JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

# Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers

Feixiong Cheng,[†] Yue Yu,[†] Jie Shen,[†] Lei Yang,[§] Weihua Li,[*,†] Guixia Liu,[†] Philip W. Lee,[†,‡] and Yun Tang[*,†]

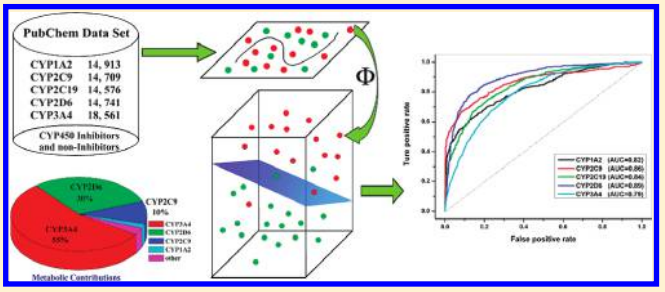
[†]Department of Pharmaceutical Sciences, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

[‡]Graduate School of Agriculture, Kyoto University, Kitashirakawa Oiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan

[§]School of Information Science & Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

Supporting Information

**ABSTRACT:** Adverse side effects of drug−drug interactions induced by human cytochrome P450 (CYP) inhibition is an important consideration, especially, during the research phase of drug discovery. It is highly desirable to develop computational models that can predict the inhibitive effect of a compound against a specific CYP isoform. In this study, inhibitor predicting models were developed for five major CYP isoforms, namely 1A2, 2C9, 2C19, 2D6, and 3A4, using a combined classifier algorithm on a large data set containing more than 24,700 unique compounds, extracted from PubChem. The combined classifiers algorithm is an ensemble of different independent machine learning classifiers including support vector machine, C4.5 decision tree, *k*-nearest neighbor, and naïve Bayes, fused by a back-propagation artificial neural network (BP-ANN). All developed models were validated by 5-fold cross-validation and a diverse validation set composed of about 9000 diverse unique compounds. The range of the area under the receiver operating characteristic curve (AUC) for the validation sets was 0.764 to 0.815 for CYP1A2, 0.837 to 0.861 for CYP2C9, 0.793 to 0.842 for CYP2C19, 0.839 to 0.886 for CYP2D6, and 0.754 to 0.790 for CYP3A4, respectively, using the new developed combined classifiers. The overall performance of the combined classifiers fused by BP-ANN was superior to that of three classic fusion techniques (Mean, Maximum, and Multiply). The chemical spaces of data sets were explored by multidimensional scaling plots, and the use of applicability domain improved the prediction accuracies of models. In addition, some representative substructure fragments differentiating CYP inhibitors and noninhibitors were characterized by the substructure fragment analysis. These classification models are applicable for virtual screening of the five major CYP isoforms inhibitors or can be used as simple filters of potential chemicals in drug discovery.




## INTRODUCTION

Nowadays, coadministration of two or more drugs is a common way for a patient during disease treatment.[1] During drug treatment, the patient has an increased risk of exposing to potential adverse drug−drug interactions (DDIs).[2] There are about two million serious adverse drug reactions reported per year in the United States, approximately 26% of which can be attributed to avoidable DDIs.[3]

The human cytochromes P450 (CYPs), a superfamily of heme-containing enzymes with about 57 isoforms, catalyze the metabolism of a variety of endogenous and xenobiotic compounds. The CYP enzymes, particularly isoforms 1A2, 2C9, 2C19, 2D6, and 3A4, are responsible for about 90% oxidative metabolic reactions.[4] Inhibition of CYP enzymes will lead to inductive or inhibitory failure of drug metabolism.[1,5,6] In the last several decades, several commercial drugs were withdrawn from the market due to adverse CYP enzymes DDIs, such as Seldane, Posicor, Hismanal, Propulsid, Lotronex, Baycol, and Seraone.[1,7] US FDA and the Pharmaceutical Research and Manufacturers of America published the guidelines for the pharmaceutical industry, urging that *in vitro* metabolic studies should be conducted early in drug discovery to determine the metabolic inhibitive properties of new chemical entities (NCEs), particularly for members of the CYP superfamily.[8]

Many researchers had attempted to develop *in vitro* screening techniques to identify potential CYP inhibitors in drug discovery.[9−11] Recently, Auld's group using an *in vitro* bioluminescent assay of quantitative high-throughput screening (qHTS) determined the $AC_{50}$ values (the compound concentration leads to 50% of the activity of an inhibition control) of more than 17,000 compounds against five recombinant CYP isoforms (1A2, 2C9, 2C19, 2D6, and 3A4).[11] These results provided a large and

**Received:** January 21, 2011
**Published:** April 14, 2011

**Table 1. Detailed Statistical Description of 24,732 Unique Compounds in PubChem Data Sets I and II for Five Major CYP Isoforms Based on the Multilabel Classification Strategy**

| data sets | CYP isoforms | number of inhibitors | number of noninhibitors | total | Tanimoto index |
|---|---|---|---|---|---|
| PubChem Data Set I[a] | 1A2 | 5663 | 6436 | 12,099 | 0.206 |
| | 2C9 | 4369 | 7761 | 12,130 | 0.208 |
| | 2C19 | 5322 | 6563 | 11,885 | 0.212 |
| | 2D6 | 2516 | 9365 | 11,881 | 0.209 |
| | 3A4 | 4637 | 6899 | 11,536 | 0.200 |
| PubChem Data Set II[b] | 1A2 | 1752 | 1052 | 2804 | 0.213 |
| | 2C9 | 609 | 1970 | 2579 | 0.220 |
| | 2C19 | 719 | 1972 | 2691 | 0.207 |
| | 2D6 | 544 | 2316 | 2860 | 0.212 |
| | 3A4 | 2070 | 4955 | 7025 | 0.113 |

[a] PubChem Data Set I was collected from the National Center for Biotechnology Information (NCBI) PubChem database AID 1851 (http://pubchem.ncbi.nlm.nih.gov/). Inhibitors: $AC_{50} \leq 10\ \mu M$; Noninhibitors: $AC_{50} > 57\ \mu M$; $AC_{50}$: Compound concentration leads to 50% of the activity of an inhibition control.[11,21] [b] PubChem Data Set II: CYP1A2 from PubChem AID 410, CYP2C9 from PubChem AID 883, CYP2C19 from PubChem AID 899, CYP2D6 from PubChem AID 891, and CYP3A4 from PubChem AID 884 and 885. Inhibitors: PubChem Activity score equal 40 to 100; Noninhibitors: PubChem Activity score equal 0.[21]

diverse bioassay database for the development of *in silico* predictive models for CYP inhibitors.[12]

In the past several years, several *in silico* models to predict CYP inhibitors had been reported. Poongavanam et al. applied the support vector machine (SVM), random forest, kappa nearest neighbor (*k*-NN), and decision tree methods to develop models to classify CYP1A2 inhibitors and noninhibitors with an overall predictive accuracy of about 75% for the internal test set.[13] Jensen et al. also reported classification models for CYP2D6 and 3A4 inhibitors using Gaussian kernel weighted k-nearest neighbor methods, but the sensitivity of test sets was only 59 and 65% for the two isoforms, respectively.[14] All reported models were limited to one to three CYP isoforms, using rather limited compound sets, and they are not very useful informative and not useful in the drug discovery research.[15−20] Developing higher prediction accuracies P450 inhibition predicting models using the diverse data set and new modeling methodologies such as combined classifiers are very urgent.

In this paper, we reported a new method to classify CYP inhibitors and noninhibitors by combining different single machine learning classifiers fused in a back-propagation artificial neural network (BP-ANN) algorithm. Inhibitors and noninhibitors classification models for five major CYP isoforms, namely 1A2, 2C9, 2C19, 2D6, and 3A4, based on a large data set of over 24,700 unique compounds with known CYP450 inhibition were developed.[11,21] The overall performance of the combined classifiers fused by BP-ANN was superior to that of three classic fusion techniques (Mean, Maximum, and Multiply). High predictive accuracies of the combined classifiers models were also obtained for a diverse validation set. The use of applicability domain improved the prediction accuracy of models. Moreover, some representative substructure fragments common to CYP inhibitors and noninhibitors were also identified via substructure fragment analysis.

## MATERIALS AND METHODS

**Data Set Collection.** The initial PubChem database (PubChem AID: 1851) in SMILES format was provided by Dr. Auld.[11,21] It contains 17,143 diverse compounds which were measured by a standard protocol under the same experimental conditions. Entries containing inorganic compounds, noncovalent complexes, and mixtures were excluded. Salts were converted to the corresponding acids or bases; water molecules were removed from hydrates. From the original list, 15,744 unique compounds (Designated as PubChem Data Set I) were extracted as training set, and the detailed statistical description of this data set is presented in Table 1.

According to the cutoff criterion of Auld's reports and PubChem BioAssay database,[11,21] a compound was assigned as a CYP inhibitor if the $AC_{50}$ (the compound concentration leads to 50% of the activity of an inhibition control) value was $\leq 10\ \mu M$, and it was considered as a noninhibitor if $AC_{50}$ was $>57\ \mu M$.[11,21] Compounds with intermediate $AC_{50}$ value (10 to 57 $\mu M$) were classified as inconclusive compounds and were excluded in this study to avoid uncertainty during models development. The PubChem ID number, SMILES, and inhibitor and noninhibitor labels of 15,744 unique compounds against CYP1A2, 2C9, 2C19, 2D6, and 3A4 are available online: http://www.lmmd.org/database.html.

In addition to the training set, a diverse validation set was also collected from the PubChem BioAssay database[21] for the purpose of verifying the robustness of prediction models. The same pretreatment of compound structure as described for the training set was applied. Based on the classification criterion of P450 inhibitor and noninhibitor in PubChem BioAssay database,[21] a compound was regarded as a CYP inhibitor if it has the PubChem activity score between 40 and 100, and as a noninhibitor if it has PubChem activity score equal to 0. Compounds with the intermediate PubChem activity score (1 to 39) were considered as inconclusive compounds and were excluded. Duplicated compounds with PubChem Data Set I were also excluded. The diverse validation set containing 8988 unique compounds was obtained. The detailed statistical description for the entire validation set (Designated as PubChem Data Set II) was presented in Table 1. The PubChem ID number, SMILES, and inhibitor and noninhibitor labels of 8988 unique compounds against five major CYP isoforms are available online: http://www.lmmd.org/database.html.

In the PubChem BioAssay database,[21] the PubChem activity score is assigned according to the fitted $IC_{50}$ value, with respect

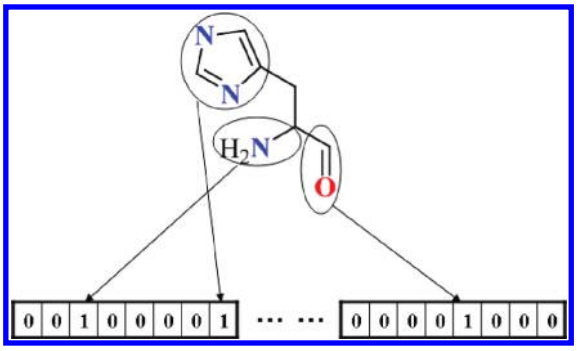



**Figure 1.** The definition of a molecular substructure pattern fingerprint. The predefined dictionary contained a SMARTS list of substructure patterns. For a SMARTS pattern, if a specified substructure is presented in the given molecule, the corresponding bit is set to "1"; conversely, it is set to "0".

to completeness of dose−response curve and efficacy of inhibition (maximum inhibition response).[15] For example, if the $IC_{50}$ of a compound is less than 40 $\mu$M, the PubChem activity score of this compound was set as >40. In order to keep the consistency of inhibitors and noninhibitors cutoff criterion between PubChem Data Sets I and II, we checked the duplicated compounds of five isoforms between them. As shown in Table S1 of the Supporting Information, 99.6% compounds for CYP1A2, 99.4% compounds for CYP2C9, 99.0% compounds for CYP2C19, 99.4% compounds for CYP2D6, and 88.0% compounds for CYP3A4 have the consistent inhibitors and noninhibitors labels among these different experimental data sets. It suggests that the classification threshold value used here is reasonable. The PubChem ID number, SMILES, and inhibitor and noninhibitor labels of all duplicated compounds (designated as inconclusive compounds) against five isoforms are available online: http://www.lmmd.org/database.html.

