Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Toxicology Research
www.rsc.org/toxicology
ROYAL SOCIETY OF CHEMISTRY
Chinese Society Of Toxicology
bts
THE BRITISH TOXICOLOGY SOCIETY

In Silico Prediction of hERG Potassium Channel Blockage by Chemical Category Approaches

Chen Zhang, Yuan Zhou, Shikai Gu, Zengrui Wu, Wenjie Wu, Changming Liu, Kaidong Wang, Guixia Liu, Weihua Li, Philip W. Lee and Yun Tang[*]

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

*Corresponding Author: Tel: +86-21-64251052; Fax: +86-21-64251033

E-mail address: ytang234@ecust.edu.cn.

ABSTRACT

The human ether-a-go-go related gene (hERG) plays an important role in cardiac action potential. It encodes an ion channel protein named Kv11.1, which is related to long QT syndrome and may cause avoidable sudden cardiac death. Therefore, it is important to assess the hERG channel blockage of lead compounds in early drug discovery process. In this study, we collected a large data set containing 1163 diverse compounds with IC_{50} values determined by patch clamp method on mammalian cell lines. The whole data set was divided into 80% as the training set and 20% as the test set. Then, five machine learning methods were applied to build a series of binary classification models based on 13 molecular descriptors, five fingerprints and molecular descriptors combining fingerprints at four IC_{50} thresholds to discriminate hERG blockers from nonblockers, respectively. Models built by molecular descriptors combining fingerprints were validated by an external validation set containing 407 compounds collected from the hERGCentral database. The performance indicated that the model built by molecular descriptors combining fingerprints yielded the best results and each threshold had its best suitable method, which means that hERG blockage assessment might depend on threshold values. Meanwhile, kNN and SVM methods were better than the others for model building. Furthermore, six privileged substructures were identified using information gain and frequency analysis methods, which could be regarded as structural alerts of cardiac toxicity mediated by hERG channel blockage.

1. Introduction

The human ether-a-go-go related gene (hERG) encodes a tetrameric potassium channel named Kv11.1, which plays an important role in cardiac action potential.[1] Inhibition of hERG channel may result in long QT Syndrome (LQTS), and cause avoidable sudden cardiac death.[2] Meanwhile, undesirable hERG related cardiotoxicity will lead to failure of drug development, and is also the main reason for drug withdrawl from the market, such as terfenadine, cisapride, sertindole, thioridazine, and grepafloxacin.[3] Therefore, it is important to assess chemical blockage of hERG channel in early drug discovery process.

In vitro evaluation of hERG binding drugs is a valuable method to identify potential hERG blockage in drug discovery.[4] Hence, a variety of *in vitro* methods are developed, including rubidium-flux assays, radioligand binding assays, electrophysiology measurements, and fluorescence-based assays. The half maximal concentration (IC_{50}) blockage of hERG channel is a surrogate marker for LQTS-related proarrhymic properties of chemicals and as a considered test for cardiac safety of drugs or drug candidates.[5] However, *in vitro* hERG binding assays are time consuming, expensive and labor-intensive. Therefore, there is an increasing demand to develop *in silico* models and improve pattern recognition methods for prediction of drug-hERG interaction.

Since the crystal structure of hERG channel has not been determined yet, ligand-based prediction models are the main tools for predicting chemical blockages of hERG channel.[2] In recent years, a series of computational models have been developed to classify hERG blockers and nonblockers. In 2002, Ekins *et al.* reported the first pharmacophore model for hERG blockers based on a training set of 15 compounds.[6] Cavalli *et al.* also presented a pharmacophore model along with CoMFA study based on a training set of 32 drugs for a series of hERG blockers, and tested by five drugs collected from literatures.[7]

In recent years, QSAR models have been widely used for prediction of hERG blockage. For example, Li *et al.* built binary classification models based on 495 compounds using support vector machine (SVM) classifier combined with pharmacophores based on GRIND descriptors in 2008.[8] The models were applied to different IC_{50} thresholds from 1 to 40 μM, and threshold at 40 μM exerted best performance with an overall accuracy up to 94% by leave-one-out cross-validation. 66 compounds from WOMBAT-PK database and PubChem hERG bioassay data set were applied as two external validation sets and achieved 72% and 73% accuracy, respectively. In 2010, Su *et al.* built a continuous partial least-squares (PLS) model and an optimized binary classification model using a set of 250 compounds.[9] The binary model achieved 91% accuracy for the training set, 83% and 77% for two external test sets, one containing 876 compounds from PubChem and the other including 106 compounds collected from literatures. In 2011, they also reported an SVM model based on 1668 PubChem bioassay compounds using the same methods.[10] The model achieved 95% and 87% accuracy for the training set and an external test set containing 365 compounds, respectively.

In 2010, Doddareddy *et al.* described *in silico* hERG models generated by 2644 compounds using linear discriminant analysis (LDA) and SVM methods. The area under curve (AUC) values of all models ranged from 0.89 to 0.94 by 5-fold cross-validation, and the SVM model was much better than the LDA model. Then, the models were experimentally validated, and worked as a pre-filtering tool to reduce the number of compounds with hERG liabilities.[11] In 2012, Wang *et al.* established binary classification models using naive Bayes (NB) classification and recursive partitioning methods based on a diverse set of 806 hERG inhibitors.[12] With molecular descriptors and ECFP_8 fingerprints, the NB model yielded 85% accuracy for the test set of 120 compounds, 89.4% and 86.1% accuracy for two additional external validation sets, WOMBAT-PK and PubChem, respectively. In 2014, Liu *et al.* also built NB classification models using similar methods based on Doddareddy's data set containing 2644 compounds.[13] The best model achieved 91% accuracy for the test set and 58% for external validation set containing 60 compounds.

Though ligand-based computational methods are widely used for prediction of hERG blockage, most published models have limitations at the number of open source data sets, qualities of data points and lack of specific and rational thresholds to distinguish hERG blockers from nonblockers. Herein, we collected a large and open source data set with IC_{50} values determined by patch clamp method on mammalian cell lines. Then, machine learning methods were applied to build binary classification models based on molecular descriptors (MD), fingerprints and their combination at four IC_{50} thresholds, respectively. The models were then validated by an external data set containing 407 chemicals that collected from the hERGCentral database. Furthermore, privileged substructures which would be significantly correlated with hERG blockage were identified by information gain (IG) and frequency analysis methods.

2. Materials and methods

2.1. Data preparation

The original chemicals with experimental IC_{50} values were collected from two publications[11, 12] and ChEMBL database (version 18, target_ID: CHEMBL240).[14, 15] Only patch clamp determined IC_{50} values on different mammalian cell lines were collected in this study, such as HEK, CHO, COS and XO (Xenopus laevis oocytes) cell lines. Since we aimed to create qualitative rather than quantitative models for hERG blockage prediction, the differences of activities among various cell lines were ignored.

After that, the whole data set was carefully prepared. Inorganic compounds, noncovalent complexes and mixtures were removed from the data set. Salts were converted to their corresponding acidic or basic forms, and water molecules were removed from the hydrates. Then, duplicates were treated using canonical SMILES by a simple principle. If duplicates had the same IC_{50} values, they were merged as one molecule; if their IC_{50} values were in one order of magnitude, the averaged IC_{50} values were used after removing duplicated molecules; otherwise, all the duplicates were removed. Finally, molecules with molecular weight greater than 40 but less than 800 were kept.[16, 17]

Since there is lack of specific thresholds to discriminate hERG blocker from nonblockers, we tested four IC_{50} values, namely 1 μM, 5 μM, 10 μM and 30 μM as thresholds to find the most suitable threshold for our data set.[12] For model building, the whole data set was randomly divided into 80% as the training set and 20% as test set, respectively. In additional, 407 chemicals with IC_{50} values were collected from hERGCentral database[18, 19] and used as an external validation set to evaluate the predictive abilities of all models. The statistics of training set, test set and external validation set were summarized in Table 1. All compounds with SMILES are available in Supplementary Materials 2.

