JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

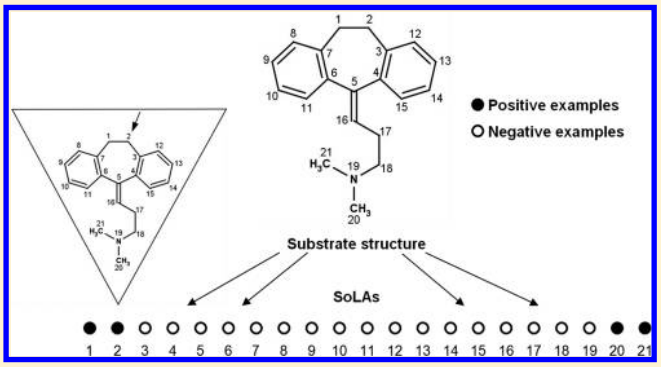
Metabolism Site Prediction Based on Xenobiotic Structural Formulas and PASS Prediction Algorithm

Anastasia V. Rudik,* Alexander V. Dmitriev, Alexey A. Lagunin, Dmitry A. Filimonov, and Vladimir V. Poroikov

Orekhovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, Building 10/8, Pogodinskaya Str., Moscow, 119121, Russia

Supporting Information

ABSTRACT: A new ligand-based method for the prediction of sites of metabolism (SOMs) for xenobiotics has been developed on the basis of the LMNA (labeled multilevel neighborhoods of atom) descriptors and the PASS (prediction of activity spectra for substances) algorithm and applied to predict the SOMs of the 1A2, 2C9, 2C19, 2D6, and 3A4 isoforms of cytochrome P450. An average IAP (invariant accuracy of prediction) of SOMs calculated by the leave-one-out cross-validation procedure was 0.89 for the developed method. The external validation was made with evaluation sets containing data on biotransformations for 57 cardiovascular drugs. An average IAP of regioselectivity for evaluation sets was 0.83. It was shown that the proposed method exceeds accuracy of SOM prediction by RS-Predictor for CYP 1A2, 2D6, 2C9, 2C19, and 3A4 and is comparable to or better than SMARTCyp for CYP 2C9 and 2D6.



INTRODUCTION

Many xenobiotics, including drugs, are metabolized by multiple enzyme systems in the human organism; the main set of enzymes is the cytochrome P450 superfamily (CYP).[1] Approximately 75% of drugs are metabolized mainly by 1A2, 2C9, 2C19, 2D6, and 3A4 cytochrome P450 isoforms.[2] These enzymes catalyze aromatic hydroxylation, N-dealkylation, O-dealkylation, S-oxidation, N-oxidation, epoxidation, and some other reactions. The active site of cytochromes contains a covalently bound heme located in the inner pocket formed by amino acid residues. CYP attacks a particular carbon atom or a heteroatom during its interaction with a substrate.[3,4] This atom is called a "site of metabolism" (SOM) or "site of attack"[5] of the substrate that leads to the formation of intermediate products, such as a carbon radical (resulting from the process of hydrogen atom abstraction) or a carbocation (the result of a one-electron transfer).[3] Further transformation of a substrate molecule into the product depends on the electronic and steric environment of the SOM in the active site of the CYP.

There are both ligand-based and structure-based approaches for the prediction of SOMs. The ligand-based approaches use data on substrates of metabolizing enzymes,[6−9] and the structure-based approaches use data on the active sites of enzymes that metabolize xenobiotics.[10,11] These two approaches can also be used in combination for the prediction of SOMs.[12]

According to another classification method, which was presented by Tarcsay and Keseru in their review,[13] SOM prediction methods can be divided into four main groups:

(1) orientation effects based on methods that predict the first substrate recognition step using information either from the protein structure or the spatial alignment of known substrates;
(2) mechanism-based methods that take into account the electronic effects of the rate-determining step by calculating the hydrogen abstraction energy in model systems;
(3) combined models that take into account orientation effects as well as electronic effects;
(4) empirical models (expert systems) that consider the outcome of the catalytic cycle as represented in the database of known biotransformations.

This classification is based on the consideration of different states in the catalytic cycle of CYP.

According to the computational techniques used for investigation in the field of SOM prediction, the approaches may be divided into six groups: (a) reactivity-based approaches, (b) fingerprint-based data mining approaches, (c) machine-learning approaches, (d) molecular interaction fields, (e) shape-focused approaches, and (f) protein−ligand docking.[12]

Received: August 9, 2013
Published: January 13, 2014

There are various models and approaches for SOM prediction for different isoforms of CYP.[13] The combination of particular methods that covers supplementary aspects of CYP interaction with substrates often provide more accurate predictions than individual approaches.[12] The combination of the mechanism-based approach with orientation effects is suggested as a straightforward and promising way to improve the prediction accuracy.[13] In general, consensus modeling is a viable method of combining different models.[14] Several different applied combinations of SOM prediction approaches have been published: combination of quantum chemical analysis with docking (MLite[15]); combination of pharmacophore-based approach, homology modeling and molecular orbital calculation;[16] and combination of quantum chemical analysis with a ligand-based model (StarDrop[17]). The machine learning-based multidescriptors approach[18,19] and precalculated density functional theory activation energies[20] were also used for SOM prediction.

A description of xenobiotic structures is very important for SOM prediction. SMARTS-defined chemical substructures[18,19] or other topological and quantum chemical descriptors were used for SOM prediction. For example, RS-Predictor represents potential SOMs through a combination of 148 topological and 392 quantum-chemical, atom-specific descriptors that are grouped together as metabolophores.[21] At the same time, topological accessibility descriptors may also be used to determine the ability of the atom to be attacked by the enzyme.[20]

SMARTCyp and RS-Predictor are able to predict SOMs for substrates of the main CYP isoforms.[22] Both methods use information about the structure of compounds and their quantum-chemical parameters. This requires additional computations or information about the active site of the enzyme and about the energy required for the interaction of the enzyme's active center with the SOMs.

Almost all of the above-mentioned methods use data about the 3D structure of the enzyme and/or the quantum-chemical characteristics of the substrate in model creation or prediction. Therefore, there is no single method of SOM prediction that is based only on 2D structural formulas of substrates.

The aim of our study was to develop a method that would allow the prediction of SOMs for the major isoforms of cytochrome P450 on the basis of 2D structural information and xenobiotic biotransformations catalyzed by P450 isoforms. This method should not require 3D structural information of enzymes and substrates and/or quantum chemical characteristics of the substrates to build a model.