**Data Description and Substructure Fragment Analysis.** The substructure pattern recognition method that we recently developed in our laboratory[22] was used for molecule description. As shown in Figure 1, each molecule was described as a binary string of structural keys. The predefined dictionary contained a SMARTS list of substructure patterns. For a SMARTS pattern, if a specified substructure is presented in the given molecule, the corresponding bit is set to "1"; conversely, it is set to "0". Two substructure dictionaries of MACCS keys and FP4 fingerprints, freely available from OpenBabel (version2.2.3, http://openbabel.org/, accessed Jan. 18, 2010),[23] were used. MACCS keys used a dictionary of MDL Public Keys,[24] which contained 166 most common substructure patterns. The dictionary of FP4 fingerprints contained 307 substructure patterns.

An advantage of fingerprints is that they can be easily translated into two-dimensional fragments. The representative substructure fragments were explored using information gain method[22] and substructure fragment analysis.[14] The frequency of a fragment in a CYP inhibition class was defined as follows

$$\text{Frequency of a fragment} = \frac{(N_{\text{fragment\_class}} \times N_{\text{total}})}{(N_{\text{fragment\_total}} \times N_{\text{class}})} \quad (1)$$

where $N_{\text{fragment\_class}}$ is the number of compounds containing the fragment in a CYP inhibition class, $N_{\text{total}}$ is the total number of compounds, $N_{\text{fragment\_total}}$ is the total number of compounds containing the fragment, and $N_{\text{class}}$ is the number of compounds in the CYP inhibition class.

**Model Building.** The entire computational workflow used in this study is presented in Figure 2, and the architecture of the classifiers combination algorithm is given in Figure 3. The support vector machine algorithm was performed by the LIBSVM2.9 package.[25] C4.5 decision tree, $k$ nearest neighbor, and naïve Bayes calculations were performed using the Orange canvas (Version 2.0b, free available on the Web site http://www.ailab.si/orange/). BP-ANN was performed using in-house MATLAB scripts in accordance with the literature.[26]

*Single Classifier Model*

*Support Vector Machine (SVM).* The SVM algorithm, originally developed by Vapnik[27] for pattern recognition, aims at minimizing the structural risk under the frame of Vapnik-Chervonenkis (VC) theory. Each molecule is represented using an eigenvector $\mathbf{t}$, and the selected patterns $t_1$, $t_2$, ..., $t_n$ are the components of $\mathbf{t}$. The category label $y$ was added in SVM training. The $i^{\text{th}}$ molecule in the data set is defined as $\mathbf{M}_i = (t_i, y_i)$, where $y_i = 1$ for the "inhibitor" category and $y_i = -1$ for the "noninhibitor" category. SVM gives a decision function (classifier)

$$f(\mathbf{t}) = \text{sgn}\left(\frac{1}{2}\sum_{i=1}^{n} \alpha_i K(\mathbf{t}_i, \mathbf{t}) + b\right) \quad (2)$$

where $\alpha_i$ is the coefficient to be trained, and $K$ is a kernel function. Parameter $\alpha_i$ is trained through maximizing the Lagrangian expression given below

$$\begin{aligned} &\underset{\alpha_i}{\text{maximize}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K(t_i, t) \\ &\text{subject to}: \sum_{y_i=1} y_i a_i = 0,\ 0 \le a_i \le C \end{aligned} \quad (3)$$

The commonly used kernel function Gaussian radial basis function (RBF) kernel was used. The different kernel parameter $\gamma$ and penalty parameter $C$ were tuned based on the training set using grid search strategy with a 5-fold cross-validation to obtain a SVM model with optimal performance.

*C4.5 Decision Tree (C4.5 DT).* C4.5 builds decision trees from a set of training data in the same way as Iterative Dichotomiser 3 (ID3). The elements of the tree generated by ID3 and C4.5 DT are either leafs or decision nodes. The leaf shows a class, and the decision node specifies the test to be implemented on an attribute value, with one branch and subtree for each possible result of the test. The detailed descriptions of C4.5 DT can be found in the original literature.[28]

*k-Nearest Neighbors (k-NN).* $k$-NN classifies objects based on the closest training examples in the feature space. The nearness is measured by a hamming distance matrix, and the standard protocol of 3-NN is implemented simply as follows: (1) to calculate the distances between an unknown object ($y$) and all the objects in the training set; (2) to select 3 objects which are most similar to object $y$ from the training set according to the calculated distances; and (3) to classify object $y$ into the group to which the majority of the 3 objects belongs.

*Naïve Bayes (NB).* Bayesian classification is a statistical method that allows the user to categorize instances in a data set based on the equal and independent contributions of their attributes.[29] For Naïve Bayes classifier, it generates the posterior probabilities which were given out directly based on the core

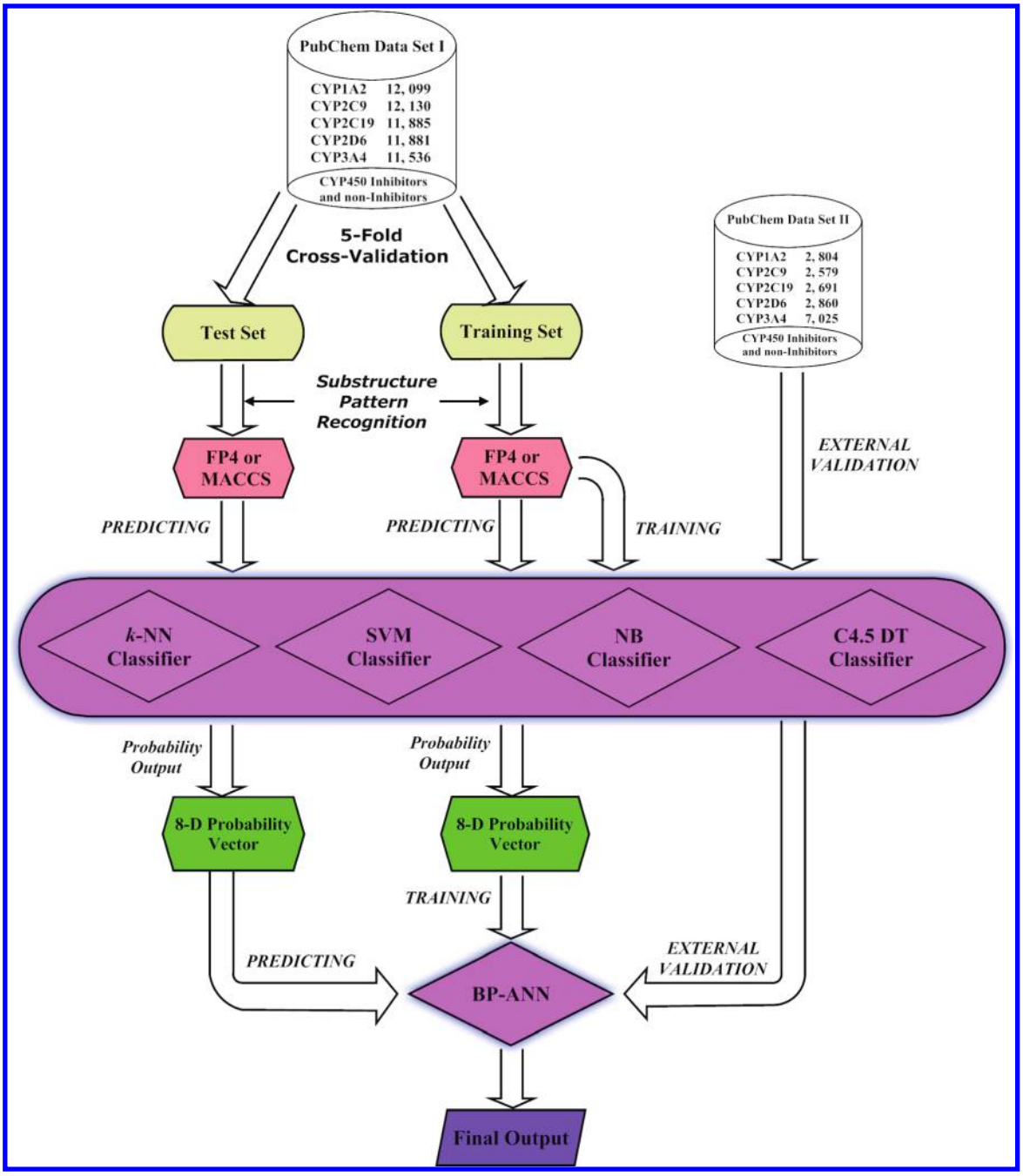



**Figure 2.** The whole work-flow for combined classifiers models building and validation as applied for CYP1A2, 2C9, 2C19, 2D6, and 3A4. SVM (Support Vector Machine), C4.5DT (C4.5 Decision Tree), *k*-NN (*k*-Nearest Neighbor), NB (Naïve Bayes), and Back-Propagation Artificial Neural Network (BP-ANN).

function of eq 4

$$P(C_i|X) = \frac{p_{C_i}p(X|C_i)}{\sum_j p_{C_j}p(X|j)} \quad (4)$$

*Combined Classifiers (CC) Model*

*Back-Propagation Artificial Neural Networks (BP-ANN).* A three-layer BP-ANN was used. The number of nodes in the input layer was decided by the combination probability outputs of several independent single classifiers. We set two neurons in the hidden layer and one neuron in the output layer. Besides, a bias of value +1 was set in the input layer and the hidden layer, respectively. The topological structure of our network is not a full-connected one. The input nodes which represent probability of the instance belonging to the +1 class only connect to the +1 neuron in the hidden layer, whereas the other input nodes connect to the −1 neuron. The hyperbolic tangent function was used as the activation function in BP-ANN

$$f(x) = a\tanh(bx) \quad (5)$$


dx.doi.org/10.1021/ci200028n |J. Chem. Inf. Model. 2011, 51, 996–1011

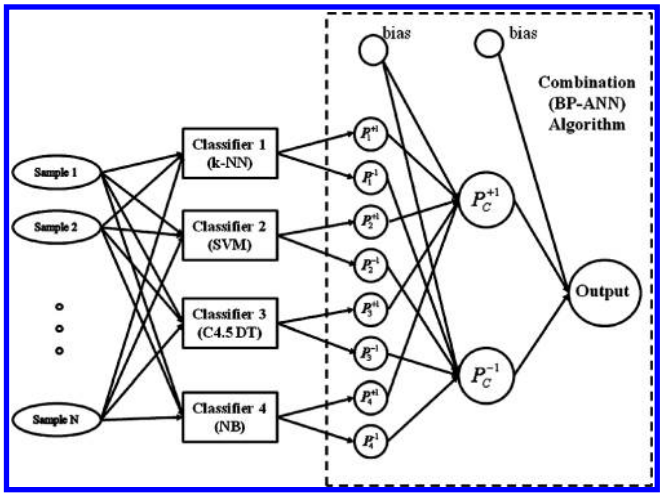



**Figure 3.** The architecture of the combined classifiers. Combined Classifiers takes a set of probability output for inhibitors (+1) or noninhibitors (−1) by single classifier (SVM or C4.5DT or *k*-NN or NB) and produces combination probability output for each class. SVM (Support Vector Machine), C4.5DT (C4.5 Decision Tree), *k*-NN (*k*-Nearest Neighbor), NB (Naïve Bayes), and Back-Propagation Artificial Neural Network (BP-ANN).

where $a$ was set to 1.72, and $b$ was set to 2/3 based on the experience values.[26]

Momentum parameter $\alpha$ was applied to accelerate the whole network's convergence, so the weight update strategy is changed to

$$\Delta w_{ji}(n) = \alpha \Delta w_{ji}(n-1) + \eta \delta_j(n) y_i(n) \quad (6)$$

where $\Delta w_{ji}(n-1)$ is the updated value of synergic weight $ij$ in last epoch, $\eta$ is the learning rate, and $\delta_j(n)$ is the local gradient of neuron $j$. $a = 0.5$ and $\eta = 0.02$ were used in this study, and the batch-learning method[26] was used in the training process. In addition, optimal synergic weights were decided by the training data. When the change rate of average error rate between two consecutive epochs was smaller than 0.001, the new connection weights were considered to be optimal. Using this stopping criterion is easy to get our network stuck in local optimal points. Thus, we repeated the training process for 10,000 times to build 10,000 independent ANN models, each time with a new randomly initiated connection weight. To maintain the maximum performance metric of overall predictive accuracy, we used a helper test set which is held out subsets (10%) from the test set to select the best ANN model. Then, the remaining test set (90%) which was not used during the model training and selection was used to evaluate the predictive power of the combined classifiers. To avoid the bias from dividing the data set, all results are obtained by averaging 10 times random dividing process.

