# Rule-Based Prediction Models of Cytochrome P450 Inhibition

Bo-Han Su,[§] Yi-shu Tu,[‡] Chieh Lin,[§] Chi-Yu Shao,[‡] Olivia A. Lin,[‡] and Yufeng J. Tseng*,[‡,§]

[‡]Graduate Institute of Biomedical Electronics and Bioinformatics and [§]Department of Computer Science and Information Engineering, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei, Taiwan 106

ⓢ *Supporting Information*

ABSTRACT: Hepatotoxicity, drug-induced liver injury, and competitive Cytochrome P-450 (CYP) isozyme binding are serious problems associated with drug use. It would be favorable to avoid or to understand potential CYP inhibition at the developmental stages. However, current in silico CYP prediction models or available public prediction servers can provide only yes/no classification results for just one or a few CYP enzymes. In this study, we utilized a rule-based C5.0 algorithm with different descriptors, including PaDEL, Mold[2], and PubChem fingerprints, to construct rule-based inhibition prediction models for five major CYP enzymes—CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4—that account for 90% of drug oxidation or hydrolysis. We also developed a rational sampling algorithm for the selection of compounds in the training data set, to enhance the performance of these CYP prediction models. The optimized models include several improved features. First, the final models significantly outperformed all of the currently available models. Second, the final models can also be used for rapid virtual screening of a large set of compounds due to their ruleset-based nature. Moreover, such rule-based prediction models can provide rulesets for structural features related to the five major CYP enzymes. The five most significant rules for CYP inhibition were identified for each CYP enzymes and discussed. An example was chosen for each of the five CYP enzymes to demonstrate how rule-based models can be used to gain insights into structural features that correspond with CYP inhibitions. A newer version of the freely accessible CYP prediction server, CypRules, is presented here as a result of the aforementioned improvements.
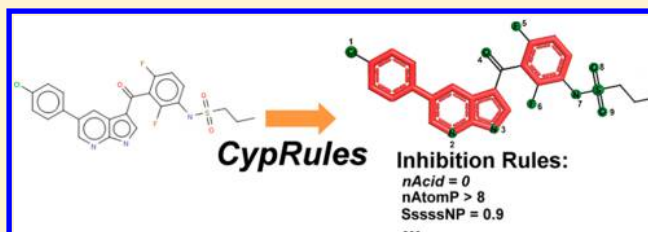
■ INTRODUCTION

In humans, more than 50 CYP enzymes have been identified. Among them, five CYP enzymes are responsible for approximately 90% of drug metabolism: CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4.[1,2] These enzymes facilitate reactions including N-, O-, and S-dealkylation, aromatic-, aliphatic-, and N-hydroxylation, and N-oxidation, sulfoxidation, deamination, as well as dehalogenation.[3] More than 900 drugs and natural chemicals were found to cause liver damage that could lead to a necessary liver transplantation operation or death.[1,4,5] In addition, hepatotoxicity and drug-induced liver injuries are among the top reasons why many drug candidates failed in clinical trials; they are also reasons why some FDA-approved drugs were recalled or withdrawn from the market. Therefore, detecting potential hepatotoxicity early in the drug development processes, using models to study CYP interactions with drug-like compounds, has been a popular topic in the past decade. Most of these types of studies have applied different machine learning approaches, including decision tree induction,[6] back-propagation artificial neural networks,[7] recursive partition,[8] Gaussian kernel weighted *k*-nearest neighbor,[9] associative neural networks,[10] and support vector machine[11−14] methodologies. However, the majority of CYP prediction models/servers were built from a small number of compounds so the applicability of these models are not optimal. Most importantly, all of these studies only provide yes/no classification results. Currently, there are no rule-based

CYP450 inhibitory classification models for virtual screening, nor rule-based predictors online.

Rule-based classification models could provide several advantages over the other models that have been described in the literature. Notably, not only can rule-based models act as screening templates like other machine learning based classification models, they can also specify rulesets of structural features that directly contribute toward specific P450 inhibitions. These rulesets provide easier interpretation and act as a guide for medicinal chemists to design or synthesize new compounds, without potentially inhibiting a specific CYP enzyme by avoiding those structural features altogether. Another advantage of using a rule-based classification model is the performance speed; as the nature of a rule-based model is generally very fast.

An issue to be considered when constructing rule-based models is the imbalanced data from large high throughput screening data sets, especially CYP2D6 data sets because the number of CYP2D6 inhibitors is very low (only 19% of the data in the CYP2D6 training sets are inhibitors). This probably explains why CYP2D6 classification models in the previous studies were the least accurate compared to other CYP models. For imbalanced data, developing a good strategy to select

representative compounds for the training set will enhance the performance of subsequent classification models.[15]

Quantitative structure–activity relationship (QSAR) approaches on P450 metabolism have been widely developed over the last two decades and extensively reviewed in several articles.[16−23] By corresponding biological CYP inhibitory activities with structural features and descriptors, QSAR analyses are advantageous in two ways: they present predictions of CYP inhibitory activity as quantitative values and they also assist in discerning the key structural features of compounds contributing to their inhibition potency. However, the QSAR-based approach works the best with analogs. Despite the fact that there have been numerous rule-based QSAR models developed over the years, still there are currently no rule-based QSAR models for the five major P450 isozyme readily available in the form of a web server, for users to publically access right away.

