JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

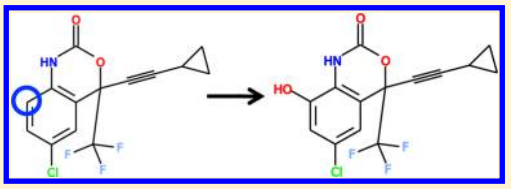
XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks

Jed Zaretzki, Matthew Matlock, and S. Joshua Swamidass*

Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63130, United States

Supporting Information

ABSTRACT: Understanding how xenobiotic molecules are metabolized is important because it influences the safety, efficacy, and dose of medicines and how they can be modified to improve these properties. The cytochrome P450s (CYPs) are proteins responsible for metabolizing 90% of drugs on the market, and many computational methods can predict which atomic sites of a molecule—sites of metabolism (SOMs)—are modified during CYP-mediated metabolism. This study improves on prior methods of predicting CYP-mediated SOMs by using new descriptors and machine learning based on neural networks. The new method, XenoSite, is faster to train and more accurate by as much as 4% or 5% for some isozymes. Furthermore, some "incorrect" predictions made by XenoSite were subsequently validated as correct predictions by revaluation of the source literature. Moreover, XenoSite output is interpretable as a probability, which reflects both the confidence of the model that a particular atom is metabolized and the statistical likelihood that its prediction for that atom is correct.



INTRODUCTION

The cytochrome P450 (CYP) enzymes are a family of proteins responsible for the metabolism of approximately 90% of FDA-approved drugs.[1,2] CYPs catalyze a number of different types of reactions—aliphatic hydroxylation and N-oxidation, for example—to alter xenobiotic molecules into more hydrophilic readily excreted molecules. The metabolism of a drug affects both efficacy and safety. If a drug is metabolized at too many locations or at too fast a rate, it will not stay in the body long enough to reach its target and have a clinical impact. Meanwhile, a drug that is metabolized into a toxic product will have a low therapeutic index and will need to be administered carefully. A classic example of this is the CYP-mediated metabolism of acetaminophen into NAPQI, a compound that leads to fatal hepatocyte damage when catalyzed in sufficient quantities.[3]

During the early stages of drug discovery, medicinal chemists are interested in knowing how each candidate lead will be metabolized by particular CYP isozymes. Knowing which atoms are oxidized by a CYP enzyme enables medicinal chemists to modify portions of the molecule to modulate its metabolism to improve its safety and efficacy. Unfortunately, it is not cost-effective to explicitly determine the CYP-mediated site(s) of metabolism (SOMs) for all viable lead candidates through experimental techniques. Consequently, there have been a number of in silico methods developed in recent years to computationally predict a compound's CYP-mediated metabolism using just its molecular structure. These methods can be applied at any stage of the drug discovery process to identify metabolically labile regions of candidate leads.[4]

This paper focuses on models that predict CYP-mediated SOMs of candidate substrates. If a molecule is metabolized by a specific CYP, SOM prediction models identify which specific atoms in the molecule are oxidized by the CYP.

In earlier work, performed by the first author at Rensselaer Polytechnic Institute in Curt Breneman's Lab, Regioselectivity-Predictor (RSP) was reported to outperform three other CYP SOM prediction models on the most comprehensive public data set of CYP substrates and metabolites.[5] RSP had an average prediction accuracy of 84% for 680 substrates distributed among nine specific CYP isozymes. The prediction accuracy of a given method was gauged using a traditional metric, whereby a substrate is considered to have been correctly predicted by a method if any of its experimentally verified SOM(s) are predicted by that method in the top two rank-positions. RSP builds a model by representing the potential SOMs contained within a set of known isozyme substrates with chemical descriptors that describe atom-level topological and electronic information; next, a customized implementation of support vector machines (SVMs) technology is used to derive a model that optimizes the ranking of experimentally verified SOMs over non-verified SOMs on a substrate by substrate basis. With performance levels higher than other SOM-prediction models for the largest set of CYP substrates and metabolites publicly available, the RSP algorithm represents a good starting point for those looking to develop new more accurate methods as others are also attempting.[6]

This work describes XenoSite, a CYP SOM prediction model based on neural networks that improves on RSP in multiple ways. XenoSite uses the substrate and descriptor sets generated by RSP in our earlier work as a starting point and makes the following enhancements: (1) New molecule-level descriptors

Received: September 5, 2013
Published: November 13, 2013

dx.doi.org/10.1021/ci400518g | J. Chem. Inf. Model. 2013, 53, 3373−3383

are developed that let machine-learning methods internally determine which atomic descriptors are the most relevant for the particular substrate being predicted. (2) Neural networks are used to build models rather than the SVM technology employed by RSP. One advantage of neural networks is that they have much quicker runtimes for training models than SVMs. A second advantage is that their SOM output is in the form of oxidation scores that can be treated as probabilities, as opposed to the SVMs of RSP that only predict a rank-ordering of the SOMs contained in the same substrate. A neural network-derived SOM score strongly correlates to the likelihood that the SOM is oxidized, while the SOM score derived from RSP rank-orderings does not. XenoSite scores therefore serve as a proxy for both the predictions confidence of the model and the accuracy of the prediction, meaning that end-users can look at the SOM scores for an entire substrate and make an informed decisions as to whether the prediction is reliable or suspect.

METHOD

Data Sets. In this work, we use a previously assembled set of 680 CYP substrates distributed across nine CYP enzymes: 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, and 3A4.[5] In addition, a human liver microsome (HLM) set is analyzed, whereby all 680 substrates and all observed metabolites, regardless of the metabolizing isoform, are considered. This HLM set does not represent all metabolic functions of a liver microsome but does represent an aggregation of known CYP metabolism.

Before any models were built, the observed metabolites for nordexfenfluramine in the 1A2 and HLM sets and sparteine in the 2D6 and HLM sets were updated based on source literature.[7,8] After models were built and their results were analyzed, errors in the source data for efavirenz in 2A6 and HLM sets (Figure 4) and verapamil in 3A4 and 2C8 sets (Figure 5) were identified and fixed. Up-to-date SDF files of the 10 substrate sets used in the work are included in the Supporting Information.

Descriptors. In a molecule, every atom capable of being metabolized in a CYP substrate is a potential SOM. Each atom is associated with a vector of numbers, with each number encoding a chemical property of that SOM; these encodings of chemical information are known as descriptors. Machine-learning algorithms then analyze these descriptor encoded SOMs to determine a scoring function that gives experimentally observed CYP-mediated SOMs high scores and non-observed SOMs low scores. We use a combination of previously defined descriptors—topological (TOP) and quantum chemical (QC) descriptors, a SMARTCyp reactivity (SCR) descriptor—in addition to a refined subset of the QC descriptors (SQC) and a new molecule-level (MOL) and fingerprint similarity (FP) descriptors. The MOL and FP descriptors are applied for the first time to SOM prediction and encode information about molecules as a whole in addition to the local atomic environment.

Previously Validated Descriptors. This study includes the 148 topological (TOP) and 392 quantum chemical (QC) descriptors developed in RSP. Full definitions of these descriptors can be found in our earlier work.[9] This study also includes DFT-derived SOM transition-state barriers encoded in SMARTCyp as an additional descriptor.[10] SMARTCyp reactivity (SCR) from version 1.5 of the software is incorporated as a SOM descriptor in XenoSite.

Subset of Quantum Chemical Descriptors. The 392 QC descriptors in RSP are highly correlated to one another, and it is possible that the signal represented in 392 descriptors of each SOM contain noise that can drown out the signal represented by descriptors in other classes. Each SOM has associated with it a set of all substrate SOMs one bond length away from the given SOM and a second set of all substrate SOMs more than one bond length from the given SOM. To calculate QC descriptors, each SOM is described by 24 specific quantum chemical properties (e.g., Fukui reactivity or nucleophilicity). For each specific chemical property the min, max, mean, norm, and sum of the property values for the atoms contained in the two associated SOM sets are calculated and mapped back to the source SOM as a 10 unique QC descriptors.

