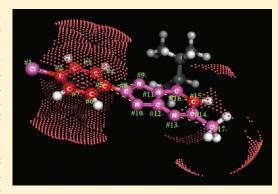# A Comprehensive Support Vector Machine Binary hERG Classification Model Based on Extensive but Biased End Point hERG Data Sets

Meng-yu Shen,[†,#] Bo-Han Su,[†,#] Emilio Xavier Esposito,[‡,§] Anton J. Hopfinger,[§,||] and Yufeng J. Tseng*[,†,⊥]

[†]Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road Taipei, Taiwan 106

[‡]exeResearch, LLC, 32 University Drive, East Lansing, Michigan 48823, United States

[§]The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, Illinois 60045, United States

[||]College of Pharmacy, MSC09 5360 1, University of New Mexico, Albuquerque, New Mexico 87131-0001, United States

[⊥]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No. 1, Sec. 4, Roosevelt Road Taipei, Taiwan 106

ABSTRACT: The human ether-a-go-go related gene (hERG) potassium ion channel plays a key role in cardiotoxicity and is therefore a key target as part of preclinical drug discovery toxicity screening. The PubChem hERG Bioassay data set, composed of 1668 compounds, was used to construct an in silico screening model. The corresponding trial models were constructed from a descriptor pool composed of 4D fingerprints (4D-FP) and traditional 2D and 3D VolSurf-like molecular descriptors. A final binary classification model was constructed via a support vector machine (SVM). The resultant model was then validated using the PubChem hERG Bioassay data set (AID 376) and an external hERG data set by evaluating the model's ability to determine hERG blockers from nonblockers. The external data set (the test set) consisted of 356 compounds collected from available literature data and consisting of 287 actives and 69 inactives. Four different sampling protocols and a 10-fold cross-correlation analysis—used in the validation process to evaluate classification models—explored the impact of the active–inactive data imbalance distribution of the PubChem high-throughput data set. Four different data sets were explored, and the one employing Lipinski's rule-of-five coupled with measures of relative molecular lipophilicity performed the best in the 10-fold cross-correlation validation of the training data set as well as overall prediction accuracy of the external test sets. The linear SVM binary classification model building strategy was applied to different combinations of MOE (traditional 2D, "21/2D", and 3D VolSurf-like) and 4D-FP molecular descriptors to further explore and refine previously proposed key descriptors, identify new significant features that contribute to the prediction of hERG toxicity, and construct the optimal SVM binary classification model from a shrunken descriptor pool. The accuracy, sensitivity, and specificity of the best model determined from 10-fold cross-validation are 95, 90, and 96%, respectively; the overall accuracy is near 87% for the external set. The models constructed in this study demonstrate the following: (i) robustness based upon performance in accuracy across the structural diversity of the training set, (ii) ability to predict a compound's "predisposition" to block hERG ion channels, and (iii) define and illustrate structural features that can be overlaid onto the chemical structures to aid in the 3D structure–activity interpretation of the hERG blocking effect.

INTRODUCTION

The human ether-a-go-go related gene (hERG) potassium is one of the major critical components associated with QT interval prolongation and development of arrhythmia called Torsades de Pointes (TdP). When the corresponding hERG potassium channel is inhibited, a fatal disorder called long QT syndrome[1−4] occurs. Chemical compounds are regularly screened for their hERG toxicity early in the drug discovery process to avoid potential cardiotoxic side effects that would remove a compound from consideration. Therefore, development of robust, sound, and expandable in silico models for predicting hERG potassium channel affinity is high on the list of current computational ADMET goals.

There are in vitro and in vivo bioassays, as well as high-throughput screens, that are widely used to assess the propensity of a compound to block the hERG potassium ion channel. The evaluation of a small number of compounds using such experimental methods is tractable. However, typically, there are several thousand compounds that need to be accurately evaluated, and these experimental methods become prohibitive in terms of both cost and time. To accelerate the drug discovery process and reduce overall costs, the development of reliable in silico hERG models can help to focus the synthesis and testing of non-hERG

blocking compounds that are promising to the therapeutic end point of interest. Many in silico hERG models, using QSAR approaches, have been published to predict if a drug candidate can block the hERG channel.[5−8] Among the applied classification methodologies[9,10] are Bayesian,[11] decision tree,[12] random forest,[13] support vector machine (SVM),[13−16] and partial least-squares (PLS).[13,17]

SVMs employ machine-learning methodologies that construct a hyperplane (a virtual division between *classes* of compounds) in high-dimensional space (a multitude of molecular descriptors) and are typically used for classification, or regression fit, upon a system of data. An advantage of SVMs is the ability to include an arbitrary large set of molecular descriptors to train the projection function (model) rather than imposing any type of limitation in the selection of the descriptors. Thus, SVMs offer the most model-building flexibility across all machine-learning methodologies.

In a previous paper, we reported a binary classification QSAR model[18] based on the genetic function approximation (GFA) methodology[19] that provides predictive performance for hERG channel blockage better than that of other published classification models.[10,20−25] Initially, a continuous QSAR model for hERG was constructed utilizing a collection of 250 compounds from the open literature having accurate and validated IC_{50} values. The continuous QSAR model was converted into a binary classification model by applying a cutoff value to delineate active and inactive compounds. The PubChem data set[26] (AID 376) was subsequently evaluated with the binary classification model as an external test set.

A combination of high-quality experimental data for the training set (IC_{50} biological end point values) and traditional QSAR classification methodology can lead to in silico models that achieve excellent predictive accuracy. To develop a better-performing in silico model for the prediction of hERG channel blockage, as compared to the previous GFA-based strategy described above, SVM binary classification was considered in the work reported here. It should be noted that the construction of a broadly reliable SVM virtual screening model requires a large number of compounds for the training set. Thus, in this study, the hERG high-throughput PubChem (AID 376) data set that was used as a test set in our previous study was employed as the training set.

PubChem (http://pubchem.ncbi.nlm.nih.gov) is an open-access repository for small molecules (chemical structures) and experimental bioassay results (biological activities). It is published and maintained by the National Institutes of Health (NIH) Molecular Libraries Roadmap Initiative and was released in 2004.[27] The PubChem BioAssay database currently contains more than 45 million biological activities for approximately 700 000 unique compounds. With the exponential growth of PubChem's biological screening data, the need for computational methodologies and strategies to mine, analyze, and employ this information rich data source has become an important consideration and goal of any computational and/or cheminformatics study.[28] In this particular study, the PubChem hERG bioassay database (AID 376) contains 1953 compounds forming a biased distribution of 250 active and 1703 inactive compounds and was employed in constructing the hERG activity modeling.

Processing highly skewed (imbalanced) data sets that have been extracted from large high-throughput databases is a significant challenge in statistical analysis and model building and has been the focus of many research articles.[29,30] The imbalanced data set problem for large binary end point data sets arises when one of the classes (active or inactive) is significantly smaller than the other class. In general, for most high-throughput data, irrespective of the type of biological end point, the positive samples (active compounds) are the minority class, and the negative samples (inactive compounds) are the majority class. The high imbalance problem becomes most acute when the goal is to reliably identify a significant class of compounds that has been sparsely sampled. The nature of imbalanced data sets often precludes the use of traditional QSAR modeling methodologies that rely on the selection of important molecular descriptors to construct a predictive linear model over a continuous end point range because the range is poorly sampled. Increasingly, scientists are focusing on developing methods and protocols that employ SVM methods[31−33] and random sampling techniques[34] to compensate for highly imbalanced data sets. However, applying these procedures to an imbalanced data set is not always successful because the measure of success is often highly coupled to fine-tuning the model creation parameters. Recently, Li et al.[35] reported a new protocol to effectively and efficiently address this imbalance problem and applied it to several PubChem high-throughput data sets, most notably in the identification of luciferase inhibitors. We initially applied the Li et al. strategy to analyze AID 376. Unfortunately, this strategy failed to select an appropriate set of samples to further develop a predictive model. The work presented herein discusses the application of various filtering methods to AID 376 that (i) devise the optimal filter criteria and, then, (ii) construct predictive models for the classification of potential hERG-blocking compounds. Additionally, we have formulated our empirical strategies so that they can be applied to other imbalanced sets of data. Specifically, binary classification hERG models were constructed using a SVM[36,37] to predict if a compound will exhibit hERG channel blockage.

■ MATERIALS AND METHODS

Data Sets. The training set has been derived from AID 376,[26] which initially contained 1953 compounds in which the biological end points are represented as the percentage of hERG blockage from a high-throughput screening. Compounds complexed with metal ions, structurally ambiguous compounds (an SDfile entry with two or more compounds), hERG activators, and compounds that are also present in our external data set—culled from the literature—were removed from the training set. A total of 1668 PubChem compounds remained and consisted of 163 active and 1505 inactive compounds. The external test set of 356 compounds was extracted from the literature (Song and Clark,[13] Yoshida et al.,[38] Thai et al.,[39] and Nisius et al.[25]) with biological end points reported as 50% inhibition concentration values (IC_{50}), as determined from in vitro assay experiments, and referred to as the "50% hERG blockage" end point. The range in the literature IC_{50} values of the compounds is from 0.001 to 10 000 μM. The compounds of AID 376, obtained as 2D molecular structures, along with the literature and in-house library of potential hERG blocking compounds, were converted into 3D structures using HyperChem 7.0.[40] The resultant 3D structures were then geometry optimized using HyperChem 7.0s MM+ force field (based on the Allinger MM2 force field[41]).

Training Data Selection Criteria. Four different training data sets were constructed by applying four different criteria to the remaining 1668 compounds from AID 376. The four data sets are as follows: (i) 1668 compounds, no filters; (ii) 1315 compounds that pass the Lipinski's rule-of-five filter; (iii) 876 compounds that pass Lipinski's rule-of-five filter and also a relative lipophilicity filter (logP; logarithm of the octanol−water partition coefficient, P); and (iv) 1010 compounds containing all of the active compounds and those inactive compounds

935

dx.doi.org/10.1021/tx200099j |*Chem. Res. Toxicol.* 2011, 24, 934–949

Table 1. Definitions of the Interaction Pharmacophore Elements, IPEs, Currently Used in the 4D-FP Paradigm

IPE code	IPE abbreviation	IPE description
0	any	all atoms in the moledule
1	np	nonpolar atoms
2	pp	polar (+) atoms
3	pn	polar (−) atoms
4	hba	hydrogen bond acceptor atoms
5	hbd	hydrogen bond donor atoms
6	aro	aromatic atoms
7	hs	nonhydrogen atoms

that pass both the Lipinski's rule-of-five and the relative lipophilicity filter. It is known that increasing the hydrophobicity of drugs also tends to increase the hERG blocking effect and vice versa.[13,24,38,42] The logP constraint discarded active compounds whose logP values are *less than* 4.1 and inactive compounds with logP values *greater than* 2.8 to match the average logP values of active and inactive compounds from our previous binary classification GFA-QSAR model for hERG blockage prediction.[18] The application of the logP constraint led to the training sets that are focused toward the physicochemical requirements of the hERG receptor. These four training sets were used to build an array of SVM models that explore a wide range of SVM parameters for the prediction of hERG blockage. The optimal model was selected from the ensemble of models and evaluated based on the G-means score.

