JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

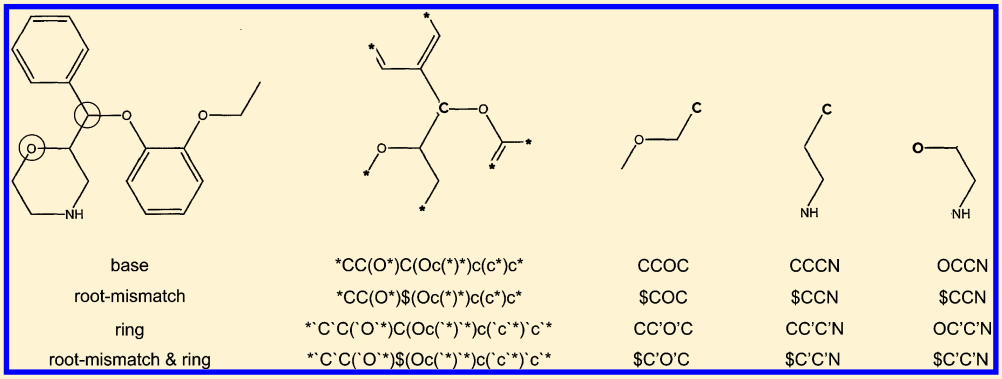
pubs.acs.org/jcim

Improved Prediction of CYP-Mediated Metabolism with Chemical Fingerprints

Jed Zaretzki, Kevin M. Boehm, and S. Joshua Swamidass*

Department of Pathology and Immunology, Washington University School of Medicine, Campus Box 1097 Whitaker Hall, St. Louis, Missouri 63130, United States

Supporting Information



ABSTRACT: Molecule and atom fingerprints, similar to path-based Daylight fingerprints, can substantially improve the accuracy of P450 site-of-metabolism prediction models. Only two chemical fingerprints have been used in metabolism prediction, so little is known about the importance of fingerprint parameters on site of metabolism predictions. It is possible that different fingerprints might yield more accurate models. Here, we study if tuning fingerprints to specific site of metabolism data sets can lead to improved models. We measure the impact of 484 specific chemical fingerprints on the accuracy of P450 site-of-metabolism prediction models on nine P450 isoform site of metabolism data sets. Using a range of search depths, we study path, circular, and subgraph fingerprints. Two different labelings, also, are considered, both standard SMILES labels and also a labeling that marks ring bonds differently than nonring bonds, enabling ortho, para, and meta positioning of substituents to be more clearly encoded. Optimal fingerprint models chosen by cross-validation performance on the full training data are, on average, 3.8% (Top-2; percent of molecules with a site of metabolism in the top two predictions) and 1.4% (AUC; area under the ROC curve) more accurate than base fingerprint models. These gains represent, respectively, a 25.6% and 16.7% reduction in error. A more rigorous assessment selects fingerprints within each cross-validation fold, sometimes selecting different fingerprints for different folds, but yielding a more reliable estimate of generalization error. In this assessment, averaging the scores from the top few fingerprints yields performances improvements of, on average, 3.0% (Top-2) and 0.7% (AUC). These gains are statistically significant and represent, respectively, a 20.1% and 8.8% reduction in error. Between different isoforms, not many consistencies were observed among the top performing fingerprints, with different fingerprints working best for different isoforms. These results suggest that there are important gains achievable in site of metabolism modeling by including and optimizing atom and molecule fingerprints. The optimal site of metabolism models determined by this approach are available for use at http://swami.wustl.edu/.

INTRODUCTION

The cytochrome P450 enzymes (P450s) are a family of enzymes responsible for the metabolism of approximately 90% of FDA approved drugs on the market.[1,2] The P450-mediated metabolism of a drug affects its clinical efficacy and safety. If a drug is metabolized too fast it will not stay in the system long enough to have a significant therapeutic impact unless it is administered at increased dosages. However, if the same drug is also metabolized by a P450—or another enzyme—into a toxic product, it may not be safe to administer at an efficacious dose. In recent years, computational models that predict which atom(s) of a lead candidate undergo CYP-mediated oxidation *in vivo*—its site(s) of metabolism—using only its molecular structure have become essential tools in the drug-discovery process.[3,4] Site-of-metabolism prediction models enable medicinal chemists to rationally design new lead candidates with improved metabolic profiles without having a deleterious effect on efficacy.

In prior work we described XenoSite, a tool for building CYP site-of-metabolism prediction models. XenoSite can identify the metabolism of 85% of the P450 substrates and metabolites

Received: September 16, 2014
Published: April 14, 2015

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

Table 1. Number of Isoform-Specific Substrates Used in This Study and Their Average Composition of P450-Oxidized Atoms (SOMs), Atoms Not Oxidized by P450s (not SOMs), and the Ratio between SOMs and not SOMs

parameter	values								
isoform	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
number of substrates	271	105	151	142	226	218	270	145	475
SOMs	1.9	1.5	1.5	1.7	1.7	1.6	1.6	1.7	1.9
not SOMs	16.0	12.3	14.5	17.2	16.8	16.7	17.0	12.2	20.1
ratio SOMs to not SOMs	11.9%	12.2%	10.3%	9.9%	10.1%	9.6%	9.4%	13.9%	9.5%

publicly available.[5] Other methods tested—some commercial and some academic—were not as accurate for the same data set—RS-Predictor[6] (84.3%), SMARTCyp[7] (82%), StarDrop[8] (76%), and Schrödinger[9] (72%).

In XenoSite, molecule and atom fingerprints, similar to path-based Daylight fingerprints, substantially improve the accuracy of P450 site-of-metabolism models.[5] These fingerprints accounted for most of the performance improvements of XenoSite over prior methods. To the best of our knowledge, this is the first time fingerprints describing atom- and molecule-level structure have been used in P450 site of metabolism modeling. Subsequently, similar results were reported, showing that fingerprints alone could be used to build reasonably performing site-of-metabolism models.[10] In both those studies, only one type of chemical fingerprint has been used in metabolism prediction, so little is known about the importance of fingerprint parameters on site-of-metabolism predictions. It is possible that different fingerprints might yield more accurate models. Here, we study if improvements in site-of-metabolism predictions are achievable by tuning fingerprints to specific site of metabolism data sets.

In this study we systematically test 484 different types of fingerprints to see which are best for building site-of-metabolism prediction models for nine P450 isoforms. We show that optimized fingerprints for specific site of metabolism data sets yield substantial performance improvements. Moreover, we make the best site-of-metabolism models in this study available on a publicly available Web site.

MATERIALS AND METHODS

Site of Metabolism Data. All variations studied here learn from a set of training data. For this training data, we use a public repository of metabolic reactions catalyzed by nine P450 isoforms for 680 distinct substrates (Table 1).[5] Each atom in the data set is marked as either a site or not-a-site of metabolism, based on whether there is published literature showing that the atom is oxidized when the substrate is incubated with a P450 isoform. Topologically equivalent atoms are given the same label; if any of the atoms are sites of metabolism they are all marked as such.