We created a subset of quantum chemical descriptors (SQC) by removing the min, max, norm, and sum descriptors from consideration during modeling. The objective in developing this descriptor set was to retain the relevant reactivity signal from the original QC descriptor class while minimizing information redundancy and potentially reducing noise in the training set. As we will see, replacing QC descriptors with SQC descriptors generally results in improved model performance.

Molecule Descriptors. All currently proposed machine-learning SOM prediction methods describe each potential SOM with descriptors that encode atom-level information but not molecule-level information. This is because the experimental response being predicted is atom-based, and models must be able to assess the chemical difference between all atoms in same molecule to determine which of them undergo CYP-mediated oxidation and which do not. However, these models have a blind-spot; the atom-level chemical rules that govern CYP-mediated metabolism are not identical for all substrates.

For example, it is possible that an individual atom's reactivity with a CYP's heme is more predictive for small molecules than large molecules. This hypothesis is supported by recent improvements in MetaSite, which employs a manually derived rule for using the size of a substrate and the size of a CYP binding pocket to weight the importance of local atom reactivity to predicting the CYP-mediated metabolism of that substrate.[11] This makes biophysical sense. Smaller substrates can rotate freely in the CYP binding pocket. Therefore, all their potential SOMs will have equal access to the oxidizing CYP heme, and local atomic reactivity will likely be the most relevant piece of chemical information to identify which of the substrate's atoms are oxidized. In contrast, larger substrates are likely to be more rigidly constrained in the binding pocket. Therefore, their potential SOMs will not all have equal access to the CYP heme, so an individual atom's reactivity may not be as important as, for example, its position in the molecule. Including molecule-level descriptors enable CYP-metabolism prediction models to take molecular size into account when determining the relative importance of local SOM reactivity to the metabolism of each specific substrate.

The example above describes just one instance where a single piece of molecule-level chemical information, molecular size, can be used to better ascertain the relative importance of a single piece of atom-level information, electronic reactivity, in order to build more accurate CYP metabolism prediction models. Given the large number (541) of atom-level descriptors employed in RSP, we hypothesized there are other less intuitive relationships between molecule-level and atom-level descriptors that can be derived to build even more accurate models.

To test this hypothesis, a small set of 15 molecular descriptors (MOL) was used to represent each potential SOM in our substrate sets, with the value of each MOL descriptor being identical for all SOMs contained in the same substrate (Table 1). These descriptors were chosen to represent chemical features suspected to strongly impact drug metabolism, such as molecular size, solubility, flexibility, and polar surface area, and were generated using the molecular operating environment (MOE).[12] As described in the Results section, neural networks are able to derive non-linear relationships that identify which atom-level descriptors are most pertinent to predicting the metabolism of a given substrate using the molecule-level descriptors that describe that substrate, resulting in more accurate CYP SOM prediction models.

Table 1. Molecule-Level Descriptors Calculated from MOE[a]

descriptor code	definition
diameter	largest vertex eccentricity
BCUT_PEOE_0	PEOE charge BCUT (0/3)
SlogP	octanol/water partition coefficient[b]
logP(o/w)	octanol/water partition coefficient[c]
logS	solubility in water
BCUT_SLOGP_0	logP BCUT (0/3)
b_1rotN	number of rotatable single bonds
KierFlex	molecular flexibility
apol	sum of atomic polarizabilities
ASA	water accessible surface area
ASA_P	total polar surface area
TPSA	topological polar surface area
vsurf_R	surface rugosity
glob	molecular globularity
vol	van der Waal's volume

[a]Complete descriptions of these descriptors can be found in the MOE documentation.[14] [b]Model was trained from 7000 molecules by Crippen et al.[13] [c]Model was trained from 1827 molecules internal to MOE.[14]

Fingerprint Descriptors. We also explored using atom-fingerprint similarity as a descriptor. In our earlier work, combining fingerprint similarity with neural networks substantially improved the performance of virtual screening methods,[15] and we hypothesized that it might also improve the accuracy of SOM prediction.

We compute the similarity between potential SOMs and experimentally verified SOMs in the training set using fingerprints and feed this information into the neural network.

The similarity between SOMs depends on atom fingerprints. We define the atom fingerprint for a specific atom to include all the paths, up to 8 bonds long, that are rooted at this atom. Atoms are labeled in a manner consisted with Daylight fingerprints in this protocol, and we also use molecule fingerprints similar to Daylight fingerprints alongside the atom fingerprints.[16]

We compute the similarity between SOMs in two ways. First, we compute the similarity as the MinMax metric[17] between their atom fingerprints. Second, we compute the similarity as the product of the MinMax metric between their atom fingerprints and the MinMax metric between their molecule fingerprints. As we will see, these similarities are surprisingly informative and experimental SOMs are found to be strongly similar to one another (Figures 7–9).

The information from these fingerprint similarities are used to generate descriptors (FP) using a weighted nearest-neighbor summary. For each potential SOM in the data set, the k most similar atoms from the training set are collected, some of which are experimental SOMs. The similarities corresponding to experimental SOMs from this list of k similarities are summed, and this sum is used as as a descriptor input to the network. With two types of similarities (based on atom or atom plus molecule fingerprints) and four values for k (1, 5, 10, or 20), this yields eight descriptors.

The FP descriptors can also be used as independent Fingerprint SOM prediction models. This is because the SOMs of a substrate can easily be ranked by their FP scores, which represents a SOM ranking according to CYP-oxidation likelihood. Just as there are eight FP descriptors there are eight Fingerprint models. The advantages Fingerprint models have over methods based on machine learning is that they are quickly applicable—as they do not requiring model optimization—and their predictions are easily explained. This is because the k most similar atoms from the training set that are used to make a Fingerprint prediction for a given SOM are already known at the time prediction is made.

Neural Networks. All models are constructed using a standard neural network with five hidden nodes calibrated through leave-one-out (LOO) cross-validation with gradient descent on the cross-entropy error.[18] LOO cross-validation in this case means that all SOMs for a single test substrate are predicted using models calibrated with all the SOMs from the remainder of the substrate set. This process is repeated with each substrate being treated as the test case once. Models created with this protocol produce output scores between 0 and 1 that can be interpreted as probabilities. For each training run, three random restarts were performed, choosing the model with the best training set accuracy prior to testing. Unique SOM prediction models have been built from each of the 10 substrate sets with SOMs represented with TOP and SCR descriptors in conjunction with different combinations of QC, SQC, MOL, and FP descriptors.

Scaled RS-Predictor Output. RSP produces a ranking of a molecule's atoms according the strength of its prediction that they are true, experimentally validated SOMs. The SOM rankings predicted by RSP in our earlier work were generated using 10 iterations of 10-fold cross-validation for all substrates in a particular data set.[5] In this framework, each molecule was treated as a test molecule 10 times and therefore has 10 independent SOM rank-orderings.

A RSP prediction score was calculated for each potential SOM from the average relative rank position of that SOM across all RSP-predicted SOM rank-orderings. The range of RSP score values is that same as the range for XenoSite scores. A SOM consistently ranked in the highest rank position will have a score of 1, and a SOM consistently ranked in the lowest rank-position will have a score of 0.

The SOM scores generated by both XenoSite and RSP are analyzed later in this work to determine whether they can be treated as probabilities (Figure 6). By this, we mean how well does the SOM score produced by a method correlate to the experimentally verified response for that SOM? Are SOMs with scores of 1 always experimentally verified to be sites of CYP-mediated metabolism and SOMs with scores of 0 always experimentally verified not to be sites of CYP-mediated metabolism?

