation of the toxicity of a compound, it is necessary to evaluate also its metabolites. Meteor uses a special language for structure description, which extends the biotransformation rules by additional data, for example, the charge and valence data, and a biotransformation base composed of data collected from databases, literature sources and consultations with professionals. Meteor also includes a knowledge base editor, which allows the users to convert the user databases to the Meteor format. This function makes Meteor operation much more efficient, as the user is not restricted to the embedded database but can use his own accumulated experimental results. The editor allows one to add information on the biotransformations and structure – activity relationships and to give references to literature sources in which these transformations are described. These sources are accessible from the knowledge base through hyperlinks formed by the editor.

The Meteor prediction consists of successive application of the biotransformation rules to the input compounds. To avoid generation of too many metabolites, each biotransformation is supplied with a qualitative probability evaluation. The probability is described by one of five values: probable, plausible, equivocal, doubted and improbable. During the prediction, these evaluations are used as absolute priorities, which in some cases reduces the number of metabolic pathways considered. The prediction stops after generation of the metabolite with a LogP value sufficient for excretion from the body. The probability for a compound to undergo biotransformation decreases with a decrease in the lipophilicity, because the ability of molecules to penetrate across membranes to undergo further biotransformations decreases. The Meteor developers warn that despite all the undertaken measures, the program generates rather large numbers of false metabolites; therefore, prediction results should be treated as only rough approximation requiring subsequent treatment and analysis. This, however, is true for almost any expert system.[146]

EAWAG-BBD Pathway Prediction System (formerly called UM-PPS[40]) is an expert system designed for predicting probable biodegradation pathways of organic compounds. It is based on the EAWAG-BBD (UM-BBD) database containing information on more than 1300 compounds, 900 enzymes, 1500 reactions and 500 microorganisms. EAWAG-BBD includes approximately 250 rules describing biotransformation reactions.

The general limitation of all systems based on the rules designed for the construction of metabolic networks is related to combinatorial explosion — the result of generation of a too large number of products at each iteration. There are several methods for solving this problem, in particular, restriction of the applicability domain of a particular biotransformation (decrease in the number of compounds to which a particular biotransformation can be applied) and removal of unlikely biotransformations.

EAWAG-PPS (UM-PPS) envisages a number of measures for preventing combinatorial explosion, which are based on prioritization of the transformation rules. For some pairs of rules, the relative priority calculated from more than 1000 reactions retrieved from EAWAG-BBD (UM-BBD) is indicated in the transformation base. During prediction, only the transformations that are most probable as regards the introduced priority are applied to the substrates; this reduces the number of predicted products by 10% – 15% with insignificant decrease in the prediction accuracy.

Also, super rules were incorporated into EAWAG-PPS (UM-PPS) for the effective control of combinatorial explosion. A super rule is a combination of several successive transformation rules representing a known metabolic pathway. An example is beta-oxidation consisting of several biotransformations. The application of super rules compresses the metabolic pathways in the tree of metabolites and reduces the number of predicted unlikely metabolites.[147]

The EAWAG-PPS (UM-PPS) system uses a relatively small number of rules for predicting a multitude of compounds; therefore the probability of each particular biotransformation is not constant but depends on a number of factors, in particular, structural details of the compounds to which the rule is applied. For increasing the prediction quality, variable probability scores for biotransformations of particular types were introduced. For this purpose, regular expressions were associated with some transformation rules. If during the prediction, a substrate structure complies with the description of the regular expression associated with the given transformation, this transformation is assigned a probability score based on the identified structural feature of the substrate.

SyGMa[148] is a system based on the rules that describe the biotransformations. Each rule is assigned an empirical probability score based on the statistical data collected from a large array of experimental results. This allows one to estimate the likelihood of formation of each of the predicted metabolites and rank metabolites according to the obtained likelihood score. A distinctive feature of the method is that the rules are modified and optimized depending on the probability scores in order to increase the prediction accuracy. The rule base was composed of the information retrieved from the BIOVIA Metabolite Database, version 2001. The reactions involving compounds with molecular mass of >900, inorganic compounds and molecules with ambiguously determined structures (containing symbols R or X) were discarded. Finally, 6187 reactions were

retrieved and 144 rules covering the major part of metabolism phases I and II were composed.

The rules chracterizing a particular biotransformation are described in SyGMa by means of the SMIRKS system. The probability scores were obtained by applying each rule to all compounds of the dataset (3144 compounds). For each rule, the number of correctly generated (using this rule) metabolites was divided by the total number of generated (using this rule) metabolites. This ratio was taken as the probability score for the metabolic rule. Some rules were refined or split into multiple rules to increase the probability scores. In order to avoid overfitting, the rules were kept as general as possible by means of special constraints. The rules for which the probability score did not exceed 0.01 were considered not predictive and were either refined or rejected.

The metabolite prediction process consists of application of all existing rules to a particular compound. If the prediction includes several steps, in each step, the whole set of generated metabolites is subjected to all existing rules; this gives the set of all predicted metabolites for the given step. If a cleavage reaction affords several products, each one is treated as a separate metabolite. The products of cleavage reactions incorporating $<15\%$ of atoms of the parent compound are considered irrelevant and removed from the metabolic tree. After completion of the metabolic network, each metabolite is assigned a probability score describing its formation on the basis of the reaction rules used to generate it. If a metabolite results from several successive reaction steps, the scores of particular steps are multiplied to give the overall score. If the same metabolite can be formed *via* several parallel reactions, the probability is calculated for each reaction, the highest value is selected and assigned to be the probability score for this metabolite. The prediction accuracy was estimated using the test set retrieved from the Biovia Metabolite Database, version 2006. About 68% of the metabolites were predicted correctly.