XenoSite Models. *Descriptors.* Five classes of descriptors are used to numerically characterize every heavy atom in the data. Topological descriptors encode information derivable from a 2D molecular structure; for example the distribution of different atom-types (C, O, N, etc.) up to three bond-lengths away from the atom being characterized. Quantum chemical descriptors for the heavy atoms of a substrate are computed using MOPAC[11] with a minimum energy conformation of its molecular structure. A SMARTCyp-derived reactivity descriptor encodes the highest transition-state barrier that must be overcome for the atom to be oxidized by a P450 heme group.[12] Molecular descriptors encode information that is specific to each substrate, such as logP and total polar surface area; all heavy atoms from the same substrate have the same molecular descriptor values.

The last class of descriptors are computed using fingerprints. Atom and molecule fingerprints are defined for each atom and molecule in the data set (a complete description is included in a following section). MinMax similarity—a variant of Tanimoto similarity that uses counts[13,14]—between atoms of different molecules in the data set is then calculated in two ways. First, similarity between atom fingerprints alone group atoms in similar environments together. Second, similarity between atom fingerprints multiplied by the similarity between the corresponding molecule fingerprints group atoms from similar molecules in similar environments together. Using both these approaches, the closest neighbor atoms of each atom in the data set are identified. For each approach four descriptors for each atom are calculated from the average class (site or not-a-site of metabolism) of its top 1, 5, 10, and 20 neighbors.

Our first implementation of XenoSite calculated topological, quantum chemical, and molecular descriptors using the commercially available Molecular Operating Environment (MOE).[15] To ease the dissemination of XenoSite to other groups, XenoSite descriptor calculation has been reimplemented in Python using open-source software packages. The functions available within these packages are different than those available within MOE, and so our reimplementation has resulted in the calculation of slightly different descriptors. Though these descriptors are different than previous implementations, they encode equivalent chemical information, and the change in XenoSite performance when using the newly calculated descriptors is small. These differences, and the exact definitions of the topological, quantum chemical, and molecular descriptors used in this work are provided in the Supporting Information.

Model Structure and Training. XenoSite uses a neural network with 10 hidden nodes to find a mapping from the descriptors to the target class value for each atom in the data set. For each training set, a neural network is used to find a mathematical mapping between the descriptors and the site-of-metabolism label. The network's weights are calibrated using gradient descent on the cross-entropy error associated with the correlation of each atom's encoded descriptors with its associated target. A leave-one-out cross-validation paradigm is employed, whereby each individual substrate in each isoform set is predicted by an independent site of metabolism model trained on the remaining substrates of same isoform set. The prediction for each heavy atom of the test substrate is a score that falls between 0 and 1, which correlates to the statistical likelihood that the atom is a site of metabolism. A more complete description is included in the original XenoSite paper.[5]

Molecule and Atom Fingerprints. Fingerprints are defined for both atoms and molecules in a similar process. First, a molecule's structure is converted to a graph, with

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

labeled edges and nodes respectively corresponding to the molecule's bonds and atoms. Labeling the graph in different ways gives rise to different fingerprints. Next, substructures are enumerated from the graph. Different algorithms, parametrized by depth, can be used to extract different types of substructures from the graph. Next, an algorithm is used to generate a canonical representation of the subgraph in string form. Different canonicalization strategies can yield different fingerprints. Finally, the number of times each substructure appears in a molecule is stored in an efficient data structure. Each of these steps are described in turn in the following sections.

In this study, we consider two labeling options, 11 molecule substructure options, 11 atom substructure options, and two canonicalization options. This defines a total of 484 combinations of options, each of which define a fingerprint and all of which were tested.

Labeling. We consider two types of labels. First, the SMILES labeling scheme labels atoms and bonds in the same way they are recorded in SMILES strings.[16,17] The base model, used in the initial XenoSite algorithm, uses this labeling. Each vertex is labeled with the corresponding atomic element (e.g., C, N, O, Br). Labels are lowercase for aromatic atoms and titlecase for nonaromatic atoms. Each edge is labeled according to the order of the corresponding bond. Double bonds are an equal sign '=', triple bonds are a hash sign '#', and both single and aromatic bonds are an empty string. Second, we modified this labeling to mark bonds with an apostrophe if they are part of a ring. This labeling enables ortho, para, and meta positioning of ring substituents to be more clearly encoded. This may improve site of metabolism predictions because positioning on rings can be important to predictions. Other labeling schemes are possible, but they are not considered in this study.

Atom Substructures. To calculate an atom fingerprint three different algorithms are used to enumerate molecular substructures centered on the atom (Figure 1). First, all paths up to a specified depth are extracted. In this study, the depth is chosen to be 4, 6, 8, or 10 bonds. The base fingerprint used in the initial XenoSite algorithm uses paths with a depth of up to eight. Path features like these are the basis Daylight fingerprints, a commonly used fingerprint in the literature.[16]

Second, circular substructures centered on each atom, up to a specified depth, are extracted. In this study, the maximum depth is chosen to be a depth of 2, 4, or 6 bonds (corresponding to substructure diameters of 5, 9, or 13 bonds). In this algorithm, each atom is a seed of a substructure that is grown by iteratively adding its neighboring atoms and bonds. Neighboring bonds are added first, so the degree of each atom is encoded in each substructure. Circle fingerprints are consistent with Extended Connectivity Fingerprints[18] and are commonly used to predict molecule activity but have not been explored for site of metabolism modeling. A more detailed description of circular fingerprints is in the literature and is apparent in the figure.

Third, all possible subgraphs containing the atom are enumerated that contain up to a specified depth number of bonds. In this study, the depth is chosen to be 4, 6, 8, or 10 bonds. By definition, subgraph substructures include all path substructures, but for branching molecules they contain additional substructures. Subgraphs are uncommonly used in chemical informatics but have been proposed.[19]

Molecule Substructures. To calculate a molecular fingerprint substructures are generated for each heavy atom of a molecule and are then combined. The same three algorithms with the same range of depth values used to calculate atom fingerprints are used to calculate molecular substructures and their corresponding molecular fingerprints. All atoms in the same molecule have the same molecular fingerprint, provided they were calculated using the same algorithm.