Table 2. XenoSite Is the Most Accurate Method for Predicting the Metabolism of the Majority of Curated CYP Substrates[a,b]

isozyme	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	HLM	
number of substrates	271	105	151	142	226	218	270	145	475	680	average
XenoSite[c]	**87.1**	**85.7**	**83.4**(1)	**88.7**(2)	**86.7**	**89.0**(3)	**88.5**	**83.5**	**87.6**	**89.4**(4)	**87.0**
RS-Predictor[d]	83.4	**85.7**(4)	**82.1**	83.8	**84.5**	**86.2**	**85.9**	**82.8**	82.3	86.2(4)	84.3
SMARTCyp	80.0	**86.0**	77.0	83.0	**84.0**	**86.0**	83.0	**82.0**	78.0		82.1
StarDrop					78.0		75.3		74.1		75.8
Schrödinger					72.1		68.1		76.4		72.2
Fingerprint[e]	66.1	63.8	64.2	65.5	68.1	69.3	74.4	64.1	71.2	75.3	68.2
random model	26.0	31.9	24.8	22.6	22.2	20.2	21.1	36.5	21.0	26.3	25.3

[a]Prediction accuracy is assessed using the top-two metric, whereby a substrate is considered to be correctly predicted if any of its experimentally validated SOMs are predicted in the top two rank-positions by the given method. To simplify the overall presentation, only one set of results is presented for both the new XenoSite method and the old RSP method, with the optimal descriptor set varying between isoform substrate sets. SOM predictions obtained from RSP, StarDrop (V4.3),[19] and Schrödinger[20] are unchanged from our prior work.[5] However the underlying experimental responses for four substrates contained in 1A2, 2C8, 2D6, 3A4, and HLM sets have been updated as noted in the Data Sets section, making slight (<0.5%) changes to the prediction accuracies for 1A2 and HLM sets previously reported for RSP. Results for SMARTCyp (V2.4) were obtained from the SMARTCyp website.[21]. [b]For each CYP, the optimal model is shown in **bold**, as are all other models found not to be statistically different using the chi-squared test test of independence. [c]Optimal XenoSite models were trained using SOMs encoded with TOP, SCR, FP, SQC, and MOL descriptors. [d]Optimal RS-Predictor models were trained using SOMs encoded with TOP, QC, and SCR descriptors. [e]Optimal fingerprint models use ATOM fingerprints and 10 nearest neighbors to gauge SOM similarity with the exceptions of the following models labels: (1) TOP, SCR, and SQC; (2) TOP, SCR, FP, and SQC; (3) TOP, QC, SCR, and FP; and (4) TOP and SCR.

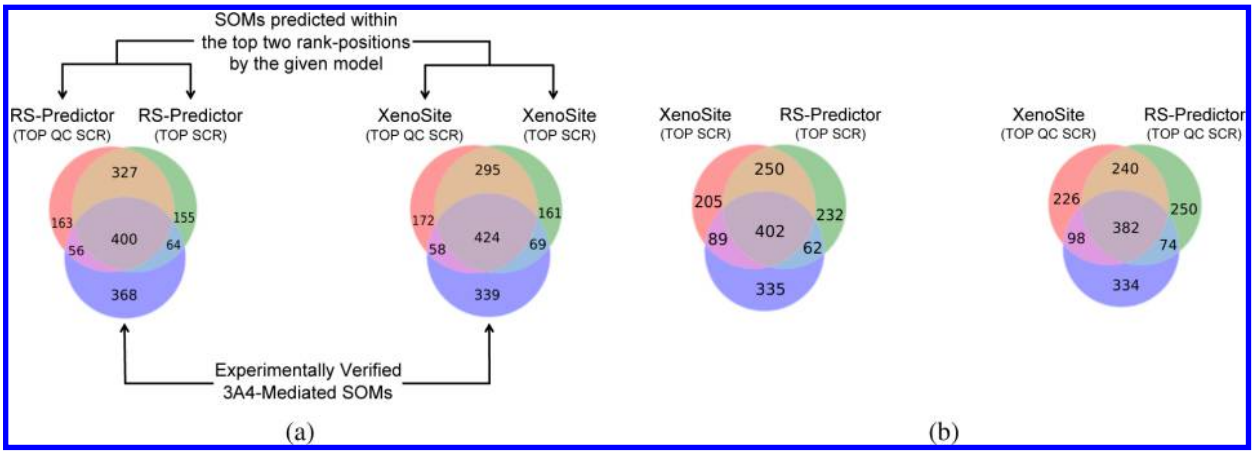


Figure 1. Number of SOMs accurately predicted by models using TOP and SCR descriptors and inaccurately by models using TOP, QC, and SCR descriptors—and vice versa—are similar whether models are calibrated using neural networks or SVMs (a). Neural networks build more accurate models than SVMs when calibrated on the same set of substrates encoded with the same set of descriptors. As 392 QC descriptors are incorporated into model optimization, greater discordances arise between the predictions made by XenoSite and RSP models (b).

We find that XenoSite scores can be treated reliably as probabilities, but that RSP scores calculated in the manner described above cannot be.

RESULTS

We evaluate XenoSite across several dimensions. First, we compare the overall prediction accuracy of our method to the accuracies reported for other methods on the same sets of substrates. Second, we evaluate the relative contribution of our newly developed descriptors to improving prediction accuracy. Third, we look at how XenoSite SOM predictions helped identify two substrates in our training and evaluation sets that had curated experimental responses that did not match existing source literature. Fourth, we assess how well atom scores generated by XenoSite and RSP can be treated as probabilities. In other words, how well does an atom's score value, which ranges from 0 to 1 and reflects the confidence of a model in its prediction, correlate to the statistical likelihood that that particular atom is actually metabolized by a given CYP? Fifth, we look at two substrates that have improved XenoSite predictions when FP descriptors are incorporated into the model (Figure 7), and we explain which chemically similar SOMs from the training set enabled those improved predictions (Figures 8 and 9). Finally, we benchmark the training time for XenoSite against those previously reported for RSP.

Overall Prediction Accuracy. The prediction accuracies of CYP SOM prediction models are commonly measured using a "top-two" metric. With this metric, a substrate is considered to be correctly predicted if any of its experimentally verified SOMs are predicted in the first- or second-rank positions by a given method. The top-two metric is used to evaluate the performance of XenoSite against previously reported performances of RSP and another academic method SMARTCyp, as well as two commercial methods from Optibrium[19] and Schrödinger,[20] for 10 distinct substrate sets (Table 2).

For nine of the ten substrate sets, optimal XenoSite models outperform optimal RSP models, which themselves outperform SMARTCyp, StarDrop, and Schrödinger models. The most significant advance of our method is for the 3A4 substrate set,