In this study, we have chosen five of the most common CYP enzymes to construct their corresponding inhibition prediction models, based on screening data for more than 16 000 compounds. The five CYP enzymes—CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4—were chosen, because together they account for approximately 90% of the drug metabolism. Our goals are (a) to build statistically high performance classification models using rule-based C5.0 algorithm, that directly provide structural information contributing to P450 inhibition in a faster and more direct manner, compared with other machine learning methods, and (b) to make these prediction models simple to use, as well as freely available on the web server, allowing users to perform quick inhibition assessment of compounds on the five major CYP enzymes. Our classification models can generate rulesets to give users insight on how chemical structures are related to P450 inhibition. We have previously published a web server, CypRules,[24] in the application notes describing the earlier version of our rule-based C5.0 algorithm prediction models, capable of predicting and providing structural rulesets for CYP inhibition, for any compound uploaded to the server. Here, we provide a detailed illustration and discussion of our methods and results. In addition, we've also designed a rational sampling algorithm to considerably improve the performance of our prediction models. We applied the rational sampling algorithm to our previous models under the published CypRules (version 1.0) web server. The optimized models, CypRules (version 2.0), presented here as a result of such improvement, significantly outperformed all of the models in all of the previous studies. Furthermore, we've identified the top five most significant CYP inhibition rules for each of the CYP models discussed here, we've also chosen known CYP inhibitors to demonstrate how rulesets from our models can relate structural features to each CYP inhibitors.

■ MATERIAL AND METHODS

Hepatotoxicity End Points. The data set used in this study was collected from the National Institutes of Health Chemical Genomics Center (NCGC) cytochrome panel assay's PubChem BioAssay database (AID 1851), using the quantitative high-throughput screening (qHTS) technique.[25,26] The detailed NCGC assay protocol is provided online (PubChem BioAssay AID 1851). The cytochrome P450 panel assay was used to determine CYP inhibitory activity through the measurement of luciferin−luciferase bioluminescence. Luciferin is a substrate for luciferase enzymes; after

luciferase was added to the assay to produce light, luciferin can be measured by luminescence. According to the protocol, the presence of inhibitors will limit the production of luciferin, therefore reduce measurable luminescence of luciferin. In this way, the inhibitory effect of these compounds on P450 isozymes were determined via the measurement of light intensity.[26]

The initial data set taken directly from the NCGC PubChem BioAssay database contains 17 143 compounds. However, not every single one of these compounds is either a known inhibitor or noninhibitor of CYP enzymes 1A2, 2C19, 2C9, 2D6, or 3A4 respectively. For this reason, the selection of appropriate compounds for each of the CYP enzymes is carefully considered. First and foremost, the compounds that are known to be inorganic, or only contain salts that provide no chemical information, were eliminated from the data set. Next, compounds that were found to be ambiguous (could be classified as both an inhibitor and noninhibitor for the one specific P450 isozyme), or have structural duplicates, were eliminated from the data set for that particular isozyme. Consequently, not all of the 17 143 compounds were screened against all five of the P450 isozymes, and the final numbers of active and inactive compounds in the refined data set for each P450 end point are listed in Table 1. In this way, the final data

Table 1. Number of Active and Inactive Compounds Screened from the PubChem BioAssay Database (AID 1851)

	CYP1A2	CYP2C19	CYP2C9	CYP2D6	CYP3A4
no. of active compounds	5891	5828	4090	2647	5177
no. of inactive compounds	6912	7062	8380	10869	7456

sets used to build the rule-based CYP 1A2, 2C19, 2C9, 2D6, and 3A4 models, respectively, are specific to each of the enzyme. At last, the compound structures in each of these data sets were screened to ensure that there are no redundant ions, then they were converted to three-dimensional structures in preparation for the following 1D, 2D, and 3D molecular descriptor calculations.

Descriptor Sets. In this study, 1D, 2D, and 3D trial descriptors were used, alone or in combination, to develop classification models for the five chosen P450 end points. The 1D and 2D descriptors were calculated by PaDEL-Descriptor, Mold[2], and PubChem fingerprints individually, and the 3D descriptors were calculated by PaDEL-Descriptor alone.

PaDEL is a software developed by the National University of Singapore; it is available free of charge.[27] The software currently calculates 1875 descriptors, 1444 of which are 1D and 2D descriptors, the remaining 431 being 3D descriptors. The descriptors were calculated with the aid of the Chemistry Development Kit and several additional programs. These additional descriptors include: atom type electrotopological state descriptors,[28] Crippen's logP and molar refractivity (MR),[29] extended topochemical atom (ETA) descriptors,[30] McGowan volume,[31] molecular linear free energy relation descriptors,[32] ring counts, and count of chemical substructures identified by Laggner.[33] For PaDEL descriptors calculation, 1D and 2D descriptors were calculated simultaneously and treated as one descriptor set (PaDEL1&2D), while 3D descriptors were treated as an independent descriptor set (PaDEL3D). Sixty-eight of the PaDEL3D descriptors were removed in our

study because the values of these descriptors in most of our compound data set cannot be calculated.

Mold[2] is software, also available free of charge, developed to enable the rapid calculation of 777 1D and 2D descriptors encoding two-dimensional chemical structure information.[34] Comparative analysis of Mold[2] descriptors with Cerius,[35] Dragon,[36] and Molconn-Z[37] on several data sets using Shannon entropy,[38] demonstrated that Mold[2] descriptors are capable of generating a similar amount of information.[34] It has been noted in the literature that when the same classification method was used, Mold[2] descriptors typically produce slightly better models compared to those generated using equivalent descriptors from commercial software packages. In addition, Mold[2] consumes lower computing power and provides a moderate amount of chemical structure information. For these reasons, Mold[2] is suitable and consequently used for the virtual screening of large databases.

PubChem fingerprints were also used to generate 1D and 2D descriptors for our data set.[39] PubChem fingerprints database consists of 881 bits of descriptors related to element counts, aromatic or nonaromatic ring counts, atom pairs, atom neighborhoods, and specific fragments.

Construction of Classification Models. All different combinations of three descriptor sets were selected to build the classification models. To evaluate their classification performance for each of the chosen P450 end point, the corresponding data set was randomly divided into a training set and a testing set of 8:2 ratio initially. Next, the training set for one end point was then trained and evaluated by the testing set, using two machine learning methods: (a) rule-based C5.0 algorithm and (b) support vector machine (SVM) respectively. Our goal is to determine, for each P450 end point, which combination of the three descriptor sets will yield the highest relative accuracy in classifying active versus inactive compounds, when one of the two machine learning methods was applied.

