

In Silico Binary Classification QSAR Models Based on 4D-Fingerprints and MOE Descriptors for Prediction of hERG Blockage

Bo-Han Su,[†] Meng-yu Shen,[†] 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, Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No.1 Sec.4, Roosevelt Road, Taipei, Taiwan 106, The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, Illinois 60045, and College of Pharmacy MSC09 5360, 1 University of New Mexico, Albuquerque, New Mexico 87131-0001, USA

Received February 27, 2010

Blockage of the human ether-a-go-go related gene (hERG) potassium ion channel is a major factor related to cardiotoxicity. Hence, drugs binding to this channel have become an important biological end point in side effects screening. A set of 250 structurally diverse compounds screened for hERG activity from the literature was assembled using a set of reliability filters. This data set was used to construct a set of two-state hERG QSAR models. The descriptor pool used to construct the models consisted of 4D-fingerprints generated from the thermodynamic distribution of conformer states available to a molecule, 204 traditional 2D descriptors and 76 3D VolSurf-like descriptors computed using the Molecular Operating Environment (MOE) software. One model is a continuous partial least-squares (PLS) QSAR hERG binding model. Another related model is an optimized binary classification QSAR model that classifies compounds as active or inactive. This binary model achieves 91% accuracy over a large range of molecular diversity spanning the training set. Two external test sets were constructed. One test set is the condensed PubChem bioassay database containing 876 compounds, and the other test set consists of 106 additional compounds found in the literature. Both of the test sets were used to validate the binary QSAR model. The binary QSAR model permits a structural interpretation of possible sources for hERG activity. In particular, the presence of a polar negative group at a distance of 6–8 Å from a hydrogen bond donor in a compound is predicted to be a quite structure-specific pharmacophore that increases hERG blockage. Since a data set of high chemical diversity was used to construct the binary model, it is applicable for performing general virtual hERG screening.

INTRODUCTION

Inhibition of the human ether-a-go-go related gene (hERG) potassium ion channel is one of the major factors related to severe cardiotoxicity leading to long QT syndrome (LQTS), and it is a predisposing factor for syncope and sudden death.^{1,2} In recent years, many major drugs, such as terfenadine, cisapride, sertindole, thioridazine, and grepafloxacin, were withdrawn from the market as a result of their cardiotoxicity. Therefore, it is important to screen and assess the hERG channel protein binding of lead compounds early in the preclinical phase of drug discovery. Early types of preclinical hERG screenings includes several computational approaches to identify potential hERG channel blockers and, thereby, reduce the time and cost expended in performing preclinical lead identification and optimization.

The crystal structure of the hERG channel is not currently available. However, it is known that the overall protein structure of hERG is closely related to the bacterial K channels, including KcsA (PDB IDs 1BL8³ and 1K4C⁴), MthK (PDB ID 1LNQ⁵), and KvAP (PDB ID 1ORQ⁶), and an eukaryotic Kv1.2 channel (PDB ID 2A79).⁷ Several studies^{8–10} have used these crystal structures as templates for homology modeling of the hERG potassium channel to provide structure-based entities to dock and predict the binding affinity of compounds to the hERG potassium channel. The results of Farid and co-workers⁸ show a good correlation (coefficient of determination, $r^2 = 0.95$) for five sertindole analogues. Rajamani et al.¹⁰ selected 27 known hERG channel binders and their composite predicted affinities are in good agreement with experiment ($r^2 = 0.82$, RMSE = 0.56). Unfortunately, these studies are based upon nonvalidated receptor models interacting with members of relatively small analog data sets possessing limited ranges in chemical diversity. Moreover, it is not clear, in general, how binding alignments and poses can be reliably selected for arbitrary, structurally diverse compounds binding to these speculative types of homology based receptor models.

In view of the limitations to performing structure-based studies because of (a) the unavailability of a hERG crystal

* To whom correspondence should be addressed. Phone: +886.2.3366.4888#529. Fax: +886 0.2.23628167. E-mail: yjtseng@csie.ntu.edu.tw.

[†] Department of Computer Science and Information Engineering, National Taiwan University.

[‡] exeResearch, LLC.

^{||} The Chem21 Group, Inc.

[⊥] University of New Mexico.

[§] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University.

structure and (b) the need to establish molecular alignments and poses for structurally diverse compounds, most in silico molecular design strategies are ligand-based and take the form of traditional 2D-QSAR and 3D-QSAR studies. Within the current ligand-based QSAR paradigm, the selection of options available to build a modeling strategy is extensive but can also be problematic. There is a wide range in the types of QSAR descriptors that can be selected and which ultimately dominate the predictive power of the resulting QSAR model. The more common molecular descriptors currently in use include 4D-FPs,¹¹ MOE,¹² VolSurf,¹³ DRAGON,¹⁴ ab initio, ADAPT,¹⁵ and winMolconn¹⁶ descriptor sets. To explore the proper selection of descriptors from the large collection available and to ultimately optimize the resultant model, various machine learning methodologies, including support vector regression,¹⁷ genetic algorithms,¹⁸ neural networks,¹⁹ recursive partition method,²³ and PLS²⁰ have been applied to ligand-design QSAR analyses.

Song and Clark,¹⁷ for example, have calculated self-computed 2D fragment-based descriptors for a training set of 71 compounds and a test set of 19 compounds, resulting in hERG QSAR models with q^2 values of 0.636 for the training set and RMSE values of 0.440 and 0.597 for the training and test sets, respectively. Yoshida and Niwa¹⁸ utilized manually descriptor-selection for 104 training compounds and 18 test compounds. The hand-built hERG QSAR model has a q^2 value of 0.67, and RMSE values of 0.763 and 0.847 for the training and test set, respectively. Seierstad and Agrafiotis¹⁹ used seven different 2D descriptors (Kier–Hall topologic indices,²¹ Ghose–Crippen atom types,²² ISIS keys, atom pairs, electrotopologic state indices, medicinal chemistry descriptors, and MOE descriptors) for an in-house large training set (not publicly available data set) of 439 compounds screened for their affinities for the hERG potassium channel to achieved a q^2 of 0.76. This model's predictive nature, as measured by RMSE, is 0.88 for a test set of 40 compounds. Recently, Gavaghan et al.²⁰ employed Selma (an AstraZeneca in-house software package), DRONE,²³ VolSurf,¹³ and their in-house structural fragment-based descriptors to build QSAR models for a training set of 436 compounds from AstraZeneca's in-house data that were signified as interesting by the IonWorks high-throughput, HT, electrophysiology assay.²⁴ The resulting hERG QSAR model has a q^2 of 0.59 and a RMSE value of 0.44 for the training set and a RMSE value of 0.46 for an in-house test set of 876 compounds that are not available in the open literature. Overall, it can be seen for the hERG models in the literature, except for Gavaghan and co-workers' model, that the "quality" of the model (as measured by q^2 , RMSE, or both) gradually decreases as the number of training set molecules increases.

