molecular pharmaceutics

Article

pubs.acs.org/molecularpharmaceutics

Development of in Silico Models for Predicting P-Glycoprotein Inhibitors Based on a Two-Step Approach for Feature Selection and Its Application to Chinese Herbal Medicine Screening

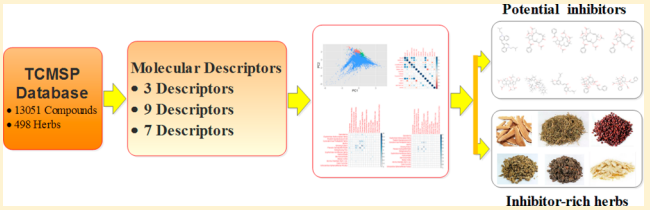
Ming Yang,[†,‡] Jialei Chen,[‡] Xiufeng Shi,[‡] Liwen Xu,[‡] Zhijun Xi,[‡] Lisha You,[†] Rui An,[*,†] and Xinhong Wang[*,†]

[†]Department of Chemistry, College of Pharmacy, Shanghai University of Traditional Chinese Medicine, Shanghai 200444, People's Republic of China

[‡]Department of Pharmacy, Longhua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai 200032, People's Republic of China

Supporting Information

ABSTRACT: P-glycoprotein (P-gp) is regarded as an important factor in determining the ADMET (absorption, distribution, metabolism, elimination, and toxicity) characteristics of drugs and drug candidates. Successful prediction of P-gp inhibitors can thus lead to an improved understanding of the underlying mechanisms of both changes in the pharmacokinetics of drugs and drug−drug interactions. Therefore, there has been considerable interest in the development of in silico modeling of P-gp inhibitors in recent years. Considering that a large number of molecular descriptors are used to characterize diverse structural moleculars, efficient feature selection methods are required to extract the most informative predictors. In this work, we constructed an extensive available data set of 2428 molecules that includes 1518 P-gp inhibitors and 910 P-gp noninhibitors from multiple resources. Importantly, a two-step feature selection approach based on a genetic algorithm and a greedy forward-searching algorithm was employed to select the minimum set of the most informative descriptors that contribute to the prediction of P-gp inhibitors. To determine the best machine learning algorithm, 18 classifiers coupled with the feature selection method were compared. The top three best-performing models (flexible discriminant analysis, support vector machine, and random forest) and their ensemble model using respectively only 3, 9, 7, and 14 descriptors achieve an overall accuracy of 83.2%−86.7% for the training set containing 1040 compounds, an overall accuracy of 82.3%−85.5% for the test set containing 1039 compounds, and a prediction accuracy of 77.4%−79.9% for the external validation set containing 349 compounds. The models were further extensively validated by DrugBank database (1890 compounds). The proposed models are competitive with and in some cases better than other published models in terms of prediction accuracy and minimum number of descriptors. Applicability domain then was addressed by developing an ensemble classification model to obtain more reliable predictions. Finally, we employed these models as a virtual screening tool for identifying potential P-gp inhibitors in Traditional Chinese Medicine Systems Pharmacology (TCMSP) database containing a total of 13 051 unique compounds from 498 herbs, resulting in 875 potential P-gp inhibitors and 15 inhibitor-rich herbs. These predictions were partly supported by a literature search and are valuable not only to develop novel P-gp inhibitors from TCM in the early stages of drug development, but also to optimize the use of herbal remedies.



KEYWORDS: *P-glycoprotein, classification models, Traditional Chinese Medicine, ADMET, virtual screening*

INTRODUCTION

P-glycoprotein (P-gp) is a member of ATP binding cassette (ABC) family that translocates a wide variety of compounds across extra- and intracellular membranes.[1−3] Multidrug resistance (MDR) phenotype during chemotherapy for cancer was reported to be associated with the overexpression of P-gp.[2,4−6] P-gp can decrease effective therapeutic concentration levels of its substrates, which is one of the major causes for the failure of treatment.[7,8] In addition, P-gp is widely present in many human tissues, such as heart, lungs, skin, and spleen, and also some epithelial cells of the gastrointestinal tract including pancreas, intestine, and liver.[2,9] In this way, P-gp was increasingly recognized as being responsible for altered ADMET (absorption, distribution, metabolism, elimination, and toxicity) characteristics of its substrates. Coadministration of P-gp inhibitors helps to enhance the bioavailability of its substrates.[3,10] Although several P-gp inhibitors were discovered earlier, many of them failed in clinical trials because of their side effects and toxicities.[2] Nature resources are believed to be

Received: June 13, 2015
Revised: August 24, 2015
Accepted: September 16, 2015
Published: September 16, 2015

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

relatively safe and low-toxic. Traditional Chinese Medicine (TCM), as an important nature resource of the novel drug design and discovery, is gaining increasingly interest to find efficient P-gp transport blockers.[11,12] In the recent years, great efforts have been made to explore this fertile ground.[13−24]

On the other hand, inhibition or induction of P-gp was reported to be one of the reasons of drug−herb interactions in humans.[25−30] P-gp's low expression (inhibition) results in toxic reactions, whereas its overexpression (induction) results in MDR. Identification of P-gp inhibitors has important implications in avoiding unwanted interactions. Biological assays can be used for transport activity assessment of P-gp inhibitors. However, these in vivo and in vitro assays are expensive and time-consuming.[31] Thus far, a variety of in silico models that can be achieved rapidly and efficiently have been reported. At present, computational approaches to predict P-gp inhibitors can be divided into three types of methods: those based on pharmacophore models,[32,33] those structure-based approaches,[34−36] and those machine learning (ML) methods based on 2D/3D molecular descriptors.[37−39] Pharmacophore models rely on the assumption that different molecules have a similar binding pattern to a receptor. Previous studies reported that the pharmacophore model comprised the hydrogen and hydrophobic bond acceptor features.[32,33] However, most of these published pharmacophore models were established on a small number of molecules. Meanwhile structure-based approaches need high-resolution structure of human P-gp, which is still absent. In general, the performance of structure-based approaches was not comparable to that of ML methods based on 2D/3D molecular descriptors, although homology modeling or murine P-gp was used instead.

In the past few decades, there has been considerable interest in applying ML approaches to identify P-gp inhibitors including unsupervised model such as kohonen self-organizing maps (SOM),[37] and supervised models such as naive bayes (NB) classifier,[40] partial least-squares discriminant analysis (PLSDA),[41] k nearest neighbor (KNN),[38,39] artificial neural networks (NNET),[36] support vector machine (SVM),[34,36,38,39] and random forest (RanForest).[38,39] Although much progress has been made, they are limited to small data sets. To our knowledge, there is no ML model established on a large amount of molecules (larger than 2000). Furthermore, the accuracy of prediction is closely related to the descriptors and statistical techniques, which should be carefully selected. To get the best prediction and easy interpretation, both the choice of the proper descriptor set and the classification method become critical. In the mean time, the problem of the model uncertainty should be considered when a ML model is used for screening new data.[42]

Considering the previous issues, this work proposed a two-step feature selection method to find a set of descriptors that could best differentiate between inhibitors and noninhibitors. Furthermore, 18 supervised ML algorithms were investigated. The main goals of this study were (1) to develop simple prediction models to identify P-gp inhibition using a large updated and structurally diverse data set, comprising 2428 compounds (1518 inhibitors and 910 noninhibitors) and (2) to apply the computational models in the screening of TCM compound database. Further, extensive comparisons of our approach with the other published methodologies show that this approach provides a competitive performance, indicating the effectiveness and advantage of our approach.

METHODS AND MATERIALS

P-gp Inhibitors Data Source. The data sets used in the experiment comprise four main resources: (1) a data set of 1273 compounds, including 797 P-gp inhibitors (multidrug resistance reversal ratio, MDRR ratio >5) and 476 non-inhibitors (MDRR ratio <4) was devived from Chen et al.,[40] (2) a data set of 1275 compounds compiled by Broccatelli et al.[43] from more than 60 literatures, including 666 inhibitors (IC_{50} < 15 μM) and 609 noninhibitors (IC_{50} > 100 μM), (3) a set of 203 compounds constructed by Zdrazil et al.,[44] including 77 inhibitors and 126 noninhibitors, (4) another 572 compounds obtained from OCHEM[45,46] database collected from more than 50 literatures. These compounds were processed as follows: (1) all compounds were represented in 2D SMILES and were prepared by Prepare Ligand Module in Discovery Studio 2.5 (DS2.5), (2) duplicates and overlapping compounds were removed, (3) those differently annotated in the different data sets or different literatures were discarded. This leads to a data set of 2428 compounds, comprising 1518 inhibitors and 910 noninhibitors (Supporting Information, Table S1).

Calculations of Molecular Descriptors and Fingerprints. Classification of P-gp inhibitors by previous studies has provided some insight into possible informative molecular descriptors. However, the descriptor sets for modeling are different, which indicate that the most informative descriptors are not fully understood. To investigate the potential influence of descriptors on classification, a large number of molecular descriptors were calculated to quantitatively define structural and physicochemical properties. All 2D molecular descriptors adopted in PaDEL[47] (version 2.20) software were used for model development. These descriptors include 489 atom-type electrotopological state descriptors, 346 autocorrelation-type descriptors, 96 burden modified eigenvalues of molecules, 91 Barysz matrix of molecules, 68 ring count-type descriptors, 43 extended topochemical atom-type descriptors, 42 information content-type descriptors, 32 Chi path descriptors, 21 topological charge descriptors, 19 molecular distance edge-type descriptors, etc.

A variety of molecular fingerprint sets generated by PaDEL, including 79 Estate fingerprints, 166 MACCS fingerprints, 881 Pubchem fingerprints, 307 Substructure fingerprints, 4860 Klekota-Roth fingerprints, 780 atom pairs 2D fingerprinters, and another 204 functional groups based fingerprints generated by checkmol package, were employed in current study to characterize the substructural features of compounds. The resulting of 8721 descriptors were calculated for all compounds. The descriptions of the descriptors are summarized in Supporting Information, Table S2.

Data Preprocessing and Splitting. There are potential advantages to perform a descriptor reduction process prior to modeling. Several procedures were applied to reduce the number of molecular descriptors.

(1) Descriptors that meet the following criteria were regarded as near-zero variance predictors.[48] The percentage of unique values out of the number of total samples is lower than 10% or the ratio of the most common value to the second most common value is larger than 19. As a result, 6817 near-zero variance descriptors were recognized and were discarded.
(2) Highly correlated descriptors indicate that they measure the same underlying information. Removing one of them

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

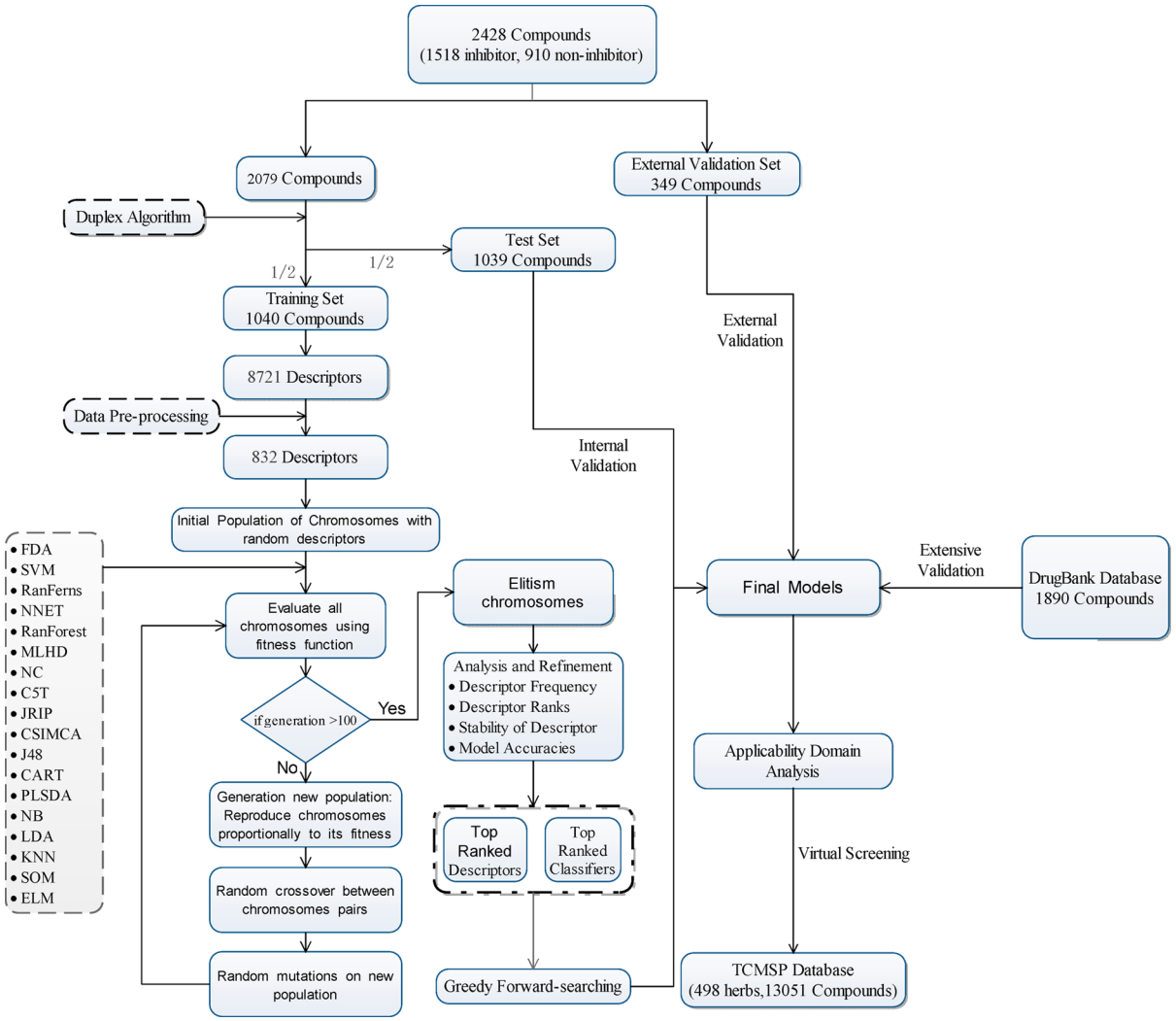


Figure 1. Overview of the analysis pipeline in this work. The proposed two-step feature selection approach is based on a GA and a GFSA. In the first step, GA combined with 18 classifiers to determine the informative descriptors and best performing classifiers. In the second step, GFSA was used to determine the optimal size of the descriptor sets according to the best performing classifiers. Then the classified models were developed based on the final optimal descriptor sets. Two independent data sets and a large publicly available database (DrugBank) were used to evaluate the models. After the applicability domain was defined, the models were used as a virtual screening tool for identifying potential P-gp inhibitors in TCMSP database containing a total of 13 051 unique compounds from 498 herbs.

should not compromise the performance of the model.[48] Then the minimum number of descriptors was removed to ensure that all pairwise correlations are below 0.85. After this step, the remaining 1904 descriptors were reduced to 832.

In addition, 28 continuous descriptors were Box-Cox transformed[48] to resolve their highly skewed distribution (Supporting Information, Table S3). Finally, 832 descriptors (417 continuous descriptors and 415 binary descriptors) were saved, and all continuous descriptors were centered and scaled to unit variance for further analysis.

The data splitting is very important to construct a reliable and robust discriminant model. To select representative samples in the development of models and ensure a sufficient number of samples for validation, 2428 compounds were divided into three subsets, of which 349 compounds from OCHEM data set were stored as external validation set. The residual 2079 compounds were divided into training and test sets using duplex algorithm.[49] The idea of duplex algorithm is based on maximizing the Euclidean distances between the iteratively selected samples. The algorithm started by selecting the two most distant samples and placed them in training set. Then the next two were assigned to test set. The process was continued until all samples in the data set were assigned. The division was performed for each individual class separately so that 50% of each class are in training and test set, separately.