Rule-Based C5.0 Algorithm. C5.0 is a decision tree generating algorithm derived from its well-known predecessor, C4.5 algorithm.[40] The preceding C4.5 algorithm was developed by Ross Quinlan[40] using the concept of information entropy. At each node on the decision tree, the C4.5 algorithm chooses an attribute from the data that most effectively divides the initial set of samples into subsets enriched in one class or the other. This splitting criterion is known as the normalized information gain (or the difference in entropy). Ultimately, the attribute which results in the highest normalized information gain is then chosen to make the subsequent decision.

Decision trees can sometimes be quite difficult to understand. An important feature of the C5.0 algorithm is its ability to generate classifiers called "rulesets" which consist of unordered collections of (relatively) simple if-then rules. The Rule-based C5.0 algorithm offers a number of other improvements on the C4.5 algorithm as well.[16] First of all, the C5.0 algorithm is significantly faster than the C4.5 algorithm.[40] It is more memory-efficient, and is capable of generating similar results compared to the C4.5 algorithm, but with considerably smaller decision trees. Last but not least, the C5.0 algorithm is able to provide the rulesets for which the predicted results were based on. The rulesets obtained from this algorithm can then be used to support further inspection, to decipher the relationship between chemical compounds and the molecular descriptors used to classify them.

Compared to our previous *CYPRules* publications,[24] we further adopted the boosting mode in the C5.0 algorithm to enhance the accuracy of our models. Under the boosting mode, the system adaptively constructs at least four rule-based models and uses these rule-based models to classify a compound as either active or inactive by the majority rules. For simplicity's sake, our revised CYPRules website will only show the five descriptors which appear with the highest frequency among the rules that correctly predicted the input compound in the majority class.

Support Vector Machine. The concept and implementation of SVM was proposed by Vapnik et al. in 1995.[41,42] It is a (kernel function based) supervised machine learning technique that is primarily used to separate compounds into binary classes. When the compounds in a training set are linearly separable, SVM will divide these compounds into two classes of molecules with a maximum margin hyperplane. The maximum margin on either side of the hyperplane is defined as the largest distance to the nearest training data points. For nonlinear cases, SVM will project feature vectors (molecular descriptors) onto a transformed high-dimensional feature space—similar to an energy landscape—and search to fit the maximum margin hyperplane in the multidimensional feature space. Finally, SVM uses the traditional training and testing sets approach; in which SVM is trained using a set of data with known classification, then applied to another data set to evaluate the trained SVM model's ability to classify other compounds.

Comparison of Classification Models. To evaluate the predictive performance of classification models, *accuracy*, *sensitivity*, and *specificity* are defined as follow:

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fn} + \text{tn} + \text{fp}} \tag{1}$$

$$\text{sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{2}$$

$$\text{specificity} = \frac{\text{tn}}{\text{tn} + \text{fp}} \tag{3}$$

Accuracy is defined as the total percentage of both active and inactive compounds correctly predicted. Sensitivity, also referred to as recall or the true-positive rate, is the percentage of active compounds correctly identified. Specificity, also known as the true-negative rate, is the percentage of inactive compounds correctly predicted. In eqs 1−3, tp is the number of true positives (active compounds that are correctly predicted), fn is the number of false negatives (active compounds that are incorrectly predicted to be inactive), tn is the number of true negatives (inactive compounds that are correctly predicted), and fp is the number of false positives (inactive compounds that are incorrectly predicted to be active). Sensitivity and specificity are good individual measures, with respect to activity and inactivity, of a model's ability to correctly classify compounds from training and testing sets. The geometric mean (*G-mean*) is calculated by combining sensitivity and specificity to obtain a single numerical value; this function provides a simple measure that indicates the extent to which a model is able to correctly predict the classification of both active and inactive compounds, as well as a convenient metric to quickly select optimal models. The G-mean value is defined as

$$\text{G-mean} = \sqrt{\text{sensitivity} \times \text{specificity}} \tag{4}$$

A good P450 prediction (classification) model should minimize the possibility of misclassifying both active and inactive compounds. Therefore, the G-mean value is a good indication of a model's performance, as both the sensitivity and specificity of the model are considered.

Sampling Algorithm. A significant novelty in our study involves the development of a sampling algorithm to better define the training and testing data sets, since randomly splitting a whole data set into a training set and a testing set is not an ideal strategy for building models and evaluating their performance. Theoretically, a sound training data set should be representative of the whole population. Therefore, the purpose for our sampling algorithm is to ensure the compounds with higher structural dissimilarity (compared to other compounds in the whole data set) are included into the training data set, and that each compound in the testing data set must have high structural similarity with at least one compound from the training data set. These criteria will ensure that the compounds chosen for the training and testing data sets will be equally representative of the known chemical diversity, so that the models built using these data sets will be able to generate more accurate prediction of CYP inhibition.

The first step in our algorithm is to calculate the pairwise structural similarities based on PubChem fingerprints using tanimoto coefficients. The arbitrary order initially assigned to these compounds will not change in the following process. Essentially, compounds with tanimoto coefficients higher than a specified threshold, in comparison with the first compound, were selected for the testing data set, whereas the first molecule was included in the training data set. The same process was repeated for the remaining compounds, until the number of testing compounds reaches our expected number. It is important to note that a compound that has been selected for the training data set will not be reassigned to the testing data set. At the end of this selection process, if the number of testing compounds is not sufficient, the program will restart again after the threshold of structural similarity is decreased by 0.1.

