

# DR-Predictor: Incorporating Flexible Docking with Specialized Electronic Reactivity and Machine Learning Techniques to Predict CYP-Mediated Sites of Metabolism

Tao-wei Huang,<sup>†</sup> Jed Zaretzki,<sup>‡</sup> Charles Bergeron,<sup>⊥</sup> Kristin P. Bennett,<sup>§</sup> and Curt M. Breneman\*,<sup>†</sup>

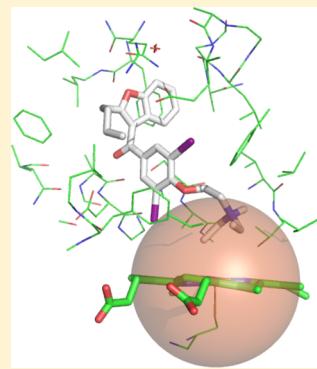
<sup>†</sup>Department of Chemistry and Chemical Biology and <sup>§</sup>Department of Mathematics, Rensselaer Polytechnic Institute, Troy, New York 12180, United States

<sup>‡</sup>Laboratory and Genomic Medicine Division, Washington University, St. Louis, Missouri 63130, United States

<sup>⊥</sup>Albany College of Pharmacy and Health Sciences, Union University, Albany, New York 12208, United States

## Supporting Information

**ABSTRACT:** Computational methods that can identify CYP-mediated sites of metabolism (SOMs) of drug-like compounds have become required tools for early stage lead optimization. In recent years, methods that combine CYP binding site features with CYP/ligand binding information have been sought in order to increase the prediction accuracy of such hybrid models over those that use only one representation. Two challenges that any hybrid ligand/structure-based method must overcome are (1) identification of the best binding pose for a specific ligand with a given CYP and (2) appropriately incorporating the results of docking with ligand reactivity. To address these challenges we have created Docking-Regioselectivity-Predictor (DR-Predictor)—a method that incorporates flexible docking-derived information with specialized electronic reactivity and multiple-instance-learning methods to predict CYP-mediated SOMs. In this study, the hybrid ligand-structure-based DR-Predictor method was tested on substrate sets for CYP 1A2 and CYP 2A6. For these data, the DR-Predictor model was found to identify the experimentally observed SOM within the top two predicted rank-positions for 86% of the 261 1A2 substrates and 83% of the 100 2A6 substrates. Given the accuracy and extendibility of the DR-Predictor method, we anticipate that it will further facilitate the prediction of CYP metabolism liabilities and aid in in-silico ADMET assessment of novel structures.



## INTRODUCTION

Cytochrome P450s (CYPs) are a superfamily of enzymes accounting for a majority of phase I drug metabolism and activation. Nearly 90% of drugs in use and clinical trials are metabolized by CYPs.<sup>1</sup> The bioavailability, pharmacokinetics, and toxicity of most drugs are greatly affected by their interaction with the CYP enzymes, especially during first-pass metabolism. Thus, prediction of likely drug metabolism pathways for each specific CYP would facilitate the efforts of medicinal chemists to optimize lead compounds in the early stages of drug discovery workflow.

Previously reported CYP site of metabolism (SOM) prediction models can be classified into two major types: ligand-based and structure-based methods. Ligand-based approaches, which include rule-based methods, molecular alignment techniques, and carefully validated qualitative structure–activity relationship (QSAR) models, can obtain quite impressive results.<sup>2–7</sup> These ligand-based models are predicated on the assumption that the local electronic environment of a given atom, namely the reactivity, largely determines the regioselectivity of reaction. The reactivity can be calculated using a variety of techniques: Rydberg et al. implemented SMARTCyp as a fast way to estimate the activation energy of hydrogen abstraction from any given ligand

atom by modeling CYP-mediated fragment oxidation with density functional theory (DFT) calculations and recording the transition state barrier for each fragment in a lookup table.<sup>2,8</sup> Alternatively, the local substrate reactivity can be approximated through the use of semiempirical Fukui indices, atomic charges, and topological descriptors.<sup>6–9</sup> For most of the above models, descriptors from various sources are correlated with experimental outcomes by using machine learning algorithms to produce models. Particularly, in RS-Predictor,<sup>6,7</sup> Zaretzki et al. employs an algorithm known as MIRank to address the problem of capturing intramolecular reactivity trends for modeling CYP regioselectivity across greatly varying chemotypes. Within this method, chemically related atoms are grouped together as a *metabolophore* and models are trained using MIRank to partially rank *metabolophores* within the same molecule. Through this approach, the propensity of atoms to be metabolized by a given CYP can be estimated based on a comparison of atoms within a given molecule, informed by the trends identified in all molecules in the training data set. Based descriptors are largely from RS-Predictor.<sup>49</sup> Since ligand properties can represent controlling aspects of ligand/CYP

Received: August 8, 2013

Published: November 5, 2013

interactions, implicit modeling has been adopted in several reactivity-based methods. For instance, the interaction between the ligand and CYP is implicitly represented by surface descriptors<sup>6,10</sup> as well as the topological position of each potential site on the substrate molecules.<sup>5–7</sup>

However effective, explicit modeling of CYP–ligand interactions are not included in ligand-based methods such as RS-Predictor. Such information can be important, especially when the active site of the CYP is relatively inflexible and preferred ligand poses play a vital role on the regioselectivity of the metabolic oxidation reaction. Additionally, ligand-based methods are unable to provide descriptions of the enzyme–substrate interactions underlying the catalytic action of a given CYP. The spatial orientation of the ligand in the active site of the CYP can be valuable for providing guidance for modifying the structure of small molecules to either enhance or reduce metabolic activity at specific sites.

Structure-based methods that incorporate docking can provide explicit binding modes that identify the accessibility of potential sites to the metabolizing heme. There are many challenges associated with predicting CYP-binding modes, which is why solely structure-based SOM prediction models are historically less accurate than ligand-based methods.<sup>11–13</sup> These difficulties may be attributed to a variety of effects, including the strong influence of ion-chelation, the high flexibility of CYP active sites, as well as cooperative binding and allosteric effects.<sup>14,15</sup> The effect of iron-chelation, which is stronger than typical intermolecular interactions, needs to be considered for developing and scoring reasonable poses, so specific scoring functions have been developed to address that need.<sup>16,17</sup> One of the reasons that CYP are promiscuous metabolizers is the high degree of flexibility around the active site, which is supported by the observed differences between crystal structures of the same CYP isozyme<sup>18,19</sup> and corroborated by the results of flexibility analysis from molecular dynamics.<sup>20</sup> Several groups<sup>21,22</sup> report flexible docking of ligands to CYP 2D6 by extracting multiple conformations from molecular dynamics simulations between the enzyme and a number of representative compounds. Their results, however, suggest that the choice of conformation for new substrates still needs exploration to obtain more accurate results. The inclusion of water molecules in protein binding sites can also have important ramifications in docking poses of substrates in the CYP active sites. Santos et al. have also reported some exploratory work on the effect of water and finds that the role of water depends on the conformation of CYP protein.<sup>23</sup> There is ambiguity as to whether or not water should be included when performing docking calculations between CYP and ligand.

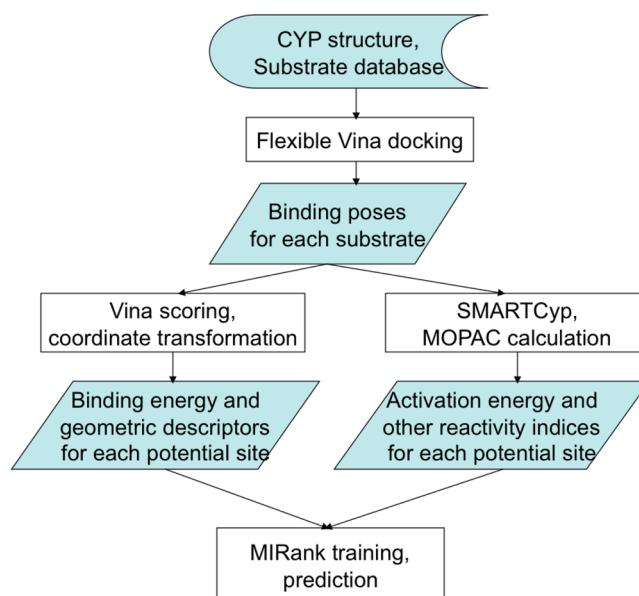
While the problems described above are important, the critical barrier to making good structure-based SOM prediction models is finding ways of extracting and integrating pose-specific information into them in a chemically meaningful way. Most of the previous studies using docking only utilized binding energies from the docked pose with lowest binding energy<sup>11,12,24,25</sup> which was then combined with reactivity scores to rank the atoms. But other factors—such as the 3D position of each putative SOM near the heme as well as the orientation of the entire substrate—may affect the formation of the actual CYP metabolite during the metabolism process. An implicit assumption in the previous studies is that substrate oxidation occurs when it is in the most energetically favorable binding mode, which is defined as the bound pose with the lowest

binding energy. In reality, the reaction can occur from a less binding favorable pose, and this situation needs to be considered when modeling CYP regioselectivity.

In this work, we address these limitations through the incorporation of flexible docking-derived descriptors into the new Docking-Regioselectivity-Predictor (DR-Predictor) algorithm. During this study, DR-Predictor was applied to isozyme-specific substrate sets of 261 CYP 1A2 substrates and 100 CYP 2A6 substrates. Among the most important CYP isoforms for drug metabolism, CYP 1A2 and CYP 2A6 have relatively rigid backbones compared to other CYP isozymes such as CYP 3A4. Because publicly available high-quality crystal structures are available, these CYPs are quite suitable for our exploratory integration of docking with reactivity models and MIRank for regioselectivity prediction. The trained DR-Predictor model was found to identify the experimentally observed SOMs for 86% of the substrates of 1A2 and 83% of the substrates of 2A6 within the top two cross-validated predictions. We also developed a new metric of averaged Spearman to evaluate models in terms of sites instead of molecules and have applied it to compare our models with the results of other methods. To interpret models, we identify the relative importance of structure and ligand-based descriptors and make chemical insights into theoretical and experimental mechanisms of CYP metabolism.

## METHOD

**Framework.** A broad overview of the DR-Predictor modeling framework is illustrated in Figure 1. In a typical run, Autodock Vina was applied to each candidate substrate using a CYP isozyme crystal structure to generate multiple binding poses both with and without water present in the active site. Conformation specific descriptors were then calculated for each putative SOM. We formulated 23 atomic descriptors that



**Figure 1.** Flowchart for DR-Predictor SOM modeling combining Autodock Vina docking and reactivity descriptors with MIRank modeling. On the basis of the docked poses, descriptors for putative SOMs were calculated. All descriptors except SMARTCyp are dependent on the 3D conformation from Vina output, as SMARTCyp is based only on the 2D structure of substrate.