SVM. The concept and implementation of SVM was proposed by Vapnik and co-workers in 1995[36,37] and is a kernel-based supervised machine-learning technique that is well-suited for the separation of compounds into two classes based upon their biological end point measures. In those cases where compounds can be separated by a direct linear functionality (a plane), SVM constructs a hyperplane that separates the two classes of molecules with a maximum margin (distance between the two groups). For cases that are not linear (nonlinearity), the SVM projects the feature vectors (molecular descriptors) onto a high-dimension feature space, similar to a potential energy landscape, and searches for an optimal linear hyperplane in the multidimensional feature space to generate the separation. The SVM model construction employs the traditional QSAR modeling training and test set approach. The SVM is trained using a data set with known classification (active or inactive in this study), and then, the resultant trained model is applied to a data set that was not used to train the model. This strategy provides an external approach to evaluate and validate the SVM model's ability to classify "new" compounds.

In SVM, given a training set $T = \{(x_i, y^i)^N\}$, where $y^i\{+1, -1\}$, the binary classification problem is transformed into the identification of a separation hyperplane with the biological end point a subspace, of dimension N-1, dividing the hyperplane into two halves derived from the inputs (end points). For a linear classification model, the hyperplane is a function of x, $f(x) = (\langle w \cdot x \rangle + b) > 0$, such that

$$y_i[f(x)] = y_i(\langle w \cdot x \rangle + b) > 0 \tag{1}$$

where w is the weight vector, which is perpendicular to the hyperplane, and b is the bias, a scalar value, that determines the offset of the hyperplane from the origin. Overall, the hyperplane function can be expressed as:

$$f(x) = \langle w \cdot x \rangle + b = 0 \tag{2}$$

the w and b are selected to maximize the margin, $2/\|w\|$, subject to

$$y_i[(w \cdot x) + b] \geq 1 \tag{3}$$

Through optimization, the SVM approach tries to define a unique separating hyperplane that partitions the training set compounds (data)

with minimum error while maximizing the margin. By introducing a Lagrangian multiplier, α_i, a unique and optimized solution can be determined as follows:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{4}$$

and

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j \tag{5}$$

The coefficients α_i are obtained by maximizing the Largrange variable, L, subject to the constraints:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \text{ and } \alpha_i > 0 \tag{6}$$

When the coefficients α_i are determined, the final hypothesis (model) is a linear combination of the training data. The decision function is expressed as follows:

$$\text{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \langle x_j \cdot x \rangle + b\right) \tag{7}$$

For nonlinear cases, the SVM can construct a hyperplane by mapping the input vector to a higher dimensional space. Thus, assuming $\Phi(x_i)$ is the function mapping x_i to a high-dimensional space, what is needed for the training set (learning) and the test set (prediction) is $\Phi(x_i) \cdot \Phi(x_j)$ instead of the mapping function Φ itself. Hence, the dot product in eq 7 can be replaced with a selected kernel function, K, to achieve a nonlinear transformation and permitting eq 7 to be written as:

$$\text{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i K(x_i, x) + b\right) \tag{8}$$

A commonly used kernel function is the Gaussian radial basis function (RBF) since it provides good overall performance. The RBF is formulated as:

$$f(u, v) = \exp(-\gamma \| u - v\|^2) \tag{9}$$

where γ is a constant and u and v are two independent variables. Selection of γ greatly influences the amount of time needed to develop a SVM model from the training set data in terms of optimizing the performance and predictive ability of the SVM model.[43] In this study, optimal models were constructed with 10 different γ values: 2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, and 2^{3}.

4D Fingerprints (4D-FPs). The theory and methodology of the universal 4D-FP descriptors have been presented in previous works.[44] The universal 4D-FPs are the eigenvalues of the molecular similarity eigenvectors determined for a given molecule based on a set of absolute molecular similarity main distance-dependent matrices (MDDM). The eigenvectors contain molecular information including a molecule's atom functional type, shape, and conformational flexibility. The atom functional types used to represent a molecule are defined using eight interaction pharmacophore elements (IPE) that are summarized in Table 1. The IPEs were initially introduced in the first 4D-QSAR paper.[45] Construction of the 4D-FP descriptor matrix for the compounds of the training set is determined by maximizing its information content. For each training set compound, the 4D-FPs (eigenvalues) are computed, and the number of eigenvalues in the IPE eigenvector of a molecule is determined for each particular IPE pair (a, b). Each molecule in the training data set is assigned $n_{max}(a, b)$ eigenvalues based upon the largest corresponding eigenvector in the training set for each IPE pair (a, b). If the number of eigenvalues in an eigenvector of a particular molecule is less than $n_{max}(a, b)$, the missing value(s) is set to zero. Each

member of the overall set of 4D-FP IPE (eigenvalues) is considered a molecular descriptor when employed in QSAR modeling studies. A total of 813 4D-FPs descriptors were used in this study including those from all, NP, PP, PN, HBA, HBD, and HS IPE types as defined in Table 1.

Molecular Operating Environment (MOE) Descriptors. The traditional 2D, 21/2D, and VolSurf-like molecular descriptors were calculated as part of this study by MOE 2008.10[46] for inclusion in the trial descriptor pool. The 2D molecular descriptors are the numerical properties evaluated from the connection tables representing a molecule and include physical properties, subdivided surface areas, atom counts, bond counts, Kier and Hall connectivity and κ shape indices, adjacency and distance matrix descriptors containing BCUT and GCUT descriptors, pharmacophore feature descriptors, and partial charge descriptors (PEOE descriptors). A 21/2D molecular descriptor is defined here as a 3D molecular property represented as an individual (singular) numerical value. In this case, the 21/2D molecular descriptors include measures of the conformational potential energy and its components, molecular surfaces, volumes and shapes, and conformation-dependent charge descriptors. All of these descriptors are dependent on the conformation of the molecule. In this study, there are 230 2D and 21/2D molecular descriptors.

The VolSurf descriptor set contains 76 molecular descriptors that are alignment independent and also not strongly dependent on molecular conformation. VolSurf-like molecular descriptors are a class of molecular descriptors that represent 3D molecular properties as a single numerical value. The compound is placed in a grid (with the exception of four VolSurf descriptors), a hydrophobic (dry) and hydrophilic (wet) probe visits each grid point, and the interaction energy between the probe and the compound is calculated. The grid points within an interaction energy range are considered an iso-contour (iso-surface), and the volume is calculated. The calculated volumes and combinations of interaction energies and volumes are used as molecular descriptors. The four nongrid VolSurf descriptors measure the molecular volume, surface area, globularity, and rugosity.

Model Evaluation. To evaluate the predictive performance of the SVM models, classification accuracy (total percentage correctly predicted), sensitivity (also referred to as recall or the true-positive rate is the percentage of active compounds correctly predicted), and specificity (also known as the false-positive rate is the percentage of inactive compounds correctly predicted) are defined as follows:

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fn} + \text{tn} + \text{fp}} \quad (10)$$

$$\text{sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (11)$$

$$\text{specificity} = \frac{\text{tn}}{\text{tn} + \text{fp}} \quad (12)$$

In eqs 9−11, tp is the number of true positives (active compounds that are correctly predicted), fn is the number of false negatives (active compounds that are incorrectly predicted to be inactive), tn is the number of true negative (inactive compounds that are correctly predicted), and fp is the number of false positive (inactive compounds that are incorrectly predicted to be active). Sensitivity and specificity are good individual measures, with respect to activity and inactivity, of a model's ability to correctly classify the compounds of the training and test sets. Combining sensitivity and specificity into a single numerical value via the geometric mean (G-mean) function provides a simple measure that indicates the extent to which a model is able to correctly predict the classification of both active and inactive compounds, as well as a convenient metric to quickly select optimal models. The G-mean value is defined as follows:

$$\text{G-mean} = \sqrt{\text{sensitivity} \times \text{specificity}} \quad (13)$$

A good hERG prediction (classification) model should minimize the possibility of misclassifying both active and inactive compounds. Therefore, the G-mean value has been selected as a criterion to emphasize the joint performance of sensitivity and specificity.

10-Fold Cross-Validation. The pruned AID 376 data set was randomly partitioned into 10 equal subsets. One of the 10 subsets was retained as the test set to validate the model. The remaining nine subsets were collectively used as the training set to construct models using the LIBSVM tool.[47] This procedure was repeated 10 times systematically using each of the 10 subsets as a unique test set. The results from the set of 10 10-fold cross-validations were then combined into a single model representation.

Additional Sampling Methods Considered for Selecting Training Sets. Li et al.[35] utilized the granular SVM; the repetitive under sampling (GSVM-RU) method[48] was applied to 147 324 luciferase inhibitors from a set of PubChem bioassays (AID, 773; AID, 1006; and AID, 1379) to overcome the data set's imbalance. The GSVM-RU method iteratively selects a subset of the majority classification set (the inactive) while retaining all of the members of the minority classification set (the active compounds of the training set). The GSVM-RU method, in its iterative mode, extracts the important compounds (those that provide a significant contribution to the model) from the data set. In each iterative round, the important inactive compounds, the support vectors (SVs), are removed from the data set of the previous round. The remaining smaller data set is then used to construct a new SVM model and to identify the important inactive compounds for the next round. Finally, all of the deleted inactive SVs from the previous iterative rounds are combined with all of the active compounds to define the training set for building the final SVM model.