*Probability Outputs.* Classical machine learning algorithms try to produce estimated target values (such as +1 or −1) instead of predictive probability ranges, which is easy to omit important detailed information of each classifier. In order to utilize more information of each single classifier, different strategies were employed to get probability output of them.

*Naïve Bayes (NB).* For the Naïve Bayes classifier, the posterior probabilities can be given out based on the eq 4.

*Support Vector Machine (SVM).* Lin and Weng have developed a Bayesian approach for SVM to generate probability estimation for each class in binary classification problems.[30] In the following we briefly described how to extend SVM for probability estimation. For SVM probability estimation, given $k$ classes of data, for any $x$, the goal is to estimate

$$p_i = p(y = i|x), i = 1, \ldots, k \quad (7)$$

First pairwise class probabilities are estimated

$$r_{ij} \approx p(y = i|y = i \text{ or } j, x) \quad (8)$$

$r_{ij}$ can be calculated by the following equation

$$r_{ij} \approx \frac{1}{1 + e^{A\hat{f} + B}} \quad (9)$$

where $A$ and $B$ are estimated by minimizing the negative log-likelihood function using the known training data and their decision values $\hat{f}$. Labels and decision values are required to be independent. Therefore a 5-fold cross-validation was conducted to obtain the decision values. Once we have $r_{ij}$, we can obtain $p_i$ by solving the following optimization problem[30]

$$\min_p \frac{1}{2} \sum_{i=1}^{k} \sum_{j:j \neq i} (r_{ji} p_i - r_{ij} p_j)^2 \text{ subject to } \sum_{i=1}^{k} p_i = 1, p_i \geq 0, \ \forall\, i \quad (10)$$

A detailed description about solving strategy can be found in Wu's work.[30]

*C4.5 Decision Tree (C4.5 DT).* For C4.5 decision tree, the Confusion Matrix and Posterior Probability Matrix were used to export the probability output.

*k-Nearest Neighbors (k-NN).* For *k*-NN algorithm, we simply defined the probability of an instance for any class as the proportion of instances which belongs to one class in all its *k*-NN. For the standard protocol of 3-NN, if two known nearest objects belong to class $y$, the probability of this object belonging to $y$ is 2/3.

*Combined Classifiers Formulation.* As above-mentioned, the whole combination process was performed on the probability output of each independent SC and BP-ANN. The decision network consists of two layers of units: input layer and output layer. First, the training data was used to train the SVM, C4.5DT, *k*-NN, and NB classifier separately. Then the training data were predicted with four SC models to obtain the probability output ($P_i^{+1}$ and $P_i^{-1}$ $i$ = 1,2,3,4). These probability outputs were used as new descriptors to develop BP-ANN models which generate the final combination decision probability ($P_C^{+1}$ *and* $P_C^{-1}$). During the prediction process, the test set was first predicted by each SC model and then put into the BP-ANN model developed by the training set. Finally, the estimated results were obtained from the developed CC models. The architecture of combined classifiers strategy is shown in Figure 3.

In addition, we also investigated three classic classifiers fusion techniques: Mean, Maximum, and Multiply.

i Mean

$$\mu_j(x) = \frac{1}{L} \sum_{i=1}^{L} d_{i,j}(x) \quad (11)$$

ii Maximum

$$\mu_j(x) = Max_i\{d_{i,j}(x)\} \quad (12)$$


dx.doi.org/10.1021/ci200028n |*J. Chem. Inf. Model.* 2011, 51, 996–1011

iii Multiply

$$\mu_j(x) = \prod_{i=1}^{L} d_{i,j}(x) \quad (13)$$

where $\mu_j(x)$ represents the final combination decision probability ($P_C^{+1}$ *and* $P_C^{-1}$) of sample $x$, and $L$ represents the independent single classifier ($i$ = 1,2,3,...,$L$.).

**Chemical Space Analysis and Models Applicability Domains.** The chemical diversity and distribution in the chemical space were explored by chemical descriptors and Tanimoto similarity. First, the drug-likeness properties of the compound, illustrated by the Lipinski's Rule-of-Five,[31] were calculated by Discovery Studio 2.1.[32] After that, the Tanimoto similarity analysis was performed.[33] Tanimoto similarity index is a classic method to explore the diversity of compounds within a chemical data set. Smaller the Tanimoto similarity index means that compounds within the data set have good diversity.

Finally, a visualization of the chemical space was examined based on the principal component analysis (PCA). The visual chemical space map was generated for the PubChem Data Sets I and II by projecting the MACCS keys on principal components (PCs). The PCs for 24,732 unique compounds were calculated, which were used to produce multidimensional scaling (MDS)[14,34] plots.

Defining model applicability domains (AD) is an active area of modern quantitative structure activity/property relationship (QSAR/QSPR) research. AD was estimated using Ambit Discovery v0.04, which can be downloaded from http://ambit.acad.bg/downloads/AmbitDiscovery/. The strategy of Ambit Discovery is to estimate the domain for a given model using molecule description data (described in the Section of **Data Set Description**) in the model training set and then to use this information to determine whether chemicals in a validation set lie within the AD of the training set. Herein, MACCS keys-based analysis was performed with PCA Data preprocessing to eliminate collinearities among model descriptors, range-based method and the 100% of training set points to determine AD. For the ranges approach, chemicals were labeled out of domain (OD) if at least one fragment count was out of range.[35]

**Models Validation.** The $k$-fold cross-validation techniques and a diverse validation set were used to evaluate all models. In a 5-fold cross-validation, the entire data set was equally divided into five cross-validation splits. Within each step of cross-validation, the model was trained on a set of four cross-validation splits together. The fifth subsample set was used as an internal validation set (test set). Moreover, a diverse validation set (PubChem Data Set II) was used to validate the generalization abilities of models.

All developed models were evaluated based on the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP represents the number of inhibitors predicted as inhibitors; TN is the number of noninhibitors predicted as noninhibitors; FP stands for the number of noninhibitors predicted as inhibitors; and FN represents the number of inhibitors predicted as noninhibitors. Furthermore, the sensitivity (SE = TP/(TP + FN)), which is the prediction accuracy of inhibitors, and the specificity (SP = TN/(TN + FP)), which is the prediction accuracy of noninhibitors, were calculated. The overall predictive accuracy ($Q$) and the Matthews correlation coefficient ($C$) were calculated using eqs 14 and 15, respectively

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (14)$$

$$C = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \quad (15)$$

The $C$ falls in the range of $-1 \le C \le +1$. A value of $C = 1$ indicates perfect agreement between predicted and experimental classes for each binary classifier, whereas $C = -1$ indicates the worst possible prediction.

In addition, the receiver operating characteristic (ROC) curve was plotted. The ROC curve was used to graphically present the model behavior in a visual way. It shows the separation ability of a binary classifier by iteratively setting the possible classifier threshold.[36]

## RESULTS

**Data Set Analysis.** PubChem Data Set I consisted of 15,744 unique compounds. Based on the multilabel classification strategy[37] of "one-versus-the-rest", the entire data sets were divided into 5663 inhibitors and 6436 noninhibitors for CYP1A2, 4369 inhibitors and 7761 noninhibitors for CYP2C9, 5322 inhibitors and 6563 noninhibitors for CYP2C19, 2516 inhibitors and 9365 noninhibitors for CYP2D6, and 4637 inhibitors and 6899 noninhibitors for CYP3A4 (Table 1). The entire data set was diverse enough as the Tanimoto index was 0.206, 0.208, 0.212, 0.209, and 0.200 for the CYP1A2, 2C9, 2C19, 2D6, and 3A4 data sets, respectively. The drug-likeness of 15,744 unique compounds was analyzed by calculating the descriptors of Lipinski's Rule-of-Five.[31] 885 compounds (5.6%) have molecular weight greater than 500; 191 compounds (1.2%) have hydrogen bond acceptors more than 10; 126 compounds (0.8%) have hydrogen bond donors more than 5; and 1452 compounds have the logP values greater than 5. In total, 13,551 compounds (86.1%) conform to Lipinski's Rule-of-Five, which indicates that the entire data set has good drug-likeness.

The same analysis was conducted for the validation set (PubChem Data Set II). The Tanimoto index for the validation set was 0.213, 0.220, 0.207, 0.212, and 0.113 for CYP1A2, 2C9, 2C19, 2D6, and 3A4, respectively. In addition, the chemical spaces of inhibitors and noninhibitors in PubChem Data Sets I and II were explored by the PCA and MDS plots techniques. As shown in Figure 4, there were diverse chemical space distributions for PubChem Data Sets I and II in the MDS plots. The total variance explained by PC1 and PC2 was 12.69 and 8.86%, respectively.

**Construction of Binary Classification Models by Single Classifier (SC).** First, binary classification models of CYP1A2, 2C9, 2C19, 2D6, and 3A4 were developed using four independent SC (SVM, C4.5DT, $k$-NN, and NB). The performance of the 5-fold cross-validation for SC models was given in Table 2. CYP1A2 models were built by the approximate balance data set (5663 inhibitors vs 6436 noninhibitor). The high performance of SE 78.6%, SP 84.7%, $Q$ 81.8%, and $C$ 0.635 was obtained for the CYP1A2 model using the SVM algorithm with MACCS keys. And the high performance of $Q$ 81.7% and $C$ 0.633 was also obtained for the CYP1A2 model using the SVM algorithm with FP4 fingerprints. For CYP2D6 models, a high performance of $Q$ 84.3 and 83.8% was obtained using the SVM algorithm with MACCS keys and FP4 fingerprints, respectively. And the

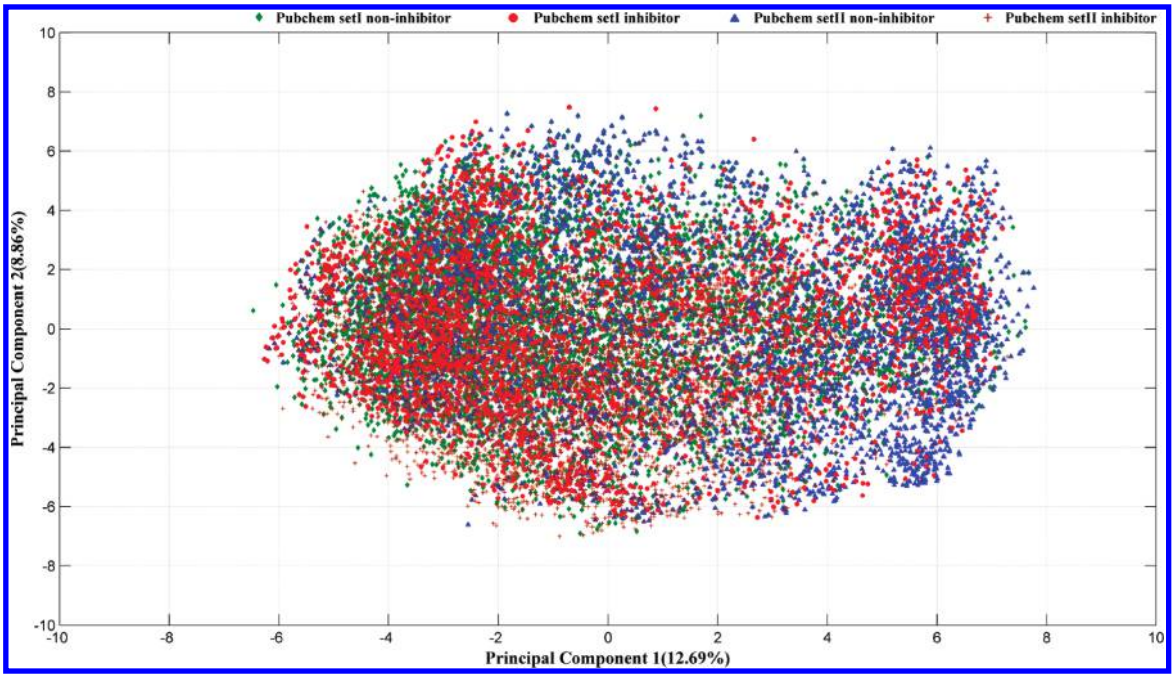



**Figure 4.** The multidimensional scaling (MDS) plot for the PubChem Data Set I (7255 inhibitors as red circles and 8489 noninhibitors as green diamonds) and PubChem Data Set II (3074 inhibitors as orange cross and 5914 noninhibitors as blue triangles) as described by the principal component analysis (PCA). The total variance explained PC1 was 12.69% and PC2 was 8.86% for entire data set. Each dot represents one of the 24,732 unique compounds.