2.2. Calculation of molecular descriptors

In our study, 13 molecular descriptors which were widely used in ADMET prediction[20, 21] were calculated and evaluated, including octanol−water partitioning coefficient AlogP, logD, molecular weight (MW), molecular solubility (logS), the number of nitrogen and oxygen atoms in the molecule (NplusO), the number of rotatable bonds (nROT), the number of rings (nR), the number of aromatic rings (nAR), the number of hydrogen bond acceptors (nHBA), the number of hydrogen bond donors (nHBD), molecular surface area (MSA), molecular polar surface area (MPSA), and molecular fractional polar surface area (MFPSA). All the descriptors were calculated using Discovery Studio version 3.5.[22]

Meanwhile, five commonly used fingerprints were calculated by PaDEL-Descriptor,[23] including Estate Fingerprint (Estate), CDK Fingerprint (FP), Substructure Fingerprint (FP4), MACCS Fingerprint (MACCS), and PubChem Fingerprint (PubChem). The detailed descriptions of these fingerprints could be found elsewhere.[23, 24] The advantage of fingerprints is that they are generated directly from chemical structures, and could be easily translated into two-dimensional fragments.[25]

2.3. Model building

The whole data set was randomly divided into 80% as the training set and 20% as test set. Using the data sets, a series of binary classification models were developed based on molecular descriptors, fingerprints or their combination at four thresholds, respectively. All models were validated by 5-fold cross-validation. Models built by molecular descriptors combining fingerprints were then applied to the external validation set to evaluate predictive abilities of the models, and the best performance threshold could be acted as the reference to discriminate hERG blockers from nonblockers.

Five machine learning methods including SVM, NB, *k*-nearest neighbors (*k*NN), random forest (RF), and classification tree (CT) were applied to develop models. The SVM algorithm was provided by the open source LIBSVM (LIBSVM2.9 package),[26] and other four methods were performed using Orange (version 2.7, freely available at http://orange.biolab.si/).

NB is a simple classification method based on the Bayes rule with strong independence assumptions for the conditional probability. Based on the equal and independent contribution of attributes, it can categorize instances in a data set. In this study, the prior probability estimates from the training set and the marginal probability are ignored, since it is the same to all of the classes.[27] Orange with the default setting was applied to perform the NB classifier in this work.

*k*NN is a non-parametric method to classify objects based on closest training examples in the feature space.[28] *k* value is a user-defined constant. An object is classified to the class that the object is assigned to the most common among its *k* nearest neighbors. A distance-weighted method is applied to weaken the impact of *k* value. The Hamming distance is an estimate of error used in telecommunication to count the number of flipped bits in a fixed-length binary word, which was commonly used in *k*NN method. Therefore, the Hamming distance was selected for distance metric. The *k* value was set to 5, which was determined after a series of *k* values were tested and compared by the performance in this study.

RF is an ensemble learning method developed by Breiman.[29] The forest is ensemble by lots of trees. In the forest, a new object from an input vector is putted down each of trees. Each tree gives a defined class, which means the tree “vector” for the class. The forest chooses the classification having the most vectors over all the trees in the forest with the trees grow up to the maximum size.

CT algorithm is used to predict the member of cases and objects in the classes of a category-dependent variable from their measurements on one or more predictor variables. In this study, information gain was opted as the first attribute selection criterion. In the pre-pruning process, the minimal instance in leaves was set to 2, and stop splitting nodes with few instances than 5. And in the post-pruning process, pruning with m-estimate was selected and m was equal to 2. Other parameters of CT in Orange was default.

The SVM method is aimed to minimize the structural risk under the frame of VC theory and it is also a widely used binary classification and regression method based on different kernel functions.[30] This method has been successfully applied in ADMET properties prediction by our group.[16, 31, 32] In this field, each chemical structure was described as binary string and treated as an eigenvector. Then the eigenvector was trained by SVM algorithm, and a decision function was given for classification. In order to obtain the optimal performance model, the Gaussian radial basis function (RBF) was applied to seek the penalty parameter C and different kernel parameter γ, using grid search strategy based on a 5-fold cross-validation.

2.4. Performance evaluation of models

All models were evaluated by the 5-fold cross-validation and validated by the external validation set to test the predictive ability. In addition, models were assessed by counting the numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) of each class, respectively. Sensitivity (SE), specificity (SP), overall predictive accuracy (Q), and Matthews correlation coefficient (C) were also calculated by following equations.[25, 33]

$$\mathrm{SE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{1}$$

$$\mathrm{SP} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \tag{2}$$

$$\mathrm{Q} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \tag{3}$$

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

The values of Q and C are two important evaluation indicators for the classification models. The above indicators were calculated for training set, test set and external validation set. The areas under the receiver operating characteristic (ROC) curve (AUC) were also calculated. The AUC value is the probability of active compounds being ranked earlier than the negative compounds, which shows the separation ability of a binary classifier iteratively setting the positive classifier threshold.

2.5. Identification of privileged substructures

The privileged structure was first defined as "a single molecular framework able to provide ligands for diverse receptors" by Evans and co-workers in 1988.[34] In toxicology research area, privileged substructures are defined as structure fragments in chemicals that are known to bring the toxicity. Herein, the privileged substructures were analyzed using information gain (IG) method and substructure frequency analysis methods.[31,35] The IG value of each substructure feature is calculated based on the information entropy theory. Features with no or lower IG values are discarded according to a predetermined threshold, and the remaining patterns compose a multidimensional vector to represent each molecule. The important substructure features which are major contributions to the classification system could be identified. In this study, 10 μM was used as the standard threshold to divide hERG blockers from nonblockers. If a substructure was more frequently presented in hERG blockers than that in hERG nonblockers, the substructure was called privileged substructure for hERG blockage. The frequency of a substurcture was defined as following:

$$Frequency\ of\ a\ substructure\ \ (F) = \frac{N_{substructure\_class} \times N_{total}}{N_{substructure\_total} \times N_{class}} \qquad (5)$$

Where $N_{substructure\_class}$ is the number of compounds containing the substructure in each class; N_{total} is the total number of compounds; $N_{substructure\_total}$ is the total number of compounds containing the substructure; and N_{class} is the number of compounds in each class.

3. Results

3.1. Data collection and chemical space analysis

After careful data preparation, a large diverse and high quality of hERG blockage database was constructed, which contained 1570 unique compounds with IC_{50} values determined by the patch clamp assay on mammalian cell lines. To our knowledge, this is the largest open source database for hERG blockage at present. To reduce the mechanism complexity of hERG blockage, the experimental IC_{50} values measured on mammalian cell lines were applied as the endpoint for model building in this study. Differences among various cell lines were ignored, since we aimed to create qualitative rather than quantitative models for hERG blockage prediction. Therefore, different cell lines IC_{50} values were collected, such as HEK, CHO, COS and XO (Xenopus laevis oocytes) cell lines.

As we known, chemical diversity is important for model building. In this study, the chemical space distribution defined by MW and AlogP of training set and test set was shown in Figure 1A, which indicated that the data set was relatively diverse. The molecular weights range from 94.11 to 780.94, and the AlogP values range from -5.23 to 11.51. The test set shares a similar chemical space of training set. To further explore the chemical diversity of the data set, Tanimoto similarity index of the entire data set was calculated using MACCS fingerprint. The Tanimoto coefficient uses the ratio of the intersecting set to the union set as the measure of similarity, which is widely used to evaluate similarities among chemicals. As shown in Figure 1B, the whole data set was separated into 100 clusters and the heat map of molecular similarity was plotted by tanimoto similarity index of cluster center molecules of the 100 clusters. The average tanimoto similarity index was 0.338, indicating that the chemical diversity of the data set was significant.

3.2. Performance of different models

Five machine learning methods were used to develope the models and the 5-fold cross-validation technique was used to evaluate the model robustness. Model performance was assessed by counting the numbers of TP, TN, FP, and FN of each class, respectively. SE, SP, Q, and C were also calculated.

At first, the binary classification models were built using the 13 molecular descriptors. Performances of descriptor-based models were summarized in Table S1 of Supplementary Materials 1. In general, the SVM method yielded the best performance among the 5 model building methods at the 4 hERG blockage thresholds, while the overall accuracy of the best model (kNN model) reached to 0.8347 for the test set among all of the models.

Then, binary classification models were built using five fingerprints. Performances of fingerprint-based models were shown in Table S2 of Supplementary Materials 1. It is easy to see from Table S2 that three fingerprints namely FP, MACCS and PubChem yielded the best predictive performances. The SVM model based on FP fingerprint at 30 μM threshold had the best predictive ability for test set and yielded overall accuracy to 0.8475.

