JOURNAL OF CHEMICAL INFORMATION AND MODELING

Machine Learning and Deep Learning

ADMET Evaluation in Drug Discovery. 19. Reliable Prediction of Human Cytochrome P450 Inhibition Using Artificial Intelligence Approaches

zhenxing Wu, Tailong Lei, Chao Shen, Zhe Wang, Dongsheng Cao, and Tingjun Hou

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.9b00801 • Publication Date (Web): 23 Oct 2019

Downloaded from pubs.acs.org on October 24, 2019

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

ADMET Evaluation in Drug Discovery. 19. Reliable Prediction of Human Cytochrome P450 Inhibition Using Artificial Intelligence Approaches

Zhenxing Wu[a], Tailong Lei[a], Chao Shen[a], Zhe Wang[a], Dongsheng Cao[c], Tingjun Hou[a,b,*]

[a]Hangzhou Institute of Innovative Medicine, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, P. R. China

[b]State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang, P. R. China

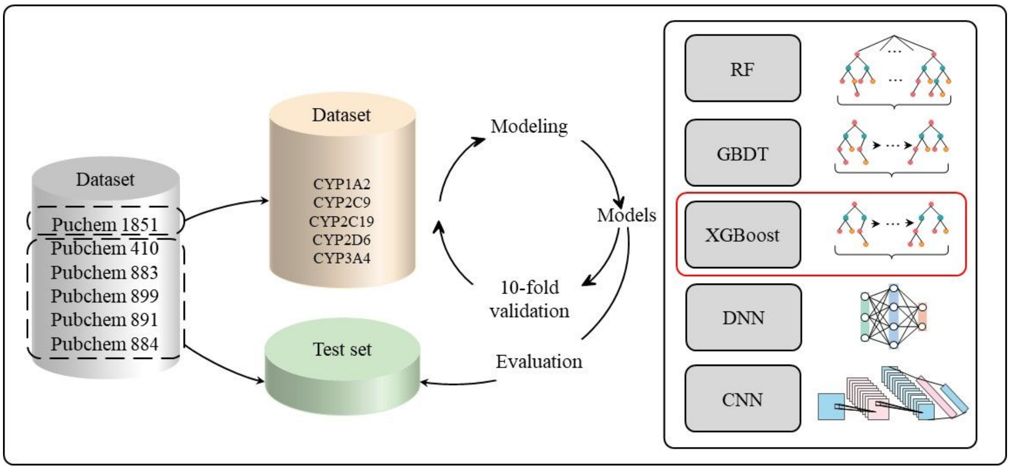
[c]Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410004, Hunan, P. R. China

Corresponding author:

Tingjun Hou

***E-mail**: tingjunhou@zju.edu.cn

Table of Contents Graphic



Abstract

Adverse effects induced by drug-drug interactions may result in early termination of drug development or even withdrawal of drugs from the market, and many drug-drug interactions are caused by inhibition of cytochrome P450 (CYP450) enzymes. Therefore, accurate prediction of the inhibition capability of a given compound against a specific CYP450 isoform is highly desirable. In this study, three ensemble learning methods, including random forest (RF), gradient boosting decision tree (GBDT) and eXtreme gradient boosting (XGBoost), and two deep learning methods, including deep neural network (DNN) and convolutional neural network (CNN), were used to develop classification models to discriminate inhibitors and non-inhibitors for five major CYP450 isoforms (1A2, 2C9, 2C19, 2D6, and 3A4). The results demonstrate that the ensemble learning models generally give better predictions than the deep learning models for the external test sets. Among all the models, the XGBoost models achieve the best classification capability (average prediction accuracy of 90.4%) for the test sets, which even outperform the previously reported model developed by multitask deep autoencoder neural network (88.5%). The Shapley Additive exPlanation (SHAP) method was then used to interpret the models and analyze the misclassified molecules. The important molecular descriptors given by our models are consistent with the structural preferences for inhibitors of different CYP450 isoforms, which may provide valuable clues to detect potential drug-drug interactions in early stage of drug discovery.

Keywords: Cytochrome P450, Ensemble learning, Deep learning, Drug-drug interaction, Shapley additive explanations

1. Introduction

More than 30 cytochrome P450 (CYP450) isoforms have been identified in human to date, and six of them (CYP1A2, 2C9, 2C19, 2D6, 3A4, and 3A5) metabolize more than 90% of drugs. Inhibition of CYP450 enzymes may lead to decreased elimination and/or changed metabolic pathways of their substrates, which is the major cause of adverse drug-drug interactions.[1, 2] In recent years, several drugs, such as mibefradil and cerivastatin, were withdrawn from the market because of adverse drug-drug interactions caused by their inhibition to CYP450.[3,4] Therefore, it is quite essential to identify undesirable CYP450 inhibition in early stage of drug discovery. The inhibition effect of a drug on different CYP450 isoforms can be detected by different experimental techniques, such as traditional single substrate assays, fluorescent probe assays with recombinant human CYP450s, and n-in-one assays.[5] Traditional *in vitro* CYP450 inhibition assays typically evaluate the inhibition of a drug on one P450 isoform at a time, and n-in-one assays, also known as CYP450 cocktail inhibition assays, can evaluate the inhibition effects of drugs on 5~8 CYP450 isoforms simultaneously by high performance liquid chromatography-tandem mass spectrometry (LC-MS/MS) or other detection techniques based on multiple probe substrates.[5] Obviously, cocktail approaches are much more efficient than traditional single probe substrate approaches, but they still have some disadvantages, such as potential interaction between probe substrates and complicated detection of probe substrates. Moreover, the experimental assays are expensive and time-consuming. Therefore, development of accurate theoretical models to predict CYP450 inhibition is highly desirable.

In the past several years, a variety of machine learning approaches have been used to develop theoretical models for *in silico* prediction of CYP450 inhibition[6-8] based on the large-scale datasets generated by high throughput *in vitro* screening of CYP450 inhibition.[9] For example, Cheng *et al.* developed single-target classification models for five CYP450 isoforms by fusing multiple machine learning classifiers through a back-propagation artificial neural network (BP-ANN), including support vector machine (SVM), C4.5 decision tree (DT), k-nearest neighbor (k-NN), and naïve Bayes, and the predictive accuracies for the test sets are 73.1%, 86.7%, 81.0%, 87.8%, and 76.0% for 1A2, 2C9, 2C19, 2D6, and 3A4, respectively.[6] After that, Sun and co-workers utilized SVM to develop classifiers for 1A2, 2C9, 2C19, 2D6, and 3A4 based on a set of customized generic atom types,[8] and the corresponding areas under the receiver operating characteristic curve (AUC) for the test sets are 0.93, 0.89, 0.89, 0.85, and 0.87, respectively. In 2015, Su *et al.* developed rule-based inhibition prediction models for 1A2, 2C9, 2C19, 2D6, and 3A4 by using a rule-based C5.0 algorithm with different descriptors, and their predictive accuracies are 79.5%, 76.8%, 86.0%, 89.8%, and 73.3%, respectively.[7] In recent years, deep learning algorithms have gained great success in drug design and ADMET predictions.[10-19] For example, in 2018, Li *et al.* developed classification models by training a multitask autoencoder deep neural network (MAE-DNN) based on a large dataset extracted from the PubChem BioAssay Database, and the predictive accuracies for the test sets are 96.8%, 86.0%, 80.9%, 89.3%, and 89.6% for 1A2, 2C9, 2C19, 2D6, and 3A4, respectively.[12] The calculation results also illustrated that the multitask model gave better predictions than the models developed by single-task autoencoder-based deep neural network (AE-based DNN) and traditional machine learning algorithms (logistic regression, SVM, C4.5 DT, and k-NN). Recently, ensemble learning approaches, especially gradient boosting decision tree (GBDT), have attracted increasing attention. GBDT builds the model by combining weak base learners into a single strong learner in an iterative way, and then generalizes them by allowing an optimization of an arbitrary differentiable loss function. Extreme gradient boosting (XGBoost), proposed by Chen and Guestrin,[20] is an efficient and scalable implementation of the gradient boosting framework, and it has been regarded as a new generation of ensemble learning algorithm. XGBoost has become the winners for several machine learning competitions in recent years.[21-23] For example, among the 29 winning solutions published at the Kaggle's blog in 2015, 17 solutions used XGBoost while only 11 used DNNs. Besides, GBDT and XGBoost have been applied in the field of drug design, such as scoring function development, ADMET prediction, etc.[24-29] For example, in 2017, Lei *et al.* developed a number of quantitative and qualitative predictions models for chemical-induced urinary tract toxicity by machine learning approaches, and the XGBoost classification model achieves the best qualitative predictions for the test set (global accuracy of 82.62%).[29] Since many chemicals are inhibitors for different CYP450 enzymes, multitask classification models have been developed to distinguish CYP450 inhibitors from non-inhibitors. For example, in 2018, Li et al. developed classification models for CYP450 inhibition by training a multitask autoencoder deep neural network (MAE-DNN). Recently, Shan et al. developed 7 different multi-label models to predict the CYP450 substrate selectivity and the NLSD-XGB models achieve the best performance with the average prediction success of 91.1%.[30] A multi-task model can predict inhibitors for multiple CYP450 enzymes simultaneously, but it should have higher complexity than single-task models. It is more suitable to construct the multi-task model for multiple relevant data sets with a small amount of data. However, the data sets for different CYP450 enzymes are relatively large. Therefore, single-task classification models may have better performance than multi-task models.

The main aim of this study was to evaluate whether ensemble learning can achieve better performance than deep learning for relatively large datasets with CYP450 inhibition data. Therefore, we developed the classification models to discriminate inhibitors and non-inhibitors for five major CYP450 isoforms (1A2, 2C9, 2C19, 2D6, and 3A4) by three ensemble learning methods, including RF, GBDT and XGBoost, and two deep learning methods, including DNN and convolutional neural network (CNN). The statistical significance of the developed models was assessed by the 10-fold cross-validation for the training sets and their actual prediction capability was validated by the external test sets.

2. Materials and Methods

2.1. Datasets

In order to compare the prediction models developed by ensemble learning and deep learning methods with that developed by multitask deep autoencoder neural network reported by Li *et al.*, the same datasets offered by Li *et al.*[12] were also used in our study.

The original datasets were extracted from the Pubchem BioAssay Database determined by six assays. The first assay (AID:1851) used an *in vitro* bioluminescent approach to determine the potency values of 17143 compounds against five CYP450 isoforms, which are included in the training set. The other five assays determined the potency values of different compounds against each CYP450 isoform, which were used as the test sets (AID: 410 for 1A2, AID: 883 for 2C9, AID: 899 for 2C19, AID: 891 for 2D6, and AID: 884 for 3A4). The redundant compounds in the training and test sets were excluded from the test sets. The original datasets were processed by Li *et al.* with the KNIME software and standardized by the ChemAxon Standardizer.[31] The compounds were categorized into two classes based on the PubChem activity scores and concentration-response curves: inhibitors (score≥40， curve class = -1.1, -1.2, -2.1) and noninhibitors (score=0, curve class = 4). The compounds that could not be identified by these two criteria were removed. More details about the datasets are shown in **Table 1**. In Li's study, 1253 PaDEL-1D&2D descriptors and 688 PubChem fingerprints calculated by the PaDEL-Descriptor software[32] were used in model training and testing. In order to make a direct comparison between the models developed in this study and those reported by Li *et al.*,[12] the datasets and descriptors offered by Li *et al.* were directly used in this study. The compound diversity of the datasets calculated by the multidimensional scaling (MDS) is shown in **Figure S1**.