VII.2. Combined approaches

As has already been noted, the prediction results of virtually any expert system are replete with false metabolites.[146] The prediction quality can be improved by using expert systems in combination with other computational methods. For example, from analysis of the 3D structures of protein–ligand complexes for the cytochrome family, one can conclude that the positions and orientations of substrates and metabolites are largely conserved. Hence, the potential metabolites should be subject to chemical and geometric constraints of the binding site, and, as a consequence, the potential metabolites and non-metabolites can be discriminated by molecular docking. This was hypothesized by Tarcsay.[149] The hypothesis was verified for the MetabolExpert system, which was used to generate the potential metabolites for some substrates. The prediction thus obtained was treated with the GLIDE molecular docking software.[150] This 'hybrid' approach, that is, combination of an expert system with molecular docking, provided 69% of correct determinations of the true metabolite among the three most probable ones.

One more example of application of the combined approach to structure determination of potential metabolites is the **MetaSite** program[151] released in 2003 by Molecular Discovery. Initially, it was designed for predicting the SOMs. The prediction of SOMs is based on the description of interaction of the cytochrome active site with numerous small organic compounds from the training set using the GRID molecular fields[152] and on the calculation of 3D descriptors (the distance between the points of interaction between the substrate biotransformation site and enzyme active site) for training set compounds. The results of calculation make it possible to determine the orientation of the molecule in the cytochrome P450 active site without resorting to molecular modelling methods. For each atom of the compound, the distance from this atom to the iron atom in the cytochrome P450 active site is calculated. The determined distance is used to estimate the accessibility of the cytochrome binding site. For each atom, the reactivity is also estimated by semiempirical calculations. The biotransformation probability for some site in a training set compound is found as the product of its accessibility and reactivity scores. This determination of the most probable sites of metabolism for the parent compound reduces the number of metabolic pathways involved in the metabolism prediction, which accelerates the prediction and increases the prediction accuracy. There is also a special version of the program, Mass-MetaSite, which allows faster identification of the metabolite structure *in vitro* from recorded mass spectra.[153]

One more example of application of the combined approach for reducing the number of false metabolites is a method for predicting small organic metabolites that are formed in enzyme-catalyzed reactions developed at the Los Alamos National Laboratory.[154] This approach is based on combination of computational chemistry and machine learning methods. The key idea is that all the enzyme-catalyzed reactions can be split into several classes, each being characterized by transformation of some functional group. Thus, a binary classifier able to sort out substrates and products for a particular type of reaction can be trained for each reaction type.

The data for classifier training and testing were selected from the KEGG databases in two stages. In 2006, 6577 reactions were selected for which 4792 unique metabolites were known. In 2008, additional 755 reactions containing 1721 unique metabolites were selected. The data of 2006 were used to classify the reactions, training of classifiers and cross-validation. The data of 2008 were used to test the classifiers. The selected reactions were split into 80 classes. The reactions belonging to the same class were characterized by identical transformation of some functional group and the identical set of reaction centres in the substrates and the products.

All transformation sites from substrates into products (for 80 selected classes of reactions) together with the surroundings were described according to the SMARTS notation. For each reaction centre pattern, a binary classifier able to discriminate between the functional groups involved and not involved in the reaction was constructed using the SVM method.

The potential reactive functional groups were represented as vectors of features. The features comprised the

properties of atoms incorporated in the reaction centre and some molecular characteristics of the substrate (shape, surface, energy and charge distribution). The atomic and molecular properties were derived from 2D and 3D structures by means of the JOELib expert system,[155] Java Chemistry Development Kit library (CDK)[156] and the MOPAC 2007 program.[157, 158] The selection of hyperparameters during the classifier training was aimed at the maximum geometrical average sensitivity and specificity values found upon cross-validation in three blocks. A classifier hyperparameter is a parameter that must be specified before training. As a result, classifiers with average sensitivity of 0.74 and average specificity of 0.87 were obtained. The sensitivity is the fraction of positive examples correctly predicted to be positives and the specificity is the fraction of negative examples correctly predicted to be negatives. This method can be used in the construction of the tree of metabolites for predicting the most probable biotransformations and selection of the most probable metabolites.

All of the methods considered above generate the tree of metabolites for some specified compound. However, the problem of prediction of metabolic pathways is not always formulated in this way. A frequently used variation of this problem is as follows: two compounds are specified, and a biotransformation chain of a minimum length to convert one compound to the other one should be found. One of the most efficient algorithms for solving the problem set up in this way was described by Nakamura *et al.*[159] Compounds are represented as structural feature vectors the elements of which correspond to the number of occurrences of some substructure in the structure of the compound described. The reactions are reflected by pairs of vectors, the first one being the characteristic substrate vector and the second one being the reaction delta (the difference between the product and substrate vectors). The information required to compute the characteristic vectors was retrieved from the KEGG databases. The characteristic compound and reaction vectors were subject to two constraints. First, all vector elements should be non-negative; therefore, if application of some reaction to the compound gives rise to negative elements of the characteristic compound vector, this reaction cannot be applied to this compound. Second, for a given reaction to be applicable to a given compound, all elements of the characteristic vector of this compound should not be smaller than the corresponding elements of the characteristic reaction substrate vector.