Guha and Schürer developed a data sampling protocol to treat imbalanced data sets and applied it to 775 compounds from the PubChem bioassays (AID, 364; AID, 463; and AID, 464), 1334 compounds in a human T-cell proliferation data set from NCGC and 103 041 animal toxicity measures from the MDL toxicity database.[34] The ratios of active to inactive compounds are 1:2, 1:12, and 1:25 for PubChem, NCGC, and MDL data sets, respectively. In this protocol, the active compounds of a data set are randomly divided into two classes; 80% of the active compounds (underrepresented) are placed in the training set, and the remaining 20% of the active compounds are assigned to the test set. Inactive compounds (overrepresented) are distributed in a similar fashion to the training and test sets. The composition of the training set is completed by randomly selecting inactive compounds until the number of inactive compounds, added to the training set, equals the number of active compounds. The random creation of a balanced training set (equal number of active and inactive compounds), and the correspondingly ratio-preserving test set is repetitively performed (30 times as reported by Guha and Schürer) to ensure that the inactive compounds are adequately sampled. Finally, the consensus predictions for the ensemble of (30) training-test set collections are analyzed together. The reported protocols used to treat imbalanced data sets, as described above, were considered in this work to develop balanced training sets of non-hERG blockers as inactives and hERG blockers as actives.

■ RESULTS AND DISCUSSION

Four different training data sets were constructed by applying four different criteria to the resultant 1668 compounds of PubChem's hERG bioassay data set along with two other data sets; one constructed in a similar fashion as Li et al.[35] and the other in a similar fashion as Guha and Schürer.[34] The four data sets formed from applying the different criteria resulted in (i) the raw data set of 1505 actives and 163 inactives (1668 total compounds) with no filters applied; (ii) 105 actives and 1210 inactives [1315] compounds that pass the Lipinski's rule-of-five

Table 2. Average Molecular Similarity for Each of the Same Type of IPE Pairs Based on 4D-Molecular Similarity for the Four Different Data Sets of This Study

IPE type		data set			
		1	2	3	4
any—any	active	0.71	0.74	0.76	0.71
	inactive	0.66	0.69	0.70	0.70
np—np	active	0.60	0.63	0.70	0.60
	inactive	0.54	0.56	0.56	0.56
pp—pp	active	0.28	0.28	0.38	0.28
	inactive	0.31	0.34	0.38	0.38
pn—pn	active	0.45	0.49	0.48	0.45
	inactive	0.47	0.52	0.54	0.54
hba—hba	active	0.41	0.46	0.50	0.41
	inactive	0.46	0.51	0.53	0.53
hbd—hbd	active	0.37	0.37	0.46	0.37
	inactive	0.34	0.35	0.36	0.36
hs—hs	active	0.73	0.76	0.79	0.73
	inactive	0.66	0.69	0.69	0.69

Table 3. Performance of SVM Models for the Training Set Adopting the logP and Lipinski's Rule-of-Five Constraints from Different r Parameters
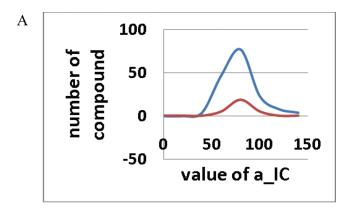
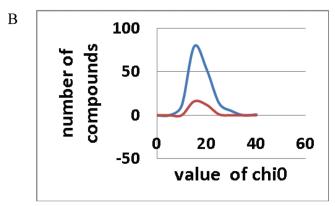| accuracy (%)	sensitivity (%)	specificity (%)	G-mean (%)	γ
95.1	**86.2**	**95.4**	**90.7**	2^{-15}
97.4	82.8	97.9	90.0	2^{-11}
95.1	82.8	95.5	88.9	2^{-13}

When applying the Lipinski's rule-of-five constraints to the raw training set, creating data set 2, the average 4D-MS similarity across all IPE types among members of data set 2 increases only 0.03 when compared to the molecular similarity for data set 1. Presumably since the molecular similarity is not significantly different between data set 1 and data set 2, the SVM protocol cannot create a significantly better hyperplane for one data set as compared to the other. Therefore, the overall accuracy of the resultant model only increases slightly for data set 2 (77%) relative to data set 1 (69%).

SVM Model Building by Adopting the logP Filter. A logP filtering constraint, suggested from our previous hERG blocking GFA modeling[18] was applied to data set 2 following the Lipinski's rule-of-five filter. This resulted in a set of 876 compounds designated as data set 3 that contains 29 active and 847 inactive compounds. The three best-performing SVM models derived from data set 3 are listed in Table 3 and ranked by their G-mean values. The five columns report the accuracy, sensitivity, specificity, G-mean, and γ values, respectively. The best SVM model constructed by this protocol achieved 95% accuracy for the 10-fold cross-validation with a G-mean of 91%. Although the tenth model in Table 3 returned a higher accuracy than the other models, the sensitivity of the defined best model is greater than that of the tenth model.

The best SVM model, constructed using data set 3, was used to virtually screen the literature data set serving as a test set. An IC_{50} cutoff value of 10 μM was adopted to divide the test set into two classes: actives and inactives. The resulting accuracy and G-mean values for this test set are 69 and 66%, respectively. Increasing the active/inactive cutoff value to 40 μM improved the hERG blockage predictions and resulted in an overall accuracy of 75% and a G-mean value of 76%. Furthermore, the 40 μM IC_{50} cutoff value, when applied to the literature test set, leads to a better predictive model than for other cutoff values. This finding demonstrates that the best SVM model built from data set 3 is reliable and robust for classifying the hERG toxicity of druglike compounds. The removal of "noisy compounds" from the imbalanced training set, by employing the physical property logP filter in addition to the Lipinksi's rule-of-five filter, effectively "cleans" the data sets and correspondingly increases the classification performance of the models, especially when employing SVM strategies.

The 4D-FP molecular similarity measures across all IPE types increase slightly, 0.07 (0.05 on average), for the active and inactive compounds, respectively, as compared to data set 1 (raw training set) and data set 3. Thus, applying this filter to condense the training set still preserves the overall molecular character found in the raw training set. Figure 1 provides the frequency distributions of the active compounds in data set 1 (blue line) and the active compounds in data set 3 (red line) for two 2D and one 21/2D MOE descriptors, namely, IC (atom information content), chi0 (atomic connectivity index), and ASA
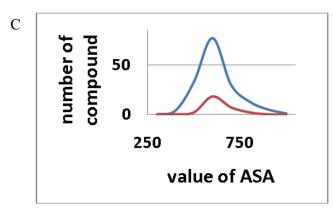
filter; (iii) 29 actives and 847 inactives [876 compounds] that pass both the Lipinski's rule-of-five filter and the logP filter; and (iv) 1010 compounds in which all 163 actives are retained but only the 847 inactives that pass both the Lipinski's rule-of-five and the logP filter are used. The effect of applying filter(s) is characterized by comparing the overall molecular similarities of each filtered [reduced] data set with the raw PubChem training set. The average molecular similarities, using the 4D-molecular similarity (4D-MS) paradigm,[45] for the four training sets were used to explore and characterize the molecular similarity between the active and the inactive compounds within a data set. The resultant molecular similarity findings are reported in Table 2. The optimal SVM models from each data set are discussed below.

SVM Model Building Using the Raw Training Set. The complete PubChem bioassay training set (the raw training set; referred to as data set 1) contains 163 hERG blocking compounds and 1505 hERG inactive compounds resulting in a ratio of 1:9, actives to inactives. The best raw training set model was selected based on the G-mean metric. The G-mean of the raw training set for the 10-fold cross-validation using a SVM was 71% along with an accuracy of 69%. This imbalanced data set, with respect to the greater number of inactive compounds, might bias the determination of the SVM hyperplane and, thereby, correspondingly adversely affect classification behavior. This reasoning is consistent with the underlying SVM algorithm that tries to "disentangle" the two classes of compounds by maximizing their distances of separation on the hyperplane.

SVM Model Building by Adopting the Lipinski Rule-of-Five Filter. As part of the exploration for treating imbalanced data sets, the Lipinski's rule-of-five filter was applied to the PubChem raw training set. A total of 1315 compounds—composed of 105 hERG blockers and 1210 hERG nonblockers—successfully passed through the Lipinski's rule-of-five filter to form data set 2. The overall predicted accuracy, using 10-fold cross-validation for the selected best hERG classification SVM model, is 77% with a G-mean of 73%. Among the 105 hERG blockers, 72 were correctly predicted, while 940 of the 1210 non-hERG blocking compounds were successfully classified.

938

dx.doi.org/10.1021/tx200099j |*Chem. Res. Toxicol.* 2011, 24, 934–949

Figure 1. Frequency distribution plots for the (A) a_IC, (B) chi0, and (C) ASA descriptors of active compounds. The blue and red lines represent the frequencies of the active compounds in data set 1 and data set 3, respectively.

(water accessible surface area). These three molecular descriptors have similar distributions for both data set 1 and data set 3. By filtering the active compounds using the logP and Lipinski's rule-of-five constraints, the original raw training set distribution is retained, but a majority of outliers are removed, which improves the ability of the classification SVM model. The outliers are compounds that can be considered noisy end points from the high-throughput data set and/or the compounds that have significantly different structural features.

SVM Model Building by Keeping All Active Compounds. As indicated in previous reports,[35] an obvious strategy for handling an imbalanced training data set problem is to find a way to adjust the proportion of the active-to-inactive samples to approach 1:1. Therefore, data set 4 was constructed, containing

1010 compounds and was composed of all 163 active compounds (hERG blockers) and 847 inactive compounds that passed Lipinski's rule-of-five (filter for data set 2) and the logP constraint (filter for data set 3). This combination of filters reduced the ratio of active-to-inactive compounds to 1:5 from an original ration of 1:9. Unexpectedly, the performance of the model from this training set was not greater than that using the model constructed with data set 3. The best data set 4 training set SVM model has an accuracy of 88% with a G-mean value of 85% for the training data set. The overall performance of the data set 4 training set is inferior to that of data set 3. The predicted accuracy and G-mean for the literature test set are 84 and 53%, respectively, when using an active/inactive cutoff value of 40 μM. On the basis of the evaluation of the best SVM model constructed from data set 4, it can be concluded that this model is not as resilient when compared to the other SVM models. In particular, the model from data set 4 has a markedly low specificity (ability to correctly predict nonactive compounds) for the literature test set.

Retaining all of the active compounds in data set 4 negatively biases the SVM optimization of the hyperplane and resulted in decreasing the hERG classification accuracy. These findings show that blindly reducing the ratio of active and inactive compounds, such as keeping all of the minority compounds as suggested in the literature,[35] is a debatable strategy for the construction of classification models. On the basis of the findings presented above, the appropriate strategy is to remove as many of the noisy features (compounds) from the data set through the careful inspection of the relevant properties within the data set.