2.2. Calculations of different molecular descriptors

Besides the PaDEL-1D&2D descriptors and PubChem fingerprints used by Li *et al.*, other several sets of molecular descriptors were calculated to explore the impact of molecular descriptors on the performance of the XGBoost prediction models, including the MOE descriptors, and two types of fingerprints (GraphOnly fingerprints and KlekotaRoth fingerprints). The SMILES representations of the compounds in the datasets were converted into 3-D structures in the Molecular Operating Environment (MOE) molecular simulation package.[33] Then the molecules were optimized by the Energy Minimize module with the MMFF94X force field and the MOE descriptors were generated.[33] The GraphOnly and KlekotaRoth fingerprints of each molecule were calculated by using the PaDEL-Descriptor software and the Morgan fingerprints were calculated by using the RDKit package in python.[34] The descriptors that have all zero values or zero variance were removed, and the vacancy values were filled with the mean of the corresponding descriptor. Then, the values of each descriptor were normalized to the range between 0 and 1 by subtracting the minimum value of the descriptor and dividing by the range. The detailed information of different sets of molecular descriptors is summarized in **Table 2.**

2.3. Models Developed by Machine Learning Approaches

In recent years, a more data-hungry machine learning algorithm, deep learning, has gained great success in the applications of drug design and discovery.[10, 12, 35-38] In 2018, Li *et al.* developed a MAE-DNN model to discriminate inhibitors and non-inhibitors for five major CPY450 isoforms, and it achieved better performance than the models developed by several traditional machine learning methods and single-task autoencoder DNN.[12] Unfortunately, ensemble learning methods, which have been regarded as the state-of-the-art solutions for many machine learning challenges,[29, 39-41] have not been used in Li's study. Therefore, three representative ensemble learning approaches, including RF, GBDT and XGBoost, were employed to develop classification models for CYP450 inhibition. The optimal parameters of the ensemble learning models were determined by using the grid searching method, which is implemented by the Scikit-learn package (version 0.19.1) in Python (version 3.6.6), and the 3-fold cross-validation was used for model optimization.

Moreover, two representative deep learning methods, including DNN and CNN, were used to train the models. The optimal parameters for the deep learning models were determined by the Tree Parzen Estimator (TPE) algorithm, which is implemented by the hyperopt package (version 0.1.1) in python. For building the CNN models, the feature data set was converted into an array of 48×48, and the insufficient data was supplemented by 0. The overall workflow to build the classification model of CYP450 inhibition based on different methods is illustrated in **Figure 1**.

Ensemble learning. An ensemble learning algorithm builds a set of base learners based on the training set and performs predictions by voting for classification or by averaging for regression based on the predictions made by individual learners.[42-45] According to different individual learner generation methods, ensemble learning methods can be roughly divided into two categories: Bagging and Boosting. Bagging algorithm generates multiple sample sets by using the bootstrap sampling method and then trains individual learners based on different sample sets. Boosting algorithm is an iterative technique that adjusts the weight of an observation predicted by the last learner. If an observation is classified incorrectly, it tries to increase the weight of this observation, and then create a stronger learner from a number of weaker learners. In our study, one bagging algorithm (RF) and two boosting algorithms (GBDT and XGBoost) were used to build the classification models.[46-48] The main hyperparameters of these three models were optimized by using the grid search method in Scikit-learn. The important hyperparameters of the XGBoost models are shown in **Table 3**, and those of the RF and GBDT models and more detailed information of the hyperparameter optimization can be found in **Tables S1 and S2** in Supporting Materials.

Random Forest (RF). RF is a bagging extension variant based on decision trees and improves the generalization of the final integration by introducing random features selection in the training process of decision trees. RF is one of the most popular algorithms in QSAR modeling.[46, 49]

Gradient Boosting Decision Tree (GBDT). GBDT, also known as MART (Multiple Additive Regression Tree), is an iterative decision tree algorithm.[45] Unlike traditional boosting methods, each calculation of GBDT is to reduce the residuals of the previously constructed tree learners rather than focus on reweighting misclassified samples. To minimize the residuals, GBDT builds a decision tree learner along the direction of the gradients of the residuals. GBDT yields the prediction results by accumulating all the trees, and the accumulation cannot be done by classification. Therefore, unlike RF, the GBDT trees are all CART regression trees instead of classification trees and these trees can only be generated serially.

eXtreme Gradient Boosting (XGBoost). XGboost is a GBDT-based machine learning method proposed in 2014.[47] Due to its extraordinary predictive power, this algorithm has been used by many winning teams for machine learning competitions. Compared with traditional GBDT, XGBoost expands the loss function into the second-order Taylor's expansion and provides more regularization options, including L1 and L2 regularizations. The penalization on the leaf nodes and column subsampling are used to balance the decline of the loss function and the model complexity in order to prevent over-fitting and reduce computational cost. In addition, XGBoost implements split finding algorithms and provides a column block structure for parallelization which enables quicker model exploration. The overall workflow of the XGBoost modeling based on different sets of descriptors is shown **in Figure 2**.

Deep Neural Network (DNN). In recent years, deep learning has won numerous contests in pattern recognition and has been also widely used in drug design.[10, 36, 37, 50] DNN is one of the typical deep learning methods which includes input layers, hidden layers and output layers. Drop out regularization, batch normalization and early stopping method were used to avoid overfitting of the DNN modeling. The number of hidden layers (1, 2, 3, 4, 5), the number of nodes of all hidden layers (64, 128, 256, 512), dropout regularization (0.25, 0.5, 0.75) and optimizers (Resprop, Adam, SGD, Adadelta) of the DNN models were optimized. The hyperparameters were optimized by using DNN (Keras with Tensorflow backend) and the TPE method from hyperopt.

Convolutional Neural Network (CNN). CNN is another typical deep learning method that includes convolutional (CONV) layers, pooling layers and fully connected (FC) layers,[51] in which the CONV layers are used to extract features from the input, while the pooling layers subsample or downsample feature maps typically with average, sum, or max pooling. Fully connected layers, however, connect every neuron in one layer to every neuron in another layer just like in a traditional ANN. The following parameters were optimized prior to the final model training: number of CONV layers (2, 3, 4), kernel size of filters ((3, 3), (5, 5)), number of filters of all CONV layers (16, 32, 64, 128), number of FC layers (1, 2, 3), number of nodes of all FC layers (64, 128, 256, 512), dropout regularization(0.25, 0.5, 0.75) and optimizers (Resprop, Adam, SGD, Adadelta). The hyperparameters were optimized by using the same strategy employed by DNN.

2.4. Evaluation of the CYP450 Models

Each model was evaluated by the following parameters: accuracy (ACC), sensitivity (SE), specificity (SP), Matthews correlation coefficient (MCC), and the area under the receiver operating characteristic curve (AUC). The more detailed definitions are as follows:

$$SE = \frac{TP}{TP + FN} \tag{1}$$

$$\mathrm{SP} = \frac{TN}{TN + FP} \tag{2}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$MCC = \frac{TP\ \times TN - FN\ \times FP}{\sqrt{(TP + FN)(TP + FP)(FN + TN)(TN + FP)}} \tag{4}$$

where TP is the number of correctly classified inhibitors, TN is the number of correctly classified noninhibitors, FP is the number of misclassified inhibitors, and FN is the number of misclassified noninhibitors. ACC represents the overall prediction accuracy of the model. MCC is a commonly used evaluation index in the binary classification model and is generally considered to be a relatively balanced index, even when the data distribution is imbalanced. AUC is the area under the ROC curve and represents the probability that a randomly chosen inhibitor is ranked higher than a randomly chosen noninhibitor.

2.5. Model Interpretation

In our study, interpretation of the models is as crucial as the prediction accuracy because it can not only provide insight into how our models may be improved but also help us to extract information that has great impact on CYP450 inhibition. However, both of ensemble learning and deep learning methods are not easy to be understood and some methods were proposed to interpret the complicated models.[52-55] Here, the Shapley Additive exPlanation (SHAP) method proposed by Lundberg *et al.* was used to interpret our models.[56, 57]

The SHAP method, falling into the class of additive feature attribution methods, has an explanation model $g(z')$ approximating to the output of the original model that is a linear function of binary variables:

$$f(x) = g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z'_i \qquad (5)$$

where $f(x)$ is the origin model, $g(z')$ is the explanation model, $z' \in \{0,1\}^M$, M is the number of molecular descriptors, φ_i is the attribution value of the i^{th} molecular descriptor named the SHAP value, and z'_i represents a molecular descriptor being observed ($z'_i = 1$) or not ($z'_i = 0$).

The following functions were used to calculate the SHAP value (φ_i):

$$f_x(S) = f\left(h_x(z^{'})\right) = E[f(x)|x_s] \qquad (6)$$

$$\varphi_i = \sum_{S \subseteq N \backslash \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(\mathrm{S} \cup \{i\}) - f_x(S)] \qquad (7)$$

where S is the set of non-zero indexes in z', $h_x(z')$ is a mapping that maps between a binary pattern of missing features represented by z' and the original function input space to evaluate the effect missing features (by setting $z' = 1$ or $z' = 0$), $E[f(x)|x_s]$ is the expected value of the function conditioned on a subset S of the input molecular descriptors, and N is the set of all molecular descriptors.

As shown above, the SHAP value (φ_i) can quantify the contribution of each molecular descriptor to a molecule being predicted as a CYP450 inhibitor, and $g(z')$ is the sum of the molecular descriptor attributions (φ_i) and approximates to the output of the original model ($f(x)$). Therefore, the SHAP method can help us understand the model more intuitively and it was implemented by SHAP (version 0.28.5) in python.