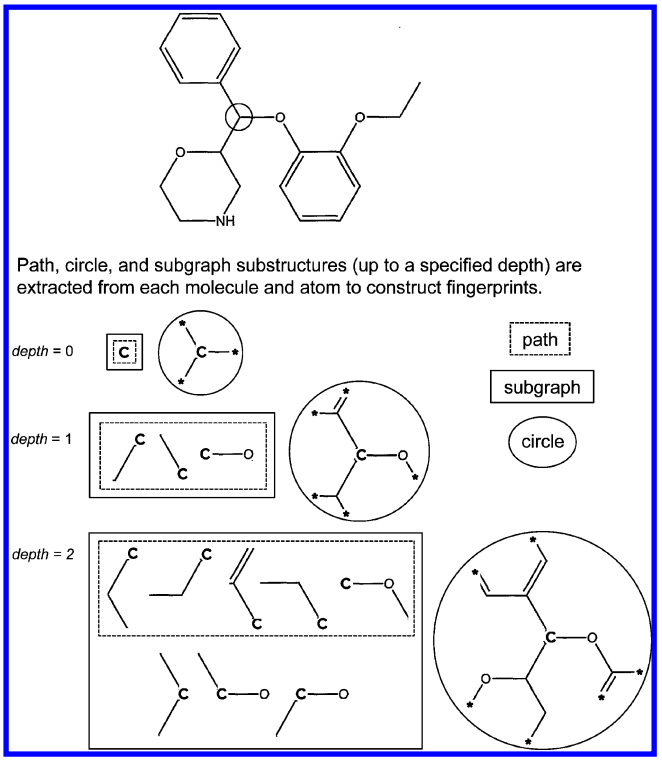


Figure 1. Different substructures are used for fingerprints. Different algorithms extract either paths, circular structures, or subgraphs. For the molecule at the top, all the substructures up to a depth of two, centered on the circled carbon atom, are shown in the figure. The paths are linear sequences of bonded atoms up to the specified depth of bonds from the starting atom. Circular substructures expand out from the root atom in all directions up to the specified depth. Subgraphs include all connected subgraphs containing the root atom with the specified number of bonds. Necessarily, all paths are included in subgraphs at a particular depth. Fingerprints include all substructures of the specified type from depth zero up to the chosen depth. Molecule fingerprints use all atoms as roots to generate substructures. Atom fingerprints use only the atom they describe as a root.

Canonicalization. Substructures are converted to strings using a canonicalization algorithm that ensures that subgraphs with identical connectivity always yield exactly the same string (Figure 2). The algorithm works by using Morgan's algorithm to assign a unique numeric ID to all vertices in the substructure's graph.[20] A depth first search across the graph, using the numeric IDs from Morgan's algorithm, prioritizes vertexes. This depth first search defines an ordering of atoms in the molecule, from which a SMILES-like representation is constructed, using parentheses to signal branches in the structure and digits to mark ring closures. For circular substructures, vertices not included in the substructure, but referenced by bonds that are included, are labeled with asterisks. The final string representation uniquely encodes the each substructure.

For atom substructures, a second canonicalization strategy was also considered than allows for the root atom to mismatch. Here, each substructure produces two strings. The first string is

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

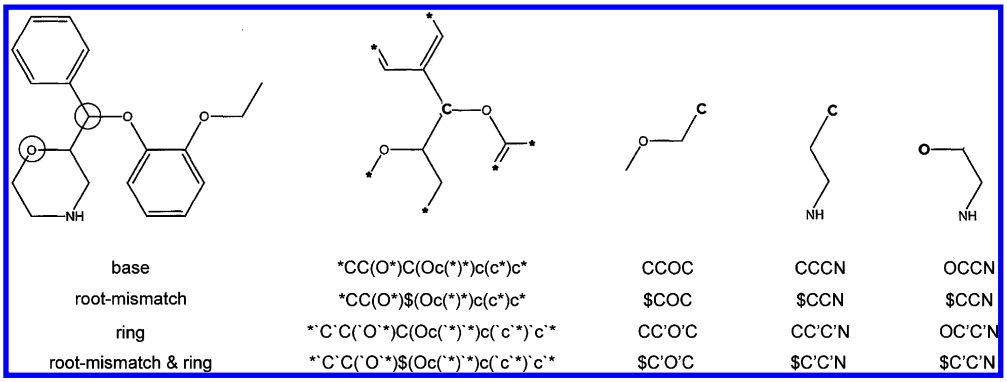


Figure 2. Canonized string representations of substructures. For the molecule at the left, three substructures rooted at the same circled carbon atom, and one at the circled oxygen atom, are displayed. The canonical string representations of both labelings (SMILES and SMILES with ring marks) and both canonicalization strategies (standard and allowing root-atom mismatches) are shown. When allowing root-atom mismatch, two strings are used to represent each substructure, both the standard string and the mismatch string. Strings follow a SMILES-like format with adjacent atoms bonded together, digits closing rings, and parentheses opening branches.

Table 2. Top-2 Accuracies of XenoSite Models Trained Using the Base Fingerprint, Optimal Fingerprint, Cross-Validated Optimal Fingerprints, and Consensus Cross-Validated Optimal Fingerprints[b]

parameter	values								
isoform	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
number of substrates	271	105	151	142	226	218	270	145	475
base fingerprint	85.6	83.8	81.5	85.9	85.8	85.8	87.0	84.1	85.1
optimal fingerprint	89.3	88.6	85.4	90.9	89.8	91.3	89.3	86.9	87.8
improvement	3.7	4.8	3.9	5.0	4.0	5.5	2.3	2.7	2.7
p-value[a]	**0.016**	**0.048**	**0.029**	**0.004**	**0.030**	**0.002**	**0.042**	0.125	**0.003**
CV fingerprint	86.0	87.6	79.5	87.3	89.8	90.8	84.4	80.7	83.2
improvement	0.04	3.8	−2.0	1.4	4.0	5.0	−2.6	−3.4	−1.9
p-value	0.414	*0.079*	0.841	0.241	**0.030**	**0.006**	0.974	0.917	0.953
consensus CV fingerprint	88.6	89.5	81.5	90.1	89.8	90.8	88.9	84.8	87.6
number of models	4	2	4	7	1	1	5	10	5
improvement	3.0	5.7	0.0	4.2	4.0	5.0	1.9	1.0	2.5
p-value	**0.037**	**0.029**	0.500	**0.017**	**0.030**	**0.006**	*0.066*	0.353	**0.007**

[a]These p-values are overly optimistic because they do not account for multiple testings comparing results of 483 different fingerprint models to the base fingerprint model. [b]P-values are computed with a one-sided paired t-test on the individual molecule predictions of the row fingerprint model and the base fingerprint model. Statistically significant p-values falling below 0.05 and 0.1 are respectively displayed in bold and italic text. The p-values for fingerprints based on their cross-validation performance (optimal fingerprint) have not been corrected for multiple-testing and are overly optimistic. The p-values for fingerprints chosen within each cross-validation fold (CV fingerprint) do not require multitest correction. The p-values for consensus fingerprint models chosen within each cross-validation fold (consensus CV fingerprint) do not require multitest correction, because only 10 closely correlated models are tested against the base model.

exactly as defined above. The second string is similar, but the root atom is labeled with a dollar sign '$'. Using a common symbol for all root atoms, in this way, enables substructures to match if their environments are similar even if their root element is different. We suspect that in some cases an atom's environment may provide important information that can extrapolate better if matches between different elements are allowed.

Data Structure. Finally, the strings representing each substructure are hashed to a numeric value and efficiently stored using lossless compression.[21] This final data structure is a fingerprint. Each atom is assigned two fingerprints: (1) a fingerprint containing the atom substructures and (2) a fingerprint containing the substructures extracted from the molecule the atom is within.