Feature Selection and Classification Models. Without further knowledge, it is difficult to examine which descriptors are best for predicting P-gp inhibitors. Hence, 832 calculated descriptors were used, resulting in high-dimensional feature vectors. Not all of the available molecular descriptors are needed for representing features between inhibitors and noninhibitors. The most informative descriptors that can differentiate among groups can be selected by using feature selection methods. The commonly used feature selection methods can be placed into two main categories: filter and wrapper approaches.[50] The wrapper approaches have gained popularity due to their higher classification accuracy. In this paper, we propose a classification framework based on a two-step wrapper-based feature selection approach. Our proposed

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

classification framework is illustrated schematically in Figure 1. The two-step feature selection approach includes (1) a genetic algorithm (GA)[51] to extract the most informative descriptors and thus greatly reduce the dimensionality of the descriptor space and (2) a greedy forward-searching algorithm (GFSA)[52] to achieve the final optimal descriptor sets for classification.

In the first step, GA combined with various classification techniques was applied on the training set. The basic premise of GA is on Darwin's natural selection principle.[51] More details about GA have been described elsewhere.[53−56] The sequence of steps of GA in current study is as follows.

(1) Initial population of chromosomes with random descriptors: Random descriptor sets were generated from the original 832 descriptors. These random descriptor sets made up a population of chromosomes. The number of descriptors selected in a chromosome is defined as chromosome size. In this study, the population size and chromosome size were set to 20 and 5, respectively.
(2) Chromosomes evaluation using fitness function: A five-fold cross-validation (CV) was applied to evaluate chromosomes' fitness during GA runs. The training set was further divided into five nonoverlapping subsets of the same size according to their categories. One of the five subsets was "held out" for validation, and then classifiers were established using the remaining subsets for calibration. Subsequently, five iterations of calibration and validation were performed, and the average of the resulting five accuracy estimates formed the five-fold CV accuracy that was used for evaluating chromosomes' fitness. For classifier design, 18 classification techniques were investigated. Table 1 lists the detailed information on these classifiers, and the parameters for constructing these classifiers were set to the default values.
(3) Design of the GA operator: Chromosomes with higher fitness would be more likely selected and put into the next generation. Then two main GA operators, crossover and mutation, were applied to generate offspring. Offspring were evaluated again, and these procedures were repeated until the maximum generation was reached. In this study, the maximum generation was set to 100, and the other GA parameters were default according to Galgo package.[57]

Table 1. Eighteen Classifiers and Their Corresponding Packages Used in the Present Work

model	description	package
FDA	Flexible Discriminant Analysis	mda[58]
SVM	Support Vector Machine with Polynomial Kernel	kernlab[59]
RanFerns	Random Ferns classifier	rFerns[60]
NNET	Neural Network	nnet[61]
RanForest	Random Forest	randomForest[62]
MLHD	Maximum Likelihood	Galgo[57]
NC	Nearest Centroid	Galgo[57]
C5T	C5.0 Tree	C50[63]
JRIP	Rule-Based Classifier	Rweka[64]
CSIMCA	SIMCA Classifier	rrcovHD[65]
J48	C4.5-like Trees	Rweka[64]
CART	Classification and Regression Trees	rpart[66]
PLSDA	Partial Least Squares Discriminate Analysis	pls[67]
NB	Naive Bayes	klaR[68]
LDA	Linear Discriminant Analysis	MASS[61]
KNN	k-Nearest Neighbors	Galgo[57]
SOM	Self-Organizing Maps	kohonen[69]
ELM	Extreme Learning Machine	elmNN[70]

At the conclusion of GA search, the single descriptor subset with the highest fitness in the last generation was recorded. However, different searches were likely to provide different subsets due to the instability of GA. Hence, the analysis of the selected descriptor spaces derived from multiple runs provided a more meaningful reduction in the descriptor space. Fifty independent runs of GA were performed in current investigation. The frequency of each descriptor selected in the final solutions during different runs was recorded to represent the importance of descriptor in a specific classifier. The descriptors were ranked according to their frequencies. Then the top-ranked descriptors will most likely contain the most informative descriptors for classification. It should be noted that the performance of combining GA with different classification techniques could be different. Only top-ranked classifiers with their top-ranked descriptors were selected for the further analysis.

One important aspect is how many descriptors from this list of top-ranked descriptors will be used for generating models. In the second step, a GFSA was used to determine the optimal size of the descriptor sets according to the top-ranked classifiers. The top-ranked descriptors derived from GA were used as an input to a GFSA, and the corresponding top-ranked classifiers served as evaluators for the classification to implement feature selection again. The procedure can be summed up as follows. The algorithm started from an empty, and a five-fold CV accuracy of the corresponding classifier was used to express the fitness value to reorder the input descriptor sets. The descriptor with the highest fitness value was first selected as an initial partial solution, and the remaining descriptors was as its children nodes. The algorithm expanded starting node, reevaluated its children, selected the best one, which became a new starting node, and the partial solution was updated accordingly. This procedure was repeated step by step until the fitness value of the partial solution was no longer increased, which meant that a resulting complete solution was obtained.

To determine the best ML algorithm for our problem, 18 classifiers coupled with GA in the first feature selection step were compared, and top-ranked classifiers were employed to establish the optimal models on their corresponding final descriptor sets. Moreover, ensemble modeling techniques were also discussed to get better models. The idea of ensemble modeling is to combine multiple inductive models for classification.

Model Evaluation. The performance of classification models based on their corresponding final descriptor sets was evaluated via 10-fold CV of the training set as well as the prediction of the two independent data set (test set and external validation set). Several statistical metrics including true positive (TP), false positive (FP), true negative (TN), false negative (FN), sensitivity (SE), specificity (SP), Matthews's correlation coefficient (MCC), Kappa statistic, and overall accuracy (OA) were used for describing the model performance. These metrics were calculated as followed:

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (1)$$

$$SE = \frac{TP}{TP + FN} \quad (2)$$

$$SP = \frac{TN}{TN + FP} \quad (3)$$

$$OA = \frac{TP + TN}{TP + FN + TN + FP} \quad (4)$$

$$Kappa = \frac{OA - EA}{1 - EA} \quad (5)$$

Where OA is the observed accuracy that equals the overall accuracy, and EA is the expected accuracy based on the marginal totals of the confusion matrix.

Applicability Domain (AD) Analysis. When applying a developed classification model, it is essential to define the AD for characterization of interpolation space in which the model can make reliable predictions.[71] Compounds in the interpolation space are believed to be predicted more suitably since they have similar characteristics and occupy similar parts of the descriptor space as the compounds on which the model was built. There are currently several different approaches for defining AD, among which projection and ML based approaches are two different methodologies. The projection approach examines the space covered by the descriptors from the training set and the new set via scatter plots using principle component analysis (PCA).[72] If the projections of these data overlap, then they have the similar characteristics. The idea of ML based approach is to predict the probability of new set being members of the training set by a classification model trained on the combination of the original training set and its randomly permuted version.[48,58] In the current study, these two AD approaches were discussed and compared. The projection approach was performed using AMBIT Discovery[73] software (version 0.04), named AMBIT approach. ML based approach was implemented in R programming language. Our ML based approach initially raised by Hastie et al.[48,58] can be described as follows:

(1) The final descriptor sets obtained by the previous feature selection steps were chosen as the predictors for building the classification model.
(2) These descriptors were randomly permuted from the training set, resulting in a shuffled version training set, which was labeled as OutSide-AD. The original training set was labeled as Inside-AD. These two data sets were combined by rows.
(3) A classification model developed by Random Forest was trained on the combined data set.
(4) An ensemble of classifiers can be built simply by repeating steps 2 and 3 100 times, which was used to predict the probability of query compounds being Inside-AD by averaging the individual classifier outputs.

The comparison of the two methods of domain analysis was performed by examining the overall accuracy from the independent test sets, which were classified as either Inside-AD or OutSide-AD for each of the method. The preferable method would have a higher accuracy on the Inside-AD compounds compared with the OutSide-AD compounds. Moreover, ΔOA was defined by the following equation to quantify the role of the compounds considered Inside-AD and OutSide-AD:

$$\Delta OA = OA_{\text{Inside}} - OA_{\text{OutSide}} \quad (6)$$

Where OA_{Inside} and OA_{OutSide} are overall prediction accuracies on the Inside-AD and OutSide-AD compounds, respectively, and ΔOA is the difference between them. The greater the value of the difference, the more preferable that defined domain is in terms of enabling the user decide the appropriateness of the model to make a prediction for a query compound.

Screening Application. The developed models were used to virtually screen the compounds from Traditional Chinese Medicine Systems Pharmacology (TCMSP[74]) database developed by Ru et al.[74] All compounds from TCMSP database were prepared by Prepare Ligand Module in DS2.5. Finally, 13 051 unique compounds from 498 herbs were submitted to virtually screen. Then, the potential P-gp inhibitors from TCM were predicted by the classification models. An inhibitor-rich herb was identified by enrichment analysis[75] using the following equation:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad (7)$$

Where N is the total number of compounds in TCMSP database, M is the total number of potential inhibitors, and n is the number of compounds in a given herb. Then P represents the probability of getting at least k potential inhibitors in a given herb by chance, which obeys hypergeometric cumulative distribution. Further, to control the family wise error rate (FWER), a corrected P-value named Q-value was calculated by the FDR method when performing multiple testing. In our case, a Q-value smaller than 0.01 demonstrates very low probability that the given herb have at least k potential inhibitors by chance; in other words, this herb may be identified as a potential inhibitor-rich herb.

Tools. All calculations were performed with R software. The packages used for implementing various ML algorithms are listed in Table 1. Galgo[57] and FSelector[76] packages were used to perform feature selection procedure.

RESULTS AND DISCUSSION

Characterization of the Data Set. The duplex algorithm resulted in 1040 training compounds (620 inhibitors, 420 noninhibitors) and 1039 test compounds (620 inhibitors, 419 noninhibitors). The external validation set contains 349 compounds (278 inhibitors, 71 noninhibitors). To explore the coverage of the chemical space of compounds, exploratory PCA was performed on the whole data set to compress and to visualize the data structure (Figure 2). Although the total variance explained by the first two principal components was low (18.4%) partly due to the elimination of highly correlated descriptors, the scores scatter plot (Figure 2) showed a trend for separation of inhibitors and noninhibitors. The noninhibitors had more dispersion of the chemical properties than the inhibitors; thus, it seems to hint the difference of prediction performance between two groups when making a classification model. On the other hand, as shown in Figure 2, the distribution of the compounds seems to be well-balanced in the training and test set over the space of the principal components, which indicates the representative ability of the

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

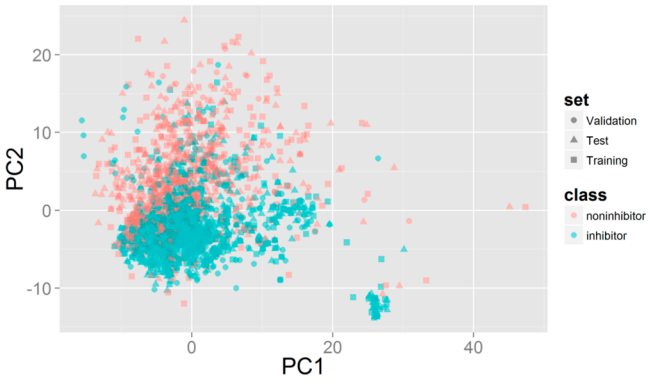


Figure 2. Score plot from PCA based on the whole data set.

compounds in both subsets during duplex algorithm for splitting. Table S4 in the Supporting Information lists the general statistics of the descriptors in the training and test set and highlights the covered range of the descriptors in both sets in a way that is well balanced.

In this study, we performed profile analysis of individual molecular property for all compounds including inhibitors and noninhibitors. Student's *t* test and Fisher's exact test were employed to measure the relevance of the continuous and binary molecular descriptors, respectively. The corresponding *p*-values were calculated and ranked, resulting in 489 statistically significant descriptors with a low *p*-value (*p*-value <0.01). The distributions of the top ten relevant descriptors between groups are shown in Figure 3. These descriptors include, in the order of relevance, MLogP, MDEC.23, topoDiameter, LipoaffinityIndex, CrippenLogP, nHother, SpMin3_Bhs, PubchemFP12, AlogP, and SpMin8_Bhi, of which four (MLogP, LipoaffinityIndex, CrippenLogP, AlogP) were related to molecular hydrophobicity, two (MDEC.23, topoDiameter) were molecular distance measurements based on graph theory, and the other continuous descriptors were electrotopological state indices (nHother) and Burden modified eigenvalues (SpMin3_Bhs, SpMin8_Bhi). As shown in Figure 3, there is a remarkable difference in the distributions of these continuous descriptors between groups. Their distribution of inhibitors is slightly overlapped with those of noninhibitors and skewed toward greater values, which means that all the top ten continuous descriptors of the inhibitor class have greater values than the noninhibitor class. The most relevant descriptor is MLogP, that is to say, P-gp inhibitors are a little more lipophilic than noninhibitors, which is consistent with Chen's results.[40] Only one fingerprint descriptor was ranked in the top ten. This PubChem fingerprint descriptor, named PubchemFP12, is an indicator for whether the number of carbon atoms in the molecule is equal to or greater than 16. The other similar indicators in PubChem fingerprint involve PubchemFP10, PubchemFP11, and PubchemFP13, which indicate that the number of carbon atoms is four, eight, and 32, respectively. PubchemFP10 and PubchemFP11 were eliminated due to their constant properties, which means that most molecules investigated in our project have at least eight carbon atoms. PubchemFP13 located a high position (40th) in the relevance ordering, and its *p*-value is 3.30 e^{-43}, indicating that there was a significant association between PubchemFP13 and P-gp activities. The odds ratio for PubchemFP12 for molecules that identified as being inhibitors was 15.95 (95% confidence interval: 11.69−22.05), which was significantly greater than that for PubchemFP13 (odds ratio = 6.49, 95% confidence interval: 4.71−9.13). These findings suggested that although more carbon atoms in the molecules seem to be discriminative, P-gp inhibitors tend to have 16−32 carbon atoms.

Feature Selection and Classifiers. In the first step feature selection procedure, 50 GA runs were conducted, in which the population size and the maximum number of generations were set at 20 and 100, respectively. Upon completion of a run, the best fitness values represented by the cross-validated performance from the last generation and its corresponding descriptor set for each evaluation function in the run were collected. The best fitness values obtained using 18 different evaluation functions for this procedure are shown in Figure 4. On average, these classifiers combined with GA achieved fitness of 0.814. SVM achieved the highest fitness of 0.826, followed by FDA and RanForest. The corrected *p*-values for each comparison by Wilcoxon signed rank test are given in the Supporting Information, Table S5. The results show that fitness values achieved by the top three classifiers involving SVM, FDA, and RanForest were significantly higher than the other classifiers, whereas ELM, NB, LDA, and PLSDA obtained significantly lower fitness values. It is obvious that the nonlinear classification models performed better than the linear models, which indicates the nonlinear relationships between the molecular characteristics and P-gp activities.

Then the top three classifiers were selected for further analysis. Figure S1 in the Supporting Information illustrates the advantage of using GA combining with these classifiers to select descriptor sets that can distinguish between groups. The performance increases across generations because of the selection of fitter individuals by GA. Figure 5 show the descriptor frequency and the color coded rank of each descriptor in previous evolutions. The 50 most frequently selected descriptors were highlighted in eight different colors. The plot shows different colors if the given descriptor has many changes in ranks. As shown in Figure 5, for the FDA classifier, the first four black descriptors are stable during 50 runs, and the following black descriptors have swap from red and green. These four descriptors are MLogP, CrippenLogP, SpMAD_Dt, and SpMin3_Bhs. Although the low ranked gray descriptors show more instable due to many changes in colors, the most frequencies of these gray descriptors were selected only once. These descriptors mainly involve autocorrelation descriptors (AATSC descriptors and GATS descriptors). Similar results were observed in the following plots. Although there is a good overlap between the top descriptors, each classifier has their own preferred descriptors. Considering the stability analysis of frequencies, we believed that the top 50 ranked descriptors would most likely contain the most informative descriptors for classification. To extensively cover the space of classifiers that were explored, 50 descriptors were left for each classifier, and the number of data dimension was greatly reduced.

