JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 1282−1291

pubs.acs.org/jcim

CypReact: A Software Tool for in Silico Reactant Prediction for Human Cytochrome P450 Enzymes

Siyang Tian,[†,¶] Yannick Djoumbou-Feunang,[‡] Russell Greiner,[*,†,¶] and David S. Wishart[†,‡]

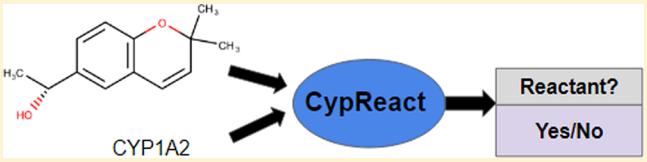
[†]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2E8

[‡]Department of Biological Science, University of Alberta, Edmonton, Alberta, Canada T6G 2E9

[¶]Alberta Machine Intelligence Institute (AMII), 2-21 Athabasca Hall, University of Alberta, Edmonton, Alberta, Canada T6G 2E8

Supporting Information

ABSTRACT: In silico metabolism prediction requires first predicting whether a specific molecule will interact with one or more specific metabolizing enzymes, then predicting the result of each enzymatic reaction. Here, we provide a computational tool, CypReact, for performing this first task of reactant prediction. Specifically, CypReact takes as input an arbitrary molecule (specified as a SMILES string or a standard SDF file) and any one of the nine of the most important human cytochrome P450 (CYP450) enzymes—CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, or CYP3A4—and accurately predicts whether the query molecule will react with that given CYP450 enzyme. Tests of CypReact, conducted over a data set of 1632 molecules (each considered a "plausible" reactant) show that it is very effective, with a (cross-validation) AUROC (area under the receiver operating characteristic curve) of 0.83−0.92. We also show that CypReact performs significantly better than other reactant prediction tools such as ADMET Predictor and (a reactant-predicting extension of) SMARTCyp, whose average AUROCs are 0.75 and 0.53, respectively. We then applied the learned CypReact models to a previously unseen set of molecules and found that our CypReact did even better and still significantly surpassed the performance of SMARTCyp and ADMET Predictor. These results suggest that CypReact could be an important component of a suite of in silico metabolism prediction tools for accurately predicting the products of Phase I, Phase II, and microbial metabolism in humans. CypReact is available at https://bitbucket.org/Leon_Ti/cypreact.



INTRODUCTION

On a daily basis, humans are exposed to hundreds or even thousands of chemicals through their routine interactions with the environment. These exposures can occur as a result of food/drug consumption, household or workplace activities, industrial or transportation activities, and even common environmental processes. Once absorbed, these chemicals usually undergo further biologically mediated transformations. These biotransformations can be beneficial or detrimental, depending on the type of chemicals (e.g., food supplements vs pesticides), the length of the exposure (short-term vs long-term), and the amount absorbed. If our bodies have absorbed or produced a toxic molecule, it is very important that it is deactivated (through various metabolic processes) and/or excreted from our body quickly.

Therefore, understanding how a molecule can be transformed or metabolized is crucial for the assessment of its bioavailability, bioactivity, and toxicology. As a result, experimental metabolite identification along with in silico metabolite prediction have become increasingly important research activities for a number of life science disciplines, including drug development, drug testing, pharmaceutics, pharmacology, toxicology, environmental monitoring, metabolomics, food science, and personalized medicine.[1]

In humans, many chemicals are extensively metabolized by cytochrome P450 (CYP450) enzymes. CYP450-mediated metabolism, which is a major component of Phase I metabolism, occurs primarily in the liver and kidneys. In humans, there are >50 known CYP450 variants (also known as CYP450 isozymes).[2,3] Among these 50 isozymes, just 9—CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, 2C19, CYP2D6, CYP2E1, and CYP3A4—are responsible for most of the known Phase I metabolism of drugs,[4] as well as the Phase I metabolism of a number of food compounds, environmental pollutants, and other xenobiotic molecules.

In silico metabolism prediction is a field of metabolite analysis that involves predicting the likely metabolites from a given starting molecule. It was initially developed in the early 1960s to help identify drug metabolites generated through Phase I metabolism based on observed mass spectrometry and/or NMR spectroscopy data.[5] Since then, in silico metabolism prediction has expanded to include not only the prediction of drug metabolism but also the prediction of environmental/microbial metabolism,[6] promiscuous enzyme metabolism,[7] and many other kinds of xenobiotic and endogenous metabolism.[8] Typically, in silico metabolism prediction can be broken down into three general steps:

Received: January 22, 2018
Published: May 8, 2018

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

1. predicting whether a molecule will react with an enzyme ("reactant" prediction) [Here, we classify an inhibitor as a non-reactant];
2. predicting where this interaction will occur ("site of metabolism" prediction); and
3. predicting the result of this interaction ("end point" prediction).

See Figure 1 for a more schematic description of these steps. A number of specific programs have been developed for certain

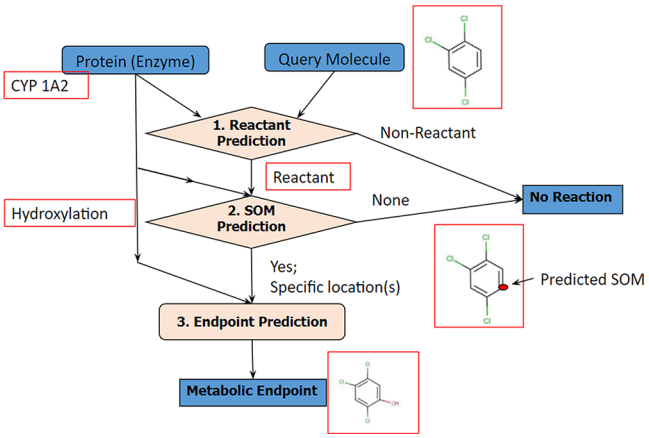


Figure 1. Overview of the overall reaction-prediction process.

individual steps in this process (or something similar to one of the above steps). For example, WhichCyp[9] predicts whether a given molecule *inhibits* a specified CYP450 enzyme—a classification task that is similar to predicting reactants (step 1). SMARTCyp[10] and MetaPrint2D[11] each take a molecule and an enzyme as input, then predict the site(s) where the interaction occurs—i.e., the site(s) of metabolism (SOM), which is similar to step 2. (We will later compare our CypReact to a modification of this SMARTCyp system.)

There are also several commercial programs, such as ADMET Predictor[12] (developed by Simulations Plus, Inc., Lancaster, California, USA) and StarDrop[13] (Optibrium Ltd., Cambridge, UK), that combine all three steps to predict which molecules are substrates of several general CYP isoforms (sometimes without identifying the specific isoform involved), where the sites of metabolism are on the target molecule and what the resulting structures of the predicted metabolites might be. (We will later compare CypReact to the relevant component of ADMET Predictor. [We did not compare to StarDrop as it is not freely available. We also wanted to compare our CypReact with RS-Predictor and MetaPrint2D. However, their web servers are very slow, and the results are difficult to process. We attempted to contact their authors, but did not receive a reply.])

While there are a few open-access in silico tools that predict the transformation products of a given compound by microbial[6] or selected endogenous enzymes,[7] there is currently no open-access tool that specifically predicts whether a substrate will react with a specific cytochrome P450 enzyme. This is surprising as CYP450 metabolism prediction is a fundamental to many xenobiotic and drug metabolism studies.