RESULTS

Here, we first measure the accuracy of fingerprints using cross-validation. Next, a more rigorous assessment tests our approach by selecting optimal fingerprints within each fold. Further improvements are studied by building consensus models that average the predictions of several models built using different fingerprints. Then, we study which fingerprint parameters are driving performance improvements. Finally, we present specific molecules to understand what types of molecules' predictions are improved.

To facilitate comparison with other published studies, the results presented in the main text are all based on variations of leave-one out cross-validation. There is sometimes concern, especially when using fingerprint-based approaches, that leave-one-out cross-validation substantially overestimates generalization accuracy. However, all results were replicated with 3-

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

Table 3. AUC Accuracies of XenoSite Models Trained Using the Base Fingerprint, Optimal Fingerprint, Cross-Validated Optimal Fingerprints, and Consensus Cross-Validated Optimal Fingerprints[b]

parameter	values								
isoform	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
number of substrates	271	105	151	142	226	218	270	145	475
base fingerprint	91.8	91.1	91.1	91.3	92.1	93.3	92.8	87.8	92.8
optimal fingerprint	92.8	92.7	92.9	93.3	93.4	94.5	93.9	89.9	93.3
improvement	1.0	1.6	1.8	2.0	1.3	1.2	1.1	2.1	0.5
p-value[a]	**0.034**	**0.007**	**0.017**	**0.033**	**0.014**	**0.022**	**0.039**	*0.062*	**0.002**
CV fingerprint	92.8	91.5	89.8	90.0	92.7	93.8	93.1	83.9	92.2
improvement	1.0	0.4	−1.3	−1.3	0.6	0.5	0.3	−3.9	−0.6
p-value	**0.034**	0.294	0.975	0.995	0.150	0.230	0.313	1.000	0.961
consensus CV fingerprint	92.8	92.1	91.4	92.1	93.0	94.2	93.7	88.4	93.1
number of models	1	3	5	9	8	5	3	8	2
improvement	1.0	1.0	0.3	0.8	0.9	0.9	0.9	0.6	0.3
p-value	**0.034**	*0.057*	0.337	0.127	*0.062*	*0.070*	**0.049**	0.249	*0.059*

[a]These p-values are overly optimistic because they do not account for multiple testings comparing results of 483 different fingerprint models to the base fingerprint model. [b]P-values are computed with a one-sided paired t-test on the individual molecule predictions of the row fingerprint model and the base fingerprint model. Statistically significant p-values falling below 0.05 and 0.1 are respectively displayed in bold and italic text. The p-values for fingerprints based on their cross-validation performance (optimal fingerprint) have not been corrected for multiple-testing and are overly optimistic. The p-values for fingerprints chosen within each cross-validation fold (CV fingerprint) do not require multitest correction. The p-values for consensus fingerprint models chosen within each cross-validation fold (consensus CV fingerprint) do not require multitest correction, because only 10 closely correlated models are tested against the base model.

fold cross-validation; performance was nearly identical, and all findings were similarly statistically significant. This is a key result because it demonstrates that our assessment is robust and not dependent on the exact cross-validation protocol.

Cross-Validated Performance. As a first assessment of each fingerprint, we measured their cross-validated accuracy. Here, each molecule is held, one at a time, from the training data, a XenoSite model is trained on the remaining data, and a prediction is made on the held out molecule. This process is repeated for all molecules in the data set, until all have model predictions assigned. The accuracy of these predictions is a good estimate of each fingerprint's accuracy on new data. Accuracy is measured in two ways. First, we measure the "Top-2" accuracy, the percentage of molecules for which at least one known site of metabolism is assigned one of the top two predictions in the molecule. Second, we measure the area under the ROC curve (AUC), a common metric used in machine learning that measures how well separated the sites of metabolism are from sites that are not metabolized. The AUC metric is computed for all molecules in the data set individually and averaged to quantify the global performance.

Using both Top-2 (Table 2) and AUC (Table 3), we find that individual fingerprints often substantially outperform the base fingerprint used in the original XenoSite model. Though no single fingerprint yielded optimal models across multiple isoforms, the optimal fingerprint for each isoform yielded considerably better results much higher than the base fingerprint initially implemented in XenoSite. Optimal fingerprint models are on average 3.8% (Top-2) and 1.4% (AUC) more accurate than base fingerprint models. These gains represent, respectively, a 25.6% and 16.7% reduction in error. For 2C19, 2C8, and 2C9, improvements are even more substantial.

For all but one isoform (2E1), these improvements over the base fingerprint are statistically significant with uncorrected p-values. The AUC p-value is computed with a a one-sided, paired t test on the pool of molecule AUCs associated with each fingerprint and the base fingerprint. The Top-2 p-value is computed with a one-sided, paired t test on the pool of 1s and 0s (encoding, respectively, correct and incorrectly predicted molecules) associated with each fingerprint and the base fingerprint. Moreover, for each isoform and metric, often multiple fingerprints led to significant improvements over the base model (Table 4).

Table 4. Number of Models with a Significantly Improved Top-2 and AUC Performances over the Base Model When Using a Paired t-Test with a Cutoff of 0.05%

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
significant by AUC	1	11	30	67	76	40	13	8	16
significant by Top-2	8	2	16	11	1	94	7	0	29

However, in these experiments, 483 different fingerprints are compared with the base fingerprint. This is a classic multiple testing scenario, where a multitest correction is needed to interpret these results. Applying either the Bonferroni correction or False Discovery Rate shows that none of these improvements are statistically significant. This is a fundamental challenge inherent to testing a large number of models with a smaller number of substrates. In addition, as models become more accurate, only small differences in performance distinguish the best models, so statistical significance between them becomes harder to achieve.

Nonetheless, a refined assessment strategy—as presented in the following sections—can overcome these limitations in some cases.

Fingerprint Selection within Cross-Validation. A more rigorous assessment selects fingerprints within each cross-validation fold, sometimes selecting different fingerprints for

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

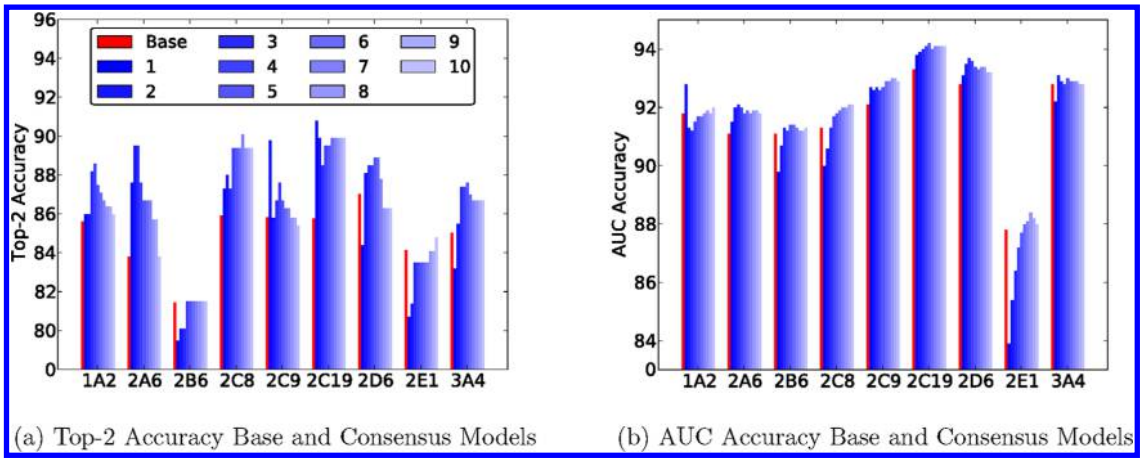


Figure 3. Top-2 (3a) and AUC (3b) accuracies of base and consensus XenoSite models. For each isoform the accuracy of the base XenoSite model is shown in red, and accuracies of consensus models derived from models with the top 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 best performances on the training set are shown from left to right in successively lighter shades of blue.

