molecular pharmaceutics

Article

pubs.acs.org/molecularpharmaceutics

ADMET Evaluation in Drug Discovery. 16. Predicting hERG Blockers by Combining Multiple Pharmacophores and Machine Learning Approaches

Shuangquan Wang,[†] Huiyong Sun,[†] Hui Liu,[†] Dan Li,[†] Youyong Li,[‡] and Tingjun Hou*[,†,§]

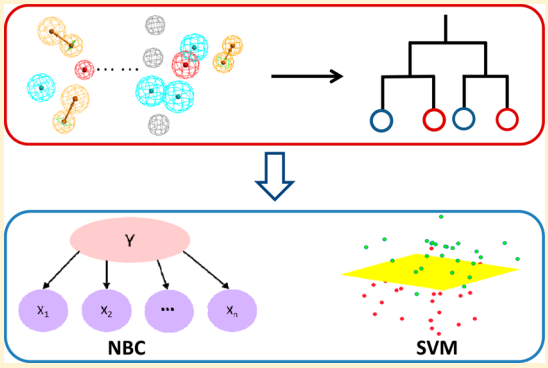
[†]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China

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

[‡]Institute of Functional Nano & Soft Materials (FUNSOM), Soochow University, Suzhou, Jiangsu 215123, China

Supporting Information

ABSTRACT: Blockade of human ether-à-go-go related gene (hERG) channel by compounds may lead to drug-induced QT prolongation, arrhythmia, and Torsades de Pointes (TdP), and therefore reliable prediction of hERG liability in the early stages of drug design is quite important to reduce the risk of cardiotoxicity-related attritions in the later development stages. In this study, pharmacophore modeling and machine learning approaches were combined to construct classification models to distinguish hERG active from inactive compounds based on a diverse data set. First, an optimal ensemble of pharmacophore hypotheses that had good capability to differentiate hERG active from inactive compounds was identified by the recursive partitioning (RP) approach. Then, the naive Bayesian classification (NBC) and support vector machine (SVM) approaches were employed to construct classification models by integrating multiple important pharmacophore hypotheses. The integrated classification models showed improved predictive capability over any single pharmacophore hypothesis, suggesting that the broad binding polyspecificity of hERG can only be well characterized by multiple pharmacophores. The best SVM model achieved the prediction accuracies of 84.7% for the training set and 82.1% for the external test set. Notably, the accuracies for the hERG blockers and nonblockers in the test set reached 83.6% and 78.2%, respectively. Analysis of significant pharmacophores helps to understand the multimechanisms of action of hERG blockers. We believe that the combination of pharmacophore modeling and SVM is a powerful strategy to develop reliable theoretical models for the prediction of potential hERG liability.



KEYWORDS: *hERG, ADMET, pharmacophore, machine learning, support vector machine, recursive partitioning, SVM*

INTRODUCTION

During cardiac depolarization and repolarization, a voltage-gated potassium channel encoded by the human ether-à-go-go related gene (hERG or Kv11.1) plays a major role in the regulation of the exchange of cardiac action potential and resting potential.[1,2] The structure of the hERG channel is a homotetramer, and each subunit contains six transmembrane helices (S1−S6). Some critical residues for the binding of hERG blockers, such as Tyr652, Phe656, and V659 located in S6, have been confirmed by a series of mutagenesis studies,[3−5] but unfortunately, the whole crystal structure of the hERG channel has not been solved currently.

The hERG blockade may cause long QT syndrome (LQTS), arrhythmia, and Torsade de Pointes (TdP), which lead to palpitations, fainting, or even sudden death.[6−8] However, to date, several important drugs, such as terfenadine, astemizole, cisapride, vardenafil, and ziprasidone, have been withdrawn from the market or severely restricted in availability due to their undesirable hERG-related cardiotoxicity.[5,9,10] Therefore, assessment of hERG-related cardiotoxicity has become an important step in the drug design/discovery pipeline.[6,11]

As the fact that hERG assays and QT animal studies are time-consuming and expensive, development of reliable and robust *in silico* models to predict potential hERG liability becomes quite important. In the past decade, a wide range of quantitative structure−activity relationship (QSAR) models for hERG liability have been reported using various machine learning approaches, such as *k*-nearest neighbor algorithm (kNN), artificial neural networks (ANN), random forest (RF), support vector machine (SVM), self-organizing mapping (SOM), recursive partitioning (RP), genetic algorithm (GA), naive Bayesian classification (NBC), etc.[5,12−17] A majority of the QSAR models are classifiers while only a few of them are quantitative regression models. Although most QSAR models showed satisfactory predictions for the training sets, these models were usually developed based on relatively small data sets, and thus their chemical coverage and extrapolability are limited. Meanwhile, because hERG blockade can be measured by various experimental techniques, it is possible

Received: May 26, 2016
Accepted: July 5, 2016
Published: July 5, 2016

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

that some available data sets are not reliable when the data from different experimental sources or protocols are mixed together. Moreover, most available QSAR models cannot be interpreted easily due to the complexity of machine learning approaches. Alternatively, structure-based approaches, including homology modeling, molecular docking, molecular dynamics (MD) simulation, and free energy calculation, have been employed to study hERG−drug interactions.[18] All structure-based studies were performed on the homology model because the hERG channel has not yet been crystallized. Unfortunately, the sequence similarity between hERG and the templates (KscA, KvaP, MthK, Kir2.2, and Kv1.2) is quite low,[5] and therefore the homology model of hERG may not be reliable and thus the receptor-based modeling for hERG blockade is interpretative rather than predictive.

To date, pharmacophore modeling has been employed to predict hERG blockade.[5,16] For example, Ekins et al. established a pharmacophore hypothesis for 15 hERG blockers, which contained four hydrophobic features and one positive ionizable moiety.[19] Aronov proposed two five-point pharmacophore hypotheses based on a data set of 194 uncharged molecules, and the two hypotheses could correctly classify 78% and 69% of the hERG blockers in the data set.[20] Leong constructed three pharmacophore hypotheses based on 26 molecules, and then developed an SVM regression model by integrating the three hypotheses of the model to predict hERG liability, which exhibited an r^2 value of 0.94 for the 13 tested molecules.[21] Garg and co-workers developed a number of pharmacophore hypotheses based on a training set of 44 molecules, but the best hypothesis with three features did not show acceptable predictive capability to the 12 tested molecules.[22] Durdagi et al. created 44 pharmacophore hypotheses based on 31 hERG blockers with diverse binding affinities, and the best model showed an average deviation of 0.29 (pIC_{50}) for the 14 hERG blockers in the test set.[23] Tan et al. proposed 47 pharmacophore models based on a training set of 53 molecules, whereas the five most representative models with high accuracy to the training set (r^2 = 0.914−0.941) did not show acceptable performance to the test set (q^2 = 0.280−0.370).[24] Kratz et al. developed seven complementary pharmacophore models by using Catalyst and LigandScout. The best Catalyst model could identify 14 out of 18 blockers in the training set and 8 out of 19 blockers in the test set, and six LigandScout models could identify 84.9% of the true blockers (ranging from 8 to 30% for individual models) with few false positives.[25] In summary, the pharmacophore hypotheses reported in the previous studies were usually generated by a limited number of molecules. Moreover, compared with traditional QSAR models, the pharmacophore models for hERG blockade usually exhibit worse predictive power. It appears to be hard to establish a common pharmacophore hypothesis to distinguish hERG blockers from nonblockers successfully.

It is well-known that many proteins relevant to ADMET (absorption, distribution, metabolism, excretion, and toxicity), such as hERG, usually have large and flexible binding pockets, and they can recognize and bind a variety of structurally diverse compounds.[26,27] Such complicated binding patterns between ADMET-related proteins and their ligands cannot be well characterized by a single pharmacophore hypothesis, and then the combination of multiple pharmacophore hypotheses, also referred to as pharmacophore ensemble, may be an effective strategy to give reliable prediction for some ADMET properties.[28−31]

In this study, in order to characterize the broad binding polyspecificity of hERG, a large number of pharmacophore hypotheses were established, and the important hypotheses were identified by Information Gain in the RP approach. Subsequently, the NBC and SVM approaches were employed to develop the classifiers by integrating multiple important pharmacophore hypotheses (Figure 1). Furthermore, the complicated interaction mechanisms of hERG blockage were discussed by the analysis of the important pharmacophores. By combining multiple pharmacophore hypotheses, we tried to develop a reliable and fast platform for the prediction of hERG liability and gain profound insights into the multimechanisms of action of hERG blockade.