There is a large divergence and corresponding uncertainty in the measures of the hERG biological activity end points across the compounds studied. For example, the IC₅₀ value of terfenadine is reported as 0.0084 μ M in the Gavaghan study while another publication reports and uses a terfenadine IC₅₀ value of 0.33 μ M,²⁵ which is about a 40 times difference. This problem of inconsistent and diverse measurements of hERG binding end points is also presented and discussed in the literature.²⁶ One consequence of this lack of data consistency is that, on average, increasing the number of compounds in a training set increases the overall data noise

and reduces the predictive performance of the resultant model. In the case of test sets, uncertainty in the data measurements makes evaluation of prediction accuracy difficult and, basically, uncertain.

In attempts to increase hERG prediction capacity, modeling strategies have included the use of molecular descriptors that contain 3D information^{27–29,26,30} and constraining hERG binding affinity estimates to classification predictions. An effective means to the meaningful treatment of data sets where the end point measures have a high component of variability is to abandon *continuous* QSAR models and to construct *classification* QSAR models irrespective of the types of descriptors employed. Here the idea is to define classes of end point measures whose ranges are larger than the uncertainty in the end point measure. The predictive reliability of the resultant QSAR model will increase at the expense of the resolution of the prediction. The development of a two-state (binary) hERG classification model has been a principal goal of the work reported here.

Recently, Li et al.²⁶ published a hERG classification model (two-states) that employed the GRIND³¹ descriptors in conjunction with support vector machine, SVM. They achieved an overall classification accuracy of 74% for the training set (283 of 343 nonblockers correctly classified and 83 of 152 blockers correctly classified) and correctly classified 48 of 66 (72%) the test set compounds. About half of the compounds in the Li et al.²³ articles consisted of binary classification values. Therefore, these compounds and biological end points have not been used in constructing any of the data sets employed in this study.

In another study, "shape signatures"³⁰ molecular descriptors were employed to build hERG blocker classification models employing the SVM model fitting and optimization methodology. The resulting QSAR classification models have an overall accuracy of 69–73% for the 130 compound training set.

A variety of molecular features and properties have been proposed from hERG binding models as being important to ligand–hERG channel interactions. Song and Clark's fragment-based models suggest that lipophilic groups, such as benzyl and chloronaphthyl, and tertiary amines, like *N*-methyl piperidine, increase hERG channel binding. Conversely, some hydrophilic groups, such as carboxylic acid, ketones, hydroxide groups, and amine or primary aliphatic amines, and the presence of benzenesulfonamides or benzenesulfonyl decreases hERG binding affinity. More generally, it appears that the hydrophobicity of drugs, commonly characterized by logP, and hydrogen-bonding groups tend to increase the hERG blocking effect. The importance of the hydrophobicity and hydrogen bonding features for hERG binding is also reported in other studies reported in the literature.^{26,32}

The Yoshida and Niwa model indicates that hydrophilic feature like total polar surface area, TPSA, and polar van der Waals surface area tend to decrease the blocking effect. In addition, it has been observed that acidic ligand groups can abolish hERG activity,³³ and some hERG pharmacophore models² suggest that a positive-charge center, such as piperidine and piperazine, is essential for hERG affinity.

In this study, a trial descriptor pool containing a novel collection of descriptors formed by combining the 4D-fingerprints¹¹ and the Molecular Operating Environment's (MOE) traditional and VolSurf-like descriptors have been

used to build hERG QSAR models. The three goals sought by including this novel collection of descriptors are to develop a binary classification hERG QSAR model that (a) is accurate for a wide range of ligand structural diversity, (b) has high predictive reliability, and (c) can be readily interpreted in terms of a mechanism of action and corresponding key ligand molecular features.

MATERIAL AND METHODS

Data Sets. A training set of 250 hERG inhibitors with measured pIC_{50} values from *in vitro* assay experiments were assembled from the combined literature reports of Li et al.,²⁶ Song et al.,¹⁷ Yoshida et al.,¹⁸ and Thai et al.³²

The training set binding affinities have been primarily determined using mammalian or nonmammalian cell lines (including HEK, CHO, COS, and XO) and provided as IC_{50} values. The IC_{50} (μM) values were then transformed to $-\log IC_{50}$ (pIC_{50}) values for use in QSAR analysis, and these end points are provided in Supporting Information Table S1. The pIC_{50} values range from 2.00 to 9.00 with 28 compounds in the 2–4 range, 56 compounds in the 4–5 range, 81 compounds in the 5–6 range, 35 compounds in the 6–7 range, 33 compounds in the 7–8 range, and 17 compounds in the 8–9 range. The distribution of pIC_{50} values over the collected data set illustrates that the training set is a relatively unbiased with respect to the dependent variable (biological end point) and, therefore, should not lead to data-biased QSAR models.

In building the training set, 105 hERG binding drugs were available with known 2D structures, and the chemical structures of the remaining compounds were retrieved from NIH's NCBI PubChem compound database as SDfiles by querying their drug name. Each SDfile was transformed into the MOL file format and checked for correctness using SciFinder Scholar database version 2007.³⁴ The resulting 3D structures were imported into HyperChem 7.0 and geometry optimization was carried out using the MM+ force field.

The continuous and binary QSAR models were evaluated and validated using two external test sets. The first test set of hERG binders was constructed from a PubChem bioassay database (AID 376)³⁵ that initially contained 1953 compounds consisting of 250 active and 1703 inactive compounds. After removal of compounds complexed with metal ions, ambiguous compounds (two molecules within a single SDF entry), hERG activators (only compounds that inhibit hERG activity were considered for this study), and one compound included in our literature data set, the scrubbed PubChem test set contained 1668 compounds consisting of 163 actives and 1505 inactives. This test set was used to validate the QSAR models. Preparation of the test set was based upon the current strategy in the design of compound libraries that focuses upon 'compound scrubbing' of a parent data set. This protocol places constraints on the selection of compounds which leads to smaller, more condensed, more applicable libraries for lead optimization against a specific target.³⁶ In this case, the following two constraints were used: (a) retain compounds that pass the Lipinski's rule of five³⁷ and (b) retain compounds based on relative lipophilicity.