different folds, and yields a more reliable estimate of generalization error. In this paradigm the fingerprint model that performs best for the training set substrates is used to predict the test substrate. Here, the performance of individual fingerprints is not assessed. Rather, this approach assesses the entire procedure, including the selection of fingerprints based on training set accuracy.

Cross-validation selected fingerprint models improve over the base model for five on the nine isoform sets using Top-2 and AUC metrics. However, only three of these improvements—1A2 (as measured by AUC) and 2C9 and 2C19 (as measured by Top-2)—are statistically significant using a one-tailed, paired *t* test. Overall, cross-validated fingerprint selection performance is more similar to the base fingerprint than to the optimal fingerprint in the prior section. The fall off in accuracy is expected. Part of the reason performance drops this substantially is because so many fingerprints are being tested. In the first assessment from the prior section, some of the best performing fingerprints are overfit to the training data. The fact that this performance drop reduces accuracy below the threshold for statistical significance is disappointing.

Consensus Models. In the prior sections, we find first that some fingerprints substantially improve over the accuracy of the base fingerprint when using leave one molecule out cross-validation. Unfortunately, in a more rigorous test that pushes the selection step into the cross-validation folds, we find that most of this improvement evaporates. Therefore, we explored the performance of consensus or ensemble models that average the predictions from the top performing fingerprint models (by either Top 2 or AUC accuracy on each cross-validation fold's training set). The consensus prediction relies on multiple optimal and near-optimal models with statistically equivalent performances. We considered averaging the predictions of models with the top 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 best performances for each fold's training set. Though the overall model accuracies are often statistically indistinguishable, each model may predict different molecules accurately. We expect that averaging out these differences will yield higher performing models that are protected from overfitting to training data.

The consensus model improves on the base fingerprint in all cases, and almost all of these improvements are statistically significant (Tables 2 and 3). In fact, consensus fingerprint models perform nearly as well for each substrate set as the optimal fingerprint model. Consensus performances are within one percent of the optimal fingerprint model for all but two isoforms (2B6 and 2E1) using a Top-2 metric and all but three isoforms using an AUC metric (2B6, 2C8, and 2E1). In one case (2A6), the consensus model improves over the optimal fingerprint model (by one percent in Top-2 performance).

The number of best performing models that result in an optimal Top-2 performing consensus model varies by isoform (Figure 3). For 2C9 and 2C19 incorporating suboptimal model predictions with the optimal model results in poorer performance. In contrast, for 2B6 and 2E1 isoforms the more models that are incorporated into the consensus predictions the better the performance. The general trend for the remaining isoforms (1A2, 2A6, 2C8, 2D6, and 3A4) is that using the predictions of the top four best performing models results in optimal or near optimal consensus predictions, which are on average two percent more accurate than base model predictions. Including predictions from six or more of the best performing models generally results in a drop in consensus model performance for those same isoforms.

In general, optimal, or near optimal, consensus results are achieved for AUC performance when the predictions of the best four models are utilized. These results do not significantly increase or decrease when the predictions of additional near optimal models are incorporated, though there are slight variations between isoforms. For 2A6, 2B6, 2C8, 2D6, 3A4, and especially 2E1, incorporating the predictions of the second and third best performing models results in a significant improvement in AUC performance over using just the optimal model. In contrast, for 1A2, utilizing additional models beyond the optimal cross-validated model results in a significant drop in consensus AUC performance, the only isoform for which this occurs.

Best Fingerprint Parameters. For the optimal consensus model of each isoform we determined which fingerprints were used to make cross-validated predictions. For the interested reader a complete listing of these fingerprints is provided in the Supporting Information. Encouragingly, the most visible trend is that the best fingerprints, meaning those with highest cross-validated accuracy for the given isoform data set, are highly conserved during cross-validation. Recall, each cross-validation fold builds its own consensus model, using its own collection of fingerprints. The consistent trend for each isoform is that there

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

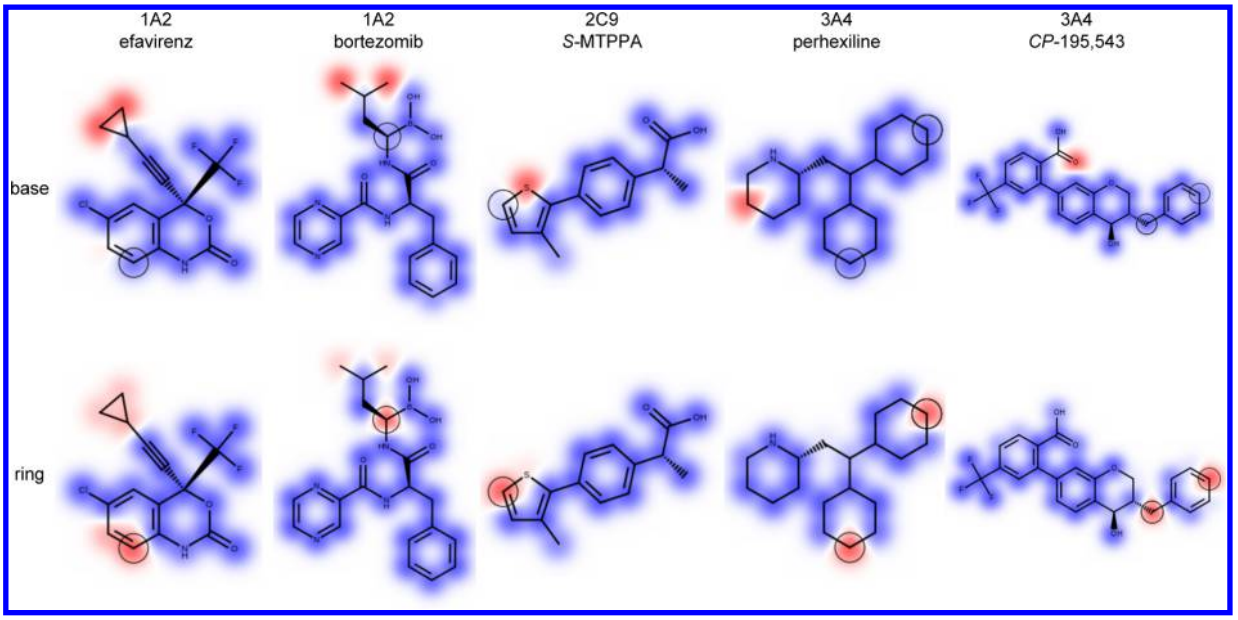


Figure 4. Examples where extending the base fingerprint with the ring SMILES encoding results in improved predictions. The observed sites of metabolism of each substrate are circled. Atoms predicted to be sites of metabolism are colored red, while those not predicted to be sites of metabolism are colored blue. The IUPAC name of *S*-MTPPA is *S*-2-[4-(3-methyl-2-thienyl)phenyl]propionic acid and of *CP*-195,543 is (+)-2-(3-benzyl-4-hydroxy-chroman-7-yl)-4-trifluoromethylbenzoic acid.