Application of Two Sampling Methods for Treating "Imbalanced" Data Sets. The GSVM-RU methodology was applied to the pruned PubChem hERG data set of 1505 inactive and 163 active compounds in an attempt to construct a reduced data set. However, after 13 iterations of removing important support vectors (compounds), all of the inactive compounds were excluded, and this sampling strategy failed. In Li et al.'s study,[35] all active compounds were reserved and applied the GSVM-RU protocol to screen an extensive data set containing 146 934 inactive compounds. After 67 iterations, their training set contained 358 inactive compounds and 390 active compounds. On the basis of this limited analysis, we can ascertain that the GSVM-RU methodology is most appropriate for treating extremely large and imbalanced data sets. Additionally, GSVM-RU may not have the separation robustness required for the application to relatively small and imbalanced data sets and is thus not suitable for use in this particular study.

Guha and Schürer's protocol was also applied in this study to the 1668 compound hERG data set to build 30 classification SVM hERG models. Again, the training set contains 130 active and 1538 inactive compounds. For the 30 proportionally equal training sets, the classification SVM models' have an average accuracy of 70% with a G-mean value of 70%. For the 338 compound test set, which contains 33 active compounds and 305 inactive compounds, the average overall accuracy predicted by the 30 SVM models reduces to 65% and a G-mean value of 71% for the test sets.

Overall, the GSVM-RU protocol is found not to be applicable for the hERG data of this study, and the Guha and Schürer's protocol yields models of lower accuracy and G-mean values when compared to our SVM model constructed using all active and inactive compounds. Thus, these findings indicate that protocols that construct both proportionally equal training sets and ratio-preserving test sets (resampling) may not produce

Table 4. Prediction Performance Measures from the Top Five Training Set SVM Models Determined Only Using 4D-FP Descriptors[a]

training set percentage (number correct/total number)				test set (40 μM) percentage (number correct/total number)			
accuracy	sensitivity	specificity	G-mean	accuracy	sensitivity	specificity	G-mean
95 (831/876)	79 (23/29)	95 (808/847)	87.0	65 (231/365)	60 (172/287)	86 (59/69)	71.6
95 (829/876)	79 (23/29)	95 (806/847)	86.9	66 (233/356)	60 (173/287)	87 (60/69)	72.4
94 (826/876)	79 (23/29)	95 (803/847)	86.7	63 (224/356)	57 (162/287)	90 (62/69)	71.2
94 (825/876)	79 (23/29)	95 (802/847)	86.7	67 (237/356)	62 (178/287))	86 (59/69)	72.8
89 (781/876)	83 (24/29)	89 (757/847)	86.0	83 (296/356)	90 (258/287)	55 (38/69)	70.4

[a] The models are ranked using the G-mean metric and have been evaluated using data set 3.

Table 5. Prediction Performance Measures from the Top Five Training Set SVM Models Determined Only Using MOE Descriptors[a]

training set percentage (number correct/total number)				test set (40 μM) percentage (number correct/total number)			
accuracy	sensitivity	specificity	G-mean	accuracy	sensitivity	specificity	G-mean
99 (865/876)	90 (26/29)	99 (839/847)	94.2	26 (94/356)	9 (26/287)	99 (68/69)	29.9
98 (862/876)	90 (26/29)	99 (836/847)	94.1	28 (98/356)	11 (30/287)	99 (68/69)	32.1
96 (843/876)	90 (26/29)	97 (817/847)	93.0	31 (111/356)	16 (45/287)	96 (66/69)	38.7
96 (840/876)	90 (26/29)	96 (814/847)	92.8	32 (115/356)	16 (47/287)	99 (68/69)	40.2
95 (834/876)	90 (26/29)	95 (808/847)	92.5	36 (128/356)	21 (61/287)	97 (67/69)	45.4

[a] The models are ranked using the G-mean metric and have been evaluated using data set 3.

significant improvement for the classification of hERG blocking compounds.

Characterization of the Contributions of Molecular Descriptors to the Classification Models. Three trial descriptor pools previously employed in building linear hERG blockage classification models[18] were employed to classify test set compounds using the SVM models. The three trial descriptor pools are as follows: (i) 4D-FPs, (ii) the MOE molecular descriptors (traditional and VolSurf-like), and (iii) the combined sets of 4D-FPs and MOE descriptors. Because data set 3 resulted in the optimal SVM model for the classification of the training set compounds, this model was also used to evaluate the hERG classification models derived from these three unique descriptor pools.

Table 4 contains the training set predictions for binary classification SVM models constructed using data set 3 when the descriptor pool is only composed 4D-FPs. The top five hERG classification models from the training set are ranked by their G-mean values, and the accuracy, sensitivity, specificity, and G-means values are listed in the first four columns, while the last four columns provide these same measures for the test set. For the best SVM model, the overall accuracy is near 95%, with a sensitivity value of 79%, a specificity value of 95%, and a G-mean value of 87%. Moreover, the G-mean value for the top five models constructed from the 4D-FPs descriptors ranges from 86 to 87%. The predicted accuracy, sensitivity, and specificity values for the test set using the top five models range from 63 to 83%, 57 to 90%, and 55 to 90%, respectively, while the G-mean values are nearly constant ranging between 70 and 73%. These findings show that the SVM models constructed using only 4D-FPs provide very good performance for both the training and the test sets.

The identical performance evaluation to that done for the 4D-FPs, and described above, was carried out for only the MOE

molecular descriptors; the results are presented in Table 5. Interestingly, the classification power of the best SVM model, based on G-mean values, achieves an accuracy of 99% (precisely 98.7%). However, only 26 out of the 287 active compounds in the test set are correctly classified by the best SVM model despite the model's ability to correctly classify inactive test set compounds (non-hERG blockers) with a specificity value of 99%. Furthermore, all top five SVM models built from only the MOE descriptors performed poorly when classifying active (hERG blocking) compounds in the test set with sensitivity values between 9 and 21%. This is a stark contrast to the sensitivity value for the training set model of 90%. These same top five models were all very good at classifying inactive compounds in the test set returning specificity values between 96 and 99%. The incongruity between the ability of the MOE-based SVM model to correctly predict training set active compounds while incorrectly classifying active compounds in the test set is a clear indication of data overfitting. That is, the model is made to correctly fit the training set but loses general predictive ability and performs poorly on nontraining set data. From this analysis, it appears that the MOE molecular descriptors of the SVM models make a vital contribution to correctly identifying inactive compounds, but they play a marginal role in capturing molecular features important to active compounds. This is an interesting conundrum because the MOE molecular descriptors were found to play an essential role in the development of successful hERG GFA classification models.

The 4D-FPs provide significant novel fitting data to the SVM methodology that is used to optimize the separation of a hyperplane that

(a) effectively divides the hERG data set into definitive classes and
(b) introduces structural sensitivity to the SVM models for both active and inactive hERG compounds. Combining

Table 6. Twenty-Nine Most Significant Molecular Descriptors Extracted from the Linear SVM Model

descriptor symbols	description of the descriptors
εx (all, NP) ($x = 28$)	pairs of all and nonpolar IPE atom types that are, on average, more than 20 Å apart within the compound
εx (all, PP) ($x = 14$)	pairs of all and polar positive IPE atoms types that are, on average, more than 12 Å apart within the compound
εx (HS, HS) ($x = 3, 22, 23$)	pairs of nonhydrogen and all nonhydrogen IPE atoms types that are, on average, 4–6 Å ($x = 3$) and greater than 20 Å ($x = 22, 23$) apart within the compound
εx (HS, NP) ($x = 31, 34, 35$)	pairs of nonhydrogen and nonpolar IPE atom types that are, on average, more than 20 Å apart within the compound
εx (NP, HBA) ($x = 1$)	pairs of nonpolar and hydrogen bond acceptor IPE atom types that are 3–4 Å apart within the compound
εx (NP, HBD) ($x = 4$)	pairs of nonhydrogen and all hydrogen bond donor IPE atoms types that are, on average, 5–7 Å apart within the compound
εx (NP, NP) ($x = 28, 34, 39, 41, 60$)	pairs of nonpolar and nonpolar IPE atom types that are, on average, more than 20 Å apart within the compound
εx (NP, PP) ($x = 10, 12, 13, 14$)	pairs of nonpolar and nonpolar IPE atom types that are 12–15 Å apart within the compound
b_ar	number of aromatic bonds within a compound
reactive	indicator of the presence of reactive groups; a nonzero value indicates that the compound contains a reactive group
a_nCl	number of chlorine atoms
a_nS	number of sulfur atoms
logP(o/w)	Log of the octanol/water partition coefficient
lip_violation	number of violations of Lipinski's rule-of-five
a_base	number of basic atoms
vsa_base	approximation to the sum of VDW surface areas of basic atoms ($Å^2$)
logP_VSAX (X = 0, 1, 2, ..., 9)	sum of the approximate accessible van der Waals surface area over all atoms i where the logP(o/w) value for atom i is in the specified range of SlogP values;[51] for example, the SlogP value of logP_VSA6 is defined in (0.20, 0.25]
PEOE_VSAX (X = +6, +5, +4, ..., +0, −0, −1, −2, ..., −6)	sum of the approximate accessible van der Waals surface area over all atoms i where the logP(o/w) value for atom i is in the specified range of PEOE (the partial equalization of orbital electronegativities method for calculating atomic partial charges) values;[52] for example, the PEOE value of PEOE_VSA-0 is defined in [-0.05, 0.00)

the 4D-FPs and MOE molecular descriptors resulted in an enhancement in the classification accuracy across a wider range of structural diverse compounds than is achieved using only single class of descriptors.

The 10-fold cross-validation protocol yields accuracy, sensitivity, specificity, and G-mean values for data set 3 of 95, 86, 95, and 91%, respectively. Corresponding test set accuracy, sensitivity, specificity, and G-mean values are, respectively, 75, 74, 78, and 76%. This is yet another indicator that the SVM models constructed using the combined 4D-FPs and MOE molecular descriptor pools for data set 3 provide the optimal binary classification model. The marriage of the molecular mechanics-based 3D molecular descriptors (4D-FPs) and the 2D and 21/2D molecular descriptors forms a trial descriptor pool yielding excellent binary classification models.