Earlier, we developed a method to predict the biotransformation of xenobiotics on the basis of structural formulas using MNA (multilevel neighborhoods of atom)[23] and RMNA (reacting multilevel neighborhoods of atom) descriptors and a fragment data set[24,25] based on the algorithm of the PASS program (prediction of activity spectra for substances).[26,27] In those studies, we used a vocabulary based on the transformation pattern, which describes the structural changes from a substrate to a product. The prediction accuracy of this method was reasonable, but it was applied to a few reaction classes and could not predict the SOMs. Consideration of the SOMs, but not the biotransformations of xenobiotics, allowed us to expand the scope of the training set and to avoid the generation of potential metabolites. We have created a method for the prediction of SOMs without using the transformation pattern vocabulary. It was based on the Bayesian-like algorithm used in PASS, but a special modification of MNA descriptors has been developed to specify SOMs.

We have prepared the training sets for the five isoforms of CYP P450 that metabolize the majority of xenobiotics:[2] 3A4, 2C9, 2C19, 2D6, and 1A2. A leave-one-out cross-validation (LOO CV) procedure was performed for each of the training sets to validate the quality of the prediction. For external validation of our method, we have performed SOM prediction for an evaluation set of the 57 cardiovascular drugs. We also compared the prediction results obtained using our method for the evaluation set with the prediction results of 2D6 and 2C9 regioselectivity provided by the SMARTCyp program (version 2.3)—a web application for SOMs prediction of cytochrome P450-mediated drug metabolism.[28]

MATERIALS AND METHODS

Initial Data Sets. The initial data set included 2415 biotransformations catalyzed by the cytochrome P450 isoforms 1A2, 2C9, 2C19, 2D6, and 3A4 CYPs for 1364 compounds. These data were collected from in vivo and in vitro experimental studies available in publications and also in the database Metabolite[29] (Table 1).

Table 1. Number of Reactions Catalyzed by Different Isoforms of Cytochrome P450

cytochrome P450 isoform	number of substrates	number of reactions
1A2	573	984
2C9	446	705
2C19	388	624
2D6	558	860
3A4	960	1629
total records	**2925**	**4802**
unique records in the data set	**1364**	**2415**

The number of records of reactions catalyzed by different P450 isoforms is twice as large as the number of unique reactions in the data set (4802 vs 2415). At the same time, the number of reactions is almost twice as large as the number of substrates (2415 vs 1364). This is because of overlap in the substrate specificity of P450 different isoforms and the fact that the same substance may undergo more than one biotransformation. For example, the PPAR agonist Muraglitazar is a substrate of four different cytochrome P450 isoforms: 3A4, 2C9, 2C19, and 2D6. According to the example given in Figure 1, specific reactions may differ or overlap.

SOM Designation. It is known that the interaction of the active site of cytochrome P450 with substrates leads to the formation of carbon radicals (as a result of the hydrogen atom abstraction process) or carbocations (as a result of an electron transfer process) and finally to the generation of an oxidized substrate. Further restructuring of the oxidized substrate leads to the formation of the final structure of the product.[3,30] On the basis of the structure of the final reaction product, it may be guessed which atom in the substrate molecule has been attacked, i.e., the SOM may be determined. During a designation of SOMs, all 2415 reactions in the initial set were analyzed to determine which atoms were attacked by the CYP P450 isoforms. Then, new structures with labeled atoms were created. The examples of transformations of known data on reactions into a substrate with labeled atoms representing SOM are shown in Table 2.

dx.doi.org/10.1021/ci400472j | J. Chem. Inf. Model. 2014, 54, 498−507

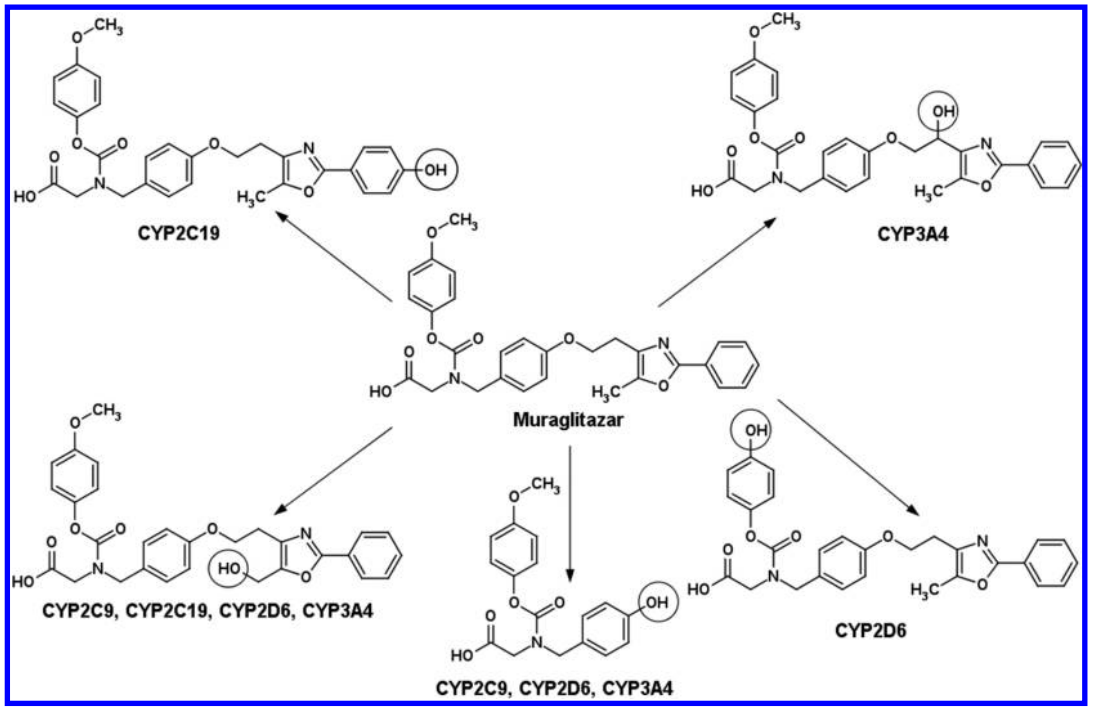


Figure 1. Known biotransformations of Muraglitazar catalyzed by different cytochrome P450 isoforms.

The Molfiles format,[31] developed by MDL (currently Accelrys) as a chemical structure file format utilized in many chemoinformatics computational tools, was used for description of the SOMs. In this format, the "Atom Block" is made up of atom lines, one line per atom describing atom coordinates, atom symbol, charge, valence, etc. There are positions in "Atom Blocks" that are not used to describe atom properties, and these were used to represent the atom's label.