Therefore, three descriptor sets and their corresponding classifiers (SVM, FDA, and RanForest) were further analysized in the second step. The GFSA starts from an empty set and progressively adds descriptors until there are no improvements in the CV performance measure. Thus, for each classifier, after feeding the top 50 ranked descriptors into the GFSA, the corresponding best descriptor subset could be selected so that the classifer maximized the CV accuracy over all possible searched subsets. Table 2 lists the final optimal descriptor subset for each classifer. The number of descriptors in the final

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

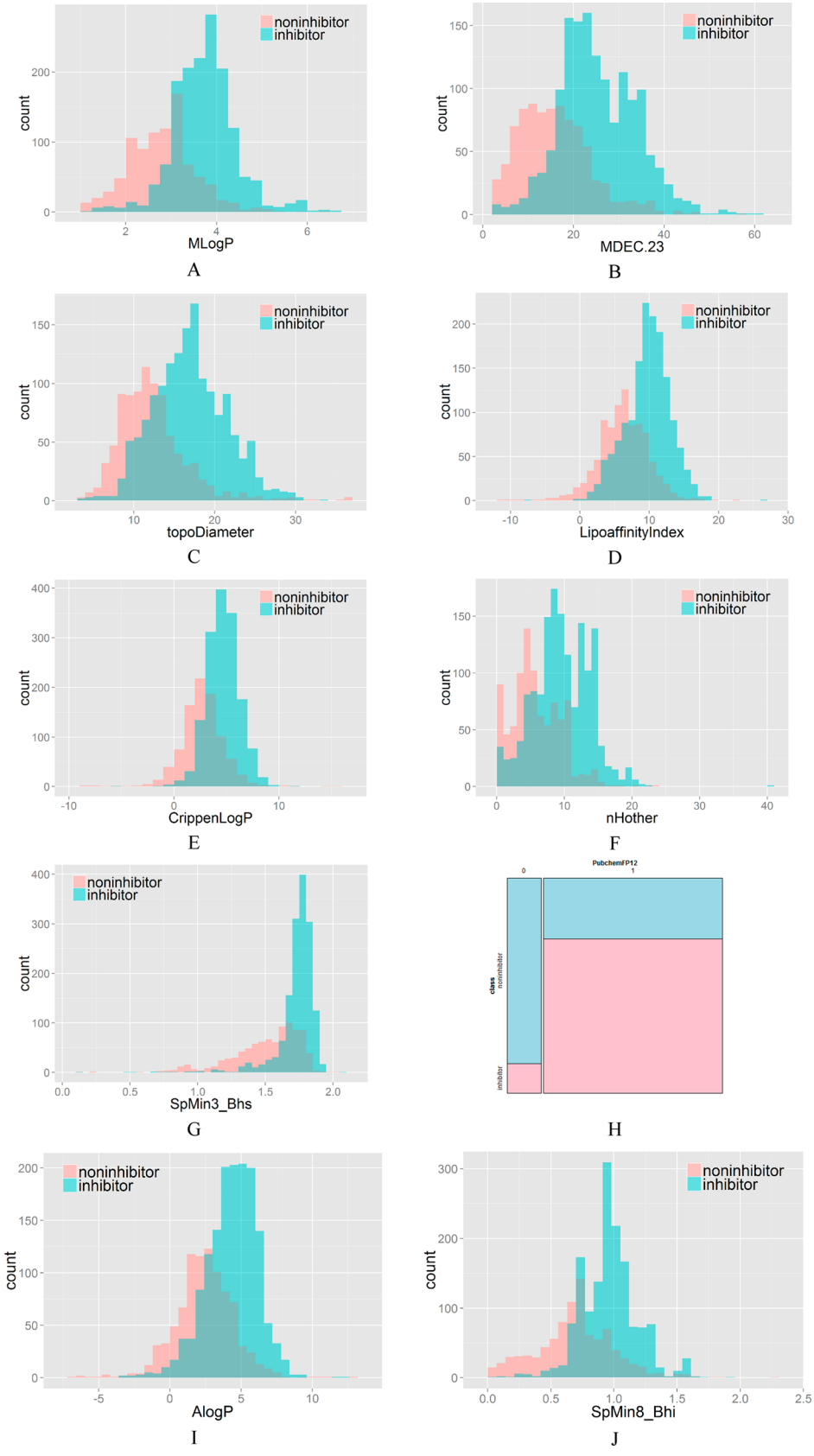


Figure 3. Distributions of top 10 molecular properties for the inhibitor and noninhibitor class. (A) MLogP, (B) MDEC.23, (C) topoDiameter, (D) LipoaffinityIndex, (E) CrippenLogP, (F) nHother, (G) SpMin3_Bhs, (H) PubchemFP12, (I) AlogP, and (J) SpMin8_Bhi.

optimal subsets for FDA, SVM, and RanForest is only 3, 9, and 7, respectively. Figure 6 shows the five-fold CV accuracy of the classifers based on the final optimal subsets, suboptimal subsets obtained by the first step selection, and the original descriptor

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

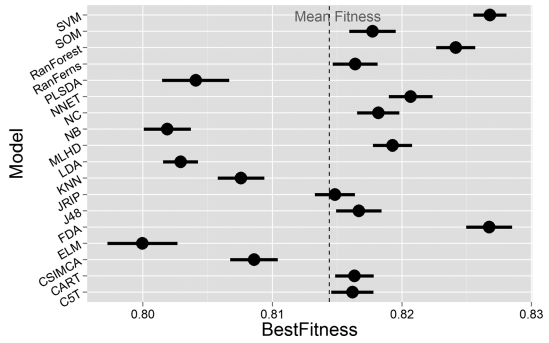


Figure 4. Performance of GA coupled with different classifiers. The best fitness values were derived from multiple GA runs. The black points denote the mean values, and the horizontal line across the point indicates the 95% confidence interval.

set. It can be seen that the descriptor subsets that underwent the two-step feature selection performed better than the descriptor subsets that underwent the GA alone. The original 832 descriptors can only achieve CV accuracy of 80.28%, 80.38%, and 80.58% for FDA, SVM, and RanForest, respectively. After the two-step selection approach was applied, the CV accuracy of classifiers was raised by 3.37%, 5.87%, and 2.98%, respectively, while the number of descriptors in the final optimal subsets is respectively only 0.36%, 1.08%, and 0.84% of the original descriptor set. This indicates that our feature selection approach successfully made a good optimization by eliminating the redundant descriptors.

Furthermore, we compared our feature selection approach with three rank-based feature selection methods: (1) a univariate statistical test (UST) implemented in the profile analysis of individual descriptors, which has been described earlier, and the descriptors were sorted by their corresponding *p*-values in ascending order; (2) ReliefF,[77] which was used to evaluate the quality of descriptors based on how well they distinguished between molecules that were near to each other; and (3) mRMR,[78] which was used to rank the descriptors by so-called minimum-redundancy-maximal-relevance criterion. These rank-based methods were employed in generating various "top *n* descriptor subset size" models based on each classifier. We started the value of *n* to be 2 and progressively added 2 at a time, until 300. Similarly, five-fold CV accuracies were calculated for different values of *n*, and the subset that yielded a maximum CV accuracy was recognized as the final subset. CV accuracy curves of different methods for each classifier can be found in Figure S2 in the Supporting Information. In general, while the number of descriptors increased, the CV accuracy increased. As a result, the best CV accuracies were achieved on the corresponding subsets, which are shown in Table 3. Results show that for all classifiers except RanForest, our method achieved highest CV accuracies on the descriptor subsets of smallest size. For RanForest, UST method raised CV accuracy only 0.57% compared with our method, whereas its final subset size is largest, which is 42-times as large as ours. Overall, UST seems more efficient than the other two rank-based methods. However, when we compared our final optimal descriptors to descriptors selected by rank-based methods, we found that they were quite different. Table S6 in the Supporting Information provides ranking information for our final optimal descriptors by the various methodologies. This analysis revealed that the rank-based methods did not identify the majority of the descriptors selected by our

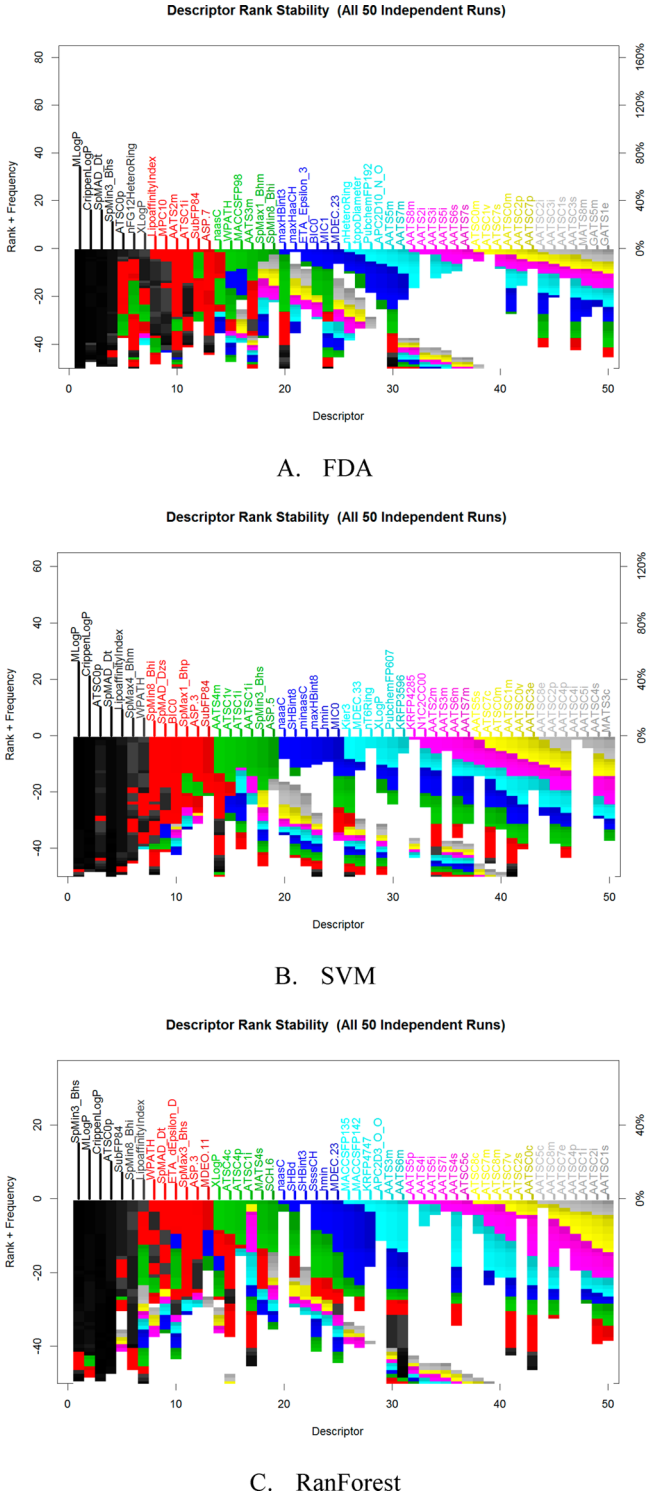


Figure 5. Rank stability of descriptors selected by GA (50 runs) coupled with different classifiers. (A) GA coupled with FDA, (B) GA coupled with SVM, and (C) GA coupled with RanForest. The most frequent 50 descriptors are coded in eight different colors. Horizontal axis denotes the descriptors ordered by rank. The top part of vertical axis denotes the descriptor frequency, and the bottom part of vertical axis denotes the color coded rank of each descriptor in previous evolutions.

approach, especially for ReliefF, and all optimal descriptors are ranked greater than 100. To ensure that all final optimal descriptors were selected, the number of descriptors selected by

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

Table 2. Optimal Descriptors Selected by Two-Step Feature Selection

classifiers	count	descriptors	descriptor type
FDA	3	MLogP	Mannhold LogP
		SpMAD_Dt	Detour Matrix Descriptor
		ATSC1i	Autocorrelation Descriptor
SVM	9	MLogP	Mannhold LogP
		SpMAD_Dt	Detour Matrix Descriptor
		LipoaffinityIndex	Atom Type Electrotopological State
		SpMax4_Bhm	Burden Modified Eigenvalues Descriptor
		SpMax1_Bhp	Burden Modified Eigenvalues Descriptor
		SubFP84	Substructure Fingerprint (Carboxylic acid)
		ATSC1i	Autocorrelation Descriptor
		SHBint8	Atom Type Electrotopological State
		PubchemFP607	Pubchem Fingerprint (N−C−C-C:C)
RanForest	7	MLogP	Mannhold LogP
		ATSC0p	Autocorrelation Descriptor
		SubFP84	Substructure Fingerprint (Carboxylic acid)
		ASP.3	PaDEL ChiPath Descriptor
		SHBint3	Atom Type Electrotopological State
		MACCSFP142	MACCS Fingerprint ($N > 1$)
		AATS5i	Autocorrelation Descriptor

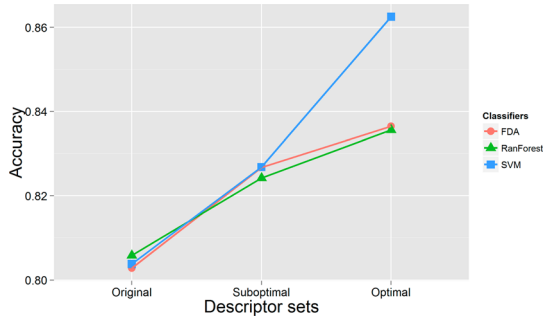


Figure 6. Performance of different classifiers on the different descriptor sets. The colored and shaped points denote the five-fold CV accuracies of the classifers based on the final optimal subsets, suboptimal subsets obtained by the first step selection, and the original descriptor set.

Table 3. Comparison of CV Accuracies of Different Feature Selection Methods for Each Classifier

methods	classifiers	number of descriptors in final subset	accuracy
UST[a]	FDA	62	0.8278
	SVM	80	0.8423
	RanForest	294	0.8413
ReliefF	FDA	100	0.8105
	SVM	288	0.8288
	RanForest	252	0.8298
mRMR	FDA	116	0.7932
	SVM	258	0.8278
	RanForest	262	0.8355
our method	FDA	3	0.8365
	SVM	9	0.8625
	RanForest	7	0.8356

[a]UST, univariate statistical test.

UST, ReliefF, and mRMR was at least 459, 650, and 756, respectively. It is understandable that these rank-based methods are all based on individual descriptor evaluation. However, the best combination of n descriptors does not necessarily contain n individually best descriptors since individual evaluation tends to ignore the interaction between descriptors, and such interactively association of descriptors could affect the classification task.

Development of Classification Models. After the completion of feature selection, the parameters of the classifiers were optimized. Since different classifiers have different tuning parameters, the tuning parameter combination coupled with the highest 10-fold CV accuracy was selected for the final model and used in conjunction with the whole training set. The following individual classification models were investigated:

(1) FDA: First-degree MARS hinge[48] functions were used, and the number of retained terms (nterm) ranged from 3−20.
(2) SVM: Inverse kernel width for the Radial Basis kernel function (sigma) was varied from 0.05−0.3 in conjunction with cost values (C) ranging from −2 to 7 in the log2 scale.
(3) RanForest: Number of trees to grow (ntree) and number of variables randomly sampled as candidates at each split (mtry) were optimized by generating models with each combination of ntree and mtry. Ntree was given in the range of 500−5000, and mtry ranged from 2−5.