The initial value of the threshold was set to 0.99. According to our analysis of the five testing sets, on average 13% of the compounds from our testing sets have Tanimoto values higher than 0.99 with one chemical in the training sets. However, this initial similarity threshold (Tanimoto = 0.99) resulted in an insufficient number of testing compounds, to be used for the construction of CYP prediction models. In order to select sufficient number of testing compounds, the final similarity threshold was adjusted accordingly, and the final threshold used was about 0.92 (Tanimoto = 0.92). This resulted in a sufficient and ideal testing set and training set ratio of 2:8 for our CYP models.

In short, the sampling algorithm was applied to construct well-defined training and testing data sets to improve the performances of CYP classification models. The overall workflow of our rule-based classification model building was illustrated in Figure 1.

■ RESULTS AND DISCUSSION

Optimized Rule-Based Classification Models for CYP Inhibition. For each cytochrome P450 end point, different combinations of the three descriptor sets (PaDEL1&2D, PaDEL3D, and Mold²) were tested in the classification models built by using two machine learning methods (C5.0 algorithm and SVM) with random sampling strategy initially. The
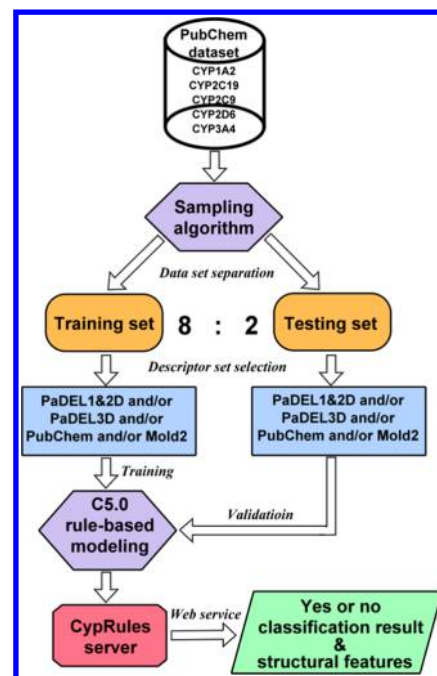

Figure 1. Overall workflow for our rule-based classification model building for prediction of five cytochrome P450 enzymes.

performance of the best models tested for each P450 testing data set, as determined by accuracy, sensitivity, specificity, and G-mean, was summarized in Table 2. The third column shows the descriptor combination used in the best models. The best C5.0 models yielded an average sensitivity and specificity of 74.7%, and 80.7%. The average sensitivity and specificity values for SVM models were 79.4% and 81.7%. The small differences between the average percentages reported for sensitivity and specificity of the two types of models, suggested that the performance of C5.0 rule-based models is comparable with the performance of SVM models. However, despite having the highest accuracy percentage reported for C5.0 models, the sensitivity for CYP2D6 end point was low (below 70%). However, C5.0 algorithm approach generated models that are more transparent and interpretable than the SVM models, the sensitivity percentages reported for some CYP data sets were low, or not of high enough performance. For this reason, sampling algorithm was applied to enhance the performance of the five CYP classification models. It should be noted that only C5.0 rule-based models were considered next; the purpose is to optimize these rulesets generating models, to provide users with a list of CYP inhibition structural characteristics, that are much easier to interpret and more user-friendly compared to SVM models.

After applying the sampling algorithms to generate a training data set that can better represent the population of the whole data set, the performance of the five optimized CYP inhibition classification models using C5.0 algorithm was summarized in Table 3. Because nearly 80% of the descriptors included in PaDEL and Mold² are abstract representation of molecular structures, such as topological indices, we further added PubChem fingerprints in our descriptor pools. If the rules generated by the C5.0 models are based more on molecular structural fragments, those rules will be more useful to the medicinal chemists, since they may be able to directly decipher which structural fragments violate CYP rules. The second column shows the combinatorial descriptor sets used in the

Table 2. Summary of the Best Models Tested for Each P450 Testing Dataset, Utilizing Different Trial Descriptor Sets with Rule-Based C5.0 and SVM Learning Methods

end point	method	descriptor usage	accuracy (%)	sensitivity (%)	specificity (%)	G-mean
CYP1A2	C5.0	PaDEL1&2D + PaDEL3D	79.5	88.9	71.5	79.7
	SVM	Mold^2	79.8	86.8	74.0	80.1
CYP2C19	C5.0	PaDEL1&2D	86.0	84.5	87.2	85.8
	SVM	PaDEL1&2D	84.3	87.5	81.7	84.6
CYP2C9	C5.0	Mold^2 + PaDEL3D	76.8	65.8	82.2	73.5
	SVM	Mold^2 + PaDEL3D	79.8	70.8	84.2	77.2
CYP2D6	C5.0	Mold^2 + PaDEL3D	89.8	58.0	90.4	72.4
	SVM	Mold^2 + PaDEL3D	91.0	71.9	95.6	82.9
CYP3A4	C5.0	Mold^2 + PaDEL3D	73.3	76.3	72.0	74.1
	SVM	Mold^2 + PaDEL3D	75.6	80.0	73.0	76.4

Table 3. Summary of the Optimized Models Tested for Each P450 End Point after Applying the Sampling Algorithm, Utilizing Different Trial Descriptor Sets with Rule-Based C5.0 Methods

end point	descriptors included	data set	accuracy (%)	sensitivity (%)	specificity (%)	G-mean
CYP1A2	PaDEL1&2D	training	96.3	94.4	97.9	96.1
		testing	93.0	92.4	93.6	93.0
CYP2C19	PaDEL1&2D + PaDEL3D + Mold^2 + Pubchem	training	84.8	86.7	83.3	85.0
		testing	84.6	80.2	88.2	84.1
CYP2C9	PaDEL1&2D + PaDEL3D + Mold^2 + Pubchem	training	99.9	99.9	99.9	99.9
		testing	81.4	82.0	81.0	81.5
CYP2D6	Pubchem	training	98.0	99.2	96.7	97.9
		testing	90.6	85.4	91.9	88.6
CYP3A4	PaDEL1&2D + PaDEL3D + Mold^2 + Pubchem	training	92.1	86.0	96.3	91.0
		testing	87.9	90.2	86.3	88.2