Applying the first constraint resulted in 353 compounds being removed from the 1668 PubChem hERG bioassay data set. The surviving 1315 compounds were passed to the next

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

IPE code	IPE abbreviation	IPE description
0	<i>any</i>	all atoms in the molecule
1	<i>np</i>	nonpolar atoms
2	<i>pp</i>	polar (+) atoms
3	<i>pn</i>	polar (−) atoms
4	<i>hba</i>	hydrogen bond acceptor atoms
5	<i>hbd</i>	hydrogen bond donor atoms
6	<i>aro</i>	aromatic atoms
7	<i>hs</i>	non-hydrogen atoms

filter. As previously discussed in the Introduction, it is well-known that the hydrophobicity (water disliking) of drugs (bioactive compounds) tends to increase the hERG blocking effect and the hydrophilicity (water liking) features of molecules tends to decrease the hERG blocking effect. This was the rationale for retaining compounds that are considered hydrophilic. The average logP values¹² of active and inactive compound (categorized using an value of 10 μM) in our literature data set were 4.1 and 2.8, respectively. The second constraint discarded active compounds whose logP values are *less than* 4.1 and inactive compounds with logP values *greater than* 2.8. The application of the second constraint led to a test set containing 876 compounds that are focused toward the physicochemical requirements of the hERG receptor. Using a focused test set provides a platform to more thoroughly explore and validate a QSAR modeling strategy.

A second test set considered in this study is the hERG data set of Nisius et al.³⁸ that contains 106 compounds. Since this was a previously published and validated data set, no constraints were applied to modify this data set.

Universal 4D-Fingerprints. The detailed formalism to compute the 4D-Fingerprints (4D-FPs) has been published in previous research articles,¹¹ and the methodology is only summarized herein. It is a method that generates a set of molecular fingerprints that are divided into pharmacophore elements designed to capture the 3D size, shape, and conformational flexibility of a molecule. These descriptors embed conformational averaged molecular information.

The first step in constructing the 4D-FPs is the generation of the conformation ensemble profile (CEP) of each molecule via molecular dynamic simulation (MDS). The universal 4D-fingerprints are the eigenvalues calculated from the absolute molecular similarity main distance-dependent matrix (MDDM) for each molecule. The MDDM captures how the shape and flexibility of a molecule are distributed for the different types of functional groups, also known as the interaction pharmacophore elements (IPEs). The atoms of a molecule are currently assigned one of the appropriate eight IPEs listed in Table 1. For each compound 36 MDDMs are constructed, from the same term or a cross-term, based on the IPE pair type from eight IPEs. An element of MDDM is defined as

$$E_{(d_{ij})} = e^{(-v\langle d_{ij} \rangle)} \quad (1)$$

where v is set to 0.25, which maximizes the difference in the sum of eigenvalues of any two compounds with the same number of a particular IPE type. The $\langle d_{ij} \rangle$ is the conformational average distance between atom i and j for a molecule over the entire CEP and defined as

$$\langle d_{ij} \rangle = \sum_k d_{ij}(k)p(k) \quad (2)$$

The term $d_{ij}(k)$ represents the distance between atom i and j of IPE u and v for a molecule in the k th conformer state; $p(k)$ denotes the thermodynamic probability of conformer state k and is evaluated from the ensemble of conformational energies.

Eigenvector and corresponding eigenvalues are derived from the diagonalization of the MDDM. When the IPE types are the same, that is, $u = v$, MDDM is an upper/lower triangular matrix and can be directly diagonalized. The resulting eigenvalues are then normalized and sorted in numerically descending order. The i th normalized eigenvalue for IPE type u of a molecule α is denoted by $\varepsilon_{uu}(\alpha)$ and evaluated using eq 3.

$$\varepsilon_{uu}(\alpha) = \frac{\varepsilon'_{uu}(\alpha)}{\text{rank}(\alpha)_u} \quad (3)$$

In eq 3, $\varepsilon'_{uu}(\alpha)$ stands for the non-normalized eigenvalue, and $\text{rank}(\alpha)_u$ is the rank of the MDDM matrix of molecule α for IPE type u .

When the IPE types are different, that is, $u \neq v$, MDDM will be rectangular because the number of u and v IPE elements can be different. Two square MDDMs can be calculated in eqs 4 and 5.

$$\text{MDDM}(u, u) = \text{MDDM}(n_u, n_v) \times \text{MDDM}(n_u, n_v)^T \quad (4)$$

$$\text{MDDM}(v, v) = \text{MDDM}(n_v, n_u) \times \text{MDDM}(n_v, n_u)^T \quad (5)$$

$\text{MDDM}(u, u)$ and $\text{MDDM}(v, v)$ possess the same rank and trace because of their equivalent set of eigenvalues. Therefore, the eigenvalues for the IPE pair $u \neq v$ are derived in eq 6.

$$\varepsilon_{uv}(\alpha) = \{[\varepsilon(\alpha)]_{\text{MDDM}(u, v)}\}^{1/2} \quad (6)$$

Up to 36 different $\varepsilon_{uv}(\alpha)$ are used as the 4D-fingerprints of a molecule α for the measurement of molecular similarity and dissimilarity. In eq 7, $D_{\alpha\beta}$ indicates the dissimilarity between molecule α and β for a specific IPE pair. Index i stands for the i th eigenvalue of one IPE pair and thus, the sum of all absolute difference of eigenvalues between molecules α and β are defined as the $D_{\alpha\beta}$.

$$D_{\alpha\beta} = \sum_i |\varepsilon(\alpha)_i - \varepsilon(\beta)_i| \quad (7)$$

The molecular similarity, $S_{\alpha\beta}$, is then given as

$$S_{\alpha\beta} = (1 - D_{\alpha\beta})(1 - \varphi) \quad (8)$$

where

$$\varphi = |\text{rank}(\alpha) - \text{rank}(\beta)| / (\text{rank}(\alpha) + \text{rank}(\beta)) \quad (9)$$

The rank of a MDDM matrix represents the number of atoms for a specific IPE pair, and hence, φ incorporates the atom size information into the similarity measurement. Molecules that are significantly different based on the number of atoms, the measure $1 - \varphi$, results in values approaching

0 for the range 0–1, and thus, $S_{\alpha\beta}$ is reduced. The range of $S_{\alpha\beta}$ and $D_{\alpha\beta}$ is between 0 and 1 because the normalized eigenvalues are adopted. $S_{\alpha\beta}$ values closer to 1 indicate higher degrees of similarity, whereas $D_{\alpha\beta}$ values approaching 1 refers to higher degrees of dissimilarity.