is one set of fingerprints used to make consensus predictions for over 90% (and often 100%) of the data set, and another set of fingerprints is used to make predictions for less than 10% (and often 2%) of the data set. This suggests an underlying robustness to our optimization procedure.

Unfortunately, the fingerprints that yield optimal consensus models vary significantly between isoforms, as well as the evaluation metric. Only for 2A6 are two of the same fingerprints used in both Top-2 and AUC consensus models. Top-2 and AUC consensus models for both 2C9 and 2B6 have just one overlapping fingerprint between them. Other than these cases, no fingerprint is used in both metric consensus models for any isoform. This indicates that the "best" fingerprint for a given isoform depends on whether the end-user is more concerned with identifying for a candidate molecule at least one of its sites of metabolism (for which we would recommend Top-2) or maximizing the predictive separation of all P450-oxidized atoms from all nonoxidized atoms (for which we would recommend AUC).

The next most visible trend is that atom fingerprints using circle features at a depth of 2 or 4 are frequently used in AUC consensus models, while atom fingerprints calculated with path features at a depth of 8 or 10 are consistently used in Top-2 consensus models. There are some exceptions; consensus Top-2 models for 2A6 and 2C19 employ circle features for atoms and for 2B6 and 3A4 employ subgraph and path features. This suggests that the local connectivity of atoms encoded in low-depth circular structures tends to give better overall separation between sites of metabolism and sites that are not metabolized. In contrast, paths with a high depth are better able to predict at least one site of metabolism of a molecule quite well but are either not as effective at identifying other sites of metabolism for that molecule or give higher scores to atoms that do not undergo P450-mediated metabolism.

There are also several weak trends in molecule-level fingerprints that may influence fingerprint selection and design. In direct contrast to our observations of atom-level fingerprints, consensus model molecule-level fingerprints rarely employ circle features. The only exceptions are consensus models for 2A6 and 2C9 using Top-2 and AUC metrics and 2E1 using the AUC metric. Other than these instances, there is a high degree of variation across isoform and metric consensus models as to which molecule substructures (path or subgraph) and depth is optimal. Similarly, the new ring extension to labels and canonicalization strategy that enables root-atom mismatches exhibit some weak trends. The ring labeling extension is included in the majority of AUC consensus models. Both variations together improve AUC performances for 2C9, 2C19, and 2D6 models, as well as the Top-2 performance of 2D6 models. Both encodings are used consistently in 2C19 AUC consensus models, while the ring encoding consistently improves 1A2 Top-2 consensus models.

Improvements to Specific Predictions. Aggregate performance increases with optimized fingerprints, and we can see these improvement in the predictions for specific molecules.

When the base fingerprint is extended with the ring SMILES labeling, more sites of metabolism located on aromatic, nonaromatic, and thiophene rings are correctly predicted (Figure 4). An especially interesting case is bortezomib, which is the only molecule in the data that contains a boron, and also undergoes P450-mediated deboronation.[22] Using ring SMILES labeling with a path feature of depth eight is the first model we have developed able to correctly predict this reaction. This shows ring SMILES labeling helps to identify not only sites of metabolism located on or near chemical rings but also sites farther removed from rings, whose metabolism is, to some degree, still influenced by them. Incorporating root-mismatch canonicalization in conjunction with ring SMILES labeling to the base fingerprint gives a tighter prediction for efavirenz, bortezomib, and *CP*-195,543, meaning the false-negative sites have lower scores, but not *S*-MTPPA and perhexiline. Still, it is the ring SMILES labeling that enables the correct prediction of all these molecules.

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

There are relatively few instances where root-mismatch canonicalization alone gives improved predictions over the base model (Figure 5). In two cases, sites of metabolism on fuzed

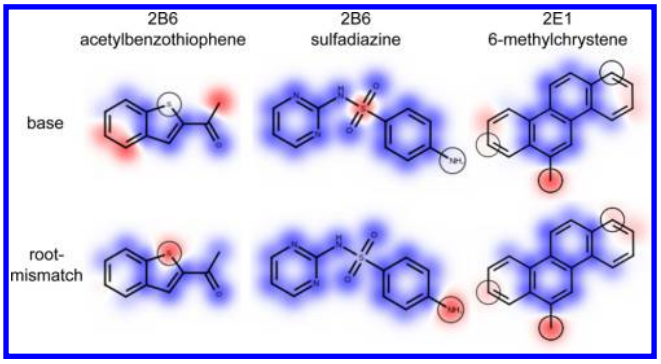


Figure 5. Examples where extending the base fingerprint with root-atom canonicalization results in improved predictions. The observed sites of metabolism of each substrate are circled. Atoms predicted to be sites of metabolism are colored red, while those not predicted to be sites of metabolism are colored blue.

ring systems were better predicted using root-mismatch canonicalization. Fuzed-ring systems with similar steric constraints and electronic environments likely have specific atomic locations vulnerable to metabolism that makes the specific element occupying those locations less important. Similarly, root-mismatch canonicalization helps to predict the benzenamine hydroxylation of sulfadiazine[23] by identifying similar methylbenzene and phenol hydroxylation reactions contained in the training set (e.g., 6-methylchrystene in the same figure).

Our results show that consensus models have improved predictions over the base model for a significant number of molecules (Table 2 and 3), nine of which are shown here (Figures 6 and 7). Consensus models appear to be better at predicting sites of aromatic ring hydroxylation, a reaction that multiple methods from different groups have had difficulty predicting in the past.[24] Illustrating this point, seven of the nine example molecules have sites of aromatic ring hydroxylations that are incorrectly predicted by the base model. In four of these instances (*o*-aminoazotoluene, propofol, licofelone, and chlorzoxazone) consensus models identify actual sites of P450-mediated aromatic ring hydroxylations that are not predicted by the base model. Meanwhile, there are four molecules (5-hydroxytriptamine, 8-hydroxy-efavirenz, licofelone, and clobazam) where the base model incorrectly predicts multiple atoms to be metabolized, none of which are actual sites of metabolism, that consensus models predict perfectly. In each of these cases the base model makes an incorrect prediction for sites of aromatic ring hydroxylation. The consensus model also improves over the base model in the prediction of isopropyl hydroxylation (flutamide and propofol).