corresponding optimized models. For CYP2C19, CYP2C9, and CYP3A4, it appears that the combination of all of the descriptor sets resulted in the best C5.0 models. It is interesting to see that the optimized CYP1A2 and CYP2D6 models featured only PaDEL1&2D descriptor set, and PubChem fingerprints, respectively. The optimized models yielded the average accuracy, sensitivity, specificity, and G-mean percentages of 94.2%, 93.2%, 94.8%, and 94.0% for training data sets, respectively. The corresponding average accuracy, sensitivity, specificity, and G-mean percentages for testing sets were 87.5%, 86.0%, 88.2%, and 87.1%. The sensitivities and specificities of C5.0 models increased by 11% and 8% for the testing data sets, after the sampling algorithm was applied. It is noted that in order to treat the imbalanced data set for the CYP2D6 isomer, we further incorporated the oversampling strategy by amplifying the CYP2D6 inhibitors in the training data set, to balance the ratio of inhibitors to noninhibitors (1:1 ratio). Without sacrificing the specificity, the sensitivity of the optimized C5.0 model for CYP2D6 was enhanced nearly 30% for testing data set. This demonstrated that the oversampling strategy combined with the appropriate selection of representative compounds in the training data sets, produced excellent CYP inhibitory classification models.

Of the five CYP models, 2C9 was the only overfitted model, since accuracy/sensitivity/specificity reported are all close to 99.9% for the training set, yet its prediction performance reduced to 81% for the testing set. However, the 2C9 model is limited by the data. Since our model includes the largest public available screening compounds, the "overfitted" 2C9 model was indeed the best model we could construct given the limitation of the current compounds in the 2C9 data set selected from all the machine learning methods and statistically boosting method we tested (best SVM model's sensitivity on the testing set was

down to 70.8%). When more chemical databases are made available, we will update our data sets and our models accordingly in the future, to again reflect the most currently known chemical space at that time.

We have updated the CypRules web server (http://cyprules. cmdm.tw/) accordingly to reflect our current optimized rule-based classification models. It should be noted that the term "optimized" models used in our study refer to the best performing models we have generated after the sampling algorithm was applied. These models are considered optimized relative to our initial "unoptimized" construction.

To investigate the variation of structural dissimilarities between the training data set and test data sets after applying the sampling algorithm, we used the Tanimoto coefficient to calculate the pairwise structural similarities within the training and testing data sets based on PubChem fingerprints. The resultant average values of structural dissimilarities (1 − similarities) for the training and testing data sets were listed in the second and third rows of Table 4, respectively. All of the differences in dissimilarities between training sets and testing sets (fourth row) were below 0.02. This demonstrated that our selected training compounds were representative compounds that enhanced performance in the training and testing set, as well as preserved structural similarities and characteristics.

Table 4. Average of Pairwise Dissimilarities within the Training and Testing Datasets for the Corresponding Optimized CYP Models after Applying Sampling Algorithm

dissimilarity	1A2	2C19	2C9	2D6	3A4
training set	0.118	0.209	0.208	0.357	0.208
testing set	0.138	0.212	0.211	0.339	0.211
difference	0.020	0.002	0.003	0.018	0.003

Table 5. Summary and Comparison of Prediction Accuracies between the Best Modeling Methodologies Tested for Each P450 End Point, from This Study and from the Published Literature

modeling methodology	accuracy	CYP1A2	CYP2C19	CYP2C9	CYP2D6	CYP3A4
C5.0 models of this study	training	**96.3% (10238)**	**84.8% (10306)**	**99.9% (9903)**	**98.0% (10814)**	**92.1% (10044)**
	testing	**93.0% (2559)**	**84.6% (2577)**	**81.4% (2475)**	**90.6% (2702)**	**87.9% (2511)**
recursive partition[8]	training	89% (306)			90% (498)	
	testing	81% (58)			89% (34)	
Gaussian kernel weighted *k*-NN method[9]	training				87% (865)	83% (1037)
	testing				88% (345)	82% (288)
associative neural networks (ASNN) method[10]	training	83% (3745)				
	testing	68% (3741)				
SVM with VHTS data[14]	training	87.5% (7208)	80.6% (6038)	82.9% (6627)	89.5% (7788)	81.0% (6800)
	testing	93.0% (7128)	89.0% (5923)	89.0% (6530)	85.0% (7761)	87.0% (6738)
combined approach based on back propagation-artificial neural network (BP-ANN)[7]	training	77.2−81.3% (12099)	72.3−78.0% (11885)	73.5−77.3% (12130)	81.7−83.7% (11881)	72.3−76.7% (11536)
	testing	59.7−73.1% (2804)	70.5−81.0% (2691)	75.4−86.7% (2579)	78.5−87.8% (2860)	66.3−76.0% (7025)

Table 6. Model Comparisons between Our Five Optimized Rule-Based Models (Rule) and WhichCyp[13] (SVM)

model	1A2		2C19		2C9		2D6		3A4	
	rule	SVM	**rule**	SVM	**rule**	SVM	**rule**	SVM	**rule**	SVM
accuracy	**93**	87	**84**	84	**81**	86	**90**	84	**88**	84
sensitivity	**92**	88	**80**	83	**81**	84	**85**	75	**90**	84
specificity	**94**	88	**88**	83	**81**	85	**91**	86	**86**	84