3. Results and Discussion

3.1. Comparison of CYP450 inhibition classification models based on different methods.

First, the binary classification models for CYP1A2, 2C9, 2C19, 2D6, and 3A4 were developed by using three representative ensemble learning methods (RF, GBDT, and XGBoost) and two representative deep learning methods (DNN and CNN) based on the PaDEL-1D&2D descriptors and PubChem fingerprints. The generalization ability of each classifier was assessed by the 10-fold cross-validation, and the actual prediction power of each classifier was validated by the predictions on the external test set. The cross-validation accuracies of different models on the five training sets are shown in **Figure 3** and the prediction accuracies of different models on the five test sets are shown in **Figure 4**. More detailed information of the performance of different models on the training and test sets are summarized in **Table S3** in Supporting Materials. Moreover, the prediction accuracies of the autoencoder-based single-task DNN and MAE-DNN models reported by Li *et al*. are also shown in **Figures 3 and 4**.[12]

According to the 10-fold cross-validation results for the training sets, the XGBoost models achieve the best performance for the 1A2 (ACC=0.905), 2C19 (ACC=0.850) and 3A4 (ACC=0.860) isoforms, and the DNN models achieve the best performance for the 2C9 (ACC=0.890) and 2D6 (ACC=0.930) isoforms. However, according to the predictions for the test sets (**Figure 4**), the XGBoost models achieve the best predictions for the 1A2, 2C9, 2C19 and 2D6 isoforms, and the corresponding accuracies are 97.4%, 90.2%, 82.3% and 92.8% respectively. The MAE-DNN model achieves the best predictions for the 3A4 isoform, but its accuracy (89.6%) is only slightly higher than that of the XGBoost model (89.4%). The predictive powers of the GBDT and RF models on the test sets, on average, are only slightly worse than those of the XGBoost models, but obviously higher than those of the other models. Compared with the ensemble learning models, except that the prediction accuracy of the MAE-DNN model is slightly higher than that of the XGBoost model for the 3A4 isoform, the performances of the deep learning models are obviously worse than those of the ensemble learning models, suggesting that ensemble learning methods may be a better choice for the development of the classification models for CYP450 inhibition than deep learning methods. Based on the average ACC values and other evaluation parameters to the test sets, the predictive powers of all the CYP450 classification models can be ranked from the best to the worst as XGBoost > GBDT > RF > MAE-DNN > CNN > SAE-DNN > DNN. In summary, considering the overall statistics and prediction accuracy, ensemble learning methods outperform deep learning methods, and the XGBoost models achieve the best predictions for the classification of CYP450 inhibitors and noninhibitors.

3.2. Comparison of XGBoost models based on different sets of descriptors.

In QSAR modeling, the structure of a molecule is encoded by molecular descriptors, and therefore selection of the most appropriate descriptors is quite critical for the development of a reliable QSAR model. Here, a number of XGBoost models were developed based on different sets of molecular descriptors, including two sets of molecular descriptors generated by MOE and PaDEL-Descriptor, three sets of molecular fingerprints (PubFP, KlebFP, GraFP and MorFP) generated by the PaDEL-Descriptor and RDKit packages in python. First, we compared the predictive powers of the XGBoost models only based on any set of molecular descriptors (MOE or PaDEL) or any set of molecular fingerprints (PubFP, GraFP, KleFP, or MorFP). As shown in **Table S4** in Supporting Materials, the XGBoost model based on PaDEL yields an average prediction accuracy of 89.8% for the five external test sets and performs better than the other XGBoost models based on MOE (88.5%), GraFP (86.6%), KlebFP (88.5%), MorFP (88.1%) or PubFP(88.4%). The accuracy of the models was similar, so MCC was used for further comparison. The average MCC values of the different XGBoost classifiers for the test sets are shown in **Figure 5**. Apparently, the XGBoost models based on molecular descriptors outperform those based on molecular fingerprints. The XGBoost model based on PubFP yields an average MCC value of 52.8% for the five external test sets and performs better than the other two models based the KleFP (51.7%), MorFP (51.0%) or GraFP (44.0%) fingerprints. Then, we developed two XGBoost models based on the combination of PubFP and molecular descriptors (PubFP+MOE and PubFP+PaDel) for comparison. As shown in **Figure 5**, the models based on PubFP+MOE or PubFP+PaDel yield higher MCC values than those only based on molecular descriptors or fingerprints, suggesting that the combination of molecular descriptors and fingerprints is a better way to characterize the chemical and structural features of the studied molecules. The average prediction accuracies of the different XGBoost classifiers for the test sets are shown in **Figure 6**. According to MCC, ACC and other evaluation parameters listed in **Table S4**, the XGBoost models based on PubFP+PaDel achieve the best predictions.

According to the statistical results shown in **Table 4**, the XGBoost models based on PubFP+Padel perform better than the MAE-DNN model for both the training and test sets. The prediction accuracies of the best XGBoost models for the test sets are 97.4%, 90.2%, 82.3%, 92.8% and 89.4% for 1A2, 2C9, 2C19, 2D6 and 3A4, respectively, and those of the MAE-DNN model are 96.8%, 86.0%, 80.9%, 89.3% and 89.6%, respectively. The ROC curves given by XGBoost models for the training and test sets are shown in **Figure S2** of Supporting Materials. The AUCs of the best XGBoost models for the test sets are 0.991, 0.814, 0.842 0.863 and 0.935 for 1A2, 2C9, 2C19, 2D6 and 3A4, respectively, and those of the MAE-DNN model are 0.982, 0.799, 0.832, 0.878 and 0.929, respectively. Therefore, according to the prediction accuracies, the XGBoost models offer obvious improvement over the MAE-DNN model for the four CYP450 isoforms (1A2, 2C9, 2C19 and 2D6), and according to the AUC values, the XGBoost models still offer obvious improvement over the MAE-DNN model for the four CYP450 isoforms (1A2, 2C9, 2C19 and 3A4).

As shown in **Table 4**, the SE values are lower than the SP values for all the five XGboost models due to the imbalanced nature of the training sets. Especially, for the 2C9 and 2D6 isoforms, the SE values of the XGBoost models are as low as 29.0% and 41.6%, respectively, which is consistent with that the fact that the training sets for CYP2C9 and CYP2D6 are much more unbalanced than those for the other three CYP450 isoforms. The models learned more about non-inhibitors because the training sets contains more non-inhibitors than inhibitors, and therefore the prediction accuracies of the model for non-inhibitors are higher than those for inhibitors.

3.3. Shapley additive explanations of XGBoost models.

In order to judge whether the XGBoost classification models are reasonable and gain a deeper insight into the models, the importance of the molecules descriptors used by the models was analyzed by the SHAP method. The top 20 representative molecular descriptors (top 20) and the corresponding SHAP values for the XGBoost models are shown in **Figure 7**. The descriptions of these representative molecular descriptors are listed in **Table S5** in Supporting Materials.

CYP1A2 model. As shown in **Figure 7A**, R_TpiPCTPC, which represents the ratio of total conventional bond order with total path count, is the most important molecular descriptor in the XGBoost model for CYP1A2. Generally, small molecules with higher degree of unsaturation have higher R_TpiPCTPC values. And according to the SHAP value and feature value of R_TpiPCTPC, a higher R_TpiPCTPC will increase the probability of a molecule to be predicted as an inhibitor by the CYP1A2 model. Similarly, higher values of the number of atoms in the largest pi system (nAtomP) and the Crippen's logP (CrippenLogP) will increase the probability while higher values of the number of basic groups (nBase) and the number of acidic groups (nAcid) will decrease the probability of a molecule to be predicted as an inhibitor. R_TpiPCTPC and nAtomP are related to the aromaticity of a molecule. CrippenLogP, nBase and nAcid are related to the solubility of a molecule. All of the above results are consistent with the Sansen's study that CYP1A2 contains a compact and closed active site suitable for accommodating lipophilic planar polyaromatic or polyheteroaromatic small molecules.[58]

CYP2C9 model. As shown in **Figure 7B**, CrippenLogP, SddsN, nBase and XlogP are the four most important molecular descriptors in the CYP2C9 model. Higher CrippenLogP and XlogP and lower SddsN and nBase will increase the probability of a molecule to be predicted as a CYP2C9 inhibitor. SddsN, which represents the sum of atom-type E-State: -N<<, is the second most important molecular descriptor. However, only eight molecules in the training set contain this structure (-N<<), which is not consistent with its importance determined by the SHAP method. In order to explore what caused this unreasonable result, the SddsN's SHAP value for each molecule was calculated. As shown in **Figure 8**, most molecules without the structure (-N<<) in the training set have the same SddsN value, but the same SddsN has completely different attributed importance. Due to the imbalanced training set for CYP2C9 (non-inhibitors are 2.3 times more than inhibitors), molecules without the structure (-N<<) are more likely to be predicted as noninhibitors by the CYP2C9 model. Thus, SddsN is a molecular descriptor that may introduce noise to the development of the CYP2C9 model. Moreover, we observed that minddsN, which represents the minimum atom-type E-state: -N<<, may be also a noisy molecular descriptor to model building. To prove the above conclusion, the CYP2C9 XGBoost model was then developed in the same way after removing these two molecular descriptors. The performances of the new model and the original model on the test sets are summarized in **Table 5**. Apparently, the new CYP2C9 model performs better for the test sets than the original one. The SE value of the model for the test set increases from 0.290 to 0.377. The importance of the representative molecular descriptors (top 20) and the corresponding SHAP values given by the new CYP2C9 model are shown in **Figure 9**. CrippenLogP, nBase and XlogP are also the most important molecular descriptors in the new CYP2C9 model, and therefore molecules with high lipophilicity are more likely to be predicted as CYP2C9 inhibitors. Higher ETA_Beta_ns (a measure of electron-richness of the molecule) will increase the probability of a molecule to be predicted as a CYP2C9 inhibitor, which is consistent with the trend that CYP2C9 favors more aromatic ring. SaaN, which represents the sum of atom-type E-State: :N:,[59] is the fifth most important molecular descriptor for the new CYP2C9 model, and a molecule with a lower SaaN is more likely to be predicted as a CYP2C9 inhibitor.

CYP2C19 model. CYP2C19 and CYP2C9 belong to the same subfamily and they share 91% amino acid sequence identity. The highly sequence similarity is also reflected by our models (**Figure 7C**). The top 3 molecular descriptors for the CYP2C9 and CYP2C19 models are identical. Similar to the CYP2C9 model, higher XLogP, CrippenLogP and ETA_Beta_ns and lower nBase and SaaN will increase the probability of a molecule to be predicted as a CYP2C19 inhibitor. The results are in good agreement with the reported trend that CYP2C19 and CYP2C9 tend to form more affinity with lipophilic ligands than the other isoforms.[12] PubchemFP594 (C-O-C-C=C), as the only molecular fingerprint among the top 20 important features, has attracted our attention. The SHAP values shown in **Figure 10** illustrate 49.0% of CYP2C19 inhibitors contain the substructure (C-O-C-C=C) while only 29.3% CYP2C19 noninhibitors contain this substructure, suggesting that those molecules with the substructure (C-O-C-C=C) are more likely to be predicted as CYP2C19 inhibitors.