The algorithm proposed by Nakamura *et al.*[159] is the problem definition in terms of the graph theory. In particular, by taking a graph in which nodes are associated with chemical compounds and edges are associated with biotransformations converting one compound to another, the problem is reduced to the search for the shortest path between some specified pair of nodes in this graph. Thus, it is possible to predict metabolic pathways and construct the metabolic network for integral evaluation of the metabolism.

Currently, systems based on biotransformation rules are most popular for generation of metabolite structures. This is due to a number of undeniable advantages over alternative methods. First of all, methods based on expert knowledge are distinguished by low computational difficulty. Indeed, the whole prediction process consists of a sequence of operations that represent the application of simple rules to chemical structures. Each operation is elementary and is not even commensurable in complexity with quantum chemical calculations. An important feature of these systems is that the researcher can intervene into generation of metabolite structures, conduct the process in the semiautomatic mode and control it in the desired way. Thus, systems operating by means of metabolite generation rules automate the routine operations, leaving the right to make key decisions to the researcher. The intervention into the system operation is done by tuning the prediction parameters, for example, the prediction depth, and by adding thresholds for metabolite cutting off. The tuning can be made both in the very beginning and after each iteration. While using these systems, the researcher can easily choose only a part of the transformation rules that correspond to the types of metabolism he is interested in. Moreover, the researcher can extend the scope of the system by adding new transformation rules into the system base. Perhaps, this simple and simultaneously flexible tuning is not inherent in any other method for metabolite prediction. Yet another fairly important feature of the systems based on rules is the possibility to associate each particular prediction with literature sources that describe the relevant types of reactions.[23] This function undoubtedly simplifies and accelerates the research.

However, despite all the listed advantages, the use of systems based on metabolite structure generation rules is faced with some problems. The major drawback of these methods is so-called combinatorial explosion, *i.e.*, the exponential growth of the number of generated metabolites.[23] In practice, this approach markedly restricts the method, and a lot of heuristic algorithms have been developed to control this phenomenon. However, when the number of predicted metabolites is decreased, many correctly predicted metabolites can also be rejected, that is, a new problem arises. One more problem is that these systems generate numerous false positive metabolites;[146] therefore, it is very important to sort out and reject exactly the false results.

VIII. Conclusion

This review considers the key computer-aided methods for estimating the metabolic pathways of xenobiotics in the human body ranging from interaction of a small organic compound with the enzyme catalyzing the biotransformation process to the generation of potential metabolite structures. It is noteworthy that, although each of the developed methods has certain advantages and limitations, by combination of computational methods, it is possible in some cases to overcome the existing drawbacks and make prediction with a reasonable accuracy. Furthermore, combined computational methods (which are probably still to be developed) allow integral evaluation of metabolism for predicting the biological properties of both xenobiotics and all possible final and intermediate metabolites.

This is supported by the fact that a recent trend in the QSAR modelling is the use of consensus models.

Several models based on different mathematical methods and/or different types of descriptors make use of the same training and validation sets, and the consensus model aggregates the predictions of all the models used. This can be done by averaging the prediction results (simple unweighted consensus) or averaging with some weight for each model (weighted consensus).[160, 161] The use of consensus models is believed to reduce the variability of particular models, resulting in more robust predictions.[162] This statement is applicable to both quantitative and qualitative (Q)SAR models.

The properties of drug candidates that are studied by *in silico* methods developed for prediction of xenobiotic metabolism can, in principle, be determined by appropriate *in vivo* and *in vitro* studies during the drug development. However, the probable termination of the drug development at late phase clinical trials due to unfavourable properties of drug candidates such as unacceptable metabolism rate, toxicity of metabolites, possible drug – drug interaction and so on boosts the application of computer-aided methods to predict these properties as early as possible in the stage of theoretical exploration and planning of the synthesis of lead compounds. Thus, at the present stage of development, the methods of predicting the metabolism of xenobiotics *in silico* are not meant to completely replace the *in vivo* and *in vitro* studies but are expected to decrease their costs and time expenses by reducing the scope and prioritization of these studies and decreasing the risk of failure in the search for new drugs.

The understanding of metabolism is a key factor in drug development; therefore, investigations in this field require a lot of effort. Although quite a few methods for computer prediction of metabolic pathways have been developed in recent years, they all are still far from being perfect; hence, studies along this line will be undoubtedly continued in the near future.

It should be noted, however, that the approaches to metabolism prediction that we considered, which are mainly based on only the structures of compounds, are fairly rough approximations. While developing QSAR models, one should also bear in mind the molecular-biological features of drug-metabolizing enzymes. On the one hand, numerous xenobiotics are metabolized by various enzyme isoforms; on the other hand, some enzyme isoforms have low specificity of interaction with xenobiotics (for example, CYP3A4 metabolizes $> 50\%$ of the known drugs). Furthermore, different isoforms can catalyze identical biotransformation reactions.

Actually, metabolism is affected by a multitude of genetic, phenotypic and ontogenetic individual factors, for example, such as being a 'slow' or 'fast' metabolizer for particular biotransformation enzyme and so on. This brings about the idea that there exists a limit for the accuracy of a prediction based on the drug structure alone.