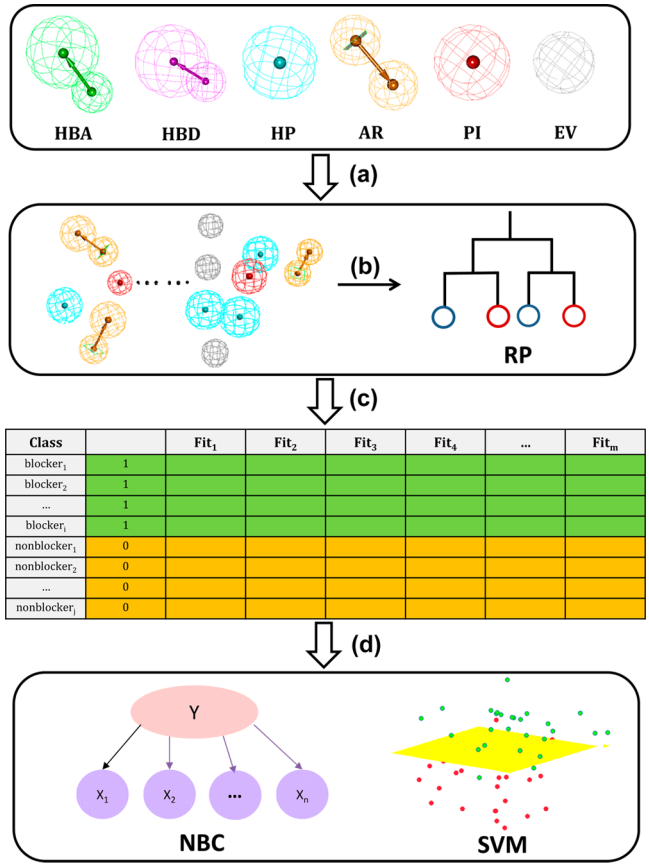


Figure 1. Pipeline of the procedure to build the NBC and SVM classifiers by integrating multiple pharmacophore hypotheses identified by PR. (a) A number of pharmacophore hypotheses were generated; (b) the molecules in the training set were mapped on the pharmacophore hypotheses, and decision trees were established by PR based on the fit values of mapping; (c) the data matrix was generated based on the fit values given by the important pharmacophores chosen by RP; (d) the classifiers of hERG blockade were generated by NBC and SVM.

MATERIALS AND METHODS

Data Set Preparation. The total data set for model construction and validation contains 587 molecules, including 527 molecules with experimental hERG blocking bioactivities (IC_{50}) and 60 hERG nonblockers without IC_{50} derived from our previous study.[32] The IC_{50} activities were mainly determined based on mammalian cell lines, such as HEK, CHO, and COS. IC_{50} data derived from a nonmammalian cell line, XO (*Xenopus laevis* oocytes), was included into the data set when no mammalian cell line data was available. If a molecule has multiple different experimental values, it would be compared with the

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

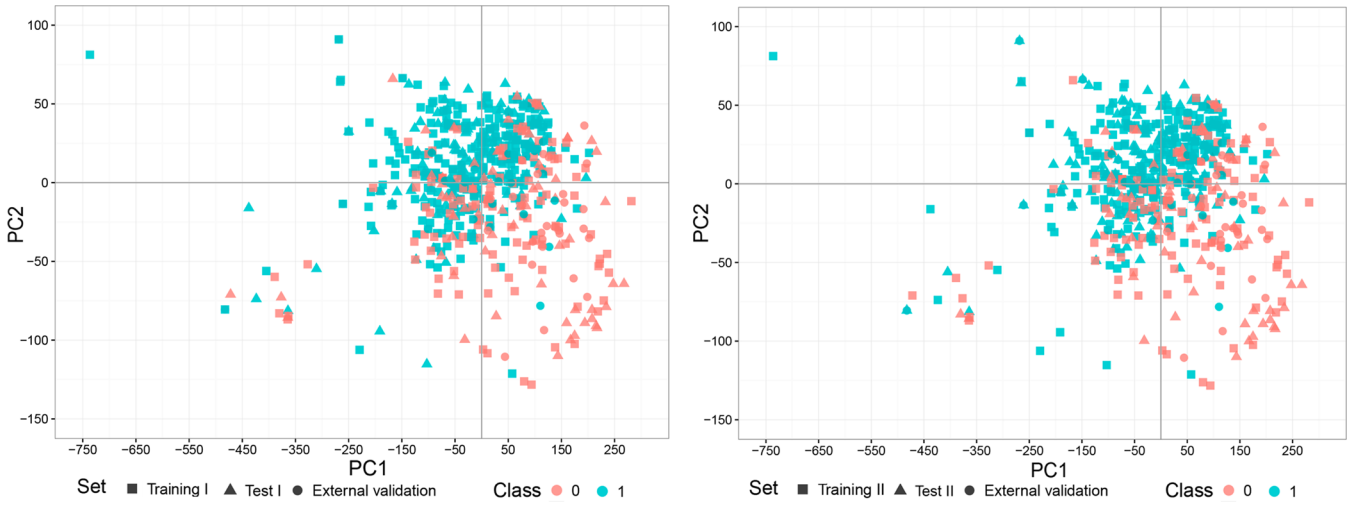


Figure 2. Chemical space distributions based on PCA for (a) training set I, test set I and External validation set and (b) training set II, test set II and External validation set.

Table 1. Performance of the RP Models for the Training and Test Sets

label	depth	model	TP	FN	FP	TN	SE	SP	PRE1	PRE2	GA	MCC
training I	2	RP1	242	41	27	82	0.855	0.752	0.900	0.667	0.827	0.586
	3	RP2	249	34	27	82	0.880	0.752	0.902	0.707	0.844	0.621
	4	RP3	233	50	15	94	0.823	0.862	0.940	0.653	0.834	0.637
	5	RP4	224	59	7	102	0.792	0.936	0.970	0.634	0.832	0.662
	6	RP5	235	48	6	103	0.830	0.945	0.975	0.682	0.862	0.714
	7	RP6	241	42	7	102	0.852	0.936	0.972	0.708	0.875	0.732
training II	2	RP7	233	39	33	87	0.857	0.725	0.876	0.690	0.816	0.574
	3	RP8	254	18	37	83	0.934	0.692	0.873	0.822	0.860	0.659
	4	RP9	248	24	34	86	0.912	0.717	0.879	0.782	0.852	0.645
	5	RP10	219	53	14	106	0.805	0.883	0.940	0.667	0.829	0.646
	6	RP11	232	40	11	109	0.853	0.908	0.955	0.732	0.870	0.723
	7	RP12	240	32	13	107	0.882	0.892	0.949	0.770	0.885	0.746
test I	2	RP1	103	26	16	50	0.798	0.758	0.866	0.658	0.785	0.539
	3	RP2	104	25	17	49	0.806	0.742	0.860	0.662	0.785	0.535
	4	RP3	96	33	14	52	0.744	0.788	0.873	0.612	0.759	0.508
	5	RP4	91	38	14	52	0.705	0.788	0.867	0.578	0.733	0.468
	6	RP5	93	36	14	52	0.721	0.788	0.869	0.591	0.744	0.484
	7	RP6	93	36	14	52	0.721	0.788	0.869	0.591	0.744	0.484
test II	2	RP7	112	28	14	41	0.800	0.745	0.889	0.594	0.785	0.513
	3	RP8	116	24	17	38	0.829	0.691	0.872	0.613	0.790	0.502
	4	RP9	114	26	16	39	0.814	0.709	0.877	0.600	0.785	0.500
	5	RP10	90	50	10	45	0.643	0.818	0.900	0.474	0.692	0.415
	6	RP11	96	44	12	43	0.686	0.782	0.889	0.494	0.713	0.423
	7	RP12	101	39	12	43	0.721	0.782	0.894	0.524	0.738	0.459
external validation	2	RP1	32	7	7	22	0.821	0.759	0.821	0.759	0.794	0.579
	3	RP2	33	6	7	22	0.846	0.759	0.825	0.786	0.809	0.608
	4	RP3	30	9	7	22	0.769	0.759	0.811	0.710	0.765	0.524
	5	RP4	29	10	5	24	0.744	0.828	0.853	0.706	0.779	0.565
	6	RP5	31	8	4	25	0.795	0.862	0.886	0.758	0.824	0.650
	7	RP6	30	9	6	23	0.769	0.793	0.833	0.719	0.779	0.557
external validation	2	RP7	29	10	6	23	0.744	0.793	0.829	0.697	0.765	0.531
	3	RP8	32	7	6	23	0.821	0.793	0.842	0.767	0.809	0.611
	4	RP9	32	7	7	22	0.821	0.759	0.821	0.759	0.794	0.579
	5	RP10	26	13	5	24	0.667	0.828	0.839	0.649	0.735	0.491
	6	RP11	29	10	4	25	0.744	0.862	0.879	0.714	0.794	0.599
	7	RP12	31	8	6	23	0.795	0.793	0.838	0.742	0.794	0.584

entry found in the PubChem's BioAssay database and the consistent experimental value was adopted.[33] Moreover, a threshold of 40 μM was used to define hERG blockers and nonblockers, where molecules with IC_{50} < 40 μM were regarded as blockers and others were regarded as nonblockers.[34,35] The purpose of this study is to develop qualitative classifiers rather than quantitative