After model tuning, for FDA, CV accuracy was largest when nterm equaled 8. Similarly, the optimal sigma and C for SVM were 0.2 and 1, respectively, and the optimal ntree and mtry for RanForest were 3500 and 2, respectively. Then these individual classification models were trained based on their corresponding optimal parameters. Thus, the probability of a compound being inhibitor or noninhibitor was given by each individual model. To further optimize model predictions, output from each individual model was aggregated in an ensemble model (EM). Then the prediction of EM was calculated by averaging the individual outputs. The default probability threshold was 0.5, which means that a compound is identified as inhibitor if its probability of being inhibitor is higher than 0.5. To assess the predictive power of the individual models and EM, all the performance measures calculated for the training set by the 10-fold CV are shown in Table 4. For the training set, we can see that all models developed in this study show the most discriminative power with MCC ranging from 0.6497−0.7237,

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

Table 4. Performance of Different Models

data set	model	*MCC*[a]	*SP*[a]	*SE*[a]	*OA*[a]	*Kappa*[a]
training	FDA	0.6497	0.7214	0.9065	0.8317	0.6422
	SVM	0.7237	0.8000	0.9129	0.8673	0.7209
	RanForest	0.6651	0.7643	0.8903	0.8394	0.6622
	EM	0.6937	0.7643	0.9129	0.8529	0.6889
test	FDA	0.6288	0.7422	0.8774	0.8229	0.6274
	SVM	0.6699	0.7780	0.8855	0.8422	0.6692
	RanForest	0.6529	0.7494	0.8919	0.8345	0.6509
	EM	0.6956	0.7780	0.9065	0.8547	0.6939
external validation	FDA	0.4676	0.7606	0.7842	0.7794	0.4445
	SVM	0.4945	0.8310	0.7590	0.7736	0.4579
	RanForest	0.4786	0.7746	0.7842	0.7822	0.4541
	EM	0.5307	0.8310	0.7914	0.7994	0.5014

[a]These metrics were obtained by the 10-fold CV for training data set.

Table 5. Performance of Monte Carlo Experiments (200 Times) for Different Models ($\bar{x} \pm s$)

data set	model	*MCC*[a]	*SP*	*SE*	*OA*	*Kappa*
test	FDA	NA	0.50 ± 0.20	0.47 ± 0.20	0.48 ± 0.10	−0.02 ± 0.18
	SVM	−0.02 ± 0.13	0.49 ± 0.13	0.49 ± 0.15	0.49 ± 0.07	−0.02 ± 0.12
	RanForest	−0.01 ± 0.06	0.49 ± 0.07	0.50 ± 0.05	0.49 ± 0.03	−0.01 ± 0.06
	EM	−0.02 ± 0.15	0.50 ± 0.16	0.48 ± 0.17	0.49 ± 0.08	−0.02 ± 0.14
external validation	FDA	NA	0.50 ± 0.20	0.48 ± 0.18	0.48 ± 0.12	−0.01 ± 0.11
	SVM	−0.02 ± 0.09	0.49 ± 0.13	0.49 ± 0.13	0.49 ± 0.09	−0.01 ± 0.08
	RanForest	−0.01 ± 0.08	0.49 ± 0.10	0.50 ± 0.07	0.50 ± 0.05	−0.01 ± 0.07
	EM	−0.01 ± 0.12	0.49 ± 0.17	0.49 ± 0.15	0.49 ± 0.10	−0.01 ± 0.09

[a]NA, not available.

OA ranging from 0.8317−0.8673, and *Kappa* ranging from 0.6422−0.7209. Among individual models, SVM achieved the highest performance metrics probably due to the fact that this algorithm used support vectors that provide a hyperplane with a maximal separation between groups, while on the independent test set, EM achieved the best predictive performance with the highest *MCC* at 0.6956, the highest *OA* at 0.8547, and the highest *Kappa* at 0.6939. Similar results were found on the external validation set; EM also achieved the best performance. However, among the individual models, RanForest achieved the highest *SE* and the highest *OA* on the external validation set, implying that it performed better in correctly identifying the positive class (inhibitor class). The performance metrics of models on the external validation set were acceptable in spite of their lower values than that on the training set and test set, and EM got slightly better prediction results compared with the individual models in most cases. The better performance of EM from our experiment is probably due to the complementary role from each of the three individual members in our EM. Indeed the improvement of the ensemble technology depends on the diversity in the ensemble, which was achieved by training the different classifiers on the corresponding different subsets of descriptors in our ensemble system.

To ensure that these results were not due to chance, Monte Carlo experiments were performed by randomly permuting class labels of training set 200 times. Each time, each classification model was retrained on its optimal descriptor set, and the statistics for model performance on the two independent test set were calculated. Table 5 shows the results of Monte Carlo experiments. It can be seen that all models achieved accuracies (*SP*, *SE*, and *OA*) of near 50% and negative *Kappa* values. These results clearly demonstrate that the predictive abilities of proposed models are not due to chance.

To further examine the effectiveness of models for practical purpose, we applied them to predict inhibitors/noninhibitors extracted from publicly available data sources. In present study, compounds from DrugBank 4.0[79] (www.drugbank.ca) were chosen as reference. DrugBank database provides many ADMET related properties of compounds, including two types of prediction probability for P-gp inhibitor/noninhibitor.[79] To obtain high confidence, those compounds whose both prediction probabilities are greater than or equal to 0.8 were extracted for investigation. After the duplicates and overlapping compounds were removed, 1890 DrugBank compounds including 1830 noninhibitors and 60 inhibitors, were evaluated by our models. Table 6 shows the predictions of our models.

Table 6. Predictions of Different Models on DrugBank Database

model	*MCC*	*SP*	*SE*	*OA*	*Kappa*
FDA	0.4977	0.9459	0.8167	0.9418	0.4461
SVM	0.6117	0.9694	0.8333	0.9651	0.5856
RanForest	0.5494	0.9612	0.8000	0.9561	0.5159
EM	0.6080	0.9705	0.8167	0.9656	0.5846

All models display good performance with higher *OA* than 0.94. All models performed better in correctly identifying at least 94% noninhibitors and 80% inhibitors. Similarly, SVM shows the best performance among the individual models, and EM achieved the best *OA* at 0.9656. Overall, our predictive models achieved better concordance with DrugBank predictions.

To explore the classification confidence, the different probability thresholds were investigated. In this way, a cutoff value (COF) was defined as

$$difpro = |prob - 0.5| \geq COF \quad (8)$$

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

Table 7. Classification Performance with Different *COF* Values

COF	model[a]	data set[b]	number of compounds	total compounds	*MCC*	*SP*	*SE*	*OA*	total *OA*	*Kappa*
0	F	T	1039	1388	0.629	0.742	0.877	0.823	0.812	0.627
		V	349		0.468	0.761	0.784	0.779		0.445
	S	T	1039	1388	0.670	0.778	0.885	0.842	0.825	0.669
		V	349		0.494	0.831	0.759	0.774		0.458
	R	T	1039	1388	0.653	0.749	0.892	0.834	0.821	0.651
		V	349		0.479	0.775	0.784	0.782		0.454
	EM	T	1039	1388	0.696	0.778	0.906	0.855	0.841	0.694
		V	349		0.531	0.831	0.791	0.799		0.501
0.1	F	T	934	1251	0.677	0.759	0.905	0.848	0.834	0.675
		V	317		0.495	0.773	0.797	0.792		0.474
	S	T	969	1295	0.725	0.806	0.911	0.869	0.848	0.724
		V	326		0.520	0.841	0.770	0.785		0.487
	R	T	949	1274	0.718	0.788	0.917	0.867	0.847	0.716
		V	325		0.489	0.776	0.791	0.788		0.466
	EM	T	937	1256	0.729	0.794	0.922	0.873	0.857	0.727
		V	319		0.541	0.818	0.806	0.809		0.517
0.2	F	T	838	1132	0.708	0.769	0.922	0.865	0.847	0.706
		V	294		0.500	0.762	0.805	0.796		0.483
	S	T	892	1195	0.768	0.829	0.930	0.890	0.864	0.767
		V	303		0.518	0.831	0.773	0.785		0.487
	R	T	840	1132	0.763	0.803	0.942	0.889	0.865	0.761
		V	292		0.510	0.810	0.791	0.795		0.482
	EM	T	828	1113	0.775	0.824	0.939	0.896	0.872	0.774
		V	285		0.523	0.797	0.805	0.804		0.501
0.3	F	T	726	993	0.759	0.817	0.931	0.890	0.866	0.758
		V	267		0.504	0.764	0.811	0.801		0.487
	S	T	751	1015	0.794	0.840	0.942	0.904	0.878	0.793
		V	264		0.537	0.821	0.798	0.803		0.513
	R	T	687	934	0.820	0.840	0.962	0.918	0.891	0.818
		V	247		0.572	0.894	0.795	0.814		0.533
	EM	T	691	941	0.839	0.883	0.951	0.928	0.897	0.839
		V	250		0.533	0.813	0.812	0.812		0.508
0.4	F	T	530	738	0.826	0.842	0.965	0.921	0.883	0.824
		V	208		0.467	0.763	0.794	0.788		0.440
	S	T	394	545	0.847	0.881	0.959	0.934	0.906	0.847
		V	151		0.527	0.700	0.868	0.834		0.522
	R	T	468	636	0.873	0.869	0.981	0.944	0.915	0.871
		V	168		0.613	0.909	0.815	0.833		0.578
	EM	T	455	616	0.898	0.896	0.984	0.956	0.927	0.896
		V	161		0.618	0.875	0.837	0.845		0.594

[a]F, FDA; S, SVM; R, RanForest. [b]T, test set; V, external validation set.

where *prob* is the probability of a compound being a certain class, which is predicted by a classification model, and *difpro* represents the absolute value of the difference between *prob* and 0.5. Those compounds whose *difpro* are lower than *COF* were discarded, and the model performance metrics were calculated on the retained compounds. For instance, if *COF* equals 0, which means that the probability threshold is set at default value of 0.5, then all compounds are included for evaluation. Table 7 shows the detailed classification performance for different *COF* values on two independent test data sets (test set and external validation set). It is apparent that increasing *COF* increased the performance of classification models, but decreased the number of compounds that the model covered. Some models achieved a "higher than 0.9" total *OA* when *COF* was 0.4; however, the number of covered compounds was approximate half of the originals. This analysis shows that the high confidence of classification is achieved at the expense of the number of covered compounds. Therefore, an appropriate balance between the confidence and the number of covered compounds should be taken into consideration when determining a *COF*.

Descriptor Importance and Contribution. As described earlier, the final optimal descriptor subsets were selected by a two-step selection procedure. After this procedure, each classifier identified a different set of important descriptors (Table 2). As shown in Table 2, the three individual models had different number of descriptors, ranging from 3−9. Together, a total of 14 unique descriptors, which belong to nine different types of molecular descriptors, were found to be important. Only one descriptor appeared in all three optimal subsets. This descriptor is MLogP, which is also found to be top-ranked in UST ranking. MLogP uses a linear model based on 13 1D-descriptors to calculate molecular lipophilicity,[80] among which the sum of lipophilic (CX-summation of weighted numbers of carbon C and halogens atoms) and hydrophilic atoms (NO-total number of N and O atoms) is two

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

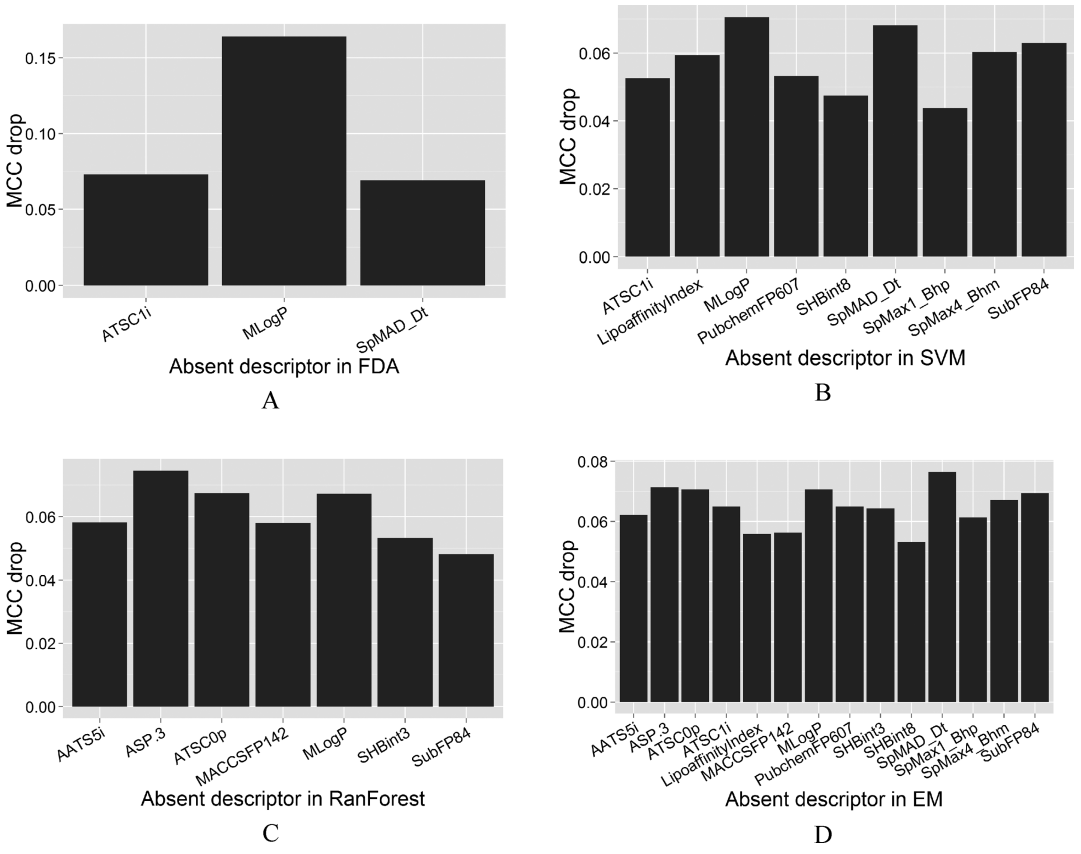


Figure 7. Loss in performance of (A) FDA, (B) SVM, (C) RanForest, and (D) EM when the specific descriptor was removed.

basic descriptors.[80] As analyzed earlier, P-gp inhibitors tend to have higher values of MLogP compared to noninhibitors. As for the other descriptors in the final optimal subsets, there were three common descriptors (SpMAD_Dt, ATSC1i, and SubFP84) in two models. SpMAD_Dt belongs to detour matrix descriptors, which are also called generalized average graph energies.[81] SpMAD_Dt is calculated by the mean absolute deviation from detour matrix. ATSC1i belongs to Moreau-Broto autocorrelation descriptors. It represents spatial autocorrelation defined on a molecular graph and describes how the property considered is distributed along the topological structure.[81] SubFP84 is a substructure fingerprint descriptor, which represents the presence of the pattern of substructure (carboxylic acid). This substructure pattern was also identified by the models developed by Klepsch et al.,[39] suggesting that this functional substructure plays an important role in determining the prediction performance of the model. Moreover, the other two different types of fingerprint descriptors were included in the optimal subsets, indicating that classification performance could be improved when combined with different fingerprint descriptors.