In conclusion, it should be noted that owing to the limited size of the review, we did not consider a number of important aspects of drug metabolism: induction of metabolic enzymes, prediction of the metabolism rate, bioavailability and half-life,[163] physiologically based pharmacokinetic and pharmacodynamic (PBPK/PBPD) modelling[11] and problems of transferability of the results of *in vitro* testing to the effects observed in *in vivo* experiments and transferability of the results obtained in animal experiments to humans.[164, 165] For the state of the art of these issues, the interested reader is referred to the publications mentioned above.

This review was prepared with the financial support of the Russian Science Foundation (Project No. 14-15-00449).

References

1. B Testa, in *The Practice of Medicinal Chemistry* (Ed. C G Wermuth) (Amsterdam: Academic Press, 2008) p. 655
2. M L Peach, A V Zakharov, R Liu, A Pugliese, G Tawa, A Wallqwist, M C Nicklaus *Future Med. Chem.* **4** 1907 (2012)
3. J L Castrejon, N Berry, S El-Ghaiesh, B Gerber, W J Pichler, B K Park, D J Naisbitt *J. Allergy Clin. Immunol.* **125** 411 (2010)
4. J Chinen, W T Shearer *J. Allergy Clin. Immunol.* **127** 336 (2011)
5. G T Arce, D R Vincent, M J Cunningham, W N Choy, A M Sarrif *Environ. Health Perspect.* **86** 75 (1990)
6. M M Moore, K Harrington-Brock *Environ. Health Perspect.* **108** (Suppl. 2) 215 (2000)
7. K M Lakin, Yu F Krylov *Biotransformatsiya Lekarstvennykh Veshchestv* (Biotransformation of Drugs) (Moscow: Meditsina, 1981)
8. A-C Macherey, P M Dansette, in *The Practice of Medicinal Chemistry* (Ed. C G Wermuth) (Amsterdam: Academic Press, 2008) p. 674
9. F P Guengerich *Chem. Res. Toxicol.* **21** 70 (2008)
10. K-M Wu *Pharmaceuticals* **2** 77 (2009)
11. *Encyclopedia of Drug Metabolism and Interactions* (1st Ed.) (Ed. A V Lyubimov) (Wiley, 2012)
12. C Ioannides, D F Lewis *Curr. Top. Med. Chem.* **4** 1767 (2004)
13. P R Ortiz de Montellano *Future Med. Chem.* **5** 213 (2013)
14. W Chen, L L Koenigs, S J Thompson, R M Peter, A E Rettie, W F Trager, S D Nelson *Chem. Res. Toxicol.* **11** 295 (1998)
15. V G Granik *Metabolizm Ekzogennykh Soedinenii* (Metabolism of Exogenous Compounds) (Moscow: Vuzovskaya Kniga, 2006)
16. P D Hansten, J R Horn *The Top 100 Drug Interactions: A Guide to Patient Management* (Edmonds, WA: H and H Publishing, 2010)
17. G N Levy, J V Ioia, E J Kuchinskas *J. Pharm. Sci.* **63** 1197 (1974)
18. M Couvaris, P Galanopoulou, C Karageorgiou, H A Jalil, D Varonos *Eur. J. Drug Metab. Pharmacokinet.* **10** 27 (1985)
19. F P Guengerich *Carcinogenesis* **21** 345 (2000)
20. E Wittwer, S E Kern *AAPS J.* **8** E348 (2006)
21. J Wang, L Urban, D Bojanic *Expert Opin. Drug Metab. Toxicol.* **3** 641 (2007)
22. S A Kulkarni, J Zhu, S Blechinger *Xenobiotica* **35** 955 (2005)
23. J Kirchmair, M J Williamson, J D Tyzack, L Tan, P J Bond, A Bender, R C Glen *J. Chem. Inf. Model.* **52** 617 (2012)
24. A V Veselovskii, B N Sobolev, M S Zharkova, A I Archakov *Biomed. Khim.* **56** 90 (2010)
25. J Kirchmair, A H Göller, D Lang, J Kunze, B Testa, I D Wilson, R C Glen, G Schneider *Nat. Rev. Drug Discov.* **14** 387 (2015)
26. L Olsen, C Oostenbrink, F S Jørgensen *Adv. Drug Deliv. Rev.* **86** 61 (2015)