Currently, nearly all existing in silico CYP450 metabolism prediction tools (corresponding to steps 2 or 3 of Figure 1) implicitly assume that every input molecule is a CYP450 substrate, as they each produce a nontrivial prediction for each molecule. This includes several methods based on molecular

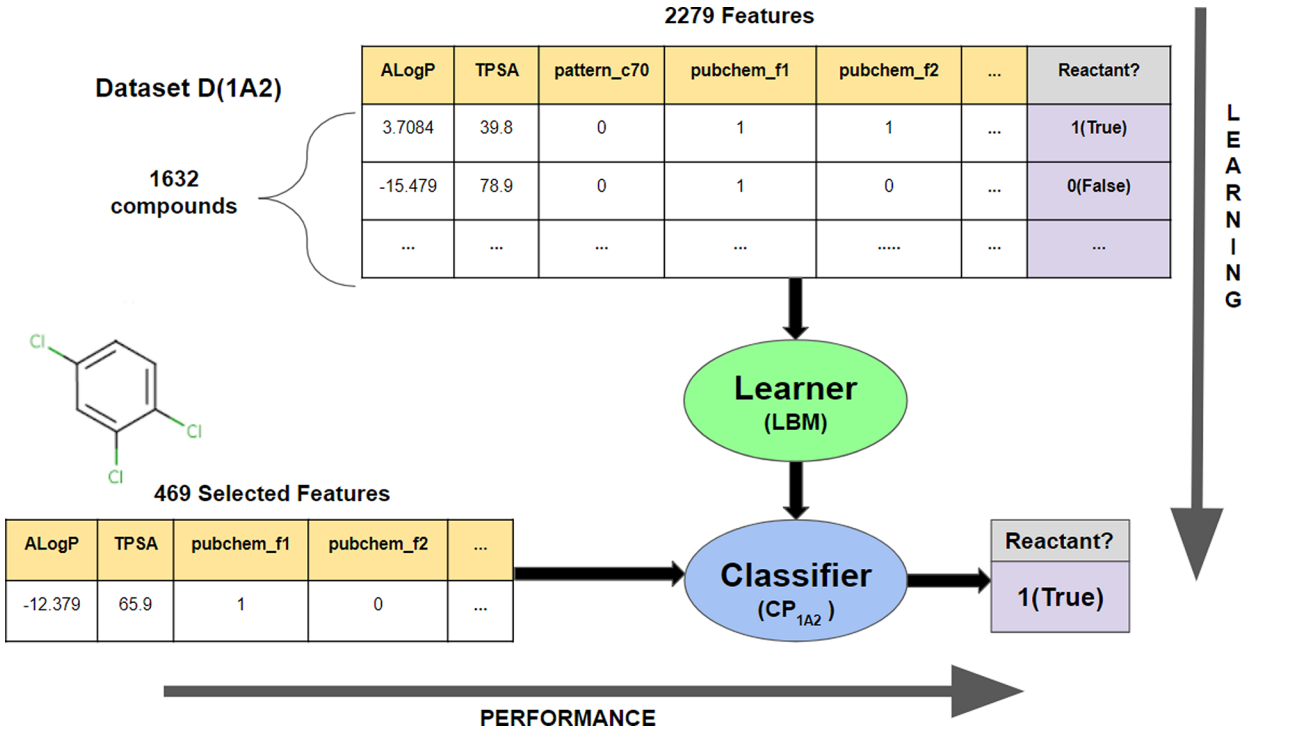


Figure 2. Basic machine learning paradigm, with learning algorithm LBM (learning base model) using the D(1A2) data set to produce a classifier CP_{1A2} (top-to-bottom), where this resulting CP_{1A2} can then make a prediction about an input molecule (left to right). Note the classifier uses a reduced set of features. Also, the data sets for the eight other isoforms are slightly different (with different "Reactant?" labels), leading to eight different classifiers.

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

pattern recognition,[14] artificial neural networks,[15] and other machine learning approaches involving molecular feature extraction and analysis.[16] We will later use a reactant-predictor variant of SMARTCyp, which computes a "score" for each location of the molecule, indicating its belief that this location is a SOM. This is challenging for our reactant-prediction task as SMARTCyp provides only a relative ranking, but we need to to make a yes/no decision, which here requires determining the cutoff score, to identify whether any of these locations should be viewed as SOMs.

Given that the vast majority of drugs and xenobiotics known today are *not* CYP450 substrates (some are inhibitors, but most are nonreactants), it is important to have reliable and accurate predictive tools that can accurately distinguish reactants from nonreactants. The many in silico CYP450 metabolism predictors that implicitly assume that every molecule is a reactant will make many mistakes. A tool that can accurately predict reactions could reduce these false positives, which in turn could also save time and money in the drug development process, as this would mean that drug metabolism researchers would not need to devote significant efforts trying to find predicted CYP450-derived drug metabolites that do not exist.

We hypothesize that machine learning tools can learn a computational model that can effectively distinguish between reactants and nonreactants—where effectiveness is measured in terms of both AUROC (area under the receiver operating characteristic curves) and a meaningful cost function, over a distribution of compounds of interest to the drug industry, environmental chemists, and metabolomic researchers. To do this, we first constructed a meaningful training (testing) sets of 1632 (169) molecules, that includes both reactants and "decoy" molecules that look like reactants but which are actually nonreactants. We then used this data set to develop and evaluate a machine learning system to produce a suite of $\{\text{CypReact}(\alpha)\}_\alpha$ "reactant classifiers", one for each of the nine main CYP450 isoforms α. Our empirical results demonstrated that these learned models were effective—with cross-validation AUROC values ranging from 0.83 to 0.92 on the training set. These were also higher than other comparable tools. Given the strong performance of our tool, we anticipate that CypReact will be a useful component for other in silico metabolism predictors as well as an important member of a new suite of open access in silico metabolism prediction tools that we are developing for accurately predicting the products of Phase I, Phase II, and microbial metabolism in humans.

MATERIALS AND METHODS

Approach. Because of the difficulty of the problem we are attempting to solve, we decided to pursue a machine learning approach, which is based on learning the relevant predictors from a large, high quality set of training data; see Figure 2. As each of the nine most important CYP450 enzymes has its own set of reactants, we built nine separate predictors—one for each CYP450 isoform. Below, we will let $\text{CypReact}(\alpha, \cdot)$ refer to the predictor for the isoform $\alpha \in \{\text{CYP1A2, CYP2A6, ..., CYP3A4}\}$, where $\text{CypReact}(\alpha, m)$ is 1 ("True") if the molecule m is a reactant to the isoform α and otherwise is 0 ("False").

Data Set Creation. CYP450 isozymes have a very broad substrate specificity and are responsible for most of the oxidative reactions seen in the Phase I metabolism of small molecule xenobiotics.[2] However, small changes in the chemical structure of a molecule can significantly alter its bioactivity or its metabolic profile.[2] Therefore, in order to train and test our models, it is very important to use a large and diverse data set that captures the molecular patterns and chemical features responsible for the specific interaction between a given CYP450 and its substrates. To be useful, this data set should include just the molecules that a biochemist would consider as possible reactants—i.e., just the molecules that a researcher would consider plausible and therefore worth sending to the resulting CypReact prediction system.