All of the eigenvalues for all of the MDDMs for all IPE pairs are utilized as the universal 4D-fingerprints for one molecule. To use all of the derived 4D-FPs as molecular descriptors in a partial least-squares (PLS) regression each molecule in the training set needs to have the same number of “effective” eigenvalues as PLS predictor variables. To accomplish this, the number of eigenvalues, $n(u, v)$ of IPE pair (u, v) for each training set molecule is calculated. The maximum number of $n(u, v)$ among all molecules is then determined as $n_{\max}(u, v)$, and assigned to the each molecule as its initial number of eigenvalues. For the case where the molecule contains the $n(u, v)$ less than $n_{\max}(u, v)$, the missing eigenvalues for $n(u, v)$ are set to zero. Hence, the total number of 4D-fingerprint descriptors of each molecule effectively becomes the sum of all $n_{\max}(u, v)$ for the 36 IPE pairs. In this study, the total number of eigenvalues is 813 for each training set compound. Each 4D-FP descriptor is normalized to a mean of zero and a unit variance (standard deviation) of one by dividing each column by the corresponding standard deviation.

For each test set molecule, the number of eigenvalues of specific IPE type (u, v) is set to $n_{\max}(u, v)$ of the training set. If the number of eigenvalues of a test set compound is greater than $n_{\max}(u, v)$, the smallest excess eigenvalues are disregarded. The test set descriptor matrix is also normalized using the same protocol used for the training set.

MOE Descriptors. In addition to the 4D-FP descriptors, a set of 280 2D and 3D descriptors from MOE¹² were included in the descriptor pool for this study. The 2D molecular descriptors are numerical features derived from the connection table representing a molecule and include physical properties, subdivided surface areas, atom counts, bond counts, Kier and Hall connectivity and Kappa Shape indices, adjacency and distance matrix descriptors containing BCUT and GCUT descriptors, pharmacophore feature descriptors, and partial charge descriptors. 3D molecular descriptors, which are dependent on the conformation of a molecule, include potential energy descriptors, surface area, volume, shape descriptors, and charge descriptors. Additionally, the VolSurf-like descriptors, which embed 3D molecular field interaction energies,³⁹ were added to the descriptor pool.

Model Selection and Statistical Analysis. Data fitting was accomplished using PLS regression analysis. This data fitting technique is useful when there are a large number of independent variables in the trial descriptor pool relative to the number of dependent variables, the pIC₅₀ end points. PLS is especially useful when there is no way of ranking the individual members (molecular descriptors) of the trial descriptor pool and/or knowing possible inter-relationships among the training set of descriptors.

The Genetic function approximation (GFA)⁴⁰ was used to optimize the QSAR models using the important descriptors as indicated by the PLS models. GFA is a multidimensional optimization procedure based on genetic algorithms⁴¹ (GA). GFA uses a GA to search the descriptor space of the possible QSAR models by evolving a population of models that best fit the training set data with respect to the lack-of-fit (LOF)

score that automatically penalizes models with too many features.⁴² Moreover, the size of a model, that is number of descriptor terms, can be controlled in a GFA optimization through the smoothing factor, a component of the LOF scoring function. As the smoothing factor is increased, the size of the corresponding QSAR model decreases.

The PLS regression was used to build a single QSAR model that contains the entire trial descriptor pool. The GFA constructed the optimal QSAR model using the fewest number of descriptors selected through the exploration of the overall molecular descriptor space. Both r^2 and the leave-one-out (LOO) cross-validated correlation coefficient, q^2 , were employed to characterize the quality of the resultant QSAR models. In this study, a cutoff pIC_{50} value, c , was selected to separate active from inactive compounds and, correspondingly, transform the continuous QSAR modeling predictions into binary classifications. For the binary classification scheme, a compound with a pIC_{50} predicted value higher than c is classified as active, whereas a predicted value lower than c is classified as inactive. This approach was taken to optimize model reliability, and to be consistent with the operational implementation of a hERG model: to discern whether or not a compound is toxic (a hERG blocker) and not to compute specific pIC_{50} hERG affinity values for the compound.

The quality of binary classification models have also evaluated in terms of *accuracy* (eq 10), *sensitivity* (eq 11), and *specificity* (eq 12).

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

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

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

In eqs 10–12, tp (true positives) and fn (false negatives) are the number of known active (toxic) compounds predicted to be active and inactive, respectively. Correspondingly, tn (true negatives) and fp (false positives) are the number of known inactive (nontoxic) compounds predicted to be inactive and active, respectively.

RESULTS

Molecular Similarity Characterization of the Training and Test Sets. The range and average molecular similarities, using 4D-fingerprints, were computed for members of both the training set and the PubChem test set. The results in Table 2 give the minimum, maximum, and average molecular similarity values of the IPE types for the 250 training set molecules and also for the 1720 PubChem test compounds. The IPE similarity values range from 0.00 (effectively no similarity) to a maximum value near 0.99 (identical molecular structures), and the average similarity values for the IPE types range from 0.28 to 0.75 for both data sets. These 4D-FP similarity measures illustrate the high diversity within each of these two data sets. Additionally, these similarity studies confirm that, in composite, the training and test sets are representative of one another. Both data sets have similar average values for all IPE types with

Table 2. Range (Minimum and Maximum) and Average Molecular Similarity for Each of the IPE Types Based on 4D-Molecular Similarity, of the Training Set, and of the PubChem Test Set

IPE type		minimum	maximum	average
<i>any-any</i>	training	0.15	0.99	0.75
	test	0.08	0.97	0.70
<i>np-np</i>	training	0.09	0.99	0.73
	test	0.00	0.97	0.56
<i>pp-pp</i>	training	0.00	0.99	0.33
	test	0.00	0.99	0.28
<i>pn-pn</i>	training	0.00	0.99	0.49
	test	0.00	0.99	0.48
<i>hba-hba</i>	training	0.00	0.99	0.46
	test	0.00	0.99	0.44
<i>hbd-hbd</i>	training	0.00	0.99	0.46
	test	0.00	0.99	0.34
<i>hs-hs</i>	training	0.10	0.99	0.73
	test	0.09	0.99	0.70

the exception of the (nonpolar–nonpolar) IPE type. Thus, the PubChem test set can be considered a validated test set for evaluating the performance of QSAR models derived from the training set.