CYP2D6 model. The training set for CYP2D6 is the most imbalanced among the five training sets. Only 13.37% of the molecules in the training set are inhibitors, thus resulting in a higher sensitivity than specificity for the CYP2D6 model. Earlier studies have shown that CYP2D6 inhibitors typically contain a basic nitrogen and a planar aromatic ring.[60] As shown in **Figure 7D**, the most important molecular descriptor, maxaaCH (maximum atom-type E-State: :CH:), is directly related to the aromatic ring. In addition, the descriptors, including SaaCH (sum of atom-type E-State: :CH:), maxHaaCH (maximum atom-type H E-State::CH:) and maxaasC (maximum atom-type E-State: :CH:), are also directly related to the aromatic ring. nBase is the second most important molecular descriptor in the CYP2D6 model. Higher nBase will increase the probability of a molecule to be predicted as a CYP2D6 inhibitor, while lower nBase is more favorable as inhibitors of the other CYP450 isoforms, which is in accordance with the fact that the basic nitrogen atoms are usually contained in CYP2D6 inhibitors. Moreover, as shown **Figure 5D**, several important molecular descriptors, including maxssNH (maximum atom-type E-State: -NH-), minsssN (minimum atom-type E-State: >N-) and maxsssN (maximum atom-type E-State: >N-), are directly related to basic nitrogen atoms. LipoaffinityIndex, the parameter that quantitatively describes the lipophilicity of a molecule, is the third most important feature in the CYP2D6 model. Higher LipoaffinityIndex value will increase the probability of a molecule to be predicted as a CYP2D6 inhibitor, which is similar to the other models. As shown in **Figure 5D**, existence of PubchemFP367 (C(~H)(~O)(~O)) will increase the probability of a molecule to be predicted as a CYP2D6 inhibitor, which is also consistent with the results reported by Li and coworkers.[12]

CYP3A4 model. As shown in **Figure 7E**, ETA_Beta (a measure of the electronic feature of a molecule) is the most important molecular descriptor in the CYP3A4 model, and a higher ETA_Beta will increase the probability of a molecule to be predicted as a CYP3A4 inhibitor. ETA_Beta contains the following structural information of a molecule: (1) the more electronegative atoms the a molecule contains, the higher the ETA value is; (2) the higher the molecular unsaturation, the higher the ETA value is,[61] suggesting that molecules with more electronegative atoms and higher molecular unsaturation may form more favorable interactions with CYP3A4, which is consistent with previous study that CYP3A4 tends to recognize molecules with more six-membered rings.[8] Similarly, IC3 is the third important molecular descriptor in the CYP3A4 model that describes molecular complexity. Higher IC3 value will increase the probability of a molecule to be predicted as a CYP2D6 inhibitor. Similar to the other models, a molecule with higher CrippenLogP will be more likely to be predicted as a CYP2C19 inhibitor. As shown in **Figure 11**, a higher PubchemFP372 (C(~H)(:C)(:N)) will increase the probability of a molecule to be predicted as a CYP3A4 inhibitor. The structural analysis shows that 31.3% of CYP3A4 inhibitors contain the substructure (C(~H)(:C)(:N)), while only 15.6% of noninhibitors contain this substructure.

3.4. Analysis of Misclassified Molecules.

The XGBoost models show the best predictive performance for the test sets (average prediction accuracy of 90.4%). However, a total of 512 molecules (355 inhibitors and 157 noninhibitors) in the test sets cannot be correctly predicted. The following reasons may be responsible for the misclassification. First, the quality of the data set may be a main source of misclassification. The compounds in the original dataset were categorized into inhibitors and noninhibitors based on relatively arbitrary criteria. It is obvious that compounds close to the classification criteria may be more likely to be misclassified. In order to prove this hypothesis, we divided the compounds in the test sets into two categories: compounds close to the classification criteria (CCCC) and compounds far from the classification criteria (CFCC). Inhibitors with PubChem activity scores between 40 and 60 are CCCC, and those with scores greater than 60 are CFCC. However, The PubChem activity scores of all non-inhibitors were 0. Therefore, max response, another experimental indicator, was used as the classification standard for non-inhibitors. In this study, noninhibitors with max response values less than 0 are CCCC, and those with response values higher than 0 are CFCC. The misclassification ratio (MR) defined by Equation 7 was used to compare the probability that CCCC and CFCC are misclassified.

$$MR = \frac{\text{the number of misclassified compounds}}{\text{total number of compounds}} \qquad (8)$$

The information of the misclassified compounds in the test sets predicted by the XGBoost models is listed in **Table 6**. For inhibitors, CCCC achieves higher MR values for the CYP1A2 (10.9%), CYP2D6 (69.0%) and CYP3A4 (43.0%) isoforms, and CFCC achieves higher MR values for the CYP2C9 (63.0%) and CYP2C19 (37.7%) isoforms. For noninhibitors, CCCC achieves higher MR values for the CYP1A2 (1.6%), CYP2C19 (17.3%), CYP2D6 (1.8%) and CYP3A4 (2.9%) isoforms and CFCC only achieves a higher MR value for CYP2C9 (4.4%). The above results are basically consistent with our hypothesis that compounds close to the classification criteria may be more likely to be misclassified than those far from the classification criteria. And this may be one of the reasons why compounds in the test sets are misclassified

As mentioned earlier, $g(z')$ is the sum of the molecular descriptor attribution and approximates to the output of the original model. The SHAP value of each molecular descriptor can also be used to analyze why molecules are misclassified. Taking (1S,2S)-2-(methylamino)-1-phenylpropan-1-ol (Pubchem SID 11111142) as an example, the molecular descriptors that push the CYP2C9 model output from the base value (the average model output over the training set) to the model output are show in **Figure 12,** suggesting that the contributions of nBase and CrippenLogP may cause the misclassification of (1S,2S)-2-(methylamino)-1-phenylpropan-1-ol to be a noninhibitor with the CYP2C9 model. Misclassified molecules in the five test sets were analyzed in the same way, and the molecules with the largest prediction errors given by the five XGBoost models are listed in **Table 7**. Apparently, the main reason why molecules are misclassified can be explained by the fact that their molecular descriptors do not conform to the model rules learned from the training set. It is worth noting that (1S,2S)-2-(methylamino)-1-phenylpropan-1-ol is the false-negative molecule with the largest prediction error for the CYP2C9 and CYP3A4 models and is also the false-negative for the other three models. As mentioned above, the CYP2C9 and CYP2C19 models are quite similar due to the high similarity between CYP2C9 and CYP2C19. Molecules with high lipophilicity while without basic groups are more likely to be predicted as inhibitors of CYP2C9 and CYP2C19. However, (1S,2S)-2-(methylamino)-1-phenylpropan-1-ol has low lipophilicity and contains a basic nitrogen so that it is misclassified as a CYP2C9 and CYP2C19 noninhibitor, and it is also misclassified as a noninhibitor by the other models due to its low lipophilicity. Similar to (1S,2S)-2-(methylamino)-1-phenylpropan-1-ol, NCGC00074720-01 also contains some basic nitrogen atoms and has low lipophilicity due to the presence of multiple hydrophilic groups, and therefore it is misclassified as a noninhibitor of CYP2C19. As shown in **Table 7**, misclassifications of the five molecules with the largest prediction errors are caused by lipophilicity, suggesting that our models may assigns too much weight to lipophilicity.

4. Conclusion

In our study, three ensemble learning methods (RF, GBDT and XGBoost) and two deep learning methods (DNN and CNN) were used to develop the classification models for five CYP450 (1A2, 2C9, 2C19, 2D6, and 3A4) isoforms. The ensemble learning models, on average, show better performance than the deep learning models (including the previously reported deep leaning models) for the training and test sets. Among all the models developed by different machine learning method, the XGBoost models yield the best prediction accuracy. In order to evaluate the effect of different molecular descriptors on model building, the XGBoost models based on different sets of molecular descriptors and fingerprints were developed, including PubFP, GraFP, KleFP, MOE, PubFP+MOE and PubFP+Padel. The XGBoost models based on PubFP+Padel give the best predictions for the external test sets (the accuracies of 97.4%, 90.1%, 82.3%, 92.8% and 89.4% for 1A2, 2C9, 2C19, 2D6 and 3A4, respectively). Moreover, the SHAP method was used to interpret the models, and the important descriptors and fingerprints were highlighted by the SHAP values. The information given by SHAP is consistent with the structural preferences for CYP450 inhibition and provides valuable clues to detect potential drug-drug interactions in early stage of drug discovery. We believe that the XGBoost models with high prediction accuracy developed in this study can be used as valuable tools in the identification of drug candidates that are unlikely to inhibit multiple CYP450 isoforms.

Acknowledgments

This work was financially supported by National Key R&D Program of China (2016YFB0201700, 2016YFA0501701), National Science Foundation of China (21575128, 81773632), and Zhejiang Provincial Natural Science Foundation of China (LZ19H300001). We are very grateful to Dr. Jianfeng Pei from Peking University for providing the datasets of CYP450 inhibition.

Supplementary Data

Table S1. The main hyperparameters for the RF models; **Table S2**. The main hyperparameters for the GBDT models; **Table S3.** Performances of different models on the training and test sets; **Table S4.** Performances of the XGBoost models based on different sets of descriptors; **Table S5**. The descriptions of the representative molecular descriptors; **Figure S1.** The multidimensional scaling (MDS) plots for the (A) CYP1A2, (B) CYP2C9, (C) CYP2C19, (D) CYP2D6, and (E) CYP3A4 datasets; **Figure S2.** The ROC curves of different XGBoost models for (A) the training sets and (B) the test sets.

References

1. Miners, J. O.; Mackenzie, P. I.; Knights, K. M., The prediction of drug-glucuronidation parameters in humans: UDP-glucuronosyltransferase enzyme-selective substrate and inhibitor probes for reaction phenotyping and in vitro-in vivo extrapolation of drug clearance and drug-drug interaction potential. *Drug Metab. Rev.* **2010**, 42, 196-208.
2. Williams, J. A.; Hyland, R.; Jones, B. C.; Smith, D. A.; Hurst, S.; Goosen, T. C.; Peterkin, V.; Koup, J. R.; Ball, S. E., Drug-drug interactions for UDP-glucuronosyltransferase substrates: A pharmacokinetic explanation for typically observed low exposure (AUC(i)/AUC) ratios. *Drug Metab. Dispos.* **2004**, 32, 1201-1208.
3. Lasser, K. E.; Allen, P. D.; Woolhandler, S. J.; Himmelstein, D. U.; Wolfe, S. N.; Bor, D. H., Timing of new black box warnings and withdrawals for prescription medications. *Jama-J. Am. Med. Assoc.* **2002**, 287, 2215-2220.
4. Backman, J.; Wang, J.; Wen, X.; Kivistö, K.; Neuvonen, P., Mibefradil but not isradipine substantially elevates the plasma concentrations of the CYP3A4 substrate triazolam. *Clin. Pharmacol. Ther.* **1999**, 66, 401-7.
5. Breimer, D. D., Interindividual variations in drug disposition. Clinical implications and methods of investigation. *Clin. Pharmacokinet.* **1983**, 8, 371-377.

6. Cheng, F. X.; Yu, Y.; Shen, J.; Yang, L.; Li, W. H.; Liu, G. X.; Lee, P. W.; Tang, Y., Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers. *J. Chem. Inf. Model.* **2011**, 51, 996-1011.

7. Su, B. H.; Tu, Y. S.; Lin, C.; Shao, C. Y.; Lin, O. A.; Tsene, Y. J., Rule-Based Prediction Models of Cytochrome P450 Inhibition. *J. Chem. Inf. Model.* **2015**, 55, 1426-1434.