The combination of molecular descriptors and fingerprints were also used to build models. The best performances models for each threshold were shown in Table 2 and the detail performances of all combinatorial classification models were summarized in Table S3 of Supplementary Materials 1. The Q values ranged from 0.6257 to 0.8393 for training set based on 5-fold cross-validation, which yielded

0.5551 to 0.8475 for the test set containing 236 compounds at different hERG blockage thresholds. The best model using SVM method based on molecular descriptors combining FP fingerprint yield the highest Q value of 0.8475 for test set.

To compare the above-mentioned three types of classification models, the average and standard deviation (SD) of Q and C values for different models on different thresholds were also calculated (Table 3). According to Table 3, we could find that models built using molecular descriptors combining fingerprints were better than others.

3.3. Performance of external validation

To test the robust and prediction ability of the models, an external validation set containing 407 compounds which was collected from the hERGCentral database was applied to the best models with various thresholds. The performances of the external validation set was shown in Table 4. According to Table 4, we could find that the Q values range from 0.5528 to 0.8550 among the four thresholds and the best model is (FP+MD)-SVM model which yielded the accuracy of 0.8550 at threshold 30 μM. Moreover, we can find that there are suitable fingerprints and model building methods for each hERG blockage threshold. It indicates that models building for hERG blockage depend on the thresholds, and hERG blockage assessments should be case by case.

3.4. Privileged substructure for hERG blockage

Besides models building, we used IG and frequency analysis methods along with FP4 fingerprint to identify privileged substructures which were involved in chemical hERG blockage. In this process, the IC_{50} equal to 10 μM was used as the threshold to discriminate hERG blockage. The higher value IG is, the more important the substructure is.

Some privileged substructures and representative compounds of hERG blockage were shown in Table 5, and the detailed IG values of all FP4 fragments were listed in Table S4 of Supplementary Materials 2. In Table 5, six substructures are listed. They are diarylthioether, tertiary mixed amine, imide acidic, amidine, arylchloride, and sulfonamide. The six fragments reflect the common features of chemicals which have the potential of hERG blockage and could be used as substructure alerts for cardiotoxicity mediated by hERG channel in drug discovery and development processes.

4. Discussion

4.1. Quality of hERG blockage database

The quality of a chemical database defines its modelability in cheminformatics or QSAR studies.[36] Three factors to affect the quality of a chemical database are data size, mechanism complexity and chemical structure diversity. In our study, the three factors were carefully checked in order to obtain a high quality database for hERG blockage modeling. All the data were collected from recent publications[11, 12] and the largest bioactive molecular database ChEMBL[14, 15] to guarantee the wide range of data points. Then, only data points with experimental IC_{50} values were kept. After that, all chemicals were carefully checked and prepared with the general criteria for chemical database quality.[37, 38] Furthermore, to check the chemical diversity of the database, MW and AlogP of the training set and test set were plotted and tanimoto similarity index based on MACCS fingerprint was also calculated.[39, 40] The results indicated that the collected hERG blockage database had a high diversity.

4.2. Relevance of molecular properties to hERG blockage

Molecular properties are important for ADMET prediction and model optimization. Herein, we systematically examined the relationships between eight molecular properties and hERG blockage. They are AlogP, logD, MW, logS, nROT, nHBA, nHBD, and MPSA. The distributions of these molecular properties for hERG blockers and nonblockers at the threshold of 10 μM were shown in Figure 2. From Figure 2, we could find that the distributions of these molecular properties were quite different between hERG blockers and nonblockers at this threshold. That is to say, these molecular properties might play key roles in hERG blockage and could be used to simply distinguish hERG blockers from hERG nonblockers. That is also the reason why we choose these properties as molecular descriptors to build models. In addition, linear correlations of these molecular properties versus IC_{50} values of the whole data set were presented in Figure 3, which also indicated that these molecular properties would have significant relationships with hERG blockage. In some cases, SAR analyses were established just using these molecular properties.[2, 41]

4.3. Comparison of different category methods

The average and standard deviation (SD) of Q and C values for models at different thresholds were used to compare the above-mentioned three types of classification models. According to Table 3, we could find that models built by molecular descriptors combining fingerprints yielded the best performance among the three molecular description methods. In other words, models based on molecular descriptors combining fingerprints reached the best results at the four hERG blockage thresholds. Therefore, these models were used for external validation data set evaluation and could also be applied in hERG blockage prediction for new chemicals or drug candidates in preclinical process of drug discovery and development.

Furthermore, comparing the five category methods, we found that *k*NN and SVM methods performed better than the others. Meanwhile, *k*NN and SVM methods also led better performance among the three different molecular description methods, which indicates that *k*NN and SVM methods would be more suitable for hERG blockage prediction in this study. The two modeling methods are easy to use than the others when apply to the actual hERG blockage assessments, since the number of independent variables in the model are easy to organize. Additionally, these results are in agreement with our previously published work that SVM algorithm is a good category method for chemical toxicity prediction.[16, 17, 40, 42, 43]

4.4. Relevance of thresholds to hERG blockage

One of the challenges for hERG blockage prediction is lack of definite threshold for discrimination of hERG blockers from nonblockers.[2, 44] To solve the problem, we chose four IC_{50} values, namely 1 μM, 5 μM, 10 μM, and 30 μM, as thresholds for model building according to published studies.[2, 12] The results revealed that for each model building method, the threshold 30 μM achieved the best performance. However, models at threshold 1 μM seemed to perform equally as those at 30 μM in comparison of overall accuracies. Hence, there are suitable category methods for each threshold of hERG blockage modeling, for example (PubChem+MD)-SVM model yielded accuracies of 0.8350, 0.8136, 0.8501 for the training, test and external validation sets at threshold 1 μM, respectively, whereas (FP+MD)-SVM model reached accuracies of 0.7832, 0.8475, 0.8550 for the three data sets at threshold 30 μM. The reason was that each threshold led to different numbers of hERG blockers and nonblockers, and the difference of positive and negative compounds had huge impacts on the performance of QSAR models.[45] That is to say, hERG blockage prediction depended on the thresholds to distinguish hERG blockers from nonblockers. Therefore, for hERG blockage prediction, there is no universal threshold for hERG blockage. We should do it case by case. And there is meaningless to discuss which is better for hERG blockage prediction despite its thresholds.

4.5. Analysis of privileged substructures for hERG blockage

To understand the common chemical features of hERG blockers, IG method along with frequency analysis were used to identify privileged substructures of hERG blockers based on FP4 fingerprint.[31] From the results of IG values and frequency analysis of all fragments in FP4 fingerprint, six privileged substructures for hERG blockers were recognized. In general, the six substructures have some common features. They are large groups and may connect to aryl groups. Large groups and aryl groups might block potassium ions to pass through the channel because there are two couples of aryl residues Tyr652 and Phe656 in the potassium channel of hERG protein structures.[46, 47] The four residues control the open and close of potassium channel and the flux of potassium ions. Hence, chemicals interacting with these residues have potential of hERG blockage. Structures containing aryl groups are easy to form interactions with these residues and block potassium transit. Such structures also have high molecular weight, hydrophobicity and large stereoscopic space. All the six privileged substructures have such features. That are why the molecular properties MW and AlogP have higher correlation with hERG blockage.

Three of the six privileged substructures imide acidic, amidine, sulfonamide are acidic and negative electronics groups, they are easy to interact with the residues by other interactions, such as hydrogen bonds and electrostatic interactions. These substructures could also capture potassium by negative electronic group and block the ion channel. The above characteristics indicate that our privileged substructure identification method is credible, and the six privileged substructures reflect the common chemical structure features actually. These finds could be a guideline for hERG blockage assessment during drug discovery and development processes.