Selection of Descriptors and Continuous QSAR Models.

The continuous hERG QSAR models were constructed using the entire training set (250 compounds) with the corresponding pIC_{50} measures. The continuous hERG QSAR models were optimized by GFA optimization, and the most relevant descriptor terms identified. Three different portions of the complete descriptor pool were considered in building the continuous hERG QSAR models: (a) 4D-FPs with logP, (b) the set of MOE descriptors (traditional and VolSurf-like), and (c) the 4D-FPs combined with MOE descriptors. Each of these three trial descriptors sets were restricted to the 200 most highly weighted descriptors of each set that were embedded in the principal components of corresponding PLS QSAR models.

The best continuous QSAR model constructed when the descriptor pool consisted of 4D-FPs and logP is given by eq 13

$$\begin{aligned} \text{pIC}_{50} = & 4.61 + 22.06\varepsilon_{112}(\text{all}, \text{all}) - 90.83\varepsilon_{25}(\text{hs}, \text{hs}) - \\ & 20.7\varepsilon_{2}(\text{np}, \text{np}) - 215.87\varepsilon_{35}(\text{hs}, \text{np}) + 0.31 \log P - \\ & 16.63\varepsilon_{1}(\text{hs}, \text{hs}) + 13.77\varepsilon_{9}(\text{np}, \text{hbd}) - \\ & 10.84\varepsilon_{6}(\text{hbd}, \text{hbd}) + 2.21\varepsilon_{2}(\text{pn}, \text{hbd}) \\ & N = 250, r^2 = 0.52, q^2 = 0.48 \quad (13) \end{aligned}$$

In eq 13, by way of example, $\varepsilon_{112}(\text{all}, \text{all})$ represents the 112th largest eigenvalue from the MDDM of $u = (\text{all})$ and $v = (\text{all})$, and $\varepsilon_{9}(\text{np}, \text{hbd})$ denotes the ninth largest eigenvalue from the MDDM of $u = (\text{np})$ and $v = (\text{hbd})$. The GFA optimization suggests that the maximum explanation of the variance, r^2 , across the training set is captured by eight universal 4D-FPs, but the r^2 is only 0.52 indicating that only 52% of the total variance of the training set can be captured. Moreover, q^2 for eq 13 is nearly the same as r^2 indicating that no specific compounds of the training set particularly bias the QSAR model.

The GFA optimized continuous QSAR model built from only the MOE descriptors (traditional descriptors and VolSurf-like descriptors) is given as follows;

$$\begin{aligned} \text{pIC}_{50} = & -2.01 - 0.005E_{\text{oop}} + 0.02\text{PEOE_VAS} + \\ & 3 - 0.001\text{DCASA} + 0.33\text{dipole} - 0.04\text{PEOE_VSA} - \\ & 5 - 5.37\text{vsurf_CW5} + 0.34\text{radius} - \\ & 2.07\text{BCUT_SMR}_0 + 0.74\text{std_dim2} \\ N = 250, r^2 = 0.60, q^2 = 0.56 \quad (14) \end{aligned}$$

In eq 14, nine MOE descriptors, including eight traditional descriptors and one VolSurf descriptor, were selected in building the GFA optimized model, and r^2 and q^2 are near 0.6. Equations 13 and 14 both have nine descriptors and comparable predictive capabilities. For eq 14 the r^2 is 0.60 indicating that only 60% of the total variance of the training set has been captured. And like eq 13, the finding that q^2 is nearly the same as r^2 indicates that no specific compounds particularly bias the QSAR model.

Finally, the 4D-FPs and MOE descriptors were combined into a single descriptor pool to construct an optimized continuous QSAR model, which is given by eq 15.

$$\begin{aligned} \text{pIC}_{50} = & 7.13 - 10.23\epsilon6(\text{all}, \text{hbd}) + 10.62\epsilon7(\text{pn}, \text{hbd}) - \\ & 15.43\epsilon1(\text{hs}, \text{hs}) + 1.9\text{vsurf_CW2} - 0.01\text{vsurf_W4} + \\ & 0.02\text{PEOE_VSA} - 4 - 0.02\text{PEOE_VSA_POL} - \\ & 0.36\text{GCUT_SLOGP}_0 + 0.62\text{std_dim2} \\ N = 250, r^2 = 0.60, q^2 = 0.56 \quad (15) \end{aligned}$$

There are three 4D-FPs, four traditional MOE descriptors, and two VolSurf-like MOE descriptors in the GFA optimized model derived from the complete descriptor pool. However, the r^2 and q^2 of eq 15 are about the same as eq 14 suggesting the continuous models (eqs 13–15) are all about equivalent in quality.

Binary QSAR Predictions Derived From the Continuous QSAR Models. The continuous QSAR models given by eqs 13–15 can be transformed into binary models to separate active from inactive compounds once a separation cutoff pIC_{50} value is established. To separate active hERG blockers from inactive compounds, different threshold values of IC_{50} have been proposed in the literature. Aronov and Goldman adopt a cutoff IC_{50} value of 40 μM ,⁴³ O'Brien and de Groot used a cutoff of 20 μM , Sun selected a cutoff values of 30 μM , respectively,^{33,44} and Fioravanzo et al.⁴⁵ selected 10 μM as a cutoff. Tobita et al.⁴⁶ and Roche et al.⁴⁷ divided their respective data sets into three classes by implementing two cutoff values; Tobita et al. used IC_{50} cutoff values of 1 and 40 μM , and Roche et al. opted for cutoff values of 1 and 10 μM . Recently Li et al.²⁶ published a binary hERG classification model, and tested the performance and accuracy of the model at various IC_{50} cutoffs of 1, 5, 10, 20, 30, and 40 μM . The model employing a 40 μM threshold showed the best performance via internal validation. By selecting a high cutoff IC_{50} value, the risk of missing potentially toxic hERG compounds is reduced, but the *accuracy* and *sensitivity* (see eqs 10 and 11) of the model may suffer. In this study, an IC_{50} cutoff value of 40 μM ($\text{pIC}_{50} = 4.39$) was chosen as the cutoff to provide an acceptable degree of safety and is also effectively the optimum cutoff for the performance of equations (constructed QSAR models) 13–15.