8. Sun, H. M.; Veith, H.; Xia, M. H.; Austin, C. P.; Huang, R. L., Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data. *J. Chem. Inf. Model.* **2011**, 51, 2474-2481.

9. Veith, H.; Southall, N.; Huang, R. L.; 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. *Nature Biotechnol.* **2009**, 27, 1050-U123.

10. Cao, C. S.; Liu, F.; Tan, H.; Song, D. S.; Shu, W. J.; Li, W. Z.; Zhou, Y. M.; Bo, X. C.; Xie, Z., Deep Learning and Its Applications in Biomedicine. *GPB* **2018**, 16, 17-32.

11. Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P. M.; Zietz, M.; Hoffman, M. M.; Xie, W.; Rosen, G. L.; Lengerich, B. J.; Israeli, J.; Lanchantin, J.; Woloszynek, S.; Carpenter, A. E.; Shrikumar, A.; Xu, J. B.; Cofer, E. M.; Lavender, C. A.; Turaga, S. C.; Alexandari, A. M.; Lu, Z. Y.; Harris, D. J.; DeCaprio, D.; Qi, Y. J.; Kundaje, A.; Peng, Y. F.; Wiley, L. K.; Segler, M. H. S.; Boca, S. M.; Swamidass, S. J.; Huang, A.; Gitter, A.; Greene, C. S., Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, 15, 20170387.

12. Li, X.; Xu, Y. J.; Lai, L. H.; Pei, J. F., Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol. Pharm.* **2018**, 15, 4336-4345.

13. Jimenez-Carretero, D.; Abrishami, V.; Fernandez-de-Manuel, L.; Palacios, I.; Quilez-Alvarez, A.; Diez-Sanchez, A.; del Pozo, M. A.; Montoya, M. C., Tox_(R)CNN: Deep learning-based nuclei profiling tool for drug toxicity screening. *PLoS Comput. Biol.* **2018**, 14, e1006238.

14. Wenzel, J.; Matter, H.; Schmidt, F., Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, 59, 1253-1268.

15. Ye, Z.; Yang, Y.; Li, X.; Cao, D.; Ouyang, D., An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction. *Mol. Pharm.* **2019**, 16, 533-541.

16. Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L., Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, 55, 2085-2093.

17. Xu, Y.; Pei, J.; Lai, L., Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction. *J. Chem. Inf. Model.* **2017**, 57, 2672-2685.

18. Ekins, S., The Next Era: Deep Learning in Pharmaceutical Research. *Pharm. Res.* **2016**, 33, 2594-2603.

19. Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S., Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharm.* **2017**, 14, 4462-4475.

20. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. arXiv e-prints2016; https://ui.adsabs.harvard.edu/\#abs/2016arXiv160302754C (accessed March 01, 2016).

21. Adam-Bourdarios, C.; Cowan, G.; Germain-Renaud, C.; Guyon, I.; Kegl, B.; Rousseau, D. The Higgs Machine Learning Challenge. In *21st International Conference on Computing in High*

Energy and Nuclear Physics; 2015; Vol. 664.

22. Volkovs, M.; Yu, G. W.; Poutanen, T. Content-based Neighbor Models for Cold Start in Recommender Systems. In *Proceedings Of the Recommender Systems Challenge Workshop 2017*; 2017.

23. Sandulescu, V.; Chiru, M. Predicting the future relevance of research institutions - The winning solution of the KDD Cup 2016. arXiv e-prints2016; https://ui.adsabs.harvard.edu/\#abs/2016arXiv160902728S (accessed September 01, 2016).

24. Ashtawy, H. M.; Mahapatra, N. R., Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment. *J. Chem. Inf. Model.* **2018**, 58, 119-133.

25. Wang, B.; Zhao, Z.; Nguyen, D. D.; Wei, G.-W., Feature functional theory-binding predictor (FFT-BP) for the blind prediction of binding free energies. *Theor. Chem. Acc.* **2017**, 136, 1-22.

26. Cang, Z.; Mu, L.; Wei, G.-W., Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **2018**, 14, e1005929.

27. Cang, Z.; Wei, G.-W., Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int. J. Numer. Meth. Bio. Eng.* **2018**, 34, e2914.

28. Duc Duy, N.; Wei, G.-W., DG-GL: Differential geometry-based geometric learning of molecular datasets. *International Journal for Numerical Methods in Biomedical Engineering* **2019**, 35, e3179.

29. Lei, T. L.; Sun, H. Y.; Kang, Y.; Zhu, F.; Liu, H.; Zhou, W. F.; Wang, Z.; Li, D.; Li, Y. Y.; Hou, T. J., ADMET Evaluation in Drug Discovery. 18. Reliable Prediction of Chemical-Induced Urinary Tract Toxicity by Boosting Machine Learning-Approaches. *Mol. Pharm.* **2017**, 14, 3935-3953.

30. Shan, X.; Wang, X.; Li, C.-D.; Chu, Y.; Zhang, Y.; Xiong, Y. I.; Wei, D.-Q., Prediction of CYP450 Enzyme-Substrate Selectivity Based on the Network-based Label Space Division Method. *Journal of chemical information and modeling* **2019**.

31. Beisken, S.; Meinl, T.; Wiswedel, B.; de Figueiredo, L. F.; Berthold, M.; Steinbeck, C., KNIME-CDK: Workflow-driven cheminformatics. *BMC Bioinf.* **2013**, 14, 257.

32. Yap, C. W., PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, 32, 1466-1474.

33. *MOE molecular simulation package*, Chemical Computing Group Inc.: Montreal, Candada, 2010.

34. Yap, C. W., PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J Comput Chem* **2011**, 32, 1466-1474.

35. Hughes, T. B.; Miller, G. P.; Swamidass, S. J., Modeling Epoxidation of Drug-like Molecules with a Deep Machine Learning Network. *ACS Cent. Sci.* **2015**, 1, 168-180.

36. Jimenez-Carretero, D.; Abrishami, V.; Fernandez-de-Manuel, L.; Palacios, I.; Quilez-Alvarez, A.; Diez-Sanchez, A.; del Pozo, M. A.; Montoya, M. C., Tox_(R)CNN: Deep learning-based nuclei profiling tool for drug toxicity screening. *Plos Comput Biol* **2018**, 14.

37. Lusci, A.; Pollastri, G.; Baldi, P., Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **2013**, 53, 1563-1575.

38. Dai, H.; Xu, Q.; Xiong, Y.; Liu, W.-L.; Wei, D.-Q., Improved Prediction of Michaelis Constants in CYP450-Mediated Reactions by Resilient Back Propagation Algorithm. *Current Drug*

Metabolism **2016**, 17, 673-680.

39. Ezzat, A.; Wu, M.; Li, X. L.; Kwoh, C. K., Drug-target interaction prediction via class imbalance-aware ensemble learning. *BMC Bioinf.* **2016**, 17, 509.

40. Lei, T. L.; Chen, F.; Liu, H.; Sun, H. Y.; Kang, Y.; Li, D.; Li, Y. Y.; Hou, T. J., ADMET Evaluation in Drug Discovery. Part 17: Development of Quantitative and Qualitative Prediction Models for Chemical-Induced Respiratory Toxicity. *Mol. Pharm.* **2017**, 14, 2407-2421.

41. Lei, T. L.; Li, Y. Y.; Song, Y. L.; Li, D.; Sun, H. Y.; Hou, T. J., ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. *J. Cheminf.* **2016**, 8, 6.

42. Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*; 2000; Vol. 1857, pp 1-15.

43. Schapire, R. E., The Strength of Weak Learnability. *Mach. Learn.* **1990**, 5, 197-227.

44. Rokach, L., Ensemble-based classifiers. *Artif. Intell. Rev.* **2010**, 33, 1-39.

45. Friedman, J. H., Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, 29, 1189-1232.

46. Breiman, L., Random forests. *Mach. Learn.* **2001**, 45, 5-32.

47. Chen, T. Q.; Guestrin, C., In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco, California, USA, 2016, pp 785-794.

48. Mason, L.; Baxter, J.; Bartlett, P.; Frean, M. Boosting algorithms as gradient descent. In International Conference on Neural Information Processing Systems, 1999; 1999; Vol. 12; pp 512-518.

49. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P., Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1947-1958.

50. Jing, Y. K.; Bian, Y. M.; Hu, Z. H.; Wang, L. R.; Xie, X. Q. S., Deep Learning for Drug Design: an Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS J.* **2018**, 20, UNSP 58.

51. Krizhevsky, A.; Sutskever, I.; Hinton, G. E., ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, 60, 84-90.

52. Datta, A.; Sen, S.; Zick, Y. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 LEEE Symposium on Security and Privacy*; 2016, pp 598-617.

53. Lipovetsky, S.; Conklin, M., Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business Industry* **2010**, 17, 319-330.

54. Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv e-prints2016.

55. Strumbelj, E.; Kononenko, I., Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **2014**, 41, 647-665.

56. Lundberg, S.; Lee, S. I. A unified approach to interpreting model predictions. arXiv e-prints2017.

57. Lundberg, S. M.; Lee, S. I. Consistent feature attribution for tree ensembles. arXiv e-prints2017.

58. Stefaan, S.; Yano, J. K.; Reynald, R. L.; Schoch, G. A.; Griffin, K. J.; C David, S.; Johnson, E. F., Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. *J. Biol. Chem.* **2007**, 282, 14348-14355.

59. Breiman, L., Bagging predictors. *Mach. Learn.* **1996**, 24, 123-140.

60. Paul, R.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon, V. R.; Oxbrow, A. K.; Lewis, C. J.; Tennant, M. G.; Sandeep, M.; Eggleston, D. S., Crystal structure of human cytochrome P450 2D6. *J. Biol. Chem.* **2006**, 281, 7614-7622.

61. Roy, K.; Ghosh, G., QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 559.

Legend of Figures

Figure 1. Overall workflow of model building based on different machine learning methods

Figure 2. Overall workflow of the XGBoost model building based on different sets of descriptors

Figure 3. Cross-validated accuracies of different models for the five training sets.

Figure 4. Accuracies of different models for the five test sets.

Figure 5. MCC values of the XGBoost models based on different sets of descriptors for the five test sets.

Figure 6. AUC values of the XGBoost models based on different sets of descriptors for the five test sets.

Figure 7. The importance of the representative molecular descriptors (top 20) and the SHAP values for each molecular descriptor given by the (A) CYP1A2 models, (B) CYP2A9 models, (C) CYP2C19 models, (D) CYP2C19 and (E) CYP3A4 models. The higher the SHAP value of a molecular descriptor, the more likely this molecule will be predicted as an inhibitor by the model. One molecule gets one dot on each descriptor's line and dots stack up to show density.

Figure 8. The SHAP value of SddsN in each molecule for the 2C19 model.

Figure 9. The importance of the representative molecular descriptors (top 20) and the SHAP values for each molecular descriptor given by the improved CYP2C9 model

Figure 10. The SHAP value of PubchemFP594 in each molecule for the CYP2C19 model.