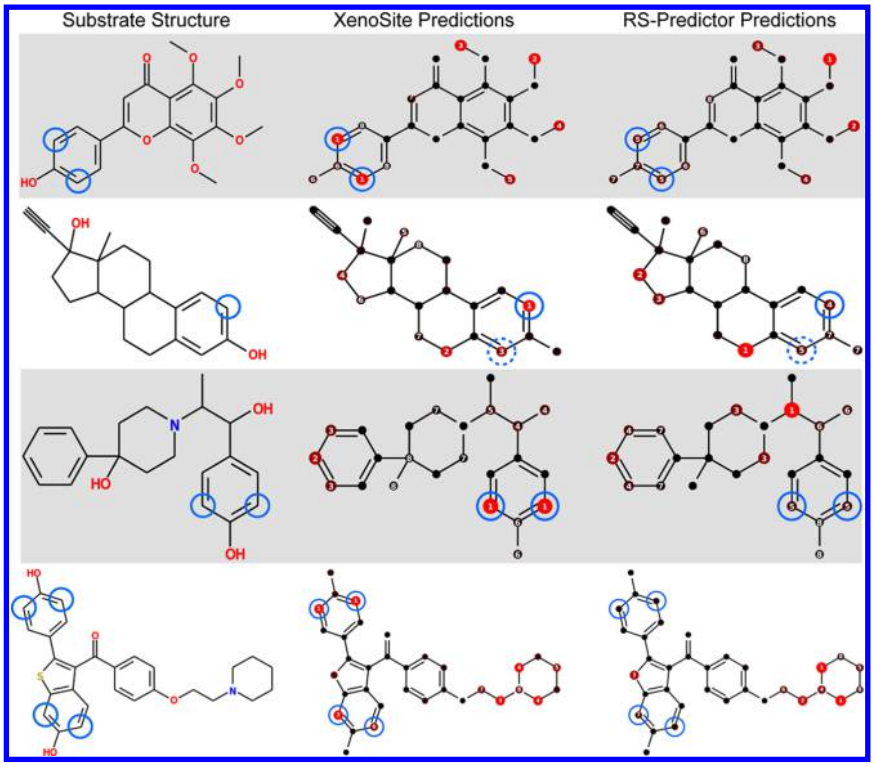


Figure 2. Substrates undergoing CYP-mediated ortho-phenol oxidation that were correctly predicted by XenoSite models and incorrectly predicted by RSP models. Both models were built using SOMs encoded with TOP and SCR descriptors. The substrate names from top to bottom are a metabolite of tangeretin, ethinylestradiol, traxoprodil, and raloxifene. 3A4-mediated metabolism and model predictions are shown for the top two substrates, and HLM-mediated metabolism and model predictions are shown for the bottom two substrates. Primary and secondary oxidations sites are, respectively, designated by solid and hashed blue circles. The radius of the prediction circle of an atom corresponds to likelihood that the atom is a verified site of CYP-mediated metabolism, which is determined by the atom's XenoSite or RSP score. For each substrate prediction, the circle colors represent prediction scores that have been normalized by value of the highest scored atom of the substrate, which is colored bright red. This display schema is used in all subsequent figures that illustrate the predictions of a given method for a given substrate.

where XenoSite models are 5.3% more accurate than RSP. The one case where XenoSite and RSP have equivalent prediction accuracies, which are slightly surpassed by those of SMARTCyp, is for the 2A6 substrate set. Of all the assembled substrate sets, the 2A6 set contains the fewest number of substrates (105), and we posit that if additional 2A6 substrates were incorporated, then XenoSite would be more effective at elucidating predictive signal than RSP. This hypothesis is supported by a general trend in the results where the number of substrates a set contains corresponds to the degree to which XenoSite outperform RSP. Of the four substrate sets that contain fewer than 152 substrates—2A6, 2B6, 2C8, and 2E1—XenoSite and RSP accuracies are within 1% of each other, with the exception of the 2C8 set where XenoSite outperforms RSP by 4.9%. In contrast, for the sets that contain greater than 217 substrates—1A2, 2C19, 2C9, 2D6, 3A4, and HLM—XenoSite models outperform RSP models by on average 3.3%.

Most methods make the same rank predictions for the same SOMs. Consider the set of 888 experimentally verified SOMs in the 3A4 substrate set, the set where XenoSite most significantly outperforms RSP (Figure 10b). Optimal XenoSite and RSP models correctly predict the same 388 (44%) verified SOMs and incorrectly predict the same 293 (33%) verified SOMs. The difference in utility between models is reflected not through their equivalent SOM predictions—correct or incorrect—but rather through the 139 (16%) SOMs correctly predicted by XenoSite and mispredicted by RSP and the 68 (7%) SOMs correctly predicted by RSP and incorrectly predicted by XenoSite. Even when models are built using the same set of 3A4 SOMs encoded with the same set of descriptors, neural networks are more effective than SVMs at correctly identifying observed sites of 3A4-mediated metabolism (Figure 1b). In addition, the discordances in the SOM predictions made by the two machine-learning algorithms become greater as SOMs are encoded with additional descriptors.

Further analysis reveals that a large number of CYP-mediated ortho-phenol oxidations are correctly identified by XenoSite models and not by RSP models (Figure 2). In a majority of these cases, models were trained using the same set of descriptors and substrates, indicating that neural networks are better than SVMs at building models able to correctly predict this type of reaction. In contrast, SVM models have a great propensity to predict the hydroxylation of carbons adjacent to nitrogens within piperidine rings. This can be a weakness of RSP, as illustrated by its predictions for raloxifene (Figure 2). However, it can also be a strength, as certain large molecules that undergo piperidine oxidation are incorrectly predicted by XenoSite models and correctly predicted by RSP (Figure 3).

Contribution of Different Descriptors. Multiple classes of descriptors were developed in this work to improve XenoSite accuracy. Molecule-level descriptors were incorporated into our models to help them better determine which atom-level descriptors are most appropriate for predicting the metabolism of a particular substrate. To aid in this endeavor certain atom-level QC descriptors were pruned in the creation of the SQC descriptor set, with the objective of retaining encoded signal

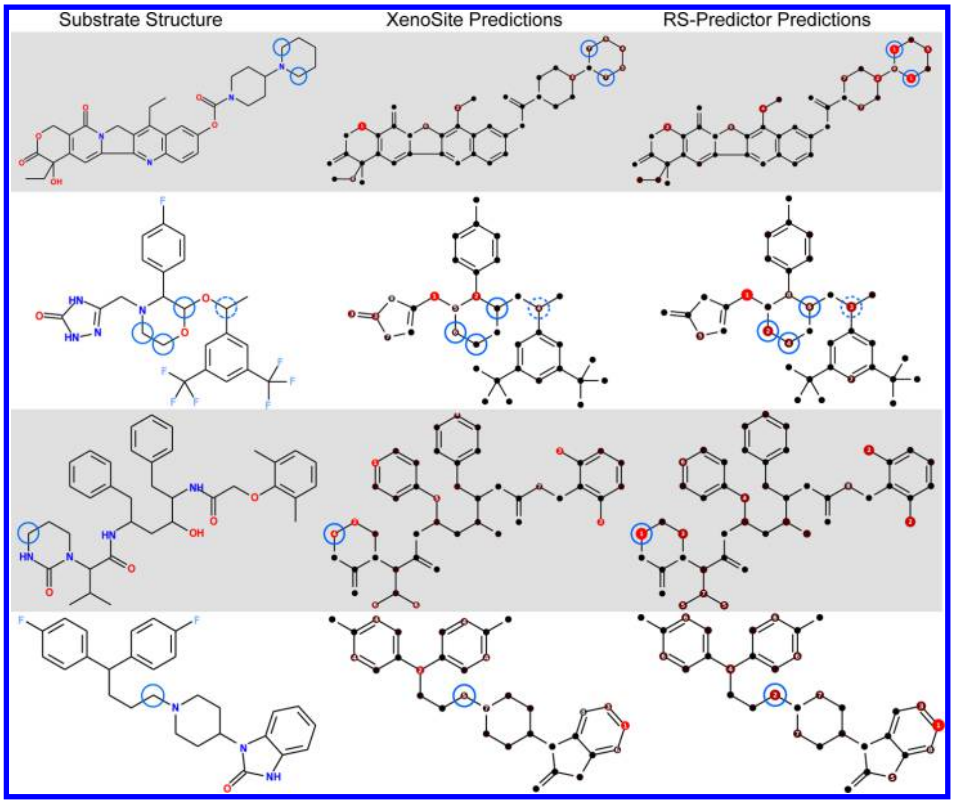


Figure 3. Substrates undergoing CYP-mediated piperadine oxidation that were correctly predicted by RSP models and incorrectly predicted by XenoSite. RSP models were built using SOMs encoded with TOP, QC, and SCR descriptors, and XenoSite models were built using SOMs encoded with TOP, SQC, SCR, MOL, and NN descriptors. The substrate names from top to bottom are irinotecan, aprepitant, abt 378, and pimozide. 3A4-mediated metabolism and model predictions are shown for the top three substrates, and 1A2-mediated metabolism and model predictions are shown for the bottom substrate. The display schema for each substrate prediction is described earlier (Figure 2).