Comparison with Other P450 Inhibition Prediction Systems. To further substantiate the overall quality of the optimized P450 classification models built in this study, five published in silico studies for P450 classification employing different methodologies, were compared to the best C5.0 models in this study. All the different P450 models considered for comparison analysis were summarized in Table 5. The first column describes each of the modeling methodologies, and the first row includes the corresponding P450 end points investigated in these studies. The second to thriteenth columns contain the predicted accuracy of the training set with the number of training set compounds in parentheses and the predicted accuracy of the testing set with the number of testing set compounds also given in parentheses for five P450 end points, respectively. The optimal C5.0 models reported within this study were placed in the first row. The five published in silico studies are listed in ascending order based on the number of compounds in each training set. Among five previous studies, relatively high training set accuracy was obtained by using the recursive partition and Gaussian kernel weighted *k*-NN modeling methods, but these models were constructed using smaller training sets. Although the prediction power of the model was greatly affected by the number of the training set molecules, the C5.0 rule-based models in general achieved the best accuracies among all the models for the training set. Table 5 is intended to simply compare the prediction accuracies of the best models published for the five P450 end points, with the best model constructed from this study. We are aware that the prediction accuracies obtained from these literature reports were based on different data sets; however, there is not an appropriate external data set since we included the largest data sets available in our training and testing data sets already, and most of previous models (only SVM and BP-ANN method in Table 5 were using the same data sets) were built from data sets containing significantly smaller number of compounds. Thus, our models were built using more chemical diverse data sets (with significantly larger number of compounds, and

incorporated much more chemical diversity to better represent the currently known chemical spaces), and the prediction accuracies of our models for the five P450 end points have not been compromised. On the contrary, the prediction accuracies of our models are higher than the best published models.

For the testing set, our C5.0 models achieved the highest accuracies for CYP1A2, CYP2D6, and CYP3A4 compared to other studies. It was also shown that VHTS data have similar testing set accuracy using SVM to our studies. SVM is indeed a powerful machine learning technique that can usually construct better models than other methods. However, most publications using SVM did not provide a program or online server for CYP inhibition prediction. Currently, the best freely accessible CYP prediction server to our knowledge is *WhichCyp* which also was built using SVM. Table 6 summarized the comparisons between our five optimized rule-based classification models and WhichCyp. Our rule-based models outperformed in most CYP isoforms compared to WhichCyp. Most importantly, it is difficult to illustrate how to predict a compound as an inhibitor or noninhibitor during the processing of SVM models. WhichCYP provides yes/no classification results only, while our rule-based models can further provide detailed structural information that can be used as guidelines to refine drug candidates.

Descriptor Terms Used in Our Models. In Table 7, the numbers of descriptor terms selected in the five optimized classification models were summarized. The second and fourth columns listed the total number of active (inhibitory) and inactive (noninhibitory) rules for each model, and the corresponding number of descriptor terms selected by C5.0 optimization processes in each model were listed in third and fifth columns. In the CYP2D6 model, the ratio of active descriptors to inactive descriptors is highly imbalanced (1:18) since the difference between the numbers of inhibitors and noninhibitor compounds in the CYP2D6 data set is highly skewed. Overall, nearly 50% of descriptors were selected in our models. It shows that the predicted capacity of the models

Table 7. Numbers of Active (Inhibitory) and Inactive (Noninhibitory) Rules and Descriptors Used in Each Model

models	active rulesets	active descriptors	inactive rulesets	inactive descriptors
CYP1A2	157	347	591	646
CYP2C19	12	59	142	214
CYP2C9	291	629	401	738
CYP2D6	24	27	1276	472
CYP3A4	181	414	276	447

would be applicable, since the model was not restricted on a small scope of descriptors. Although a portion of the used descriptors in our models are not easy to understand, for each model, we have selected and introduced five significant rules that can be used as guidelines in the correct classification of each CYP inhibitions in the following sections. The five rules having highest frequencies used to classify a compound as a CYP inhibitor both in training and testing data set were identified as our five significant rules. Moreover, we selected one well-known CYP inhibitor in each model for interpretations of corresponding significant rules. Since most compounds are expected to be noninhibitors, the noninhibitory rules were not specifically shown and discussed in this study. The interpretations of the CYP1A2, CYP2C19, CYP2C9, and CYP3A4 models were included in the supplementary file.

Interpretations of CYP2D6 Model. For CYP2D6, the best selected descriptor set in the optimized C5.0 model contains only PubChem fingerprints. Five significant rules that assist the CYP2D6 model to classify a compound as a CYP2D6 inhibitor were shown in Figure 2A. We take Berberine (DrugBank ID: DB04115), a well-known CYP2D6 inhibitor, as an example to explain these rules. Berberine was correctly classified as an inhibitor by the CYP2D6 C5.0 prediction model and the five identified significant rules are the dominant rules for Berberine as well. Since these PubChem fingerprints can be easily understood by their original descriptor names, the definitions of these functional groups were ignored. For the rules listed in Figure 2A, "=1" means the presence of the rule in a compound whereas "=0" means the absence of the rule. We first focused on the two presences of the rules. For the second descriptor (propylene) listed in Figure 2A, if a compound contains aromatic rings, the compound must contain the propylene fragment. Berberine includes three aromatic rings which increases the probability that the compound will be a CYP2D6 inhibitor and the features are colored in red in Figure 2A. According to the fourth rule, the substructure, 1,3-dioxolane from Berberine contains one "C(~H)(~O)(~O)" group colored blue (Figure 2A) which also enhances the CYP2D6 inhibitory propensity. Regarding the absence of rules, Berberine contains no rings of size 3, no "O(~H)(~S)" fragments, and only one unsaturated nonaromatic heteroatom-containing ring size of six included in Berberine. Overall, Berberine was predicted as an inhibitor because of its aromatic rings, 1,3-dioxolane, low occurrences of unsaturated non-aromatic heteroatom-containing rings of size 6, and absence of heteroatom-containing ring of size 3, or "O(~H)(~S)".