Extraction of Features from Linear SVM Model for Descriptor Reduction. Two procedures were adopted to identify the effective model features (molecular properties) and reduce the large descriptor set during SVM model construction: (i) linear SVM modeling and (ii) the independent variable deletion method. Linear SVM modeling deduces the important molecular descriptors by using linear combinations of the descriptors during classification of

the samples (compounds). Applying the linear SVM procedure to data set 3, with the trial descriptor pool containing both the 4D-FPs and the MOE molecular descriptors, led to 29 significant descriptors being identified. This set of 29 significant descriptors contains 19 4D-FP descriptors: one [all, NP], one [all, PP], three [HS, HS], three [HS, NP], one [NP, HBA], one [NP, HBD], five [NP, NP], and four [NP, PP] in terms of IPE types. The significant 2D and 21/2D molecular descriptors are as follows: b_ar, reactive, a_nCl, a_nS, logP(o/w), lip_violation, a_base, vsa_base, logP_VSA6, and PEOE_VSA-0. Descriptions of the 29 significant molecular descriptors are given in Table 6.

To build linear combinations of relevant descriptors, the 19 significant 4D-FP descriptors were expanded to include all eigenvalues (4D-FP) for the corresponding IPE pair types. PEOE_VSA-0 and logP_VSA6 were extended to include the set of 14 subdivided PEOE_VSA descriptors (from −6 to +6) and 10 subdivided logP_VSA (from 0 to 9) molecular descriptors. This was done because the respective "sister" descriptors measure similar molecular properties as the 29 identified significant descriptors, and thus, there is the possibility that they may contain correlated and important information that will contribute to the classification model. The resulting set of 18 linear

Table 7. Most Highly Weighted (Significant) Descriptors, Listed in Descending Order of Their Weights, for the Prediction of hERG Toxicity as Determined by Linear SVM Modeling

descriptor name	weight
logP(o/w)	6.9
b_ar	2.2
a_nCl	1.5
SlogP_VSA6	1.1
ε^* (np, np)	1.0
$\varepsilon1$ (np, hba)	−0.03
$\varepsilon4$ (np, hbd)	−0.43

combinations of 51 molecular descriptors was used to construct a nonlinear SVM model. The accuracy, sensitivity, specificity, and G-mean for this model applied to the training set are 96, 90, 96, and 93%, respectively, and 74, 73, 75, and 74%, respectively, for the test set.

Interestingly, the classifying performance for the training and test sets for this nonlinear SVM model is roughly equivalent to the best SVM model constructed using the combined 4D-FPs and MOE molecular descriptors for data set 3. It can be interpreted that the 51 molecular descriptors capture the important information needed to classify whether or not a compound will potentially block the hERG channel and be termed cardiotoxic. To keep the molecular descriptor pools distinct within these discussions, the linear combination descriptors have been titled Linear MOE and Linear 4D-FPs.

The significant molecular descriptors have been classified according to their weighted contributions to the linear SVM model. The weight of each molecular descriptor in the linear SVM model indicates the contribution that the descriptor makes in the classification of molecules with respect to hERG toxicity. A positive weight (coefficient) for a molecular descriptor increases its contribution to the classification model and thus increases a compound's predicted hERG cardiotoxicity. Conversely, a negative weight decreases the descriptor's overall contribution to the classification model and correspondingly decreases a compound's predicted cardiotoxicity. Overall, seven dominant molecular descriptors were identified, listed in Table 7 in descending order based on the absolute value of their weight. Five of the seven dominant descriptors are activity-enhancing 2D and 21/2D molecular descriptors that relate to a compound's hydrophobicity, and two descriptors are activity-decreasing 4D-FP molecular descriptors. The descriptors that are activity-enhancing (add to a compound's hERG blocking potential) are MOE molecular descriptors, while those that are activity-decreasing are 4D-FPs. This finding is at odds with the behavior of SVM models constructed solely from MOE descriptors that performed poorly in the prediction of active hERG blocking compounds in a blind test set.

A discussion of the important molecular descriptors can be simplified by dividing the descriptors into three categories, namely: (i) physical properties, logP(o/w) and SlogP_VSA6; (ii) atom and bond counts, b_ar and a_nCl; and (iii) spatial shape and size, 4D-FPs that represent nonpolar−nonpolar, nonpolar−hydrogen bond acceptor, and nonpolor−hydrogen bond donor interactive pharmacophore elements (atom types).

Physical Property Descriptors. LogP(o/w), the log of the octanol/water partition coefficient, is the most highly weighted

molecular descriptor and indicates a compound's ability to pass from aqueous environments through hydrophobic membrane barriers. As logP(o/w) of a compound increases, the potency of a compound to be a hERG blocker also significantly increases. The physical property descriptor SlogP_VSA6 is the summation of the water accessible van der Waals surface area for atoms whose contribution to the Crippen−Wildman logP estimate is between 0.2 and 0.25. Increases in the SlogP_VSA6 value indicates that the number of aromatic halides, aliphatic heteroatoms, and amines functional groups have increased. Hence, logP(o/w) and SlogP_VSA6 capture a nearly common physical phenomenon, and thus, these two descriptors can be combined. Moreover, a related molecular descriptor, SlogP_VSA7 (aromatic bridgeheads, quaternary aromatics, C=C aromatics, acids, and ionized nitrogen atoms functional groups), has been reported to have a large impact on increasing hERG blockage.[42] In composite, these molecular descriptors illustrate that the hydrophobic properties of a compound have a significant impact on whether or not the compound has the potential to block the hERG channel.

Atom Counts and Bond Counts Descriptors. The b_ar molecular descriptor counts the number of aromatic bonds and suggests that as the number of aromatic bonds increases for a compound, there is an increasing likelihood the compound is a hERG blocker. The b_ar descriptor is related to the logP(o/w), SlogP_VSA6, and SlogP_VSA7 descriptors since a compound's hydrophobic nature increases as the number of aromatic bonds increases. The other atom count molecular descriptor, a_nCl, is the number of chlorine atoms present in the compound. Aromatic chlorines are quite hydrophobic and given the atomistic definitions of functional groups that contribute to SlogP_VSA6 and thus logP(o/w). The chance of a compound being predicted as cardiotoxic (hERG active) increases as the number of chlorine atoms in a compound increases. Song and Clark[13] have reported that lipophilic fragments, such as benzyl and the chloronaphthyl, increase hERG channel binding. Hence, the identification in this study of b_ar and a_nCl as key descriptors correlates with Song and Clark's findings. Interestingly, the individual correlation coefficients (r) of logP(o/w) versus SlogP_VSA6 ($r = 0.28$), logP(o/w) versus b_ar ($r = 0.42$), and SlogP_VSA6 versus b_ar ($r = 0.18$) are not very strong and indicate that these descriptors are not very intercorrelated and thus provide distinct hydrophobic information.

Spatial Shape and Size Descriptors. The 4D-FP descriptors are classified as spatial shape and size descriptors. There are many (np, np) IPE types whose absolute weight is close to 1.0. Therefore, all of the (np, np) IPE types were included in the analysis. The IPE eigenvector ε^*(np, np) suggests that as the number of nonpolar atoms in a compound increases so does the potential for that compound to possess hERG blocking activity. The two significant 4D-FPs that have negative weights are $\varepsilon1$(np, hba) and $\varepsilon4$(np, hbd). The smaller the value of the eigenvalue index, the number x in εx(IPE, IPE), the less is the average distance separating the IPE pairs. For the $\varepsilon1$(np, hba) 4D-FP, the eigenvalue index of numerical value one indicates that nonpolar and hydrogen bond acceptor atoms are as close as possible to one another, around 3.0 Å. It also implies that the greater the number of pairs of nonpolar and hydrogen bond acceptors atoms in close proximity to one another in a compound, the less the propensity for hERG potency. The corresponding 4D-FP descriptor to $\varepsilon1$(np, hba) is $\varepsilon4$(np, hbd), and it also decreases the possibility of a compound being classified as a hERG blocker. $\varepsilon4$(np, hbd)

Table 8. Prediction Performance Measures for Each Classification Model Constructed from the Reduced Set of 4D-FPs Formed by the Deleted One Type of IPE

model	training set percentage (number correct/total number)				test set (40 μm cutoff) percentage (number correct/total number)			
	accuracy	sensitivity	specificity	G-means	accuracy	sensitivity	specificity	G-means
No_HBA	98 (857/876)	86 (25/29)	98 (832/847)	92.0	67 (239/356)	63 (181/287)	84 (58/69)	72.8
No_HBD	95 (828/876)	86 (25/29)	95 (803/847)	90.4	86 (307/356)	88 (253/287)	78 (54/69)	83.1
No_HS	98 (856/876)	90 (26/29)	98 (830/847)	93.7	74 (262/356)	71 (205/287)	83 (57/69)	76.8
No_NP	98 (858/876)	90 (26/29)	98 (832/847)	93.8	74 (262/356)	71 (204/287)	84 (58/69)	77.3
No_PN	97 (850/876)	90 (26/29)	97 (824/847)	93.4	74 (262/356)	72 (208/287)	78 (54/69)	75.3
No_PP	99 (865/876)	90 (26/29)	99 (839/847)	94.2	69 (247/356)	67 (191/287)	81 (56/69)	73.5

Table 9. List of the Prediction Performance Metrics of the SVM Models Constructed from the Different Sets of Descriptors

descriptors	training result percentage				testing result (40 μM) percentage			
	acc	sen	spe	G-means	acc	sen	spe	G-means
linear MOE + linear 4D-FPs	96	90	96	92.8	74	73	75	74.3
linear MOE + linear 4D-FPs + selected 4D-FPs	96	83	97	89.4	70	68	80	73.6
linear MOE + all 4D-FPs	97	83	91	89.7	64	57	91	72.2
all MOE + linear 4D-FPs + selected 4D-FPs	98	86	99	92.3	67	65	77	70.6
all MOE + linear 4D-FPs	99	86	99	92.4	67	65	75	70.1
all MOE + selected 4D-FPs	96	90	96	92.6	87	90	74	81.4

represents pairs of nonpolar and hydrogen bond donor atoms that are, on average, from 3 to 5 Å apart within the compound.
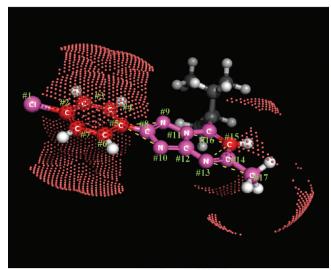
To confirm the significance of a 4D-FP descriptor of a particular IPE type in a SVM model, 4D-FP descriptors containing the IPE type were removed in a round-robin fashion from the trial set of 4D-FP descriptors, and a new SVM model was constructed using data set 3 as the training set. Table 8 lists the predicted performance of each SVM model using all of the MOE descriptors and the reduced set of 4D-FP as a trial descriptor pool. The model name in the first column indicates the IPE type of 4D-FP descriptors that were deleted, and the predicted accuracy, sensitivity, specificity, and G-mean are listed in the second through the fifth columns for the training set and for the literature test set in the sixth through ninth columns. The predicted G-mean value is 76% for the literature test set using the best SVM model constructed employing all of the 4D-FPs and MOE descriptors as a trial descriptor pool and data set 3. The G-mean values for the literature test set decrease when the HBA (G-mean = 73%), PP (G-mean = 74%), and PN (G-mean = 75%) IPE types are excluded from the 4D-FPs trial descriptor pool. This decreasing trend of G-mean values indicates that the molecular information encompassed in these 4D-FPs IPE types contains important information needed for hERG blockage classification. 4D-FPs containing these three IPE types are termed selected 4D-FPs for the remainder of this study.