Figure 11. The SHAP value of PubchemFP372 in each molecule for the CYP3A4 model.

Figure 12. Molecular descriptors of (1S,2S)-2-(methylamino)-1-phenylpropan-1-ol that push output given by the CYP2C9 model from the base value (the average model output over the training set). The descriptors pushing the prediction to be an inhibitor are colored red, and those pushing the prediction to be a noninhibitor are colored blue.

Table 1. Information of the datasets.

Isoform	Datasets	No. of inhibitors	No. of noninhibitors	No.
CYP1A2	training set	3559	5929	9488
	test set	107	477	584
CYP2C9	training set	2552	6833	9385
	test set	69	596	665
CYP2C19	training set	4450	5608	10058
	test set	142	569	711
CYP2D6	training set	1351	8752	10103
	test set	77	671	748
CYP3A4	training set	3070	5985	9055
	test set	537	1845	2382

Table 2. Number of the molecular descriptors

Isoform	PubFP	MorFP	KleFP	GraFP	MOE	PaDel	PubFP+MOE	PubFP+ PaDel
1A2	687	1024	3181	1022	338	1252	1025	1936
2C9	687	1024	3163	1022	338	1252	1025	1937
2C19	686	1024	3200	1022	338	1252	1024	1936
2D6	683	1024	3232	1022	338	1244	1021	1926
3A4	688	1024	3249	1023	338	1251	1026	1925

Abbreviations: PubFP, Pubchem fingerprints; MorFP, Morgan fingerprints; KleFP, KlekotaRoth fingerprints; GraFP, GraphOnly fingerprints; Des(MOE), molecular descriptors calculated by MOE; PaDel, 1D&2D descriptors calculated by PaDel; PubFP+MOE, Pubchem fingerprints and molecular descriptors calculated by MOE; PubFP+PaDel, Pubchem fingerprints and 1D&2D descriptors calculated by PaDel.

Table 3. The main hyperparameters for the XGBoost models

Hyperparameter	CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP3A4
learning_rate	0.005	0.005	0.005	0.1	0.005
n_estimators	19620	8900	5000	174	5850
max_depth	6	8	7	5	5
min_child_weight	2	1	5	2	3
gamma	0.06	0.02	0	0.02	0.23
colsample_bytree	0.6	0.65	0.8	0.85	0.8
subsample	0.9	0.6	0.8	0.8	0.8
reg_alpha	0.01	1	5e-5	1	5e-6
reg_lambda	0.1	0.01	0.5	0.01	1

Table 4. Performances of the XGBoost models for the training and test sets

Isoform	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.905	0.862	0.930	0.796	0.962	0.974	0.916	0.987	0.913	0.991
2C9	0.876	0.733	0.930	0.680	0.931	0.902	0.290	0.973	0.354	0.814
2C19	0.850	0.843	0.856	0.697	0.917	0.823	0.669	0.861	0.493	0.842
2D6	0.909	0.434	0.982	0.542	0.877	0.928	0.416	0.987	0.537	0.863
3A4	0.860	0.757	0.913	0.683	0.931	0.894	0.618	0.975	0.677	0.935

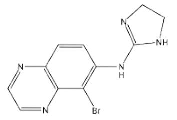
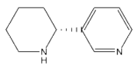
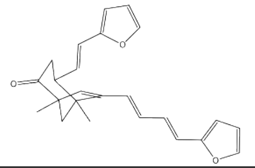
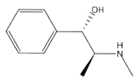
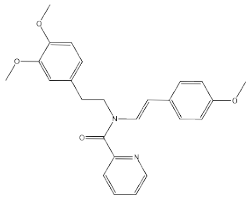
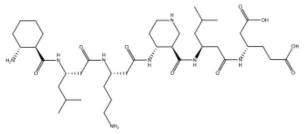
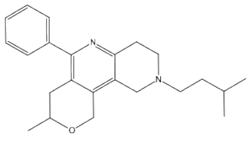
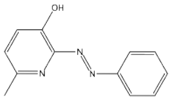
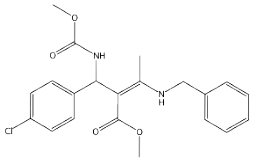
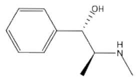
Table 5. Performances of the new CYP2C9 model and the original model for the test set.

Model	ACC	SE	SP	MCC	AUC
2C9	0.902	0.290	0.973	0.354	0.814
2C9_Del	0.901	0.377	0.961	0.395	0.826

Table 6. General information about the misclassified compounds in the test sets predicted by the XGBoost models.

Models	Inhibitor/Noninhibitor	CCCC/CFCC	Number of misclassified compounds	Number of all compounds	MR
CYP1A2	Inhibitor	CCCC	6	55	0.109
		CFCC	3	52	0.058
	Noninhibitor	CCCC	6	381	0.016
		CFCC	0	96	0
CYP2C9	Inhibitor	CCCC	9	15	0.6
		CFCC	34	54	0.63
	Noninhibitor	CCCC	11	322	0.034
		CFCC	12	274	0.044
CYP2C19	Inhibitor	CCCC	12	44	0.237
		CFCC	35	98	0.357
	Noninhibitor	CCCC	61	353	0.173
		CFCC	18	216	0.083
CYP2D6	Inhibitor	CCCC	20	29	0.69
		CFCC	25	48	0.521
	Noninhibitor	CCCC	7	392	0.018
		CFCC	3	279	0.011
CYP3A4	Inhibitor	CCCC	101	235	0.43
		CFCC	104	302	0.344
	Noninhibitor	CCCC	46	1594	0.029
		CFCC	2	251	0.008

Table 7. The misclassified compounds with the largest prediction errors in the test sets predicted by the XGBoost models.

Model	Structure	Pubchem SID	Misclassified Cause	$\lvert g(z')\rvert$	Classification Results
CYP1A2 Model		11111939	R_TpiPCTPC nAtomP	2.79	FP
		11113089	CrippenLogP nAtomP	7.74	FN
CYP2C9 Model		26752657	CrippenLogP maxaaO	2.52	FP
		11111142	nBase CrippenLogP	8.81	FN
CYP2C19 Model		26752579	CrippenLogP	3.55	FP
		11114366	CrippenLogP	-6.30	FN
CYP2D6 Model		26752358	nBase maxaaCH	1.44	FP
		11111832	SpMin6_Bhs maxaaCH	4.38	FN
CYP3A4 Model		26752612	ETA_Beta	3.46	FP
		11111142	ETA_Beta nBase	6.49	FN

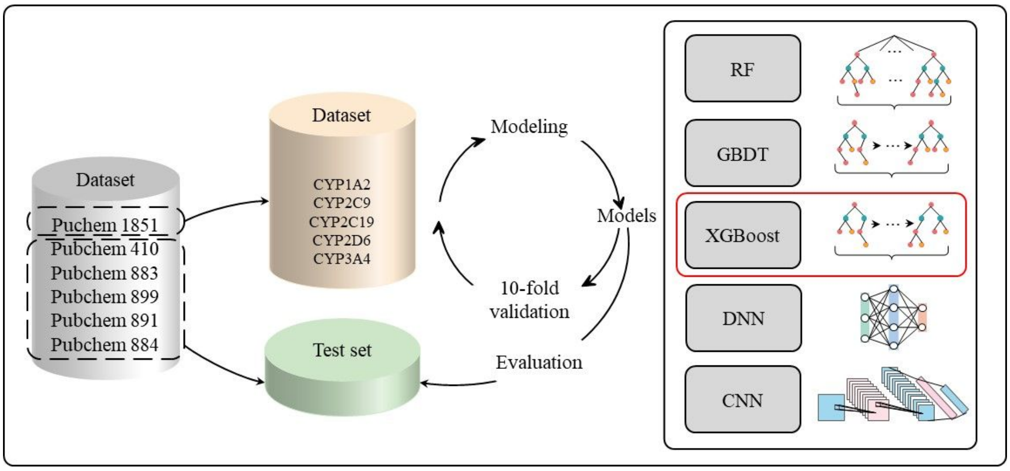

Dataset
Puchem 1851
Pubchem 410
Pubchem 883
Pubchem 899
Pubchem 891
Pubchem 884
Dataset
CYP1A2
CYP2C9
CYP2C19
CYP2D6
CYP3A4
Test set
Modeling
Models
10-fold validation
Evaluation
RF
GBDT
XGBoost
DNN
CNN

Figure 1

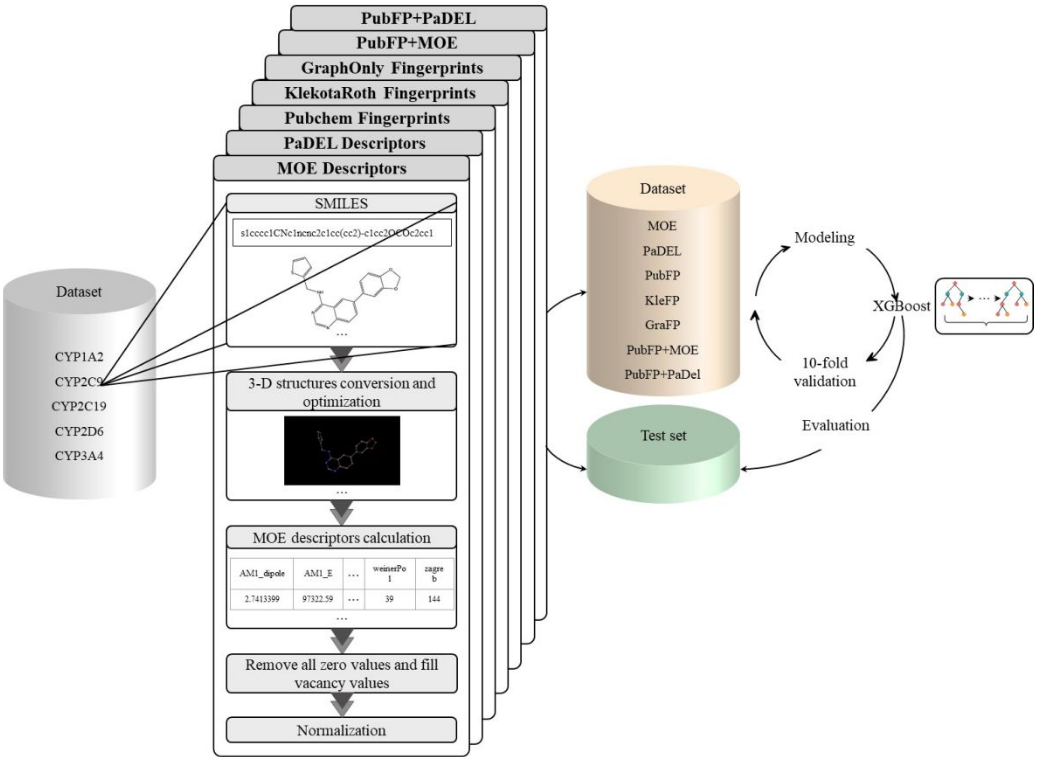

PubFP+PaDEL
PubFP+MOE
GraphOnly Fingerprints
KlekotaRoth Fingerprints
Pubchem Fingerprints
PaDEL Descriptors
MOE Descriptors
Dataset
CYP1A2
CYP2C9
CYP2C19
CYP2D6
CYP3A4
SMILES
s1cccc1CNc1ncnc2c1cc(cc2)-c1cc2OCOc2cc1
...
3-D structures conversion and optimization
...
MOE descriptors calculation
AM1_dipole
AM1_E
...
weinerPol
zagreb
2.7413399
97322.59
...
39
144
...
Remove all zero values and fill vacancy values
Normalization
Dataset
MOE
PaDEL
PubFP
KleFP
GraFP
PubFP+MOE
PubFP+PaDel
Modeling
XGBoost
10-fold validation
Evaluation
Test set