5. Conclusions

In this study, we focused on three aspects of hERG blockage: data quality, hERG blockage thresholds and prediction models. We built a hERG blockage database containing 1570 compounds, which were the largest available hERG blockage database until now. Besides, we constructed a workflow to obtain high quality data. Therefore, all data points in our hERG blockage have high quality with simple mechanisms and high diversity. Then, hERG blockers from nonblockers were distinguished by four common used thresholds and series binary classification models were built based on thirteen molecular descriptors, five common used fingerprints and molecular descriptors combining fingerprints using five machine learning methods, respectively. After systemic evaluated all models, we found that each threshold had its best category methods and hERG blockage assessments depend on its thresholds. The method could be used for hERG blockage assessment and solved the three challenges of hERG blockage studies. Models developed in this study will provide critical information and useful tools for hERG blockage assessment of new drug candidates. Furthermore, IG method combining frequency analysis were applied to identify six privileged substructures of hERG blockers. The six privileged substructures reflect the common chemical structure features and explained the mechanisms hERG blockage of compounds. They could be treated as alert substructures for hERG blockage assessments during safety evaluation process of drug discovery.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (Grant 81373329), the 863 Project (Grant 2012AA020308) and the 111 Project (Grant B07023).

Supporting information available

The performance of the descriptor-based models and fingerprint-based models were summarized in Table S1 and S2 and the performance of all combinatorial classification models were summarized in Table S3 of Supplementary Materials 1. The details of all compounds with SMILES were available in Supplementary Materials 2 and IG values of each pattern in the FP4 fingerprint were also listed in Table S4 of Supplementary Materials 2.

References

1 M. C. Sanguinetti and M. Tristani-Firouzi, hERG potassium channels and cardiac arrhythmia, *Nature*, 2006, **440**, 463-469.

2 S. Wang, Y. Li, L. Xu, D. Li and T. Hou, Recent developments in computational prediction of HERG blockage, *Curr. Top. Med. Chem.*, 2013, **13**, 1317-1326.

3 H. Laverty, C. Benson, E. Cartwright, M. Cross, C. Garland, T. Hammond, C. Holloway, N. McMahon, J. Milligan, B. Park, M. Pirmohamed, C. Pollard, J. Radford, N. Roome, P. Sager, S. Singh, T. Suter, W. Suter, A. Trafford, P. Volders, R. Wallis, R. Weaver, M. York and J. Valentin, How can we improve our understanding of cardiovascular safety liabilities to develop safer medicines?, *Br. J. Pharmacol.*, 2011, **163**, 675-693.

4 G. E. Kirsch, E. S. Trepakova, J. C. Brimecombe, S. S. Sidach, H. D. Erickson, M. C. Kochan, L. M. Shyjka, A. E. Lacerda and A. M. Brown, Variability in the measurement of hERG potassium channel inhibition: effects of temperature and stimulus pattern, *J. Pharmacol. Toxicol. Methods*, 2004, **50**, 93-101.

5 S. Polak, B. Wisniowska and J. Brandys, Collation, assessment and analysis of literature in vitro data on hERG receptor blocking potency for subsequent modeling of drugs' cardiotoxic properties, *J. Appl. Toxicol.*, 2009, **29**, 183-206.

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

7 A. Cavalli, E. Poluzzi, F. De Ponti and M. Recanatini, Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K(+) channel blockers, *J. Med. Chem.*, 2002, **45**, 3844-3853.

8 Q. Li, F. S. Jorgensen, T. Oprea, S. Brunak and O. Taboureau, hERG classification model based on a combination of support vector machine method and GRIND descriptors, *Mol. Pharm.*, 2008, **5**, 117-127.

9 B. H. Su, M. Y. Shen, E. X. Esposito, A. J. Hopfinger and Y. J. Tseng, In Silico Binary Classification QSAR Models Based on 4D-Fingerprints and MOE Descriptors for Prediction of hERG Blockage, *J. Chem. Inf. Model.*, 2010, **50**, 1304-1318.

10 M. Y. Shen, B. H. Su, E. X. Esposito, A. J. Hopfinger and Y. J. Tseng, A comprehensive support vector machine binary hERG classification model based on extensive but biased end point hERG data sets, *Chem. Res. Toxicol.*, 2011, **24**, 934-949.

11 M. R. Doddareddy, E. C. Klaasse, Shagufta, A. P. Ijzerman and A. Bender, Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases, *ChemMedChem*, 2010, **5**, 716-729.

12 S. Wang, Y. Li, J. Wang, L. Chen, L. Zhang, H. Yu and T. Hou, ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage, *Mol. Pharm.*, 2012, **9**, 996-1010.

13 L. L. Liu, J. Lu, Y. Lu, M. Y. Zheng, X. M. Luo, W. L. Zhu, H. L. Jiang and K. X. Chen, Novel Bayesian classification models for predicting compounds blocking hERG potassium channels, *Acta Pharmacol. Sin.*, 2014, **35**, 1093-1102.

14 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Res.*, 2012, **40**, D1100-1107.

15 P. Czodrowski, hERG me out, *J. Chem. Inf. Model.*, 2013, **53**, 2240-2251.

16 F. Cheng, Y. Ikenaga, Y. Zhou, Y. Yu, W. Li, J. Shen, Z. Du, L. Chen, C. Xu, G. Liu, P. W. Lee and Y. Tang, In silico assessment of chemical biodegradability, *J. Chem. Inf. Model.*, 2012, **52**, 655-669.

17 C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P. W. Lee and Y. Tang, In silico prediction of chemical Ames mutagenicity, *J. Chem. Inf. Model.*, 2012, **52**, 2840-2847.

18 F. Du, H. Yu, B. Zou, J. Babcock, S. Long and M. Li, hERGCentral: a large database to store, retrieve, and analyze compound-human Ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development, *Assay Drug Dev. Technol.*, 2011, **9**, 580-588.

19 J. J. Babcock, F. Du, K. Xu, S. J. Wheelan and M. Li, Integrated analysis of drug-induced gene expression profiles predicts novel hERG inhibitors, *Plos One*, 2013, **8**, e69513.

20 T. Hou and J. Wang, Structure-ADME relationship: still a long way to go?, *Expert. Opin. Drug Metab. Toxicol.*, 2008, **4**, 759-770.

21 T. Hou, Y. Li, W. Zhang and J. Wang, Recent developments of in silico predictions of intestinal absorption and oral bioavailability, *Comb. Chem. High Throughput Screening*, 2009, **12**, 497-506.

22 Discovery Studio, Version 3.5, Accelrys Inc., San Diego, CA, USA, 2012.

23 C. W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, *J. Comput. Chem.*, 2011, **32**, 1466-1474.

24 J. Klekota and F. P. Roth, Chemical substructures that enrich for biological activity, *Bioinformatics*, 2008, **24**, 2518-2525.

25 F. Cheng, Y. Yu, Y. Zhou, Z. Shen, W. Xiao, G. Liu, W. Li, P. W. Lee and Y. Tang, Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds, *J. Chem. Inf. Model.*, 2011, **51**, 2482-2495.

26 C.-C. Chang and C.-J. Lin, LIBSVM, version 2.9. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed November 28, 2011).

27 H. Sun, A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing, *J. Med. Chem.*, 2005, **48**, 4031-4039.

28 S. Geva and J. Sitte, Adaptive nearest neighbor pattern classification, *IEEE Trans. Neural Netw.*, 1991, **2**, 318-322.

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

30 C. Cortes and V. Vapnik, Support-Vector Networks, *Mach. Learn.*, 1995, **20**, 273-297.

31 J. Shen, F. Cheng, Y. Xu, W. Li and Y. Tang, Estimation of ADME properties with substructure pattern recognition, *J. Chem. Inf. Model.*, 2010, **50**, 1034-1041.

32 F. Cheng, Y. Yu, J. Shen, L. Yang, W. Li, G. Liu, P. W. Lee and Y. Tang, Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers, *J. Chem. Inf. Model.*, 2011, **51**, 996-1011.

33 P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen and H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics*, 2000, **16**, 412-424.

34 B. E. Evans, K. E. Rittle, M. G. Bock, R. M. DiPardo, R. M. Freidinger, W. L. Whitter, G. F. Lundell, D. F. Veber, P. S. Anderson, R. S. Chang and et al., Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists, *J. Med. Chem.*, 1988, **31**, 2235-2246.

35 B. F. Jensen, C. Vind, S. B. Padkjær, P. B. Brockhoff and H. H. Refsgaard, In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors, *J. Med. Chem.*, 2007, **50**, 501-511.