On the basis of the two reduced descriptor pools, linear MOE and 4D-FPs and selected 4D-FPs, a series of SVM models were built by combining, or extending, these descriptor sets. Table 9 has the data set 3 training set and the literature test set accuracy (acc), sensitivity (sen), specificity (spe), and G-means values for the best SVM models resulting from each of these combinations of trial descriptor sets. The linear MOE + linear 4D-FPs model correctly classifies compounds as active or inactive hERG blockers at a relatively constant level (training set sensitivity = 90% and specificity = 96%; test set sensitivity = 73%; and specificity = 75%).

The addition of the selected-4D-FPs to the linear MOE + linear 4D-FPs model reduces the correct classification of active compounds but increases the ability to classify inactive compounds. With the exception of SVM models constructed from the all MOE + selected 4D-FPs trial descriptor pool, the ability of a SVM model to correctly classify active and inactive compounds in the training and test sets changes markedly with the trial descriptor pool used to construct the model. For most of the combinations of descriptor sets used as trial descriptor pools, the resultant model's ability to correctly classify inactive (non-hERG blockers) increased while its ability to classify potential hERG blockers (the goal of this study) was reduced. The all MOE + selected 4D-FPs SVM model matched the active (90%) and inactive (96%) predictability of the linear MOE + linear 4D-FPs SVM model for the training set and surpassed the active (87 vs 74%) classification performance of the test set with a modest reduction in correctly predicting inactive compounds (74 vs 75%). Overall, the reduced trial descriptors pool consisting of all of the MOE descriptors and the set of selected 4D-FP descriptors leads to the construction of a SVM model from data set 3 that is the best hERG classification SVM model, and this model is discussed in detail below.

Interpretation of Descriptors and Exploration of the Best SVM Model. An operational bonus that comes from the molecular descriptors used to construct the SVM models along with the corresponding model analysis is that pharmacophores and physicochemical properties can be extracted from the SVM models. The pharmacophores and physicochemical properties, in turn, can then be used as guidelines to design and refine drug candidates that should not be hERG active. To effectively illustrate this capability, a strong hERG blocker and a very inactive compound have been selected from the PubChem data set. The compound NCIStruc1_001728, shown in Figure 2 in its 2D chemical structure and 3D (lowest energy conformation) ball-and-stick rendering, is a potent hERG channel blocker with
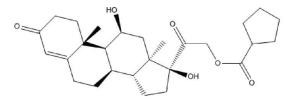
NCIStruc1_001728
% hERG blockade=90%

Figure 2. NCIStruc1_001728 (2D depiction and lowest energy conformation). The structural features, based upon the descriptors given in Table 6, which contribute to increasing hERG blockage, are shown in red. These descriptors that increase hERG activity are portrayed in different shades of red to reflect their relative importance with dark red representing the most significant descriptor. The red dots depict the surface regions of the nonpolar atoms. The numbers along the yellow dotted line indicate interactions between specified pairs of atoms. Atoms of particular significance are highlighted with green numbers.

NSC17245
% hERG blockade=2%

Figure 3. NSC17245 (2D depiction and lowest energy conformation). The structural features, based upon the descriptors given in Table 6, which contribute to decreasing hERG activity, are portrayed in different shades of blue to reflect their relative importance with dark blue, indicating the most highly weighted feature. The red dots define the surface regions contributed by the nonpolar atoms, and the numbers along the yellow dotted lines indicate the distances between the specified pairs of atoms. Atoms of particular significance are highlighted with green numbers. Atoms #9, #10, #13, #15, #24, #25, #26, and #27 define structural features that modestly contribute to increasing hERG binding affinity and, for this reason, are colored red.

an experimental end point (HTS derived) of 90%.[26] The important structural features, based upon the descriptors listed in Table 7, responsible for increasing and decreasing the hERG blockage potency of the compound, are highlighted. All structural features colored red in Figure 2 represent a constructive descriptor term (see Table 7) and an increase in hERG activity, while structural features colored blue indicate the destructive descriptors. The shade of red or blue of the structural feature indicates the weight, or significance, of the corresponding descriptor term. The darker the shading, the more influence the descriptor of the structural feature imparts based on the SVM model weightings.

Although the molecular descriptor logP(o/w) cannot be directly represented as part of the structure, this physical property (hydrophobicity) is similar to that of the descriptor SlogP_VSA6. Thus, the physical characteristics/properties of logP(o/w) and SlogP_VSA6 have been merged together in their portrayal in Figure 2. SlogP_VSA6 defines that portion of the solvent accessible surface area (calculated from a graphical representation of the compound where the surface area is based on the van der Waals radii of the atoms) where the SlogP contribution values of the atoms are between 0.20 and 0.25. This area roughly corresponds to the effective solvent accessible
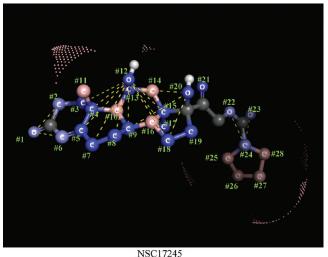
surface area from all of the nonpolar atoms within a compound. The red dots in Figure 2 depict this nonpolar solvent accessible surface area of the molecule, and these regions significantly increase hERG affinity.

The through-space distances between carbon atoms (nonpolar) and nitrogen atoms (hydrogen bond acceptors) for atom pairs #5 and #9, #5 and #10, #15 and #13, and #17 and #13 are approximately 2 Å apart. These atomic distances and interactions are embodied in the negative descriptor, $\varepsilon 1(np, hba)$, which reduces hERG affinity (Table 7), and are correspondingly colored blue.

NCIStruc1_001728 has 16 aromatic bonds from the three aromatic rings. These bonds are colored red and pink and contribute to a significant positive descriptor term b_ar (Table 7). NCIStruc1_001728 also has many nonpolar atoms, including atoms #2, #3, #4, #5, #6, #7, #15, and #17, which are in close proximity to one another and satisfy the least significant positive descriptor, $\varepsilon^*(np, np)$. Because atoms #2, #3, #4, #5, #6, #7, and #15 are part of two constructive descriptors, $\varepsilon^*(np, np)$ and b_ar, these atoms are colored in red, while atoms that only represent one hERG blocking descriptor are colored pink. NCIStruc1_001728 also contains a chlorine atom (atom #1), and on the basis of the importance of the number of chlorine atoms that a compound possesses, molecular descriptor a_nCl increases the possibility of hERG binding. Thus, it is no surprise

that NCIStruc1_001728 is predicted by this study's best SVM model to be a potent hERG channel blocker due to its rich possession of molecular features corresponding to the descriptors [b_ar, n_aCl, SlogP_VSA6, and $\varepsilon^*(np, np)$] that contribute significantly and positively to the compound's affinity to the hERG channel.

NSC17245, shown in Figure 3 in an identical fashion to the renderings in Figure 2, is considered to be an inactive compound having a hERG blockage percentage of only 2% based on PubChem HTS screening data. NSC17245 has four hydrogen bond acceptor atoms (atoms #1, #20, #21, and #23) and two hydrogen bond donor atoms (atoms #12 and #20). This inactive compound has no aromatic bonds. Thus, most of the atoms within rings are classified as nonpolar atoms except for those bonded to oxygen atoms. There are many nonpolar atoms in close proximity to one another (#10, #11, #14, #16, #25, #26, #27, and #28), thus satisfying the least significant (see Table 7) constructive descriptor, $\varepsilon^*(np, np)$, and are colored pink. The remaining nonpolar atoms contribute to the negative descriptors, $\varepsilon1(np, hba)$ and $\varepsilon4(np, hbd)$, and are consequently colored blue. In particular, the oxygen atoms (#1, #22, and #23) are in close proximity to the carbons atoms (#2, #6, and #24). These distance features, associated with the $\varepsilon1(np, hba)$ descriptor, are indicated by yellow dotted lines and diminish the predicted hERG binding affinity of the molecule. Furthermore, many nonpolar atoms of NSC17245 are 3—5 Å away from two oxygen atoms (#12 and #20) and satisfy the descriptor $\varepsilon4(np, hbd)$ that further reduces the compound's hERG blocking potential. There are two sets of nonpolar—hydrogen bond donor atom-pair interactions: One interaction is between the oxygen atom #12 and the nonpolar

atoms #3, #4, #5, #7, #8, #9, #15, #17, and #18, and the other is between the oxygen atom #20 and the nonpolar atoms #8, #12, and #17. These molecular features are the sources of the primary contribution to the most significant negative descriptor, $\varepsilon4(np, hbd)$, and are also shown in yellow as dotted line interaction distances.
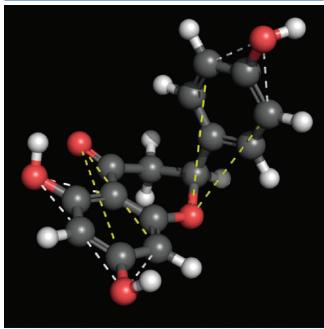
There are two impact levels of structural features having decreasing effects on the predicted hERG binding potency. They are correspondingly colored, based on significance, as light blue for features related to $\varepsilon1(np, hba)$, and dark blue for those structural features associated with the $\varepsilon4(np, hbd)$ descriptor; shown in Figure 3. The large contributions of these two descriptors to reduce NSC17245's hERG blocking potential can also be inferred from Figures 2 and 3 by the relatively small nonpolar surface area for NSC17245 as compared to NCIStruc1_001728. Structural features that decrease hERG activity are dominant in NSC17245, and thus, this compound is not a hERG blocker.