We built a data set with 1632 compounds, including 679 known CYP450 reactants from the set provided by Zaretzki et al.,[15] each of which is metabolized by at least one of the nine CYP450 isozymes. [One of the 680 molecules reported in the set was a duplicate (phenanthrene) and was, therefore, removed from the set.] To provide a sufficiently large and relevant training set, we manually collected an additional set of 1053 nonreactant compounds that were "plausible" metabolites—i.e., small molecules that are structurally similar to known substrates, in terms of structural classification, functional classification, and size. (Of these, we use 953 in the training dataset and 100 in the testing dataset; see Table 1), We included these 1053 nonreactant "decoys" to enrich the existing set of "Zaretzki et al. nonreactants" [Recall that only some of those 679 molecules will react with any specific CYP450 isoform; see Table 1.] and to span a greater portion of the relevant chemical space of small molecules. These compounds include known drugs, pesticides, food compounds, pollutants, endogenous metabolites, and a variety of other compounds that, while plausible CYP450 reactants, are all known *not* to be metabolized by any of the nine selected CYP450 isozymes. We extracted these nonreactants from various databases, including the Human Metabolome Database,[17] the KEGG database,[18] DrugBank,[19] and the PubChem database.[20] In selecting the set of nonreactants, we explicitly avoided molecules that are obviously not metabolized by CYP450 isozymes—e.g., glycerolipids, glycerophospholipids, sphingolipids, inorganic compounds.[3,21] To be robust, the CypReact performance system handles these molecules separately, using a simple rule-based filter; see Figure 3.

Table 1. Data Distribution of the Nine CYP450 Isoforms[a]

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
Training Dataset Data Distribution										
#Reactant	271	105	151	142	226	218	270	145	475	679
#Non-Reactants	1361	1527	1481	1490	1406	1414	1362	1487	1157	953
#R / #Total	0.17	0.06	0.09	0.09	0.14	0.13	0.17	0.09	0.29	0.42
Hold-out Testing Dataset Data Distribution										
#Reactants	24	6	4	12	28	20	21	6	32	69
#Non-Reactants	100	100	100	100	100	100	100	100	100	100

[a]The light cyan colored rows correspond to the training datasets; note these datasets contain the same set of 1632 instances for each CYP450 isoform but different labels. The Hold-Out Testing Datasets (in yellow) have different reactant sets but the same nonreactant set.

We formed a training set for each of the nine selected CYP450 isozymes, consisting of the same 1632 compounds, but with different reactant/nonreactant labels, as a given compound might be a reactant for one CYP soform, but not for another. For instance, the anti-inflammatory drug, amodiaquine (DrugBank ID DB00613) is labeled as a reactant for CYP2C8, CYP2C19, CYP2D6, and CYP3A4 but as a nonreactant for CYP1A2, CYP2A6, CYP2B6, CYP2C9, and CYP2E1. As different CYP450 isoforms react with different molecules, the class distribution (reactant vs nonreactant) varied from one CYP450 isozyme to another. Table 1 shows

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

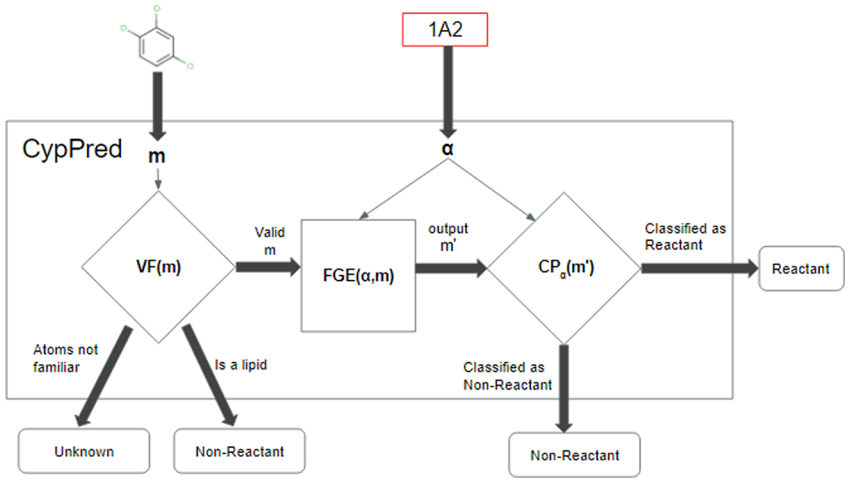


Figure 3. Components of the CypReact performance process.

the number of reactants, and nonreactants, for each of the nine data sets, as well as the union over all nine, labeled "All". We will let $D(\alpha)$ denote the data set associated with the "isoform" $\alpha \in \{1A2, 2A6, ..., 3A4, All\}$.

Feature Generation. Standard machine learning algorithms assume that each instance is described as a vector, whose components are values of certain "features". Here, we want to identify which properties or features associated with a molecule m are useful for determining whether m is a reactant versus a nonreactant.

We first performed several standardization operations to each of the 1632 compounds, to produce a precise description of each molecule. This involved removing salts, explicitly adding hydrogen atoms and generating a geometrically correct 3D structure for each molecule. Here, we used the Molconvert command-line tool from ChemAxon's Marvin Suite.[22]

Our LBM learning algorithm then considered a set of 2279 features for each molecule—selected based on their reported effect on the metabolism and the bioavailability of small molecules.[15,23,24] This included 36 physicochemical properties (such as molecular weight, and XLogP—each computed using the Chemistry Development Kit (CDK)[25]) and 2243 structure-based features, which includes the MACCS 166 fingerprint,[26] and 881 PubChem fingerprints.[20] Additionally, LBM used a ClassyFire[27] fingerprint, which consists of 1196 structural features encoded in the SMARTS language.[28] These include (1) functional group/chemical class definitions provided by ClassyFire, (2) structural patterns reported by the literature to correlate with reactivity to, or inhibition of CYP450 isozymes, (3) structural patterns of length 3−18 atoms obtained by mining the chemical structures of known CYP450 reactants and nonreactants, and (4) the MACCS 322 fingerprint (provided by Sud et al.[29]). The MACCS 166 fingerprint and the PubChem fingerprint are calculated using MACCSFingerprinter and the PubchemFingerprinter modules of the CDK library, respectively. The ClassyFire fingerprint was computed using the SMARTSQueryTool module of the CDK library. While the physicochemical properties were represented as numerical features, the structural features were represented as binary features to express the presence "1" or absence "0" of a specific structural feature within the molecule of interest.

Feature Selection. *Feature selection* is a technique, often used in machine learning, to select a subset of the features that the learner will use, to produce a classifier that uses only these features. Once identified, this makes the training phase faster and more efficient (as it involves fewer features) while also reducing the chance that the learner will overfit, as this means the learned model will involve relatively few parameters.

Recall that we initially selected 2279 features that are potentially useful for our task—e.g., the number of hydrogen bond acceptors, the sum of atomic polarizabilities, etc. However, some features contribute very little information. For example, while fingerprint features in general are potentially useful for our task, certain ones had values that were the same for all the molecules in the data set. As such features do not distinguish any molecules from one another, they of course cannot help in classification. Moreover, different features may have different degrees of importance for predicting the substrate specificity for each of the nine CYP450s—e.g., features that are critical to CYP1A2's substrate specificity, might be irrelevant to CYP2B6's substrate specificity.

Hence, in order to reduce the chance of overfitting and also to improve the computational efficiency, for each $D(\alpha)$ data set, our learning algorithm computed the information gain[30] of each feature with respect to the "reactant/nonreactant" label. This measures how important that feature is, for the given isoform, α. It then removed the features that appeared to be relatively uninformative—specifically removing all of the features with an information gain less than a threshold γ, which was learned by internal cross-validation; see below. Hence each CYP450 has its own unique feature set. Table 2 provides the numbers of features for each CYP450 reactant predictor.