The predicted *accuracy*, *sensitivity*, and *specificity* values are 89%, 97%, and 48%, respectively, when using the pIC_{50} cutoff value of 4.39 for eq 13. The binary classification quality improves for eq 14's *specificity*, relative to eq 13, when using the pIC_{50} cutoff value of 4.39. The predicted

Table 3. Performance of the Optimized Analog Binary Qsar Models to Eq 15 Resulting from Using Different Gfa Smoothing Factors

smoothing factor	accuracy (%)	sensitivity (%)	specificity (%)	no. of terms
0.15	90	96	60	9
0.17	90	96	60	9
0.2	90	97	55	8
0.4	90	96	58	10
0.6	88	96	45	9
0.8	91	96	63	9
1	89	95	55	9
1.2	91	97	60	8

accuracy, *sensitivity*, and *specificity* value are 89%, 95%, and 58%, respectively, for model 14. When binary classification modeling, with the cutoff value of 4.39, is applied to eq 15, the predicted *accuracy*, *sensitivity*, and *specificity* value are 90%, 96%, and 60%, respectively and the predictive power is slightly greater than the MOE descriptor-based model. While the two models from eqs 14 and 15 are comparable, eq 15's is based upon 4D-FPs and MOE descriptors. Hence, eq 15 was adopted for further consideration because of the 1D, 2D, 3D, and 4D information that is embedded into this model, and its prediction power, especially *specificity*, is slightly better than that of eq 14. Additionally, the inclusion of the 4D-FP descriptors, as discussed below, permits the extraction of several 3D-pharmacophore features that are characteristic of hERG binding.

Determination of the Best QSAR Models. The model derived from the descriptor pool of the 4D-fingerprints and MOE descriptors, was further explored, refined and optimized with respect to the different parameter and settings employed during the GFA and PLS model building and optimization process. Equation 15 was built using a GFA smoothing factor of 0.17. Different smoothing factor values were applied in this part of the study to explore model stability and to avoid overfitting the QSAR model. The results of varying the GFA smoothing factor are presented in the Table 3 with the first column, SF, containing the selected smoothing factor, the second through fourth columns reporting the accuracy, sensitivity, and specificity percentages, respectively found for the resulting analog binary QSAR models to eq 15. These results are based on using a pIC_{50} cutoff value of 4.39 for each of the optimized binary QSAR models reported in Table 3. The right-most column of Table 3 indicates the number of descriptor terms in the corresponding optimized model. The findings reported in Table 3 indicate that the overall quality of this family of QSAR models is not particularly sensitive to the choice in smoothing factor value. In turn, this finding suggests that these models are quite stable with respect to the PLS fitting metric. The eighth model in Table 3 was selected as the overall best model because of its combined high accuracy, sensitivity, and specificity, and for the variety of structural information contained in the eight descriptors of the model. The continuous form of this optimal binary QSAR model is given by

$$\begin{aligned} \text{pIC}_{50} = & 9.52 - 14.61\epsilon6(\text{all}, \text{hbd}) - 17.52\epsilon1(\text{hs}, \text{hs}) + \\ & 16.43\epsilon7(\text{pn}, \text{hbd}) - 4.36\text{vsurf_HL1} + 2.06\text{vsurf_CW2} - \\ & 0.11a_{\text{acc}} - 0.02\text{PEOE_VSA_POL} - \\ & 0.004E_{\text{SOL}} \quad N = 250, r^2 = 0.58, q^2 = 0.54 \quad (16) \end{aligned}$$

Table 4. Statistics Fit and Performance of Ten Classification Models Derived from the 80% Randomly Selected Training Set Compounds of the Literature Data Set^a

part a								
model	training set		test set			complete literature data set		
	r^2	q^2	Acc (%)	Sen (%)	Spe (%)	Acc (%)	Sen (%)	Spe (%)
1	0.61	0.57	86	98	25	87	95	43
2	0.65	0.59	88	100	40	89	96	52
3	0.62	0.58	86	95	44	88	95	50
4	0.62	0.58	88	100	25	87	97	38
5	0.62	0.57	82	95	13	88	96	48
6	0.64	0.59	82	90	38	88	94	58
7	0.66	0.62	76	81	50	89	94	63
8	0.67	0.63	88	95	50	85	93	45
9	0.64	0.58	86	93	56	89	95	58
10	0.63	0.59	82	97	27	90	97	55
average	0.64	0.59	84	94	37	88	95	51
best model (eq 16)	0.58	0.54				91	97	60

part b						
model	PubChem all testing			filtered PubChem testing		
	Acc (%)	Sen (%)	Spe (%)	Acc (%)	Sen (%)	Spe (%)
1	67	64	68	87	83	87
2	59	70	58	75	90	75
3	68	64	68	84	86	84
4	69	59	70	81	79	81
5	65	60	65	80	90	79
6	71	66	71	89	90	89
7	71	61	72	86	79	86
8	66	60	67	85	83	85
9	65	66	65	81	86	80
10	68	59	70	89	83	90
average	67	63	67	83	85	84
best model (eq 16)	65	74	64	83	97	82

^a Part a lists the r^2 and q^2 of the 80% training set and the accuracy (Acc), sensitivity (Sen), and specificity (Spe) for the 20% test set, as well as the complete literature data set, respectively. Part b lists the accuracy (Acc), sensitivity (Sen) and specificity (Spe) for the all PubChem and filtered PubChem data set, respectively.

The r^2 and q^2 of eq 16 are slightly lower than those of eq 15. However, the optimized binary prediction accuracy of 91%, sensitivity of 97%, and specificity of 60% for eq 16 are slightly better than the accuracy of 90%, sensitivity of 96%, and specificity of 60% for eq 15 when it is used as a binary model.