DISCUSSION

XenoSite was initially developed using a fingerprint that was known to work well in other domains but had not been optimized for P450 site of metabolism modeling. Our comprehensive investigation of the performance 484 fingerprints across nine P450 isoforms identifies several fingerprints that significantly outperform the base fingerprint that XenoSite initially used. Optimal fingerprint models chosen by cross-validation performance on the full training data are, on average, 3.8% (Top-2; percent of molecules with a site of metabolism in the top two predictions) and 1.4% (AUC; area under the ROC curve) more accurate than base fingerprint models. These gains represent, respectively, a 25.6% and 16.7% reduction in error. A more rigorous assessment selects fingerprints within each cross-validation fold, sometimes selecting different fingerprints for different folds, and yields a more reliable estimate of generalization error. Consensus models were built by averaging the predictions of the top performing cross-validated fingerprint models. Consensus models improve on the base model, on average, 3.0% (Top-2) and 0.7% (AUC). These gains are statistically significant and represent, respectively, a 20.1% and 8.8% reduction in error.

Unfortunately, there are no strong patterns governing which fingerprints are best, each isoform uses different fingerprints in

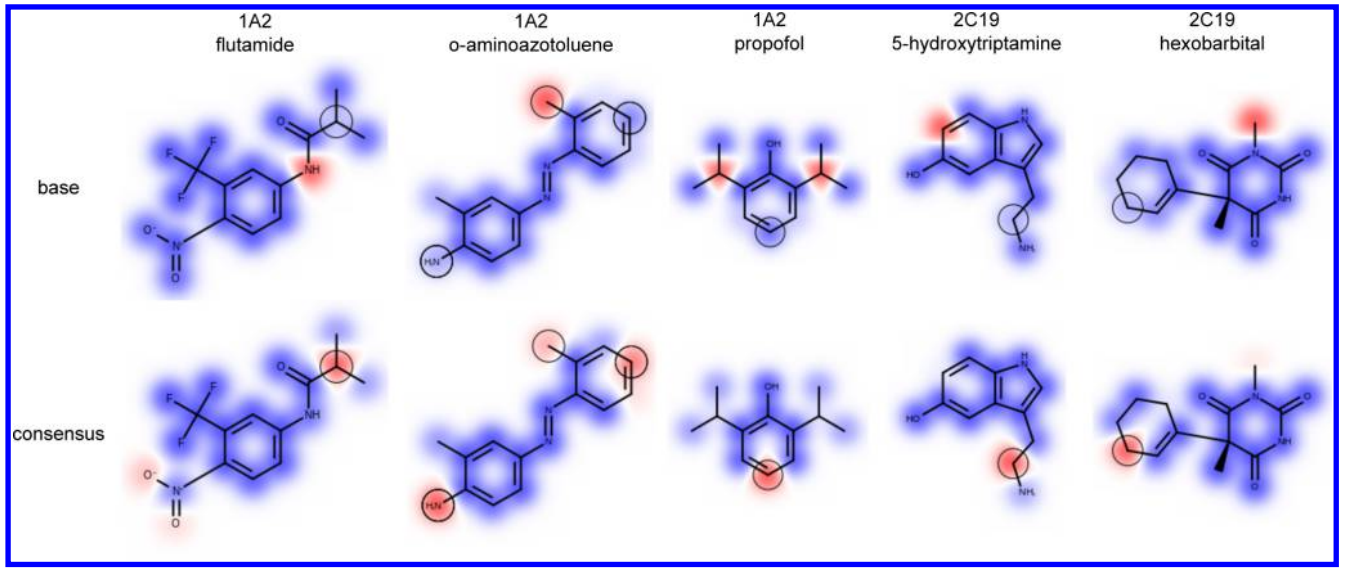


Figure 6. Examples where consensus models have improved predictions over the base model. The observed sites of metabolism of each substrate are circled. Atoms predicted to be sites of metabolism are colored red, while those not predicted to be sites of metabolism are colored blue.

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

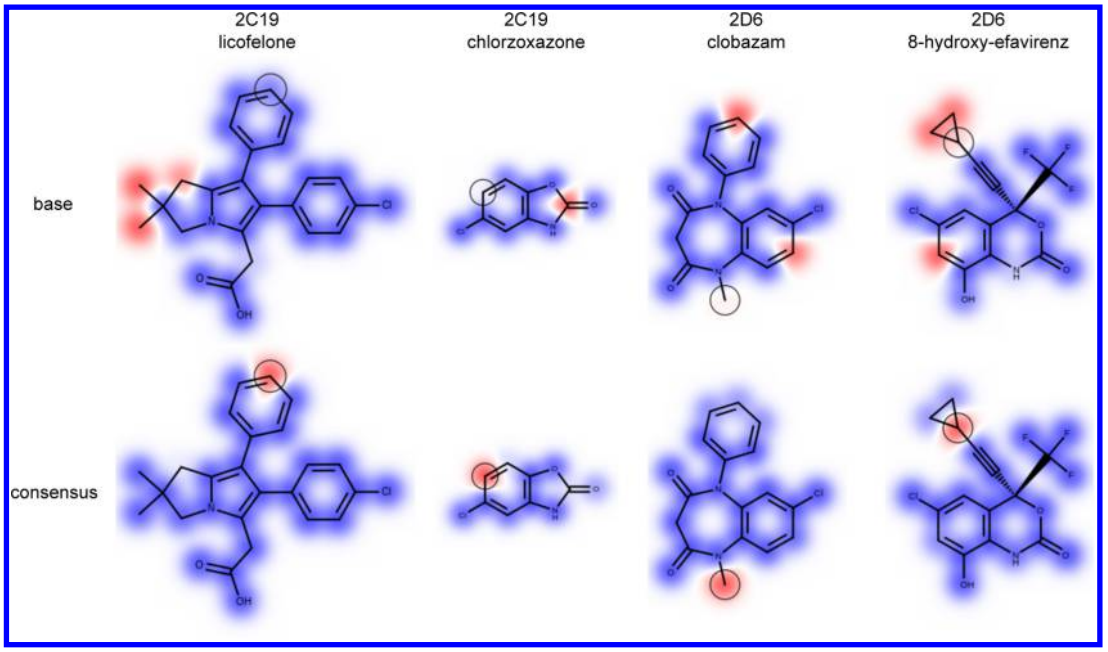


Figure 7. Examples where consensus models have improved predictions over the base model. The observed sites of metabolism of each substrate are circled. Atoms predicted to be sites of metabolism are colored red, while those not predicted to be sites of metabolism are colored blue.