36 A. Golbraikh, E. Muratov, D. Fourches and A. Tropsha, Data set modelability by QSAR, *J. Chem. Inf. Model.*, 2014, **54**, 1-4.

37 D. Fourches, E. Muratov and A. Tropsha, Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research, *J. Chem. Inf. Model.*, 2010, **50**, 1189-1204.

38 D. Young, T. Martin, R. Venkatapathy and P. Harten, Are the Chemical Structures in Your QSAR Correct?, *QSAR Comb. Sci.*, 2008, **27**, 1337-1345.

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

40 C. Zhang, F. Cheng, L. Sun, S. Zhuang, W. Li, G. Liu, P. W. Lee and Y. Tang, In silico prediction of chemical toxicity on avian species using chemical category approaches, *Chemosphere*, 2015, **122**, 280-287.

41 A. Aronov, Predictive in silico modeling for hERG channel blockers, *Drug Discov. Today*, 2005, **10**, 149-155.

42 X. Li, L. Chen, F. Cheng, Z. Wu, H. Bian, C. Xu, W. Li, G. Liu, X. Shen and Y. Tang, In silico prediction of chemical acute oral toxicity using multi-classification methods, *J. Chem. Inf. Model.*, 2014, **54**, 1061-1069.

43 L. Sun, C. Zhang, Y. J. Chen, X. Li, S. L. Zhuang, W. H. Li, G. X. Liu, P. W. Lee and Y. Tang, In silico prediction of chemical aquatic toxicity with chemical category approaches and substructural alerts, *Toxicol. Res.*, 2015, **4**, 452-463.

44 J. Heijman, N. Voigt, L. G. Carlsson and D. Dobrev, Cardiac safety assays, *Curr. Opin. Pharmacol.*, 2014, **15**, 16-21.

45 A. V. Zakharov, M. L. Peach, M. Sitzmann and M. C. Nicklaus, QSAR modeling of imbalanced high-throughput screening data in PubChem, *J. Chem. Inf. Model.*, 2014, **54**, 705-712.

46 D. J. Leishman and Z. Rankovic, Drug Discovery vs hERG, *Top. Med. Chem.*, 2014, **9**, 225-259.

47 B. O. Villoutreix and O. Taboureau, Computational investigations of hERG channel blockers: New insights and current predictive models, *Adv. Drug Del. Rev.*, 2015, **86**, 72-82.

Tables

Table 1 The statistics of chemicals in the training set, test set and external validation set.

data sets	thresholds (μM)					total
	(0, 1]	(1, 5]	(5, 10]	(10, 30]	(30, ∞）	
training set	233	214	115	159	206	927
test set	60	60	33	36	47	236
external validation set	57	97	90	149	14	407
total	350	371	238	344	267	1570

Table 2 Performance of classification models for the taining set and test set using different modeling methods based on molecular descriptors (MD) combining fingerprints

thresholds (μM)	modeling methods	training set					test set				
		Q	SE	SP	AUC	C	Q	SE	SP	AUC	C
1	(Estate+MD)-*k*NN	0.7993	0.5622	0.8790	0.8144	0.4534	0.7712	0.6000	0.8295	0.8326	0.4167
	(Estate+MD)-SVM	0.8091	0.4721	0.9222	0.8306	0.4482	0.7881	0.4167	0.9148	0.8389	0.3847
	(FP+MD)-*k*NN	0.8220	0.5837	0.9020	0.8402	0.5086	0.8008	0.6667	0.8466	0.8366	0.4957
	(FP+MD)-SVM	0.8177	0.4807	0.9308	0.8529	0.4724	0.8305	0.6333	0.8977	0.8414	0.5436
	(FP4+MD)-*k*NN	0.8004	0.5279	0.8919	0.8065	0.4444	0.8136	0.6167	0.8807	0.8312	0.5030
	(FP4+MD)-SVM	0.8393	0.4893	0.9568	0.8680	0.5342	0.8263	0.5500	0.9205	0.8472	0.5129
	(MACCS+MD)-*k*NN	0.8058	0.5579	0.889	0.8177	0.4656	0.8051	0.6500	0.8580	0.8255	0.4975
	(MACCS+MD)-SVM	0.8123	0.4592	0.9308	0.8264	0.4534	0.8220	0.5500	0.9148	0.8511	0.5028
	(PubChem+MD)-*k*NN	0.8079	0.5837	0.8833	0.8314	0.4784	0.7754	0.6500	0.8182	0.8397	0.4445
	(PubChem+MD)-SVM	0.8350	0.5150	0.9424	0.8570	0.5250	0.8136	0.5667	0.8977	0.8416	0.4879
5	(Estate+MD)-*k*NN	0.7422	0.7427	0.7417	0.8152	0.4842	0.7288	0.7750	0.6810	0.7955	0.4583
	(Estate+MD)-SVM	0.7131	0.6913	0.7333	0.7771	0.4250	0.7331	0.7250	0.7414	0.7691	0.4663
	(FP+MD)-*k*NN	0.7249	0.7271	0.7229	0.7927	0.4497	0.7500	0.7667	0.7328	0.8243	0.4998
	(FP+MD)-SVM	0.7400	0.7025	0.7750	0.8120	0.4791	0.7627	0.7750	0.7500	0.8349	0.5252
	(FP4+MD)-*k*NN	0.7368	0.7315	0.7417	0.7950	0.4731	0.6102	0.7167	0.5000	0.6546	0.2221
	(FP4+MD)-SVM	0.7389	0.7047	0.7708	0.8079	0.4769	0.7415	0.7333	0.7500	0.8004	0.4833
	(MACCS+MD)-*k*NN	0.7586	0.7539	0.7625	0.8095	0.5163	0.7288	0.7500	0.7069	0.8062	0.4574
	(MACCS+MD)-SVM	0.7691	0.7293	0.8063	0.8286	0.5377	0.7585	0.7583	0.7586	0.8261	0.5169
	(PubChem+MD)-*k*NN	0.7487	0.7517	0.7458	0.8032	0.4972	0.7627	0.7833	0.7414	0.8243	0.5253
	(PubChem+MD)-SVM	0.7605	0.8416	0.6356	0.8270	0.4901	0.7712	0.7750	0.7672	0.8425	0.5422
10	(Estate+MD)-*k*NN	0.7304	0.7989	0.6247	0.7817	0.4292	0.7542	0.8627	0.5542	0.8030	0.4416
	(Estate+MD)-SVM	0.7152	0.7972	0.5890	0.7867	0.3943	0.7627	0.8497	0.6024	0.7947	0.4670
	(FP+MD)-*k*NN	0.7433	0.8096	0.6411	0.8109	0.4566	0.7542	0.8039	0.6627	0.8335	0.4641

	(FP+MD)-SVM	0.7411	0.8327	0.6000	0.7876	0.4400	0.7924	0.8693	0.6506	0.8122	0.5351
	(FP4+MD)-*k*NN	0.7336	0.8185	0.6027	0.8039	0.4323	0.7331	0.8301	0.5542	0.8147	0.3985
	(FP4+MD)-SVM	0.7476	0.8345	0.6137	0.8215	0.4616	0.7500	0.8301	0.6024	0.7729	0.4420
	(MACCS+MD)-*k*NN	0.7325	0.806	0.6192	0.7960	0.4325	0.7203	0.9000	0.5345	0.8052	0.4682
	(MACCS+MD)-SVM	0.7584	0.8310	0.6466	0.8268	0.4869	0.6737	0.9000	0.4397	0.8063	0.3839
	(PubChem+MD)-*k*NN	0.7498	0.8167	0.6466	0.8160	0.4700	0.7415	0.8366	0.5663	0.7935	0.4178
	(PubChem+MD)-SVM	0.7605	0.8416	0.6356	0.8270	0.4901	0.7797	0.8693	0.6145	0.8197	0.5036
30	(Estate+MD)-*k*NN	0.7843	0.8821	0.4417	0.7522	0.3433	0.8347	0.9365	0.4255	0.7638	0.4223
	(Estate+MD)-SVM	0.7961	0.9570	0.2330	0.7334	0.2829	0.8347	0.9735	0.2766	0.7554	0.3764
	(FP+MD)-*k*NN	0.7875	0.8946	0.4126	0.7540	0.3371	0.7881	0.8730	0.4468	0.7579	0.3251
	(FP+MD)-SVM	0.7832	0.9501	0.1990	0.7325	0.2246	0.8475	0.9894	0.2766	0.7988	0.4355
	(FP4+MD)-*k*NN	0.7918	0.8877	0.4563	0.7418	0.3654	0.7966	0.9048	0.3617	0.7469	0.2994
	(FP4+MD)-SVM	0.7994	0.9612	0.2330	0.8002	0.2943	0.839	0.9471	0.4043	0.7631	0.4274
	(MACCS+MD)-*k*NN	0.8015	0.9015	0.4515	0.7607	0.3846	0.8220	0.9259	0.4043	0.7899	0.3802
	(MACCS+MD)-SVM	0.7853	0.9417	0.2379	0.7842	0.2510	0.8347	0.9577	0.3404	0.7855	0.3939
	(PubChem+MD)-*k*NN	0.7972	0.896	0.4515	0.7821	0.3750	0.8051	0.9101	0.3830	0.7562	0.3293
	(PubChem+MD)-SVM	0.7853	0.932	0.2718	0.7421	0.2675	0.8432	0.9577	0.3830	0.8338	0.4345