Table 3. Value of Novel Descriptor Sets[a]

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	HLM	Average
FP SQC MOL	87.1	85.7	79.5	86.7	85.9	86.7	88.5	83.5	87.6	88.7	86.0
FP SQC	86.0	82.9	82.1	88.7	85.8	87.2	86.7	81.4	87.0	87.7	85.6
FP MOL	83.0	84.8	80.1	85.2	85.8	86.2	88.2	82.8	84.2	88.5	84.9
SQC MOL	86.0	82.9	79.5	87.3	83.2	87.6	86.3	80.7	86.5	87.5	84.8
FP	84.1	81.9	80.8	84.5	86.3	87.2	86.7	82.1	84.8	88.1	84.7
SQC	85.6	80.0	83.4	82.4	84.5	85.3	87.4	81.4	84.2	89.0	84.3
MOL	83.0	81.0	78.8	88.0	86.3	84.9	85.2	81.4	85.5	88.5	84.3
–	84.9	81.0	78.8	84.5	85.4	84.4	86.3	80.7	84.8	89.4	84.0

[a]All results below are for XenoSite models trained using baseline TOP and SCR descriptors, in addition to the descriptors in the listed sets. Results are grouped by whether they use 0, 1, 2, or 3 of the new types of descriptors. Within each group, the optimal performance is underlined.

while reducing information redundancy. A comprehensive investigation of different combinations of atom-level (TOP, SCR, QC, and SCR) and molecule-level (MOP and FP) descriptors were explored (Table 3; Supporting Information).

FP descriptors are the most informative of the novel descriptors, followed by SQC and then MOL. For use as baseline of comparison, the XenoSite model is trained using the same descriptor set employed by RSP (TOP and SCR). The average performances of XenoSite and RSP methods using these descriptors are 84% and 83.1%, respectively. The addition of the three newly developed descriptor sets (FP, SQC, MOL) to XenoSite modeling results in an improved average performance of 86%. Adding just two of the three new descriptor sets results in improved performances ranging between 84.8−85.6%, depending on which two sets are added; similarly, adding just one of the new descriptor sets results in improved performances ranging between 84.3−84.7%. The FP descriptors yield the biggest performance gains.

Predictions on Specific Molecules. In two instances, investigating molecules that were correctly predicted by one method and not another led to the identification of mistakes in the curated data. XenoSite models correctly predicted the metabolism of efavirenz into 8-hydroxyefavirenz, a reaction that RSP did not predict, but they did not predict the metabolism of 8-hydroxyefavirenz into 8,14-hydroxyefavirenz, a reaction that RSP correctly predicts (Figure 4). The predictions of multiple

dx.doi.org/10.1021/ci400518g | J. Chem. Inf. Model. 2013, 53, 3373−3383

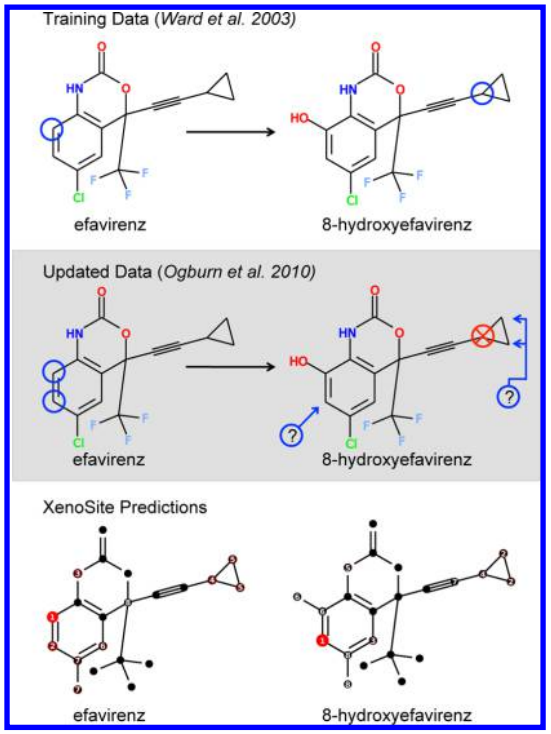


Figure 4. CYP-mediated SOMs of efavirenz and 8-hydroxyefavirenz as reported by Ward et al. are circled in blue in the top panel.[24] Subsequent studies by Ogburn et al. identified an additional metabolite of efavirenz and confirmed that while CYPs do metabolize 8-hydroxyefavirenz, the previously reported site of cyclopropyl oxidation is not correct, as indicated by red circles with Xs through them. The sites hypothesized by Ogburn et al. as the most likely to be oxidized by the CYPs are designated by blue circles filled with a question mark. These hypotheses are confirmed by predictions of HLM XenoSite models on efavirenz and 8-hydroxyefavirenz trained, respectively, with TOP, SCR, and SQC and TOP, SCR, MOL, and FP descriptors. The display schema for each substrate prediction is described earlier (Figure 2).

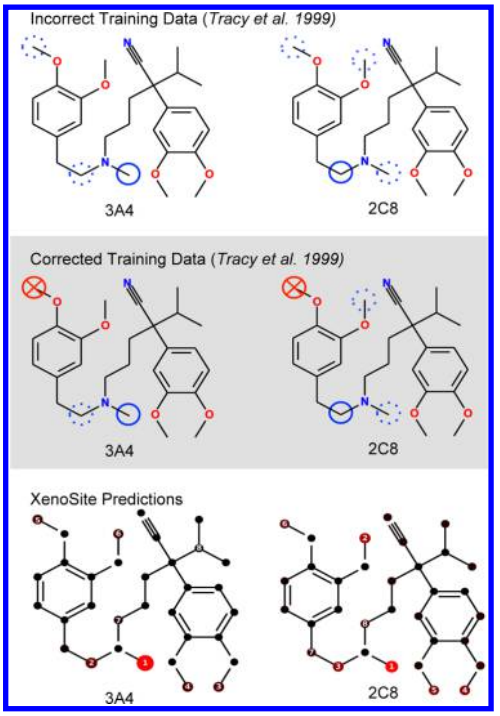


Figure 5. Primary and secondary 3A4- and 2C8-mediated SOMs of verapamil as initially curated from Tracy et al. are, respectively, circled in solid and hashed blue lines in the top panel. Subsequent re-examination of Tracy et al. revealed that previous curation had incorrectly marked potential sites of O-dealkylation—designated by red circles with Xs through them—as secondary SOMs. 3A4 and 2C8 XenoSite models that predict the correctly curated metabolism of verapamil were, respectively, trained with TOP, SCR, SQC, and MOL and TOP, SCR, QC, and MOL descriptors. The display schema for each substrate prediction is described earlier (Figure 2).

XenoSite models on efavirenz and its 8-hydroxylated metabolite were analyzed to determine why the starting substrate was correctly predicted, but its metabolite, which has a similar structure to the starting compound, was incorrectly predicted. XenoSite models consistently scored C7, C15, and C16 atoms highly, despite the fact that they were not recorded SOMs. This prompted further investigation of the literature, where a new source paper by Ogburn et al. identified these three atoms as likely sites of oxidation.[22]

A similar situation occurred with verapamil, where XenoSite models correctly identify primary sites of 3A4- and 2C8-mediated metabolism, but do not predict secondary sites of metabolism that were correctly identified by RSP (Figure 5). Examining the source literature of verapamil metabolism revealed that these secondary SOMs are in fact not metabolized by 3A4 and 2C8 isozymes,[23] consistent with the XenoSite's predictions.

These two examples demonstrate that XenoSite is able to (1) identify errors in the curated source data and (2) identify which metabolic reactions for novel therapeutics are consistently predicted across multiple models, increasing the likelihood of those predictions being correct.

Probabilistic Predictions. The evaluation of a SOM prediction model often focuses on the accuracy of its predictions and not on the confidence the model has in those predictions. This is an important point, for while a given model may have a certain level of predictive accuracy for a large set of substrates, that accuracy does not necessarily translate to the prediction of a particular substrate. End-users of SOM prediction models benefit not only from having more accurate models, but from knowing which model predictions are reliable, and which are not. In this section we investigate the correlation between the predictive score of a model and whether that prediction is correct.