27. A Zakharov, A Lagunin, in *Application of Computational Techniques in Pharmacy and Medicine* (*Challenges and Advances in Computational Chemistry and Physics* Vol. 17) (Springer, 2014) p. 325
28. H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, P E Bourne *Nucleic Acids Res.* **28** 235 (2000)
29. B C Monk, T M Tomasiak, M V Keniya, F U Huschmann, J D A Tyndall, J D O'Connell, R D Cannon, J G McDonald, A Rodriguez, J S Finer-Moore, R M Stroud *Proc. Natl. Acad. Sci. USA* **111** 3865 (2014)
30. G Papadatos, J P Overington *Future Med. Chem.* **6** 361 (2014)
31. M Davies, M Nowotka, G Papadatos, N Dedman, A Gaulton, F Atkinson, L Bellis, J P Overington *Nucleic Acids Res.* **43** W612 (2015)
32. M Davies, N Dedman, A Hersey, G Papadatos, M D Hall, L Cucurull-Sanchez, P Jeffrey, S Hasan, P J Eddershaw, J P Overington *Bioinformatics* **31** 1695 (2015)
33. Y Wang, J Xiao, T O Suzek, J Zhang, J Wang, S H Bryant *Nucleic Acids Res.* **37** W623 (2009)
34. NCBI Resource Coordinators *Nucleic Acids Res.* **43** D6 (2015)
35. S Kim, P A Thiessen, E E Bolton, S H Bryant *Nucleic Acids Res.* **43** W605 (2015)
36. V Law, C Knox, Y Djoumbou, T Jewison, A C Guo, Y Liu, A Maciejewski, D Arndt, M Wilson, V Neveu, A Tang, G Gabriel, C Ly, S Adamjee, Z T Dame, B Han, Y Zhou, D S Wishart *Nucleic Acids Res.* **42** 1091 (2014)
37. T Liu, Y Lin, X Wen, R N Jorissen, M K Gilson *Nucleic Acids Res.* **35** D198 (2007)
38. D S Wishart, T Jewison, A C Guo, M Wilson, C Knox, Y Liu, Y Djoumbou, R Mandal, F Aziat, E Dong, S Bouatra, I Sinelnikov, D Arndt, J Xia, P Liu, F Yallou, T Bjorndahl, R Perez-Pineiro, R Eisner, F Allen, V Neveu, R Greiner, A Scalbert *Nucleic Acids Res.* **41** D801 (2013)
39. M Kanehisa, S Goto, Y Sato, M Furumichi, M Tanabe *Nucleic Acids Res.* **40** D109 (2012)
40. J Gao, L B M Ellis, L P Wackett *Nucleic Acids Res.* **38** D488 (2010)
41. S Preissner, K Kroll, M Dunkel, C Senger, G Goldsobel, D Kuzman, S Guenther, R Winnenburg, M Schroeder, R Preissner *Nucleic Acids Res.* **38** D237 (2010)
42. M F Hoffmann, S C Preissner, J Nickel, M Dunkel, R Preissner, S Preissner *Nucleic Acids Res.* **42** D1113 (2014)
43. D Cao, J Wang, R Zhou, Y Li, H Yu, T Hou *J. Chem. Inf. Model.* **52** 1132 (2012)
44. T Kalliokoski, C Kramer, A Vulpetti, P Gedeck *PLoS One* **8** e61007 (2013)
45. C Kramer, T Kalliokoski, P Gedeck, A Vulpetti *J. Med. Chem.* **55** 5165 (2012)
46. O A Tarasova, A F Urusova, D A Filimonov, M C Nicklaus, A V Zakharov, V V Poroikov *J. Chem. Inf. Model.* **55** 1388 (2015)
47. M Waldman, R Fraczkiewicz, R D Clark *J. Comput.-Aided Mol. Des.* **29** 897 (2015)
48. A Cherkasov, E N Muratov, D Fourches, A Varnek, I I Baskin, M Cronin, J Dearden, P Gramatica, Y C Martin, R Todeschini, V Consonni, V E Kuz'min, R Cramer, R Benigni, C Yang, J Rathman, L Terfloth, J Gasteiger, A Richard, A Tropsha *J. Med. Chem.* **57** 4977 (2014)
49. O A Raevsky *Russ. Chem. Rev.* **68** 505 (1999)
50. Ya A Ivanenkov, E V Bovina, K V Balakin *Russ. Chem. Rev.* **78** 465 (2009)
51. D A Filimonov, V V Poroikov *Ros. Khim. Zh.* **50** (2) 66 (2006)
52. R Barone, M Chanon, in *Handbook of Chemoinformatics* (Ed. J Gasteiger) (Weinheim: Wiley-VCH, 2003) p. 1428
53. D Filimonov, V Poroikov, Yu Borodina, T Gloriozova *J. Chem. Inf. Comput. Sci.* **39** 666 (1999)
54. A V Rudik, A V Dmitriev, A A Lagunin, D A Filimonov, V V Poroikov *J. Chem. Inf. Model.* **54** 498 (2014)
55. Yu Borodina, A Rudik, D Filimonov, N Kharchevnikova, A Dmitriev, V Blinova, V Poroikov *J. Chem. Inf. Comput. Sci.* **44** 1998 (2004)
56. J D Tyzack, H Y Mussa, M J Williamson, J Kirchmair, R C Glen *J. Cheminf.* **6** 29 (2014)
57. J Zaretzki, C Bergeron, P Rydberg, T-w Huang, K P Bennett, C M Breneman *J. Chem. Inf. Model.* **51** 1667 (2011)
58. M Zheng, X Luo, Q Shen, Y Wang, Y Du, W Zhu, H Jiang *Bioinformaics* **25** 1251 (2009)
59. S B Singh, L Q Shen, M J Walker, R P Sheridan *J. Med. Chem.* **46** 1330 (2003)
60. A R Leach, J Bradshaw, D V S Green, M M Hann, J J Delany *J. Chem. Inf. Comput. Sci.* **39** 1161 (1999)
61. D Weininger *J. Chem. Inf. Comput. Sci.* **28** 31 (1988)
62. A N Tevyashova, E N Olsufyeva, M N Preobrazhenskaya *Russ. Chem. Rev.* **84** 61 (2015)
63. L G Dezhenkova, V B Tsvetkov, A A Shtil *Russ. Chem. Rev.* **83** 82 (2014)
64. N P E Vermeulen *Curr. Top. Med. Chem.* **3** 1227 (2003)
65. A Long, J D Walker *Environ. Toxicol. Chem.* **22** 1894 (2003)
66. P Vasanthanathan, O Taboureau, C Oostenbrink, N P E Vermeulen, L Olsen, F S Jørgensen *Drug Metab. Dispos.* **37** 658 (2009)
67. D Schuster, T M Steindl, T Langer *Curr. Top. Med. Chem.* **6** 1627 (2006)
68. D F V Lewis *InflammoPharmacology* **11** 43 (2003)
69. L Terfloth, B Bienfait, J Gasteiger *J. Chem. Inf. Model.* **47** 1688 (2007)
70. R E White, M-B McCarthy *Arch. Biochem. Biophys.* **246** 19 (1986)
71. D F V Lewis, C Ioannides, D V Parke *Biochem. Pharmacol.* **50** 619 (1995)
72. D F V Lewis *J. Chem. Technol. Biotechnol.* **76** 237 (2001)
73. N K Mishra, S Agarwal, G P S Raghava *BMC Pharmacol.* **10** 8 (2010)
74. G A Bakken, P C Jurs *J. Chem. Inf. Comput. Sci.* **41** 1255 (2001)
75. D F Lewis, S Modi, M Dickins *Drug Metab. Drug Interact.* **18** 221 (2001)
76. D A Filimonov, A A Lagunin, T A Gloriozova, A V Rudik, D S Druzhilovskii, P V Pogodin, V V Poroikov *Chem. Heterocycl. Compd.* **50** 444 (2014) [*Khim. Geterotsikl. Soedin.* 483 (2014)]
77. D Horvath, B Mao, R Gozalbes, F Barbosa, S L Rogalski, in *Chemoinformatics in Drug Discovery* (Ed. T I Oprea) (Weinheim: Wiley-VCH, 2005) p. 117
78. S Toba, J Srinivasan, A J Maynard, J Sutter *J. Chem. Inf. Model.* **46** 728 (2006)
79. B C Jones, G Hawksworth, V A Horne, A Newlands, J Morsman, M S Tute, D A Smith *Drug Metab. Dispos.* **24** 260 (1996)
80. D Dong, B Wu, D Chow, M Hu *Drug Metab. Rev.* **44** 192 (2012)
81. S-L Mo, W-F Liu, Y Chen, H-B Luo, L-B Sun, X-W Chen, Z-W Zhou, K B Sneed, C G Li, Y-M Du, J Liang, S-F Zhou *Comb. Chem. High Throughput Screen.* **15** 36 (2012)
82. C G Wermuth, C R Ganellin, P Lindberg, L A Mitscher *Pure Appl. Chem.* **70** 1129 (1998)