Figure 2

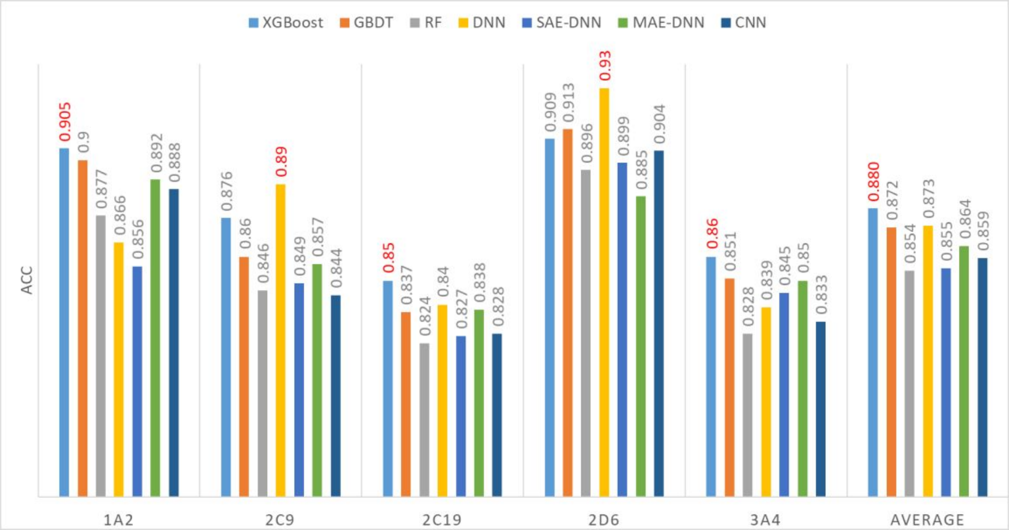

XGBoost
GBDT
RF
DNN
SAE-DNN
MAE-DNN
CNN
ACC
1A2
0.905
0.9
0.877
0.866
0.856
0.892
0.888
2C9
0.876
0.86
0.846
0.89
0.849
0.857
0.844
2C19
0.85
0.837
0.824
0.84
0.827
0.838
0.828
2D6
0.909
0.913
0.896
0.93
0.899
0.885
0.904
3A4
0.86
0.851
0.828
0.839
0.845
0.85
0.833
AVERAGE
0.880
0.872
0.854
0.873
0.855
0.864
0.859

Figure 3

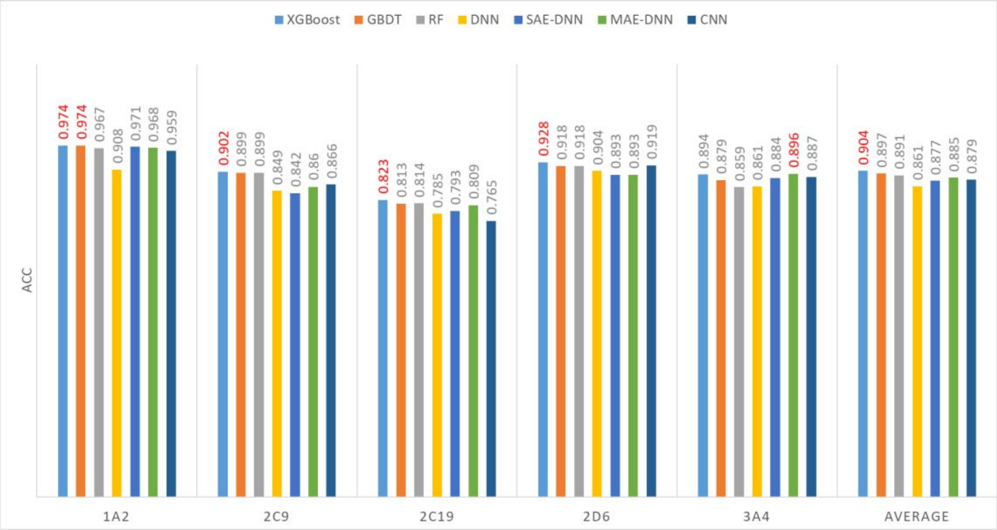

XGBoost
GBDT
RF
DNN
SAE-DNN
MAE-DNN
CNN
ACC
0.974
0.974
0.967
0.908
0.971
0.968
0.959
0.902
0.899
0.899
0.849
0.842
0.86
0.866
0.823
0.813
0.814
0.785
0.793
0.809
0.765
0.928
0.918
0.918
0.904
0.893
0.893
0.919
0.894
0.879
0.859
0.861
0.884
0.896
0.887
0.904
0.897
0.891
0.861
0.877
0.885
0.879
1A2
2C9
2C19
2D6
3A4
AVERAGE

Figure 4

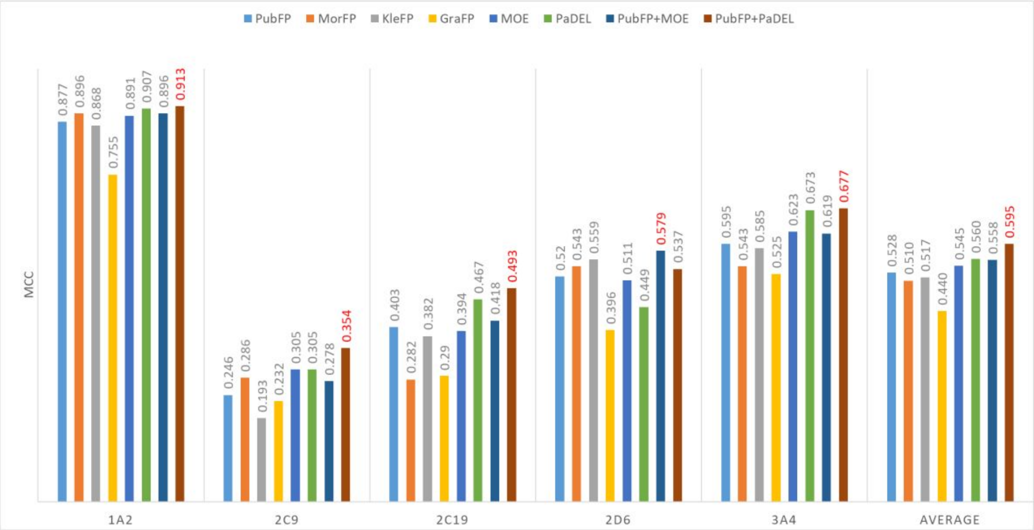

■ PubFP ■ MorFP ■ KleFP ■ GraFP ■ MOE ■ PaDEL ■ PubFP+MOE ■ PubFP+PaDEL
MCC
0.877
0.896
0.868
0.755
0.891
0.907
0.896
0.913
0.246
0.286
0.193
0.232
0.305
0.305
0.278
0.354
0.403
0.282
0.382
0.29
0.394
0.467
0.418
0.493
0.52
0.543
0.559
0.396
0.511
0.449
0.579
0.537
0.595
0.543
0.585
0.525
0.623
0.673
0.619
0.677
0.528
0.510
0.517
0.440
0.545
0.560
0.558
0.595
1A2
2C9
2C19
2D6
3A4
AVERAGE

Figure 5

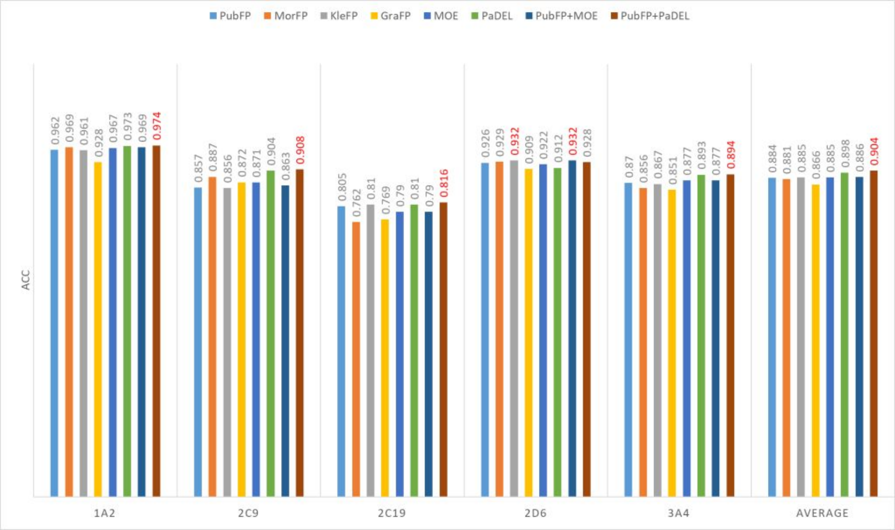
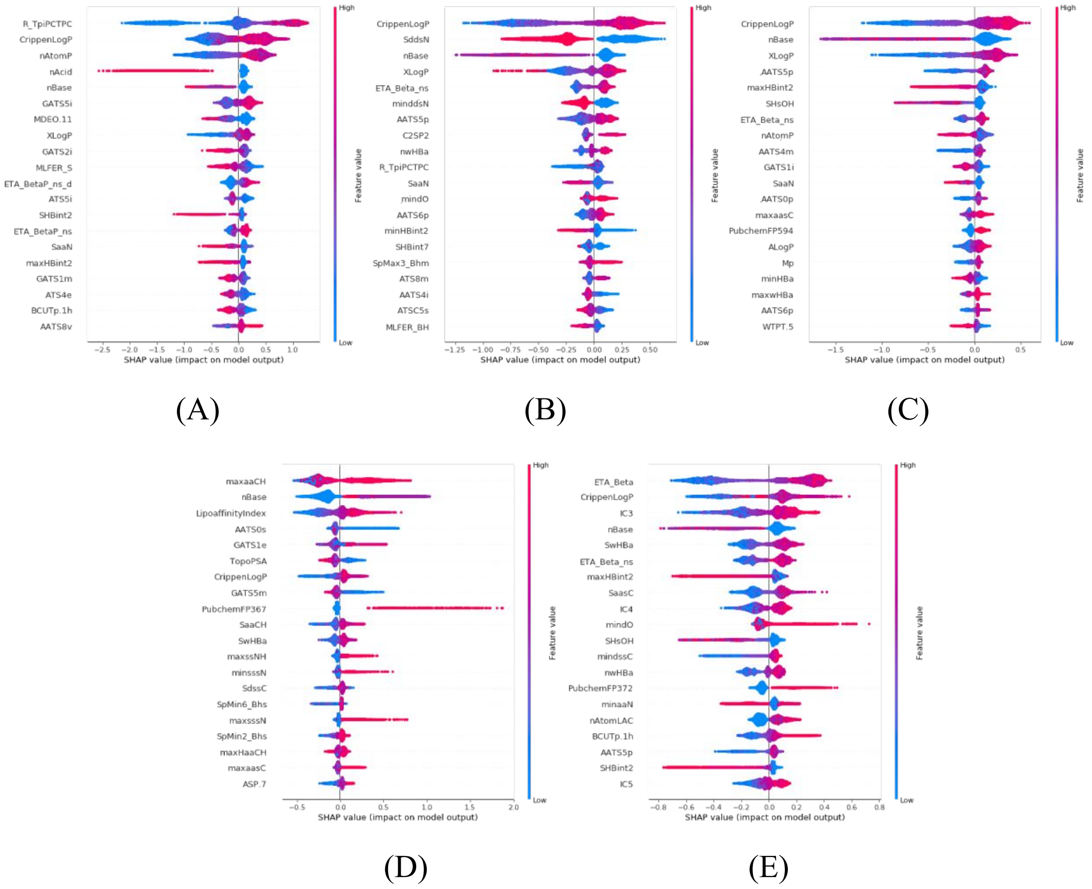