Creation of Data Sets. The final data sets (training and evaluation sets) consist of structures with one labeled atom (SoLA). The number of SoLAs is equal to the number of non-hydrogen atoms in a substrate. Data sets consist of both positive and negative examples with respect to the result of interaction with a particular enzyme. If SoLA represents the chemical structure where a labeled atom is a SOM of the particular enzyme, then this SoLA is the positive example. Otherwise, the SoLA is considered a negative example. For example, 21 SoLAs were generated for amitriptyline (Figure 2). The interaction of amitriptyline with 2D6 leads to the appearance of the four metabolites that were generated after the attack of CYP 2D6 on the four atoms (nos. 1, 2, 20, and 21). So, SoLAs with labeled atoms in the appropriate positions (nos. 1, 2, 20, and 21) are considered positive examples. In Figure 2, all SoLAs represented by a circle and a number in the lower string indicate the atom number that was labeled. SoLAs from positive examples are illustrated as black circles, and SoLAs from negative examples are illustrated as white circles.

Training and Evaluation Sets. The total amounts of positive and negative examples in training sets, which were obtained from 2415 biotransformation reactions for five isoforms, are shown in Table 3.

Table 3 shows that the total number of negative examples was 10 times greater than the number of positive examples. In turn, the number of positive examples for each of the cytochrome P450 isoforms in Table 3 is less than the number of reactions in Table 1. It is observed that duplicate SoLAs are generated in cases with symmetrical atoms in the substrate's structure, but only unique SoLAs were included in the training set.

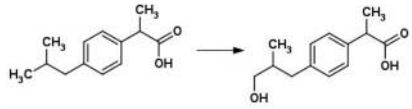
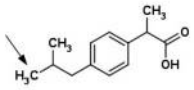
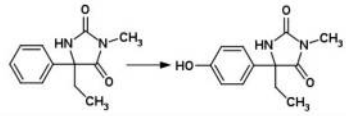
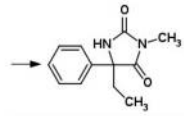
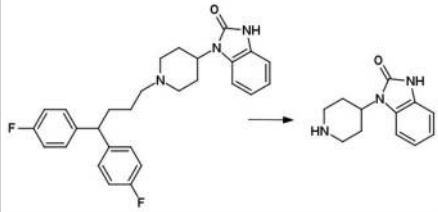
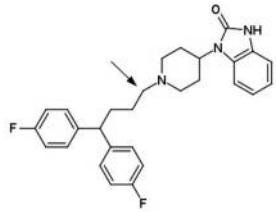
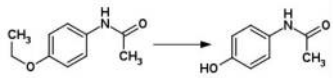
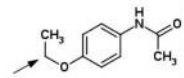
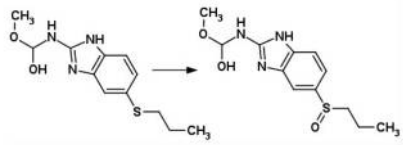
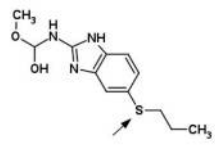
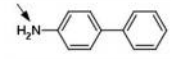
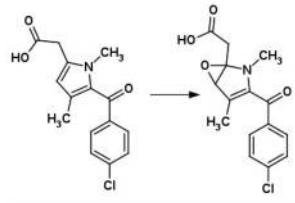
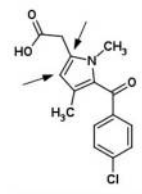
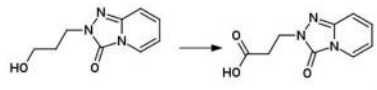
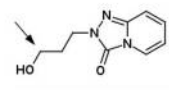
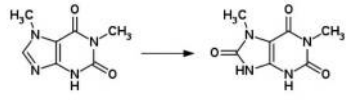
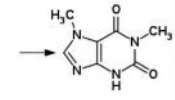
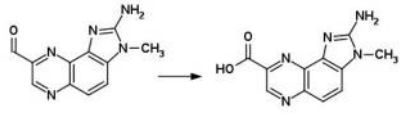
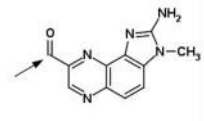
The knowledge of biotransformations is especially important for drugs administered orally because they are initially metabolized in the liver. The majority of cardiovascular drugs are used through oral administration. Thus, prediction of the SOMs for cardiovascular compounds is very important at early stages of RD of new chemical entities studied as candidates for treatment of cardiovascular disorders. Cardiovascular drugs belong to diverse therapeutic and chemical classes; therefore, they are a good representative evaluation set for verification of the new approach for SOMs prediction. So, for the creation of evaluation sets, we selected 57 cardiovascular drugs with known data on biotransformations from all drugs represented by code "C" (cardiovascular system) in the Anatomical Therapeutic Chemical Classification System.[32] The data on structures and known SOMs of drugs from evaluation sets are represented in the Supporting Information. The appropriate positive and negative SoLAs were generated for the evaluation sets. Thereby, five evaluation sets were prepared according to the five cytochrome P450 isoforms: 1A2, 2C9, 2C19, 2D6, and 3A4. ATC codes for substrates and SOMs distributions from metabolizing enzymes from the evaluation sets are shown in Table 4.

Training sets 6−10 (see Table 5) were generated on the basis of compounds that did not include compounds from the evaluation sets. A comparison of data in Tables 3 and 5 shows that the evaluation sets roughly correspond to 10% of the training sets.

Labeled Multilevel Neighborhoods of Atom Descriptors. MNA (multilevel neighborhoods of atom) descriptors,[23] which were first developed for prediction of the biological activity, were used as a basis for the creation of descriptors for SOM prediction. They are based on the molecular structure representation, which includes hydrogen atoms according to the valences and partial charges of atoms and does not specify bond types. A modification of MNA descriptors called labeled

dx.doi.org/10.1021/ci400472j | J. Chem. Inf. Model. 2014, 54, 498−507

Table 2. Examples of SOMs Labeled for Different Types of Biotransformations[a]

Name of reaction	N	Example of reaction	Substrate with labeled SOM (indicated by an arrow)
Aliphatic Hydroxylation	694		
Aromatic Hydroxylation	540		
N-Dealkylation	464		
O-Dealkylation	304		
S-Oxidation	107		
N-Oxidation	89		
Epoxidation	84		
Alcohol Oxidation	66		
Oxo group formation	19		
Aldehyde Oxidation	16		
Unclassified	32		
Total:	**2415**		

[a]N is the number of appropriate biotransformations in the data sets.

multilevel neighborhoods of atom (LMNA) descriptors was created to describe atoms attacked by the enzyme. Thus, all SoLAs in data sets are represented by the sets of LMNA descriptors.