The internal training set predictions, based on a cutoff IC_{50} value of 40 μM , for all 250 training molecules using eq 16 are presented as part of Table S1 in Supporting Information. The first column of Table S1 is the drug name as reported in each reference, and the second and third columns list the observed and predicted pIC_{50} values (using eq 16). The corresponding binary values are listed in columns 4 (observed) and 5 (predicted). A binary value of 1 in Table S1 indicates the compound is active, and a value of 0 indicates that the compound is inactive with respect to hERG blockage.

The 40 μM IC_{50} cutoff used in this study was, as discussed above, chosen based upon a previous investigation²⁶ where it was demonstrated that an IC_{50} of 40 μM was the optimum value to classify the hERG toxicity as active or inactive. However, an alternate cutoff value of 10 μM has been explored in this study, and the observed and predicted binary results are listed in columns 6 and 7 of Table S1. This choice in the cutoff value also leads to high accuracy, sensitivity, and specificity values of 85%, 97%, and 85%, respectively.

The good results found in using two significantly different cutoff values illustrates that the underlying QSAR model is very stable and reliable. The cutoff value of 40 μM ($pIC_{50} = 4.39$) was employed in making validation predictions for the large external data set even though the 10 μM ($pIC_{50} = 5.00$) cutoff appears to perform better for eq 16. This was done because the cutoff of 4.39 is consistent with, and comparable to, other reported studies that employed the 4.39 ($IC_{50} = 40 \mu M$) cutoff for hERG blockage activity.

Although leave-one-out cross-validation was utilized to evaluate the best PLS models, there is residual debate on whether the performance of our model is stable with respect to the training set. The literature data set of 250 compounds was randomly divided into a training set (80% of the compounds; 200 compounds, 168 active, 32 inactive) and a test set (the remaining 20% of the compounds; 50 compounds, 42 active, 8 inactive). The PLS model was derived from the training set, and the best classification model was chosen based on the performance of the model in predicting the end points of the test set. While the compounds were randomly selected, the selection process was constrained to maintain the same ratio of active to inactive compounds as found in the complete literature set. Ten training and test set groups were constructed using this protocol. Table 4a lists the r^2 and q^2 of the PLS models constructed using the

Table 5. Relative Importance (Weight) of the Descriptor Terms (Molecular Descriptors) of Eq 16 as Measured by the Product of the Regression Coefficient and the Range (Maximum Minus Minimum Value) for Each Descriptor Value in the Training Set^a

descriptor name	relative importance of descriptors over the training set for eq 16
$\epsilon 1(hs,hs)$	-4.77
PEOE_VSA_POL	-3.18
<i>vsurf_CW2</i>	2.82
<i>E_SOL</i>	-2.28
<i>a_acc</i>	-1.58
<i>vsurf_HL1</i>	-1.48
$\epsilon 6(all,hbd)^b$	-1.31
$\epsilon 7(pn,hbd)^b$	1.16

^a The sign of the relative weight reflects the sign of the regression coefficient. ^b An inter-related pair of descriptors, see text.

ten randomly selected training sets. Also included in Table 4a are the accuracy (Acc), sensitivity (Sen) and specificity (Spe) for the test sets and the entire data set. The average values for the statistics of performance from these ten models, the 80–20 models, are reported in the second to last row. The results for the suggested (best) model (eq 16) are shown in the last row of Table 4a. All of the models based on randomly selected training sets are observed to result in slightly higher r^2 and q^2 than the best model because of the smaller number of compounds in these training sets. Overall, the training and test set models retain a high accuracy and sensitivity for the entire literature data set and the test set. However, the specificity values of several models are relatively low when compared to the average specificity value. These lower specificity values are likely the result of low molecular similarity between the inactive compounds of the training and test sets. The average accuracy, sensitivity, and specificity values from the predictions for the complete literature set, using the ten models, are only slightly lower than those reported using eq 16.

Statistical and Mechanistic Interpretation of the 4D-FPs and MOE Descriptors of the Best Model. Three 4D-FPs, two VolSurf-like MOE descriptors, and three traditional MOE descriptors terms form the optimal binary QSAR model, eq 16. In terms of a statistical interpretation of eq 16, the constant value (y-intercept) of 9.52 indicates that descriptor terms, in composite, contribute to decreasing the predicted pIC_{50} values since the maximum pIC_{50} value in the training set is only 9.00. This can also be inferred from the fact that the regression coefficients are negative for all but two of the descriptor terms in eq 16, and the large majority of the molecular descriptor values are positive.

One metric to measure a descriptor's relative importance in a QSAR model is to compute the product of the magnitude of a regression coefficient times the range in values adopted by its descriptor across the training set. This metric is a measure of the variability of dependent variable of the QSAR upon the descriptor. Table 5 lists the relative importance of each of the descriptors in eq 16, based on this metric. The descriptors are listed in descending order (descriptors with the largest magnitude first) of their relative importance in Table 5. The data in Table 5 shows that the $\epsilon 1(hs,hs)$ molecular descriptor (4D-FP) is the most important factor and $\epsilon 7(pn,hbd)$ is the least important term of the optimal QSAR model. A full discussion on the physical features these

Table 6. Classification of the Descriptors in Eq 16 by Their General Physicochemical Property Type

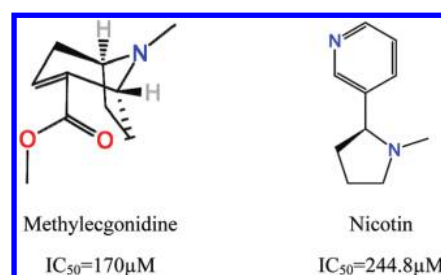
class name	descriptor names
six spatial shape and size descriptors	$\epsilon 6(all,hbd)$ $\epsilon 1(hs,hs)$ $\epsilon 7(pn,hbd)$ <i>vsurf_HL1</i> <i>vsurf_CW2</i>
two thermodynamic descriptors	<i>E_SOL</i> PEOE_VAS_POL
one electronic descriptor	<i>a_acc</i>

descriptors encode is presented in the Spatial Shape and Size Descriptors section below.

The descriptors comprising eq 16, regardless of their relative importance, can be divided into spatial shape and size, thermodynamic, and electronic descriptors. Table 6 lists the descriptors of eq 16 using this classification.