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

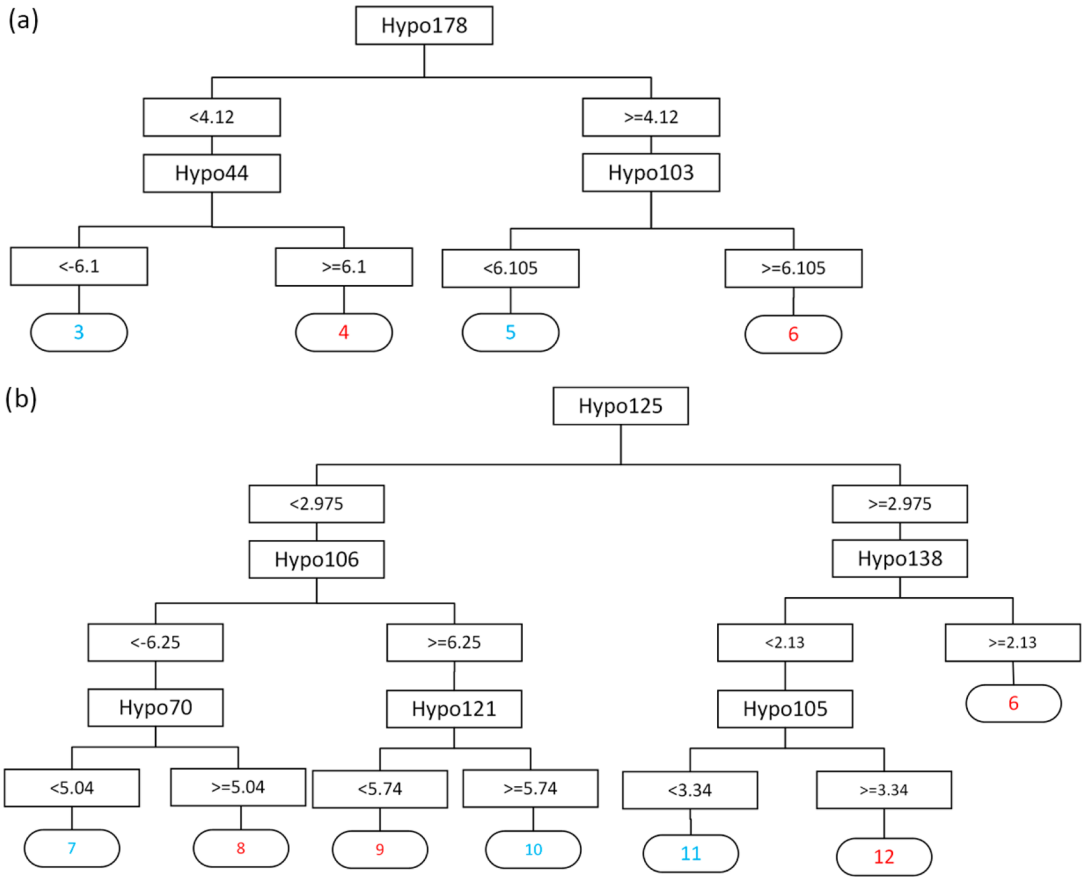


Figure 3. (a) The decision tree at a tree depth of 2 established based on training set I, and (b) the decision three at a tree depth of 3 established based on training set II. The predicted blockers are labeled in red and nonblockers in blue.

regression models, and therefore some noise of the experimental data may be tolerated.

The whole data set was divided into training set (392) and test set (195) with the ratio of 2:1. The 352 molecules with IC_{50} values and 40 nonblockers without IC_{50} in the training set were extracted from the whole data set by using the *Find Diverse Molecules* protocol in Discovery Studio 2.5 (DS 2.5), respectively,[36] which guarantees that the selected molecules in the training set have the largest diversity evaluated by the Tanimoto coefficients based on the LCFP_8 fingerprints and log *P*.[32,37] The pipeline of the data set partition is shown in Figure S1. To give a comparison, 352 bioactive molecules and 40 nonblockers were randomly selected from the whole data set to construct the training set, and the remaining molecules were used as the test set. The two training sets constructed by structural diversity and random selection were referred to as training sets I and II, and the corresponding test sets were named test sets I and II. Additionally, to validate the prediction performance of models, 68 molecules reported by Li et al. were employed as the external validation set.[35]

Pharmacophore Generation. Pharmacophore hypotheses were derived by using the *3D QSAR Pharmacophore Generation* protocol in DS2.5.[36,38] In the generation of pharmacophore hypotheses, IC_{50} was converted to pIC_{50}. In order to generate diverse pharmacophore hypotheses to characterize the complicated binding patterns of the hERG blockers, the following strategy was employed. First, 40 molecules with experimental hERG activities were randomly chosen from the training set, and they were used to generate pharmacophore hypotheses. The above pharmacophore building process was repeated 200 times, and then diverse pharmacophore hypotheses could be generated.

The low-energy conformations of each molecule were generated by the FAST algorithm implemented in the *Conformation Generation protocol* in DS2.5.[36] The maximum number of conformations per molecule was limited to 255, and the maximum energy threshold was set to 20 kcal/mol. The maximum number of features in each hypothesis was set to 5, and five types of pharmacophoric features were used, including hydrogen-bond acceptor (HBA), hydrogen-bond donor (HBD), hydrophobic center (HP), positive ionizable center (PI), and ring aromatic center (RA). In addition, the excluded volume (EV) feature was used to describe the shape of the binding pocket of hERG, and the maximum number of EV was set to 10. The other parameters for pharmacophore generation were set to the default values.

Recursive Partitioning Analysis. For the training set, 200 randomly selected subsets were generated. It should be noted that the pharmacophore hypotheses could not be successfully generated for some certain subsets, and finally 190 and 189 pharmacophore hypotheses were generated for the training sets I and II, respectively. Each hypothesis was labeled by a serial number from 1 to 200 according to the number of the corresponding subset for pharmacophore generation. Each molecule in the training set was mapped onto each pharmacophore by using the *Ligand Pharmacophore Mapping protocol* in DS2.5,[36] and the match of the mapping was measured by a fit score, which can determine how well a molecule is mapped onto a pharmacophore. The higher the fit score, the better the match.

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

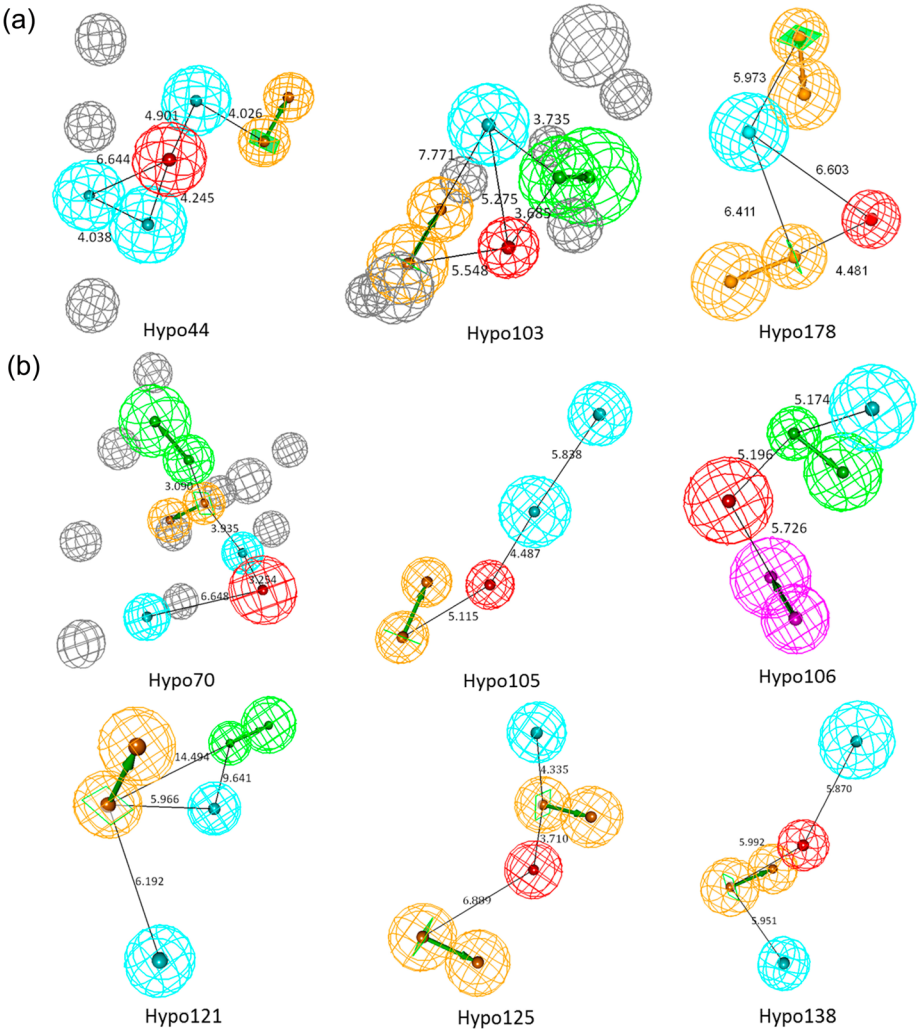


Figure 4. Pharmacophore hypotheses in (a) the decision tree shown in Figure 3a and (b) the decision tree shown in Figure 3b. The HBA, HBD, HP, PI, AR, and EV pharmacophoric features are colored in green, magenta, cyan, red, orange, and gray, respectively.

If a molecule could not be mapped onto a pharmacophore successfully, a penalty of −1 was assigned. In total, 190 and 189 fit scores were generated for each molecule in the training sets I and II, and they were used as the independent variables (X). The response variable Y was set to 1 for blockers and 0 for nonblockers.