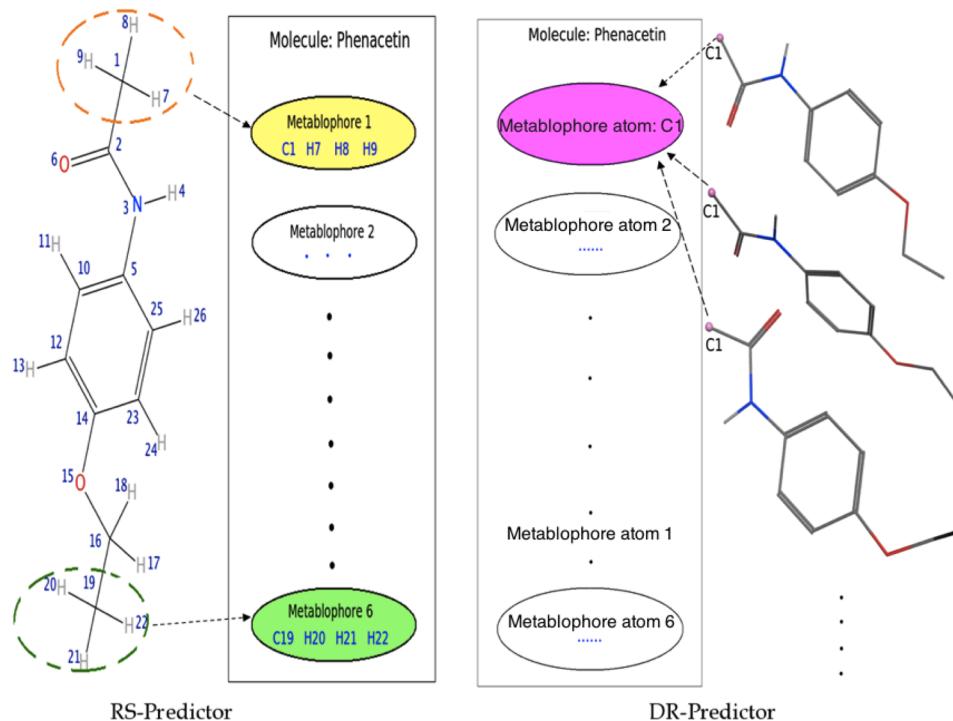
**Table 1.** Definitions for 23 Descriptors Based on Docking and 31 Descriptors of Chemical Reactivity in Our Models

category	descriptor ID	description
Autodock Vina docking descriptors	bindE	binding energy between flexible CYP and test ligand pose
	bindE_rigid	binding energy between rigid CYP without water and test ligand pose
	bindE_flex_wt	binding energy between flexible CYP with water and test ligand pose
	bindE_heme	binding energy between heme and test ligand pose
	bindE_resi	binding energy between specified residues (5) and test ligand pose
	atm_bindE	binding energy between flexible CYP and reactive atom
	atm_bindE_rigid	binding energy between rigid CYP without water and reactive atom
	atm_bindE_heme	binding energy between heme and reactive atom
	atm_bindE_resi	binding energy between specified residues (5) and reactive atom
	gauss_2	Gauss component 2 from Vina scoring in flexible docking
	repulsion	repulsion term from Vina scoring in flexible docking
	hydrophobic	hydrophobic interaction term from Vina scoring in flexible docking
	h_bond	hydrogen bond term from Vina scoring in flexible docking
	strDev_resi	geometric difference in rmsd between initial structure and final structure for flexible residues (4)
	dist	distance between iron in the heme and the reactive atom
	theta	cosine of angle between reactive atom-iron line and axis perpendicular to the heme plane (Figure 4)
	phi	cosine of angle between reactive atom-iron line and axis within the heme plane (Figure 4)
	dist_centerMol	distance between iron in the heme and the center of the molecule with reactive atom
	theta_centerMol	cosine of angle between center-iron line and axis perpendicular to the heme plane (Figure 4)
	phi_centerMol	cosine of angle between center-iron line and axis within the heme plane (Figure 4)
SMARTCyp	actE	activation energy calculated by SMARTCyp
	charge	atomic charge based on AM1
	density	atomic electron density
	electr	Fukui atomic electrophilicity index
	nucleo	Fukui atomic nucleophilicity index
	fukui	Fukui one-electron reactivity index
	cosmo_charge	atomic charge calculated with COSMO method
	area	surface area for the base atom
	sum_HeleF	sum of Fukui electrophilicity of hydrogens next to the base atom
	sum_Hchg	maximum Fukui electrophilicity of hydrogens next to the base atom
	max_Hchg	maximum AM1 charges of hydrogens next to the base atom
	min_Hfk	minimum Fukui index of hydrogens next to the base atom
	sum_Hchg_cosmo	sum of COSMO charges of hydrogens next to the base atom
	max_Hchg_cosmo	maximum of COSMO charges of hydrogens next to the base atom
	max_Harea	maximum of surface area of hydrogens next to the base atom
	sum_Harea	sum of surface area of hydrogens next to the base atom
	max_nb_nucF	maximum nucleophilicity index of heavy atoms next to the base atom
	min_nb_chg	minimum charge of heavy atoms next to the base atom
	max_nb_fuk	maximum Fukui reactivity index of heavy atoms next to the base atom
	max_nb_area	maximum surface area of heavy atoms next to the base atom
	min_NbChg_cosmo	minimum COSMO charge of heavy atoms next to the base atom

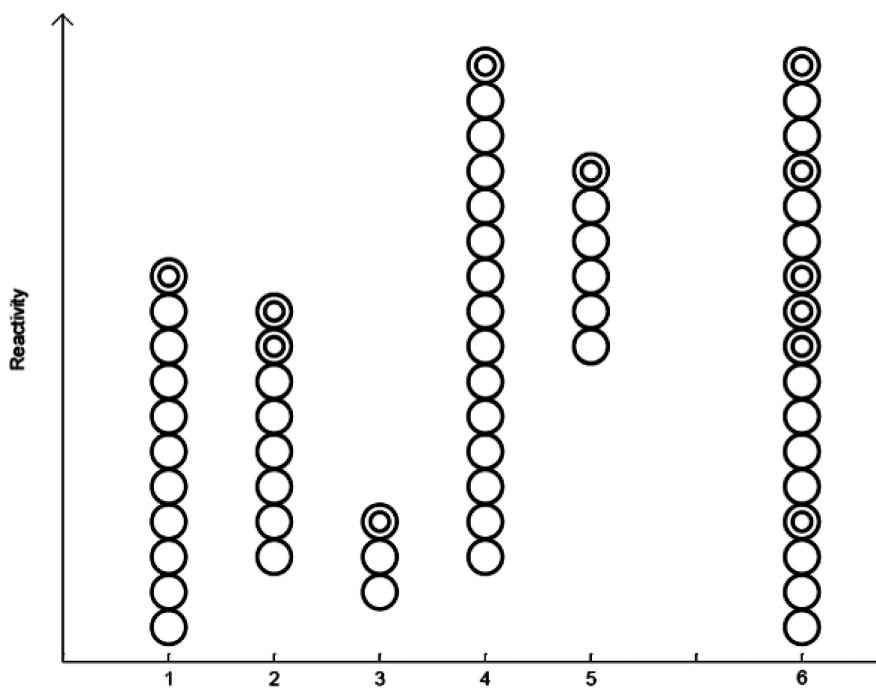
semiempirical atomic reactivity descriptor based on MOPAC calculation

characterize various types of interactions with the enzyme, as well as represent the geometrical position of the molecule within the active site. We also calculated 31 reactivity descriptors including interpretable AM1-derived descriptors and highly informative SMARTCyp activation energies for each potential SOM site. These two sets of descriptors are described in Table 1. On the basis of these descriptors and specific structure of our problem (substrate, conformation, and atoms), we apply MIRank with 10 repeats of a 10-fold cross-validation (CV) schema. Ten cross-validated SOM rankings were obtained for each compound and merged into single consensus ranking using techniques of rank-aggregation described in our previous work.<sup>6,7</sup>

The hybrid structure-ligand DR-Predictor method is an extension of our previously reported ligand-based algorithm RS-Predictor.<sup>6,7</sup> Both methods treat individual molecules as competitions between putative SOMs, with each SOM being represented as a metabolophore: a collection of atoms encoded by several chemical descriptors. The main difference between DR- and RS-Predictor is in how these metabolophores are constructed. The first point of difference in metabolophores between the two methods is the composition of the descriptor sets used: RS-Predictor depends on local and topological descriptors to *implicitly* describe enzyme–substrate interactions, while DR-Predictor *explicitly* incorporates docking-based pose descriptors from Vina. The second difference in metabolophore construction is illustrated in Figure 2; an RS-



**Figure 2.** Illustration of difference in metabolophore representation in DR-Predictor relative to RS-Predictor. In RS-Predictor (left), due to uncertainties as to which atom is oxidized by the CYP heme, atoms C1, H7, H8, and H9 of Phenacetin are grouped into metabolophore 1, which is considered a potential site of metabolism of that compound. Similarly, atoms C19, H20, H21, and H22 are grouped as another potential SOM. In DR-Predictor (right), metabolophore atom C1 from different docking conformations of Phenacetin are grouped together as a unit as a potential SOM within Phenacetin.



**Figure 3.** Illustration of the fate of different metabolophores in five different molecules, with the sixth column representing a superimposed set. In each case, the double circle highlights metabolophores that undergo oxidative degradation. The sixth column shows that the characteristics of metabolophores cannot be correctly ranked together in an overall sense.

Predictor metabolophore is constructed as a hierarchical cluster of all chemically related atoms associated with a specific CYP-mediated reaction. Within that paradigm, a candidate SOM is represented by topologically equivalent heavy atoms, their

attached hydrogens, and certain other heavy atoms.<sup>6</sup> A DR-Predictor metabolophore is a set of descriptor vectors of a potential site in various 3D poses obtained from docking. Each metabolophore instance describes a potential site of reaction in

a specific conformation close enough to the heme to be oxidized. During training, metabolophores are labeled as either *experimental sites* or *nonexperimental sites* within each compound, depending on whether metabolism is observed at that site. Metabolophores serve as the basic modeling unit of MIRank, a specialized version of the SVM ranking algorithm whose error function was designed to optimize the rankings of oxidized metabolophores over nonoxidized metabolophores on a substrate-by-substrate basis.