To evaluate the importance and individual contribution of each descriptor, first, we used univariate statistical tests described earlier to check whether the associations were significant. As mentioned previously, although the maximum *p*-value based rank of descriptors in the optimal subsets was 459, its value was still lower than 0.01, indicating that there were significant associations between all descriptors of the optimal subsets and P-gp activities. The detailed rank information was shown in Table S6. Second, the individual contribution of each descriptor was derived by assessing the loss in performance when the effect of the descriptor was negated. In this case, each descriptor was taken out of the optimal subsets in turn, and the remaining descriptors were used to establish the classification model for prediction. Then the performance metrics on the union of independent test sets were calculated, and a substantial drop in performance was indicative of an important descriptor. The changes in the performance of *MCC* are presented in Figure 7. Taking *MCC* drop as the measurement, the discriminative power in FDA is MLogP > ATSC1i > SpMAD_Dt, in SVM is MLogP > SpMAD_Dt > SubFP84 > SpMax4_Bhm > LipoaffinityIndex > PubchemFP607 > ATSC1i > SHBint8 > SpMax1_Bhp, and in RanForest is ASP.3 > ATSC0*p* > MLogP > AATS5i > MACCSFP142 > SHBint3 > SubFP84. Compared with the other descriptors in the individual models, it is clear that MLogP was the most important descriptor because lacking of this descriptor from the model caused the largest performance decrease in most cases. It is particularly surprising that the ranks of contribution of the same descriptors were different in different models. For example, ATSC1i preceded SpMAD_Dt in FDA, while the opposite result was observed in SVM. The difference of *MCC* drop between two descriptors in FDA and SVM is only 0.004 and 0.015, respectively. These findings suggest that these two descriptors may have approximative contributions to the prediction performance apart from the different mathematical mechanism of models. Similar results were found in EM, and the contribution sequence was somewhat different from that of individual models. EM is considerably more complex due to its black-box property. The loss based assessment does not enlighten the modeler as to the exact form of the relationship. Therefore, a drop in performance of a specific descriptor is a

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

conditional value by implicitly taking into account the complementary effect of the other descriptors. Figure 8 is the

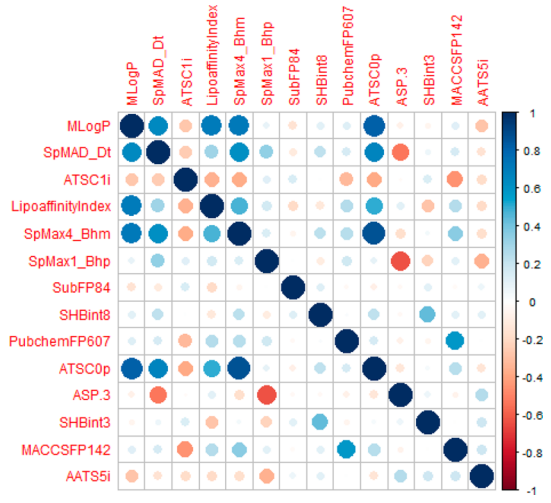


Figure 8. Visualization of the correlation matrix of the optimal descriptor set.

graphical display of the correlation matrix of descriptors. It can be seen that there were four descriptors that were highly correlated with MLogP in EM subset. While there was no descriptors that showed high correlation with ASP.3 in the same individual model set, which means that ASP.3 shared a little information with other descriptors. That is a possible explanation for why ASP.3 takes the important position in the contribution sequences of RanForest and EM, whereas it is only ranked 417 in the univariate statistical analysis. These findings suggest that, for the present problem, any information useful in classification provided by the descriptor can be encoded in other descriptors that are highly correlated.

To better understand how descriptors influence outcomes, we then constructed a quantitative description of the descriptors for classification models. Probability response curves for the optimal descriptors were evaluated to determine the strength and nature of the relationship between P-gp activities and each relevant descriptor. For the continuous descriptor, we altered the value from minimum to maximum, and the remaining explanatory descriptors obtained the medians, and the probability of being inhibitor class was predicted by each model. For the binary fingerprint descriptor, the probability predicted by EM was only investigated when the continuous descriptors obtained the different quantile values. Figure S3 in the Supporting Information reports the probability response curves for all optimal descriptors. The figure shows that the effects of the different descriptors on the probabilities are also different. It is apparent that the probabilities were more sensitive to changes in MLogP. All models give the higher probabilities when MLogP ranges from 4−5, which means that compounds with MLogP in this range have high probability of being inhibitors. The probability response curves of SpMAD_Dt were somewhat complecated. It can be seen that there were two segments in which FDA gave the high probabilities for this descriptor. While the curve of SVM looks like a parabolic curve, and the averaged outputs given by EM seem to be more influenced by FDA, resulting in a similar curve of FDA. In general, the probabilities of being inhibitor class given by models are high when SpMAD_Dt ranges from 20−30. Since SpMAD_Dt was absent from the RanForest subset, the probability given by RanForest was at a constant value. It should be noted that this constant probability value was obtained when the other descriptors were set at the medians. As shown in the graphs, these constant probability values in the probability response curves are high, which suggests that, for our problem, compounds with the median characteristics are likely to be the inhibitor class. As for the binary fingerprint descriptors, the probabilities increase when the quantile values of the other descriptors increase. It is obvious that SubFP84 influenced the change in the probability more significantly than did the other fingerprint descriptors. It can be seen that the probability response curve of the absence of SubFP84 (carboxylic acid) characteristic is mostly above that of the presence. In particular, the difference of the probabilities between two curves is enlarged when the other descriptors

Table 8. Comparisons with Previously Published P-gp Inhibitor Classification Models

model	feature selection	*DM*[a]	*TC*[b]	performance	refs
PLSDA	none	94 Volsurf descriptors	325	Training accuracy = 88.7%; Test accuracy = 72.4%	41
NB	none	177 atom typing descriptors and fingerprints	609	Test accuracy = 82.2%	83
RP, NB	none	Fingerprints and molecular properties	1273	Training: accuracy = 81.7%; Test: accuracy = 81.2%	40
PLSDA, LDA	none	16 VolSurf+ Molecular descriptors and pharmacophoric descriptors	1275	Training:accuracy = 88.0%; Test1: accuracy = 85.0%; Test2: accuracy = 86.0%	82
KNN, SVM, RF	Wrapper subset selection	26 WSE bins based on 204 checkmol fingerprints	1935	Training accuracy = 81.0%; Test accuracy = 75.0%	38
ECP model	Forward-stepwise Logistic regression	16 qualified descriptors out of 87 numerical structural descriptors	1275	Training accuracy = 78.06%; Test accuracy = 82.06%	84
SOM, NNET	Stepwise discriminant analysis	11 out of more than 700 molecular descriptors	206	Total accuracy: 80.8% for the inhibitors	37
SVM	Backward selection	87 out of 1081 descriptors, 3 out of 87 descriptors	1275	Training: accuracy = 84.0%; Test: accuracy = 86.8%	34
SVM, KNN, RF	BestFirst algorithm	46 out of 535 molecular descriptors	1954	Test1: accuracy = 82.0%; Test2: accuracy = 73.0%	39
18 Classifiers	Two-step feature selection approach: a genetic algorithm and a forward greedy search algorithm	3 descriptors for FDA; 9 descriptors for SVM; 7 descriptors for RanForest; 14 descriptors for EM	2428	Training: accuracy = 86.7%; Test1: accuracy = 85.5%; Test2: accuracy = 79.9%	Ours

[a]Descriptors for modeling. [b]Total compounds.

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

Table 9. Statistics for AD Analysis by Different Approaches

		number of compounds			*OA*		
approach	data set	inside	outside	model	inside	outside	Δ*OA*
AMBIT	Test	1029	10	FDA	0.8251	0.6000	0.2251
				SVM	0.8416	0.9000	−0.0584
				RanForest	0.8348	0.8000	0.0348
				EM	0.8542	0.9000	−0.0458
	ExternalValidation	345	4	FDA	0.7797	0.7500	0.0297
				SVM	0.7797	0.2500	0.5297
				RanForest	0.7826	0.7500	0.0326
				EM	0.8000	0.7500	0.0500
	DrugBank	1740	150	FDA	0.9466	0.8867	0.0599
				SVM	0.9626	0.9933	−0.0307
				RanForest	0.9529	0.9933	−0.0404
				EM	0.9638	0.9867	−0.0229
ML[a]	Test	1002	37	FDA	0.8293	0.6486	0.1807
				SVM	0.8513	0.5946	0.2567
				RanForest	0.8413	0.6486	0.1927
				EM	0.8623	0.6486	0.2137
	ExternalValidation	319	30	FDA	0.7994	0.5667	0.2327
				SVM	0.8119	0.3667	0.4452
				RanForest	0.7962	0.6333	0.1629
				EM	0.8182	0.6000	0.2182
	DrugBank	1689	201	FDA	0.9556	0.8259	0.1297
				SVM	0.9751	0.8805	0.0946
				RanForest	0.9692	0.8458	0.1234
				EM	0.9751	0.8855	0.0896

[a]ML, machine learning based approach.

obtained a greater quantile value than 0.5 (median). To summarize, compounds with SubFP84 (carboxylic acid) characteristic have lower probability of being the inhibitor class in most cases. The curves of the other two fingerprint descriptors show little influence on the change in the probability between whether the corresponding structural characteristic is present or not. However, these fingerprint descriptors are important in improving the performance of classification by previous loss-based assessment. The knowledge behind it remains enigmatic and needs to be further explored in the future. The possible explanation may be that there are multiple-interaction effects between descriptors when the model makes predictions.

Comparison with Other Published Classification Models. In this section, we compared our results of the prediction performance with those of published literature. There are many computational models including the regression models and the classification models for predicting P-gp inhibitors. The present work focuses on the classification task, and then the classification performances compared with other research are listed in Table 8. Although some of the literature also explored the structure-based approach, the best performance was usually given by the ML method. It is very obvious to observe that the published models were all developed on no more than 2000 compounds, and the present work was based on the largest data set. The main data sources for previous reported studies were from Chen et al.[40] and Broccatelli et al.[82] Recently, Poongavanam et al.[38] and Klepsch et al.[39] developed models on the combination of these two data sources; however, their performances were not comparable to ours in terms of both the size of data set and prediction accuracy. As shown in Table 8, apparently, it indicates that the current research achieved the higher accuracy and the simplest models, illustrating that our two-step feature selection method outperforms the other approaches embodied in the predictive models for identifying P-gp inhibitors. Furthermore, 18 classifiers coupled with feature selection method were evaluated in this work, which facilitated to the direct comparison between different modeling technologies and yielded the appropriate algorithms for the current problem. In summary, our approach contributes to the better prediction performance compared with others probably for the following reasons: (1) the extraction of more useful and complementary predictors from the large pool of multiple descriptors; (2) the use of better performance ML algorithm from many different classifiers to develop classification models; and (3) the use of an efficient two-step feature selection approach to eliminate irrelevant and noisy descriptors and to choose better combinations of informative descriptors that contribute to the predictions.

Applicability Domain Analysis. It should be noted that AD analysis is necessary to examine a model's predictive reliability for practical purpose. The predictive reliability is the extent to which a classification model is an extrapolation. As for an individual prediction of a given compound, the reliability can be measured in relation to the chemical domain that is represented by molecular descriptors for building a classification model. As mentioned earlier, two AD analysis methods including AMBIT and ML were used to evaluate the extent to which the classification model was applicable to predict a specific compound. The unique 14 molecular descriptors in the final optimal subsets were used for the AD analysis. For AMBIT approach, the default method setting was used. This method calculates the range of descriptors of a specific compound and examined whether it locates in a bounding box, which is a *p*-dimentional hyper-rectangle defined on the basis of maximum and minimum values for the principal components after PCA.

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

For ML approach, we generated 100 randomly permuted training sets and built an ensemble RanForest classification model to predict the probability that the compounds were from the training data. The default cutoff value of the predicted probability to distinguish Inside-AD and OutSide-AD was 0.5, so that compounds with a predicted probability ≤0.5 were classified as OutSide-AD and vice versa. In the present work, the predictions given by all classification models on the two independent test set and the publicly available data source (DrugBank[79]) were used for the comparison. Results generated from these two strategies are shown in Table 9. AMBIT approach identified a smaller number of Outside-AD compounds compared to ML approach, and its ΔOA values were negative in many cases, indicating that it was difficult to differentiate between Inside-AD and OutSide-AD by the classification performance. As is obvious from Table 9, ΔOA values of ML approach were all positive on all data sets and were greater than the corresponding ones of AMBIT approach in most cases. According to the hypothesis that the most useful domain would classify inside-compounds more correctly than outside-compounds, ML approach was an appropriate tool for defining the AD for our problem.

Furthermore, the effect of the cutoff values of the AD probability predicted by ML approach on the model performance was investigated. The model performances on the Inside-AD compounds with the different cutoff values are shown in Table S7 of the Supporting Information. It can be seen that increasing the cutoff value reduced the number of compounds regarded as Inside-AD and improved the classification performance in most cases. Like the previous tactics for an alternative *COF* value, there are also a trade-off between the predictive reliability and the size of AD. However, it is surprising that the best performance was not obtained at the highest cutoff value for the external validation set. These findings suggest that improvements in predictions can be achieved by the benefit of enabling better cutoff values.

Screening Application. After demonstrating the prediction ability of our models for identifying P-gp inhibitors, we applied them to virtually screen the compounds from TCMSP database. After molecule preparation step, 13 051 unique compounds, among which only 96 compounds were present in modeling data sets from 498 herbs, were obtained and evaluated. First, we performed profile analysis of TCM compounds to inspect the coverage of the chemical space. A PCA model derived from the final optimal descriptors was used. Figure 9 shows the PCA score plot. The first two components accounted for 32.1% and 20.4% of the variance in the descriptor matrix, respectively. As seen in Figure 9, there is no clearly defined separation between the TCM compounds and the modeling compounds on the major source of variation, which means that the TCM compounds share the most of the chemical space with the modeling compounds. Figure S4 in the Supporting Information shows the distributions of the final continuous descriptors for these compounds. It is obvious that the range of each descriptor of the TCM compounds covers that of the modeling compounds. The distributions are slightly different between groups, although they peak at a similar position in some cases. For example, the MLogP distribution of TCM compounds is highly overlapped with that of the modeling compounds and slightly skewed to lower MLogP values, which are prevalent in noninhibitors. However, there is also a small peak at its higher value position, which is prevalent in inhibitors. These results suggest that TCM compounds have different and more diverse distributions than either inhibitors or noninhibitors in modeling data sets. We believe that some TCM compounds are likely to be promising compounds with potential to inhibit P-gp due to their similar chemical characteristics of inhibitors.

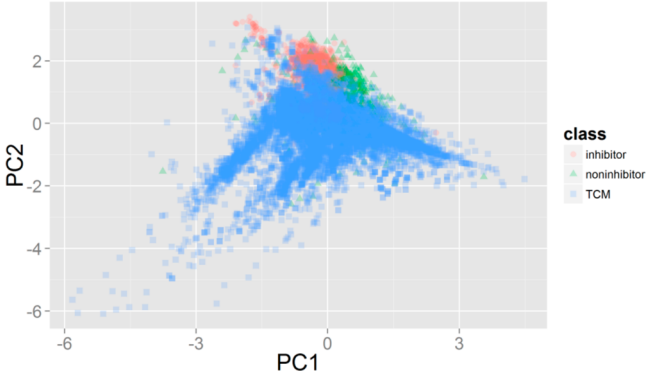


Figure 9. Score plot from PCA based on the combination of TCMSP database and modeling data set.