Then, the RP approach was employed to develop classifiers to distinguish the blockers from nonblockers.[32] RP creates a decision tree that strives to continuously classify molecules by splitting them into subsets based on independent properties and terminates when the stopping criteria are satisfied. The decision trees were created by the *Create Recursive Partitioning Model* protocol in DS2.5.[36] The minimum samples per node was set to 7. The *Gini* index was adopted by the split method in RP models. The 5-fold cross-validation was used to evaluate the statistical significance of each model, and the other parameters were set to the default values. In order to evaluate the impact of the maximum tree depth, different RP models were built by changing the tree depth from 2 to 7 (tree depth higher than 7 was not considered because higher level did not change the decision tree any more).

Integration of Multiple Pharmacophores by SVM and NBC. The decision tree splits the compounds into different subsets by grouping those that can be mapped onto specific pharmacophore hypotheses. The pharmacophores chosen by a decision tree are more significant than the others. Moreover, the important pharmacophores chosen by a decision tree are independent of each other, and thus a single tree may account for the multimechanisms of action of hERG blockers.

Then, the fit values given by these important pharmacophores were used as the independent variables for the SVM and NBC calculations. The effectiveness of SVM and NBC for binary classification has been extensively validated by our previous studies.[32,39−45] SVM is a popular statistical method for classification and regression. The theory of SVM is based on the structural risk minimization (SRM) principle that constructs a hyperplane with the largest distances to the nearest training data points of any class. The detailed description of SVM can be found in previous studies.[44,46] Here, the radial basis function (RBF) was used as the kernel function for the classification. Grid searching was used to optimize two parameters, cost (c) and gamma (γ), which were designed to exponentially grow against 2, namely 2^n, which goes from −15 to 15 and −15 to 5, respectively, with a step of 0.5. Because the numbers of blockers and nonblockers are not balanced, a higher weight ($k\_$) of 2 was applied to the nonblocker class. The statistical significance of each model was tested by 5-fold cross-validation. The SVM models were developed by using the e1071 package in R.[47]

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

Table 2. Performance of Naive Bayesian Classifiers for the Training and Test Sets[a]

label	level	model	SE	SP	PRE1	PRE2	GA	MCC	AUC
training I	2	NBC1	0.859	0.706	0.884	0.658	0.816	0.553	0.783
	3	NBC2	0.777	0.771	0.898	0.571	0.776	0.507	0.778
	4	NBC3	0.774	0.780	0.901	0.570	0.776	0.511	0.769
	5	NBC4	0.887	0.651	0.869	0.689	0.821	0.548	0.761
	6	NBC5	0.866	0.670	0.872	0.658	0.811	0.532	0.772
training II	2	NBC6	0.838	0.733	0.877	0.667	0.806	0.557	0.799
	3	NBC7	0.904	0.617	0.842	0.740	0.816	0.551	0.779
	4	NBC8	0.912	0.608	0.841	0.753	0.819	0.555	0.790
	5	NBC9	0.835	0.742	0.880	0.664	0.806	0.560	0.791
	6	NBC10	0.857	0.733	0.879	0.693	0.819	0.581	0.793
test I	2	NBC1	0.822	0.758	0.869	0.685	0.800	0.566	0.847
	3	NBC2	0.775	0.788	0.877	0.642	0.779	0.541	0.821
	4	NBC3	0.814	0.742	0.861	0.671	0.790	0.544	0.825
	5	NBC4	0.822	0.727	0.855	0.676	0.790	0.540	0.819
	6	NBC5	0.829	0.697	0.843	0.676	0.785	0.523	0.820
test II	2	NBC6	0.743	0.764	0.889	0.538	0.749	0.465	0.820
	3	NBC7	0.807	0.764	0.897	0.609	0.795	0.537	0.811
	4	NBC8	0.779	0.764	0.893	0.575	0.774	0.504	0.813
	5	NBC9	0.814	0.800	0.912	0.629	0.810	0.576	0.829
	6	NBC10	0.800	0.800	0.911	0.611	0.800	0.559	0.819
external validation	2	NBC1	0.846	0.724	0.805	0.778	0.794	0.576	0.815
	3	NBC2	0.846	0.759	0.825	0.786	0.809	0.608	0.850
	4	NBC3	0.846	0.759	0.825	0.786	0.809	0.608	0.855
	5	NBC4	0.821	0.759	0.821	0.759	0.794	0.579	0.865
	6	NBC5	0.795	0.724	0.795	0.724	0.765	0.519	0.865
external validation	2	NBC6	0.769	0.828	0.857	0.727	0.794	0.591	0.832
	3	NBC7	0.795	0.793	0.838	0.742	0.794	0.584	0.839
	4	NBC8	0.795	0.862	0.886	0.758	0.824	0.650	0.848
	5	NBC9	0.795	0.724	0.795	0.724	0.765	0.519	0.834
	6	NBC10	0.795	0.759	0.816	0.733	0.779	0.551	0.826

[a]Levels of 2, 3, 4, 5, and 6 represent the models established based on the pharmacophore hypotheses chosen by the decision trees at depths of 2, 3, 4, 5, and 6, respectively.

NBC is a probabilistic classification method based on Bayes's theorem. Bayesian statistics not only considers the likelihood of a model but also takes into consideration the complexity of the model. Compared with other machine learning approaches, NBC can deal with large-scale data, trains quickly, and is tolerant of random noise. The mathematical details of NBC and the procedure to create a naive Bayesian classifier were described previously.[16,39,40,48] The number of bins to divide continuous variables was set to 10, and the other parameters were set to the default values. 5-fold cross-validation was used to assess the statistical significance of each model. The NBC classifiers were created by the *Create Bayesian Model* protocol in DS2.5.[36]

Validation of Classifiers. Each classifier was evaluated by the following parameters: true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity [SE = TP/(TP + FN)], specificity [SP = TN/(TN + FP)], prediction accuracy for blockers [PRE1 = TP/(TP + FP)], prediction accuracy for nonblockers [PRE2 = TN/(TN + FN)], global accuracy (GA, eq 1), and Matthews correlation coefficient (MCC, eq 2):

$$GA = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$MMC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(FN + TN)(TN + FP)}} \quad (2)$$

The above parameters were calculated for both training and test sets.

RESULTS AND DISCUSSION

The distribution of the experimental hERG blocking activities (pIC_{50}) is shown in Figure S2. Figure 2 shows the chemical spaces of the training and test sets characterized by the scattered distributions of the first and second principal components derived from the principal component analysis (PCA) based on eight molecular descriptors. These descriptors are extensively used in ADMET predictions,[49,50] including octanol−water partition coefficient (log *P*), number of rotatable bonds, number of rings, topological polar surface area (TPSA), number of H-bond acceptors, number of H-bond donors, number of heavy atoms, and molecular weight. The cumulative proportion of the first two components can explain >99% of the variance. As shown in Figure 2, the chemical spaces of the test sets are roughly distributed within the scope of the training sets, and therefore it is reliable to predict the hERG liability for the test sets by using the classifiers built from the training sets.

Important Pharmacophores Identified by RP. For the training set, the fit values of mapping provided by the 190 or 189 pharmacophore hypotheses were used as the input matrix for RP to establish decision trees. Compared with the "blind operations" of many other machine learning approaches, the results of RP can be easily converted to a series of simple hierarchical rules. The pharmacophores as the splitting descriptors for a decision