LBM also normalized each feature f_i in each $D(\alpha)$ data set: Assume the values of f_i in $D(\alpha)$ are $\{x_i^j\}_j$. First let $b_i = \max_j\{x_i^j\}$ (respectively, $s_i = \min_j\{x_i^j\}$) be the maximum (respectively,

Table 2. Number of Features Selected by CypReact with Respect to Each CYP450 Enzyme[a]

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
# of features	469	421	274	563	536	509	495	263	934	1082
γ ‡	0.0075	0.001	0.0075	0.001	0.005	0.005	0.0075	0.0075	0.001	

[a]Note the "All" value corresponds to the union of the features over all nine isoforms. ‡$\gamma \in \{0.001, 0.005, 0.0075, 0.01, 0.03\}$ is the information gain threshold, found in the cross-validation process, used to find the number of features to use.

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

minimum) of these values. It then replaced each x_i^j with its normalized value

$$\hat{x}_i^j \leftarrow \frac{x_i^j - s_i}{b_i - s_i}$$

which is by construction in the range [0, 1].

Each D(α) data set uses these features to describe each molecule. (We will soon see that the FGE(α, m) process translates a molecule m, in SMILES or structure (sdf) format, into a vector of values for these features.)

Cost-Sensitive Learner. Most machine learning algorithms are designed to work best when the data set is relatively balanced—i.e., when the number of positive and negative cases (here, reactants vs nonreactants) is nearly the same. Our data set is, however, very imbalanced, as the number of reactants (~11%) is much less than the number of nonreactants (~89%). This is intentional, as it reflects the performance task that we anticipate for most of the scientists using our reactant predictor. In particular, we expect that very few of the molecules they will consider will actually be reactants. For instance, of the more than 400 000 known natural products, metabolites, and drugs, less than 10 000 molecules have been tested, of which fewer than 1000 are actually CYP450 reactants. In addition to this imbalance, we anticipate most users will consider false negatives (predicting a reactant to be a nonreactant) to be worse than false positives (predicting a nonreactant to be a reactant). Such users will prefer tools that rarely predict a reactant to be a nonreactant, even if this means (as an unavoidable side-effect) that those tools incorrectly predict several nonreactants to be reactants. After all, each false positive means the researcher may need to do a bit of extra work (e.g., run an extra experiment), before finding this mistake. However, each false negative means the researcher will (probably) just ignore this molecule, which might mean s/he may not bother to look for a metabolite. In the world of drug research, not knowing about a reaction means the researcher may miss a potential toxic metabolite or a potentially beneficial drug byproduct.

To emphasize the importance of false negatives over false positives, LBM uses a *cost sensitive learner*,[31] which involves a base learner (for instance, a support vector machine[32] or a neural network) and a cost matrix (such as Table 3[right]). It trains the base learner, seeking a classifier that minimizes the *total weighted cost*, which is the dot product of the given cost matrix and the confusion matrix, where a confusion matrix presents the number of each type of classification results produced by the classifier $C(\cdot)$ on the test data D—in particular, the number of true positives, false positives, false negatives, and true negatives; see Table 3[left]. (Note that "Reactant" is considered "True" and "Nonreactant" is considered "False".) A cost matrix presents the cost of each of these types of classification results as seen in Table 3[right]. Note that true positives and true negatives each cost 0 while the cost of each false positive is set to 1, and the cost of each false negative is set to β.

Table 3. Confusion Matrix of Classifier $C(\cdot)$ on Dataset D (Left) and Cost Matrix (Right)

Truth⇒ Prediction⇓	R	N
R	#True_Positives	#False_Positives
N	#False_Negatives	#True_Negatives

Truth⇒ Prediction⇓	R	N
R	0.0	1.0
N	β	0.0

Given this cost matrix, the "(Weighted) Cost" of a classifier $C(\cdot)$, based on its confusion matrix on a set of test data D, simplifies to the sum of the number of false positives, plus β times the number of false negatives.

$$\text{Cost}_\beta(C(\cdot), D) = \frac{(1 \times \#\text{False\_Positives} + \beta \times \#\text{False\_Negatives})}{|D|} \quad (1)$$

(We divide by the number of instances, $|D|$, to "normalize" the cost.)

Hence, this parameter β quantifies the trade-off between false-positives to false-negatives. For example, standard machine learning algorithms try to minimize the total (unweighted) number of mistakes, which is the sum of the number of false positives and false negatives. Hence, they implicitly assume that $\beta = 1$. As noted above, this is not appropriate here. Setting $\beta = 3.1$ means the learning algorithm would rather mistakenly claim that three nonreactants are reactants, rather than claim one reactant is a nonreactant.

To determine the appropriate value for β, we consulted with experts in the field, who collectively suggested we use a β between 3 and 7. Our subsequent sensitivity studies (e.g., using Cost Curves; see below) showed that the resulting classifiers were not particularly sensitive to the precise value in that range. We therefore selected the midpoint $\beta = 5$—that is, our system treats each false negative as five times as bad as a false positive. (While this paper focuses on this setting, our code-base allows the user to set this β parameters as s/he wishes.)

Our learning algorithm LBM($\cdot$) takes as input a labeled data set, here D_α (see top portion of Table 1), and implicitly the cost matrix shown in Table 3[right] and returns a classifier. This learned classifier, called CP_α [This CP_α represents the function CP(α, $\cdot$).], takes a representation of a molecule and returns {1, 0} (and occasionally "Unknown"; see below). We will see that this CP_α is the main part of the CypReact(α, $\cdot$) system, but there are also several other important components; see Figure 3.

For each isoform α, using the data set D(α), LBM considers five candidate base learners for the cost sensitive classifier: support vector machine SVM, logistic regression LR,[33] decision tree DT,[34] random forest RF,[35] and an ensemble method ES[36] that returns the majority class of the learned weak classifiers. Given the various parameter settings for some learners, there are 31 different learners+parameters. LBM first identifies the best base learner and also the best setting for its parameters, as well as the best threshold $\gamma \in \{0.001, 0.005, 0.0075, 0.01, 0.03\}$ for the feature selection process, by running an internal cross-validation process on its given entire data set D(α). This process involved dividing the given data set into five disjointed subsets. It then trains each of these learners on four of these five subsets, to produce 155 = 31 × 5 models (one for each of pair of [base_learner + parameter, value of γ]). It then evaluates each of these models on the remaining subset, which produced a single score (eq 1) for each of the models. It does this five times, each time holding out a different subset and, then, computes the average score (over these five iterations) for each of the 155 base_learner + parameter + γ settings. For each D_α, LBM found that the most accurate method was RF (random forest) for $\alpha \in$ {CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C19, CYP2E1, CYP3A4} and ES (ensemble methods) for $\alpha \in$ {CYP2C9, CYP2D6}. Table 2 shows the number of

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

Table 4. Five-Fold Cross-Validation (Top, in Cyan; Average ± Standard Deviation) and Hold-Out Testing (Bottom, Yellow) Weighted Cost of the CypReact, SMARTCyp, ADMET Predictor, and MajorityClassifier Models, for Each CYP450 Enzyme[a]