Second, the probabilities of being the inhibitor class of TCM compounds were predicted by the developed models. To generate high-confidence prediction results, a TCM compound could be recognized as a potential inhibitor only when both of the following criteria were met: (1) the probability that this query compound was a member of the modeling data set predicted by the AD model was greater than or equal to 0.6; (2) the probabilities that this query compound was classified as the inhibitor class predicted by the developed classification models were all greater than or equal to 0.6. In this case, there were 11 556 Inside-AD compounds, which covers 88.5% of the TCM compounds, and consequently, a total of 875 compounds were identified as the potential P-gp inhibitors. The complete list of the predicted inhibitors with detailed information is available in the Supporting Information, Table S8. Figure 10 shows the 2D structures of top 10 compounds with the higher predicted probabilities (the average value of probability >0.95) of being the inhibitor class. Most of these candidate inhibitors are terpenoids. For example, euphornin A, euphornin D, epieuphoscopin B, and euphoscopin B are diterpenes, and these four compounds are all from *Euphorbiae helioscopiae herba*. Two compounds (triptofordin D1 and triptofordin D2) found in *Tripterygii radix* are sesquiterpenes, and one compound (nimbolidin C) from *Toosendan fructus* is triterpenoids. These compounds have the similar stem nucleus. Besides, one peptidic compound (sanjoinine D) and one lignanoid compound (sophoradochromene) respectively from *Ziziphi spinosae semen* and Sophorae *Tonkinensis radix et rhizome* were both found in the top 10 position. The remaining compound (cyanine) is a quinoline alkaloid from four herbs (*Rose rugosae flos*, *Mori fructus*, *Lycii fructus*, and *Perilla frutescens*) in TCMSP database. Subsequently, the structure-inhibitory potency relationship was investigated by maximum common substructure (MSC) analysis. In this way, a cosine similarity between each top 10 compound and each known inhibitor in the modeling data sets was calculated based on the optimal descriptors for measuring structural similarities. Then MSC analysis implemented by fmcsR package[85] was performed between the query compound and its most similar inhibitor. Table 10 shows the results of similarity and MSC analysis. It is obvious that there exists such a known inhibitor that is highly

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

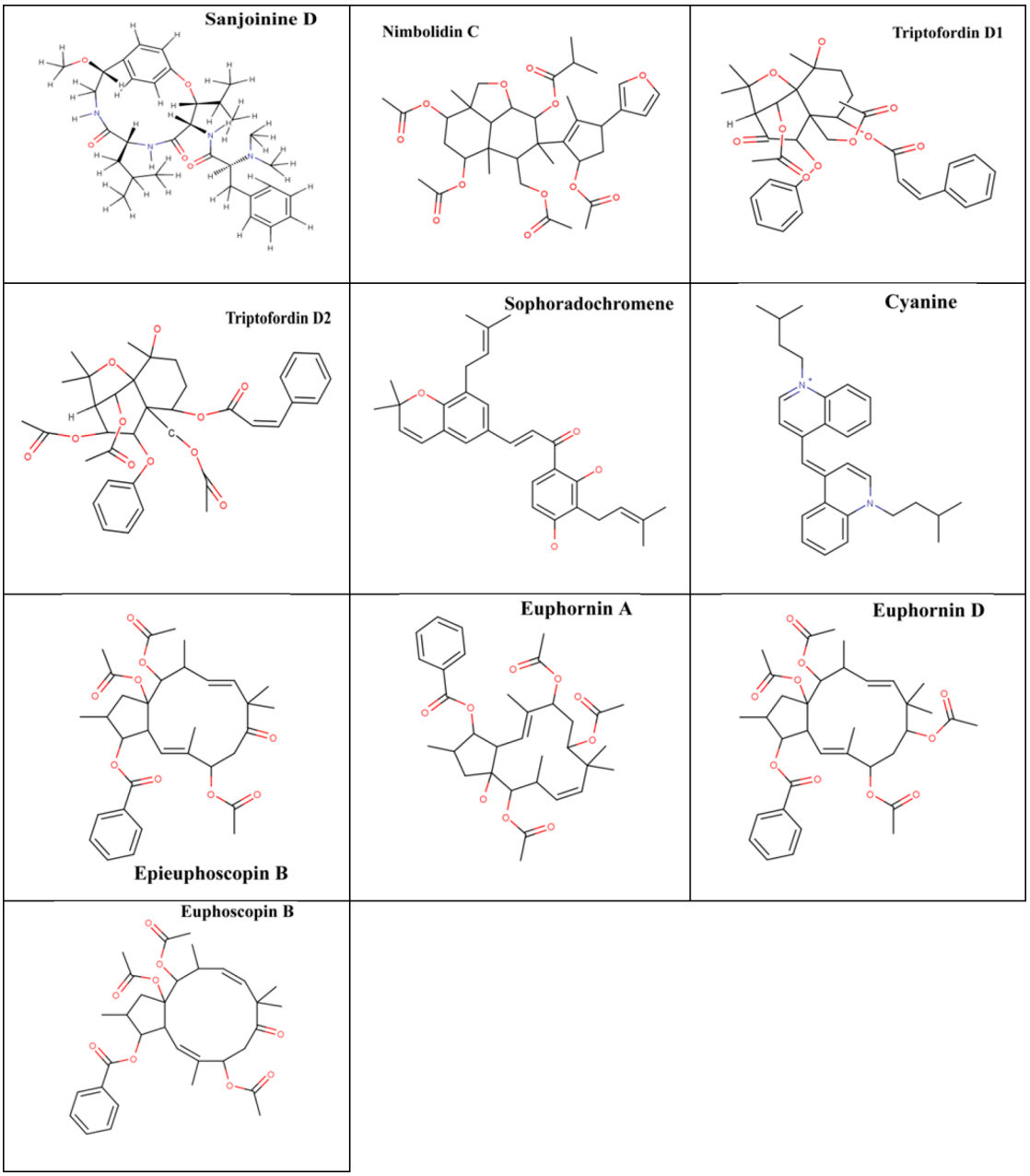


Figure 10. Top 10 compounds from TCMSP database identified as P-gp inhibitors by virtual screening.

similar to the query compound. The similarity values range from 0.893−0.976. Some compounds share the most similar inhibitor due to the similar structural characteristics in the descriptor space. For example, triptofordin D1, triptofordin D2, and nimbolidin C share the same similar inhibitor, while their MSC are somewhat different. Although the relationship between MSC and P-gp properties is unclear, MSC gives the possible explanation for the fact that a higher similarity between a query compound and a known inhibitor may cause a higher probability of being an inhibitor for the query compound.

After enrichment analysis, there were 15 herbs identified as potential inhibitor-rich herbs (with *Q*-value <0.01). Table 11 gives the detailed information on enrichment analysis. As seen in Table 11, the proportion of inhibitors to the total number of known compounds in the inhibitor-rich herb is at least 14.9%. Fifteen inhibitor-rich herbs contain a total of 322 unique inhibitors, which account for 36.8% of the total potential inhibitors in TCMSP. However, the number of inhibitor-rich herbs occupies only 3.0% of total herbs in the database. There are several inhibitor-rich herbs that share the inhibitors. Figure 11 shows the number of compounds and inhibitors shared between inhibitor-rich herbs. We can see that *Gynostemmae pentaphylli herba* and *Panacis quinquefolii radix* share six inhibitors, which account for 60% of the common compounds between them. Similar results can be also found in the other inhibitor-rich herb pairs such as *Ginseng folium*/*Panacis quinquefolii radix* and *Ginseng folium*/*Gynostemmae pentaphylli herba*. These findings demonstrate that different herbs possibly exhibit similar activities due to the large number of shared active compounds. It should be noted that the *Q*-value of enrichment analysis is sensitive to the number of known compounds of both the specific herb and the database. Therefore, our results of enrichment analysis are limited to the current version of TCMSP database.

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

Table 10. Similarity and MSC Analysis between Top 10 Potential Inhibitors in TCMSP Database and Known Inhibitors in Modeling Dataset. Superscripts Indicate (a) Cosine Similarity and (b) MSC, Maximum Common Substructure

Query compound	Known inhibitor	Similarity[a]	MSC[b]
Triptofordin D2		0.976	
Triptofordin D1		0.961	
Nimbolidin C		0.967	
Euphornin D		0.927	
Euphornin A		0.962	
Cyanine		0.955	
Sanjoinine D		0.893	
Epieuphoscopin B		0.965	
Euphoscopin B		0.965	
Sophoradochromene		0.896	

To verify the predictions supported by evidence from the published literature, we carried out a literature search in PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and CNKI (http://www.cnki.net/) to determine to what extent the inhibitor-rich herbs have been observed to possess the activities predicted by our models. However, there are only limited data related to in vivo or in vitro assays on P-gp properties of inhibitor-rich herbs reported. Despite the fact that evidence that supports our predictions is weak, there is still some literature covering our predictions. We consider the literatures that reported the herbal compound hitting our predictions as the direct evidence. However, the inhibitory action to P-gp reported by the extract or decoction of inhibitor-rich herbs is regarded as the indirect evidence. From Table 11, we can see that *Schisandrae sphenantherae fructus* aroused more interest to be investigated on the P-gp properties. Many literatures have given the indirect evidence that its extract has a strong inhibitory action to P-gp, resulting in the improvement of the systemic exposure of paclitaxel[97] and talinolol[20] when coadministered. Elisa et al.[99] reported that a series of diterpenes from *Euphorbiae helioscopiae herba* had high potency for inhibiting P-gp. Among these diterpenes, epieuphoscopin B is also within the top 10 compounds with the higher predicted probabilities of being the inhibitor class. These findings also

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

Table 11. Inhibitor-Rich Herbs Identified by Enrichment Analysis[a]

herb	*k*	*M*	*n*	*N*	proportion	Q-value	LE[b]
Ganoderma	53	875	242	11556	21.9%	0.0000	I,[86−88] D[89−91]
Gynostemmae pentaphylli herba	47	875	202	11556	23.3%	0.0000	D[92,93]
Schisandrae sphenantherae fructus	19	875	39	11556	48.7%	0.0000	I[15,18−20,94−97]
Kansui radix	15	875	31	11556	48.4%	0.0000	I[98]
Euphorbiae helioscopiae herba	21	875	79	11556	26.6%	0.0000	D[99]
Thalictri glandulosissimi et	20	875	75	11556	26.7%	0.0000	I,[91] D[89]
Stemonae radix	24	875	110	11556	21.8%	0.0001	I,[91] D[89]
Meliae cortex	14	875	43	11556	32.6%	0.0001	I,[91] D[89]
Ginseng folium	25	875	126	11556	19.8%	0.0004	I,[100] D[93,101−103]
Myrrha	41	875	276	11556	14.9%	0.0011	I,[104] D[105]
Notoginseng flos	14	875	56	11556	25.0%	0.0024	I,[91,100] D[89,101,102]
Euphorbiae humifusae herba	11	875	38	11556	28.9%	0.0030	
Panacis quinquefolii radix	26	875	153	11556	17.0%	0.0030	I,[100] D[92,93,101−103,106]
Calyx Ccucumis	6	875	12	11556	50.0%	0.0038	
Alisma orientale (Sam.) Juz.	12	875	46	11556	26.1%	0.0038	I,[107] D[108]

[a] *k*, potential inhibitors in a specific herb; *M*, the total number of potential inhibitors identified in TCMSP database; *n*, the number of compounds in a specific herb; *N*, the total number of compounds in TCMSP database. [b] LE, literature evidence: I, indirect; D, direct.

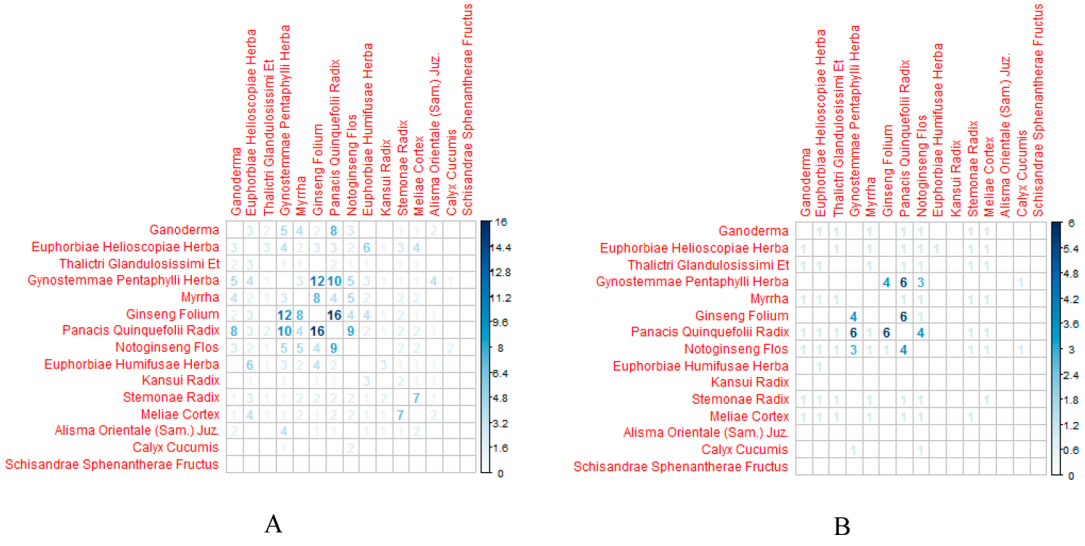


Figure 11. Number of (A) compounds and (B) inhibitors shared between inhibitor-rich herbs.

indirectly support the evidence that the other three analogues in the top 10 compounds, euphornin A, euphornin D, and euphoscopin B, are potential inhibitors of P-gp due to their similar structures. *Ginseng* is one of the most commonly used herbs with antifatigue,[109] immunity-enhancing,[110] and anticancer effects.[111,112] *Ginseng folium* and *Ginseng* are derived from the same species and share most active constituents. Ginsenosides are regarded as the major active constituents. Many of them showed inhibitory effects on P-gp such as 20(S)-ginsenoside Rh2,[22,100] 20(S)-ginsenoside Rg3,[101,106] 20(S)-protopanaxadiol,[103] Ginsenoside Rd, Re, Rb1, and Rg1.[93] The predictions of these ginsenosides were mostly consistent with the published reports. These results suggest that *Ginseng* can improve the activity of existing chemotherapy for cancer treatment because of the potential inhibitory effects on P-gp.

Because of the lack of pertinent information about most of the potential TCM P-gp inhibitors with structures differing from any known P-gp inhibitors, further well-designed studies are warranted to address the validations. In general, the models we have proposed in this paper can be applied for virtual screening purposes. Considering the necessity of the knowledge on P-gp properties of drugs to optimize the use of herbal remedies, the predictions of our models represent a valuable resource for generation of useful hypothesis and experimental validation of novel P-gp inhibitors.

Limitations and Future Work. Although our framework makes encouraging identification, it has some limitations and shares common weakness associated with classification task. First, because of the lack of a standard data set for model building, our models were developed on the data coming from various literature sources. This may be a limitation for the combining data because the data from different literatures were determined in different assays, resulting in the risk of erroneous judgment. Zdrazil et al.[44] suggested that a more robust combining data set could be achieved by using a tailored threshold for every assay, so such a large classification data set, covering diverse structural molecules, is still encouraged to yield more reliable predictions. Second, although large-scale chemical 2D structural descriptors were calculated, 2D structurally similar compounds may have quite different shapes in 3D structure. This means that it may be not good enough to build more predictive models based on only 2D structural

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

descriptors. Broccatelli et al.[82] developed a competitive model using molecular interaction fields that capture 3D structure features. Therefore, incorporating both descriptor features will be of interest for the classification task in future. Third, our two-step feature selection method is a wrapper based approach, which is computationally expensive. Moreover, to achieve the stable ranks of the *n* most frequent descriptors in the first step, multiple runs of GA are required. This could help to decide how many descriptors to be selected for the next step. However, the exact number of runs is not easy to decide. More runs may result in a more stable and different descriptor ranking at a cost of more computation time. We are planning to design reranking strategy to reduce the computation time and to improve stability of the selected descriptors in future. On the other hand, the selected descriptor subset largely depends on the selection of the classifier. Our results demonstrate that the feature selection method coupled with different classifiers can achieve different solutions. There was no clear advantage of one over the other. The detailed relationship among these solutions remains unclear and needs to be further investigated to describe structural characteristics of P-gp properties. Fourth, AD problem addressed by ML based approach was proved to be more appropriate for our practical use, while the selection of descriptors for AD analysis may be a limitation. Max et al.[48] recommend that the AD model could be built on a subset or the whole of original predictors. Since we have reported three individual and one ensemble models, it is possible to question our AD approach of using the union of the individual predictor sets. Building AD models on each predictor set may allow more discrimination. The point is that the proposed approach achieved substantially good discrimination between more reliable and less reliable predictions. We are not claiming that the approach present here is necessarily optimal. One can achieve better discrimination for specific moleculars with specific models by the other AD methodologies. Finally, although we have shown the impact of different probability cutoffs of both prediction and AD on the performance, we have determined cutoffs heuristically for the screening application. This may be a limitation because different cutoffs may alter the number of potential inhibitors identified in the database. As mentioned previously, this may result in a different inhibitor-rich herb set. Besides, the exact spectrum of compounds of most herbs is not defined. Our enrichment results are thus limited to both the current database and the proposed approach. Nevertheless, *Q*-values can be used as a criteria to rank the extent of inhibitor-rich herbs. On the other hand, the models presented here are "preliminary screening" tools. Therefore, a positive prediction by the models may imply a strong potential inhibitor and needs to be further confirmed in future.