**Table 2. Performance of the 5-Fold Cross-Validation for Five Major CYP Isoforms Using the Single Classifier[a]**

| | | MACCS | | | | FP4 | | | |
|---|---|---|---|---|---|---|---|---|---|
| CYP isoforms | methods | SE (%) | SP (%) | Q (%) | C | SE (%) | SP (%) | Q (%) | C |
| 1A2 | SVM | 78.6 | 84.7 | 81.8 | 0.635 | 79.4 | 83.8 | 81.7 | 0.633 |
| | C4.5 DT | 73.0 | 77.1 | 75.1 | 0.501 | 74.2 | 77.9 | 76.1 | 0.521 |
| | *k*-NN | 80.7 | 74.8 | 77.5 | 0.554 | 84.7 | 64.7 | 74.1 | 0.501 |
| | NB | 75.1 | 75.1 | 75.1 | 0.501 | 76.9 | 70.4 | 73.4 | 0.472 |
| 2C9 | SVM | 61.6 | 86.9 | 77.8 | 0.505 | 59.0 | 86.7 | 76.8 | 0.480 |
| | C4.5 DT | 56.1 | 78.9 | 70.7 | 0.356 | 59.5 | 79.2 | 72.1 | 0.390 |
| | *k*-NN | 65.5 | 77.8 | 73.4 | 0.429 | 72.2 | 69.0 | 70.2 | 0.397 |
| | NB | 61.5 | 70.7 | 67.4 | 0.315 | 68.7 | 68.5 | 68.6 | 0.359 |
| 2C19 | SVM | 75.4 | 81.2 | 78.6 | 0.567 | 74.8 | 80.2 | 77.7 | 0.551 |
| | C4.5 DT | 66.2 | 76.1 | 71.7 | 0.425 | 68.0 | 75.8 | 72.3 | 0.440 |
| | *k*-NN | 63.0 | 80.8 | 72.8 | 0.447 | 51.6 | 86.0 | 70.6 | 0.405 |
| | NB | 67.5 | 68.9 | 68.3 | 0.363 | 75.6 | 66.6 | 70.6 | 0.420 |
| 2D6 | SVM | 39.0 | 96.3 | 84.3 | 0.457 | 36.9 | 96.4 | 83.8 | 0.441 |
| | C4.5 DT | 47.9 | 87.5 | 79.1 | 0.361 | 46.2 | 88.3 | 79.4 | 0.359 |
| | *k*-NN | 51.0 | 88.7 | 80.7 | 0.408 | 58.0 | 82.7 | 77.5 | 0.379 |
| | NB | 53.9 | 78.6 | 73.4 | 0.295 | 48.5 | 86.1 | 78.2 | 0.346 |
| 3A4 | SVM | 65.2 | 85.8 | 77.5 | 0.525 | 62.8 | 85.3 | 76.2 | 0.497 |
| | C4.5 DT | 63.1 | 75.5 | 70.5 | 0.386 | 63.3 | 77.1 | 71.6 | 0.406 |
| | *k*-NN | 69.4 | 74.8 | 72.6 | 0.437 | 74.0 | 64.6 | 68.3 | 0.378 |
| | NB | 69.3 | 60.0 | 63.7 | 0.287 | 67.9 | 64.4 | 65.8 | 0.317 |

[a] Performances of classification model were given as follows: SE (sensitivity), SP (specificity), *Q* (overall predictive accuracy), and *C* (Matthews correlation coefficient). SVM (support vector machine), C4.5 DT (C4.5 decision tree), *k*-NN (*k*-nearest neighbor), and NB (Naïve Bayes).

similarly high *Q* values of 77.8, 78.6, and 77.5% for CYP2C9, 2C19, and 3A4, respectively, were obtained using the SVM algorithm with MACCS keys, which were evaluated by 5-fold cross-validation.

Compared with SVM performance, C4.5 DT, *k*-NN, and NB performed worse. For example, the *Q* value of CYP1A2 models was 81.8% for SVM with MACCS keys. But the *Q* value was only 75.1, 77.5, and 75.1% for C4.5 DT, *k*-NN, and NB, respectively,

Table 3. Performance of the 5-Fold Cross-Validation for Five Major CYP Isoforms Using the Combined Classifiers (CC)[a]

| | | MACCS | | | | FP4 | | | |
|---|---|---|---|---|---|---|---|---|---|
| CYP isoforms | methods | SE (%) | SP (%) | Q (%) | C | SE (%) | SP (%) | Q (%) | C |
| 1A2 | CC-I | 79.5 | 82.4 | 81.1 | 0.620 | 78.5 | 81.8 | 80.2 | 0.603 |
| | CC-II | 80.0 | 82.5 | 81.3 | 0.626 | 77.1 | 83.9 | 80.7 | 0.612 |
| | CC-III | 79.2 | 83.1 | 81.3 | 0.625 | 77.2 | 83.6 | 80.6 | 0.610 |
| | CC-IV | 78.8 | 83.0 | 81.1 | 0.620 | 77.1 | 83.7 | 80.6 | 0.610 |
| | CC-V | 77.9 | 78.3 | 78.1 | 0.561 | 75.5 | 78.9 | 77.3 | 0.544 |
| | Mean | 79.1 | 80.9 | 80.0 | 0.600 | 81.4 | 77.1 | 79.1 | 0.584 |
| | Maximum | 79.3 | 77.9 | 78.5 | 0.571 | 82.5 | 72.6 | 77.2 | 0.551 |
| | Multiply | 79.6 | 79.0 | 79.3 | 0.585 | 83.0 | 74.1 | 78.3 | 0.571 |
| 2C9 | CC-I | 60.9 | 85.4 | 76.6 | 0.479 | 59.9 | 84.6 | 75.7 | 0.460 |
| | CC-II | 63.3 | 85.1 | 77.3 | 0.498 | 61.1 | 84.4 | 76.0 | 0.468 |
| | CC-III | 63.9 | 84.6 | 77.2 | 0.496 | 61.0 | 84.5 | 76.1 | 0.469 |
| | CC-IV | 63.2 | 84.8 | 77.0 | 0.493 | 60.8 | 84.8 | 76.1 | 0.470 |
| | CC-V | 59.0 | 83.5 | 74.7 | 0.438 | 56.8 | 83.1 | 73.7 | 0.414 |
| | Mean | 64.2 | 82.4 | 75.8 | 0.471 | 68.0 | 79.5 | 75.4 | 0.471 |
| | Maximum | 64.6 | 79.5 | 74.2 | 0.441 | 69.1 | 75.9 | 73.5 | 0.441 |
| | Multiply | 65.1 | 81.3 | 75.4 | 0.466 | 69.6 | 77.0 | 74.4 | 0.457 |
| 2C19 | CC-I | 72.7 | 80.4 | 77.0 | 0.533 | 73.6 | 77.9 | 75.9 | 0.514 |
| | CC-II | 72.9 | 82.1 | 78.0 | 0.553 | 72.7 | 80.7 | 77.1 | 0.536 |
| | CC-III | 74.3 | 81.0 | 78.0 | 0.554 | 72.9 | 80.2 | 76.9 | 0.533 |
| | CC-IV | 72.7 | 82.1 | 77.9 | 0.552 | 73.3 | 80.3 | 77.1 | 0.538 |
| | CC-V | 70.2 | 77.0 | 74.0 | 0.473 | 70.5 | 77.3 | 74.2 | 0.479 |
| | Mean | 65.5 | 75.7 | 71.1 | 0.414 | 65.5 | 77.8 | 72.3 | 0.437 |
| | Maximum | 69.2 | 78.5 | 74.3 | 0.479 | 64.9 | 80.2 | 73.4 | 0.458 |
| | Multiply | 68.0 | 78.3 | 73.7 | 0.466 | 63.9 | 80.6 | 73.1 | 0.452 |
| 2D6 | CC-I | 41.0 | 95.0 | 83.6 | 0.444 | 40.1 | 94.9 | 83.3 | 0.433 |
| | CC-II | 39.7 | 95.6 | 83.8 | 0.447 | 39.0 | 95.2 | 83.3 | 0.428 |
| | CC-III | 39.6 | 95.5 | 83.7 | 0.442 | 39.7 | 95.3 | 83.5 | 0.438 |
| | CC-IV | 37.6 | 96.1 | 83.7 | 0.439 | 38.8 | 95.3 | 83.3 | 0.430 |
| | CC-V | 37.1 | 95.0 | 82.8 | 0.408 | 37.0 | 94.7 | 82.5 | 0.398 |
| | Mean | 47.1 | 93.4 | 83.6 | 0.461 | 45.6 | 93.3 | 83.2 | 0.445 |
| | Maximum | 49.0 | 90.5 | 81.7 | 0.422 | 48.5 | 91.0 | 82.0 | 0.427 |
| | Multiply | 48.3 | 92.3 | 82.9 | 0.448 | 48.0 | 91.8 | 82.5 | 0.437 |
| 3A4 | CC-I | 65.3 | 83.4 | 76.1 | 0.497 | 63.7 | 82.3 | 74.8 | 0.470 |
| | CC-II | 64.6 | 84.9 | 76.7 | 0.509 | 62.2 | 83.7 | 75.1 | 0.473 |
| | CC-III | 64.6 | 84.5 | 76.5 | 0.504 | 63.2 | 83.1 | 75.1 | 0.475 |
| | CC-IV | 64.6 | 84.5 | 76.5 | 0.504 | 61.4 | 84.4 | 75.2 | 0.475 |
| | CC-V | 61.4 | 81.8 | 73.6 | 0.442 | 59.5 | 81.0 | 72.4 | 0.416 |
| | Mean | 70.0 | 79.0 | 75.4 | 0.490 | 68.9 | 77.3 | 74.0 | 0.461 |
| | Maximum | 71.1 | 74.6 | 73.2 | 0.452 | 69.1 | 74.5 | 72.3 | 0.432 |
| | Multiply | 70.7 | 77.7 | 74.9 | 0.481 | 69.6 | 75.1 | 72.9 | 0.443 |

[a] Performances of classification model were given as follows: SE (sensitivity), SP (specificity), Q (overall predictive accuracy), and C (Matthews correlation coefficient). CC-I (SVM+C4.5DT+*k*-NN+NB), CC-II (SVM+*k*-NN), CC-III (SVM+C4.5DT), CC-IV (SVM+NB), and CC-V (C4.5DT+*k*-NN+NB). SVM (support vector machine), C4.5DT (C4.5 decision tree), *k*-NN (*k*-nearest neighbor), NB (Naïve Bayes).

using the same data description method. Similar performance occurred for the CYP2C9, 2C19, 2D6, and 3A4 models. This is in agreement with previous studies, in which the SVM algorithm is superior to other machine learning algorithms for the classification of CYP substrates and inhibitors[38] and toxicity prediction.[39] Although the Q and C values of SVM were higher than those of C4.5 DT, *k*-NN, and NB for CYP2C9 and 2D6, the SE of SVM was anomaly low. For CYP2D6, the SE value was only 39.0% using the SVM algorithm with MACCS keys. In contrast, the SE value was 51.0 and 53.9% using *k*-NN and NB classifiers, respectively, with the same data description. The similar performance also occurred for the CYP2C9 models. This may be caused by the unbalanced data set for CYP2C9 and 2D6. The ratio of positive examples to negative examples was 0.56:1 for CYP2C9 and 0.26:1 for CYP2D6, respectively, as given in Table 1. Therefore, we were particularly interested in the performance of combining different single classifier algorithm in the context of the unbalance data sets.

**Construction of Binary Classification Models by Combined Classifiers (CC).** In this work, five kinds of new CC were designed based on the BP-ANN fusion rules. They include one kind of four CC: SVM+C4.5DT+*k*-NN+NB (CC-I); three kinds of two CC: SVM+*k*-NN (CC-II), SVM+C4.5DT (CC-III), SVM+NB (CC-IV); and one kind of three CC: C4.5DT+*k*-NN+NB (CC-V). In addition, SVM, C4.5DT, *k*-NN, and NB fused by three classic classifiers fusion techniques (Mean, Maximum, and Multiply) were also evaluated. The performance of different CC was given in Table 3, evaluated by the 5-fold cross-validation.