Next, we demonstrated that how to use theses rulesets as a guide for medicinal chemists to design or modify an inhibitor without potentially inhibiting a specific CYP enzyme by avoiding those structural alerts (Figure 2A) altogether. Changing one of the rules described in Figure 2A by structural modification of Berberine can make it a noninhibitor. A
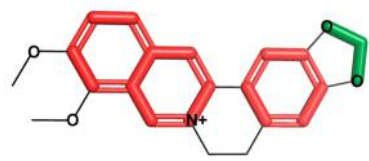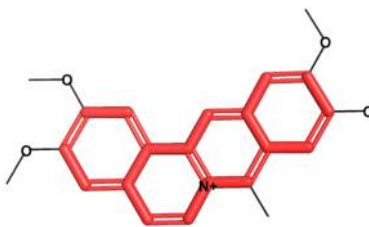

Figure 2. (A) Berberine: an example of the specific rulesets and calculated values provided by the optimized CYP2D6 rule-based C5.0 classification model, which correctly predicted Berberine as a CYP2D6 inhibitor. (B) Coralyne Sulfoacetate: an example of CYP2D6 noninhibitor which is similar to the structure of Berberine. (The rules listed for Coralyne Sulfoacetate is same as for Berberine.)

minimal change that we can apply without altering the main core structure of Berberine is to change the alert of 1,3-dioxolane. After deleting the carbon atom between the two oxygens on the 1,3-dioxolane of Berberine, the modified Berberine then mismatched the CYP2D6 inhibitory ruleset. A very similar structure to Berberine, Coralyne Sulfoacetate (CID: 6419900), was observed and shown in Figure 2B. The structure of Coralyne Sulfoacetate not only satisfied the condition we discussed above without the structure alert of 1,3-dioxolane, but also was actually a CYP2D6 noninhibitor according to the NCGC Cytochrome panel assay (AID: 1851). Our best CYP2D6 rule-based model also successfully predicted Coralyne Sulfoacetate as a CYP2D6 noninhibitor. As a result, we can infer that the modified Berberine which violate the alert of 1,3-dioxolane could become a CYP2D6 noninhibitor.

Limitations of our Models. Our final C5.0 rule-based CYP prediction models would be more appropriately described as a "multiple rules, 1 compound" type of model. The medicinal chemists can inspect the structural or property alerts contributing to the CYP inhibition for a query compound predicted by the CypRules. However, a portion of our used descriptors which belong to PaDEL or Mold[2] descriptor pools are the abstract representation of compounds, such as molecular topological indices. Some of these descriptors are

not easy to understand and could limit applicability of the system. But, one should keep in mind, although the molecular structural fragments such as PubChem Fingerprints can facilitate directly modification on the molecular structure to improve the CYP inhibition profile, the traditional structural fragments cannot moderately fit on the CYP inhibition data whereas our reported CYP1A2 C5.0 models based on the combination of PaDEL1D and PaDEL2D descriptors indeed resulted in a better performance. Actually, most of the Mold[2] and PaDEL descriptors in optimal rulesets of models selected by our C5.0 system can be interpreted and correlated with physical meanings for structural modification of CYP profile.

CONCLUSIONS

The major findings and conclusions drawn from this study are summarized here: (1) Rule-based C5.0 models have explanatory ability for each P450 end point compared to the SVM models. (2) It was shown that by using a sampling method with the rule-based C5.0 algorithm, we could improve each model's explanatory ability, both on the training set as well as on the testing set. (3) A CYP inhibition prediction model was built, featuring the advantages of the C5.0 algorithm that provides chemical information and rulesets for further inspection. (4) The updated CypRules web server not only predicted the inhibition of P450 end points but also provided structural rulesets that contribute to inhibition. (5) Five significant CYP inhibition rules for each model were suggested and can give users some insight on how chemical structures are related to P450 inhibition. We have updated the CypRules website built by our new optimized models as version 2. The user interface on website of our CypRules version 2 is same as version 1. The main difference between the two versions is that the prediction power of version 2 has improved considerably after the sampling algorithm, described in this study, was applied. The average accuracy, sensitivity, and specificity increased 6%, 6%, and 8%, respectively, for the five CYP data sets.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Interpretations of CYP1A2, CYPC19, CYP2C9, and CYP3A4 models, Figures S1–S4, Tables S1–S4, and additional references. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00130.

AUTHOR INFORMATION

Corresponding Author
*Voice: +886.2.3366.4888 #529. Fax: +886.2.23628167. E-mail: yjtseng@csie.ntu.edu.tw.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