CONCLUSIONS

The in silico prediction of P-gp inhibitors has been a challenge for drug designers. Although many models have been developed during the past decade, there have been very few strictly validated models with defined AD that were developed using a large number of chemically diverse compounds. In the present work, we have compiled the extensive available data set of P-gp inhibitors and noninhibitors extracted from various literature sources. This work is motivated by the belief that the predictive performance can be affected by the different combination of predictors and modeling techniques. For this reason, our work has explored 18 classifiers and employed an efficient two-step feature selection approach to characterize the descriptors that are the most relevant for the prediction. As a result, we have achieved the simpler and more interpretable models that provide a competitive and in some cases better performance compared with existing models. The top three best-performing individual models (FDA, SVM, and RanForest) and their ensemble model were reported. The experiment results have indicated that our informative descriptors and models worked well in the discrimination between inhibitors and noninhibitors. The best model proposed here achieves an overall accuracy of 86.73% for the training set using a 10-fold CV procedure, an overall accuracy of 85.47% for the test set, and an overall accuracy of 79.94% for external validation set. Besides, the extensive validation by DrugBank database further confirmed our predictive results. Using ML based AD approach results in a more thoroughly defined domain compared to AMBIT approach for our problem. Inside-AD compounds are more likely to be correctly predicted than OutSide-AD ones. The definition of AD has made our models more practical and applicable. The initial analysis of the TCMSP database including a total of 13 051 unique compounds from 498 herbs revealed that TCM compounds have more diverse chemical characteristics and may be a promising source to develop novel P-gp inhibitors. Then virtual screenings of 11 556 Inside-AD herbal compounds in TCMSP database were carried out. The proposed models identified 875 potential P-gp inhibitors and 15 inhibitor-rich herbs. The obtained results for the presented virtual screening of some possible candidates were also externally supported by the literature. Because of the limited publication data, we anticipate that our models will be used as helpful tools for hypothesis-driven experimental studies on novel P-gp inhibitors. The results of virtual screening are valuable not only to identify potential inhibitors for drug development from TCM, but also to address some broader issues, which are of interest in herbal remedies. In addition, the resources (data sets, algorithms, and virtual screening predictions) in this project are freely available for public use and improvement.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.molpharmaceut.5b00465.

All molecules and their structures; descriptors calculated in the present work; skewness of Box-Cox transformation for 28 continuous descriptors; statistics of continuous/binary descriptors for both data sets; corrected *p*-values for each comparison between classifiers by a Wilcoxon signed rank test; ranking information on our optimal descriptors by the rank-based feature selection methods; classification performance with different cutoff values of AD probability; predictions by the developed models for Inside-AD compounds in TCMSP database; fitness evolution in different classifiers; accuracy curves of different rank-based feature selection methods for each classifier; probability response curves for the optimal descriptors; distributions of continuous optimal descriptors for the combination of TCMSP database and modeling data set (PDF)

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

AUTHOR INFORMATION

Corresponding Authors

*E-mail: wxh6020@163.com.

*E-mail: anruimw@126.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors are grateful to the anonymous reviewers and the editors for their helpful comments and suggestions, which substantially improved the quality of this paper. This study was supported by the Innovation Program of Shanghai Municipal Education Commission (15ZZ066) and Budget Project of Shanghai Municipal Education Commission (2014YSN08).

REFERENCES

(1) Hennessy, M.; Spiers, J. P. A primer on the mechanics of P-glycoprotein the multidrug transporter. *Pharmacol. Res.* **2007**, *55* (1), 1−15.

(2) Srivalli, K. M. R.; Lakshmi, P. K. Overview of P-glycoprotein inhibitors: a rational outlook. *Braz. J. Pharm. Sci.* **2012**, *48* (3), 353−367.

(3) Han, H. K. Role of Transporters in Drug Interactions. *Arch. Pharmacal Res.* **2011**, *34* (11), 1865−1877.

(4) Yang, Z.; Wang, J. R.; Niu, T.; Gao, S.; Yin, T.; You, M.; Jiang, Z. H.; Hu, M. Inhibition of P-glycoprotein leads to improved oral bioavailability of compound K, an anticancer metabolite of red ginseng extract produced by gut microflora. *Drug Metab. Dispos.* **2012**, *40* (8), 1538−44.

(5) Nooter, K.; Sonneveld, P. Multidrug-Resistance (Mdr) Genes in Hematological Malignancies. *Cytotechnology* **1993**, *12* (1−3), 213−230.

(6) Sun, J.; Luo, C.; Wang, Y. J.; He, Z. G. The holistic 3M modality of drug delivery nanosystems for cancer therapy. *Nanoscale* **2013**, *5* (3), 845−859.

(7) Hu, Z. P.; Yang, X. X.; Ho, P. C. L.; Chan, S. Y.; Heng, P. W. S.; Chan, E.; Duan, W.; Koh, H. L.; Zhou, S. F. Herb-drug interactions - A literature review. *Drugs* **2005**, *65* (9), 1239−1282.

(8) Marchetti, S.; Mazzanti, R.; Beijnen, J. H.; Schellens, J. H. M. Concise review: Clinical relevance of drug-drug and herb-drug interactions mediated by the ABC transporter ABCB1 (MDR1, P-glycoprotein). *Oncologist* **2007**, *12* (8), 927−941.

(9) Liu, J. Y.; Lee, K. F.; Sze, C. W.; Tong, Y.; Tang, S. C. W.; Ng, T. B.; Zhang, Y. B. Intestinal absorption and bioavailability of traditional Chinese medicines: a review of recent experimental progress and implication for quality control. *J. Pharm. Pharmacol.* **2013**, *65* (5), 621−633.

(10) Hermann, R.; von Richter, O. Clinical Evidence of Herbal Drugs As Perpetrators of Pharmacokinetic Drug Interactions. *Planta Med.* **2012**, *78* (13), 1458−1477.

(11) Zhou, S. F.; Lim, L. Y.; Chowbay, B. Herbal modulation of P-glycoprotein. *Drug Metab. Rev.* **2004**, *36* (1), 57−104.

(12) Chieli, E.; Romiti, N. Kidney proximal human tubule HK-2 cell line as a tool for the investigation of P-glycoprotein modulation by natural compounds. *B Latinoam Caribe Pl* **2008**, *7* (6), 281−294.

(13) Sun, M.; Xu, X. L.; Lu, Q. H.; Pan, Q. R.; Hu, X. Schisandrin B: A dual inhibitor of P-glycoprotein and multidrug resistance-associated protein 1. *Cancer Lett.* **2007**, *246* (1−2), 300−307.

(14) Liao, Z. G.; Liang, X. L.; Zhu, J. Y.; Zhao, G. W.; Guan, Y. M.; Cao, Y. C.; Zhao, L. J. Transport Properties of Puerarin and Effect of Extract of Radix Angelicae dahuricae on Puerarin Intestinal Absorption Using In Situ and In Vitro Models. *Phytother. Res.* **2014**, *28* (9), 1288−1294.

(15) Liang, Y.; Zhou, Y. Y.; Zhang, J. W.; Rao, T.; Zhou, L. J.; Xing, R.; Wang, Q.; Fu, H. X.; Hao, K.; Xie, L.; Wang, G. J. Pharmacokinetic Compatibility of Ginsenosides and Schisandra Lignans in Shengmaisan: From the Perspective of P-Glycoprotein. *PLoS One* **2014**, *9* (6), e98717.

(16) Li, W. D.; Zhang, B. D.; Wei, R.; Liu, J. H.; Lin, Z. B. Reversal effect of Ganoderma lucidum polysaccharide on multidrug resistance in K562/ADM cell line. *Acta Pharmacol. Sin.* **2008**, *29* (5), 620−7.

(17) Li, L.; Pan, Q.; Han, W.; Liu, Z.; Li, L.; Hu, X. Schisandrin B prevents doxorubicin-induced cardiotoxicity via enhancing glutathione redox cycling. *Clin. Cancer Res.* **2007**, *13* (22), 6753−6760.

(18) Huang, M.; Jin, J.; Sun, H.; Liu, G. T. Reversal of P-glycoprotein-mediated multidrug resistance of cancer cells by five schizandrins isolated from the Chinese herb Fructus Schizandrae. *Cancer Chemother. Pharmacol.* **2008**, *62* (6), 1015−1026.

(19) Zhou, S. F.; He, J. L.; Zhou, Z. W.; Yin, J. J.; He, C. Q.; Yu, Y. Schisandra chinensis regulates drug metabolizing enzymes and drug transporters via activation of Nrf2-mediated signaling pathway. *Drug Des., Dev. Ther.* **2014**, *9*, 127−146.

(20) Fan, L.; Mao, X. Q.; Tao, G. Y.; Wang, G.; Jiang, F.; Chen, Y.; Li, Q.; Zhang, W.; Lei, H. P.; Hu, D. L.; Huang, Y. F.; Wang, D.; Zhou, H. H. Effect of Schisandra chinensis extract and Ginkgo biloba extract on the pharmacokinetics of talinolol in healthy volunteers. *Xenobiotica* **2009**, *39* (3), 249−254.

(21) Chula, S.; Hang, L.; Yinying, B.; Jianning, S.; Shi, R. The effects of notoginsenoside R(1) on the intestinal absorption of geniposide by the everted rat gut sac model. *J. Ethnopharmacol.* **2012**, *142* (1), 136−43.

(22) Zhang, J. W.; Zhou, F.; Wu, X. L.; Gu, Y.; Ai, H.; Zheng, Y. T.; Li, Y. N.; Zhang, X. X.; Hao, G.; Sun, J. G.; Peng, Y.; Wang, G. J. 20(S)-Ginsenoside Rh2 Noncompetitively Inhibits P-Glycoprotein In Vitro and In Vivo: A Case for Herb-Drug Interactions. *Drug Metab. Dispos.* **2010**, *38* (12), 2179−2187.

(23) Yi, H. J.; Oh, J. H.; Lee, Y. J. Absence of Drug Interaction Between Hwang-Ryun-Hae-Dok-Tang and Phenolsulfonphthalein. *Arch. Pharmacal Res.* **2010**, *33* (12), 2025−2031.

(24) Yang, J. M.; Ip, S. P.; Xian, Y. F.; Zhao, M.; Lin, Z. X.; Yeung, J. H. K.; Chan, R. C. Y.; Lee, S. S.; Che, C. T. Impact of the Herbal Medicine Sophora flavescens on the Oral Pharmacokinetics of Indinavir in Rats: The Involvement of CYP3A and P-Glycoprotein. *PLoS One* **2012**, *7* (2), e31312.

(25) Meijerman, I.; Beijnen, J. H.; Schellens, J. H. M. Herb-drug interactions in oncology: Focus on mechanisms of induction. *Oncologist* **2006**, *11* (7), 742−752.

(26) Mazzari, A. L. D. A.; Prieto, J. M. Herbal medicines in Brazil: pharmacokinetic profile and potential herb-drug interactions. *Front. Pharmacol.* **2014**, *5*, 162.

(27) Hu, M.; Wang, D. Q.; Xiao, Y. J.; Mak, V. W. L.; Tomlinson, B. Herb-Drug Interactions: Methods to Identify Potential Influence of Genetic Variations in Genes Encoding Drug Metabolizing Enzymes and Drug Transporters. *Curr. Pharm. Biotechnol.* **2012**, *13* (9), 1718−1730.

(28) Hu, M.; Fan, L.; Zhou, H. H.; Tomlinson, B. Theranostics meets traditional Chinese medicine: rational prediction of drug-herb interactions. *Expert Rev. Mol. Diagn.* **2012**, *12* (8), 815−830.

(29) Gurley, B. J. Pharmacokinetic Herb-Drug Interactions (Part 1): Origins, Mechanisms, and the Impact of Botanical Dietary Supplements. *Planta Med.* **2012**, *78* (13), 1478−1489.

(30) Gouws, C.; Steyn, D.; Du Plessis, L.; Steenekamp, J.; Hamman, J. H. Combination therapy of Western drugs and herbal medicines: recent advances in understanding interactions involving metabolism and efflux. *Expert Opin. Drug Metab. Toxicol.* **2012**, *8* (8), 973−984.

(31) Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Li, Z. R.; Han, L. Y.; Lin, H. H.; Chen, Y. Z. Machine learning approaches for predicting compounds that interact with therapeutic and ADMET mated proteins. *J. Pharm. Sci.* **2007**, *96* (11), 2838−2860.

(32) Reis, M.; Ferreira, R. J.; Serly, J.; Duarte, N.; Madureira, A. M.; Santos, D. J. V. A.; Molnar, J.; Ferreira, M. J. U. Colon Adenocarcinoma Multidrug Resistance Reverted by Euphorbia Diterpenes: Structure-Activity Relationships and Pharmacophore Modeling. *Anti-Cancer Agents Med. Chem.* **2012**, *12* (9), 1015−1024.