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

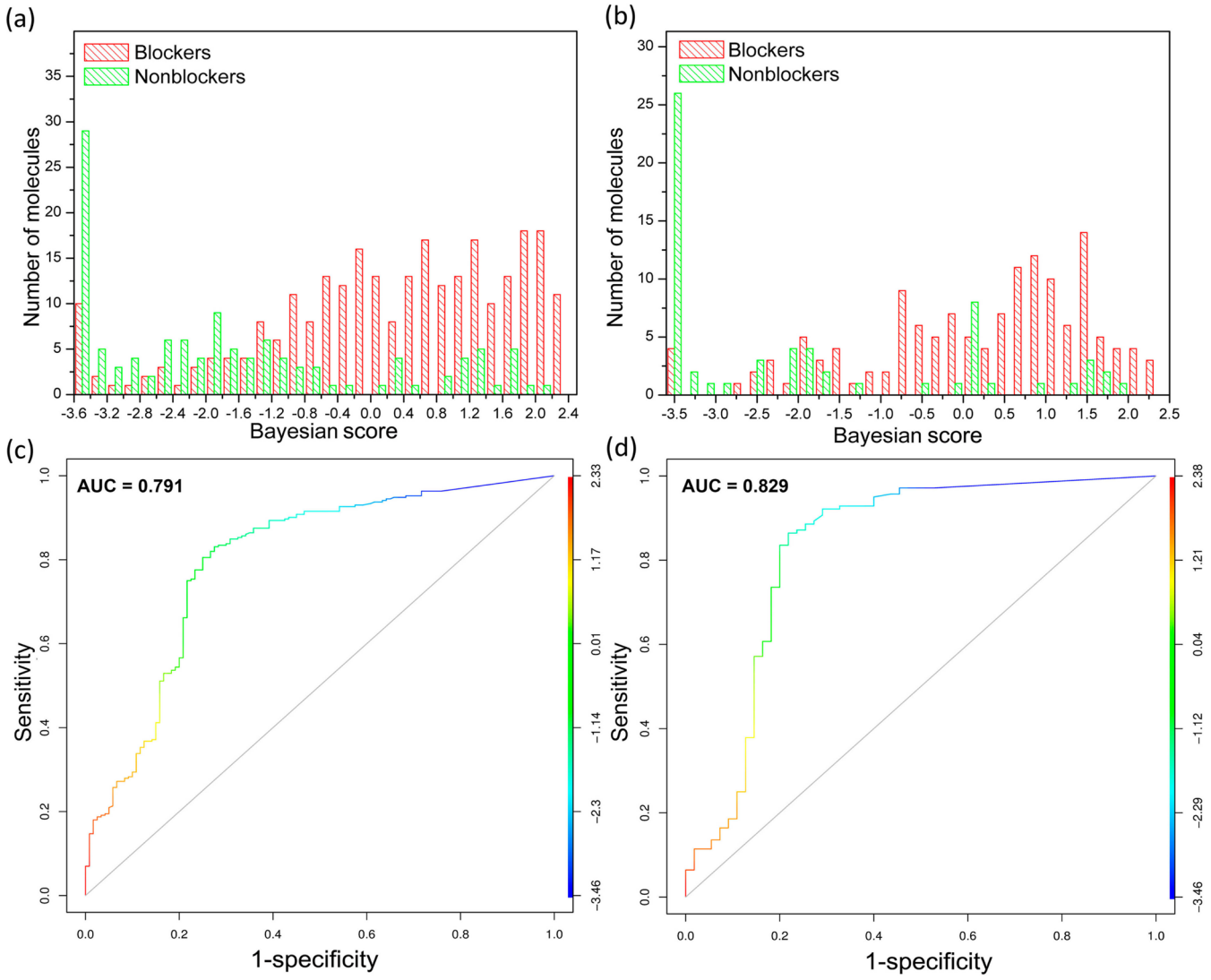


Figure 5. Distributions of the Bayesian scores predicted by the Bayesian classifier based on the pharmacophores chosen by the decision three at a tree depth of 5 for (a) training set II and (b) test set II. The ROC curves for (c) training set II and (d) test set II. The color of the right stripe displays the different cutoff to split the blockers and nonblockers.

tree should be more significant than the others. Moreover, during the RP modeling, the tree depth was increased from 2 to 7, and the prediction accuracies of the established models at different tree depths were compared according to their prediction accuracies to the test sets.

The prediction results of the RP models are summarized in Table 1. For training set I, the RP model at a tree depth of 2 (RP1 model) does not achieve the highest prediction accuracy for training set I (82.7%) but it achieve the highest accuracy for test set I (78.5%). The decision tree established based on training set I at a tree depth of 2 is depicted in Figure 3a. For training set II, the RP model at a tree depth of 3 (RP8 model) reaches the highest prediction accuracy for test set II (79%). The decision tree established based on training set II at a tree depth of 3 is shown in Figure 3b. It can be observed that the prediction accuracies for the test sets decrease slightly at the higher three depths (5−7).

Based on the theory of RP, the pharmacophores chosen by a decision tree have the higher weights to split the blockers and nonblockers. The important pharmacophores chosen by the decision tree at a tree depth of 2 for training set I include Hypo44, Hypo103, and Hypo178 (Figure 4a), and those chosen by the decision tree at a tree depth of 3 for training set II include Hypo70, Hypo105, Hypo106, Hypo121, Hypo125, and Hypo138 (Figure 4b). Apparently, two chemical features, including HP and RA, can be found in all of the nine important pharmacophores, highlighting the importance of the hydrophobic interactions upon the binding of hERG blockers. This observation is consistent with the experimental mutagenesis data, which suggests that Tyr652 and Phe656 in the S6 helix can form favorable hydrophobic interactions with high affinity blockers.[19,21,23,51]

As shown in Figure 3, the most important discriminant pharmacophores are Hypo178 and Hypo125. Both of these two pharmacophores have four chemical features, including two aromatic rings (RA), a hydrophobic center (HP), and a positive ionizable center (PI). These chemical features were also observed in most pharmacophore models for hERG blockers reported in previous studies.[19,21,25] So far, the published pharmacophore hypotheses typically have the following common features: (1) hydrophobic (HP) and/or aromatic centers (RA), which can form hydrophobic and/or $\pi-\pi$ stacking interactions with Tyr652 and Phe656 of hERG; (2) a protonated nitrogen (PI), which is

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

Table 3. Performance of SVM Classifiers for the Training and Test Sets[a]

label	level	model	SE	SP	PRE1	PRE2	GA	MCC	AUC
training I	2	SVM1	0.883	0.670	0.874	0.689	0.824	0.558	0.871
	3	SVM2	0.922	0.587	0.853	0.744	0.829	0.552	0.867
	4	SVM3	0.940	0.578	0.853	0.788	0.839	0.576	0.879
	5	SVM4	0.936	0.578	0.852	0.778	0.837	0.569	0.885
	6	SVM5	0.943	0.596	0.859	0.802	0.847	0.597	0.899
training II	2	SVM6	0.882	0.675	0.86	0.717	0.819	0.567	0.799
	3	SVM7	0.897	0.708	0.875	0.752	0.839	0.616	0.872
	4	SVM8	0.897	0.708	0.875	0.752	0.839	0.616	0.876
	5	SVM9	0.901	0.700	0.872	0.757	0.839	0.615	0.898
	6	SVM10	0.897	0.717	0.878	0.754	0.842	0.623	0.898
test I	2	SVM1	0.860	0.697	0.847	0.719	0.805	0.562	0.832
	3	SVM2	0.891	0.652	0.833	0.754	0.810	0.565	0.819
	4	SVM3	0.891	0.621	0.821	0.745	0.800	0.539	0.813
	5	SVM4	0.915	0.621	0.825	0.788	0.815	0.573	0.838
	6	SVM5	0.907	0.652	0.836	0.782	0.821	0.587	0.842
test II	2	SVM6	0.843	0.709	0.881	0.639	0.805	0.536	0.759
	3	SVM7	0.836	0.727	0.886	0.635	0.805	0.542	0.785
	4	SVM8	0.829	0.745	0.892	0.631	0.805	0.548	0.82
	5	SVM9	0.850	0.745	0.895	0.661	0.821	0.575	0.839
	6	SVM10	0.843	0.745	0.894	0.651	0.815	0.566	0.845
external validation	2	SVM1	0.872	0.724	0.810	0.808	0.809	0.606	0.769
	3	SVM2	0.872	0.690	0.791	0.800	0.794	0.576	0.821
	4	SVM3	0.897	0.690	0.795	0.833	0.809	0.608	0.836
	5	SVM4	0.872	0.690	0.791	0.800	0.794	0.576	0.805
	6	SVM5	0.897	0.690	0.795	0.833	0.809	0.608	0.821
external validation	2	SVM6	0.795	0.759	0.816	0.733	0.779	0.551	0.773
	3	SVM7	0.846	0.724	0.805	0.778	0.794	0.576	0.813
	4	SVM8	0.846	0.690	0.786	0.769	0.779	0.545	0.814
	5	SVM9	0.846	0.724	0.805	0.778	0.794	0.576	0.824
	6	SVM10	0.846	0.724	0.805	0.778	0.794	0.576	0.828

[a]Levels of 2, 3, 4, 5, and 6 represent the models established based on the pharmacophore hypotheses chosen by the decision trees at depths of 2, 3, 4, 5, and 6, respectively.

important for achieving high hERG potency but not necessary to all blockers.[16,20,24] For training set I, 296 molecules could be successfully mapped onto Hypo178, while 241 molecules could be mapped onto Hypo125 for training set II. However, the blocking activities of a number of molecules could not be well predicted by the fit values given by pharmacophore mapping. For example, the predicted activity of amsacrine given by Hypo178 has a high deviation ($\Delta pIC_{50} = 6.54$) from the experimental data. Apparently, a single pharmacophore hypothesis cannot effectively distinguish blockers from nonblockers, and multiple pharmacophore hypotheses are necessary for the development of reliable hERG prediction models.