The best SVM model was built using all of the MOE molecular descriptors and selected 4D-FP molecular descriptors applied to the compounds that form data set 3. For the 876 compound training set, 27 of 29 active compounds were correctly classified, and 815 compounds of 847 nonblockers were correctly categorized. Only two of the 29 active compounds from the PubChem data set, NSC7814 and Fentichlor, were misclassified. Interestingly, NSC7814, depicted in Figure 4, was also misclassified in our previous study.[18] This compound is only weakly active and possesses structural features—based on the descriptors given in Table 7—that should reduce hERG potency. These structural features include hydrophilic groups (benzenesulfonyl, hydroxide groups, and amine functional groups) that are chemical fragments associated with non-hERG blocking compounds. The other misclassified compound, Fentichlor, is also a weakly active hERG blocking compound and contains two hydrogen bond donors, two hydrogen bond acceptors, and many nonpolar atoms in aromatic rings that, collectively, increase the values of descriptors $\varepsilon1(np, hba)$ and $\varepsilon4(np, hbd)$, which, in turn, should diminish hERG blockage as suggested in our current and previous[18] studies. Hence, the optimal SVM model of this study predicts that Fentichlor is a non-hERG blocking compound.

For the inactive compounds, 32 of the 847 data set 3 compounds are misclassified as active hERG blockers by the best SVM model. Across the 32 misclassified compounds, there is

NSC7814
% hERG blockade=30%

Fentichlor
% hERG blockade=36%

Figure 4. Chemical structures of two weak hERG binders from the PubChem training set that were misclassified by the best SVM model constructed in this study.

CID: 411628
% hERG blockade=16%

CID: 411503
% hERG blockade=4%

CID: 250569
% hERG blockade=4%

CID: 246464
% hERG blockade=7%

Figure 5. Four examples of hERG nonblockers in the PubChem test set that are misclassified by the best SVM model of this study yet contain significant hERG binding features based upon the model.

945

dx.doi.org/10.1021/tx200099j |*Chem. Res. Toxicol.* 2011, 24, 934–949

an average of eight aromatic bonds per compound, and only seven of the misclassified compounds do not have any aromatic bonds. This study has found that the number of aromatic bonds in a compound is a molecular feature often associated with hERG binding affinity. Thus, it is possible that these 32 compounds are, in fact, weak hERG blockers. Four examples of false positive predictions for compounds containing a large number of aromatic bonds are shown in Figure 5. In addition to a significant

Naringenin
$IC_{50}=36.5 \mu M$

Figure 6. Example of the hERG blockers in the literature test set that are misclassified by the best SVM model of this study yet contain significant structural features corresponding to decreasing hERG binding. Among these features are hydrogen bond donor or acceptor groups at critical distances from nonpolar (np) atoms. The dotted lines indicate the separation distances between these pairs of atoms.

number of the bonds being aromatic in these four compounds, they also contain tertiary amino, benzyl, and hydrophobic groups. These functional groups have been implicated as important features for hERG channel blocking.[13]

For the literature test set of 356 compounds (287 hERG blockers and 69 non-hERG blockers), 30 false positive predictions are made by the best SVM model. To better understand why Naringenin, one of the false positive compounds, is incorrectly classified as an active hERG-blocker, its molecular features have been explored and indicated as either aiding or preventing hERG blockage. Naringenin contains three hydrogen bond acceptors, two hydrogen bond donors, and two aromatic rings resulting in a large set of atom pairs between np and hbd/hba IPE types. Naringenin is shown in Figure 6. The yellow dotted line indicates that one nonpolar atom and one hydrogen bond donor atom are 3−5 Å apart, while the white dotted line represents one nonpolar atom and one hydrogen bond acceptor atom that are packed closely together. Overall, Naringenin contains a significant number of structural features that specifically contribute to both $\varepsilon 1(np, hba)$ and $\varepsilon 4(np, hbd)$, which lead to the classification as an inactive hERG blocker using the best SVM model.

For the non-hERG blockers, 18 out of 69 compounds were misclassified as hERG binders (false positives) using the best SVM model. Among the 18 misclassified compounds, the average number of aromatic bonds per compound is approximately 11, and three of these compounds also have chlorine atoms, which increases the likelihood of a compound binding to the hERG channel; see Table 7. Most of these 18 outliers also contain nitrogen atoms that are part of piperazine, tertiary amino, benzyl, and/or hydrophobic functional group, which are molecular fragments that have been suggested as important features for active hERG compounds.[13] Moreover, these 18 compounds have an average IC_{50} value of 105 μM, suggesting that they may, in fact, be weak hERG blockers. Four of these misclassified compounds contain functional groups that are typically associated with hERG blockage and are shown in Figure 7.

Model Comparisons. To further substantiate the overall quality of the best hERG classification model of this study, 10 published in silico studies of hERG classification, employing different methodologies, have been compared to the best SVM
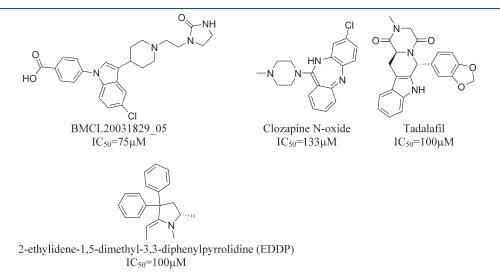

BMCL20031829_05
$IC_{50}=75\mu M$

Clozapine N-oxide
$IC_{50}=133\mu M$

Tadalafil
$IC_{50}=100\mu M$

2-ethylidene-1,5-dimethyl-3,3-diphenylpyrrolidine (EDDP)
$IC_{50}=100\mu M$

Figure 7. Four examples of hERG nonblockers of the literature test set that are misclassified by the best hERG binding SVM model of this study and contain significant structural features associated with hERG binding (see text).

Table 10. Listing of the Comparisons in Performance of Several hERG Models Including the Best SVM Model of This Study, Which Is Given in Row 1

modeling methodology	accuracy of training set predictions (number of compounds)	accuracy of testing set predictions (number of compounds)
SVM of this study	96% (876)	87% (356)
PLS (traditional and hologram QSAR)[23]	83—87 (55)	83% (95)
shape signatures[21]	69—73% (83)	85—95% (21)
fragment-based—evolutionary algorithm[20]	87—89% (70—100)	85—90% (22—24)
recursive partition[22]	96% (100)	93—96% (55)
binary QSAR model[42]	83—87% (150—223)	78—86% (58)
supervised neural network[10]	93% (244)	82% (72)
similarity-based method[25]	76% (275)	80% (500[a])
GFA binary QSAR model (40 μM cutoff)[18]	86% (356)	83% (876)
SVM with GRIND descriptors[24]	70—86%[a] (495)	73% (1877)
SVM with atom descriptors[14]	92%: ROC[b] (977)	94% (66)

[a] The reported methods contains linear and nonlinear models at different threshold values; 86% accuracy is for the linear SVM model at a 1 μM threshold, and 72% is an approximate overall accuracy of the nonlinear SVM model at a 30 μM threshold since the precise values are not stated in the reference. [b] ROC, receiver operating characteristic.

model of this study. The different hERG models considered in this comparison analysis are summarized in Table 10. The first column of Table 10 describes each of the modeling methodologies, the second column contains the predicted accuracy of training set with the number of training set compounds in parentheses, and the last column has the predicted accuracy of that test set with the number of test set compounds also given in parentheses. The optimal SVM model reported within this study has been placed in the first row of Table 10. The 10 comparison studies are listed in ascending order based on the number of compounds in each training set. The second through fifth studies of Table 10 have good training set accuracy, but these models are based upon relatively small training sets. The sizes of the training

sets for the four models listed in rows six through nine are all similar. Across these four studies, the overall accuracy of the training set is slightly less than the accuracy of the best SVM model from this study. The binary QSAR and supervised neural network models both lack sufficiently large test sets to permit a comprehensive validation of their respective predictive performances. The overall accuracies of the last four models, and their corresponding methodologies, in Table 10 are comparable to the best SVM model of this study (row 1) for both training and test sets. The molecular similarity-based hERG study is based upon a training set covering a quite limited chemical space. Additionally, it is not clear from the Nisius and Goller study[25] which fitting protocols were applied and what is the accuracy of the reported classifications (the accuracy of this strategy is limited to the chemical space of the applied training set). The SVM study using GRIND descriptors was applied to a large in-house training set, but the test set has a lower accuracy, by approximately 14%, than the best SVM model presented in this study. It would appear that the combination of 4D-FPs and MOE molecular descriptors may provide a more robust set of molecular information than do the GRIND descriptors.

The best SVM model of this study possesses better accuracy values for both the training and the test sets when compared to our previous GFA binary QSAR model. The SVM training set was the GFA's test set, and the GFA's training set was the SVM's test set. However, for the equivalent literature data set (the SVM test set), the SVM prediction accuracy is slightly higher than that of the GFA study. Thus, overall, the accuracy of the SVM model of this study is comparable to the GFA model.

The last model listed in Table 10 used atom descriptors[49] to construct a SVM model that was applied to an in-house training set of 977 compounds. The receiver operating characteristic, ROC, accuracy[50] of this SVM classification is 92%, which is comparable to the SVM model of this study. However, the external validation test set used in for the last model listed in Table 10 only contained 66 compounds, while the 977 training set compounds were divided into subsets with active to inactive ratios of 9:1, 1:1, and 1:9 for the training and test sets. Overall, the performance of the best SVM model of this study is slightly better than that of the SVM model constructed with atom descriptors (the last entry in Table 10).

■ CONCLUSIONS

The robust predictive performance of the best hERG blocking classification SVM model found in this study derives from our ability to accommodate the high-throughput imbalance of inactive versus active compounds found in current hERG data sets, and the universality of the molecular descriptors employed. However, the most novel aspect of this SVM model, as well as the GFA model that we previously reported for hERG classification, is the capacity to portray the key descriptors as 3D molecular features that contribute to both increasing and decreasing hERG potency. This ability to structurally characterize the key molecular features permits an investigator to readily view and access hERG features in a trial drug candidate.

Various molecular filter protocols were designed using Lipinski's rule-of-five and logP values to focus and balance the numbers of active and inactive hERG compounds comprising trial training sets. The most obvious way to manage a considerably large (more than 10 000 compounds) and imbalanced data set is to retain all of the minority compounds and filter the majority compounds. Unfortunately, this protocol did not work for the data sets investigated in this study. The best SVM model found in this work was based on

filtering the entire data set using Lipinski's rule-of-five and logP. To avoid the problem of overfitting when constructing the SVM model, the linear SVM modeling and deletion procedure was applied to reduce the size of the trial descriptor pool. The optimal reduced trial descriptor pool contains all of the MOE descriptors, but only a selected set of 4D-FP descriptors, which, in turn, indicates that a methodically chosen reduced trial descriptor pool provides a preferred approach to maximize the correct classification of compounds for hERG toxicity.