Our analysis shows that the output from XenoSite models can be interpreted as a probability (Figure 6a). In other words, atoms assigned a prediction score of 0.8 have an 80% chance of being an experimentally verified SOM. Likewise, atoms with a score of 0.2 have a 20% probability of being an experimentally verified SOM. This feature—output interpretable as a probability—is a step forward in the field, providing a natural way of assessing the confidence of a model's predictions. It is also not a surprising finding, as neural networks are designed specifically to produce output interpretable in this manner.[18,25]

In contrast, raw RSP scores do not correlate well with probability (Figure 6b). This too is not surprising, as the SVMs of RSP only seek to rank atoms appropriately within a molecule, without constraints on the scale of the resulting predictions. To be sure, there are ways of rescaling these scores into probabilities,[26] as is also possible for any method that yields quantitative output. Rescaling scores to probabilities, however, is not common practice in SOM research because the metric commonly used to evaluate SOM predictions—any

dx.doi.org/10.1021/ci400518g | J. Chem. Inf. Model. 2013, 53, 3373−3383

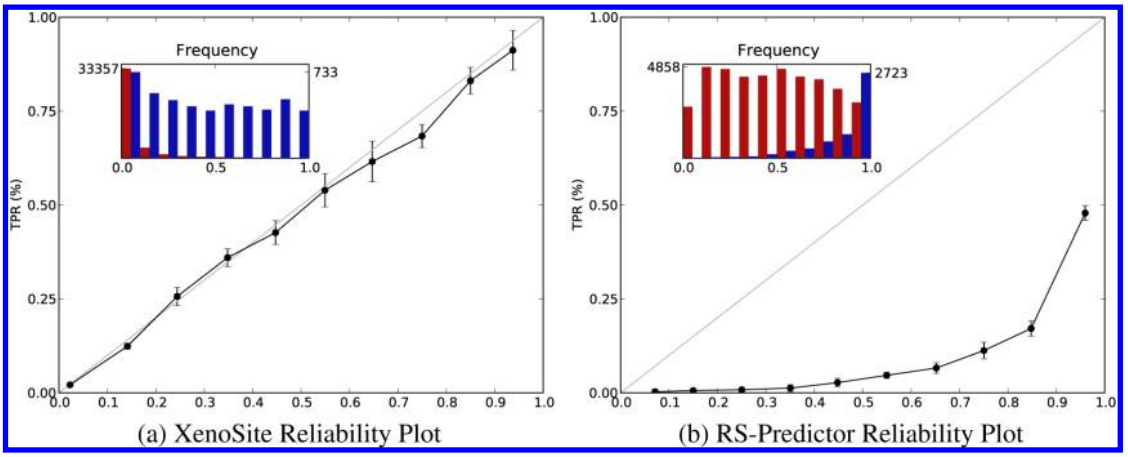


Figure 6. Correlation plots between predicted SOM scores and true-positive SOM prediction rate for the given method across all SOMs in all substrate sets. Panel b was constructed using RSP models built with SOMs encoded with TOP and SCR descriptors. The *x*-axis of each graph constitutes SOMs placed into bins of size 0.1 based on their normalized predicted scores. The *y*-axis of each line-graph designates the true-positive prediction rate of each bin. The histogram subfigures show the number of both experimentally observed (blue) and non-observed (red) SOMs in each bin. The subfigure *y*-axes show the number of SOMs contained in the most heavily populated observed bin (blue) on the right and the most heavily populated non-observed bin (red) on the left. These figures show that prediction scores of XenoSite models are reliable indicators of the accuracy of those predictions, while RSP scores are not.

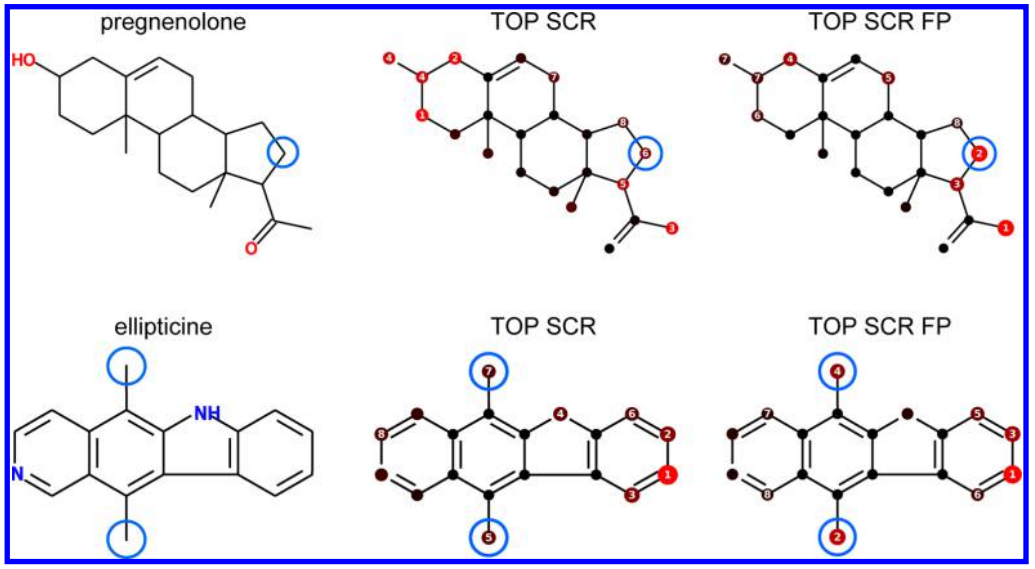


Figure 7. Fingerprint descriptors improve XenoSite predictions. The SOMs of pregnenolone by 3A4 and ellipticine by 2C9 are marked with blue circles. XenoSite models trained with SOMs encoded with TOP and SCR descriptors do not successfully identify the CYP-mediated metabolism of either substrate. Encoding SOMs with FP, TOP, and SCR descriptors results in accurate XenoSite prediction of both substrates. The 10 nearest neighbor SOMs of the observed ring hydroxylation of pregnenolone using the Fingerprint method are shown in Figure 8. The 11 nearest neighbor SOMs of the observed hydroxylation of ellipticine using the Fingerprint method are shown in Figure 9. The display schema for each substrate prediction is described earlier (Figure 2).

experimentally verified SOM predicted within the top two rank-positions—is insensitive to the exact value of predictions and only depends on their rank relative to each other.

XenoSite models, therefore, represent an important advance over RSP models and other SOM predictors beyond being more accurate. Users can look at the predictions made by XenoSite models and know which ones are reliable and trustworthy and which ones are not. In contrast, users looking at the top predicted rankings of RSP cannot be as confident in identifying which predictions may be reliably trusted.

Explaining Predictions. A disadvantage of methods based on machine learning is that their predictions often lack transparency. It is difficult to derive a set of simple chemical rules to explain CYP-oxidation from the weights of hundreds of descriptors optimized over hundreds of substrates. The lack of explanation for a given prediction is a significant obstacle to overcome for the continued advancement of existing methods dependent on machine learning. How can one know what improvements should be made to a method if they do not know exactly why the method incorrectly predicts the oxidation of some substrates and correctly predicts the oxidation of others?

The main advantage of the Fingerprint method is that it has an explanation for each prediction that it makes. Each SOM being predicted has a set of nearest neighbor SOMs determined from the training set through the evaluation of fingerprint similarities. Looking at the molecule that contains the SOM being predicted in conjunction with the molecules that contain

dx.doi.org/10.1021/ci400518g | J. Chem. Inf. Model. 2013, 53, 3373−3383

its nearest neighbor SOMs can provide insights as to how and why certain predictions get made (Figures 7–9).

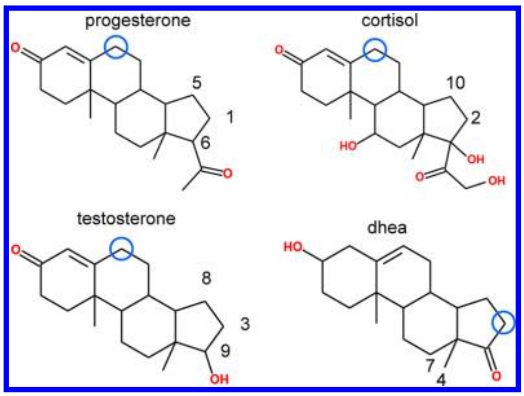


Figure 8. Top 10 most similar SOMs to the site of 3A4-mediated ring hydroxylation of pregnenolone (Figure 7) are designated by number in the figure above. Similarity between the observed SOM of pregnenolone and the SOMs contained the other 474 substrate of 3A4 are assessed using the atom and molecule fingerprints of the Fingerprint method. Primary and secondary sites of 3A4-mediated metabolism of the substrates containing the most similar SOMs are respectively designated with solid and hashed blue circles.