Table 5. Fingerprints We Recommend Using To Build Consensus XenoSite Models[a]

isoform	LOO CV accuracy	ring	root-atom mismatch	atom substructure	atom depth	molecule substructure	molecule depth
1A2	92.8	n	y	circle	4	subgraphs	4
2A6	92.7	y	y	circle	2	circle	6
	92.7	y	y	circle	2	circle	4
	92.7	y	n	circle	2	circle	4
	92.7	y	y	circle	2	circle	2
2B6	92.9	y	y	paths	10	subgraphs	2
	92.8	y	n	circle	4	paths	6
	92.8	y	n	subgraphs	2	subgraphs	2
	92.7	n	n	circle	4	paths	4
2C8	93.3	y	y	circle	2	subgraphs	4
	93.1	n	y	circle	4	paths	8
	93.1	y	n	circle	4	circle	6
	93.1	y	y	circle	4	circle	6
2C9	93.4	y	n	circle	2	circle	6
	93.4	y	y	circle	2	circle	6
	93.4	y	n	circle	2	circle	4
	93.3	n	y	subgraphs	2	circle	6
2C19	94.5	y	y	circle	4	paths	6
	94.5	y	y	circle	4	subgraphs	4
	94.5	y	y	circle	6	subgraphs	4
	94.5	y	y	circle	4	subgraphs	4
2D6	93.9	n	y	circle	6	circle	2
	93.8	n	y	circle	4	paths	10
	93.8	n	y	circle	4	subgraphs	8
	93.7	n	y	circle	4	paths	8
2E1	89.9	y	n	circle	4	circle	2
	89.9	n	y	subgraphs	4	circle	6
	89.8	n	y	circle	6	subgraphs	8
	89.8	y	n	circle	4	circle	4
3A4	93.3	n	y	paths	6	subgraphs	4
	93.3	y	n	paths	6	subgraphs	4
	93.3	n	n	paths	6	subgraphs	8
	93.3	y	y	paths	6	paths	8

[a]Models using these fingerprints are available for public use at our Web site.

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

its optimal consensus models. There is a large variation in the types and number of fingerprints in optimal consensus models for each isoform. This result indicates that, although significant performance gains are achievable by tuning fingerprints, the best fingerprints need to be chosen separately for every data set.

Moreover, the optimal fingerprints depend on whether Top-2 or AUC accuracy is used to choose them. In only four instances was the same chemical fingerprint used in both Top-2 and AUC consensus models of the same isoform. It is possible that better strategies for selecting fingerprints to use in the consensus model could optimize AUC and Top-2 accuracy, but this was not considered in the current study and left for future work. Until then, these results suggest that the ideal model depends on choosing which performance metric is most important in practice. In our view, AUC is a better metric because it considers the predictions of all atoms in a model instead of just the top performing known site of metabolism.

Therefore, our final recommendation is that the four best performing fingerprints for each isoform using the AUC metric should be used to build a consensus model (Table 5). We make an exception with 1A2 to only recommend a single optimal fingerprint, because incorporating suboptimal fingerprint predictions always results in poorer AUC performance. Our rational for choosing fingerprints that optimize AUC over Top-2 accuracy is that we believe a better overall separation between the predictions of metabolized sites versus not-metabolized sites is more important to end-users than predicting a single site of metabolism correctly. In addition, there is significant variation in Top-2 performance across different isoforms depending on the number of models used to make consensus predictions, while AUC accuracies are relatively stable once the predictions of at least four optimal AUC models are used. A Web server has been created to let end-users use the consensus XenoSite fingerprint models built with these recommendations to predict the P450-mediated metabolism of a set of input molecules.

■ ASSOCIATED CONTENT

Supporting Information

This includes (1) a brief description of topological, quantum chemical, molecular, and SMARTCyp descriptors used in this work; (2) a table comparing base XenoSite model performances using descriptors calculated in previous work from proprietary software and in this work from open-source software; a replication of key results using 3-fold cross-validation; and (4) a comprehensive list of all fingerprints used to calculate predictions for optimal consensus models. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/ci5005652.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: swamidass@gmail.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Matt Matlock for help with fingerprint calculation, Tyler Hughes for help with generating figures, and Na Le Dang for editing.

■ REFERENCES

(1) Nebert, D. W.; Russell, D. W. Clinical Importance of the Cytochromes P450. *Lancet* **2002**, *360*, 1155−1162.
(2) Guengerich, F. P. Cytochrome P450s and Other Enzymes in Drug Metabolism and Toxicity. *AAPS J.* **2006**, *8*, E101−E111.
(3) 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.
(4) Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical Regioselectivity Models for Human Cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173−3184.
(5) Zaretzki, J.; Matlock, M.; Swamidass, S. J. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. *J. Chem. Inf. Model.* **2013**, *53*, 3373−3383.
(6) 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.
(7) SMARTCyp Web Service - About. http://www.farma.ku.dk/smartcyp/about.php (accessed June 20, 2013).
(8) *StarDrop*, Version 4.3; Optibrium Ltd.: Cambridge, United Kingdom, 2009.
(9) *P450 SOM Prediction*, Version 1.0; Schrödinger LLC: New York, NY, 2011.
(10) Tyzack, J. D.; Mussa, H. Y.; Williamson, M. J.; Kirchmair, J.; Glen, R. C. Cytochrome P450 Site of Metabolism Prediction from 2D Topological Fingerprints Using GPU Accelerated Probabilistic Classifiers. *J. Cheminform.* **2014**, *6*, 29.
(11) Stewart, J. J. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−103.
(12) 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.
(13) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093−1110.
(14) Azencott, C.-A.; Ksikes, A.; Swamidass, S. J.; Chen, J. H.; Ralaivola, L.; Baldi, P. One-to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties. *J. Chem. Inf. Model.* **2007**, *47*, 965−974.
(15) *MOE*, Version 2009.10; Chemical Computing Group: Montreal, Canada, 2009.
(16) James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; Daylight Chemical Information Systems; 2005.
(17) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES.2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(18) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(19) Liu, P.; Agrafiotis, D. K.; Rassokhin, D. N. Power Keys: A Novel Class of Topological Descriptors Based on Exhaustive Subgraph Enumeration and Their Application in Substructure Searching. *J. Chem. Inf. Model.* **2011**, *51*, 2843−2851.
(20) Morgan, H. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(21) Baldi, P.; Benz, R. W.; Hirschberg, D. S.; Swamidass, S. J. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 2098−2109.
(22) Uttamsingh, V.; Lu, C.; Miwa, G.; Gan, L.-S. Relative Contributions of the Five Major Human Cytochromes P450, 1A2, 2C9, 2C19, 2D6, and 3A4, to the Hepatic Metabolism of the Proteasome Inhibitor Bortezomib. *Drug. Metab. Dispos.* **2005**, *33*, 1723−1728.
(23) Pfeifer, T.; Tuerk, J.; Fuchs, R. Structural Characterization of Sulfadiazine Metabolites Using H/D Exchange Combined with Various MS/MS Experiments. *J. Am. Soc. Mass. Spectrom.* **2005**, *16*, 1687−1694.

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982

(24) 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.

DOI: 10.1021/ci5005652
J. Chem. Inf. Model. 2015, 55, 972−982