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

(33) Langer, T.; Eder, M.; Hoffmann, R. D.; Chiba, P.; Ecker, G. F. Lead identification for modulators of multidrug resistance based on in silico screening with a pharmacophoric feature model. *Arch. Pharm.* **2004**, *337* (6), 317−327.
(34) Tan, W.; Mei, H.; Chao, L.; Liu, T. F.; Pan, X. C.; Shu, M.; Yang, L. Combined QSAR and molecule docking studies on predicting P-glycoprotein inhibitors. *J. Comput.-Aided Mol. Des.* **2013**, *27* (12), 1067−1073.
(35) Kothandan, G.; Gadhe, C. G.; Madhavan, T.; Choi, C. H.; Cho, S. J. Docking and 3D-QSAR (quantitative structure activity relationship) studies of flavones, the potent inhibitors of p-glycoprotein targeting the nucleotide binding domain. *Eur. J. Med. Chem.* **2011**, *46* (9), 4078−4088.
(36) Ghandadi, M.; Shayanfar, A.; Hamzeh-Mivehroud, M.; Jouyban, A. Quantitative structure activity relationship and docking studies of imidazole-based derivatives as P-glycoprotein inhibitors. *Med. Chem. Res.* **2014**, *23* (11), 4700−4712.
(37) Wang, Y. H.; Li, Y.; Yang, S. L.; Yang, L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J. Chem. Inf. Model.* **2005**, *45* (3), 750−757.
(38) Poongavanam, V.; Haider, N.; Ecker, G. F. Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors. *Bioorg. Med. Chem.* **2012**, *20* (18), 5388−5395.
(39) Klepsch, F.; Vasanthanathan, P.; Ecker, G. F. Ligand and Structure-Based Classification Models for Prediction of P-Glycoprotein Inhibitors. *J. Chem. Inf. Model.* **2014**, *54* (1), 218−229.
(40) Chen, L.; Li, Y. Y.; Zhao, Q.; Peng, H.; Hou, T. J. ADME Evaluation in Drug Discovery. 10. Predictions of P-Glycoprotein Inhibitors Using Recursive Partitioning and Naive Bayesian Classification Techniques. *Mol. Pharmaceutics* **2011**, *8* (3), 889−900.
(41) Crivori, P.; Reinach, B.; Pezzetta, D.; Poggesi, I. Computational models for identifying potential P-glycoprotein substrates and inhibitors. *Mol. Pharmaceutics* **2006**, *3* (1), 33−44.
(42) Sahlin, U. Uncertainty in QSAR Predictions. *Atla-Altern Lab Anim* **2013**, *41* (1), 111−125.
(43) Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T. I.; Cruciani, G. A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. *J. Med. Chem.* **2011**, *54* (6), 1740−51.
(44) Zdrazil, B.; Pinto, M.; Vasanthanathan, P.; Williams, A. J.; Balderud, L. Z.; Engkvist, O.; Chichester, C.; Hersey, A.; Overington, J. P.; Ecker, G. F. Annotating Human P-Glycoprotein Bioassay Data. *Mol. Inf.* **2012**, *31* (8), 599−609.
(45) Oprisiu, I.; Novotarskyi, S.; Tetko, I. V. Modeling of non-additive mixture properties using the Online CHEmical database and Modeling environment (OCHEM). *J. Cheminf.* **2013**, *5* (1), 4.
(46) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011**, *25* (6), 533−54.
(47) Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466−74.
(48) Max, Kuhn; Johnson, K. *Applied Predictive Modeling*; Springer: New York, 2013.
(49) Alsberg, B. K.; Winson, M. K.; Kell, D. B. Improving the interpretation of multivariate and rule induction models by using a peak parameter representation. *Chemom. Intell. Lab. Syst.* **1997**, *36* (2), 95−109.
(50) Goodarzi, M.; Dejaegher, B.; Heyden, Y. v. Feature Selection Methods in QSAR Studies. *J. AOAC Int.* **2012**, *95* (3), 636.
(51) Pourbasheer, E.; Aalizadeh, R.; Ardabili, J. S.; Ganjali, M. R. QSPR study on solubility of some fullerenes derivatives using the genetic algorithms - Multiple linear regression. *J. Mol. Liq.* **2015**, *204*, 162−169.
(52) Chen, M. *Advances in Greedy Algorithms*; In-Teh: Austria, 2008.
(53) Leardi, R. Application of genetic algorithm-PLS for feature selection in spectral data sets. *J. Chemom.* **2000**, *14* (5−6), 643−655.
(54) Jalali-Heravi, M.; Kyani, A. Application of genetic algorithm-kernel partial least square as a novel nonlinear feature selection method: Activity of carbonic anhydrase II inhibitors. *Eur. J. Med. Chem.* **2007**, *42* (5), 649−659.
(55) Guo, J. M.; White, J.; Wang, G. X.; Li, J.; Wang, Y. L. A genetic algorithm for optimized feature selection with resource constraints in software product lines. *J. Syst. Software* **2011**, *84* (12), 2208−2221.
(56) Cho, H. W.; Kim, S. B.; Jeong, M. K.; Park, Y.; Ziegler, T. R.; Jones, D. P. Genetic algorithm-based feature selection in high-resolution NMR spectra. *Expert Syst. Appl.* **2008**, *35* (3), 967−975.
(57) Trevino, V.; Falciani, F. GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics* **2006**, *22* (9), 1154−1156.
(58) Hastie, T.; Tibshirani, R.; J, F. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: California, 2008.
(59) Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab − An S4 Package for Kernel Methods in R. *J. Stat. Soft.* **2004**, *11* (9), 1−20.
(60) Kursa, M. B. rFerns: An Implementation of the Random Ferns Method for General-Purpose Machine Learning. *J. Stat Softw* **2014**, *61* (10), 1−13.
(61) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S.*, 4th ed.; Springer: New York, 2002.
(62) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2* (3), 18−22.
(63) Kuhn, M.; Weston, S.; Coulter, N.; Culp, M. *C5.0 Decision Trees and Rule-Based Models*, 2015. http://cran.r-project.org/web/packages/C50/ (accessed Mar 9, 2015).
(64) Hornik, K.; Buchta, C.; Zeileis, A. Open-source machine learning: R meets Weka. *Computation Stat* **2009**, *24* (2), 225−232.
(65) Todorov, V.; Filzmoser, P. An Object-Oriented Framework for Robust Multivariate Analysis. *J. Stat. Softw.* **2009**, *32* (3), 1−47.
(66) Therneau, T.; Atkinson, B.; Ripley, B. *Recursive Partitioning and Regression Trees*, 2015. http://cran.r-project.org/web/packages/rpart/ (accessed Feb 11, 2015).
(67) Mevik, B.-H.; Wehrens, R.; Liland, K. H. *Partial Least Squares and Principal Component regression*, 2015. http://cran.r-project.org/web/packages/pls/ (accessed Aug 22, 2015).
(68) Weihs, C.; Ligges, U.; Luebke, K.; Raabe, N. *klaR Analyzing German Business Cycles*; Baier, D., Decker, R., Thieme, L. S., Eds.; Springer: Berlin, 2005.
(69) Wehrens, R.; Buydens, L. M. C. Self- and super-organizing maps in R: The kohonen package. *J. Stat. Softw.* **2007**, *21* (5), 1−19.
(70) Huang, G. B.; Zhou, H.; Ding, X.; Zhang, R. Extreme Learning Machine for Regression and Multiclass Classification. *Ieee T Syst. Man Cy B* **2012**, *42* (2), 513−529.
(71) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Atla-Altern Lab Anim* **2005**, *33* (5), 445−459.
(72) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17* (5), 4791−4810.
(73) Jeliazkova, N.; Jeliazkov, V. AMBIT RESTful web services: an implementation of the OpenTox application programming interface. *J. Cheminf.* **2011**, *3*, 18.
(74) Ru, J. L.; Li, P.; Wang, J. N.; Zhou, W.; Li, B. H.; Huang, C.; Li, P. D.; Guo, Z. H.; Tao, W. Y.; Yang, Y. F.; Xu, X.; Li, Y.; Wang, Y. H.; Yang, L. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J. Cheminf.* **2014**, *6*, 13.

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

(75) Vimaleswaran, K. S.; Tachmazidou, I.; Zhao, J. H.; Hirschhorn, J. N.; Dudbridge, F.; Loos, R. J. F. Candidate genes for obesity-susceptibility show enriched association within a large genome-wide association study for BMI. *Hum. Mol. Genet.* **2012**, *21* (20), 4537−4542.
(76) Romanski, R.; Kotthoff, L. *FSelector: Selecting attributes*, 2014. http://cran.r-project.org/web/packages/FSelector/ (accessed Oct 25, 2014).
(77) Robnik-Sikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn* **2003**, *53* (1−2), 23−69.
(78) Peng, H. C.; Long, F. H.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *Ieee T Pattern Anal* **2005**, *27* (8), 1226−1238.
(79) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y. F.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B. S.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **2014**, *42* (D1), D1091−D1097.
(80) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More Than 96,000 Compounds. *J. Pharm. Sci.* **2009**, *98* (3), 861−893.
(81) Todeschini, R.; Consonni, V. *Molecular descriptors for chemoinformatics*; WILEY-VCH: Weinheim, Germany, 2009.
(82) Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T. I.; Cruciani, G. A Novel Approach for Predicting P-Glycoprotein (ABCB1) Inhibition Using Molecular Interaction Fields. *J. Med. Chem.* **2011**, *54* (6), 1740−1751.
(83) Sun, H. M. A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* **2005**, *48* (12), 4031−4039.
(84) Pan, X. C.; Chao, L.; Tan, W.; Yang, L.; Podraza, R.; Mei, H. Emerging chemical patterns applied to prediction of P-glycoprotein inhibitors. *Chemom. Intell. Lab. Syst.* **2014**, *137*, 140−145.
(85) Wang, Y.; Backman, T. W. H.; Horan, K.; Girke, T. fmcsR: mismatch tolerant maximum common substructure searching in R. *Bioinformatics* **2013**, *29* (21), 2792−2794.
(86) Schmidt, M.; Polednik, C.; Roller, J.; Hagen, R. Cytotoxicity of herbal extracts used for treatment of prostatic disease on head and neck carcinoma cell lines and non-malignant primary mucosal cells. *Oncol. Rep.* **2012**, *29* (2), 628−36.
(87) Sadava, D.; Still, D. W.; Mudry, R. R.; Kane, S. E. Effect of Ganoderma on drug-sensitive and multidrug-resistant small-cell lung carcinoma cells. *Cancer Lett.* **2009**, *277* (2), 182−189.
(88) Jiang, D. H. *The MDR reversal effect of Ganoderma Lucidum extracts on human gastric cancer SGC7901/ADR cells*. M.S. Thesis, Shaanxi Normal University, 2012.
(89) Rubis, B.; Polrolniczak, A.; Knula, H.; Potapinska, O.; Kaczmarek, M.; Rybczynska, M. Phytosterols in physiological concentrations target multidrug resistant cancer cells. *Med. Chem.* **2010**, *6* (4), 184−90.
(90) Jiang, Z. J.; Jin, T. T.; Gao, F.; Liu, J. W.; Zhong, J. J.; Zhao, H. Effects of Ganoderic acid Me on inhibiting multidrug resistance and inducing apoptosis in multidrug resistant colon cancer cells. *Process Biochem.* **2011**, *46* (6), 1307−1314.
(91) Eid, S. Y.; El-Readi, M. Z.; Eldin, E. E.; Fatani, S. H.; Wink, M. Influence of combinations of digitonin with selected phenolics, terpenoids, and alkaloids on the expression and activity of P-glycoprotein in leukaemia and colon cancer cells. *Phytomedicine* **2013**, *21* (1), 47−61.
(92) Zhao, Y.; Bu, L.; Yan, H.; Jia, W. 20S-Protopanaxadiol Inhibits P-Glycoprotein in Multidrug Resistant Cancer Cells. *Planta Med.* **2009**, *75* (10), 1124−1128.
(93) Pokharel, Y. R.; Kim, N. D.; Han, H. K.; Oh, W. K.; Kang, K. W. Increased Ubiquitination of Multidrug Resistance 1 by Ginsenoside Rd. *Nutr. Cancer* **2010**, *62* (2), 252−259.
(94) Qin, X. L.; Chen, X.; Wang, Y.; Xue, X. P.; Wang, Y.; Li, J. L.; Wang, X. D.; Zhong, G. P.; Wang, C. X.; Yang, H.; Huang, M.; Bi, H. C. In Vivo to In Vitro Effects of Six Bioactive Lignans of Wuzhi Tablet (Schisandra Sphenanthera Extract) on the CYP3A/P-glycoprotein-Mediated Absorption and Metabolism of Tacrolimus. *Drug Metab. Dispos.* **2014**, *42* (1), 193−199.
(95) Pan, Q. R.; Lu, Q. H.; Zhang, K.; Hu, X. Dibenzocyclooctadiene lingnans: a class of novel inhibitors of P-glycoprotein. *Cancer Chemother. Pharmacol.* **2006**, *58* (1), 99−106.
(96) Liang, Y.; Zhou, Y. Y.; Zhang, J. W.; Liu, Y. N.; Guan, T. Y.; Wang, Y.; Xing, L.; Rao, T.; Zhou, L. J.; Hao, K.; Xie, L.; Wang, G. J. In vitro to in vivo evidence of the inhibitor characteristics of Schisandra lignans toward P-glycoprotein. *Phytomedicine* **2013**, *20* (11), 1030−1038.
(97) Jin, J.; Bi, H. C.; Hu, J. Q.; Zeng, H.; Zhong, G. P.; Zhao, L. Z.; Huang, Z. Y.; Huang, M. Effect of Wuzhi Tablet (Schisandra sphenanthera extract) on the Pharmacokinetics of Paclitaxel in Rats. *Phytother. Res.* **2011**, *25* (8), 1250−1253.
(98) Sun, Y. B.; Li, G. F.; Tang, Z. K.; Wu, B. Y. [Modulation on the P-glycoprotein in the jejunum by combined use of Glycyrrhiza inflata and Kansui]. *Yao Xue Xue Bao* **2010**, *45* (4), 510−6.
(99) Barile, E.; Borriello, M.; Di Pietro, A.; Doreau, A.; Fattorusso, C.; Fattorusso, E.; Lanzotti, V. Discovery of a new series of jatrophane and lathyrane diterpenes as potent and specific P-glycoprotein modulators. *Org. Biomol. Chem.* **2008**, *6* (10), 1756−1762.
(100) Shi, J.; Cao, B.; Zha, W. B.; Wu, X. L.; Liu, L. S.; Xiao, W. J.; Gu, R. R.; Sun, R. B.; Yu, X. Y.; Zheng, T.; Li, M. J.; Wang, X. W.; Zhou, J.; Mao, Y.; Ge, C.; Ma, T.; Xia, W. J.; Aa, J. Y.; Wang, G. J.; Liu, C. X. Pharmacokinetic interactions between 20(S)-ginsenoside Rh2 and the HIV protease inhibitor ritonavir in vitro and in vivo. *Acta Pharmacol. Sin.* **2013**, *34* (10), 1349−1358.
(101) Yang, L. Q.; Wang, B.; Gan, H.; Fu, S. T.; Zhu, X. X.; Wu, Z. N.; Zhan, D. W.; Gu, R. L.; Dou, G. F.; Meng, Z. Y. Enhanced oral bioavailability and anti-tumour effect of paclitaxel by 20(s)-ginsenoside Rg3 in vivo. *Biopharm. Drug Dispos.* **2012**, *33* (8), 425−436.
(102) Kim, S. W.; Kwon, H.; Chi, D. W.; Shim, J. H.; Park, J. D.; Lee, Y. H.; Pyo, S.; Rhee, D. K. Reversal of P-glycoprotein-mediated multidrug resistance by ginsenoside Rg(3). *Biochem. Pharmacol.* **2003**, *65* (1), 75−82.
(103) Wang, W.; Wu, X.; Wang, L.; Meng, Q.; Liu, W. Stereoselective property of 20(S)-protopanaxadiol ocotillol type epimers affects its absorption and also the inhibition of P-glycoprotein. *PLoS One* **2014**, *9* (6), e98887.
(104) Zhu, N.; Rafi, M. M.; DiPaola, R. S.; Xin, J.; Chin, C. K.; Badmaev, V.; Ghai, G.; Rosen, R. T.; Ho, C. T. Bioactive constituents from gum guggul (Commiphora wightii). *Phytochemistry* **2001**, *56* (7), 723−7.
(105) Nabekura, T.; Yamaki, T.; Ueno, K.; Kitagawa, S. Effects of plant sterols on human multidrug transporters ABCB1 and ABCC1. *Biochem. Biophys. Res. Commun.* **2008**, *369* (2), 363−8.
(106) Park, J. D.; Kim, D. S.; Kwon, H. Y.; Son, S. K.; Lee, Y. H.; Baek, N. I.; Kim, S. I.; Rhee, D. K. Effects of ginseng saponin on modulation of multidrug resistance. *Arch. Pharmacal Res.* **1996**, *19* (3), 213−218.
(107) Fong, W. F.; Wang, C.; Zhu, G. Y.; Leung, C. H.; Yang, M. S.; Cheung, H. Y. Reversal of multidrug resistance in cancer cells by Rhizoma Alismatis extract. *Phytomedicine* **2007**, *14* (2−3), 160−5.
(108) Wang, C.; Zhang, J. X.; Shen, X. L.; Wan, C. K.; Tse, A. K. W.; Fong, W. F. Reversal of P-glycoprotein-mediated multidrug resistance by Alisol B 23-acetate. *Biochem. Pharmacol.* **2004**, *68* (5), 843−855.
(109) Kim, H. G.; Cho, J. H.; Yoo, S. R.; Lee, J. S.; Han, J. M.; Lee, N. H.; Ahn, Y. C.; Son, C. G. Antifatigue Effects of Panax ginseng CA Meyer: A Randomised, Double-Blind, Placebo-Controlled Trial. *PLoS One* **2013**, *8* (4), e61271.
(110) Kang, S.; Min, H. Ginseng, the 'Immunity Boost': The Effects of Panax ginseng on Immune System. *J. Ginseng Res.* **2012**, *36* (4), 354−368.
(111) Ji, Y. X.; Rao, Z. Y.; Cui, J. L.; Bao, H. Y.; Chen, C. Y.; Shu, C. Y.; Gong, J. R. Ginsenosides Extracted from Nanoscale Chinese White Ginseng Enhances Anticancer Effect. *J. Nanosci. Nanotechnol.* **2012**, *12* (8), 6163−6167.

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713

(112) Lee, S. D.; Yoo, G.; Chae, H. J.; In, M. J.; Oh, N. S.; Hwang, Y. K.; Hwang, W. I.; Kim, D. C. Lipid-Soluble Extracts as the Main Source of Anticancer Activity in Ginseng and Ginseng Marc. *J. Am. Oil Chem. Soc.* **2009**, *86* (11), 1065−1071.

DOI: 10.1021/acs.molpharmaceut.5b00465
Mol. Pharmaceutics 2015, 12, 3691−3713