On the basis of these three levels of structure in regioselectivity prediction data sets—atoms, metabolophores, and compounds—MIRank performs partial ranking of metabolophores within each compound.<sup>26,27</sup> This contrasts with other approaches<sup>5</sup> that treat regioselectivity as a classification problem while measuring performance using a metric that suggests rank-ordering of the sites (e.g., top-2). Figure 3 exemplifies this concept. Columns 1–5 represent different compounds, and circles represent potential sites of metabolism. The goal of the model should be, for each compound, to identify an experimental site of metabolism (i.e., a site of high reactivity, represented by double-circles) and not to find a classification line that separates experimental sites from unreactive nonexperimental sites. Column 6, a superposition of Columns 1–5, demonstrates the difficulty in classifying experimental and nonexperimental sites when molecules possessing a wide range of reactivities coexist in a data set.

A more subtle difference in metabolophore definition between DR- and RS-Predictor is the conformation used to calculate descriptors for each composite atom. In DR-Predictor, we use conformations explicitly derived from docking-generated poses. In DR-Predictor, a set of descriptors for each potential site is calculated with various docking conformations, as shown in Figure 2. In RS-Predictor, a single minimum energy conformation of molecule is used to calculate all descriptors for atoms in that molecule. Zaretzki et al. tried to use ligand-based stochastic searching to generate multiple conformations for MOPAC reactivity descriptor calculations.<sup>6</sup> The descriptors were averaged according to Boltzmann's distribution law, and standard MIRank modeling was performed. The results showed that there was no benefit derived by using multiple conformations over a single minimum energy conformation. A likely reason for this is that stochastic conformation generation has no enzyme structure information and, therefore, introduces noise into the descriptor set. The way DR-Predictor utilizes multiple instance aspects of MIRank allows it to calculate electronic information for each pose, thereby improving accuracy.

**Data Set.** The collection of CYP 1A2 and CYP 2A6 substrates and metabolites were assembled as part of our prior work.<sup>7</sup> Structures of these molecules were generated using the 2009 version of the Molecular Operating Environment (MOE) program<sup>28</sup> and compared with corresponding structures in PubChem.<sup>29</sup> For each molecule, experimental sites were annotated based on the experimental data given in each original reference, consistent with the data format used previously.<sup>5–7</sup> For DR-Predictor modeling, topologically equivalent heavy atoms are always grouped into the same metabolophore, so if one of them is labeled as an observed site of oxidation, all of them are. The molecules whose potential sites are all experimental sites and molecules where the SOM is ambiguous (i.e., somewhere on the benzene ring) are removed. This decision was made because high quality data is preferred,

and the size of the resulting data sets is large enough to support good statistical modeling. On the basis of these criteria, 10 substrates of CYP 1A2 and 5 substrates of CYP 2A6 were removed from consideration,<sup>7</sup> and the final data sets are composed of 261 substrates for CYP 1A2 and 100 molecules for 2A6. Substrate structures and metabolites are provided in the Supporting Information.

**Validation of Autodock Vina.** Autodock Vina is open source software used primarily for docking calculations between a given receptor and a set of potential ligands. The accuracy and speed of Vina is superior to those of prior versions,<sup>30</sup> and its speed can be further increased by using multicore computers. In our work, Autodock Vina 1.1.2 was used for all docking calculation. To ascertain whether Vina was suitable for gauging CYP binding modes, we perform redocking experiments using crystal structures of known CYP–ligand complexes. In redocking experiments, we extract ligand from binding complexes of crystal structures and perform docking between that ligand and active site. In our experiments, only crystal structures of human CYP isoforms with the ligand bound in the active site were used.<sup>31</sup> Their PDB IDs of CYP isoforms used were: 2HI4 (1A2), 2FDV (2A6), 2FDU (2A6), 2FDY (2A6), 2FDW (2A6), 1Z11 (2A6), 2NNJ (2C8), 1R9O (2C9), 2V0M (3A4), 2J0D (3A4), 2NNI (2C8), 2NNH (2C8), 1Z10 (2A6), and 1W0G (3A4). We collected one extra crystal structure from an additional reference: 3NXU (3A4).<sup>32</sup> These 15 crystal structures of CYP–substrate binding complexes were downloaded from the Brookhaven Protein Data Bank.<sup>33</sup>

If there are multiple chains with the same substrate in the crystal structures, only complexes from chain A were used in model development. The ligand structure was then abstracted from each of the bound complexes. If an active site was found to contain multiple ligands, as is the case for 2NNH (2C8), the ligand closest to the heme was chosen for redocking. Hydrogen atoms were then added and new 3D coordinates were generated using the MMFF94x force field using Pybel<sup>34</sup> within Openbabel 2.2.3.<sup>35</sup> Nonpolar hydrogens were deleted and the pdbqt file, which is the ligand input file format for Autodock, was generated using the utilities of ADT 1.5.4.<sup>36</sup> The CYP protein structure and water molecules from the original crystal structure were unchanged. Polar hydrogens were added and the pdbqt file was written using ADT functions.

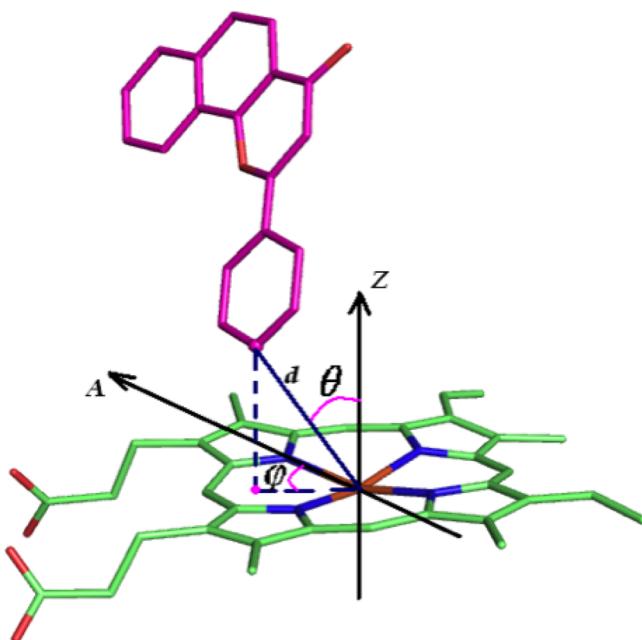
The parameters for Vina docking were set as follows: the size of the docking box was 30 Å in each dimension; the center of the docking box was set to be the geometric center of the ligand in the original crystal structure; the number of CPUs was extended to 7; other parameters were the default values within Vina.

#### Vina Docking and Docking Descriptor Generation.

Each substrate was built and optimized with MMFF94x force field in MOE 2009.<sup>28</sup> Openbabel 2.2.3<sup>35</sup> was employed to convert each substrate set from SDF into pdb format and to add polar hydrogens. This pdb file was then parsed into individual pdb files for each substrate which were then converted to pdbqt files using utilities of AutoDockTools 1.5.4 (ADT).<sup>36</sup> Crystal structures of 2HI4 and 2FDV were used as enzyme structures for docking, because 2HI4 was the only publicly available crystal structure for CYP 1A2 and 2FDV shares the same protein structure with other crystal files of CYP 2A6. For docking with flexible residues, both rigid and flexible parts of the protein structure were written to pdbqt files with ADT 1.5.4. Multiple investigations were made into the docking benefits of making certain side chains in the active-site fixed or

flexible, as well as the benefits of docking with or without active site waters present. Full details of these experiments are described later in the Results and Discussion section and provided in Table 3. For our optimal models, LEU382, LEU497, VAL227, and ILE117 were chosen to be flexible residues in CYP 1A2; LEU370, LEU241, ILE300, and LEU296 were set to be flexible residues in CYP 2A6. The specific parameters used when docking were as follows: the size of docking box was 20 Å in each dimension, which is large enough for substrates in our experiments to be docked in; the center of the docking box was, in angstroms, (3.456, 20, 20.65) for 1A2 and (55.095, 78.335, 59.952) for 2A6; the number of CPUs was set to 7; the energy restraint was extended to 8 kcal/mol; the number of modes is 100; other parameters were set to the default values within Vina.

On the basis of docked poses from Vina, metabolophores were constructed as sets of potential sites that fall within 6.5 Å of the iron. A cutoff distance larger than 5 Å was used in the previous work,<sup>24,25</sup> because the distance itself is one descriptor for evaluating the effects of accessibility on metabolism liability. For each potential site, 31 descriptors were calculated to assess the accessibility and interactions of a putative pose with CYP. A brief definition of each descriptor is given in Table 1. Generally, a descriptor belongs to one of two categories: a geometric characterization or a binding energy term. Spatial information including  $\cos(\theta)$  and  $\cos(\phi)$  within the unique defined coordinate system as shown in Figure 4 is calculated for each



**Figure 4.** Illustration for 3D spatial descriptors abstracted from poses from Autodock Vina. The Z axis is perpendicular to the heme plane, and the A axis is the line within the heme plane and crosses through iron and nitrogen ND of the heme. The coordinate system is uniquely defined on the heme, since the heme is not symmetric.

atom and for the centroid of substrate. For flexible residues, the structural difference is compared between the initial state and the final state after docking using an rmsd metric. The components of the binding energy including the *Gauss* terms, the repulsion and hydrogen interactions between ligands and CYP were extracted from Vina scoring output. There are two terms of *Gauss* in Vina's scoring function, which are highly

correlated, so we only used one of them for the DR-Predictor modeling. Certain highly informative descriptors, like binding energy, were calculated under different docking scenarios: one using a rigid protein with no waters and the other using a rigid backbone with flexible side chains and the crystal waters retained. Other evaluations of binding energy were realized by generating pdbqt input files for both parts and using the Vina score\_only functionality. For instance, for calculating interaction energy between heme and ligand, we extracted the ligand structure from the Vina output file and the heme from CYP pdbqt file. We then wrote new pdbqt files for both of them and called Vina with the "score\_only" parameter set. The above process was scripted in Python using packages including MolKit 1.22.4.1, ADT 1.5.4, and Pybel.<sup>38,34</sup>

**Revamping Reactivity Descriptors.** DR-Predictor calculates 23 atomic descriptors for chemical reactivity based on 3D substrate structures from Vina-generated poses. The most important feature in this class is the activation energy descriptor derived for each atom from the application of SMARTCyp 2.0<sup>8</sup> to the given ligand. Only those atoms that have a SMARTCyp reactivity value are included in models as candidate sites. The derivation of SMARTCyp energies are based on 2D molecular structures so each instance of the candidate SOM in the metabolophore has the same activation energy descriptor value.