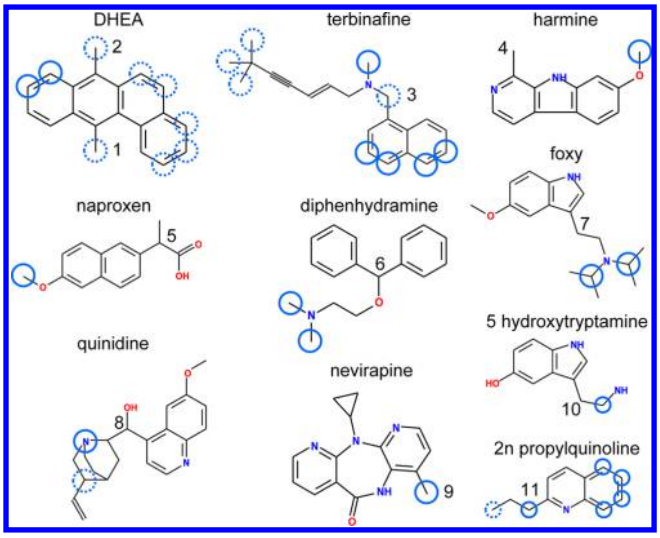


Figure 9. Top 11 most similar SOMs to the bottom site of 2C9-mediated hydroxylation of ellipticine (Figure 7) are designated by number in the figure above. Similarity between the top observed SOM of ellipticine and the SOMs contained the other 225 substrate of 2C9 are assessed using the atom and molecule fingerprints of the Fingerprint method. Primary and secondary sites of 2C9-mediated metabolism of the substrates containing the most similar SOMs are, respectively, designated with solid and hashed blue circles.

Training Time. XenoSite models train significantly faster than RSP models when both methods are applied to the same data set (Table 4). Runtimes were assessed for three substrate sets using SOMs encoded with either TOP and SCR, or TOP, QC, and SCR descriptors using 10 iterations of 10-fold cross-validation.

When SOMs are represented with TOP and SCR descriptors—where the 149 descriptors are predominantly whole numbers—XenoSite models are built on average 4 times faster than RSP models. However, when SOMs are represented with TOP, SCR, and QC descriptors—where the 392 QC descriptors are predominantly long integers—XenoSite models are built on average just 1.3 times faster than RSP models. Increasing the number of substrates being optimized does not significantly increase the per-substrate runtime of either method.

Table 4. Runtimes for Training XenoSite and RS-Predictor Models from the Same Sets of CYP Substrates Represented with the Same Two Sets of Descriptors

method and descriptor set	isozyme and number of substrates		
XenoSite	2A6 (105)	2C9 (226)	3A4 (475)
TOP SCR (149)	2.2 h	7.2 h	18.5 h
TOP QC SCR (541)	10.8 h	1.4 d	3.2 d
RS-Predictor			
TOP SCR (149)	9.0 h	1.1 d	3.0 d
TOP QC SCR (541)	13.7 h	1.7 d	4.6 d

DISCUSSION

This work describes XenoSite, a method for building models that predict CYP-mediated sites of metabolism (SOMs) for drug-like molecules. The predictive accuracies of these models surpass the accuracies reported for other methods for nine distinct CYP substrate sets. XenoSite improves on a previously reported method, RS-Predictor (RSP), by (1) incorporating two types of molecule-level descriptors and (2) building predictive models using neural networks rather than SVMs. Neural networks build models 4 times faster than SVMs when applied to the same set of SOMs encoded with the same set of descriptors. In addition, SOM scores generated by XenoSite models can be treated as probabilities, while those of RSP cannot. This means that XenoSite SOM score represents the confidence of the model that the SOM is oxidized and that that score value is strongly correlated to the statistical likelihood that the prediction of that SOM is accurate. An alternative Fingerprint model is described that predicts SOM oxidation likelihood from the observed CYP-mediated oxidation of chemically similar SOMs. This method differers from XenoSite in that oxidative similarities between SOMs are determined through fixed binary fingerprints rather than through descriptor and machine-learning tuned models.

Optimal XenoSite models are on average 87% accurate across all analyzed substrate sets, a level of performance 3% higher than that of the previously existing optimal method RSP. This increase in performance comes from representing putative SOMs with two new classes of molecule-level descriptors and pruning the descriptor composition of a previously developed atom-level descriptor class to remove noise while retaining signal; no single one of these improvements is responsible for the total increase in predictive accuracy.

These results support our initial hypothesis that encoding molecule-level information for each SOM would allow models to better determine the relative importance of atomic-level information in order to more accurately predict the CYP-mediated oxidation of that SOM. In this work, only a small number (15) of molecule-level descriptors were explored, and our findings suggest that a more comprehensive investigation into different classes of molecule-level descriptors would be a viable research path. One option would be to investigate the 1497 molecule-level DRAGON descriptors used by Yap et. al to build CYP substrate and inhibitor classification models.[27]

Creating a consensus prediction for a substrate by combining the predictions made for it by different models is another viable

dx.doi.org/10.1021/ci400518g | J. Chem. Inf. Model. 2013, 53, 3373–3383

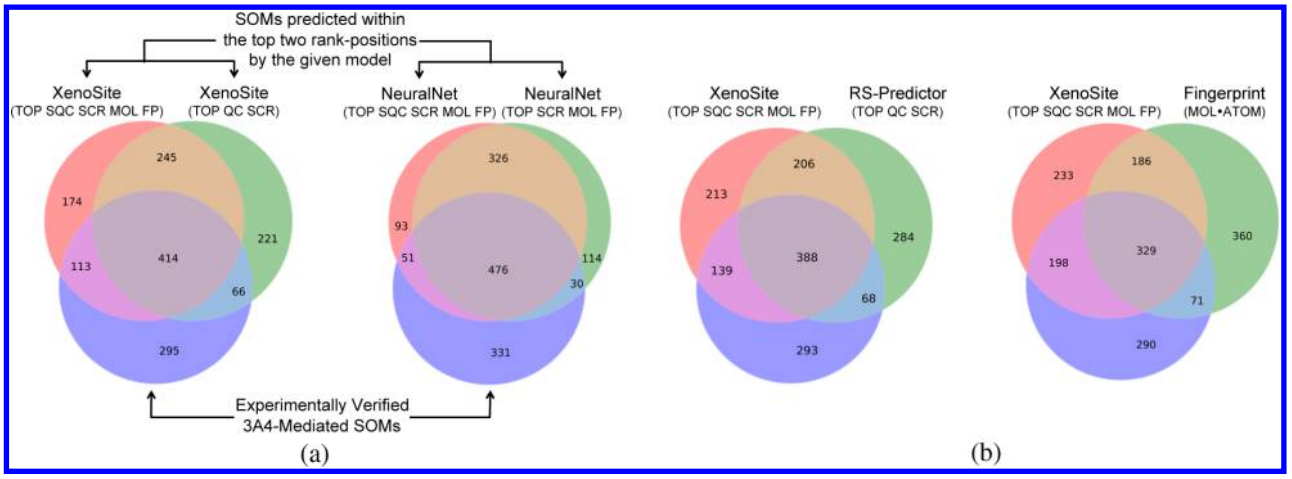


Figure 10. Ensemble scoring techniques can be developed with the goal of producing an ensemble model able to identify all experimentally verified SOMs that are correctly predicted by at least one of its constituent models. Ensemble methods derived from XenoSite and RS-Predictor or Fingerprint methods (b) have greater predictive potential than those derived from two different XenoSite models (a). However, aggregating the scores of two models will be easier if they stem from the same method, as they will be in the same format and probability scale.

path. There are a number of cases for which non-optimal models correctly predict the oxidation of substrates that were incorrectly predicted by optimal models. SOMs that were correctly predicted by one model and not another represent valuable targets for continued investigations (Figure 10). If two models each identify different sets of experimentally verified SOMs correctly, it should be possible to integrate them into single ensemble model able to identify the experimentally verified SOMs of both sets. Merging the prediction scores of different models should be easier if they are derived using neural networks because each score generated by a neural network can be treated on the same probability scale (Figure 10a). In contrast, successfully aggregating the predictions of RSP or Fingerprint methods with those of the XenoSite method, while more technically involved, has the greatest potential to build the most accurate ensemble model (Figure 10b).