Table 3 Average and standard deviation (SD) of overall accuracy (Q) and C values for different models based on different thresholds

thresholds (μM)	descriptors	training set				test set			
		Q		C		Q		C	
		Average	SD	Average	SD	Average	SD	Average	SD
1	MD*	0.7666	0.0346	0.3131	0.1266	0.7559	0.0431	0.3270	0.0981
	Fingerprints	0.7689	0.0465	0.3430	0.1224	0.7741	0.0432	0.3846	0.1148
	Fingerprints+MD	0.7797	0.0404	0.3923	0.0805	0.7715	0.0488	0.4081	0.0828
5	MD	0.6522	0.0716	0.3177	0.1183	0.6390	0.0738	0.3016	0.1084
	Fingerprints	0.6902	0.0405	0.3813	0.0804	0.6920	0.0607	0.3872	0.1200
	Fingerprints+MD	0.6998	0.0429	0.3991	0.0843	0.6914	0.0577	0.3876	0.1088
10	MD	0.6677	0.0369	0.2876	0.0962	0.6797	0.0154	0.2754	0.0707
	Fingerprints	0.6957	0.0380	0.3550	0.0801	0.7214	0.0355	0.4027	0.0708
	Fingerprints+MD	0.7050	0.0404	0.3810	0.0791	0.7191	0.0403	0.3984	0.0721
30	MD	0.7819	0.0217	0.3165	0.0202	0.7983	0.0294	0.3017	0.0459
	Fingerprints	0.7736	0.0294	0.2757	0.0720	0.8059	0.0300	0.3225	0.0688
	Fingerprints+MD	0.7810	0.0237	0.3303	0.0524	0.8007	0.0340	0.3353	0.0758

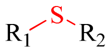
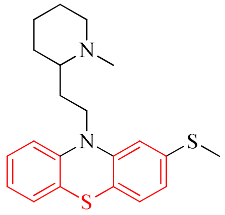
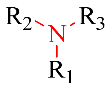
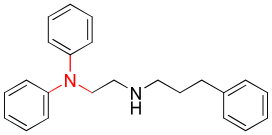
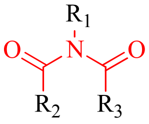
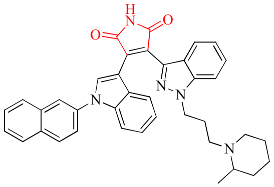
* MD represents molecular descriptors.

Table 4 Performance of classification models for the external validation set based on molecular properties (MD) combining fingerprints

thresholds(μM)	modeling methods	Q	SE	SP	AUC	C	TP	TN	FP	FN
1	(Estate+MD)-*k*NN	0.7445	0.3333	0.8114	0.5714	0.1236	19	284	66	38
	(Estate+MD)-SVM	0.8329	0.1930	0.9371	0.6550	0.1654	11	328	22	46
	(FP+MD)-*k*NN	0.7641	0.4912	0.8086	0.6744	0.2460	28	283	67	29
	(FP+MD)-SVM	0.8059	0.3333	0.8829	0.7033	0.2116	19	309	41	38
	(FP4+MD)-*k*NN	0.7961	0.4211	0.8571	0.7050	0.2503	24	300	50	33
	(FP4+MD)-SVM	0.8354	0.1228	0.9514	0.6984	0.1094	7	333	17	50
	(MACCS+MD)-*k*NN	0.7789	0.6667	0.7971	0.7753	0.3635	38	279	71	19
	(MACCS+MD)-SVM	0.8452	0.3158	0.9314	0.7857	0.2820	18	326	24	39
	(PubChem+MD)-*k*NN	0.7150	0.4737	0.7543	0.6826	0.1767	27	264	86	30
	(PubChem+MD)-SVM	0.8501	0.2456	0.9486	0.7179	0.2504	14	332	18	43
5	(Estate+MD)-*k*NN	0.6069	0.6364	0.5889	0.6279	0.2185	98	149	104	56
	(Estate+MD)-SVM	0.5921	0.6169	0.5771	0.6320	0.1881	95	146	107	59
	(FP+MD)-*k*NN	0.5528	0.5649	0.5455	0.5948	0.1071	87	138	115	67
	(FP+MD)-SVM	0.6093	0.5519	0.6443	0.6389	0.1922	85	163	90	69
	(FP4+MD)-*k*NN	0.5921	0.5974	0.5889	0.6308	0.1809	92	149	104	62
	(FP4+MD)-SVM	0.6413	0.5909	0.6719	0.6610	0.2577	91	170	83	63
	(MACCS+MD)-*k*NN	0.5799	0.6623	0.5296	0.6552	0.1869	102	134	119	52
	(MACCS+MD)-SVM	0.6437	0.6753	0.6245	0.6786	0.2909	104	158	95	50
	(PubChem+MD)-*k*NN	0.5971	0.6364	0.5731	0.6345	0.2032	98	145	108	56
	(PubChem+MD)-SVM	0.6364	0.5844	0.6680	0.6426	0.2474	90	169	84	64
10	(Estate+MD)-*k*NN	0.6216	0.7172	0.4785	0.6314	0.1997	175	78	85	69
	(Estate+MD)-SVM	0.5995	0.7746	0.3374	0.6307	0.1236	189	55	108	55
	(FP+MD)-*k*NN	0.5872	0.6926	0.4294	0.5972	0.1249	169	70	93	75
	(FP+MD)-SVM	0.5971	0.7336	0.3926	0.6029	0.1329	179	64	99	65

	(FP4+MD)-*k*NN	0.6216	0.7213	0.4724	0.6081	0.1982	176	77	86	68
	(FP4+MD)-SVM	0.6192	0.7459	0.4294	0.6356	0.1835	182	70	93	62
	(MACCS+MD)-*k*NN	0.6118	0.7254	0.4417	0.6259	0.1727	177	72	91	67
	(MACCS+MD)-SVM	0.6241	0.7828	0.3865	0.6520	0.1838	191	63	100	53
	(PubChem+MD)-*k*NN	0.6044	0.7295	0.4172	0.5950	0.1529	178	68	95	66
	(PubChem+MD)-SVM	0.6020	0.7008	0.4540	0.6289	0.1579	171	74	89	73
30	(Estate+MD)-*k*NN	0.8084	0.8193	0.5000	0.7600	0.1497	322	7	7	71
	(Estate+MD)-SVM	0.8501	0.8626	0.5000	0.7870	0.1851	339	7	7	54
	(FP+MD)-*k*NN	0.7764	0.7913	0.3571	0.6160	0.0660	311	5	9	82
	(FP+MD)-SVM	0.8550	0.8753	0.2857	0.7201	0.0872	344	4	10	49
	(FP4+MD)-*k*NN	0.7961	0.8066	0.5000	0.7197	0.1387	317	7	7	76
	(FP4+MD)-SVM	0.8182	0.8346	0.3571	0.7172	0.0926	328	5	9	65
	(MACCS+MD)-*k*NN	0.8157	0.8244	0.5714	0.7817	0.1842	324	8	6	69
	(MACCS+MD)-SVM	0.8526	0.8702	0.3571	0.7846	0.1203	342	5	9	51
	(PubChem+MD)-*k*NN	0.7887	0.8041	0.3571	0.6511	0.0733	316	5	9	77
	(PubChem+MD)-SVM	0.8182	0.8372	0.2857	0.7561	0.0600	329	4	10	64