dx.doi.org/10.1021/ci400472j | J. Chem. Inf. Model. 2014, 54, 498−507

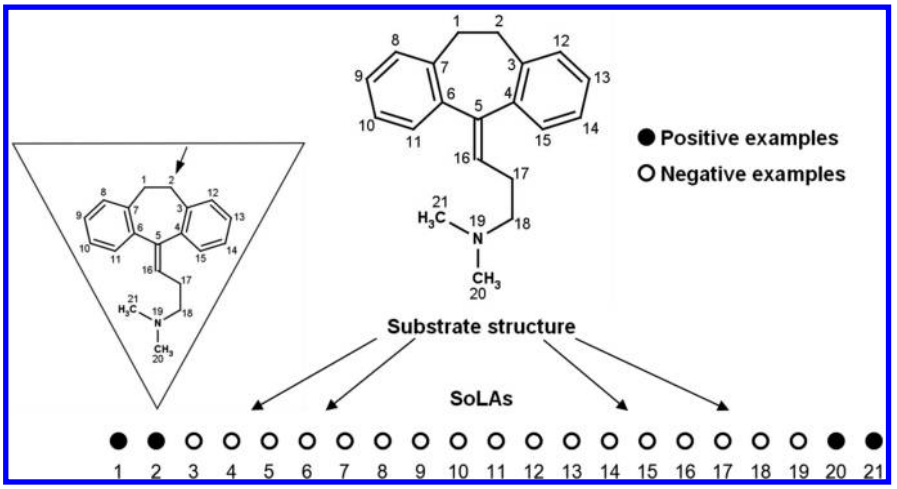


Figure 2. Schematic representation of SoLAs generated for amitriptyline (black SoLAs describe 2D6 SOMs). The number in the upper string indicates atom number, which was labeled in appropriate SoLAs.

Table 3. Characteristics of the Training Sets

data set no.	isoforms	positive examples	negative examples
1	1A2	888	9375
2	2D6	726	8366
3	2C9	600	7471
4	2C19	517	6329
5	3A4	1410	18621

The zero-level LMNA descriptor for atom **A** is the atom's symbol with mark of "been labeled in SoLA":

$$\mathbf{D_0} = [*]\mathbf{A}$$

where "*" is a mark for a labeled atom.

The *N*th level of LMNA descriptor for atom **A** is created using an iterative procedure and has the following structure:

$$D_N = [-][*]A(D_{N-1}(B_1)D_{N-1}(B_2)...D_{N-1}(B_k))$$

where "−" is a mark added to nonring atoms (added to the second level of LMNA), and $\mathbf{D}_{N-1}(\mathbf{B}_i)$ is the $(N-1)$-level LMNA descriptor for atom **A**'s *i*th immediate neighbor.

An example of LMNA descriptors for four atoms (nos. 18−21) in the SoLA, which represents amitriptyline with labeled atom no. 21, is given in Figure 3.

Table 5. Characteristics of the Training Sets That Did Not Include Compounds from the Evaluation Sets

data set no.	isoform	positive examples	negative examples
6	1A2	839	9420
7	2D6	651	8212
8	2C9	544	7284
9	2C19	486	6357
10	3A4	1302	18420

Each of the atoms in the SoLA is described using LMNA descriptors in the same way, and a set of LMNA descriptors is generated. The complete list of LMNA descriptors for SoLA, which represents amitriptyline with labeled atom no. 21, is provided in the Supporting Information. This SoLA (Figure 3) is described by 28 unique LMNA descriptors (including first and second levels of LMNA descriptors).

Algorithm of SOM Estimation and Validation of SOM Prediction. Each SoLA in a training set is described by a set of LMNA descriptors. It is considered to be a positive example if the labeled atom related with the known SOM for a particular enzyme, E_k, and a negative example in the opposite case.

Table 4. ATC Codes of Drugs and the Number of Known SOMs for the Five Major Cytochrome P450 Isoforms

isoform	number positive/ negative examples[a]	number of drugs	ATC codes of substrates
1A2	49/496	18	C01BA01; C01BA03; C01BB01; C01BB02; C01BC03; C01BD01; C02CC04; C03BD01; C03CA04; C03DA04; C03DB02; C04AD03; C07AA05; C07AG02; C08CA04; C08DA01; C09CA04; C10AA06
2C9	51/635	24	C01BA01; C01BA02; C01BA03; C01EB03; C01EB16; C02CC04; C02KX01; C03CA04; C03CC02; C03DA04; C05AA01; C07AA05; C07AG02; C08CA04; C08DB01; C08EA02; C09CA01; C09CA03; C09CA04; C09CA05; C10AA04; C10AA06; C10AA07; C10AA08
2C19	30/367	13	C01BA01; C01BD01; C01EB16; C02CC04; C03DA04; C04AX20; C07AA05; C07AA06; C07AB02; C07AG02; C08CA04; C08EA02; C08EX02
2D6	99/1154	27	C01BA02; C01BA03; C01BA04; C01BB04; C01BC03; C01BC04; C01BG07; C01EB11; C02CC04; C03CA04; C03DA04; C04AX20; C07AA05; C07AA06; C07AA15; C07AA19; C07AB02; C07AB07; C07AB12; C07AG02; C08CA04; C08DB01; C08EA02; C08EX02; C09CA04; C10AA04; C10AA06
3A4	95/1246	35	C01BA01; C01BA03; C01BB01; C01BC03; C01BD01; C01BD04; C01EB17; C01EB18; C02CC04; C02KX01; C03DA04; C05AA01; C05AA04; C05AA05; C07AA05; C07AB07; C07AG02; C08CA04; C08CA05; C08CA07; C08CA12; C08CA15; C08DA01; C08DB01; C08EA02; C08EX02; C09CA01; C09CA04; C09CA05; C10AA01; C10AA02; C10AA04; C10AA05; C10AA06; C10AA08

[a]Number of positive and negative SoLAs generated for evaluation sets according to the scheme represented on Figure 2.

dx.doi.org/10.1021/ci400472j | J. Chem. Inf. Model. 2014, 54, 498−507

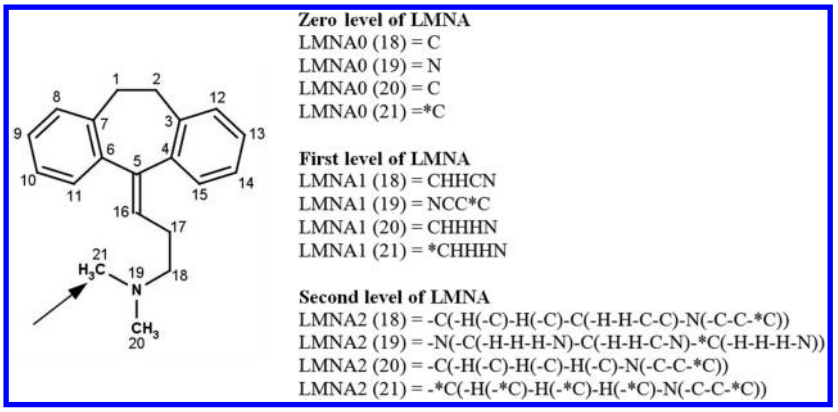


Figure 3. Example of LMNA descriptors for four atoms (nos. 18−21) in the SoLA, which represents amitriptyline with labeled atom no. 21.