The remaining descriptors represent a pruning and slight expansion of the 392 quantum chemical descriptors in RS-Predictor. These descriptors were derived from AM1MOPAC Hamiltonian, which was applied to each Vina generated pose. MOPAC<sup>37</sup> is a semiempirical quantum chemistry program that has been applied for reactivity calculation in several similar cases.<sup>6,9</sup> Though 392 reactivity descriptors were calculated based on MOPAC in our previous work,<sup>6</sup> we calculated only 22 atomic reactivity descriptors because these descriptors can be explicitly related to CYP oxidation mechanisms, and Zaretzki et al. found that a small variance in reactivity can account for the majority of signal variance of the model.<sup>7</sup> The input *mop* files for MOPAC2007 were generated by in-house python scripts with the aid of the pybel package. The keywords for MOPAC calculation were: MMOK VECTORS ALLVEC AM1 EPS=78.4 COSWRT 1SCF EF XYZ MULLIK Charge=n, where n equals the molecular formal charge determined by pybel. We specified "1SCF" to do only one SCF calculation without geometric optimization because there are multiple poses for each substrate and the time cost would not be justified if all the structures were to be optimized. For closer approximation, we extended the calculation of several important properties with alternative COSMO method in MOPAC, as listed in Table 1.

**New Metric for Gauging the Quality of SOM Predictions.** The standard top-*k* metric for evaluating the quality of CYP SOM predictions is the percentage of molecules which have at least one experimentally observed site predicted in the top *k* rank positions by a given model, where *k* is usually 1, 2, or 3.<sup>38</sup> On the basis of the top-*k* metric, Zaretzki et al.<sup>6</sup> developed a LIFT top-*k* metric to rate the relative importance of accurately predicting a candidate substrate based upon the statistical likelihood of randomly choosing one of the observed SOMs of that molecule in *k* guesses. The LIFT top-*k* metric also indicates the correctness in terms of the number of correctly predicted molecules but favors correct prediction on molecules of larger size. However, a flaw exists in both of these metrics in that models that predict one experimental site correctly but do not predict others would be considered successful. This could be misleading when comparing the

Table 2. Comparison of DR-Predictor with RS-Predictor, Jung's Method, and SMARTCyp

CYP 1A2				CYP 2A6			
		molecular percentage	LIFT		averaged Spearman	molecular percentage	LIFT
top-1	DR-Predictor	<b>0.71</b>	<b>0.66</b>	0.48	0.65	0.57	0.42
	RS-Predictor	0.69	0.65	<b>0.49</b>	<b>0.73</b>	<b>0.60</b>	<b>0.54</b>
	Jung <sup>a</sup>	0.60	0.57	0.43	0.67	0.57	0.50
	SMARTCyp	0.64	0.60	0.45	0.69	0.58	<b>0.54</b>
top-2	DR-Predictor	<b>0.86</b>	<b>0.82</b>	0.49	0.83	<b>0.76</b>	<b>0.49</b>
	RS-Predictor	0.83	0.80	<b>0.50</b>	<b>0.84</b>	<b>0.76</b>	0.48
	Jung	0.73	0.70	0.38	0.77	0.66	0.45
	SMARTCyp	0.76	0.72	0.40	0.78	0.66	0.44
top-3	DR-Predictor	<b>0.93</b>	<b>0.91</b>	<b>0.45</b>	<b>0.90</b>	<b>0.85</b>	0.41
	RS-Predictor	0.91	0.88	<b>0.45</b>	0.89	0.81	<b>0.42</b>
	Jung	0.79	0.75	0.36	0.80	0.71	0.41
	SMARTCyp	0.82	0.79	0.34	0.81	0.74	0.34

<sup>a</sup>The model is built with Jung's method, while the calculation is realized through our modeling with Vina and SMARTCyp.

results of different SOM prediction methods. In order to reduce the potential for comparative errors, a new metric was formulated to evaluate the performance of models that can more accurately assess how many metabolism sites that each one of them predicts correctly.

For this purpose, we utilize the Spearman rank correlation coefficient (or simply "Spearman correlation") for assessing the accuracy of regioselectivity of models in predicting experimental sites of metabolism. The Spearman correlation measures the extent to which actual and predicted ranks correlate.<sup>39</sup> It ranges from -1 of complete reverse correlation between actual and prediction ranks, to 1 of perfect prediction. The Spearman coefficient equals to 0 when actual and predicted ranks are not correlated at all. In our work, predicted and experimental ranks of a molecule, denoted as variables  $X$  and  $Y$ , respectively, only contain partial ranking information, as all SOMs rank 1 and all non-SOMs rank 0. For instance, in the top-2 models, the prediction values of top 2 ranking positions are set to be 1 and the predicted values of other positions are set to be 0. The Spearman coefficient is considered to be a good metric for evaluating general ranking problems and can be used for our partial ranking tasks. The Spearman coefficient is calculated by Pearson's correlation coefficient with ranks of  $X$  and  $Y$ , denoted as  $x$  and  $y$ :

$$R_{\text{Spearman}}(X, Y) = R_{\text{Pearson}}(x, y) \\ = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

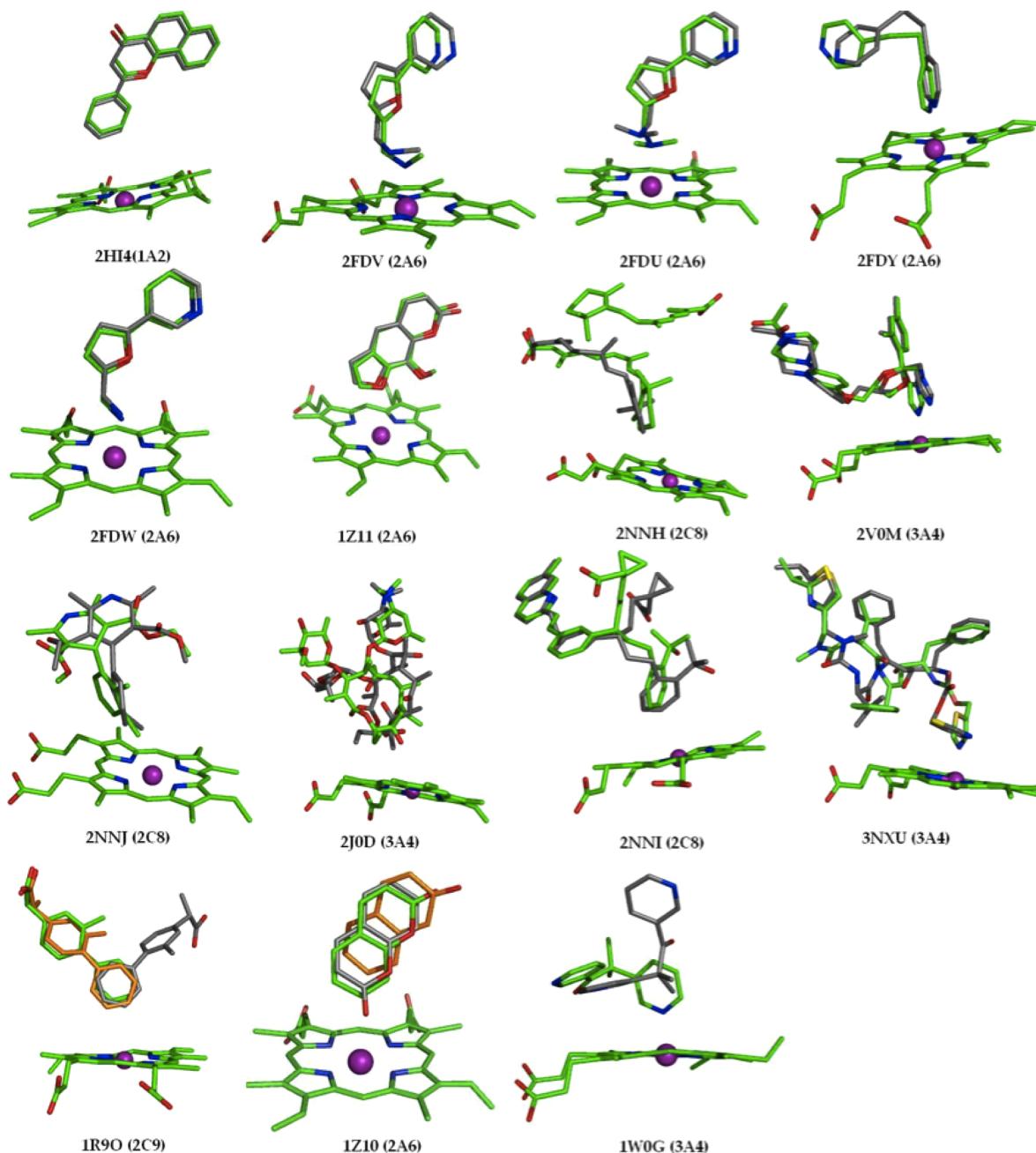
Where,  $\bar{x}$  and  $\bar{y}$  are the average values of variables  $x$  and  $y$  and  $n$  is the number of metabolophores in a molecule. If tied ranks exist—namely there is the same rank to different items in the variable—the rank of these items are all set to the averaged rank. Since  $R_{\text{Spearman}}$  above is only designed for one molecule, we average the coefficients for multiple molecules and denote it as averaged Spearman.

## RESULTS AND DISCUSSION

**Evaluation of DR-Predictor against Previously Released Methods.** The performance of DR-Predictor and three other methods using various metrics are shown in Table 2. Top-1, Top-2 and Top-3 of these models refer to the number of ranking postions to be classified as SOMs. The evaluation algorithms for molecular percentage and LIFT are the same

with our previous publications.<sup>6,7</sup> The result columns designating molecular percentage, LIFT and Average Spearman, reflect different aspects of performance. The molecular percentage reflects the number of molecules correctly predicted with standard Top-k metric. LIFT also reflects the number of molecules correctly predicted but assign a larger score for molecules with larger size. Average Spearman serves as a metric for evaluating how many sites are correctly predicted. Jung's method is implemented according to Jung's Reactivity-Binding-Based kinetics.<sup>24</sup> In Jung's theory, the probability of a SOM being metabolized is determined by the sum of its binding energy and activation energy. While we use Jung's method, the techniques we have employed are unique to this work. The binding energy is from Vina docking calculation and the activation energy is directly from SMARTCyp calculation. The binding energy from Vina converted from kcal/mol to kJ/mol and the sum of binding energy and activation energy is in kJ/mol. RS-Predictor results come from models representing SOMs with topological and SMARTCyp reactivity descriptors as described in our prior work.<sup>6,7</sup> SMARTCyp results are obtained by applying software SMARTCyp 2.0<sup>8</sup> as a ligand-based method.