Classifier	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
5-fold CV results										
CypReact	0.313 ± 0.05	0.207 ± 0.03	0.278 ± 0.03	0.290 ± 0.05	0.359 ± 0.07	0.343 ± 0.06	0.296 ± 0.02	0.247 ± 0.05	0.289 ± 0.05	0.218
ADMET†	0.347	0.331	0.369	0.430	0.400	0.393	0.309	0.339	0.478	0.408
SmartCyp -React					0.682 ± 0.03		0.740 ± 0.05		0.702 ± 0.01	0.629‡
Majority Classifier	0.830	0.322	0.463	0.435	0.692	0.668	0.827	0.444	1.455	2.496
Hold-out testing results										
CypReact	0.177	0.038	0.077	0.143	0.141	0.217	0.099	0.104	0.152	0.183
ADMET†	0.298	0.320	0.288	0.375	0.500	0.475	0.190	0.311	0.333	0.497
SmartCyp -React					1.032		0.831		0.752	0.669‡
Majority Classifier	0.935	0.098	0.098	0.536	1.094	0.833	0.868	0.146	1.154	2.450

[a]Recall that smaller values of weighted cost are better. †ADMET is the abbreviation for ADMET Predictor. ‡These results are based on only the three isoforms that SMARTCyp covers: CYP2C9, CYP2D6, and CYP3A4.

features selected, for each isoform. Note that both of these base learners, RF and ES, involve consensus voting.[37] LBM then ran the selected base learner on the entire D(α) data set, which generated the model we will use—called CP$_\alpha$.

Implementation (See Figure 3). Recall our CypReact tool was trained on only compounds that were "plausible" CYP450 substrates—the set of 1632 summarized above. As noted, our training data intentionally did not include any molecule from classes of compounds that are obviously not CYP450 reactants—which means we ignored very large and hydrophobic molecules such as lipids (glycerolipids, glycerophospholipids, and sphingolipids) as well as inorganic compounds. We also noted that the training set included only molecules that contain only the following atoms: {H, C, O, N, S, F, Cl, P, Br, I}, which means we know the preprocessing can correctly handle those atoms.

To make our system more robust, we want to allow users to enter any molecule. For most molecules, CypReact will be able to make an accurate assessment. But for some—e.g., the ones that include atoms that did not appear in any molecule in the training set—we cannot be as confident. We therefore wrote a molecular filter program, called VF(m), that makes a three-way decision, for any molecule m:

1. If m is in an excluded class (currently, any lipid), VF returns "No" (not a reactant) and exits.
2. If m includes any atom that is not "familiar" (i.e., not in the list above), VF returns "Unknown" and exits.
3. Otherwise, m is considered valid, and VF passes it to the main part of the CypReact process, to be labeled.

If the molecule m is valid (3 above), it will be passed to the FGE(α, m) function [FGE stands for feature generation and extraction.], which will re-express m as a set of values associated with molecular features relevant to the CYP α (such as "PubChem fingerprints"[20]). The resulting description, m', will be input into the trained CP$_\alpha$ model and classified. Our implementation is written in Java using the WEKA[38] APIs.

Related Systems. In general, a good way to understand how well a system works is to compare its performance to that of other similar systems. Below we describe two systems: one that performs the same task as our CypReact and another that performs a similar function.

ADMET Predictor. ADMET Predictor (Simulations Plus, Inc., Lancaster, California, USA) is a commercial software tool for predicting properties of chemical compounds, including whether a molecule is a reactant for a specific CYP450 enzyme—i.e., the same function as CypReact. We can therefore compare our tool directly to ADMET Predictor. (Of course, as we do not know the data set on which ADMET Predictor was trained, we do not know whether that training set included our test set; this means we do not know whether our estimate of ADMET Predictor's accuracy is optimistic as we may be testing its performance on its training set.)

Reactant-Predictor Variant of SMARTCyp. We also compare our tool with a reactant predictor variant of SMARTCyp,[10] which is a site-of-metabolism (SOM) predictor. In general, SMARTCyp(α, m, s) generates a score for a site s of a given molecule m, for any of three isoforms $\alpha \in$ {CYP3A4, CYP2D6, CYP2C9}, where lower scores means SMARTCyp thinks it more likely that that site will be a SOM. We can use SMARTCyp to produce a tool that predicts whether a given molecule is a reactant: Given that a molecule is a reactant if and only if at least one of its sites is a SOM, we create a tool SMARTCyp-React(α, m) that predicts whether m is a reactant of the isoform α, which is True whenever SMARTCyp$_\tau$(α, m, s) is below some learned threshold τ, for any site s.

We use a learning algorithm to learn τ by internal cross-validation—i.e., the learning algorithm considers various different thresholds to determine the threshold that has the best score. It then uses external cross-validation to estimate the weighted cost of SMARTCyp-React$_{\tau^*}$(α, ·), with this best τ^*.

All Variants of the Predictors. Some users may just want to know whether a molecule will react with any CYP450 isoform but do not care which one. We therefore consider the CypReact-All variant that predicts that a given molecule m as an "All-reactant" if and only if CP$_\alpha$ predicts it is a reactant, for at least one of the nine CYP450 isoforms α. (Note this uses the nine already trained {CP$_\alpha$} models—n.b., it does not train a new CP$_{All}$ model to optimize the weighted cost.) We used the same approach to create a combined model for ADMET

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

Predictor-All, over all nine isoforms, and also for SMARTCyp-React-All, over its three isoforms: CY2C9, CYPD6, and CYP3A4.

RESULTS AND DISCUSSION

Evaluation Criterion. As mentioned above, for each CYP isoform α, we first ran the LBM(D(α)) learning process to find the best model $CP_\alpha(\cdot)$, based on all of the training data. To evaluate the quality of this learned model, for each isoform α we then used a evaluation algorithm that ran this LBM($\cdot$) process five more times, as a form of external cross-validation.[34] That evaluation algorithm divided D = D(α) into five subsets, then it ran the entire LBM($\cdot$) process on four of these five subsets—recall that this LBM process will run internal cross-validation to identify the best base learner. (Note this might lead to different base learners and different values of γ, in different iterations.) It then ran the resulting learned classifier on the hold-out subset. It repeated this process five times and reported the average score. Note this means our evaluation algorithm will run each base learner + parameter (e.g., SVM) at least five times for the external cross-validation and another $5 \times 5 = 25$ times for the internal cross-validation runs, each time on a slightly different subset of the D(α) data set.

Average Weighted Cost. Based on the discussion above, our goal is to optimize the weighted cost (eq 1); this section reports those scores, for each of our various classifiers: CypReact, MajorityClassifier (which just returns "No, not a reactant" for each molecule, and so serves as a baseline), SMARTCyp-React (for the three CYP isoforms {CYP2C9, CYP2D6, CYP3A4} that it considers), and ADMET Predictor for all nine isoforms. Notice we also consider the "All" situation (see below). These results appear in the top (cyan-colored) portion of Table 4 and Figure 4. Note that the lower score

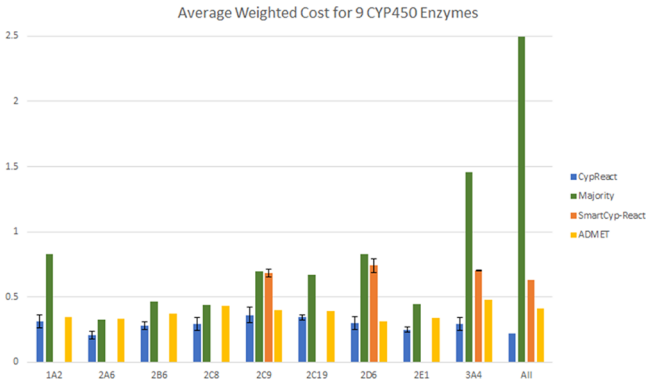


Figure 4. Average weighted cost for CypReact, SMARTCyp-React, and ADMET Predictor (lower is better).

means better performance: a perfect result is 0, and the weighted cost of the baseline (MajorityClassifier) varies from 0.322 to 1.455. Paired two-sided t tests showed that each CYP450 predictor in CypReact is statistically significantly better than the baseline, at $p < [1.91E{-}5, 1.56E{-}3, 1.02E{-}4, 1.89E{-}3, 1.68E{-}4, 9.1E{-}6, 1.95E{-}5, 2.83E{-}5, 8.29E{-}7]$ over the nine CYPs (in order shown in Table 4). After applying Bonferroni correction, we can claim that all are significantly ($p < 0.0056$) better than the baseline. We also see that our CypReact is statistically better than SMARTCyp-React, for $\alpha \in$ {CYP2C9, CYP2D6, CYP3A4}, at $p < [3.38E{-}6, 8.63E{-}7, 1.06E{-}7]$.