On the basis of the SoLA representation using the set of m LMNA descriptors $\{D_1, D_2, ..., D_m\}$, the following values are calculated for cytochrome P450 isoforms E_k SOMs:

$$B_k = \frac{S_k - S_{0k}}{1 - S_k S_{0k}}$$

$$S_k = \sin\left[\frac{1}{m}\sum \arcsin(2P(E_k|D_i) - 1)\right]$$

$$S_{0k} = 2P(E_k) - 1$$

where $P(E_k)$ is the *a priori* probability that the labeled atom in SoLA is a SOM for the enzyme E_k, and $P(E_k|D_i)$ is the *conditional* probability that the labeled atom in SoLA is a SOM for the enzyme E_k if the descriptor D_i belongs to a set of LMNA descriptors of SoLA.

If $P(E_k|D_i) = 1$ for all descriptors of SoLA, then $B_k = 1$. If $P(E_k|D_i) = 0$ for all descriptors of SoLA, then $B_k = -1$. If there is no relationship between descriptors of SoLA and the fact that labeled atom in the SoLA is a SOM (i.e., $P(E_k) \approx P(E_k|D_i)$), then $B_k = 0$.

The simplest frequency estimations of probabilities $P(E_k)$, $P(E_k|D_i)$, are given by

$$P(E_k) = \frac{N_k}{N},\ P(E_k|D_i) = \frac{N_{ik}}{N_i}$$

where N is the total number of SoLAs in the training set, N_k is the number of SoLAs in which the labeled atom is a SOM for enzyme E_k, N_i is the number of SoLAs contained in the descriptor D_i, and N_{ik} is the number of positive SoLAs (where the labeled atom is a SOM for enzyme E_k) contained in the descriptor D_i.

During the training procedure, each SoLA is excluded from the training set, and the *B*-value is calculated for it; thus, the leave-one-out cross-validation (LOO CV) procedure is performed. Using calculated *B*-values for all SoLAs, the distribution of the *B*-values for both positive examples ($P_t(B)$) and negative examples ($P_f(B)$) are created.

During the prediction of SOMs for a new compound, the set of all possible SoLAs with the appropriate LMNA descriptors is generated for a new compound. The predictions of SOMs for new compounds were created on the basis of the prediction results of all SoLAs generated for the compound. Each SoLA relates to one appropriate SOM. The probabilities P_t and P_f are calculated for each SoLA of a new compound. P_t is the probability that labeled atom in the SoLA is the SOM of the appropriate enzyme. P_f is the probability that the labeled atom in SoLA is not the SOM of the appropriate enzyme. The ΔP value is calculated as $\Delta P = P_t - P_f$.

The prediction results are directly related to the LMNA descriptors in the training set. Some of the SoLAs of a new compound may contain LMNA descriptors that are different from the LMNA descriptors belonging to the training set. Such LMNA descriptors are not taken into account during the estimation of possible SOMs.

Invariant accuracy of prediction (IAP) criterion, similar to AUC (the area under the receiver operating characteristic curve),[33] was used to estimate the accuracy of the created method. Mathematically, IAP[26,34,35] values equal the probability that the ΔP estimation has a higher value for a randomly selected positive example (SoLAs in which labeled atom is a SOM, ΔP_+) than for a randomly selected negative example (SoLAs in which labeled atom is not a SOM, ΔP_-):

$$\text{IAP} = \text{probability}\{\Delta P_+ > \Delta P_-\}$$

IAP is calculated as

$$\text{IAP} = \frac{\text{num of}\{\Delta P_+ > \Delta P_-\}}{N_+ \cdot N_-}$$

where num of$\{\Delta P_+ > \Delta P_-\}$ is the number of cases where ΔP for positive SoLAs exceeds the ΔP value for negative SoLAs. Thus, all pairs of SoLAs from the evaluation set are compared. N_+ and N_- are the number of all positive examples and all negative examples in the set, respectively. Detailed information about IAP statistics was published earlier.[26,34]

For an estimation of the accuracy of the created method, we also applied three metrics (Top-1, Top-2, Top-3), which were used earlier by the authors of SMARTCyp and RS-Predictor.[21] The meaning of these metrics is the following: a molecule is considered correctly predicted if any experimental SOM (for each enzyme considered separately) is ranked as first (Top-1), first or second (Top-2), or first or second or third (Top-3) within the consensus SOM regioselectivity ranking. The atoms in compounds in the evaluation sets are arranged according to ΔP values. Top-1, Top-2, and Top-3 metrics may be used for the assessment of the prediction ranking results, but the results are strongly dependent on the size of the molecule. IAP criterion overcomes this drawback of Top-metrics.

RESULTS AND DISCUSSION

It was shown that the average IAP of SOM predictions for the developed method calculated for training sets by the LOO CV

dx.doi.org/10.1021/ci400472j | J. Chem. Inf. Model. 2014, 54, 498−507

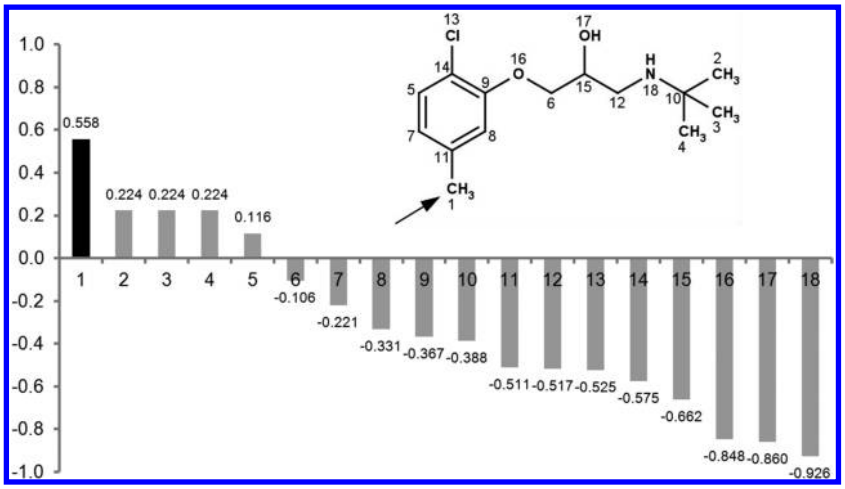


Figure 4. Predictions of the 2D6 SOMs for bupranolol. The known SOM is labeled by an arrow on the structure and it is colored black in the diagram.