Overall DR-Predictor has higher prediction accuracy than the other three methods but not in a statistically significant amount. Specially, in regard to all CYP 1A2 models, metrics including molecular percentage and LIFT suggest better performance for DR-Predictor over RS-Predictor, Jung, and SMARTCyp. As more sites are included in the prediction evaluation, the prediction accuracy of DR-Predictor models rises in terms of molecular percentage and LIFT, but not in terms of averaged Spearman. It suggests that the 1A2 models likely predict molecules right only in top-3, with a cost of high false positive of sites. For the 2A6 data set, the performance of DR-Predictor and RS-Predictor is quite close in the top-2 and top-3 models with regard to molecular percentage and LIFT. Since the size of CYP 2A6 data set is relatively small, one percentage difference only indicates one molecule difference in prediction. RS-Predictor shows greater score over other methods within top-1 models. We also note that ligand-based methods including RS-Predictor and SMARTCyp perform better than both structure-integrated methods including DR-Predictor and Jung's in top-1 models. One possible reason is that the molecular size of 2A6 substrates is generally small, so the interaction between CYP 2A6 and the substrates may not play a significant role on the metabolism reaction.



**Figure 5.** Redocking results for available human CYP crystal structures. The ID in each complex is the pdb ID in the pdb bank and the CYP isoform is stated in the following parentheses. The ligand and heme in the crystal structures are shown in the green stick, and the first pose from docking is shown in the gray stick. For 1R9O and 1Z10, the poses of the orange stick, represent the results from Vina redocking most similar to the crystal pose.

The results in Table 2 and Jung's work<sup>24</sup> supports Jung's method as an effective technique for integrating docking and reactivity calculations for regioselectivity prediction. The sum of activation energy and binding energy makes it concise enough for further understanding and analysis of the prediction results. But other factors such as position of potential sites in the active site and the interaction of ligand with specific residues may also affect the metabolism reaction. Another problem we identify is that the sum of activation energy and binding energy may not necessarily differentiate atoms and create degeneracy issues, since many atoms share the binding energy of the ligand and activation energy may also be the same. This does not necessarily mean Jung's modeling theory is incorrect. One reason for the degeneracy issue may be the calculation of both

activation energy and binding energy is an approximation method and more accurate methods might help. However, based on current progress of calculating both terms, incorporating more factors in DR-Predictor may facilitate the differentiation of SOMs and inactive sites of the substrates.

Though accuracy of SMARTCyp is lower than DR-Predictor, it is the essential foundation of DR-Predictor algorithm. The activation energy of SMARTCyp is an effective component for DR-Predictor. Its approximation method on local atomic reactivity is useful because we cannot calculate activation energy by DFT for all compounds due to the high cost of computational time and resources. The energy calculation of SMARTCyp is fast since it does not involve any semiempirical calculation. On the other hand, RS-Predictor provides DR-

Table 3. Comparison of Different Docking Scenarios Affecting the Final Regioselectivity Results

CYP isoforms	scenario	flexible residues	water in receptor	molecular percentage	LIFT	averaged Spearman
1A2	I	ASP313, THR124, LEU382, VAL227	deleted	0.84	0.81	0.46
	II	LEU382, LEU497, VAL227, ILE117	deleted	0.84	0.82	0.46
	III	LEU382, LEU497, VAL227, ILE117	kept	0.85	0.8	0.48
	IV	LEU382, LEU497, VAL227, ILE117	both	0.86	0.82	0.5
2A6	I	PHE107, PHE209, ILE306, THR305	deleted	0.76	0.61	0.44
	II	LEU370, LEU241, ILE300, LEU296	deleted	0.78	0.67	0.47
	III	LEU370, LEU241, ILE300, LEU296	kept	0.8	0.7	0.5
	IV	LEU370, LEU241, ILE300, LEU296	both	0.83	0.76	0.54

Predictor the basic modeling framework, especially the application of multiple instance learning, though different paradigms are constructed.

**CYP Redocking with Vina.** To ensure that Autodock Vina is a reliable tool for generating accurate substrate–CYP complexes, we conducted redocking experiments for available human cytochrome P450 crystal structures with inhibitors when we performed modeling. Though the protein is fixed in redocking, the results can demonstrate the effectiveness of the scoring function and conformation sampling of Vina for CYP docking. Redocking results for 15 human CYP holo structures can be seen in Figure 5. The heme and original ligand from crystal structures are shown in the green sticks, while the first pose from Vina prediction is shown as gray sticks. The PDB ID of each crystal structure is annotated below the binding complex, with specific CYP isoenzymes in the parentheses. For 8 out of 15 CYP binding complexes (shown in the top 2 rows of Figure 5), the first predicted pose from Vina aligns quite well with the ligand crystal pose. Specifically, there are two retinoic acid molecules in the active site of CYP 2C8, 2NNH. Even with promiscuity, Vina’s output aligns with the experimental pose close to the heme, while the second ligand in the crystal structure is removed in the docking process. Another factor to be considered is the role of water in 2HI4. Since  $\alpha$ -naphthoflavone ( $\alpha$ NF) in the CYP 1A2 active site forms a hydrogen bond with structural water, as indicated by previous work,<sup>25</sup> Vina gets the first pose right for 2HI4 if we keep the water. When we do docking without the water molecule, the second predicted pose can align well with the native pose.

When the size of the ligand is relatively large, as shown in the complexes in the third row, the majority of Vina’s output is still consistent with the experimental substrate. For instance, ritonavir is a complex and large inhibitor binding with CYP 3A4 in 3NXU. Even the isopropyl group and thiazol group are misplaced in the predicted pose, the main chain of ritonavir and three major chemical groups align quite well with those in the experimental pose. In the bottom row, the first predicted pose from Vina is not quite the same as the native pose, but we can still see Vina’s performance with further analysis. For 1Z10, Vina reproduces the ligand poses with the second ranked poses (shown in the orange sticks), whose binding energy is only 0.6 kcal/mol lower than that of first predicted pose. The worst case is for 1W0G crystal structure, as no pose from Vina output is the same or approximately the same as the original metyrapone pose in crystal structure. Overall, the results indicate that Vina is suitable for the CYP docking problem in terms of scoring function.

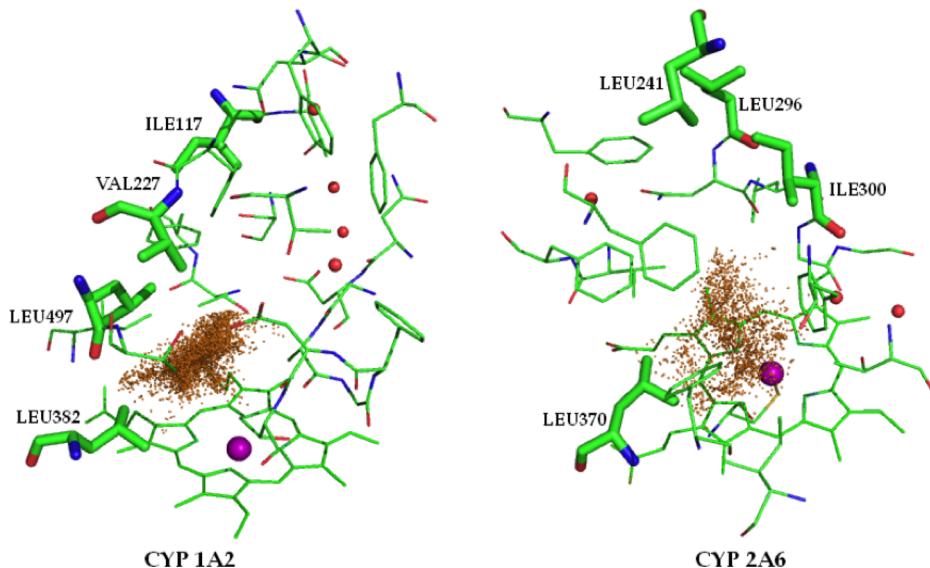
Vina’s scoring function address the iron-ligation effect quite well by setting metals as hydrogen bonding donors and assigning larger weights for hydrogen bonding than other intermolecular interactions such as hydrophobic interaction.

For ligand–CYP interaction, the strength of iron chelation is much stronger than other intermolecular interactions such as hydrophobic interactions. Besides, the flexibility of side chains and even main chains for CYP enzymes would lessen the importance of the ligand side chain interaction and thus increase the importance of the ligation interaction. Therefore, iron chelation as a particular term for CYP docking would contribute a great part to the positioning of different substrate atoms near the heme.<sup>16,17</sup> This may account partially for the unsatisfactory results for previous CYP metabolism site predictions with docking alone.<sup>11,12</sup> The scoring functions in their docking suite might not characterize the iron-ligation effect for polar atoms such as nitrogen and oxygen in substrates.<sup>40</sup> Since the close environment around heme is quite hydrophobic, the hydrogen bonding term in Vina serves more as an iron-ligation term in CYP docking. The predicted pose of 2FDW (2A6) ligand, which also contains iron-ligation nitrogen, aligns almost exactly with the original pose. Besides, for binding complexes of 2FDV (2A6), 2FDU (2A6), and 2V0M (3A4), there is only slight difference of the nitrogen-group position of ligand between prediction pose and original pose. These observations support Vina’s effective weight on iron-ligation.