Table 5 Some privileged substructures for hERG blockage based on information gain (IG) and frequency analysis

Description	SMARTS	General Structure	IG	Frequency in positive (hERG blockers)	Frequency in negative (hERG nonblockers)	Representative Compound
Diarylthioether	[c][SX2][c]	 R_1=aryl R_2=aryl	0.0018	1.4097	0.3461	 IC_{50} = 0.363µm
Tertiary mixed amine	[NX3H0+0,NX4H1+;$([N]([c])([C])[#6]);!$([N]*~[#7,#8,#15,#16])]	 R_1=alkyl, aryl R_2=alkyl, aryl R_3=alkyl, aryl	0.0014	1.2056	0.6719	 IC_{50} = 0.605µm
Imide acidic	[#6X3;$([H0][#6]),$([H1])](=[OX1])[#7X3H1][#6X3;$([H0][#6]),$([H1])](=[OX1])	 R_1=H R_2=alkyl, aryl	0.0017	1.4639	0.2596	 IC_{50} = 0.19µm

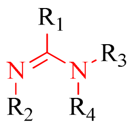
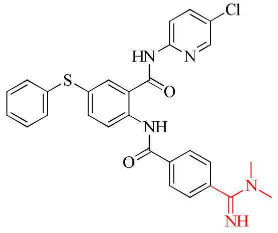
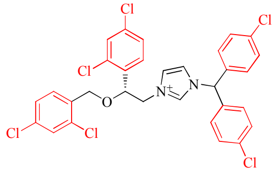
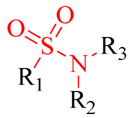
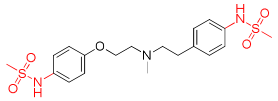
		R_3=alkyl, aryl				
Amidine	[NX3;!$(NC=[O,S])][CX3;$([CH]),$([C][#6])]=[NX2;!$(NC=[O,S])]		0.0044	1.3663	0.4154	
Amidine	[#7X3v3;!$(N([#6X3]=[#7X2])C=[O,S])][CX3R0;$([H1]),$([H0][#6])]=[NX2v3;!$(N(=[#6X3][#7X3])C=[O,S])]	R_1=alkyl, aryl R_2=H, alkyl, aryl R_3=H, alkyl, aryl R_4=H, alkyl, aryl	0.0056	1.4458	0.2884	IC_{50} = 0.089μm
Arylchloride	[Cl][c]	Aryl—Cl	0.0026	1.1725	0.7247	 IC_{50} = 0.03μm
Sulfonamide	[SX4;$([H1]),$([H0][#6])](=[OX1])(=[OX1])[#7X3;$([H2]),$([H1][#6;!$(C=[O,N,S])]),$([#7]([#6;!$(C=[O,N,S])])[#6;!$(C=[O,N,S])])]	 R_1=alkyl, aryl R_2=H, alkyl, aryl R_3=H, alkyl, aryl	0.0008	1.2037	0.6750	 IC_{50} = 0.01μm

Figure Captions

Figure 1. Diversity analysis of hERG blockage data set. (A) Chemical space defined by MW and AlogP for training set and test set. (B) Heat map of molecular similarity plotted by tanimoto similarity index of cluster center molecules of the 100 clusters using MACCS fingerprint.

Figure 2. Distributions of eight molecular properties of AlogP, logD, MW, logS, nROT, nHBA, nHBD, MPSA for hERG blockers and nonblockers chemicals at the threshold of 10 μM. Nonlinear Gaussian curve fitting was appled to analyse frequency distribution of these molecular properties.

Figure 3. Correlations of eight representative chemical descriptors AlogP, logD, MW, logS, nROT, nHBA, nHBD, MPSA versus IC_{50} values of 1163 chemicals.

Figure 1

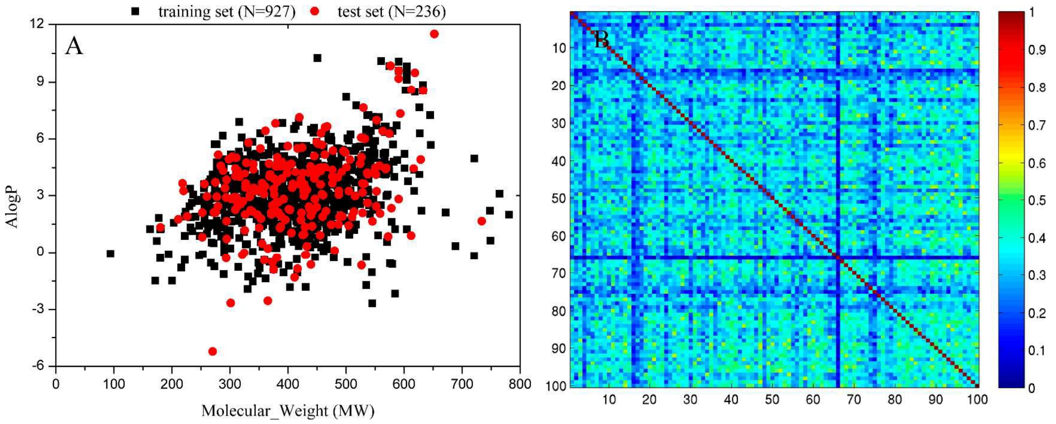

training set (N=927)
test set (N=236)
A
AlogP
12
9
6
3
0
-3
-6
0
100
200
300
400
500
600
700
800
Molecular_Weight (MW)
B
10
20
30
40
50
60
70
80
90
100
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Figure 2