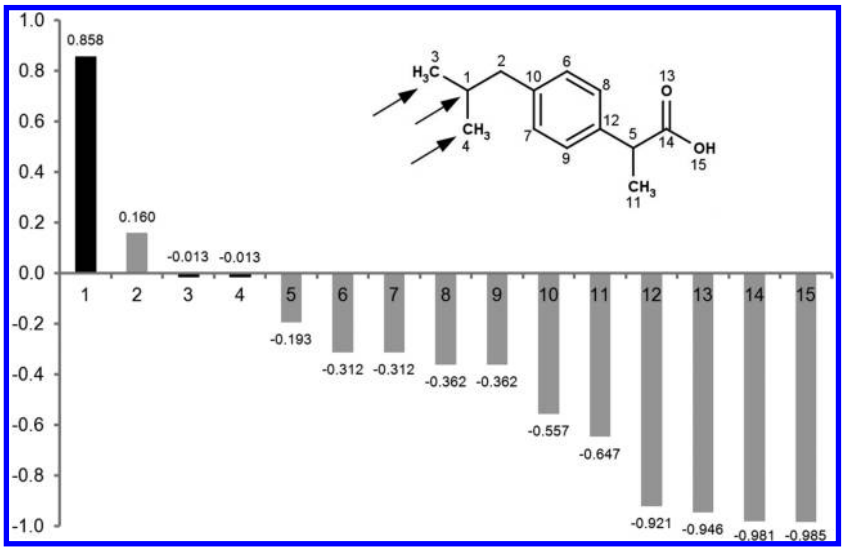


Figure 5. Predictions of CYP 2C9 SOMs for ibuprofen. The known SOMs are labeled by arrows on the structure and colored black in the diagram.

procedure was 0.89. This accuracy shows that this method is reasonable to use in the prediction of SOMs for new compounds.

For example, only one SOM is known for Bupranolol (ATC code C07AA19) if it is attacked by the 2D6 isoform (Figure 4).[36] The results of the SOM prediction for Bupranolol using data set 7 (containing data on the metabolism of CYP 2D6 substrates) are shown in Figure 4. The *X* axis represents the atom numbers in descending order of $\Delta P = P_t - P_\beta$; the known SOM is colored black. The atom numbers in the Bupranolol structure correspond to the atom numbers on the *X* axis.

On the basis of the reverse application of SOM designation rules (see the first record in Table 2), the C-hydroxylation reaction for position no. 1 is expected and the appropriate hydroxylated product will be the most probable.

Another example is ibuprofen (ATC code C01EB16). Two SOMs are known for it in the case of the CYP 2C9 isoform.[37,38] Our method does not take into account the spatial and stereochemical features of molecules, and we consider carbon atoms no. 3 and 4 (Figure 5) to be identical, so ibuprofen has the following SOMs: atoms no. 1, 3, and 4, which are indicated by arrows in Figure 5. The result in Figure 4 demonstrates an "ideal situation", where the atoms with the highest value of ΔP are the SOMs, but in reality, not all SOMs may be predicted from the highest ΔP value. For example, the results of SOM predictions for ibuprofen using data set 8 (containing data on metabolism of the CYP 2C9 substrates) are shown in Figure 5. The *X* axis represents the atom numbers in descending order of ΔP. The known SOMs are colored black. The atom numbers in the ibuprofen structure correspond to the atom numbers on the *X* axis.

In this example, the SOM for atom no. 1 was predicted to have the highest values of ΔP; however, for atoms no. 3 and 4, which are also SOMs, the values of ΔP are smaller. Negative values of ΔP for atoms no. 3 and 4 mean that the majority of LMNA descriptors for the structure containing the appropriate SOMs are found in the negative examples of the training set. The main reasons for this behavior are (1) the incompleteness of the training data and (2) errors in the training data. Nevertheless, the calculated metrics for ibuprofen in this case are Top1 = 1, Top2 = 1, Top3 = 1, IAP = 0.95.

By comparing the accuracy of SOM predictions for bupranolol (one SOM) and ibuprofen (two SOMs), it can be noted that Top-1, Top-2, and Top-3 do not properly represent

dx.doi.org/10.1021/ci400472j | J. Chem. Inf. Model. 2014, 54, 498−507

the accuracy of the SOM predictions for the ibuprofen molecule. The fact that the second real SOM for ibuprofen has the third rank is not taken into account by the Top-1, Top-2, and Top-3 calculation. The IAP represents a better estimation of the accuracy of the SOM prediction for molecules with multiple SOMs because it takes into account all of the SOMs of a molecule. Thus, the IAP metric, in contrast to the Top-metrics, takes into account the fact that not all SOMs may be predicted with the highest rank.

We compared the results of the prediction obtained using our method, SMARTCyp, and RS-WebPredictor because these methods are freely available web-services and calculate ranks for the atoms in the substrates. SOMs for the same cardiovascular compounds, which were used in the evaluation sets, were predicted by SMARTCyp (version 2.3) for 2C9 and 2D6 isoforms and RS-WebPredictor (version 1.0) for five studied CYP isoforms. We calculated the IAP value, and Top-1, Top-2, and Top-3 metrics to estimate the accuracy of SMARTCyp and RS-WebPredictor (Table 6).

Table 6. Results of Predictions for the Evaluation Sets

method	ID of training data set	isoform	Top-1	Top-2	Top-3	IAP
SMARTCyp	V2.3	2D6	0.67	0.78	0.86	0.86
	V2.3	2C9	0.63	0.75	0.92	0.91
RS-WebPredictor	1.0	1A2	0.37	0.37	0.37	*a*
	1.0	2D6	0.21	0.29	0.36	*a*
	1.0	2C9	0.17	0.21	0.25	*a*
	1.0	2C19	0.23	0.31	0.38	*a*
	1.0	3A4	0.31	0.34	0.40	*a*
full training sets						
LMNA and PASS based method	1	1A2	0.79	0.89	0.94	0.94
	2	2D6	0.76	0.93	0.96	0.92
	3	2C9	0.83	0.88	0.96	0.96
	4	2C19	0.92	1	1	0.94
	5	3A4	0.89	0.91	0.91	0.95
training sets excluding data from evaluation sets						
	6	1A2	0.47	0.68	0.68	0.83
	7	2D6	0.43	0.61	0.68	0.83
	8	2C9	0.58	0.75	0.88	0.87
	9	2C19	0.46	0.62	0.85	0.76
	10	3A4	0.69	0.77	0.8	0.85

[a]The prediction results did not allow calculating IAP.

We calculated IAP, Top-1, Top-2, and Top-3 metrics of the proposed method for the evaluation sets based on two types of training sets (without (full training sets) and with exclusion of the data from the evaluation sets). After excluding from the training sets all compounds belonging to the evaluation sets, the average accuracy of the SOM predictions decreased from 0.94, 0.84, 0.92, and 0.95 to 0.83, 0.53, 0.69, and 0.78 for IAP, Top-1, Top-2, and Top-3, respectively (Table 6). The prediction results for all compounds from the evaluation sets are shown in the Supporting Information.