NBC and SVM Classifiers Based on Multiple Pharmacophores. In order to develop more reliable classification models for hERG liability, two powerful machine learning approaches (NBC and SVM) were used to develop classifiers based on multiple pharmacophores. The fit values given by the important pharmacophores chosen by RP were used to generate the data matrix for model construction. First, NBC was employed to integrate the predictions given by the multiple pharmacophores chosen by RP at five different tree depths (2−6). The performances of the Bayesian classifiers are summarized in Table 2. For training set II, based on the predictions from the 12 pharmacophores chosen by the decision tree at a tree depth of 5 (Figure S3), the Bayesian classifier (NBC9 model) achieves the best prediction accuracy with the global accuracy of 0.81, sensitivity of 0.814, specificity of 0.80, MCC of 0.576, and AUC of 0.829 for test set II. The Bayesian scores and receiver operating characteristic (ROC) curves given by NBC9 for training set II and test set II are shown in Figure 5. The histograms show that the blockers tend to have more positive values, while the nonblockers tend to have more negative values. For training set I, based on the three pharmacophores chosen by the decision tree at a tree depth of 2 (Figure 3a), the Bayesian classifier (NBC1 model) has the global accuracy of 0.8, sensitivity of 0.822, specificity of 0.758, MCC of 0.566, and AUC of 0.847 for test set I. Based on the same training and test sets, the performances of the best Bayesian classifiers are only slightly better than those of the RP classifiers. The theory of naive Bayesian supposes that probability distribution obeys conditional independence assumption, that is to say, the input variables are conditionally independent, but the predictions given by some similar pharmacophores generated based on a set of similar molecules are possibly relevant to each other. Therefore, some relevant input variables may reduce the global accuracy of NBC. Moreover, data unbalance may have unfavorable impact on the reliability and robustness of the NBC models. For all the Bayesian classifiers, the prediction accuracies for the nonblockers are lower than those for the blockers, which may be attributed to the low number of the nonblockers in the training set.

Similarly, SVM was employed to integrate the predictions given by the important pharmacophores chosen by RP. Compared with NBC, SVM has intrinsic advantage to avoid the conditional independence problem involved in NBC. The performance of the

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

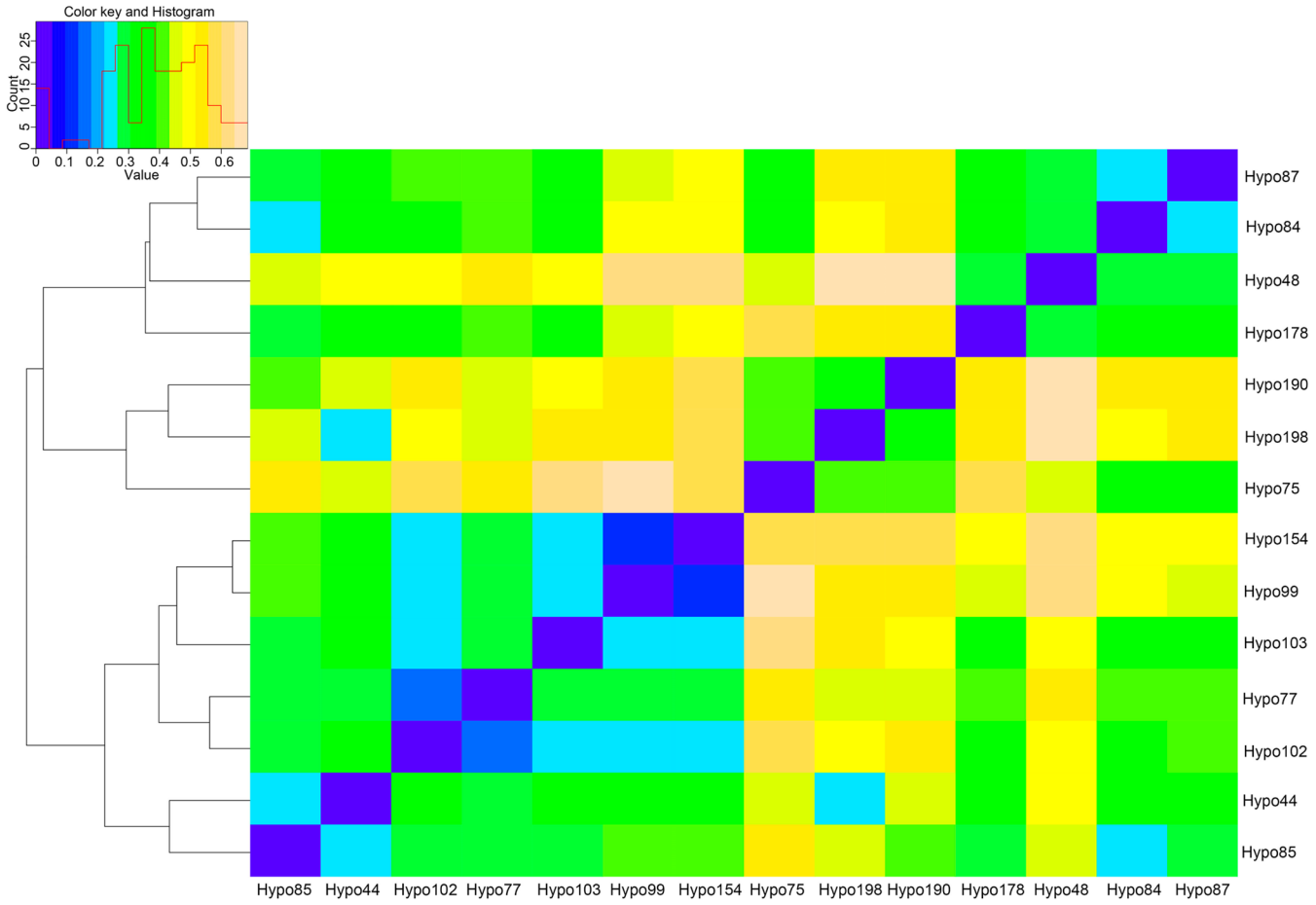


Figure 6. Hierarchical clustering of the 14 pharmacophores in the SVM5 model established based on training set I.

SVM classifiers is summarized in Table 3. For training set I, the best SVM classifier (SVM5 model; cost = 1 and gamma = 0.0174) based on the decision tree at a tree depth of 6 has a sensitivity of 0.943, specificity of 0.596, MCC of 0.597, and global accuracy of 0.847 for training set I, and a sensitivity of 0.907, specificity of 0.652, MCC of 0.587, and global accuracy of 0.821 for test set I, and a global accuracy of 0.809 for the external validation set. For training set II, the best SVM classifier based on the decision tree at a tree depth of 5 (SVM9 model) has a sensitivity of 0.85, specificity of 0.745, MCC of 0.575, and global accuracy of 0.821 for test set II. The prediction accuracies of the best SVM classifiers are obviously better than those of the best Bayesian classifiers.

By analyzing the average performance of the three different machine learning approaches for test set I and test set II, SVM exhibits better predictive capability than the other two approaches (Figure S4a). The performance of the SVM models based on the pharmacophore hypotheses chosen by the decision trees with different tree depths was then discussed. With the increase of the decision tree depth from 2 to 6, the overall prediction power improves significantly (Figure S4b–d). To all the models shown in Tables 2 and 3, the MCC and SP values are relatively low, which may be explained by the following reasons. First, the unbalanced nature of blockers versus nonblockers has great impact on MCC, where the MCC values would decrease dramatically with the increase of false positives or false negatives.[52] Second, the method of pharmacophore mapping may be more suitable in predicting hERG blockers than nonblockers based on the unbalanced data. Consequently, the prediction accuracy for the blockers is obviously higher than that of the nonblockers.

Clustering Analysis of Important Pharmacophores. The SVM classifier (SVM5) based on the decision tree at a depth of 6 established from training set I has the best predictive capability. As shown in Figure S5, the pharmacophore hypotheses used by SVM5 include Hypo44, Hypo48, Hypo75, Hypo77, Hypo84, Hypo85, Hypo87, Hypo99, Hypo102, Hypo103, Hypo154, Hypo178, Hypo190, and Hypo198 (Figure S6). Different pharmacophores may imply the multiple binding mechanisms of action of hERG blockers. In order to explore the structural difference of these 14 pharmacophores, the distance between each pair of pharmacophores was calculated, and the distance matrix was hierarchically clustered and presented as a dendrogram (Figure 6). The distance is defined as a function of the number of common pharmacophoric features and the RMSD (root-mean-squared displacement) between the matching features. As shown in Figure 6, these pharmacophores can be roughly divided into four classes. The detailed descriptions of the four representative pharmacophore hypotheses (Hypo44, Hypo75, Hypo103, and Hypo178) chosen from the four clusters are summarized in Tables S2–S5. It can be observed that different pharmacophores are strongly inclined to certain subsets in the training set. For example, as shown in Figure 7, four representative molecules (clofilium phosphate, neferine, lidoflazine, GBR12909) were mapped perfectly onto Hypo44, Hypo75, Hypo103, and Hypo178, respectively. Apparently, each pharmacophore can characterize the binding features of certain molecules and represent one binding mode. That is to say, the

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855–2866

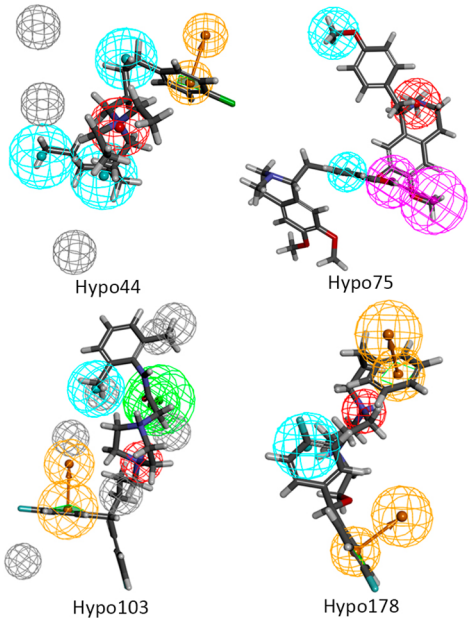


Figure 7. Mapping of four molecules (clofilium phosphate, neferine, lidoflazine, and GBR 12909) onto Hypo44, Hypo75, Hypo103, and Hypo178.

multimechanisms of action of hERG blockers may be effectively characterized by the four representative pharmacophore hypotheses, and therefore the integration of multiple pharmacophores is necessary to improve the prediction accuracies for the whole data set.