The final column of Table 4 shows that CypReact-All performs better than ADMET Predictor-All and SMARTCyp-React-All.

Jaccard Scores. Another obvious measure to deal with imbalanced data is the Jaccard score, which is intersection over union, with respect to the minority class:

$$\text{Jaccard} = \frac{\#\text{True\_Positives}}{\#\text{True\_Positives} + \#\text{False\_Positives} + \#\text{False\_Negatives}}$$

The top (cyan-colored) portion of Table 5 reports the Jaccard score for each of these classifiers; note these are the same classifiers discussed above—i.e., each is still trained to optimize the weighted loss function.

A simple paired t test shows that CypReact is statistically significantly better than the baseline, at $p < [4.17E{-}6, 2.60E{-}4, 2.36E{-}5, 1.46E{-}4, 4.41E{-}5, 5.01E{-}6, 3.23E{-}5, 6.44E{-}6, 3.25E{-}6]$ over the nine CYPs. CypReact is also statistically better than SMARTCyp-React, for all three isoforms considered, at $p < [4.90E{-}6, 5.25E{-}6, 1.54E{-}7]$.

The final column of Table 5 shows that CypReact-All performs better than ADMET Predictor-All and SMARTCyp-React-All, in terms of this criterion as well.

Cost Curves. Above, we motivated the use of a cost-sensitive learner and suggested we learn classifiers that optimize eq 1, with $\beta = 5$. Below we show the confusion matrix for the CypReact classifier for the CYP2D6 isoform (see Table 3):

The previous sections evaluated this classifier, using the evaluation function eq 1, with $\beta = 5$—which we will write as eq 1 [$\beta = 5$]. We can also consider evaluating simple variants of this classifier, and others, with respect to other values of β.

To be more precise: the core component of each learned CypReact system actually returns a score for each input molecule m; the β value is used to set a threshold $\tau(\beta) = \frac{1}{\beta+1} \in [0, 1]$, for determining whether that molecule should be labeled Reactant—here m is labeled "Reactant" if that score is larger than $\tau(\beta)$ and otherwise, "NonReactant". Equation 2 corresponds to the performance-time value of $\beta = 5$; we clearly produce different confusion matrices for other values of β.

Truth⇒ Prediction⇓	R	N
R	$\#\text{True\_Positives}_{\beta=5} = 235$	$\#\text{False\_Positives}_{\beta=5} = 308$
N	$\#\text{False\_Negatives}_{\beta=5} = 35$	$\#\text{True\_Negatives}_{\beta=5} = 1054$

(2)

This idea motivates "Cost Curves":[39] a curve of (x, y) pairs, where each x-value corresponds (indirectly) to a value of β, and the y-value measures how well this fixed classifier does, with respect to this β. The orange curve in Figure 5 corresponds to the CypReact(2D6, $\cdot$) classifier, based on the points (x_β, y_β), computed as

$$x_\beta = \frac{p(R)M(N|R)}{p(R)M(N|R) + [1 - p(R)]M(R|N)} = \frac{0.17\beta}{0.17\beta + 0.83 \times 1} \quad (3)$$

$$y_\beta = y_\beta(C) = \text{FN}(C)x_\beta + \text{FP}(C)(1 - x_\beta) \quad (4)$$

where in general

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

Table 5. Five-Fold Cross-Validation (Top, Cyan; Average ± Standard Deviation) and Hold-Out Testing (Bottom, Yellow) Jaccard Score of the CypReact, SMARTCyp, and ADMET Predictor Models, for Each CYP450 Enzyme[a]

Classifier	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
5-fold CV results										
CypReact	0.389 ± 0.03	0.275 ± 0.05	0.282 ± 0.03	0.251 ± 0.04	0.302 ± 0.04	0.304 ± 0.02	0.406 ± 0.04	0.306 ± 0.02	0.545 ± 0.03	0.687
ADMET†	0.379	0.157	0.201	0.157	0.286	0.278	0.448	0.211	0.463	0.506
SmartCyp -React					0.092 ± 0.03		0.164 ± 0.03		0.296 ± 0.01	0.369‡
Hold-out testing results										
CypReact	0.605	0.455	0.364	0.556	0.651	0.567	0.714	0.375	0.593	0.690
ADMET†	0.488	0.150	0.118	0.231	0.385	0.298	0.621	0.147	0.437	0.459
SmartCyp -React					0.094		0.143		0.248	0.331‡

[a]We did not show the Majority Classifier as it was 0.0 for all isoforms. Recall that larger values of Jaccard score are better. †ADMET is the abbreviation for ADMET Predictor. ‡These results are based only on the three isoforms that SMARTCyp covers: CYP2C9, CYP2D6, and CYP3A4.

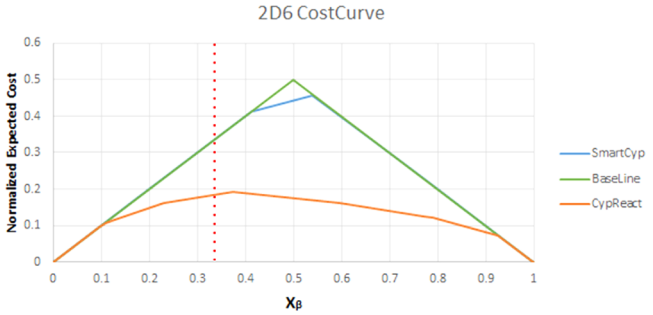


Figure 5. CostCurves for CypReact(2D6, ·) in orange, SMARTCyp-React(2D6, ·) in blue, and the baseline in green (covering much of SMARTCyp-React(2D6, ·)). The red vertical dashed line corresponds to $\beta = 5$ here. We see that CypReact dominates SMARTCyp-React over all x_β values—which means for all misclassification costs, β.

- $p(R)$ is the ratio of reactants over all instances (which corresponds to the bottom cyan-colored row of Table 1, "#R/#Total"—and so is 0.17 for our data set).
- $M(N|R)$ is the misclassification cost of predicting an instance with real label Reactant as NonReactant—which recall we defined as β—and the other misclassification cost $M(R|N)$, here is set to 1
- $\mathrm{FN}(C) = \frac{\#\,\text{False\_Negatives}}{\#\,\text{False\_Negatives} + \#\,\text{True\_Positives}}$ is the false negative rate for this classifier—which, using eq 2, is $\frac{35}{35+235} \approx 0.13$ for C = CypReact(2D6, ·) and $\mathrm{FP}(C) = \frac{\#\,\text{False\_Positives}}{\#\,\text{False\_Positives} + \#\,\text{True\_Negatives}}$ is the false positive rate—here $\frac{308}{308+1054} \approx 0.23$

(Here, we include C as an argument of y_β, FN, and FP, to show its dependence.)