83. O F Guner, J P Bowen *Curr. Top. Med. Chem.* **13** 1327 (2013)
84. K-M Thai, T-D Ngo, T-D Tran, M-T Le *Curr. Top. Med. Chem.* **13** 1002 (2013)
85. L Koymans, N P E Vermeulen, S A B E Van Acker, J M Te Koppele, J J P Heykants, K Lavrijsen, W Meuldermans, G M Donne-Op den Kelder *Chem. Res. Toxicol.* **5** 211 (1992)
86. G R Strobl, S von Kruedener, J Stöckigt, F P Guengerich, T Wolff *J. Med. Chem.* **36** 1136 (1993)
87. M J de Groot, M J Ackland, V A Horne, A A Alex, B C Jones *J. Med. Chem.* **42** 4062 (1999)
88. B Bordás, T Kőmtves, A Lopata *Pest Manage. Sci.* **59** 393 (2003)
89. L E Korhonen, M Rahnasto, N J Mähönen, C Wittekindt, A Poso, R O Juvonen, H Raunio *J. Med. Chem.* **48** 3808 (2005)
90. C W Locuson, H Suzuki, A E Rettie, J P Jones *J. Med. Chem.* **47** 6768 (2004)
91. N S Zefirov, V A Palyulin, E V Radchenko *Dokl. Akad. Nauk* **352** 630 (1997)
92. K H Kim, G Greco, E Novellino, in *3D QSAR in Drug Design.: Recent Advances* Vol. 3 (Eds H Kubinyi, G Folkers, Y Martin) (Dordrecht: Kluwer Academic, 2002) p. 257
93. Y Yao, W-W Han, Y-H Zhou, Z-S Li, Q Li, X-Y Chen, D-F Zhong *Eur. J. Med. Chem.* **44** 854 (2009)
94. N V Belkina, V S Skvortsov, A S Ivanov, A I Archakov *Vopr. Med. Khim.* **44** 464 (1998)
95. J K Yano, M R Wester, G A Schoch, K J Griffin, C D Stout, E F Johnson *J. Biol. Chem.* **279** 38091 (2004)
96. M R Wester, J K Yano, G A Schoch, C Yang, K J Griffin, C D Stout, E F Johnson *J. Biol. Chem.* **279** 35630 (2004)
97. P Rowland, F E Blaney, M G Smyth, J J Jones, V R Leydon, A K Oxbrow, C J Lewis, M G Tennant, S Modi, D S Eggleston, R J Chenery, A M Bridges *J. Biol. Chem.* **281** 7614 (2006)
98. S Sansen, J K Yano, R L Reynald, G A Schoch, K J Griffin, C D Stout, E F Johnson *J. Biol. Chem.* **282** 14348 (2007)
99. G Brändén, T Sjögren, V Schnecke, Y Xue *Drug Discov. Today* **19** 905 (2014)
100. E E Scott, J R Halpert *Trends Biochem. Sci.* **30** 5 (2005)
101. R Meech, P I Mackenzie *Arch. Biochem. Biophys.* **356** 77 (1998)
102. M Ouzzine, J Magdalou, B Burchell, S Fournel-Gigleux *J. Biol. Chem.* **274** 31401 (1999)
103. M J Miley, A K Zielinska, J E Keenan, S M Bratton, A Radominska-Pandya, M R Redinbo *J. Mol. Biol.* **369** 498 (2007)
104. G Sliwoski, S Kothiwale, J Meiler, E W Lowe Jr *Pharmacol. Rev.* **66** 334 (2014)
105. A R Leach *Molecular Modelling: Principles and Applications* (2nd Ed.) (Dorchester: Pearson Education, 2001)
106. K V Shaitan, in *Stochastic Dynamics of Reacting Biomolecules* (Eds W Ebeling, L Schimansky-Geier, Yu M Romanovsky) (Singapore: World Scientific, 2003) p. 285
107. Y Shan, E T Kim, M P Eastwood, R O Dror, M A Seeliger, D E Shaw *J. Am. Chem. Soc.* **133** 9181 (2011)
108. D B Kitchen, H Decornez, J R Furr, J Bajorath *Nat. Rev. Drug Discov.* **3** 935 (2004)
109. P P Nandekar, A T Sangamwar *Expert Opin. Drug Discov.* **7** 771 (2012)
110. P Vasanthanathan, J Hritz, O Taboureau, L Olsen, F S Jørgensen, N P E Vermeulen, C Oostenbrink *J. Chem. Inf. Model.* **49** 43 (2009)
111. P Vasanthanathan, L Olsen, F S Jørgensen, N P E Vermeulen, C Oostenbrink *Drug Metab. Dispos.* **38** 1347 (2010)