**Comparison of Different Docking Scenarios.** We use different docking scenarios to investigate the effect of flexible residues and water on the accuracy of DR-Predictor in Table 3 with top-2 models. Three metrics including molecular percentage, LIFT, and averaged Spearman coefficient consistently show that scenario IV can yield better regioselectivity results than other scenarios. Here scenario II for both CYP isoforms is set to be without water, while the original water in CYP crystal structure is kept in scenario III. Scenario IV is the model combining poses from both scenario II and scenario III. In scenario IV, the MIRank algorithm determines whether poses including water are selected for use in the final prediction model. This illustrates a significant advantage of DR-Predictor models because choice of retaining water molecules is no longer left to the discretion of the modeler. Previous exploration of setting water molecules for CYP docking suggests that there may not be a consistent rule but more exceptions for determining which water molecules to be kept for docking.<sup>23</sup> This is quite reasonable because the position of water may depend on the conformation of the CYP protein, which is complicated by the flexibility and the structure and position of each substrate. Though it is hard to determine which water molecules to conserve, the presence of water may affect the binding between substrates and CYP not only in terms of the binding energy but also the spatial position of each atom in the CYP active site. This can be solved, as shown in our work, with multiple instance learning algorithms.<sup>41</sup> The algorithm automatically selects the conformations, which may

Table 4. Averaged Weights' Importance and Their Variation in Both 1A2 and 2A6 Models

CYP 1A2			CYP 2A6		
descriptors	averaged weights	weights std	descriptors	averaged weights	weights std
actE	-0.3307	0.0171	actE	-0.3664	0.0112
nucleo	0.1561	0.0064	min_nb_chg	-0.1882	0.0065
atm_bindE_VAL227	-0.1136	0.0075	nucleo	0.1543	0.0043
dist	-0.111	0.006	cosmo_chg	0.1173	0.0042
phi	-0.1066	0.0024	max_Harea_cosmo	0.1152	0.0053
atm_gauss_1	0.0954	0.0016	sum_HeleF	0.1012	0.0027
bindE_VAL227	-0.0846	0.008	dist	-0.0997	0.0043
max_nb_nucF	0.0841	0.0027	min_Hfuk	-0.093	0.0022
min_nb_chg	-0.084	0.0036	max_nb_nucF	0.0928	0.0027
theta_centerMol	-0.0771	0.0075	dist_centerMol	0.0915	0.0043
atm_hydrophobic	0.0759	0.0058	atm_bindE_THR305	0.0872	0.0032
max_nb_den	0.0739	0.0005	charge	0.0871	0.0034
sum_HeleF	0.0708	0.0017	electr	-0.085	0.0035
cosmo_area	0.0675	0.0019	atm_hydrophobic	0.0842	0.0037
charge	0.0645	0.0007	bindE_ILE300	0.0779	0.005
fukui	0.0571	0.0021	sum_Harea_cosmo	0.0777	0.0026
sum_Hchg_cosmo	-0.0494	0.001	bindE_LEU241	-0.0749	0.0035
dist_centerMol	0.0458	0.0037	rmsd_LEU370	0.0738	0.0044
sum_Harea_cosmo	0.0428	0.0009	bindE_heme	0.0722	0.0036
bindE	-0.0427	0.0038	atm_gauss_1	0.0684	0.0026
phi_centerMol	0.0415	0.002	atm_bindE_LEU296	0.066	0.0028
sum_Hchg	0.0409	0.001	atm_bindE_LEU241	-0.0605	0.0034
rmsd_VAL227	-0.0399	0.0024	theta_centerMol	0.0562	0.0025
density	0.0392	0.0009	density	0.0556	0.0021
atm_bindE_THR321	-0.0385	0.0033	fukui	-0.047	0.0016
max_Harea_cosmo	-0.0378	0.005	phi_centerMol	-0.0454	0.0044
repulsion	-0.0363	0.0035	atm_repulsion	-0.0411	0.0046
h_bond	-0.0335	0.0016	rmsd_LEU241	-0.0374	0.0038
cosmo_chg	0.0324	0.004	repulsion	-0.0365	0.0039
electr	-0.0321	0.0021	cosmo_area	0.0356	0.0041
min_chg_cosmo	0.0319	0.0014	bindE_LEU370	-0.0307	0.0045
bindE_THR321	-0.0313	0.0017	max_nb_den	0.0249	0.0015
bindE_LEU382	-0.0301	0.002	rmsd_res3	0.0231	0.0051
max_nb_fuk	-0.0287	0.0018	bindE_LEU296	-0.0228	0.0024
atm_bindE_heme	-0.0283	0.0009	atm_gauss_2	-0.0194	0.0015
atm_bindE_ILE117	0.0265	0.0006	max_nb_fuk	0.0194	0.0025
atm_bindE	0.0263	0.0044	atm_bindE_heme	-0.0161	0.0026
bindE_heme	-0.0231	0.0019	gauss_1	-0.0147	0.0008
rmsd_LEU382	0.0225	0.0015	max_nb_area	0.0147	0.0012
rmsd_ILE117	-0.0225	0.0021	atm_h_bond	-0.0136	0.0004
rmsd_LEU497	-0.0222	0.0006	atm_bindE	0.0135	0.0031
theta	-0.0215	0.0003	phi	-0.0119	0.0016
atm_bindE_LEU382	-0.0204	0.0011	hydrophobic	0.0105	0.001
min_Hfuk	0.0198	0.0015	atm_bindE_LEU370	0.0103	0.0021
atm_gauss_2	0.0195	0.0019	h_bond	-0.0082	0.0042
max_nb_area	0.0195	0.0009	min_chg_cosmo	-0.0082	0.0015
atm_repulsion	-0.0186	0.0012	min_nb_eleF	0.0075	0.002
hydrophobic	-0.0158	0.0015	sum_Hchg_cosmo	-0.0055	0.0016
max_Hchg_cosmo	-0.0152	0.0019	theta	0.0038	0.0008
max_Hchg	-0.0095	0.0026	bindE	0.004	0.0013
atm_bindE_LEU497	-0.0055	0.0019	bindE_THR305	-0.0029	0.001
gauss_1	-0.005	0.0002	max_Hchg_cosmo	0.0032	0.0007
atm_h_bond	-0.0048	0.0006	max_Hchg	0.0023	0.0015
bindE_LEU497	0.0044	0.0008	rmsd_LEU296	0.0003	0.0024
min_nb_eleF	0.0043	0.0005	sum_Hchg	-0.0004	0.0007
bindE_ILE117	-0.0035	0.0008	atm_bindE_ILE300	0	0.2254



**Figure 6.** Distribution of SOMs in the CYP active site from docking. The heme and rigid residues in the crystal structures are shown in the green line, and the flexible residues in the docking experiments are highlighted in the green stick. The iron of the heme is shown as the purple ball. Each yellow represents an occurrence of experimental SOM close to the iron in docking pose.

come from docking with water or docking without water for our work and fits the descriptors of the instances with experimental responses.

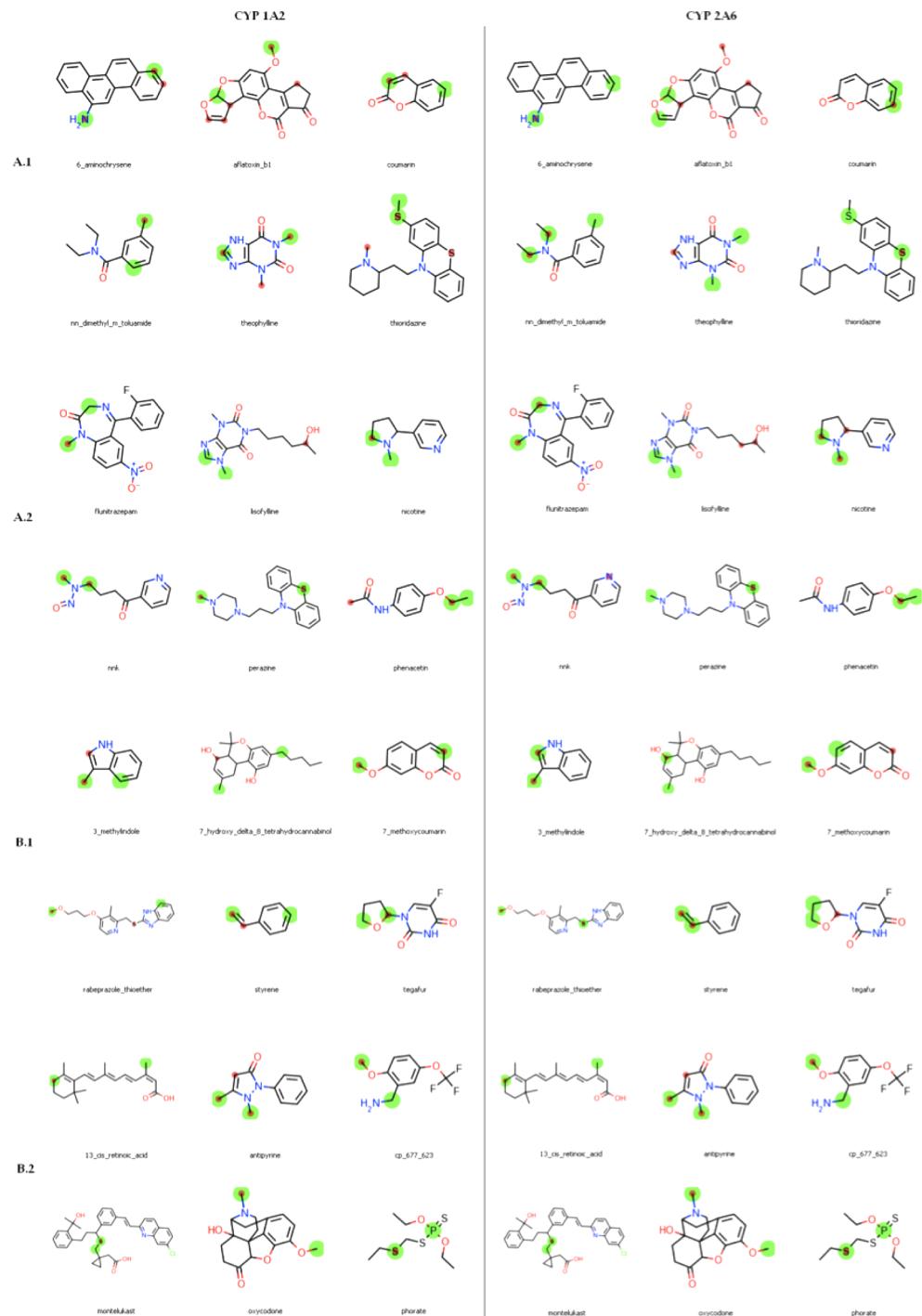
Even though CYP 1A2 and CYP 2A6 are considered two of the most rigid CYP enzymes and their active sites are relatively small, flexible residues are still necessary. First, the promiscuity property for CYP 2A6, indicated by experiments,<sup>42</sup> reflects its local flexibility. In addition, the substrates we collect here for both enzymes are quite structurally diverse, and some of them like montelukast are quite large compared to the active sites of the CYP enzyme.