Analysis of Misclassified Molecules. By analyzing the prediction results, we observed that 16 molecules in test set I could not be correctly predicted by any NBC or SVM classifier. The 9 misclassified hERG blockers and 7 misclassified nonblockers are shown in Figure 8. We believe that the following reasons may explain the misclassification.

First, the misclassifications may be caused by the intrinsic drawback of the approach used in this study. The pharmacophore hypotheses can describe the spatial arrangement of the important structural features favorable for hERG binding, but they cannot precisely characterize the shape and chemical environment of the binding sites of hERG. As shown in Figure S7, all of the 7 misclassified nonblockers can be mapped onto at least one pharmacophore hypothesis. In these 9 misclassified blockers, only one (atropine) has the charged nitrogen. The blockers may not be well distinguished from the nonblockers because most pharmacophores have positive ionizable centers (Figure S6). Apparently, the successful mapping of a molecule onto a pharmacophore does not mean that this molecule can form favorable interaction with hERG because this molecule may form an unfavorable atom bump with the nearby residues that are not characterized by the hypotheses. All of the 7 misclassified nonblockers have aromatic rings and charged nitrogens, which are therefore more likely to be mapped onto the hydrophobic centers, ring aromatic centers, and positive ionizable centers in the pharmacophores. Second, the noise of the experimental data is another source of misclassification. Here, an arbitrary cutoff of 40 μM was used to define blockers and nonblockers. In the 9 misclassified hERG blockers, dihydropyrrole has an IC_{50} of 25 μM, which is quite close to the cutoff. It is understandable

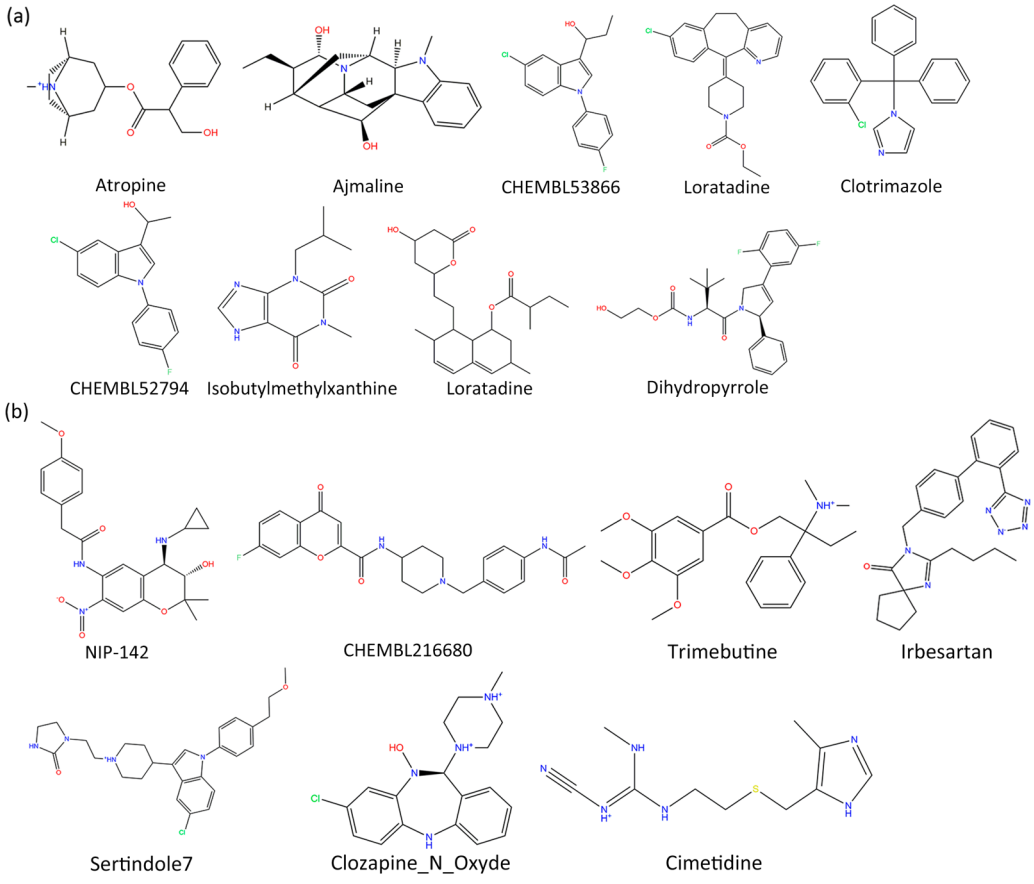


Figure 8. Structures of (a) the misclassified 9 blockers and (b) 7 nonblockers.

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866

that dihydropyrrole was easily misclassified as a nonblocker. Actually, in some previous studies, dihydropyrrole was also regarded as a nonblocker.[35,53] Similarly, two of the 7 misclassified nonblockers, Nip-142 and CHEMBL216680, have relatively low IC_{50} (44 μM and 47.7 μM), and these two compounds were easily predicted as blockers.

CONCLUSIONS

As a necessary step for safety evaluation of drugs, the assessment of hERG blockage is extremely important. In this study, based on a relatively large data set of 587 hERG blockers and nonblockers, pharmacophore modeling and machine learning approaches were combined to establish classification models to distinguish hERG active from inactive compounds. First, a large number of pharmacophore hypotheses were generated and an ensemble of pharmacophore hypotheses defined by a decision tree were identified by RP. Then, the NBC and SVM approaches were employed to construct classification models by integrating multiple representative pharmacophore hypotheses identified by RP. The best SVM model achieved prediction accuracies of 84.7% for the training set and 82.1% for the test set. The improved prediction capability of the integrated classification models suggests that the broad binding polyspecificity of hERG can only be well characterized by multiple pharmacophores. Moreover, the analysis of the important pharmacophores provides a deeper insight into the multimechanisms of action of hERG blockage. In summary, the multimechanisms of hERG blockage can be characterized by several important pharmacophores, and the combination of pharmacophores and machine learning methods can provide a powerful way to evaluate hERG blockage.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.molpharmaceut.6b00471.

List of the molecular structures and activities used in this study (XLSX)

Validation parameters, weights, tolerances, three-dimensional coordinates, pipeline of the data set partition, and additional tables and figures (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: tingjunhou@zju.edu.cn. Tel: +86-571-88208412.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was supported by the National Science Foundation of China (21575128) and the National Major Basic Research Program of China (2016YFA0501701 and 2016YFB0201700).

REFERENCES

(1) Smith, P. L.; Baukrowitz, T.; Yellen, G. The inward rectification mechanism of the HERG cardiac potassium channel. *Nature* **1996**, *379* (6568), 833−836.

(2) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **2006**, *440* (7083), 463−469.

(3) Vandenberg, J. I.; Perry, M. D.; Perrin, M. J.; Mann, S. A.; Ke, Y.; Hill, A. P. hERG K+ channels: structure, function, and clinical significance. *Physiol. Rev.* **2012**, *92* (3), 1393−1478.

(4) Recanatini, M.; Poluzzi, E.; Masetti, M.; Cavalli, A.; De Ponti, F. QT prolongation through hERG K+ channel blockade: current knowledge and strategies for the early prediction during drug development. *Med. Res. Rev.* **2005**, *25* (2), 133−166.

(5) Villoutreix, B. O.; Taboureau, O. Computational investigations of hERG channel blockers: New insights and current predictive models. *Adv. Drug Delivery Rev.* **2015**, *86*, 72−82.

(6) Witchel, H. J. The hERG potassium channel as a therapeutic target. *Expert Opin. Ther. Targets* **2007**, *11* (3), 321−36.

(7) Brugada, R.; Hong, K.; Dumaine, R.; Cordeiro, J.; Gaita, F.; Borggrefe, M.; Menendez, T. M.; Brugada, J.; Pollevick, G. D.; Wolpert, C. Sudden death associated with short-QT syndrome linked to mutations in HERG. *Circulation* **2004**, *109* (1), 30−35.

(8) Curran, M. E.; Splawski, I.; Timothy, K. W.; Vincen, G. M.; Green, E. D.; Keating, M. T. A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome. *Cell* **1995**, *80* (5), 795−803.

(9) Brown, A. Drugs, hERG and sudden death. *Cell Calcium* **2004**, *35* (6), 543−547.

(10) Aronov, A. M. Predictive in silico modeling for hERG channel blockers. *Drug Discovery Today* **2005**, *10* (2), 149−155.

(11) Raschi, E.; Vasina, V.; Poluzzi, E.; De Ponti, F. The hERG K+ channel: target and antitarget strategies in drug development. *Pharmacol. Res.* **2008**, *57* (3), 181−195.

(12) Bains, W.; Basman, A.; White, C. HERG binding specificity and binding site structure: evidence from a fragment-based evolutionary computing SAR study. *Prog. Biophys. Mol. Biol.* **2004**, *86* (2), 205−233.