112. E Stjernschantz, C Oostenbrink *Biophys. J.* **98** 2682 (2010)
113. L Perić-Hassler, E Stjernschantz, C Oostenbrink, D P Geerke *Int. J. Mol. Sci.* **14** 24514 (2013)
114. U Bren, C Oostenbrink *J. Chem. Inf. Model.* **52** 1573 (2012)
115. C de Graaf, C Oostenbrink, P H J Keizers, T van der Wijst, A Jongejan, N P E Vermeulen *J. Med. Chem.* **49** 2417 (2006)
116. C de Graaf, P Pospisil, W Pos, G Folkers N P E Vermeulen *J. Med. Chem.* **48** 2308 (2005)
117. V H Teixeira, V Ribeiro, P J Martel *Biochim. Biophys. Acta* **1804** 2036 (2010)
118. J Hritz, A de Ruiter, C Oostenbrink *J. Med. Chem.* **51** 7469 (2008)
119. M L Danielson, P V Desai, M A Mohutsky, S A Wrighton, M A Lill *Eur. J. Med. Chem.* **46** 3953 (2011)
120. S L Moors, A M Vos, M D Cummings, H Van Vlijmen, A Ceulemans *J. Med. Chem.* **54** 6098 (2011)
121. J Jung, N D Kim, S Y Kim, I Choi, K-H Cho, W S Oh, D N Kim, K T No *J. Chem. Inf. Model.* **48** 1074 (2008)
122. C de Graaf, N P E Vermeulen, K A Feenstra *J. Med. Chem.* **48** 2725 (2005)
123. K Korzekwa, W Trager, M Gouterman, D Spangler, G H Loew *J. Am. Chem. Soc.* **107** 4273 (1985)
124. A T Pudzianowski, G H Loew, B A Mico, R V Branchflower, L R Pohl *J. Am. Chem. Soc.* **105** 3434 (1983)
125. P N D'yachkov *Kvantovokhimicheskie Raschety v Izuchenii Mekhanizma Deistviya i Toksichnosti Chuzherodnykh Veshchestv (Itogi Nauki i Tekhniki. Toksikologiya)* [Quantum Chemical Calculations in the Study of the Mechanism of Action and Toxicity of Foreign Substances (Advances in Science and Engineering. Toxicology) Vol.16] (Moscow: Izd. VINITI, 1990)
126. M J S Dewar, W Thiel *J. Am. Chem. Soc.* **99** 4899 (1977)
127. J P Jones, M Mysinger, K R Korzekwa *Drug Metab. Dispos.* **30** 7 (2002)
128. M J S Dewar, E G Zoebisch, E F Healy, J J P Stewart *J. Am. Chem. Soc.* **107** 3902 (1985)
129. P N D'yachkov, N V Kharchevnikova, A V Dmitriev, A V Kuznetsov, V V Poroikov *Int. J. Quantum Chem.* **107** 2454 (2007)
130. M Hennemann, A Friedl, M Lobell, J Keldenich, A Hillisch, T Clark, A H Glller *ChemMedChem* **4** 657 (2009)
131. P Rydberg, D E Gloriam, J Zaretzki, C Breneman, L Olsen *ACS Med. Chem. Lett.* **1** 96 (2010)
132. R G Parr, W Yang *Density-Functional Theory of Atoms and Molecules* (New York, Oxford: Oxford University Press, 1989)
133. R P Sheridan, K R Korzekwa, R A Torres, M J Walker *J. Med. Chem.* **50** 3173 (2007)
134. A Rudik, A Dmitriev, A Lagunin, D Filimonov, V Poroikov *Bioinformatics* **31** 2046 (2015)
135. C L DeVane, C B Nemeroff *Clin. Pharmacokinet.* **40** 509 (2001)
136. L J Ereshefsky *J. Clin. Psychiatry* **57** (Suppl. 11) 12 (1996)
137. C B Nemeroff, B Kinkead, J Goldstein *J. Clin. Psychiatry* **63** (Suppl. 13) 5 (2002)
138. K-Y Li, X Li, Z N Cheng, H D Li *Methods Find. Exp. Clin. Pharmacol.* **27** 83 (2005)
139. G Klopman, M Dimayuga, J Talafous *J. Chem. Inf. Comput Sci.* **34** 1320 (1994)
140. J Talafous, L M Sayre, J J Mieyal, G Klopman *J. Chem. Inf. Comput. Sci.* **34** 1326 (1994)
141. F Darvas, in *QSAR in Environmental Toxicology – II* (Ed. K L E Kaiser) (Dordrecht: Springer-Verlag, 1987) p. 71