Table 6 shows that the obtained prediction accuracy for SMARTCyp and our method were comparable and considerably higher than the accuracy of SOMs predictions given by RS-WebPredictor. An average IAP value for a SOM prediction given by our method and training sets no. 1−5 for all evaluation sets was 0.94, particularly the regioselectivity of 2D6 is 0.92 (the result of SMARTCyp is 0.86), and the regioselectivity of 2C9 is 0.96 (the result of SMARTCyp is 0.91). When structures from the evaluation set were excluded from the training set (it is approximately 10% information from experimental SOMs), the average IAP of SOM predictions for five isoforms decreased to 0.83. This decrease in accuracy may be explained by the reduced number of original LMNA descriptors in the training sets after the deletion of data related with evaluation sets. For SMARTCyp, it is analogous to excluding a part of fragment-based energy rules, which are used to calculate the SMARTCyp prediction rank.[20] It is known that some of the compounds from the 2D6 evaluation sets are used to create the prediction model of SMARTCyp.[39,40]

SMARTCyp, RS-WebPredictor, and our method predict the site of metabolism directly from the 2D structure of a molecule on the basis of precalculated models. The SMARTCyp model for the prediction of CYP 3A4 SOMs uses a sophisticated algorithm, which is based on the calculation of a reactivity descriptor and an accessibility descriptor. The reactivity descriptor (an estimation of the energy required for CYP to react at this position) is calculated for each atom by matching SMARTS patterns to a lookup table of energies in kilojoules per mole. For creating an atom reactivity library, it is necessary to calculate the transition state energies using density functional theory, group calculations by fragments, and average energies.[20] The SMARTCyp model for prediction of the SOMs of CYP 2D6 and CYP 2C9 requires additional descriptors: the distance from the end of the molecule and the largest distance to a protonated nitrogen atom.[28,41]

RS-WebPredictor is a freely available service for SOM predictions of nine CYP isoforms which is based on the RS-Predictor algorithm.[42] It uses isozyme-specfic SOM prediction models from sets of known substrates and metabolites.[21,22] The SMARTCyp-derived reactivity descriptor and 148 topological descriptors are used for SOM description. MIRank (multiple-instance ranking—a generalization of support vector machines) is used to optimize the ranking of observed SOMs over nonobserved SOMs.[22]

In contrast to SMARTCyp and RS-Predictor, our method uses models created only on the basis of 2D structures with labeled SOMs without energy estimation and does not take into account any information about the binding cavity of the enzyme. The proposed method can be used for any type of enzyme or reaction in which SOMs are described by the set of reactions. It requires only information about experimentally observed SOMs for the creation of "structure−SOM" relationships. Therefore, to improve the quality of the prediction models created by our method, it is necessary to increase only the quality of the training sets by including compounds from different chemical classes.

Prediction of SOMs can significantly facilitate the analysis of data related to the metabolism of xenobiotics at early RD stages of drug development. This method opens the way for predicting toxic metabolites and helping to develop pro-drugs.

CONCLUSIONS

Cytochromes P450 are the main human biotransformation enzymes; therefore, computational predictions of interactions with cytochromes P450 can increase the efficiency and decrease the cost and time of drug development. We have created a method for the prediction of the sites of metabolism, which uses only structural formulas of xenobiotics. It is based on the a set of enzyme substrates, which are converted to SoLAs (structure with one labeled atom) and the PASS algorithm.

dx.doi.org/10.1021/ci400472j | J. Chem. Inf. Model. 2014, 54, 498−507

Currently, the proposed method was applied to five P450 isoforms: 1A2, 2C9, 2C19, 2D6, and 3A4, which metabolize the majority of xenobiotics, but it may also be used for any type of substrate transformation if the appropriate training set exists.

Our study shows that IAP metrics are more appropriate for the evaluation of the prediction accuracy for molecules with multiple SOMs because it takes into account the results from predictions for all potential SOMs of molecules. The proposed algorithm is very fast (taking approximately 50 ms for one molecule on a modern PC with a 2.6 GHz processor and Windows OS). It was shown on external test sets that it has an accuracy of SOM prediction higher than RS-Predictor models for CYP 1A2, 2D6, 2C9, 2C19, and 3A4 and comparable to or better than SMARTCyp models for the prediction of SOMs for CYP 2C9 and 2D6.

■ ASSOCIATED CONTENT

Supporting Information

Complete list of LMNA descriptors for SoLA representing amitriptyline with labeled atom no. 21, prediction results for CYP1A2, prediction results for CYP2C9, prediction results for CYP2C19, prediction results for CYP2D6, prediction results for CYP3A4. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rudik_anastassia@mail.ru.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The study was partially supported by RFBR grant 12-04-31670 and by AstraZeneca Award “Vanguard of Knowledge”.

■ ABBREVIATIONS

PASS, prediction of activity spectra for substances; MNA, multilevel neighborhoods of atom; LMNA, labeled multilevel neighborhoods of atom; RMNA, reacting multilevel neighborhoods of atom; SOM, site of metabolism; SoLA, structure with one labeled atom; LOO CV, leave-one-out cross-validation; IAP, invariant accuracy of prediction; AUC, area under the ROC curve; PPAR, peroxisome proliferator-activated receptor

■ REFERENCES

(1) Lewis, D. F.; Ito, Y. Human cytochromes P450 in the metabolism of drugs: new molecular models of enzyme-substrate interactions. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4*, 1181−1186.

(2) Williams, J. A.; Hyland, R.; Jones, B. C.; Smith, D. A.; Hurst, S.; Goosen, T. C.; Peterkin, V.; Koup, J. R.; Ball, S. E. Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab. Dispos.* **2004**, *32*, 1201−1208.

(3) Guengerich, F. P. Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.* **2001**, *14*, 611−650.

(4) Isin, E. M.; Guengerich, F. P. Complex reactions catalyzed by cytochrome P450 enzymes. *Biochim. Biophys. Acta* **2007**, *1770*, 314−329.

(5) Brown, C. M.; Reisfeld, B.; Mayeno, A. N. Cytochromes P450: a structure-based summary of biotransformations using representative substrates. *Drug. Metab. Rev.* **2008**, *40*, 1−100.

(6) Hsiao, Y. W.; Petersson, C.; Svensson, M. A.; Norinder, U. A pragmatic approach using first-principle methods to address site of metabolism with implications for reactive metabolite formation. *J. Chem. Inf. Model.* **2012**, *52*, 686−695.

(7) Sato, K.; Yamazoe, Y. Unimolecular and bimolecular binding system for the prediction of CYP2D6-mediated metabolism. *Drug. Metab. Dispos.* **2012**, *40*, 486−496.