(13) Roche, O.; Trube, G.; Zuegge, J.; Pflimlin, P.; Alanine, A.; Schneider, G. A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *ChemBioChem* **2002**, *3* (5), 455−459.

(14) Broccatelli, F.; Mannhold, R.; Moriconi, A.; Giuli, S.; Carosati, E. QSAR modeling and data mining link torsades de pointes risk to the interplay of extent of metabolism, active transport, and hERG liability. *Mol. Pharmaceutics* **2012**, *9* (8), 2290−2301.

(15) Sinha, N.; Sen, S. Predicting hERG activities of compounds from their 3D structures: Development and evaluation of a global descriptors based QSAR model. *Eur. J. Med. Chem.* **2011**, *46* (2), 618−630.

(16) Wang, S.; Li, Y.; Xu, L.; Li, D.; Hou, T. Recent developments in computational prediction of HERG blockage. *Curr. Top. Med. Chem.* **2013**, *13* (11), 1317−1326.

(17) Jing, Y.; Easter, A.; Peters, D.; Kim, N.; Enyedy, I. J. In silico prediction of hERG inhibition. *Future Med. Chem.* **2015**, *7* (5), 571−586.

(18) Dempsey, C. E.; Wright, D.; Colenso, C. K.; Sessions, R. B.; Hancox, J. C. Assessing hERG pore models as templates for drug docking using published experimental constraints: the inactivated state in the context of drug block. *J. Chem. Inf. Model.* **2014**, *54* (2), 601−612.

(19) Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *J. Pharmacol. Exp. Ther.* **2002**, *301* (2), 427−434.

(20) Aronov, A. M. Common pharmacophores for uncharged human ether-a-go-go-related gene (hERG) blockers. *J. Med. Chem.* **2006**, *49* (23), 6917−6921.

(21) Leong, M. K. A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. *Chem. Res. Toxicol.* **2007**, *20* (2), 217−226.

(22) Garg, D.; Gandhi, T.; Mohan, C. G. Exploring QSTR and toxicophore of hERG K+ channel blockers using GFA and HypoGen techniques. *J. Mol. Graphics Modell.* **2008**, *26* (6), 966−976.

(23) Durdagi, S.; Duff, H. J.; Noskov, S. Y. Combined receptor and ligand-based approach to the universal pharmacophore model development for studies of drug blockade to the hERG1 pore domain. *J. Chem. Inf. Model.* **2011**, *51* (2), 463−474.

(24) Tan, Y.; Chen, Y.; You, Q.; Sun, H.; Li, M. Predicting the potency of hERG K+ channel inhibition by combining 3D-QSAR pharmacophore and 2D-QSAR models. *J. Mol. Model.* **2012**, *18* (3), 1023−1036.

(25) Kratz, J. M.; Schuster, D.; Edtbauer, M.; Saxena, P.; Mair, C. E.; Kirchebner, J.; Matuszczak, B.; Baburin, I.; Hering, S.; Rollinger, J. M.

Experimentally validated HERG pharmacophore models as cardiotoxicity prediction tools. *J. Chem. Inf. Model.* **2014**, *54* (10), 2887−901.
(26) Stoll, F.; Goeller, A. H.; Hillisch, A. Utility of protein structures in overcoming ADMET-related issues of drug-like compounds. *Drug Discovery Today* **2011**, *16* (11−12), 530−538.
(27) Chen, L.; Li, Y. Y.; Yu, H. D.; Zhang, L. L.; Hou, T. J. Computational models for predicting substrates or inhibitors of P-glycoprotein. *Drug Discovery Today* **2012**, *17* (7−8), 343−351.
(28) Xu, Y.; Liu, X.; Li, S.; Zhou, N.; Gong, L.; Luo, C.; Luo, X.; Zheng, M.; Jiang, H.; Chen, K. Combinatorial pharmacophore modeling of organic cation transporter 2 (OCT2) inhibitors: insights into multiple inhibitory mechanisms. *Mol. Pharmaceutics* **2013**, *10* (12), 4611−4619.
(29) Ding, Y.-L.; Shih, Y.-H.; Tsai, F.-Y.; Leong, M. K. In silico prediction of inhibition of promiscuous breast cancer resistance protein (BCRP/ABCG2). *PLoS One* **2014**, *9* (3), e90689.
(30) Xu, Y.; Liu, X.; Wang, Y.; Zhou, N.; Peng, J.; Gong, L.; Ren, J.; Luo, C.; Luo, X.; Jiang, H. Combinatorial Pharmacophore Modeling of Multidrug and Toxin Extrusion Transporter 1 Inhibitors: a Theoretical Perspective for Understanding Multiple Inhibitory Mechanisms. *Sci. Rep.* **2015**, *5*, 13684.
(31) Leong, M. K.; Chen, H.-B.; Shih, Y.-H. Prediction of promiscuous p-glycoprotein inhibition using a novel machine learning scheme. *PLoS One* **2012**, *7* (3), e33829.
(32) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharmaceutics* **2012**, *9* (4), 996−1010.
(33) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A. PubChem's BioAssay database. *Nucleic Acids Res.* **2012**, *40* (D1), D400−D412.
(34) Aronov, A. M.; Goldman, B. B. A model for identifying HERG K+ channel blockers. *Bioorg. Med. Chem.* **2004**, *12* (9), 2307−2315.
(35) Li, Q.; Jørgensen, F. S.; Oprea, T.; Brunak, S.; Taboureau, O. hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharmaceutics* **2008**, *5* (1), 117−127.
(36) *Discovery Studio, version 2.5*; Accelrys Inc.: San Diego, CA, 2009.
(37) Waring, M. J.; Johnstone, C. A quantitative assessment of hERG liability as a function of lipophilicity. *Bioorg. Med. Chem. Lett.* **2007**, *17* (6), 1759−1764.
(38) Mannhold, R.; Kubinyi, H.; Folkers, G.; Langer, T.; Hoffmann, R. D. *Pharmacophores and pharmacophore searches*; John Wiley & Sons: 2006; Vol. 32.
(39) Shi, H.; Tian, S.; Li, Y.; Li, D.; Yu, H.; Zhen, X.; Hou, T. Absorption, Distribution, Metabolism, Excretion, and Toxicity Evaluation in Drug Discovery. 14. Prediction of Human Pregnane X Receptor Activators by Using Naive Bayesian Classification Technique. *Chem. Res. Toxicol.* **2015**, *28* (1), 116−125.
(40) Chen, L.; Li, Y.; Zhao, Q.; Peng, H.; Hou, T. 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) Tian, S.; Wang, J.; Li, Y.; Xu, X.; Hou, T. Drug-likeness analysis of traditional Chinese medicines: prediction of drug-likeness using machine learning approaches. *Mol. Pharmaceutics* **2012**, *9* (10), 2875−2886.
(42) Hou, T.; Zhang, W.; Wang, J.; Wang, W. Predicting drug resistance of the HIV-1 protease using molecular interaction energy components. *Proteins: Struct., Funct., Genet.* **2009**, *74* (4), 837−846.
(43) Hou, T.; Zhang, W.; Case, D. A.; Wang, W. Characterization of domain−peptide interaction interface: a case study on the amphiphysin-1 SH3 domain. *J. Mol. Biol.* **2008**, *376* (4), 1201−1214.
(44) Hou, T.; Wang, J.; Li, Y. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* **2007**, *47* (6), 2408−2415.
(45) Hou, T.; Li, N.; Li, Y.; Wang, W. Characterization of domain−peptide interaction interface: prediction of SH3 domain-mediated protein−protein interaction network in yeast by generic structure-based models. *J. Proteome Res.* **2012**, *11* (5), 2982−2995.
(46) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM T. Intel. Syst. Tec.* **2011**, DOI: 10.1145/1961189.1961199.
(47) Meyer, D. Support vector machines: The interface to libsvm in package e1071. 2004.
(48) Liu, L.-l.; Lu, J.; Lu, Y.; Zheng, M.-y.; Luo, X.-m.; Zhu, W.-l.; Jiang, H.-l.; Chen, K.-x. Novel Bayesian classification models for predicting compounds blocking hERG potassium channels. *Acta Pharmacol. Sin.* **2014**, *35* (8), 1093−1102.
(49) Hou, T.; Wang, J.; Zhang, W.; Wang, W.; Xu, X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006**, *13* (22), 2653−2667.
(50) Hou, T.; Wang, J. Structure - ADME relationship: still a long way to go? *Expert Opin. Drug Metab. Toxicol.* **2008**, *4* (6), 759−770.
(51) Du, L.; Li, M.; You, Q. The interactions between hERG potassium channel and blockers. *Curr. Top. Med. Chem.* **2009**, *9* (4), 330−338.
(52) Wei, Q.; Dunbrack, R. L., Jr The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One* **2013**, *8* (7), e67863.
(53) Sun, H. An accurate and interpretable Bayesian classification model for prediction of hERG liability. *ChemMedChem* **2006**, *1* (3), 315−322.

DOI: 10.1021/acs.molpharmaceut.6b00471
Mol. Pharmaceutics 2016, 13, 2855−2866