To perform docking with flexible residues with Autodock Vina or other versions of Autodock, it is necessary to specify which residues are allowed to be flexible. Since it is not practically possible to combinatorially examine the flexibility of all residues for docking in a cross-validation fashion, we manually select a few residues to conduct an exploratory comparison experiment. We limit the number of flexible residues to four because statistically three flexible residues can account for 85% flexibility of active sites<sup>43</sup> and the active sites of 1A2 and 2A6 are relatively small. Too many flexible residues not only cost computation time but also may lower the accuracy of docking results by creating noise for flexibility of some irrelevant residues. These residues are chosen collectively based on three sources. The first one is the temperature factor of each residue in crystal structures. Temperature factor, representing mean square fluctuations of atoms relative to their average position,<sup>44</sup> can reflect the mobility of residues to some extent. Most flexible residues chosen here rank top in terms of temperature factors around the active site of CYP. Second, we refer to previous work related to CYP 1A2 and CYP 2A6.<sup>20,25</sup> In regard to CYP 2A6, Skopalík et al.<sup>20</sup> mention that several residues including PHE209, THR305, ILE366, ILE208, and PHE117 may play an important role in the path between active site and bulk outside the protein. For CYP 1A2, through virtual screening, Oostenbrink's group identifies the residues interacting most with the substrates, which are aromatic residues, THR124 and ASP313.<sup>25</sup> We also consider the general ranking of residue flexibility from statistical analysis:<sup>43</sup> Lys > Arg, Gln,

Met > Glu, Ile, Leu > Asn, Thr, Val, Tyr, Ser, His, Asp > Cys, Trp, Phe.

As shown in Table 3, two sets of flexible residues are presented across four scenarios for both CYP 1A2 and CYP 2A6. The flexible residues in scenario I of both CYP 1A2 and CYP 2A6 diversely consist of nonpolar, polar, and aromatic residues, while residues in other scenarios are purely nonpolar residues. For both 1A2 and 2A6, docking models with nonpolar flexible residues perform better than models with multiple types of flexible residues. This could be especially seen by the slightly higher accuracy of scenario II over scenario I, where the only difference is the setting of flexible residues. The better performance of setting nonpolar residues flexible here may stem from the strong interactions, like hydrogen bonds, between polar residues and other residues in the active sites of both CYP 1A2 and CYP 2A6. Besides, most CYP substrates are quite hydrophobic; therefore, they may have more binding interaction with nonpolar residues in the active site. As for aromatic residues like phenylalanine, they consist of a low number of rotatable bonds and are in the lowest ranking position of flexible residues.<sup>43</sup>

**Analysis of Descriptor Importance.** The average weights and standard deviation of each descriptor across 10 models are presented in Table 4. The weight represents the importance of the descriptors to determining the metabolic liability of each atom. Average and standard deviation for these weights are calculated on 10 models of 10-fold validation. The table is ranked by the absolute value of each descriptor's weight but the weights are shown with signs that reflect its positive or negative correlation to regioselectivity. For instance, the averaged weight for *actE* in CYP 1A2 model is -0.3307, and it indicates the lower the activation energy is, the more likely the atom would be an SOM. Generally, the weight for most of descriptors is stable, as the standard deviation is smaller in scale than its averaged value. The descriptors with the lowest absolute weight (the last thirteen descriptors for CYP 2A6) are small with average value but relatively large with standard deviation. They can be considered the noise of the models as the value of their importance is small and varies too much.



**Figure 7.** Prediction examples for molecules predicted from CYP 1A2 and CYP 2A6 substrate sets in left and right columns respectively. Predicted sites of metabolism by DR-Predictor are highlighted in green, while red highlighted atoms are experimental sites of metabolism. Molecules in panel A have different experimental SOMs. Within panel A, molecules in panel A.1 have different predictions for 1A2 and 2A6 models, while molecules in A.2 are have the same predictions for the two models. Molecules in panel B have the same experimental SOMs for 1A2 and 2A6 enzymes. In panel B, molecules in B.1 are predicted with different SOMs by our two models, while molecules in B.2 are predicted with same SOMs.

These most important descriptors are from both reactivity and docking, validating our assumption that both chemical reactivity and protein–ligand interactions<sup>45</sup> are important for CYP metabolism. Various descriptors from docking show their importance for both CYP 1A2 and CYP 2A6. The relative position of a given SOM within the binding site is one of the most important aspects of the docking-derived descriptors—*dist*, *phi* for CYP 1A2 and *dist* for CYP 2A6. The position of

each atom also determines its interaction energy with CYP protein including its specific residues, and the importance of the position is indicated by high ranking of *atm\_BindE\_VAL227* for CYP 1A2 model and *atm\_BindE\_THR305* for CYP 2A6 model. On the basis of the ranking of these docking descriptors, we also can see the importance of some residues. For instance, descriptors concerning interaction with VAL227 rank relatively high for CYP 1A2 model. The spatial distribution of all SOMs

from all docking poses can be seen from Figure 6. We can see that the distribution is not quite even in either active site, and this explains that the position of atoms in active sites do affect whether they would be metabolized or not.

Besides interaction between potential SOM and CYP, the property of the whole substrate molecule and its interaction with the enzyme are also important. The position of the molecular center is also influential, indicated by the high ranking of descriptors: *theta\_centerMol*, *dist\_centerMol*, *phi\_centerMol*, and *bindE* for CYP 1A2; *dist\_centerMol*, *bindE\_LEU300*, and *bindE\_LEU241* for CYP 2A6. These descriptors characterize the essential interaction between ligand and protein that account for rate limiting steps of CYP-mediated metabolism reaction, such as activation of oxygen.

The activation energy calculated from SMARTCyp—*actE*—is the most important descriptor for both CYP models. Beyond *actE*, other AM1-derived reactivity descriptors strongly influence the models. Descriptors including *nucleo* and *charge* are positively correlated with the metabolic liability of each atom, while *electro* is negatively correlated with this liability with a large weight in 2A6 models. The importance of reactivity descriptors should vary between N-dealkylation, O-dealkylation, aliphatic hydroxylation, aromatic hydroxylation, and other different oxidation mechanisms,<sup>5</sup> as the mechanism for different oxidation reactions is different.<sup>46</sup> Our models are capable of selecting commonly important reactivity descriptors for various mechanisms, as supported by the top 3 reactivity descriptors in the table including activation energy and atomic Fukui nucleophilicity. Other than *actE*, atomic Fukui indices are alternative approaches to modeling the atomic reactivity. In our 1A2 and 2A6 models, the importance of Fukui nucleophilicity of base atoms is consistent with general CYP oxidation mechanisms. According to CYP oxidation mechanisms, reaction intermediates including both compound I (Cpd I) and compound 0 (Cpd 0) are currently considered candidates for oxidizing the substrates.<sup>47</sup> Cpd I is electrophile, under which the Fukui electrophilicity of SOM is important. And hydrogen Fukui electrophilicity might play an important role when the nucleophile Cpd 0 is dominant for oxidation. On the basis of frontier molecular orbital theory, atoms with high Fukui indices would probably be attacked by a soft nucleophilic/electrophilic compound, while the reactivity of hard atoms is largely determined by the partial charge.

Another important category of descriptors from MOPAC calculation is the surface area properties, whose importance is also supported by previous work.<sup>2,5</sup> The surface area descriptors of the base atoms and their connected hydrogen atoms are in the top 6 ranking for CYP 1A2 and 2A6, indicated by higher ranking of *cosmo\_area* and *max\_Harea\_cosmo* relatively. Surface descriptors are positively correlated to SOM-prediction; this is expected, as greater area/accessibility of a given SOM mean that that SOM has a greater likelihood to reach the catalytic-heme and be oxidized.

**Comparison of Prediction Across CYP Isoforms.** In this section, we will look at some example molecules metabolized by both 1A2 and 2A6. While both 1A2 and 2A6 active sites are small and structurally similar,<sup>48</sup> we find that 1A2 is much more promiscuous. Between 261 1A2 molecules and 100 2A6 molecules, there are 60 overlap molecules. Figure 7 shows the examples of these 60 molecules based our top-2 predictions. We enclose the molecules into two panels: panel A in which molecules have different experimental SOMs and panel B in which molecules have the same experimental SOMs. Within

panel A, two subpanels are defined: panel A.1 in which molecules have different predictions for 1A2 and 2A6 models and panel A.2 in which molecules have the same predictions for the two models. Similarly, in panel B, molecules in B.1 are predicted with different SOMs by our two models, while molecules in B.2 are predicted with same SOMs. Within each molecule, experimental SOMs are highlighted by small red dots and predicted SOMs are highlighted with large green dots. The molecules that CYP 1A2 model fails to predict accurately are the following: *lisofylline* and *7\_hydroxyl\_delta\_8\_tetrahydrocannabinol*. The CYP 2A6 model fails only on *aflatoxin\_b1*, *lisofylline*, and *tegafur*.

For molecules in panel A, at least one of their experimental SOMs differ between 1A2 and 2A6 metabolism. This suggests the necessity of building a specific model for each CYP enzyme and including factors describing enzyme–ligand interaction within the model. Because the chemical reactivity of each putative SOM would be the same and it is the differential interaction of the ligand with these binding pockets that differentiates the CYP-mediated metabolism of each substrate. In panel A.1, DR-Predictor shows isoform selectivity by correctly predicting molecules with different atoms across CYP 1A2 and CYP 2A6. It is reported that aromatic hydroxylation sites are more difficult to predict,<sup>9</sup> but DR-Predictor correctly predicts the site of *6\_aminochrysene* aromatic hydroxylation metabolized by CYP 1A2. For most substrates in panel A.2, the experimental SOMs between CYP 1A2 and CYP 2A6 are not exactly the same. The difference is mostly due to the fact that one enzyme metabolizes one or more sites than the other enzyme. For instance, metabolism of *flunitrazepam* shares, as highlighted in Figure 7, the same site of methyl group on nitrogen between CYP 1A2 and CYP 2A6. In these cases, DR-Predictor models correctly predict the common SOMs between two enzymes and achieve satisfactory accuracy.

The substrates in panel B have the same experimental SOMs for both CYP 1A2 and 2A6. The predicted SOMs of molecules in panel B.1 vary between CYP 1A2 and CYP 2A6 models, while the predicted SOMs for molecules in panel B.2 are the same. Results in B.2 are reasonable because we should obtain the same prediction results if the experimental results are the same for these CYP substrates. In B.1, the experimental results suggest the same binding effect on the molecules, but our docking software may predict different binding poses between CYP 1A2 and CYP 2A6, while the difference is 2A6 metabolizes an additional ring-hydroxylation. DR-Predictor predicts some molecules in panel B that are relatively large in size and hard to predict. For instance, in *montelukast*, the sulfur atom as the SOM is in the middle of the whole large molecule.