Seven key descriptors consisting of four MOE and three 4D-FP descriptors were found for the optimum GFA binary hERG QSAR classification model. The GFA binary hERG classification model is comparable in predictive performance to the best SVM model developed in the present study. The significance of these seven key descriptors has been further confirmed in this current study by the construction of a SVM binary hERG classification model, using only these seven descriptors, which is able to outperform other models created with significantly larger descriptor pools.

Application of the linear SVM methodology in this study revealed six additional key descriptors governing hERG blocking potency that can easily be interpreted, and visually represented, for training and test set compounds. Increasing the hydrophobicity of a compound is the most vital molecular property for increasing the hERG potency of a compound, and the logP and SlogP_VSA6 descriptors best capture this molecular property. The number of nonpolar atoms, the number of aromatic bonds and the number of chlorine atoms are also identified as positive features to increase hERG-blocking potency. Only two descriptors are found to be significant to decrease the hERG-blocking ability of compounds, namely, $\varepsilon^*(np, np)$ and $\varepsilon 1(np, hba)$.

What cannot be discerned from the SVM models developed in this study, as well as from virtually all previous hERG models based upon structurally diverse training sets, is whether or not the descriptors in the models are capturing one, or more, mechanisms of action (e.g., modes of binding to the hERG channel). This nondiscriminatory limitation, with respect to mechanism of action, is an inherent drawback to building QSAR models from noncongeneric data sets.

■ AUTHOR INFORMATION

Corresponding Author
*Tel: +886.2.3366.4888, #529. Fax: +886.2.23628167. E-mail: yjtseng@csie.ntu.edu.tw.

Author Contributions
#Equal contribution to this work.

■ REFERENCES

(1) Brown, A. M. (2004) Drugs, hERG and sudden death. *Cell. Physiol. Biochem.* 35, 543–547.

(2) Pearlstein, R. A., Vaz, R. J., kang, J., Chen, X. L., Preobrazhenskaya, M., Shchekotikhin, A. E., Korolev, A. M., Lysenkova, L. N., Miroshnikova, O. V., Hendrix, J., and Rampe, D. (2003) Characterization of hERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorg. Med. Chem. Lett.* 13, 1829–1835.

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

(4) Sanguinetti, M. C., Jiang, C., Curran, M. E., and Keating, M. T. (1995) A mechanistic link between an inherited and an acquired cardiac arrhythmia: hERG encodes the IKr potassium channel. *Cell 81*, 299–307.

(5) Aptula, A., and Cronin, M. (2004) Prediction of hERG K+ blocking potency: Application of structural knowledge. *SAR QSAR Environ. Res. 15*, 399–411.

(6) Cianchetta, G., Li, Y., Kang, J., Rampe, D., Fravolini, A., Cruciani, G., and Vaz, R. (2005) Predictive models for hERG potassium channel blockers. *Bioorg. Med. Chem. Lett. 15*, 3637–3642.

(7) Coi, A., Massarelli, I., Murgia, L., Saraceno, M., Calderone, V., and Bianucci, A. (2006) Prediction of hERG potassium channel affinity by the CODESSA approach. *Bioorg. Med. Chem. 14*, 3153–3159.

(8) Obrezanova, O., Csanyi, G., Gola, J. M. R., and Segall, M. D. (2007) Gaussian processes: A method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model. 47*, 1847–1857.

(9) Chen, X., Li, H., Yap, C., Ung, C., Jiang, L., Cao, Z., Li, Y., and Chen, Y. (2007) Computer prediction of cardiovascular and hematological agents by statistical learning methods. *Cardiovasc. Hematol. Agents Med. Chem. 5*, 11–19.

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

(11) Sun, H. (2006) An accurate and interpretable bayesian classification model for prediction of HERG liability. *ChemMedChem 1*, 315–322.

(12) Gepp, M., and Hutter, M. (2006) Determination of hERG channel blockers using a decision tree. *Bioorg. Med. Chem. 14*, 5325–5332.

(13) Song, M., and Clark, M. (2006) Development and evaluation of an in silico model for hERG binding. *J. Chem. Inf. Model. 46*, 392–400.

(14) Jia, L., and Sun, H. (2008) Support vector machines classification of hERG liabilities based on atom types. *Bioorg. Med. Chem. 16*, 6252–6260.

(15) Leong, M. (2007) A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. *Chem. Res. Toxicol. 20*, 217–216.

(16) Tobita, M., Nishikawa, T., and Nagashima, R. (2005) A discriminant model constructed by the support vector machine method for hERG potassium channel inhibitors. *Bioorg. Med. Chem. Lett. 15*, 2886–2890.

(17) Gavaghan, C. L., Arnby, C. H., Blomberg, N., Strandlund, G., and Boyer, S. (2007) Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J. Comput.-Aided. Mol. Des. 21*, 189–206.

(18) Su, B.-H., Shen, M.-y., Esposito, E. X., Hopfinger, A. J., and Tseng, Y. J. (2010) In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. *J. Chem. Inf. Model. 50*, 1304–1318.

(19) Rogers, D., and Hopfinger, A. J. (1994) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci. 34*, 854–866.

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

(21) Chekmarev, D. S., Kholodovych, V., Balakin, K. V., Ivanenkov, Y., Ekins, S., and Welsh, W. J. (2008) Shape signatures: New descriptors for predicting cardiotoxicity In silico. *Chem. Res. Toxicol. 21*, 1304–1314.

(22) Dubus, E., Ijjaali, I., Petitet, F., and Michel, A. (2006) In silico classification of hERG channel blockers: A knowledge-based strategy. *ChemMedChem 1*, 622–630.

(23) Keseru, G. M. (2003) Prediction of hERG potassium channel affinity by traditional and hologram qSAR methods. *Bioorg. Med. Chem. Lett. 13*, 2773–2775.

(24) Li, Q., Jørgensen, F. S., Oprea, T., Brunak, S., and Taboureau, O. (2008) HERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharmaceutics 5*, 117–127.

(25) Nisius, B., and Goller, A. H. (2008) Similarity-based classifier using topomers to provide a knowledge based for hERG channel inhibition. *J. Chem. Inf. Model. 49*, 247–256.

(26) PubChem. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?cid=376 (November 18, 2008).

(27) Zerhouni, E. (2003) The NIH Roadmap. *Science 302*, 63–72.

(28) Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009) PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res. 37*, W623–633.

(29) Japkowicz, N., and Stephen, S. (2002) The class imbalance problem: a systematic study. *Intell. Data Anal. 6*, 429–449.

(30) Chawla, N., Japkowicz, N., and Kolcz, A. (2004) Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. 6*, 1–6.

(31) Akbani, R., Kwek, S., and Japkowicz, N. (2004) Applying support vector machines to imbalanced datasets. *Machine Learn.: ECML* 39–50.

(32) Wu, G., and Chang, E. Y. (2005) KBA: Kernel boundary alignment considering imbalanced data distribution. *IEEE Trans. Knowledge Data Eng. 17*, 786–795.

(33) Raskutti, B., and Kowalczyk, A. (2004) Extreme re-balancing for SVMs: a case study. *SIGKDD Explor. 6*, 60–69.

(34) Guha, R., and Schurer, S. C. (2008) Utilizing high throughput screening data for predictive toxicology models: Protocols and application to MLSCN assays. *J. Comput.-Aided Mol. Des. 22*, 367–384.

(35) Li, Q., Wang, Y., and Bryant, S. H. (2009) A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics 25*, 3310–3316.

(36) Vapnik, V. N. (1998) *Statistical Learning Theory*, Wiley, New York.

(37) Vapnik, V. N. (2000) *The Nature of Statistical Learning Theory*, Springer, New York.

(38) Yoshida, K., and Niwa, T. (2006) Quantitative structure-activity relationship studies on inhibition of HERG potassium channels. *J. Chem. Inf. Model. 46*, 1371–1378.

(39) Thai, K., and Ecker, G. (2008) A binary QSAR model for classification of hERG potassium channel blockers. *Bioorg. Med. Chem. 16*, 4107–4119.

(40) Mitcheson, J. S. (2008) HERG potassium channels and the structural basis of drug-induced arrhythmias. *Chem. Res. Toxicol. 21*, 1005–1010.

(41) Allinger, N. L. (1977) Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V_1 and V_2 torsional terms. *J. Am. Chem. Soc. 99*, 8127–8134.

(42) Thai, K. M., and Ecker, G. F. (2008) A binary QSAR model for classification of hERG potassium channel blockers. *Bioorg. Med. Chem. 16*, 4107–4119.

(43) Fan, R.-E., Chen, P.-H., and Lin, C.-J. (2005) Working set selection using the second order information for training SVM. *J. Machine Learn. Res. 6*, 1889–1918.

(44) Senese, C. L., Duca, J., Pan, D., Hopfinger, A. J., and Tseng, Y. J. (2004) 4D-fingerprints, universal QSAR and QSPR descriptors. *J. Chem. Inf. Comput. Sci. 44*, 1526–1539.

(45) Duca, J. S., and Hopfinger, A. J. (2001) Estimation of molecular similarity based on 4D-QSAR analysis: Formalism and validation. *J. Chem. Inf. Comput. Sci. 41*, 1367–1387.

(46) MOE (2008) *Molecular Operating Environment*, Chemical Computing Group, Inc., Montreal, Canada.

(47) Chang, C.-C. Lin, C.-J. (2003) *LIBSVM: A Library for Support Vector Machines*.

(48) Yuchun Tang, Y.-Q. Z., Chawla, N. V., and Krasser, S. (2009) SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst., Man, Cybernetics—Part B: Cybernetics 39*, 281–288.

(49) Sun, H. (2004) A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *J. Chem. Inf. Comput. Sci. 44*, 748–757.

(50) Linden, A. (2006) Measuring diagnostic and predictive accuracy in disease management: An introduction to receiver operating characteristic (ROC) analysis. *J. Eval. Clin. Pract. 12*, 132–139.

(51) Wildman, S. A., and Crippen, G. M. (1999) Prediction of physiochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci. 39*, 868–873.

(52) Gasteiger, J., and Marsili, M. (1980) Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron 36*, 3219.