142. B Testa, P Jenner *Drug Metabolism: Chemical and Biochemical Aspects* (Ed. H Oelschläger) (New York: Marcel Dekker, 1976)
143. O G Mekenyan, S D Dimitrov, T S Pavlov, G D Veith *Curr. Pharm. Des.* **10** 1273 (2004)
144. S Ekins, S Andreyev, A Ryabov, E Kirillov, E A Rakhmatulin, A Bugrim, T Nikolskaya *Expert Opin. Drug Metab. Toxicol.* **1** 303 (2005)
145. C A Marchant, K A Briggs, A Long *Toxicol. Mech. Methods* **18** 177 (2008)
146. A Boobis, U Gundert-Remy, P Kremers, P Macheras, O Pelkonen *Eur. J. Pharm. Sci.* **17** 183 (2002)
147. K Fenner, J Gao, S Kramer, L Ellis, L Wackett *Bioinformatics* **24** 2079 (2008)
148. L Ridder, M Wagener *ChemMedChem* **3** 821 (2008)
149. Á Tarcsay, R Kiss, G M Keserű *J. Comput.-Aided Mol. Des.* **24** 399 (2010)
150. R A Friesner, J L Banks, R B Murphy, T A Halgren, J J Klicic, D T Mainz, M P Repasky, E H Knoll, M Shelley, J K Perry, D E Shaw, P Francis, P S Shenkin *J. Med. Chem.* **47** 1739 (2004)
151. G Cruciani, E Carosati, B De Boeck, K Ethirajulu, C Mackie, T Howe, R Vianello *J. Med. Chem.* **48** 6970 (2005)
152. P J Goodford *J. Med. Chem.* **28** 849 (1985)
153. B Bonn, C Leandersson, F Fontaine, I Zamora *Rapid Commun. Mass Spectrom.* **24** 3127 (2010)
154. F Mu, C J Unkefer, P J Unkefer, W S Hlavacek *Bioinformatics* **27** 1537 (2011)
155. R Guha, M T Howard, G R Hutchison, P Murray-Rust, H Rzepa, C Steinbeck, J Wegner, E L Willighagen *J. Chem. Inf. Model.* **46** 991 (2006)
156. C Steinbeck, Y Han, S Kuhn, O Horlacher, E Luttmann, E Willighagen *J. Chem. Inf. Comput. Sci.* **43** 493 (2003)
157. J J P Stewart *J. Comput.-Aided Mol. Des.* **4** 1 (1990)
158. J J P Stewart *J. Mol. Model.* **13** 1173 (2007)
159. M Nakamura, T Hachiya, Y Saito, K Sato, Y Sakakibara *BMC Bioinf.* **13** (Suppl. 17) S8 (2012)
160. M Hewitt, M T D Cronin, J C Madden, P H Rowe, C Johnson, A Obi, S J Enoch *J. Chem. Inf. Model.* **47** 1460 (2007)
161. A Lagunin, D Filimonov, A Zakharov, W Xie, Y Huang, F Zhu, T Shen, J Yao, V Poroikov *QSAR Comb. Sci.* **28** 806 (2009)
162. S Geman, E Bienenstock, R Doursat *Neural Comput.* **4** 1 (1992)
163. A V Zakharov, M L Peach, M Sitzmann, I V Filippov, H J McCartney, L H Smith, A Pugliese, M C Nicklaus *Future Med. Chem.* **4** 1933 (2012)
164. E Nielsen, G Østergaard, J Ch Larsen *Toxicological Risk Assessment of Chemicals: A Practical Guide* (New York: CRC Press, 2008)
165. P Perel, I Roberts, E Sena, P Wheble, C Briscoe, P Sandercock, M Macleod, L E Mignini, P Jayaram, K S Khan *Br. Med. J.* **334** 197 (2007)