As shown in Table 3, the reasonably high performance was obtained for all five major CYP isoforms using CC evaluated by 5-fold cross-validation. For CYP1A2, the *Q* value was 81.1, 81.3, 81.3, 81.1, and 78.1% for the CC-I, CC-II, CC-III, CC-IV, and CC-V, respectively, with MACCS keys. The similarly high *C* value of 0.603 for CC-I, 0.612 for CC-II, 0.610 for CC-III, 0.610 for CC-IV, and 0.544 for CC-V were obtained with FP4 fingerprints. For CYP2C9, 2C19, 2D6, and 3A4, the similarly high performances were obtained using CC in Table 3. Compared to CC-I, CC-II, CC-III, CC-IV, the performance of CC-V was the worst. For example, the *C* value of 2C19 classification models was 0.533, 0.553, 0.554, and 0.552 for CC-I, CC-II, CC-III, and CC-IV, respectively, which were higher than 0.473 for CC-V. The similar phenomenon occurred for CYP1A2, 2C9, 2D6, and 3A4 models. However, there are no obvious differences between the CC-I, CC-II, CC-III, and CC-IV. It showed that CC performed very well if they included the SVM classifier. The detailed summaries of TP, TN, FP, and FN in 5-fold cross-validation were presented in Table S2 of the Supporting Information.

As shown in Table 3, the overall performance of CC fused by BP-ANN was higher than that of Mean, Maximum, and Multiply. For Mean, Maximum, and Multiply, the prediction with the simple average probability or the highest probability was selected. Thus, the potential drawback is that a good prediction can be overridden by many bad predictions. For CC fused by BP-ANN, the individual class membership probabilities from the SC models were used as the new descriptors to build new BP-ANN classification models. So our developed new CC models fused by BP-ANN can avoid some drawback of Mean, Maximum, and Multiply and outperform better than these classic fusion techniques. Our finding is in agreement with Kramer's reports.[40]

**Assessment of Generalization Abilities.** Generalization ability of a model decides the usefulness and reliability of models. Herein, a diverse validation set (PubChem Data Set II) containing 8988 unique compounds was further used to examine the performance of SC and CC models. First, the data sets of 5663 inhibitors and 6436 noninhibitors for CYP1A2, 4369 inhibitors and 7761 noninhibitors for CYP2C9, 5322 inhibitors and 6563 noninhibitors for CYP2C19, 2516 inhibitors and 9365 noninhibitors for CYP2D6, and 4637 inhibitors and 6899 noninhibitors for CYP3A4 in PubChem Data Set I were used as the new training sets to develop CYP1A2, 2C9, 2C19, 2D6, and 3A4 new global classification models, respectively. Then, new developed global models were further validated by the PubChem Data Set II to test the generalization abilities of models.

The performances of SC and CC models for the validation set without applying AD were summarized in Tables 4 and 5. The high *Q* values ranged from 86.2 to 87.6% were obtained for CYP2D6 using CC with MACCS keys. A similarly high predictive accuracy was obtained for CYP1A2, 2C9, 2C19, and 3A4,

**Table 4. Overall Predictive Accuracy (*Q*) of Five Major CYP Isoforms Validation Sets with Full Coverage (100% - No Applicability Domain)[a]**

| | | Q (%) | | | | |
|---|---|---|---|---|---|---|
| data description | methods | 1A2 | 2C9 | 2C19 | 2D6 | 3A4 |
| MACCS | SVM | 68.0 | 86.6 | 80.3 | 87.8 | 74.9 |
| | C4.5 DT | 68.3 | 78.1 | 77.5 | 82.6 | 70.6 |
| | *k*-NN | 69.0 | 81.5 | 80.7 | 84.5 | 70.4 |
| | NB | 61.0 | 69.4 | 70.5 | 80.6 | 69.3 |
| | CC-I | 71.3 | 86.5 | 80.6 | 87.5 | 75.1 |
| | CC-II | 72.0 | 86.7 | 80.8 | 87.5 | 74.9 |
| | CC-III | 72.3 | 86.3 | 80.5 | 87.5 | 74.6 |
| | CC-IV | 73.1 | 86.4 | 80.4 | 87.6 | 75.0 |
| | CC-V | 69.9 | 83.6 | 80.4 | 86.2 | 74.2 |
| | Mean | 68.1 | 84.0 | 76.6 | 87.2 | 73.6 |
| | Maximum | 67.0 | 81.5 | 79.7 | 86.0 | 72.8 |
| | Multiply | 67.1 | 82.7 | 79.5 | 86.8 | 73.7 |
| FP4 | SVM | 67.6 | 84.2 | 80.5 | 86.4 | 75.6 |
| | C4.5 DT | 65.8 | 78.8 | 77.2 | 78.5 | 72.0 |
| | *k*-NN | 69.7 | 78.4 | 77.9 | 81.1 | 66.3 |
| | NB | 59.7 | 75.4 | 73.5 | 80.2 | 70.9 |
| | CC-I | 70.7 | 84.5 | 80.2 | 85.6 | 75.8 |
| | CC-II | 71.2 | 84.6 | 81.0 | 86.1 | 75.8 |
| | CC-III | 70.8 | 84.6 | 80.7 | 85.8 | 76.0 |
| | CC-IV | 71.7 | 84.1 | 80.7 | 85.6 | 75.4 |
| | CC-V | 70.2 | 83.2 | 79.0 | 84.5 | 75.0 |
| | Mean | 67.2 | 81.9 | 77.5 | 85.6 | 74.8 |
| | Maximum | 66.5 | 80.0 | 78.1 | 83.9 | 74.0 |
| | Multiply | 67.3 | 81.1 | 77.9 | 85.2 | 73.9 |

[a] Performances of classification models were given as follows: Overall predictive accuracy (Q). CC-I (SVM+C4.5DT+*k*-NN+NB), CC-II (SVM+*k*-NN), CC-III (SVM+C4.5DT), CC-IV (SVM+NB), and CC-V (C4.5DT+*k*-NN+NB), SVM (support vector machine), C4.5DT (C4.5 decision tree), *k*-NN (*k*-nearest neighbor), and NB (Naïve Bayes).

as presented in Table 4. For the classification problem, the measurement of the area under the receiver operating characteristic curve (AUC) was highly recommended.[59] The AUC value of the validation sets for different CC and SVM models were given in Table 5. The AUC value was 0.764 to 0.815 for CYP1A2, 0.837 to 0.861 for CYP2C9, 0.793 to 0.842 for CYP2C19, 0.839 to 0.886 for CYP2D6, and 0.754 to 0.790 for CYP3A4, respectively, using the BP-ANN combined classifiers. The detailed summaries of TP, TN, FP, and FN for the validation sets were presented in Table S3 of the Supporting Information. These results indicated that the reasonable high prediction accuracies were obtained here using our new developed CC models.

**Fragments Characteristics between Inhibitors and Noninhibitors.** To further explore the structural features of inhibitors and noninhibitors against CYP isoforms selectivity, IG method, and substructure fragment analysis[14,22] were performed on the 24,732 unique compounds by combining PubChem Data Sets I and II (Table 1) using FP4 fingerprints. The representative substructure fragments characterizing inhibitors and noninhibitors against different CYP isoforms and the frequency of fragment occurrence were identified as shown in Table 6 and Table S4.

As shown in Table 6, the patterns of Alcohol, Primary_alcohol, Secondary_alcohol, Tertiary_alcohol, 1,2-Diol, 1,2-Aminoalcohol,

**Table 5. Area under the Receiver Operating Characteristic Curve (AUC) Value of Five Major CYP Isoforms Validation Sets with Full Coverage (100% - No Applicability Domain)[a]**

| | | AUC | | | | |
|---|---|---|---|---|---|---|
| data description | methods | 1A2 | 2C9 | 2C19 | 2D6 | 3A4 |
| MACCS | CC-I | 0.809 | 0.857 | 0.838 | 0.880 | 0.780 |
| | CC-II | 0.806 | 0.861 | 0.842 | 0.880 | 0.777 |
| | CC-III | 0.814 | 0.850 | 0.839 | 0.872 | 0.775 |
| | CC-IV | 0.815 | 0.853 | 0.829 | 0.886 | 0.785 |
| | CC-V | 0.776 | 0.837 | 0.829 | 0.860 | 0.754 |
| | Mean | 0.801 | 0.845 | 0.756 | 0.881 | 0.778 |
| | Maximum | 0.789 | 0.813 | 0.670 | 0.874 | 0.765 |
| | Multiply | 0.795 | 0.829 | 0.704 | 0.879 | 0.772 |
| | SVM | 0.814 | 0.854 | 0.841 | 0.880 | 0.783 |
| FP4 | CC-I | 0.805 | 0.842 | 0.798 | 0.860 | 0.790 |
| | CC-II | 0.798 | 0.850 | 0.815 | 0.860 | 0.783 |
| | CC-III | 0.794 | 0.827 | 0.812 | 0.839 | 0.787 |
| | CC-IV | 0.805 | 0.837 | 0.793 | 0.842 | 0.792 |
| | CC-V | 0.764 | 0.844 | 0.819 | 0.841 | 0.777 |
| | Mean | 0.782 | 0.848 | 0.774 | 0.856 | 0.790 |
| | Maximum | 0.772 | 0.836 | 0.674 | 0.843 | 0.784 |
| | Multiply | 0.778 | 0.842 | 0.729 | 0.851 | 0.788 |
| | SVM | 0.803 | 0.832 | 0.820 | 0.848 | 0.785 |

[a] Performances of classification models were given as follows: CC-I (SVM+C4.5DT+*k*-NN+NB), CC-II (SVM+*k*-NN), CC-III (SVM+C4.5DT), CC-IV (SVM+NB), and CC-V (C4.5DT+*k*-NN+NB), SVM (support vector machine), C4.5DT (C4.5 decision tree), *k*-NN (*k*-nearest neighbor), and NB (Naïve Bayes).

and Alpha_Amino_acid were presented more frequently in noninhibitors than inhibitors against five major CYP isoforms. The patterns of Aryl_chloride and Aryl_bromide were presented more frequently in inhibitors than noninhibitors against five CYP isoforms, which is consistent with the known preference for planar, polyaromatic substrates of CYP isoforms, particularly for CYP1A2.[41] The patterns of Primary_aliph_amine and Ammonium were associated with compounds showing low or noninhibitive activities against five isoforms. Yet, secondary_aliph_amine and tertiary_aliph_amine showed some isoform-specific behavior for CYP2D6. For example, secondary_aliph_amine and tertiary_aliph_amine were presented more frequently in noninhibitors than inhibitors against CYP1A2, 2C9, 2C19, and 3A4, but they were presented more frequently in inhibitors than noninhibitors against CYP2D6, which is in agreement with the known preference of CYP2D6 for substrates containing basic, protonatable nitrogen atoms.[42] The substructure fragment of Carboxylic_acid was presented more frequently in noninhibitors than inhibitors against five CYP isoforms. There is a minor difference for patterns of Carboxylic_ester. Carboxylic_ester was presented more frequently in noninhibitors than inhibitors against CYP1A2 and 2C19, but it was presented more frequently in inhibitors than noninhibitors against CYP2C9, 2D6, and 3A4. The pattern of Oximether was strongly correlated with noninhibitory activities against CYP1A2, 2C9, and 2D6. In contrast, it was more frequently in inhibitors than noninhibitors against CYP3A4. The pattern of Alkyl_imide was also strongly correlated with noninhibitory activities against CYP1A2, 2C9, 2C19, and 2D6, yet it was equally correlated with inhibitory and noninhibitory activities against CYP3A4. The pattern of Nitrile was more frequently in inhibitors than noninhibitors against CYP1A2, 2C9, and 3A4. In contrast, it was weakly correlated with noninhibitory activities against CYP2C19 and 2D6.