## CONCLUSION

In this paper, we present DR-Predictor for CYP metabolism prediction by integrating docking and reactivity calculation with machine learning algorithms. We apply our method to substrate sets of two pharmaceutically important CYP enzymes: 261 substrates of CYP 1A2 and 100 substrates of CYP 2A6. DR-Predictor identifies experimental SOMs for 86% CYP 1A2 substrates and 84% CYP 2A6 substrates in the top two rank-positions. For 1A2 data set, DR-Predictor yields results with higher accuracy than other models in various metrics, while DR-Predictor performs equivalently well with other methods for 2A6 data set. In order to guide improvement of CYP metabolism prediction in the future, we develop the metric of averaged Spearman in terms of site numbers, instead of

molecule numbers, to evaluate our models. We believe average Spearman would let users of regioselectivity software to be more aware of the performance of different models account every metabolism sites. In addition, key issues for CYP docking including residue flexibility and water setting are well addressed with our models.

DR-Predictor is the first method to integrate binding information from docking and chemical reactivity with machine learning algorithms. Specifically, DR-Predictor utilizes multiple instance learning algorithm MIRank to compare SOMs of the same substrate based on conformation-specific descriptors. In this approach, we automatically weight importance of different descriptors from docking and chemical reactivity calculation. On the basis of model weights, we can identify important docking and reactivity descriptors such as spatial position of putative SOMs in the active site and activation energy from SMARTCyp. The important descriptors identified here may further help chemists to understand CYP metabolism reaction with more factors. Though other CYP enzymes such as CYP 3A4 are more flexible than CYP 1A2 and 2A6, our method could still be extended to those enzymes, provided the availability of high quality of crystal structures and techniques addressing global flexibility.

## ■ ASSOCIATED CONTENT

### Supporting Information

SD file for both CYP 1A2 and CYP 2A6 substrate structures are available. Both experimental SOMs and top-2 predicted SOMs are included as attributes for each substrate in the SD file. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: brenec@rpi.edu.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors like to thank Eli Lilly for funding this project. We thank Michael Krein and Nagamani Sukumar for wonderful discussion about QSAR modeling and chemical reactivity calculation. We also appreciate Mike Krein provides support of computer administration.

## ■ REFERENCES

- Burton, J.; Ijjaali, I.; Petitet, F.; Michel, A.; Vercauteren, D. Virtual screening for cytochromes p450: successes of machine learning filters. *Comb. Chem. High Throughput Screening* **2009**, *12*, 369–382.
- Rydberg, P.; Gloriam, D.; 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.
- Sykes, M.; McKinnon, R.; Miners, J. Prediction of metabolism by cytochrome P450 2C9: alignment and docking studies of a validated database of substrates. *J. Med. Chem.* **2008**, *51*, 780–791.
- Singh, S.; Shen, L.; Walker, M.; Sheridan, R. A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *J. Med. Chem.* **2003**, *46*, 1330–1336.
- Sheridan, R.; Korzekwa, K.; Torres, R.; Walker, M. Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173–3184.
- Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T.; Bennett, K.; Breneman, C. 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.
- Zaretzki, J.; Rydberg, P.; Bergeron, C.; Bennett, K.; Olsen, L.; Breneman, C. RS-Predictor models augmented with SMARTCyp reactivities: robust metabolic regioselectivity predictions for nine CYP isozymes. *J. Chem. Inf. Model.* **2012**, *52*, 1637–1659.
- Rydberg, P.; Olsen, L. Ligand-based site of metabolism prediction for cytochrome P450 2D6. *ACS Med. Chem. Lett.* **2011**, *3*, 69–73.
- 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.
- Hennemann, M.; Friedl, A.; Lobell, M.; Keldenich, J.; Hillisch, A.; Clark, T.; Gölle, A. CypScore: quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory. *ChemMedChem* **2009**, *4*, 657–669.
- Prusis, P.; Afzelius, L. Improvement of site of metabolism predictions for CYP3A4 by using discriminant analysis of compound preference of CYP3A4 X-Ray structural conformers and subsequent docking. *QSAR Comb. Sci.* **2009**, *28*, 865–868.
- Prusis, P.; Afzelius, L. reduced vdW radius improves site of metabolism predictions using X-Ray structure of CYP2D6. *QSAR Comb. Sci.* **2009**, *28*, 891–894.
- Campagna-Slater, V.; Pottel, J.; Therrien, E.; Cantin, L.; Moitessier, N. Development of a computational tool to rival experts in the prediction of sites of metabolism of xenobiotics by P450s. *J. Chem. Inf. Model.* **2012**, *52*, 2471–2483.
- Guengerich, F.; Isin, E. Mechanisms of cytochrome P450 reactions. *Acta Chim. Slov.* **2008**, *55*, 7–19.
- Isin, E.; Guengerich, F. Substrate binding to cytochromes P450. *Anal. Bioanal. Chem.* **2008**, *392*, 1019–1030.
- Kirton, S.; Murray, C.; Verdonk, M.; Taylor, R. Prediction of binding modes for ligands in the cytochromes P450 and other heme-containing proteins. *Proteins Struct. Funct. Bioinf.* **2005**, *58*, 836–844.
- Röhrig, U.; Grosdidier, A.; Zoete, V.; Michielin, O. Docking to heme proteins. *J. Comput. Chem.* **2009**, *30*, 2305–2315.
- Williams, P.; Cosme, J.; Vinković, D.; Ward, A.; Angove, H.; Day, P.; Vonrhein, C.; Tickle, I.; Jhoti, H. Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* **2004**, *305*, 683–686.
- Ekoos, M.; Sjögren, T. Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13682–13687.
- Skopalík, J.; Anzenbacher, P.; Otyepka, M. Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *J. Phys. Chem. B* **2008**, *112*, 8165–8173.
- Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J. Med. Chem.* **2008**, *51*, 7469–7477.
- Moors, S.; Vos, A.; Cummings, M.; Van Vlijmen, H.; Ceulemans, A. Structure-based site of metabolism prediction for cytochrome P450 2D6. *J. Med. Chem.* **2011**, *54*, 6098–6105.
- Santos, R.; Hritz, J.; Oostenbrink, C. Role of water in molecular docking simulations of cytochrome P450 2D6. *J. Chem. Inf. Model.* **2009**, *50*, 146–154.
- Jung, J.; Kim, N.; Kim, S.; Choi, I.; Cho, K.; Oh, W.; Kim, D.; No, K. Regioselectivity prediction of CYP1A2-mediated phase I metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 1074–1080.
- Vasanthanathan, P.; Hritz, J.; Tabouret, O.; Olsen, L.; Steen Jørgensen, F.; Vermeulen, N.; Oostenbrink, C. Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. *J. Chem. Inf. Model.* **2008**, *49*, 43–52.
- Bergeron, C.; Zaretzki, J.; Breneman, C.; Bennett, K. *Multiple instance ranking*. Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, July 5–9, 2008.
- Bergeron, C.; Moore, G.; Zaretzki, J.; Breneman, C.; Bennett, K. Fast bundle algorithm for multiple-instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1068–1079.

- (28) MOE, version 2009.10; Chemical Computing Group: Montreal, Canada, 2009.
- (29) Bolton, E.; Wang, Y.; Thiessen, P.; Bryant, S. PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217–241.
- (30) Trott, O.; Olson, A. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.
- (31) Wang, J.; Zhang, C.; Chou, K.; Wei, D. Structure of cytochrome P450s and personalized drug. *Curr. Med. Chem.* **2009**, *16*, 232–244.
- (32) Sevrioukova, I.; Poulos, T. Structure and mechanism of the complex between cytochrome P4503A4 and ritonavir. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 18422–18427.
- (33) Bernstein, F.; Koetzle, T.; Williams, G.; Meyer, E.; Brice, M.; Rodgers, J.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.* **1978**, *185*, 584–591.
- (34) O’Boyle, N.; Morley, C.; Hutchison, G. Pybel: a Python wrapper for the OpenBabel Cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*.
- (35) O’Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: an open chemical toolbox. *J. Cheminf.* **2011**, *3*, 1–14.
- (36) Morris, G.; Huey, R.; Lindstrom, W.; Sanner, M.; Belew, R.; Goodsell, D.; Olson, A. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791.
- (37) Stewart, J. *MOPAC 2007*, version 7, 290 W; Stewart Computational Chemistry, Colorado Springs, CO, 2007,
- (38) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* **2005**, *48*, 6970–6979.
- (39) Boslaugh, S.; Watters, P. *Statistics in a nutshell: A desktop quick reference*; O'Reilly Media, Incorporated: Sebastopol, CA, 2008; pp 183–184.
- (40) Friesner, R.; Banks, J.; Murphy, R.; Halgren, T.; Klicic, J.; Mainz, D.; Repasky, M.; Knoll, E.; Shelley, M.; Perry, J.; Shaw, D.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- (41) Dietterich, T.; Lathrop, R.; Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **1997**, *89*, 31–71.
- (42) Harrelson, J.; Atkins, W.; Nelson, S. Multiple-ligand binding in CYP2A6: probing mechanisms of cytochrome P450 cooperativity by assessing substrate dynamics. *Biochemistry* **2008**, *47*, 2978.
- (43) Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Side-chain flexibility in proteins upon ligand binding. *Proteins Struct. Funct. Bioinf.* **2000**, *39*, 261–268.
- (44) Yuan, Z.; Zhao, J.; Wang, Z. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng.* **2003**, *16*, 109–114.
- (45) Zhou, D.; Afzelius, L.; Grimm, S.; Andersson, T.; Zauhar, R.; Zamora, I. Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metab. Dispos.* **2006**, *34*, 976–983.
- (46) Meunier, B.; de Visser, S.; Shaik, S. Mechanism of oxidation reactions catalyzed by cytochrome P450 enzymes. *Chem. Rev.* **2004**, *104*, 3947–3980.
- (47) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations. *Chem. Rev.* **2010**, *110*, 949–1017.
- (48) Sansen, S.; Yano, J.; Reynald, R.; Schoch, G.; Griffin, K.; Stout, C.; Johnson, E. Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. *J. Biol. Chem.* **2007**, *282*, 14348–14355.
- (49) Zaretzki, J.; Matlock, M. K.; Swamidass, S. J. XenoSite: Accurately predicting CYP-mediated sites of metabolism with neural networks. *J. Chem. Inf. Model.*, 10.1021/ci400518g.