(8) Zheng, M.; Luo, X.; Shen, Q.; Wang, Y.; Du, Y.; Zhu, W.; Jiang, H. Site of metabolism prediction for six biotransformations mediated by cytochromes P450. *Bioinformatics* **2009**, *25*, 1251−1258.

(9) Sykes, M. J.; McKinnon, R. A.; Miners, J. O. Prediction of metabolism by cytochrome P450 2C9: alignment and docking studies of a validated database of substrates. *J. Med. Chem.* **2008**, *51*, 780−791.

(10) Tarcsay, A.; Kiss, R.; Keseru, G. M. Site of metabolism prediction on cytochrome P450 2C9: a knowledge-based docking approach. *J. Comput. Aided. Mol. Des.* **2010**, *24*, 399−408.

(11) Santos, R.; Hritz, J.; Oostenbrink, C. Role of water in molecular docking simulations of cytochrome P450 2D6. *J. Chem. Inf. Model.* **2010**, *50*, 146−154.

(12) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.* **2012**, *52*, 617−648.

(13) Tarcsay, Á.; Keseru, G. M. In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin. Drug Metab. Toxicol.* **2011**, *7*, 299−312.

(14) Kuncheva, L. I. *Combining pattern classifiers: methods and algorithms*; J. Wiley: Hoboken, NJ, 2004.

(15) Oh, W. S.; Kim, D. N.; Jung, J.; Cho, K. H.; No, K. T. New combined model for the prediction of regioselectivity in cytochrome P450/3A4 mediated metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 591−601.

(16) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed N-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062−4070.

(17) *StarDrop*, version 5.0; Optibrium: Cambridge, U.K., 2011.

(18) Hasegawa, K.; Koyama, M.; Funatsu, K. Quantitative prediction of regioselectivity toward cytochrome P450/3A4 using machine learning approaches. *Mol. Inf.* **2010**, *29*, 243−249.

(19) Mu, F.; Unkefer, C. J.; Unkefer, P. J.; Hlavacek, W. S. Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. *Bioinformatics* **2011**, *27*, 1537−1545.

(20) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett.* **2010**, *1*, 96−100.

(21) Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T. W.; Bennett, K. P.; Breneman, C. M. RS-Predictor: A new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4. *J. Chem. Inf. Model.* **2011**, *51*, 1667−1689.

(22) Zaretzki, J.; Rydberg, P.; Bergeron, C.; Bennett, K. P.; Olsen, L.; Breneman, C. M. RS-Predictor Models Augmented with SMARTCyp Reactivities: Robust Metabolic Regioselectivity Predictions for Nine CYP Isozymes. *J. Chem. Inf. Model.* **2012**, *52*, 1637−1659.

(23) Filimonov, D.; Poroikov, V.; Borodina, Yu.; Gloriozova, T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666−670.

(24) Borodina, Yu. V.; Rudik, A. V.; Filimonov, D. A.; Kharchevnikova, N. V.; Dmitriev, A. V.; Blinova, V. G.; Poroikov, V. V. A New Statistical Approach to Predicting Aromatic Hydroxylation Sites. Comparison with Model-Based Approaches. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1998−2009.

(25) Borodina, Yu. V.; Sadym, A. V.; Filimonov, D. A.; Blinova, V. G.; Dmitriev, A. V.; Poroikov, V. V. Predicting Biotransformation Potential from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1636−1646.

(26) Filimonov, D. A.; Poroikov, V. V. In *Chemoinformatics Approaches to Virtual Screening*; Varnek, A., Tropsha, A., Ed.; RSC Publishing: Cambridge (UK), 2008; pp 182−216.

(27) Lagunin, A.; Filimonov, D.; Poroikov, V. Multi-targeted natural products evaluation based on biological activity prediction with PASS. *Curr. Pharm. Des.* **2010**, *16*, 1703−1717.
(28) Rydberg, P.; Gloriam, D.; Olsen, L. The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* **2010**, *26*, 2988−2989.
(29) Accelrys Metabolite. http://accelrys.com/products/databases/bioactivity/metabolite.html (accessed Aug 1, 2013).
(30) *Organic Chemistry of Enzyme-Catalyzed Reactions*, revised 2nd ed.; Silverman, R., Ed.; Academic Press: UK, 2002.
(31) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(32) The Anatomical Therapeutic Chemical (ATC) classification. http://www.whocc.no/ (accessed Aug 1, 2013).
(33) Swets, J. Measuring the accuracy of diagnostic systems. Review. *Science* **1988**, *240*, 1285−1293.
(34) Poroikov, V. V.; Filimonov, D. A.; Borodina, Yu. V.; Lagunin, A. A.; Kos, A. Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349−1355.
(35) Sobolev, B. N.; Filimonov, D. A.; Lagunin, A. A.; Zakharov, A. V.; Koborova, O. N.; Kel, A.; Poroikov, V. V. Functional classification of proteins based on projection of amino acid sequences: application for prediction of protein kinase substrates. *BMC Bioinformatics* **2010**, *11*, 313.
(36) Appanna, G.; Tang, B. K.; Mller, R.; Kalow, W. A sensitive method for determination of cytochrome P4502D6 activity in vitro using bupranolol as substrate. *Drug. Metab. Dispos.* **1996**, *24*, 303−306.
(37) Hamman, M. A.; Thompson, G. A.; Hall, S. D. Regioselective and stereoselective metabolism of ibuprofen by human cytochrome P450 2C. *Biochem. Pharmacol.* **1997**, *54*, 33−41.
(38) Lonsdale, R.; Houghton, K. T.; Żurek, J.; Bathelt, C. M.; Foloppe, N.; de Groot, M. J.; Harvey, J. N.; Mulholland, A. J. Quantum mechanics/molecular mechanics modeling of regioselectivity of drug metabolism in cytochrome P450 2C9. *J. Am. Chem. Soc.* **2013**, *135*, 8001−8015.
(39) Moors, S. L.; Vos, A. M.; Cummings, M. D.; Van Vlijmen, H.; Ceulemans, A. Structure-based site of metabolism prediction for cytochrome P450 2D6. *J. Med. Chem.* **2011**, *54*, 6098−6105.
(40) Rydberg, P.; Olsen, L. Ligand-based site of metabolism prediction for cytochrome P450 2D6. *ACS Med. Chem. Lett.* **2011**, *3*, 69−73.
(41) Rydberg, P.; Olsen, L. Predicting drug metabolism by cytochrome P450 2C9: comparison with the 2D6 and 3A4 isoforms. *ChemMedChem.* **2012**, *7*, 1202−1209.
(42) Zaretzki, J.; Bergeron, C.; Huang, T. W.; Rydberg, P.; Swamidass, S. J.; Breneman, C. M. RS-WebPredictor: a server for predicting CYP-mediated sites of metabolism on drug-like molecules. *Bioinformatics* **2013**, *29*, 497−498.