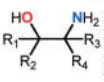
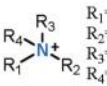
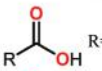
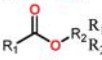
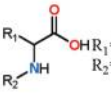
## DISCUSSION

**Comparing Single Classifier (SC) with Combined Classifiers (CC).** Although there exists many machine learning algorithms, there are a few single algorithms which have both the accuracy and robustness to handle the challenge of a real world problems.[43] Previous studies showed that the accuracy of classifications by combining independent single classifier was higher than that of any single classifier.[40,44] The combined strategies and ensemble modeling have been successfully applied in some research fields, such as consensus docking scoring,[45,46] the similarity fusion approach,[47] consensus or ensemble QSAR/QSPR models,[40,48−50] and high-throughput screening (HTS) data analysis and screening.[51]

The objective of this prospective methodological study was to explore the suitability of combined classifiers modeling tools for P450 inhibition prediction. Four reliable and independent SC, namely SVM, C4.5 DT, *k*-NN, and NB, were fused by the BP-ANN algorithm. Five different CC including CC-I, CC-II, CC-III, CC-IV, and CC-V were designed and evaluated here. Hunag and Suen also found that a ANN combinator compared favorably to other combination methods in pattern recognition problems.[52] Recently, Tulyakov et al. reviewed classifiers combination methods and its theoretical basis.[53] When developing ANN fusion CC models, generalization is an issue because the ANN model is often easier to overfit the data and generalize poorly on new data when training data are insufficient or unevenly distributed.[54] In order to avoid this problem, first we employed Momentum parameter $\alpha$ to accelerate the entire network's convergence. Then we repeated the training process for 10,000 times to avoid ANN stuck in local optimal point. As shown in Tables 4 and 5, the reasonable high prediction accuracies were obtained for the validation sets. It indicated that all combined classifiers models developed here had good generalization abilities.

The results of systematic comparisons among different SC and CC models were plotted in Figure 5, which reported the distribution of overall predictive accuracy rate from averaging 10 times realization for CYP1A2 validation set (more details can be found in Table S3 of the Supporting Information). As shown in Figure 5, the overall predictive accuracy of CC-I, CC-II, CC-III, and CC-IV outperformed than any SC, Mean, Maximum, and Multiply. On average, CC-IV achieved the highest accuracy (73.1%) followed by CC-III (72.3%), CC-II (72.0%), and CC-I (71.3%), respectively. SVM performed relatively worse with an average accuracy rate of 68.0%. NB behaved worst with the lowest average accuracy rate of 61.0%. The systematic comparisons among different SC and CC models for CYP2C9, 2C19, 2D6, and 3A4 validation sets were also plotted in Figure S1 of the Supporting Information. As shown in Figure S1, the overall accuracy rate of CC marginally outperformed the best SC of SVM and the other three classic fusion techniques. For the CYP2D6 validation set, the overall predictive accuracy of CC models fused by BP-ANN was equal to the SVM model. However, the overall predictive accuracy rate of the CYP1A2 validation set using CC improved 5% than SVM. When analyzing the detail performance of CYP1A2 and 2D6 validation sets, we

**Table 6. Occurrence and Frequency of 20 Representative Substructure Fragments in the PubChem Data Sets I and II[a]**

| No | Description | General Structure | 1A2 | | 2C9 | | 2C19 | | 2D6 | | 3A4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $N_I$ | $N_{non-I}$ | $N_I$ | $N_{non-I}$ | $N_I$ | $N_{non-I}$ | $N_I$ | $N_{non-I}$ | $N_I$ | $N_{non-I}$ |
| 1 | Quaternary_carbon |  $R_1$=alkyl, aryl $R_2$=alkyl, aryl $R_3$=alkyl, aryl $R_4$=alkyl, aryl | 552 (0.59) | 1333 (1.41) | 425 (0.70) | 1373 (1.15) | 419 (1.13) | 1374 (0.97) | 479 (0.63) | 1346 (1.26) | 925 (1.00) | 1626 (1.00) |
| 2 | Alcohol |  R=alkyl | 280 (0.35) | 1324 (1.64) | 219 (0.35) | 1640 (1.33) | 285 (0.75) | 1553 (1.07) | 273 (0.36) | 1570 (1.45) | 589 (0.63) | 2007 (1.21) |
| 3 | Primary_alcohol |  R=alkyl, aryl | 120 (0.43) | 440 (1.56) | 98 (0.49) | 491 (1.26) | 67 (0.53) | 544 (1.12) | 86 (0.36) | 484 (1.45) | 183 (0.58) | 696 (1.24) |
| 4 | Secondary_alcohol |  $R_1$=alkyl, aryl $R_2$=alkyl, aryl | 157 (0.28) | 955 (1.71) | 103 (0.23) | 1233 (1.40) | 187 (0.70) | 1105 (1.08) | 148 (0.27) | 1178 (1.52) | 374 (0.55) | 1505 (1.25) |
| 5 | Tertiary_alcohol |  $R_1$=alkyl, aryl $R_2$=alkyl, aryl $R_3$=alkyl, aryl | 52 (0.34) | 257 (1.66) | 37 (0.34) | 284 (1.34) | 47 (0.71) | 270 (1.07) | 60 (0.44) | 269 (1.40) | 100 (0.54) | 412 (1.26) |
| 6 | Primary_aliph_amine |  R=alkyl | 129 (0.49) | 402 (1.51) | 34 (0.18) | 525 (1.42) | 63 (0.53) | 506 (1.12) | 58 (0.24) | 517 (1.54) | 63 (0.30) | 528 (1.40) |
| 7 | Secondary_aliph_amine |  $R_1$=alkyl $R_2$=alkyl | 324 (0.97) | 345 (1.03) | 119 (0.52) | 560 (1.25) | 139 (0.50) | 531 (1.35) | 193 (1.44) | 454 (0.87) | 188 (0.73) | 522 (1.15) |
| 8 | Tertiary_aliph_amine |  $R_1$=alkyl $R_2$=alkyl $R_3$=alkyl | 731 (0.72) | 1320 (1.28) | 307 (0.45) | 1721 (1.28) | 435 (0.54) | 1518 (1.33) | 880 (2.21) | 1036 (0.68) | 829 (0.90) | 1708 (1.05) |
| 9 | 1,2-Aminoalcohol |  $R_1$=H, alkyl, aryl $R_2$=H, alkyl, aryl $R_3$=H, alkyl, aryl $R_4$=H, alkyl, aryl | 79 (0.50) | 236 (1.49) | 32 (0.30) | 288 (1.36) | 98 (1.56) | 204 (0.85) | 46 (0.34) | 278 (1.47) | 49 (0.32) | 380 (1.39) |
| 10 | Ammonium |  $R_1$=H, alkyl, aryl $R_2$=H, alkyl, aryl $R_3$=H, alkyl, aryl $R_4$=alkyl, aryl | 18 (0.32) | 96 (1.68) | 2 (0.05) | 116 (1.49) | 19 (0.87) | 86 (1.03) | 5 (0.10) | 113 (1.64) | 9 (0.19) | 119 (1.46) |
| 11 | 1,2-Diol |  $R_1$=H, alkyl, aryl $R_2$=H, alkyl, aryl $R_3$=H, alkyl, aryl $R_4$=H, alkyl, aryl | 49 (0.22) | 409 (1.78) | 18 (0.08) | 626 (1.47) | 48 (0.38) | 557 (1.16) | 32 (0.12) | 586 (1.62) | 147 (0.46) | 733 (1.30) |
| 12 | Oximether |  $R_1$=H, alkyl, aryl $R_2$=H, alkyl, aryl $R_3$=alkyl, aryl | 37 (0.18) | 372 (1.81) | 90 (0.50) | 438 (1.25) | 115 (1.05) | 411 (0.99) | 114 (0.53) | 402 (1.33) | 252 (1.31) | 281 (0.83) |
| 13 | Carboxylic_acid |  R=H, alkyl, aryl | 279 (0.41) | 1100 (1.59) | 165 (0.35) | 1235 (1.33) | 66 (0.22) | 1368 (1.20) | 121 (0.21) | 1269 (1.56) | 88 (0.14) | 1652 (1.49) |
| 14 | Carboxylic_ester |  $R_1$=H, alkyl, aryl $R_2$=alkyl, aryl | 637 (0.83) | 914 (1.17) | 864 (1.53) | 808 (0.73) | 240 (0.68) | 1454 (1.08) | 836 (1.27) | 753 (0.81) | 1112 (1.27) | 1320 (0.85) |
| 15 | Tertiary_amide |  $R_1$=alkyl, aryl $R_2$=alkyl, aryl $R_3$=alkyl, aryl | 492 (0.59) | 1181 (1.40) | 642 (1.19) | 955 (0.90) | 258 (0.79) | 1316 (1.06) | 608 (0.99) | 875 (1.01) | 1043 (1.53) | 849 (0.70) |
| 16 | Alkyl_imide |  $R_1$, $R_2$= alkly $R_3$=H, alkyl, aryl | 62 (0.32) | 329 (1.67) | 63 (0.32) | 511 (1.35) | 62 (0.54) | 496 (1.12) | 94 (0.40) | 470 (1.42) | 237 (1.00) | 420 (1.00) |
| 17 | Alpha_Amino_acid |  $R_1$=H, alkyl, ary $R_2$=H, alkyl, aryl | 43 (0.36) | 200 (1.54) | 16 (0.16) | 276 (1.43) | 23 (0.38) | 265 (1.16) | 24 (0.21) | 257 (1.56) | 14 (0.12) | 315 (1.50) |
| 18 | Nitrile |  R=H, alkyl, aryl | 513 (1.35) | 250 (0.65) | 304 (1.26) | 410 (0.87) | 131 (0.89) | 581 (1.03) | 299 (1.02) | 411 (0.99) | 361 (1.25) | 437 (0.86) |
| 19 | Aryl_chloride |  R=aryl | 1163 (1.34) | 585 (0.67) | 832 (1.49) | 813 (0.75) | 457 (1.30) | 1232 (0.92) | 1097 (1.54) | 625 (0.62) | 732 (1.07) | 1160 (0.96) |
| 20 | Aryl_bromide |  R=aryl | 251 (1.39) | 111 (0.61) | 214 (1.77) | 143 (0.61) | 98 (1.22) | 289 (0.94) | 264 (1.68) | 115 (0.52) | 173 (1.16) | 240 (0.91) |

[a] $N_I$ is the number of inhibitors in inhibitor class with specified pattern t. $N_{non-I}$ is the number of noninhibitors in noninhibitor class with specified pattern t. The data in square bracket represent the frequency of a fragment in inhibitor or noninhibitor class, respectively.

found that SVM performed as a weak classifier for CYP1A2 but it performed as a strong classifier for CYP2D6. It showed that CC fused by BP-ANN can obtain the higher performance when combining several weak classifiers, but it only marginally improved or at least retained the performance of the best SC when combining the strong and weak classifiers. This phenomenon was also emerged in 5-fold cross-validation. In 5-fold cross-validation, SVM classifier all performed a very strong classifier for

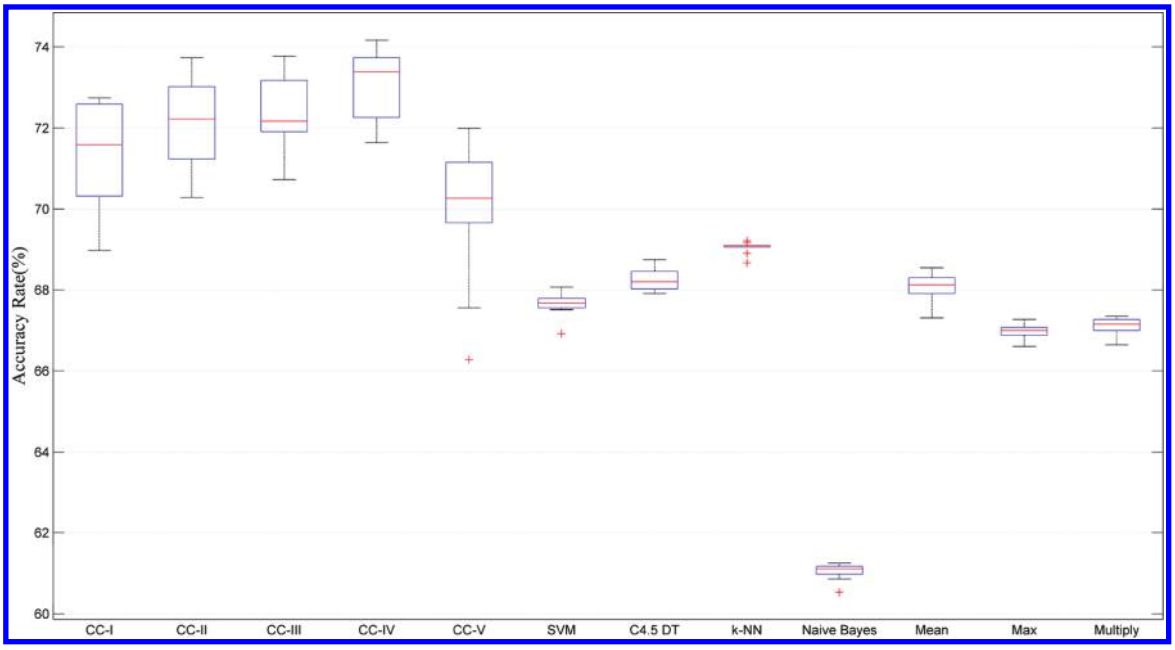



**Figure 5.** Box plot shows the minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), and maximum of the overall predictive accuracy rate from averaging 10 times realization on CYP1A2 validation set with MACCS keys. SVM (Support Vector Machine), C4.5DT (C4.5 Decision Tree), *k*-NN (*k*-Nearest Neighbor), NB (Naïve Bayes). CC-I (SVM+C4.5DT+*k*-NN+NB), CC-II (SVM+*k*-NN), CC-III (SVM+C4.5DT), CC-IV (SVM+NB), CC-V (C4.5DT+*k*-NN+NB), Max (Maximum).