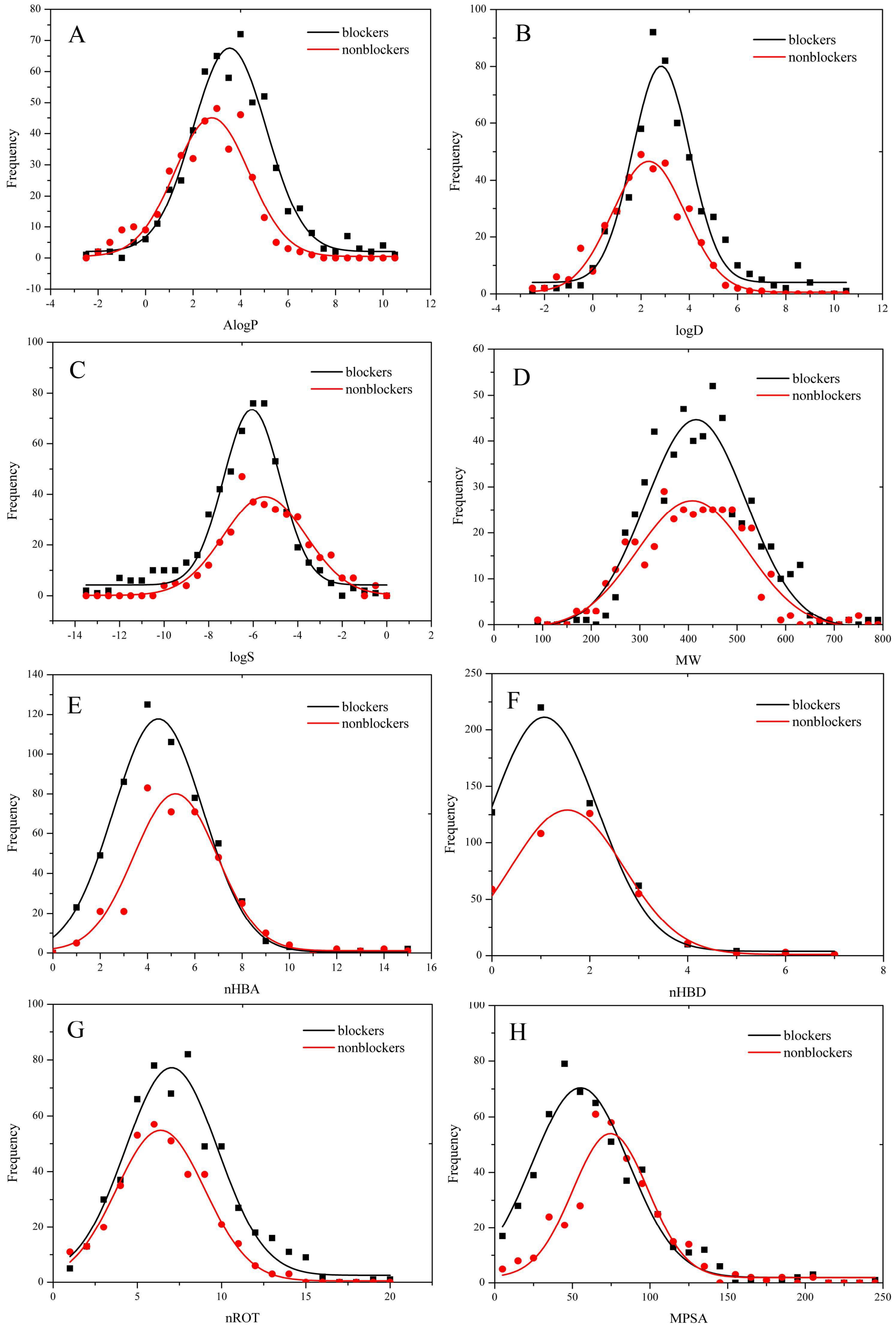
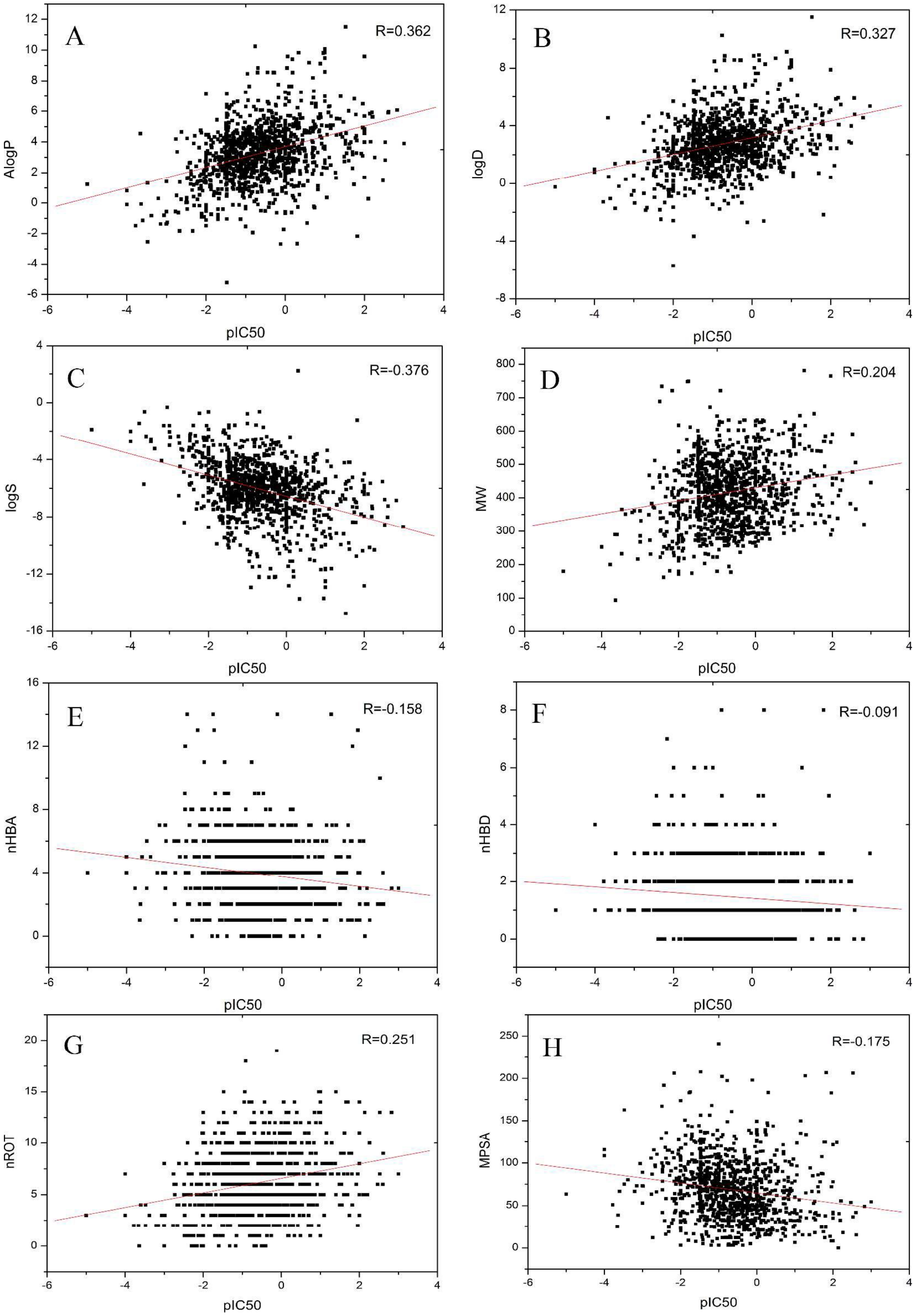
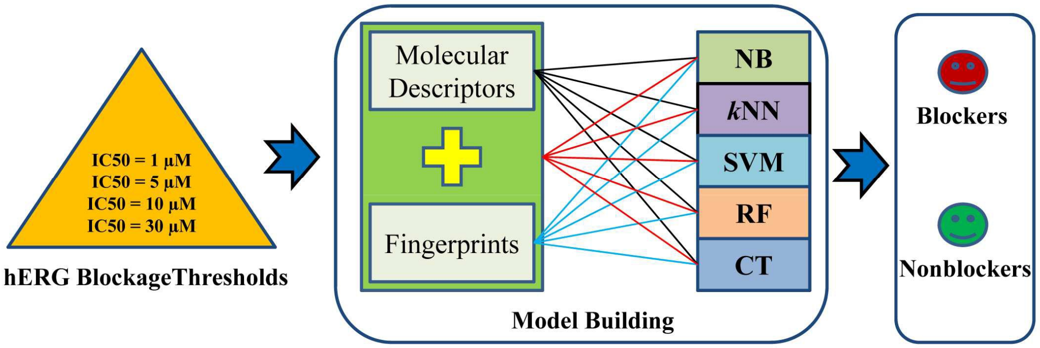

A
blockers
nonblockers
Frequency
AlogP
-10
0
10
20
30
40
50
60
70
80
-4
-2
0
2
4
6
8
10
12
B
blockers
nonblockers
Frequency
logD
0
20
40
60
80
100
-4
-2
0
2
4
6
8
10
12
C
blockers
nonblockers
Frequency
logS
0
20
40
60
80
100
-14
-12
-10
-8
-6
-4
-2
0
2
D
blockers
nonblockers
Frequency
MW
0
10
20
30
40
50
60
0
100
200
300
400
500
600
700
800
E
blockers
nonblockers
Frequency
nHBA
0
20
40
60
80
100
120
140
0
2
4
6
8
10
12
14
16
F
blockers
nonblockers
Frequency
nHBD
0
50
100
150
200
250
0
2
4
6
8
G
blockers
nonblockers
Frequency
nROT
0
20
40
60
80
100
0
5
10
15
20
H
blockers
nonblockers
Frequency
MPSA
0
20
40
60
80
100
0
50
100
150
200
250

Figure 3


A
R=0.362
AlogP
pIC50
B
R=0.327
logD
pIC50
C
R=-0.376
logS
pIC50
D
R=0.204
MW
pIC50
E
R=-0.158
nHBA
pIC50
F
R=-0.091
nHBD
pIC50
G
R=0.251
nROT
pIC50
H
R=-0.175
MPSA
pIC50

Graphical Abstract

Series models of hERG blockage were built using five machine learning methods based on 13 molecular descriptors, five types of fingerprints and molecular descriptors combining fingerprints at four blockage thresholds.


IC50 = 1 µM
IC50 = 5 µM
IC50 = 10 µM
IC50 = 30 µM
hERG BlockageThresholds
Molecular Descriptors
Fingerprints
NB
*k*NN
SVM
RF
CT
Model Building
Blockers
Nonblockers