The Fingerprint method has relatively low predictive accuracy (68.2%) averaged across all substrate sets. This is to be expected, as the fingerprints and MinMax metric used to gauge SOM similarity were developed for other applications and were not optimized in any way for CYP-metabolism prediction. Now that fingerprint-based descriptors have been shown to improve model performance, experimenting with alternative fingerprints and similarity metrics is a justifiable research path that could easily yield a more effective Fingerprint model and more effective descriptors. Though the XenoSite predictions are, on average, more accurate (87.0%), the Fingerprint method has the advantage of being transparent; in other words, the SOMs and substrates that are directly responsible for making each fingerprint prediction are readily assessable (Figures 7–9). In this work, only the end-scores predicted by the Fingerprint method were incorporated into XenoSite. Another viable way to utilize the nearest neighbor information would be to calibrate a model using only substrates that fall within similarity cutoff value to the substrate being predicted rather than using all known substrates for the catalyzing isozyme.

The Fingerprint method in this work—with an average accuracy of 68.2%—is just a first step toward explaining the machine-learning-based predictions of the XenoSite method—with an average accuracy of 87%. There is no guarantee that the nearest neighboring substrates to the one being predicted have a greater impact on model calibration than others. In addition, there is no guarantee that substrates that are similar in fingerprint space have similar CYP-mediated metabolism, which is why incorporating fingerprint-derived oxidation scores as descriptor values occasionally decreases prediction accuracy. Finally, the fingerprints currently being used do not encode nontopological information used to build XenoSite models.

Experimenting with creating substrate-specialized training sets is a much more viable research path when using neural networks, which train models significantly faster than SVMs (Table 4). On average, neural networks build models 4 and 1.3 times faster than SVMs using SOMs encoded with TOP and SCR, and TOP, QC, and SCR descriptor sets, respectively. In addition to facilitating model generation with training sets of different substrates, the superior runtimes of neural networks make them a better tool to quickly investigate the utility of new descriptor classes. However, our findings indicate that descriptors with numerical integer values will engender a less significant time-benefit to building models with neural network rather than SVM technologies, as opposed to descriptors with whole or binary numerical values. Further investigations can also be made into different techniques for performing model optimization with neural networks. The implementation used in this work—using five hidden nodes calibrated through leave-one-out cross-validation with gradient descent on cross-entropy error—is a generic neural network solution that has not been tailored to the intricacies of SOM-prediction modeling.

■ ASSOCIATED CONTENT

Supporting Information

Full results for all XenoSite models trained on all substrates sets using different combinations of TOP, SCR, QC, SQC, MOL, and FP descriptors. Definitions for TOP and QC descriptor sets. A brief discussion of how SCR descriptors are extracted from a newly released version of SMARTCyp effect XenoSite model performances. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: swamidass@gmail.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Professor Curt Breneman for access to computers and software and Michael Browning 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) Zimmerman, H. J.; Maddrey, W. C. Acetaminophen (paracetamol) hepatotoxicity with regular intake of alcohol: Analysis of instances of therapeutic misadventure. *Hepatology* **1995**, *22*, 767−773.
(4) 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.
(5) 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.
(6) Huang, T.-w.; Zaretzki, J.; Bergeron, C.; Bennett, K. P.; Breneman, C. M. DR-Predictor: Incorporating flexible docking with specialized electronic reactivity and machine learning techniques to predict CYP-mediated sites of metabolism. *J. Chem. Inf. Model.* **2013**, DOI: 10.1021/ci4004688.
(7) Haritos, V. S.; Ching, M. S.; Ghabrial, H.; Gross, A. S.; Taavitsainen, P.; Pelkonen, O.; Battaglia, S. E.; Smallwood, R. A.; Ahokas, J. T. Metabolism of dexfenfluramine in human liver microsomes and by recombinant enzymes: Role of CYP2D6 and 1A2. *Pharmacogenet. Genomics* **1998**, *8*, 423−432.
(8) Ebner, T.; Meese, C. O.; Eichelbaum, M. Mechanism of cytochrome P450 2D6-catalyzed sparteine metabolism in humans. *Mol. Pharmacol.* **1995**, *48*, 1078−1086.
(9) 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.
(10) 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.
(11) Cruciani, G.; Baroni, M.; Benedetti, P.; Goracci, L.; Fortuna, C. G. Exposition and reactivity optimization to predict sites of metabolism in chemicals. *Drug Discovery Today: Technol.* **2012**, *10*, e155−e165.
(12) *MOE*, version 2009.10; Chemical Computing Group: Montreal, Canada, 2009.,
(13) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(14) MOE 2008.10 QuaSAR-Description. http://www.cadaster.eu/sites/cadaster.eu/files/challenge/descr.htm (accessed June 21, 2013).
(15) Swamidass, S. J.; Azencott, C.-A.; Lin, T.-W.; Gramajo, H.; Tsai, S.-C.; Baldi, P. Influence relevance voting: an accurate and interpretable virtual high throughput screening method. *J. Chem. Inf. Model.* **2009**, *49*, 756−766.
(16) Butina, D. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.
(17) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093−1110.
(18) Baldi, P.; Brunak, S. *Bioinformatics: The Machine Learning Approach*; MIT Press: Cambridge, MA, 2001
(19) *StarDrop*, version 4.3; Optibrium Ltd.: Cambridge, U.K., 2009.
(20) *P450 SOM Prediction*, version 1.0; Schrödinger LLC: New York. 2011,
(21) About SMARTCyp. http://www.farma.ku.dk/smartcyp/about.php (accessed June 20, 2013).
(22) Ogburn, E. T.; Jones, D. R.; Masters, A. R.; Xu, C.; Guo, Y.; Desta, Z. Efavirenz primary and secondary metabolism in vitro and in vivo: Identification of novel metabolic pathways and cytochrome P450 2A6 as the principal catalyst of efavirenz 7-hydroxylation. *Drug Metab. Dispos.* **2010**, *38*, 1218−1229.
(23) Tracy, T. S.; Korzekwa, K. R.; Gonzalez, F. J.; Wainer, I. W. Cytochrome P450 isoforms involved in metabolism of the enantiomers of verapamil and norverapamil. *Br. J. Clin. Pharmacol.* **1999**, *47*, 545−552.
(24) Ward, B. A.; Gorski, J. C.; Jones, D. R.; Hall, S. D.; Flockhart, D. A.; Desta, Z. The cytochrome P450 2B6 (CYP2B6) is the main catalyst of efavirenz primary and secondary metabolism: implication for HIV/AIDS therapy and utility of efavirenz as a substrate marker of CYP2B6 catalytic activity. *J. Pharmacol. Exp. Ther.* **2003**, *306*, 287−300.
(25) Dybowski, R.; Roberts, S. Confidence Intervals and Prediction Intervals for Feed-Forward Neural Networks. In *Clinical Applications Artificial Neural Networks*; Cambridge University Press: Cambridge, U.K., 2001; Chapter 13, pp 298−326.
(26) Tao, Q.; Wu, G.-W.; Wang, F.-Y.; Wang, J. Posterior probability support vector machines for unbalanced data. *IEEE Trans. Neural Networks* **2005**, *16*, 1561−1573.
(27) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982−992.