With a little algebra, using eq 1, we find that

$$y_\beta(C) = \frac{\mathrm{cost}_\beta(C, D)}{p(R)M(N|R) + [1 - p(R)]M(R|N)} = \frac{\mathrm{cost}_\beta(C, D)}{0.17\beta + 0.83 \times 1} \quad (5)$$

which is why it is often called "Normalized Expected Cost". Now notice that the denominator does not depend on the classifier, which means a classifier that optimizes eq 1 will be optimizing this $y_\beta(C)$ value.

Note the x values are independent of the classifier itself and so can vary independently. This allows us to compare different classifiers, over a range of different β-values, to see when each classifier is best. [In addition, we could consider other "label distributions": While our training data set had a 17-to-83 mix of Reactants to NonReactants (see the bottom cyan-colored row of Table 1), we could alternatively consider a data set that had a 20-to-80 mix, 50-to-50, or whatever, by varying the $p(R)$ value. However, we did not do this here.] This is why we consider the full range of values $x_\beta \in [0, 1]$ for the x-axis and, then, use eq 4 to compute the associated normalized expected cost y_β (which is related to Cost(·); see eq 5). In operation, the user would first identify the Cost Matrix (Table 3), which here means stating the β value. That user would then use eq 3 to compute the x_β value and, then, adjust the classifier to this value of β—call it C^β—which updates the classifier's confusion matrix, which is then used to determine the associated $y_\beta(C^\beta)$ cost.

We can also see how well other classifiers would perform over the entire range of β values, which induces values for both x_β values $\in [0, 1]$ and then y_β, based on x_β and the confusion matrix (based on β). We can consider some trivial classifiers: The "JustSayN" classifier just returns NonReactant for each instance; it is easy to see that, for any x, its normalized-expected-cost (i.e., its y-value) will be the $y = x$ line. There is no reason for any classifier to ever be above this line—i.e., if for any x_β value, a classifier $C(\cdot)$ had a cost that was above this $y_\beta = x_\beta$ line, it would be silly to use $C(\cdot)$, as we would get a better score by just ignoring that $C(\cdot)$ classifier and instead using the JustSayN classifier.

Similarly, the cost curve for the JustSayR classifier, which just returns Reactant, would trivially be the $y = 1 - x$ line. Again, there is no reason to consider a classifier that is above that line. We consider the minimum of these two lines to be the "Baseline"—shown as the green line in Figure 5—and, for any classifier, will only show the cost-curve portion that appears below this curve.

The blue line in Figure 5 shows the curve for SMARTCyp-(Cyp2D6, ·). We see that it matches the Baseline for much of the domain $x_\beta \in [0, 1]$, dipping below only around $x_\beta \in (0.41, 0.54)$. Moreover, we see that our CypReact(2D6, ·) system is strictly better (that is, smaller) than SMARTCyp(Cyp2D6, ·) for many x_β values, and it is never worse.

This suggests that one should prefer the CYP2D6 model of CypReact over the one of SMARTCyp as CypReact is always at least as good, and often better. (While it did not happen here,

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

the curves for different classifiers could cross—meaning there would be a region of x_β-values where classifier #1 is best and another where classifier #2 is best. Here, once we knew the β value for the target domain, we could compute the x_β value and, then, find which classifier is best here—that is, use $C_\beta = \arg\min_C\{y_\beta(C)\}$.)

We also found that CypReact is similarly superior to SMARTCyp-React for CYP3A4 and CYP2C9; see the Cost Curves for CYP3A4 and CYP2C9 in the Supporting Information.

ROC and AUC. CostCurves allow the user to decide, for each β, which specific classifier to use—meaning one might use one 2D6 classifier for $\beta = 5$, here corresponding to the value $x_\beta = 0.51$, but another classifier for $\beta = 8$ (leading to $x_\beta = 0.62$). If one just wanted to use a single classifier, we could evaluate a classifier based on its AUROC value, which essentially measures how well its performance "on average", over the entire range of β values. In general, a classifier's ROC curve is a set of (x, y) points, where here x is the FalsePositiveRate and y is the TruePositiveRate, as you vary some natural parameter. Note that the shape of the ROC curve for a perfect classifier is essentially a Gamma "Γ", while the baseline is a diagonal line ("/") with a slope of one. This means the AUROC of a perfect classifier is 1.0 and of the baseline is 0.5.

Figure 6 shows the ROC curves for CypReact and SMARTCyp-React for 2D6, as well as the baseline "random guess" classifier. We see that CypReact performs much better than SMARTCyp-React here—with AUROC of 0.872 versus 0.490. Table 6 shows the AUROC values for all nine isoforms, showing they range from 83% to 92% for CypReact and from 49% to 60% for SMARTCyp. (The Supporting Information presents the ROC curves for the CypReact classifier for the other eight CYP450 isoforms and for SMARTCyp where relevant.)

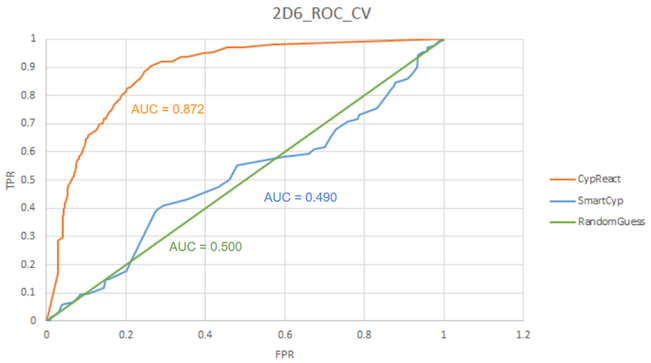


Figure 6. ROC curve of CypReact and SMARTCyp-React for CYP2D6. (Note we did not take the convex hull, to better illustrate the shapes.)

Table 6. Area under ROC of CypReact on the Nine CYP450 Isoforms

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
CypReact	86%	84%	86%	84%	83%	83%	87%	87%	92%
SmartCyp-React					51%		49%		60%
ADMET Predictor	79%	77%	74%	68%	74%	75%	81%	75%	75%

Results on a New Data Set. After computing the cross-validation scores on the training/testing set, LBM then learned nine CypReact models, each based on *all* 1632 molecules, then tested these learned models on new, disjoint data sets—one for each isoform α. We produced these data sets by first identifying 69 new molecules that were reactants to at least one isoform and combining them with 100 molecules that are known to be nonreactants to all 9 isoforms; see the bottom three rows (colored yellow) in Table 1.

The lower (yellow-colored) portions of Tables 4 and 5 shows the results of these learned CypReact algorithms on these validation sets—showing (respectively) average weighted cost and Jaccard scores. It also presents the results of SMARTCyp-React and ADMET Predictor on these data sets.

These results confirm that CypReact works extremely well, and in particular, better than the other CYP450 reaction prediction systems considered.

CONCLUSION

CypReact is a family of CYP450 reactant-predictors that contains nine subtools, each built for one CYP450 enzyme individually. Each CypReact classifier is trained to minimize the average weighted cost score for its associated CYP450 isoform, based on a weighted cost that penalizes each false negative five times more than each false positive. Our empirical results show that our classifiers exhibit very good weighted cost scores, and AUROC scores—here ranging from 83% to 92%—and that they significantly outperform SMARTCyp-React and ADMET Predictor.