CYP1A2, 2C9, 2C19, 2D6, and 3A4. So our developed CC models only obtained the performance of the best single classifier of SVM in 5-fold cross-validation.

As presented in Tables S2 and S3 of the Supporting Information, the overall performance of CC models marginally outperform or equally perform SVM. However, the SE value of CC models was higher than SVM models. It is important to highlight that sensitivity is the most important parameter in a classification model. In fact, the low sensitivity value indicates the low ability of a model to recognize the inhibitors from diverse compounds. Why CC models can improve the SE value? The key step of combined classifiers is to get probability output of each SC model. In this study, the SC can be viewed as an institute that could convert the MACCS keys or FP4 fingerprints data into new descriptors (new probability outputs) toward a more reasonable orientation, i.e., all the spatial position of each chemical instance are rearranged using the SC. The classifiers combination algorithm based on BP-ANN fusion rules can automatically account for the strengths and score ranges of the SC. So our developed CC models were expected to perform more balance than any SC models. Admittedly, we still have to bear a risk of introducing more errors into the final output.

It is worth noting that the overall performance of CC-V was better than C4.5 DT, *k*-NN, and NB. As shown in Table 4, for CYP3A4 validation set, the overall predictive accuracy was 74.2% for CC-V models, which was higher than 70.6% of C4.5 DT, 70.4% of *k*-NN, and 69.3% of NB. And overall predictive accuracy of 74.2% for the CC-V model was near the 74.9% of the SVM model. The similar phenomenon was performed for CYP1A2, 2C9, and 2C19 validation sets. In addition, we also compared the performance of the single BP-ANN models with our new developed combined classifiers models. New single BP-ANN classification models were developed using PubChem Data Set I and validated by PubChem Data Set II with MACCS keys. As shown in Table S5 of the Supporting Information, the overall performance of single BP-ANN classifier models was anomaly lower than CC models. It further proved that CC strategies can improve the performance of weak classifiers. And combining some different weak classifiers can obtain the performance of a strong single classifier. This is consistent with the idea that the combinational or ensemble models are usually more reliable than its component models.[49,55−57]

**Role of the Applicability Domain (AD).** It is well-known that the predictive reliability is a very important issue given the fact that any QSAR/QSPR model is characterized by its applicability domain (AD). If the test molecule is too far away from a training set member as defined by the user based on a combination of distance and similarity matrix of choice, the performance of the model is usually poor.[58] Herein, the chemical spaces of the training set and the validation set were first explored by MDS plots.[14,34] As shown in Figure 4, the entire data set covered diverse chemical space, and the chemical domain of the validation set (PubChem Data Set II) was basically located in the training set (PubChem Data Set I), which ensured the reasonable high prediction accuracies of the validation set.

The number of validation chemicals determined to be in domain (ID) and out of domain (OD) based on AD analysis were presented in Table 7. For the classification problem, the measurement of the area under the ROC was highly recommended.[59] The AUC values of different CC models with MACCS keys for ID and OD chemicals were calculated and given in Table 8. As expected, the AUC value of ID chemicals was higher than OD chemicals. The detailed summaries of SE, SP, *Q*, and *C* values for ID and OD chemicals were given in Table S6 of the Supporting Information. As shown in Table S6 of the Supporting Information, the *C* value of CYP2D6 OD chemicals was only 0.346 using CC-I with MACCS keys, which was significantly lower than the *C* value 0.576 of ID chemicals. The

*C* value of CYP2C19 OD chemicals were only 0.276 using CC-I with MACCS keys, which was significantly lower than the *C* value of 0.480 for ID chemicals. The similar low performance occurred in CYP1A2, 2C9, and 3A4 OD chemicals. Overall, we concluded that the use of the AD method can lead to improvements in predictive accuracy for the validation sets, although the improvement came at the expense of lower chemistry space coverage. Recently, Didziapetris et al. built the CYP3A4 inhibitor and noninhibitor classification models and applied the reliability index to explore the AD. The higher predictive accuracies were obtained after applying the reliability index.[15] Herein, we did not systematically investigate different AD assessment methods, as it was beyond the range of this article. Our groups are actively investigating this important issue.

**Diversity of Data Set.** Another critical point for developing QSAR models is the diversity of the training set, especially for global models. ADMET predicting models limited their applications when traditionally developed by the small data sets from the literature or combining data sets from different groups. *In vitro* screening conditions often differed in different experimental laboratories and published literatures, which often resulted in unnecessary errors. In addition, some data sets only covered a small region of chemical space focused on drug-like molecules.[60] Recently, several local classification models for P450 substrates and inhibitors were developed, which only built based on the FDA data set or some drug-likeness molecules reported in the literature.[15−20,37] Although high predictive accuracies were obtained for these models, the extremely finite chemical space limited their applications when confronting the real world chemical space, such as virtual screening molecules in drug discovery. Thus, it is important to develop models using high quality diverse data set tested in the same screening condition from various sources, such as public databases (PubChem, BindingDB (http://www.bindingdb.org/)) and pharmaceutical and biotechnology companies.[60] Unlike previous studies, we first took advantage of a high-quality data set containing 15,744 unique compounds with known inhibition of $AC_{50}$ value tested in the same experimental condition by Auld's group,[11] which is believed to be so far the largest public one. High quality P450 inhibition classification models were built using our new developed combined classifiers. As shown in Tables 4 and 5, the reasonable high accuracies were obtained for the diverse validation sets. These global classification models are applicable for virtual screening of the five major CYP isoforms inhibitors or can be used as simple filters of potential chemicals in drug discovery.

**Table 7. Numbers of Chemicals Were Determined To Be In Domain (ID) and Out of Domain (OD) in the Validation Sets Using Application Domain Assessment Methods**[a]

| | in domain (ID) | | | out of domain (OD) | | |
|---|---|---|---|---|---|---|
| CYP isoforms | $N_I$ | $N_{non\text{-}I}$ | total | $N_I$ | $N_{non\text{-}I}$ | total |
| 1A2 | 1717 | 1023 | 2740 | 35 | 29 | 64 |
| 2C9 | 595 | 1914 | 2509 | 14 | 56 | 70 |
| 2C19 | 698 | 1837 | 2535 | 21 | 135 | 156 |
| 2D6 | 530 | 2218 | 2748 | 14 | 98 | 112 |
| 3A4 | 1953 | 4743 | 6696 | 117 | 212 | 329 |

[a] $N_I$ represents the number of inhibitors. $N_{non\text{-}I}$ represents the number of noninhibitors.

**Features and Physical Meanings of the Substructure Fragments.** Interpretation of QSAR/QSPR models is the most important issue. In this study, some representative substructure fragments characterizing inhibitors and noninhibitors against CYP1A2, 2C9, 2C19, 2D6, and 3A4 were identified by combining information gain and substructure fragment analysis. As given in Table 6, the presence of Carboxylic_acid, Primary_alcohol, Secondary_alcohol, Tertiary_alcohol, 1,2-Diol, 1,2-Aminoalcohol, and Alpha_Amino_acid were frequently associated with noninhibitors against five major CYP isoforms. It indicated that the existence of these fragments is unfavorable for CYP1A2, 2C9, 2C19, 2D6, and 3A4 inhibition. Arylchloride or Arylbromide were frequently associated with inhibitors against five isoforms;[41] secondary_aliph_amine and tertiary_aliph_amine showed some isoform-specific behavior for CYP2D6, which is consistent with the known preference of CYP2D6 for substrates containing basic, protonatable nitrogen atoms.[42] Recently, Jensen et al. also searched the frequent structural fragments of noninhibitors and inhibitors for CYP2D6 and 3A4 based on the small data sets. They found that carboxyl acid fragments were more frequent in noninhibitors than inhibitors of both CYP2D6 and 3A4. Auld et al. also identified some key structural features based on 16,144 unique compounds against CYP1A2, 2C9, 2C19, 2D6, and 3A4.[11] Our results were in agreement with Jensen's[14] and Auld's findings.[11] The substructure fragment analysis can characterize the important fragments favorable or unfavorable for P450 inhibition, but they cannot characterize the spatial arrangement of these important fragments if multiple important fragments are

**Table 8. Area under the Receiver Operating Characteristic Curve (AUC) Value of In Domain (ID) and Out of Domain (OD) Chemicals Using Combined Classifiers with MACCS Keys**[a]

| | AUC (in domain) | | | | | AUC (out of domain) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| methods | 1A2 | 2C9 | 2C19 | 2D6 | 3A4 | 1A2 | 2C9 | 2C19 | 2D6 | 3A4 |
| CC-I | 0.803 | 0.858 | 0.829 | 0.875 | 0.779 | 0.714 | 0.723 | 0.754 | 0.732 | 0.739 |
| CC-II | 0.805 | 0.863 | 0.824 | 0.886 | 0.777 | 0.766 | 0.732 | 0.753 | 0.733 | 0.762 |
| CC-III | 0.805 | 0.855 | 0.827 | 0.888 | 0.784 | 0.756 | 0.723 | 0.744 | 0.731 | 0.759 |
| CC-IV | 0.818 | 0.855 | 0.836 | 0.878 | 0.780 | 0.761 | 0.721 | 0.759 | 0.722 | 0.748 |
| CC-V | 0.787 | 0.850 | 0.829 | 0.862 | 0.756 | 0.644 | 0.702 | 0.740 | 0.732 | 0.701 |
| Mean | 0.802 | 0.849 | 0.760 | 0.886 | 0.781 | 0.692 | 0.713 | 0.629 | 0.742 | 0.731 |
| Maximum | 0.795 | 0.819 | 0.675 | 0.881 | 0.770 | 0.682 | 0.700 | 0.541 | 0.735 | 0.726 |
| Multiply | 0.796 | 0.832 | 0.709 | 0.883 | 0.775 | 0.687 | 0.710 | 0.577 | 0.740 | 0.729 |

[a] CC-I (SVM+C4.5DT+*k*-NN+NB), CC-II (SVM+*k*-NN), CC-III (SVM+C4.5DT), CC-IV (SVM+NB), and CC-V (C4.5 DT+*k*-NN+NB). SVM (Support Vector Machine), C4.5DT (C4.5 Decision Tree), *k*-NN (*k*-Nearest Neighbor), and NB (Naïve Bayes).

found simultaneously in a chemical. Anyways, these meaningful substructure fragments identified here can potentially provide scaffold modification for exploring potential metabolic relation toxicological profiles, such as DDIs problems in the early drug discovery phase.