REFERENCES

(1) Lynch, T.; Price, A. The Effect of Cytochrome P450 Metabolism on Drug Response, Interactions, and Adverse Effects. *Am. Fam. Physician* **2007**, *76* (3), 391−396.
(2) Wilkinson, G. R. Drug therapy - Drug metabolism and variability among patients in drug response. *N. Engl. J. Med.* **2005**, *352* (21), 2211−2221.
(3) Smith, H. S. Opioid metabolism. *Mayo Clin. Proc.* **2009**, *84* (7), 613−624.
(4) Friedman, E. S.; Grendell, H. J.; McQuaid, R. K. *Current diagnosis & treatment in gastroenterology*; Lang Medical Books/McGraw-Hill: New York, 2003.
(5) Pandit, A.; Sachdeva, T.; Bafna, P. Drug-Induced Hepatotoxicity: A Review. *J. Appl. Pharm. Sci.* **2012**, *2* (5), 233−243.
(6) Hammann, F.; Gutmann, H.; Baumann, U.; Helma, C.; Drewe, J. Classification of cytochrome P450 activities using machine learning methods. *Mol. Pharmaceutics* **2009**, *6* (6), 1920−1926.
(7) Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Classification of Cytochrome P450 Inhibitors and Non-inhibitors Using Combined Classifiers. *J. Chem. Inf. Model.* **2011**, *51* (5), 996−1011.
(8) Burton, J.; Ijjaali, I.; Barberan, O.; Petitet, F.; Daniel, P.; Vercauteren, A. Recursive Partitioning for the Prediction of Cytochromes P450 2D6 and 1A2 Inhibition: Importance of the Quality of the Dataset. *J. Med. Chem.* **2006**, *49*, 6231−6240.
(9) Jensen, B. F.; Vind, C.; Padkjær, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted k-Nearest Neighbor and Extended Connectivity Fingerprints, Including Structural Fragment Analysis of Inhibitors versus Noninhibitors. *J. Med. Chem.* **2007**, *50*, 501−511.
(10) Novotarskyi, S.; Sushko, Iurii; Koerner, R.; Pandey, Anil Kumar; Tetko, I. V. A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. *J. Chem. Inf. Model.* **2011**, *51*, 1271−1280.
(11) Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro, S. Comparison of Multilabel and Single-Label Classification Applied to the Prediction of the Isoform Specificity of Cytochrome P450 Substrates. *J. Chem. Inf. Model.* **2009**, *49* (11), 2588−605.
(12) Mishra, N. K.; Agarwal, S.; Raghava, G. P. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacol.* **2010**, *10*, 8.
(13) Rostkowski, M.; Spjuth, O.; Rydberg, P. WhichCyp: prediction of cytochromes P450 inhibition. *Bioinformatics* **2013**, *29* (16), 2051−2052.
(14) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Huang, R. Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data. *J. Chem. Inf. Model.* **2011**, *51* (10), 2474−2481.
(15) Chang, C. Y.; Hsu, M. T.; Esposito, E. X.; Tseng, Y. J. Oversampling to Overcome Overfitting: Exploring the Relationship between Data Set Composition, Molecular Descriptors, and Predictive Modeling Methods. *J. Chem. Inf. Model.* **2013**, *53* (4), 958−971.
(16) Sridhar, J.; Liu, J. W.; Foroozesh, M.; Stevens, C. L. K. Insights on Cytochrome P450 Enzymes and Inhibitors Obtained Through QSAR Studies. *Molecules* **2012**, *17* (8), 9283−9305.
(17) Roy, K.; Roy, P. P. QSAR of cytochrome inhibitors. *Expert Opin. Drug Metab. Toxicol.* **2009**, *5* (10), 1245−1266.
(18) Lewis, D. F. V.; Lake, B. G.; Dickins, M. Quantitative structure-activity relationships (QSARs) in inhibitors of various cytochromes P450: The importance of compound lipophilicity. *Journal of Enzyme Inhibition and Medicinal Chemistry* **2007**, *22* (1), 1−6.
(19) Lewis, D. F. V.; Modi, S.; Dickins, M. Structure-activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metab. Rev.* **2002**, *34* (1−2), 69−82.
(20) Ekins, S.; De Groot, M. J.; Jones, J. P. Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome P450 active sites. *Drug Metab. Dispos.* **2001**, *29* (7), 936−944.

(21) Li, H. Y.; Sun, J.; Fan, X. W.; Sui, X. F.; Zhang, L.; Wang, Y. J.; He, Z. G. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J. Comput.-Aided Mol. Des.* **2008**, *22* (11), 843−855.

(22) Gleeson, M. P.; Davis, A. M.; Chohan, K. K.; Paine, S. W.; Boyer, S.; Gavaghan, C. L.; Arnby, C. H.; Kankkonen, C.; Albertson, N. Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models. *J. Comput.-Aided Mol. Des.* **2007**, *21* (10−11), 559−73.

(23) Miller, G. P. Advances in the interpretation and prediction of CYP2E1 metabolism from a biochemical perspective. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4* (8), 1053−64.

(24) Shao, C.-Y.; Su, B.-H.; Tu, Y.-S.; Lin, C.; Lin, O. A.; Tseng, Y. J. CypRules: A rule-based P450 inhibition prediction server. *Bioinformatics* **2015**, *31*, 1869.

(25) Veith, H.; Southall, N.; Huang, R.; James, T.; Fayne, D.; Artemenko, N.; Shen, M.; Inglese, J.; Austin, C. P.; Lloyd, D. G.; Auld, D. S. Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat. Biotechnol.* **2009**, *27* (11), 1050−5.

(26) NIH Chemical Genomics Center PubChem Assay ID 1851: Cytochrome panel assay with activity outcomes. https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1851 (accessed May 10, 2014).

(27) YAP, C. W. Software News and Update PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466.

(28) Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Model.* **1995**, *35*, 1039−1045.

(29) Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Model.* **1999**, *39* (5), 868−873.

(30) Roy, K.; Ghosh, G. QSTR with Extended Topochemical Atom Indices. 2. Fish Toxicity of Substituted Benzenes. *J. Chem. Inf. Model.* **2004**, *44* (2), 559−567.

(31) Abraham, M. H.; McGowan, J. C. The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. *Chromatographia* **1987**, *23*, 243−6.

(32) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Model.* **1999**, *39* (5), 835−845.

(33) Laggner, C. SMARTS Patterns for Functional Group Classification. http://code.google.com/p/semanticchemistry/source/browse/wiki/ (2013).

(34) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48* (7), 1337−1344.

(35) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178−180.

(36) Talete Dragon. http://www.talete.mi.it/products/dragon_description.htm (accessed July 15, 2015).

(37) Haney, D. N.; Hall, L. H. Molconn-Z. http://www.edusoft-lc.com/molconn/mconpubs.html (accessed July 15, 2015).

(38) Shannon, C. E. A Mathematical Theory of Communication. *Mobile Computing and Communications Re* **2001**, *5*, 3−55.

(39) Modi, S.; Li, J.; Malcomber, S.; Moore, C.; Scott, A.; White, A.; Carmichael, P. Integrated in silico approaches for the prediction of Ames test mutagenicity. *J. Comput.-Aided Mol. Des.* **2012**, *26* (9), 1017−1033.

(40) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers, 1993.

(41) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: Berlin, 1995.

(42) Vapnik, V. N. *Statistical Learning Theory*; Wiley-Interscience: New York, 1998.