PubFP
MorFP
KleFP
GraFP
MOE
PaDEL
PubFP+MOE
PubFP+PaDEL
ACC
0.962
0.969
0.961
0.928
0.967
0.973
0.969
0.974
0.857
0.887
0.856
0.872
0.871
0.904
0.863
0.908
0.805
0.762
0.81
0.769
0.79
0.81
0.79
0.816
0.926
0.929
0.932
0.909
0.922
0.912
0.932
0.928
0.87
0.856
0.867
0.851
0.877
0.893
0.877
0.894
0.884
0.881
0.885
0.866
0.885
0.898
0.886
0.904
1A2
2C9
2C19
2D6
3A4
AVERAGE

Figure 6



(A) (B) (C)

(D) (E)

Figure 7

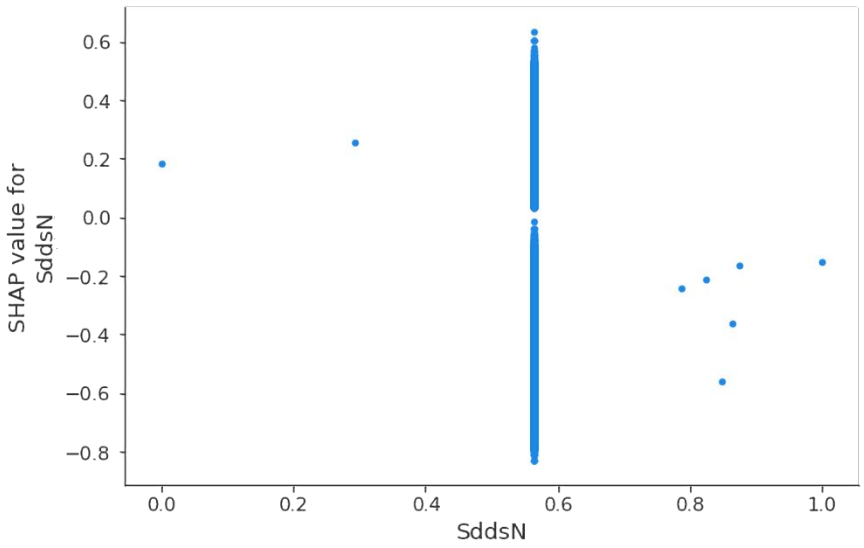

SHAP value for SddsN
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0.0
0.2
0.4
0.6
0.8
1.0
SddsN

Figure 8

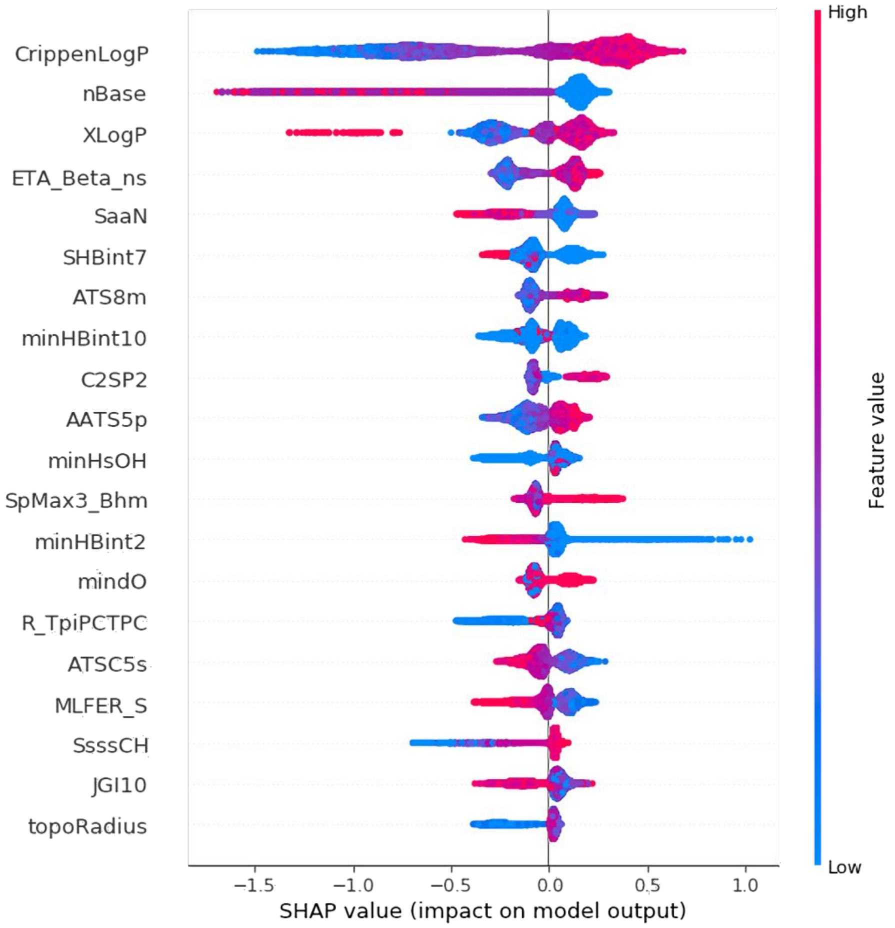

CrippenLogP
nBase
XLogP
ETA_Beta_ns
SaaN
SHBint7
ATS8m
minHBint10
C2SP2
AATS5p
minHsOH
SpMax3_Bhm
minHBint2
mindO
R_TpiPCTPC
ATSC5s
MLFER_S
SsssCH
JGI10
topoRadius
−1.5
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Low
Feature value

Figure 9

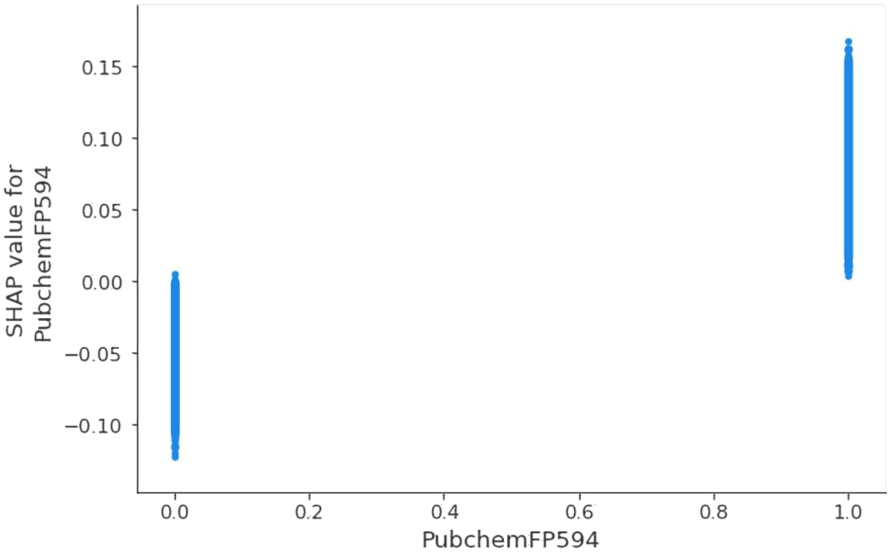

0.15
0.10
0.05
0.00
−0.05
−0.10
SHAP value for PubchemFP594
0.0
0.2
0.4
0.6
0.8
1.0
PubchemFP594

Figure 10

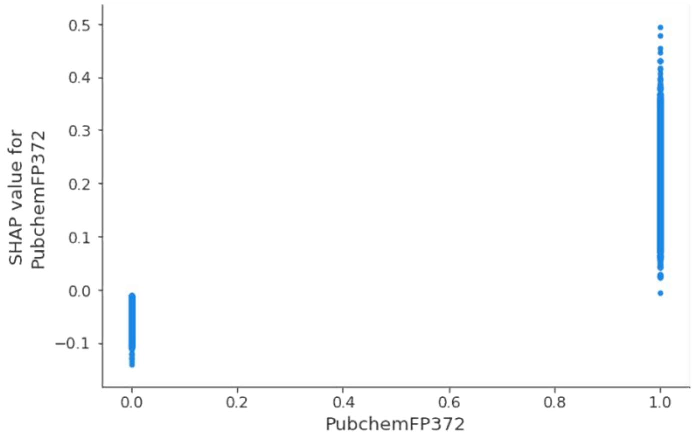

0.5
0.4
0.3
0.2
0.1
0.0
−0.1
SHAP value for PubchemFP372
0.0
0.2
0.4
0.6
0.8
1.0
PubchemFP372

Figure 11

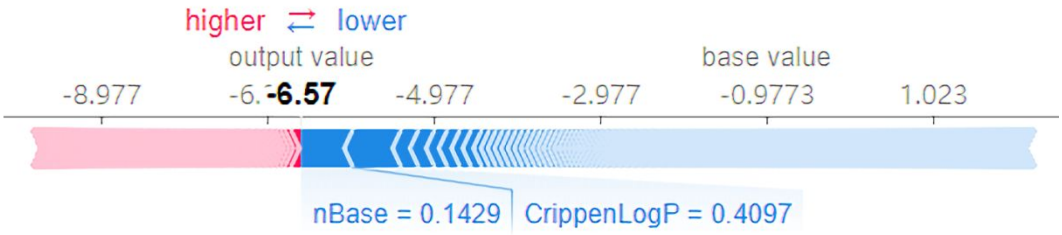

higher ⇄ lower
output value
base value
-8.977
-6.
-6.57
-4.977
-2.977
-0.9773
1.023
nBase = 0.1429
CrippenLogP = 0.4097

Figure 12