**Comparison with Literature.** A direct comparison of our results with previous studies is inappropriate, because the data sets and data description methods were different between the various models. Nevertheless, a simple comparison of the model statistics could provide some basic information about the accuracy of the various CYP inhibition predicting methodologies. In this study, we built inhibitors and noninhibitors classification models for five major CYP isoforms (1A2, 2C9, 2C19, 2D6, and 3A4) based on the largest data set (more than 20,000 compounds). These data sets not only cover diverse chemical space (Figure 4) but also have good drug-likeness based on Lipinski's Rule-of-Five (seeing **Data Set Analysis**). Recently, Vasanthanathan et al. built the CYP1A2 inhibitors and noninhibitors classification models using the different machine learning methods.[13] The overall predictive accuracy of the best SVM models was 73 to 76% for the internal test set and only 67% for the external validation set of 89 drug molecules. In this study, the overall predictive accuracy of CYP1A2 models was 78.1 to 81.3% for the 5-fold cross-validation and 69.9 to 72.3% for the diverse validation set using the combined classifiers, which was higher than those of Vasanthanathan's models.[13] As shown in Tables 4 and 5, the reasonable high overall predictive accuracies for the validation set were obtained using the CC method. However, the specificity was higher than sensitivity for most models, which also occurred in the 5-fold cross-validation. For example, in the validation set of CYP2D6, the range of the SE value was from 53.1 to 63.1%, but the range of the SP value was from 91.6 to 94.7% using CC with MACCS keys (Table S3 of the Supporting Information). The similar performance occurred for CYP2C9, 2C19, and 3A4. Jensen et al. also found a similar phenomenon that specificity value was 94% for both CYP2D6 and 3A4 models in the internal test set. However, the SE values were only 59 and 65% for CYP2D6 and 3A4 models, respectively.[14]

## CONCLUSIONS

In this study, combined classifiers models were developed to predict inhibition of CYP1A2, 2C9, 2C19, 2D6, and 3A4. For the first time, the prediction models were systematically developed for five major CYP isoforms based on the largest data set of more than 24,700 unique compounds. The range of overall predictive accuracies was 77.2 to 81.3%, 73.5 to 77.3%, 72.3 to 78.0%, 81.7 to 83.7%, and 72.3 to 76.7% for CYP1A2, 2C9, 2C19, 2D6, and 3A4, respectively, using combined classifiers, evaluated by the 5-fold cross-validation. The generalization ability of models was further validated by the diverse validation set of about 8900 unique compounds. The range of AUC values for the validation sets was 0.764 to 0.815 for CYP1A2, 0.837 to 0.861 for CYP2C9, 0.793 to 0.842 for CYP2C19, 0.839 to 0.886 for CYP2D6, and 0.754 to 0.790 for CYP3A4, respectively, using the newly developed combined classifiers. The overall performance of our newly developed combined classifiers fused by BP-ANN was superior to that of three classic fusion techniques of Mean, Maximum, and Multiply, and the use of applicability domain can improve the prediction accuracy. Moreover, some representative substructure patterns were also identified to characterize inhibitors and noninhibitors against the five major CYPs isoforms. We will make all of combined classifiers models and software code available to interested scientists upon request and will collaborate toward establishing a publicly available Web server for predicting P450 inhibition.

In conclusion, the models developed here are reasonable robust and accuracy, which can be applicable for virtual screening of large databases and for predicting potential metabolic relation toxicological profiles, i.e., drug−drug interactions, caused by CYP1A2, 2C9, 2C19, 2D6, and 3A4 inhibition in the early stage of drug discovery.

The PubChem ID number, SMILES, and inhibitor and noninhibitor labels of 24,732 unique compounds against CYP1A2, 2C9, 2C19, 2D6, and 3A4 are available online: http://www.lmmd.org/database.html.

## ASSOCIATED CONTENT

**Supporting Information.** Tables S1−S6 and Figure S1. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

**Corresponding Author**

*Phone: +86-21-64251052. Fax: +86-21-64253651. E-mail: whli@ecust.edu.cn (W.L.), ytang234@ecust.edu.cn (Y.T.).

## ACKNOWLEDGMENT

We thank Dr. Douglas Auld (Genomic Assay Technologies, NIH Chemical Genomics Center, USA) in sharing their data sets with us. This work was supported by the National Natural Science Foundation of China (Grant No. 21072059), Shanghai Natural Science Foundation (Grant No. 10ZR1407000), the Fundamental Research Funds for the Central Universities (Grant No. WY1014010), the Program for New Century Excellent Talents in University (Grant No. NCET-08-0774), the 111 Project (Grant No. B07023), and the National S&T Major Project of China (Grant No. 2009ZX09501-001).

## REFERENCES

(1) Wienkers, L. C.; Heath, T. G. Predicting in vivo drug interactions from in vitro drug discovery data. *Nat. Rev. Drug Discovery* **2005**, *4*, 825–833.

(2) du Souich, P. In human therapy, is the drug-drug interaction or the adverse drug reaction the issue?. *Can. J. Clin. Pharmacol.* **2001**, *8*, 153–161.

(3) Lazarou, J.; Pomeranz, B. H.; Corey, P. N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **1998**, *279*, 1200–1205.

(4) 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 (AUCi/AUC) ratios. *Drug Metab. Dispos.* **2004**, *32*, 1201–1208.

(5) Friedman, M. A.; Woodcock, J.; Lumpkin, M. M.; Shuren, J. E.; Hass, A. E.; Thompson, L. J. The safety of newly approved medicines: do recent market removals mean there is a problem?. *JAMA* **1999**, *281*, 1728–1734.

(6) Zhang, L.; Zhang, Y.; Huang, S. M. Scientific and regulatory perspectives on metabolizing enzyme-transporter interplay and its role in drug interactions: challenges in predicting drug interactions. *Mol. Pharmaceutics* **2009**, *6*, 1766–1774.

(7) Lasser, K. E.; Allen, P. D.; Woolhandler, S. J.; Himmelstein, D. U.; Wolfe, S. M.; Bor, D. H. Timing of new black box warnings

and withdrawals for prescription medications. *JAMA* **2002**, *287*, 2215–2220.

(8) Bjornsson, T. D.; Callaghan, J. T.; Einolf, H. J.; Fischer, V.; Gan, L.; Grimm, S.; Kao, J.; King, S. P.; Miwa, G.; Ni, L.; Kumar, G.; McLeod, J.; Obach, R. S.; Roberts, S.; Roe, A.; Shah, A.; Snikeris, F.; Sullivan, J. T.; Tweedie, D.; Vega, J. M.; Walsh, J.; Wrighton, S. A. The conduct of in vitro and in vivo drug-drug interaction studies: a Pharmaceutical Research and Manufacturers of America (PhRMA) perspective. *Drug Metab. Dispos.* **2003**, *31*, 815–832.

(9) Hutzler, M.; Messing, D. M.; Wienkers, L. C. Predicting drug-drug interactions in drug discovery: where are we now and where are we going?. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 51–58.

(10) Lin, J. H.; Lu, A. Y. Role of pharmacokinetics and metabolism in drug discovery and development. *Pharmacol. Rev.* **1997**, *49*, 403–449.

(11) 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*, 1050–1055.

(12) Shukla, S. J.; Huang, R.; Austin, C. P.; Xia, M. The future of toxicity testing: a focus on in vitro methods using a quantitative high-throughput screening platform. *Drug Discovery Today* **2010**, *15*, 997–1007.

(13) Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P.; Olsen, L.; Jorgensen, F. S. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* **2009**, *37*, 658–664.

(14) Jensen, B. F.; Vind, C.; Padkjaer, S. B.; Brockhoff, P. B.; Refsgaard, H. H. 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.

(15) Didziapetris, R.; Dapkunas, J.; Sazonovas, A.; Japertas, P. Trainable structure-activity relationship model for virtual screening of CYP3A4 inhibition. *J. Comput.-Aided. Mol. Des.* **2010**, *24*, 891–906.

(16) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982–992.

(17) Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. Generalized proteochemometric model of multiple cytochrome p450 enzymes and their inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 1840–1850.

(18) Eitrich, T.; Kless, A.; Druska, C.; Meyer, W.; Grotendorst, J. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. *J. Chem. Inf. Model.* **2007**, *47*, 92–103.

(19) Hammann, F.; Gutmann, H.; Baumann, U.; Helma, C.; Drewe, J. Classification of cytochrome p(450) activities using machine learning methods. *Mol. Pharmaceutics* **2009**, *6*, 1920–1926.

(20) Dagliyan, O.; Kavakli, I. H.; Turkay, M. Classification of cytochrome P450 inhibitors with respect to binding free energy and pIC50 using common molecular descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 2403–2411.

(21) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids. Res.* **2009**, 1–11.

(22) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.* **2010**, *50*, 1034–1041.

(23) Open Babel (version2.2.3). http://openbabel.org/ (accessed Jan. 18, 2010).

(24) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(25) Chang, C. C.; Lin., C.-J. LIBSVM: a library for support vector machines. ≪http://www.csie.ntu.edu.tw/∼cjlin/libsvm≫ (accessed Jan. 18, 2010).

(26) Simon, H. *Neural Networks and Learning Machines*, 3rd ed.; Inc.: Pearson Education: 2009; ISBN 978-0-13-147139-9.

(27) Corinna, C.; Vladimir, V. Support-Vector Networks. *Machine. Learn.* **1995**, *20*, 273–297.

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

(29) Watson, P. Naive Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model.* **2008**, *48*, 166–178.

(30) Ting, F. W.; Chin, J. L.; Ruby, C. W. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. Machine. Learn. Res.* **2004**, *5*, 975–1005.

(31) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.

(32) Accelrys Soft ware Inc., Discovery Studio Modeling Environment, Release 2.1; Accelrys, Inc.: San Diego, CA, USA, 2004. http://accelrys.com/ (accessed May 26, 2010).

(33) John, M. B.; Geoffrey, M. D.; Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(34) Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **1964**, *29*, 115–129.

(35) Boethling, R. S.; Costanza, J. Domain of EPI suite biotransformation models. *SAR QSAR Environ. Res.* **2010**, *21*, 415–443.

(36) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412–424.

(37) 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*, 2588–2605.

(38) Mishra, N. K.; Agarwal, S.; Raghava, G. P. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacol.* **2010**, *10*, 8.

(39) Cheng, F.; Shen, J.; Yu, Y.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere* **2011**, *82*, 1636–1643.

(40) Kramer, C.; Beck, B.; Clark, T. Insolubility classification with accurate prediction probabilities using a MetaClassifier. *J. Chem. Inf. Model.* **2010**, *50*, 404–414.

(41) Lewis, D. F.; Eddershaw, P. J.; Dickins, M.; Tarbit, M. H.; Goldfarb, P. S. Structural determinants of cytochrome P450 substrate specificity, binding affinity and catalytic rate. *Chem. Biol. Interact.* **1998**, *115*, 175–199.

(42) Lewis, D. F. V. A Guide to Cytochrome P450 Structure and Function; Taylor & Francis: London, 2001.

(43) Lee, D. S. *Theory of Classifier Combination: The Neural Network Approach*. Ph.D Thesis, SUNY at Buffalo, 1995.

(44) Tulyakov, S.; Jaeger, S.; Govindaraju, V.; Doermann, D. Review of Classifier Combination Methods. *Studies in Computational Intelligence*; 2008; Vol. 90, pp 361-386.

(45) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548.

(46) Teramoto, R.; Fukunishi, H. Supervised consensus scoring for docking and virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 526–534.

(47) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

(48) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.

(49) Dutta, D.; Guha, R.; Wild, D.; Chen, T. Ensemble feature selection: consistent descriptor subsets for multiple QSAR models. *J. Chem. Inf. Model.* **2007**, *47*, 989–997.

(50) Culp, M.; Johnson, K.; Michailidis, G. The ensemble bridge algorithm: a new modeling tool for drug discovery problems. *J. Chem. Inf. Model.* **2010**, *50*, 309–316.

(51) Simmons, K.; Kinney, J.; Owens, A.; Kleier, D. A.; Bloch, K.; Argentar, D.; Walsh, A.; Vaidyanathan, G. Practical outcomes of

applying ensemble machine learning classifiers to High-Throughput Screening (HTS) data analysis and screening. *J. Chem. Inf. Model.* **2008**, *48*, 2196–2206.

(52) Huang, Y. S.; Liu, K.; Suen, C. Y. A neural network approach for multiclassier recognition systems. In *Proceedings of the Fourth International Workshop on Frontiers in Handwriting Recognition*, Taiwan, Dec 1994; pp 235−244.

(53) Tulyakov, S.; Jaeger, S.; Govindaraju, V.; Doermann, D. Review of Classifier Combination Methods. *Studies in Computational Intelligence*; Springer-Verlag: Berlin, Heidelberg, 2008; Vol. 90, pp 361−386, ISSN: 1860-949X.

(54) Robert, H. N. *Neurocomputing*; Addison-Wesley: 1990.

(55) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* **2008**, *48*, 766–784.

(56) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1971–1978.

(57) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903–911.

(58) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26*, 1315–1326.

(59) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.

(60) Gupta, R. R.; Gifford, E. M.; Liston, T.; Waller, C. L.; Hohman, M.; Bunin, B. A.; Ekins, S. Using open source computational tools for predicting human metabolic stability and additional absorption, distribution, metabolism, excretion, and toxicity properties. *Drug Metab. Dispos.* **2010**, *38*, 2083–2090.