While our CypReact family of classifiers work extremely well, there is still room for improvement. Due to the relatively low number of reactants in the training data set, our predictors $\{CP_\alpha\}$ may be imperfect. Further, as the cost matrix imposes a high cost of predicting a reactant as a nonreactant, the precision of our various predictors (with respect to predicting reactants) may be low. To improve CypReact in the future, we plan (1) to collect more molecules (both reactants and relevant non-reactants) and (2) to identify more useful chemical and molecular features and use feature engineering techniques to create other, perhaps more effective, features.

Currently, most existing in silico metabolism prediction tools assume that the input molecule is a CYP450 substrate. Given that the vast majority of molecules, and even known drugs and xenobiotics, are *not* CYP450 substrates, simply assuming every molecule is a reactant may lead to far too many errors in terms of downstream in silico metabolism prediction. This may lead growing doubt about the utility and effectiveness of in silico metabolism predictors. It is therefore important to have predictive tools that can distinguish reactants from non-reactants—e.g., to use as a filter. CypReact is currently the only such publicly available tool. We believe that the development and use of CypReact will help improve the utility of in silico metabolism predictors. We also believe that CypReact will find applications not only in the field of drug metabolism research but also in the fields of metabolomics, exposure science, and food and nutrition research. Indeed, CypReact has already been used in a metabolomics pipeline to help identify novel food-derived terpenoids (manuscript in preparation).

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00035.

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291

Cost Curves and ROC Curves for the other eight cytochrome P450 isoforms (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: rgreiner@ualberta.ca.

ORCID

Siyang Tian: 0000-0002-7298-2520

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from NSERC (Natural Sciences and Engineering Research Council of Canada), AMII (Alberta Machine Intelligence Institute), CIHR (Canadian Institutes of Health Research), and Genome Canada. We also thank the developers of ADMET Predictor and SMARTCyp for developing superb software tools and for letting us use them for our research.

REFERENCES

(1) Van De Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise. *Nature reviews. Drug discovery* **2003**, *2*, 192.

(2) Delaney, K. A.; Kleinschmidt, K. C. In *Goldfrank's Toxicologic Emergencies, 9e*; Nelson, L. S., Lewin, N. A., Howland, M. A., Hoffman, R. S., Goldfrank, L. R., Flomenbaum, N. E., Eds.; The McGraw-Hill Companies: New York, NY, 2011.

(3) Furge, L. L.; Guengerich, F. P. Cytochrome P450 enzymes in drug metabolism and chemical toxicology: An introduction. *Biochem. Mol. Biol. Educ.* **2006**, *34*, 66−74.

(4) Nebert, D. W.; Russell, D. W. Clinical importance of the cytochromes P450. *Lancet* **2002**, *360*, 1155−1162.

(5) Pan, Z.; Raftery, D. Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. *Anal. Bioanal. Chem.* **2007**, *387*, 525−527.

(6) EAWAG-BBD Pathway Prediction System. http://eawag-bbd.ethz.ch/predict (accessed 2017-09-20).

(7) Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt, L. J.; Hanson, A. D.; Fiehn, O.; Tyo, K. E.; Henry, C. S. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminf.* **2015**, *7*, 44.

(8) Anzenbacher, P.; Anzenbacherová, E. Cytochromes P450 and metabolism of xenobiotics. *Cell. Mol. Life Sci.* **2001**, *58*, 737−747.

(9) Rostkowski, M.; Spjuth, O.; Rydberg, P. WhichCyp: prediction of cytochromes P450 inhibition. *Bioinformatics* **2013**, *29*, 2051−2052.

(10) Rydberg, P.; Gloriam, D. E.; Olsen, L. The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* **2010**, *26*, 2988−2989.

(11) Adams, S. E. Molecular similarity and xenobiotic metabolism. Ph.D. thesis, University of Cambridge, 2010.

(12) ADMET Predictor. Simulations Plus, Inc., Lancaster, California, USA; https://www.simulations-plus.com/software/admetpredictor/metabolism/ (accessed 2018-04-03).

(13) StarDrop. https://www.optibrium.com/stardrop/ (accessed 2017-05-21).

(14) Elloumi, M.; Iliopoulos, C.; Wang, J. T.; Zomaya, A. Y. *Pattern Recognition in Computational Molecular Biology: Techniques and Approaches*; John Wiley & Sons, 2015.

(15) 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.

(16) Molecular Networks. isoCYP−Predictions of the predominant isoform of human cytochrome P450 substrates. www.molecular-networks.com/products/isocyp (accessed 2017-06-03).

(17) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. HMDB 3.0-the human metabolome database in 2013. *Nucleic Acids Res.* **2012**, *41*, D801−D807.

(18) KEGG Database. http://www.genome.jp/kegg/kegg1.html (accessed 2017-08-03).

(19) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−D1097.

(20) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem substance and compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(21) Gunstone, F. D.; Harwood, J. L.; Dijkstra, A. J. *The lipid handbook with CD-ROM*; CRC press, 2007.

(22) ChemAxon's Marvin Suite. 2017; https://www.chemaxon.com/download/marvin-suite/ (accessed 2017-11-25).

(23) Ioannides, C. *Cytochromes P450: role in the metabolism and toxicity of drugs and other xenobiotics* **2008**, DOI: 10.1039/9781847558428.

(24) Wilson, A. G. *New Horizons in Predictive Drug Metabolism and Pharmacokinetics* **2015**, DOI: 10.1039/9781782622376.

(25) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.* **2017**, *9*, 33.

(26) BIOVIA: The keys to understanding MDL keyset technology. http://accelrys.com/products/pdf/keys-to-keyset-technology.pdf (accessed 2017-11-10).

(27) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminf.* **2016**, *8*, 61.

(28) SMARTS−A Language for Describing Molecular Patterns.http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 2017-01-25).

(29) Sud, M. Mayachemtools: an open source package for computational drug discovery. *J. Chem. Inf. Model.* **2016**, *56*, 2292−2297.

(30) Quinlan, J. R. Induction of decision trees. *Machine learning* **1986**, *1*, 81−106.

(31) Elkan, C. The foundations of cost-sensitive learning. *International joint conference on artificial intelligence*; 2001; pp 973−978.

(32) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their applications* **1998**, *13*, 18−28.

(33) Hosmer, D. W., Jr; Lemeshow, S.; Sturdivant, R. X. *Applied logistic regression*; John Wiley & Sons, 2013; Vol. *398*.

(34) Alpaydin, E. *Introduction to machine learning*; MIT press, 2014.

(35) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5−32.

(36) Dietterich, T. G. Ensemble methods in machine learning. *Multiple classifier systems* **2000**, *1857*, 1−15.

(37) Ballabio, D.; Biganzoli, F.; Todeschini, R.; Consonni, V. Qualitative consensus of QSAR ready biodegradability predictions. *Toxicol. Environ. Chem.* **2016**, *99*, 1193−1216.

(38) Frank, E.; Hall, M. A.; Witten, I. H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical machine learning tools and techniques"*; Morgan Kaufmann, 2016.

(39) Drummond, C.; Holte, R. C. Cost curves: An improved method for visualizing classifier performance. *Machine learning* **2006**, *65*, 95−130.

DOI: 10.1021/acs.jcim.8b00035
J. Chem. Inf. Model. 2018, 58, 1282−1291