Spatial Shape and Size Descriptors. All of the 4D-FPs are classified as spatial shape and size descriptors. $\epsilon 1(hs,hs)$ is the most dominate molecular descriptor term in eq 16, and this suggests that the greater the number of non-hydrogen atoms that are tightly packed near to one another in a compound, the compound's hERG blocking activity is reduced. That is, the smaller the value of the eigenvalue index, number, x , in ϵx , the closer the distance between the IPE pair as can be inferred from eqs 1 and 2. An eigenvalue index of one means the nonbonded, non-hydrogen atoms are as close as possible to one another. Unfortunately, using this relatively nonspecific structural feature is difficult. This descriptor can be interpreted as indicating that having a relatively large number of heavy (polar) atoms, such as nitrogen, oxygen, sulfur, and chlorine, clumped together, within a molecule, is a good way to avoid hERG activity. Methylecgonidine and nicotine, shown in Figure 1, are two examples of compounds that possess a high percentage of polar atoms closely packed that are also low in hERG potency.

Perhaps the most interesting molecular feature from eq 16 is the pair of 4D-FP molecular descriptors ($-14.61\epsilon 6(all,hbd) + 16.43\epsilon 7(pn,hbd)$); *all* refers to all IPE types, *hbd* represents the hydrogen bond donor IPE, and *pn* indicates the polar-negative IPE. This pair of descriptors actually forms a single descriptor since $\epsilon 6$ and $\epsilon 7$ are equivalent eigenvalue terms (eg., they have nearly the same eigenvalue index number, x , in ϵx) representing a common separation distance between the two stated IPE types. The larger the value of the eigenvalue index, the greater is the separation distance between the IPE pair as can be inferred from eqs 1 and 2. In

**Figure 1.** Methylecgonidine and nicotin are two examples of compounds that possess a high percentage of polar atoms closely packed that should decrease hERG binding based upon the $\epsilon A1(hs,hs)$ term of eq 16.

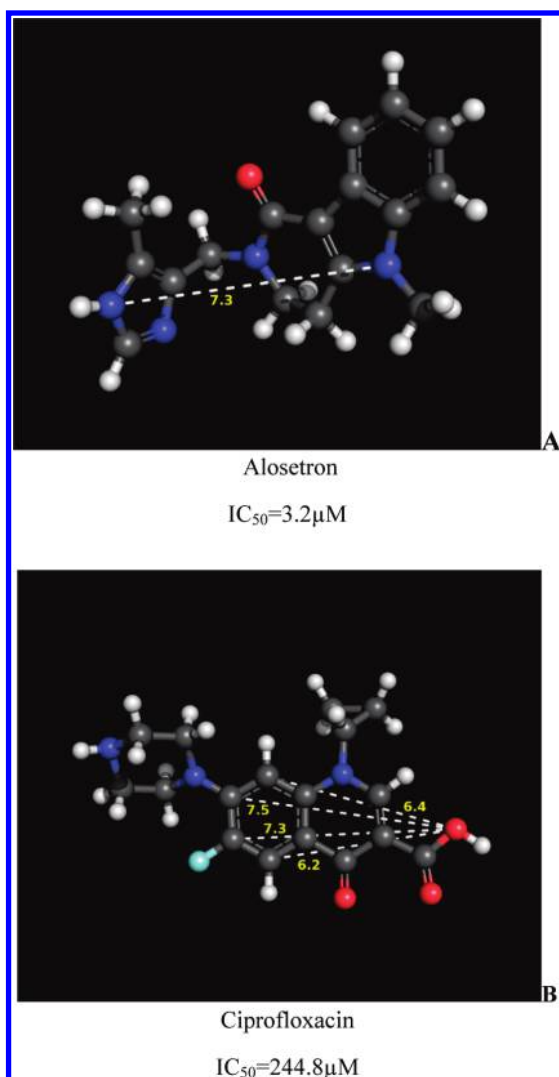


Figure 2. Two examples of compounds which possess a hydrogen bond donor (*hbd*) 6–8 Å from (A) a negative-polar (*np*) atom and (B) non-negative-polar atoms, respectively. The dotted lines and numbers indicate the separation distances between these pairs of atoms in Ångstroms.

this case, the meaning of this pair of descriptor terms is that all IPE pairs, other than *pn* and *hbd*, which are separated by 6–8 Å will decrease pIC_{50} . But (*pn*, *hbd*) IPE pairs 6–8 Å apart will increase pIC_{50} . Thus, eq 16 shows that compounds having polar-negative groups located 6–8 Å from hydrogen bond donors may increase the propensity for hERG binding. These same two descriptors are found in identical fashion but with different regression coefficients, in eq 15. An example of a compound from the training set that contains the (*pn*, *hbd*) IPE pair is alosetron, which is shown in its global minimum conformation in Figure 2A. Alosetron possesses a high hERG activity ($IC_{50} = 3.2 \mu M$) and has one hydrogen bond donor (*hbd*) and one polar-negative atom (*pn*) separated by a distance of 7.8 Å, depicted by the yellow dotted line in Figure 2A. An example of a compound with low hERG activity and containing one hydrogen bond donor atom but with polar-positive and aromatic atoms that are 6–8 Å apart is ciprofloxacin ($IC_{50} = 244.8 \mu M$), which is shown in Figure 2B.

$vsurf_CW2$ and $vsurf_HL1$ are molecular surface and volume descriptors based on the 3D molecular interaction fields constructed through the calculated interaction energy

between a compound in its low-energy conformation and a molecular probe. Thus, $vsurf_CW2$ and $vsurf_HL1$ can be classified as shape and size descriptors. Specifically, $vsurf_CW2$ is the fraction of a molecule's hydrophilic surface area compared to the molecule's entire molecule surface area. For eq 16, as the fractional amount of hydrophilic surface area increase, the predicted pIC_{50} values are also expected to increase. This relationship between hydrophilic surface area and pIC_{50} is at odds with the hypothesis that increasing the hydrophilic groups in a compound decreases a compound's hERG binding affinity. However, it needs to be remembered that there are other descriptor terms in eq 16, such as E_{sol} , where increasing hydrophilic character is predicted to decrease hERG binding. In other words, it can be misleading to decouple the descriptors in eq 16 or any QSAR model from one another in trying to postulate a mechanism of action from the QSAR.

$vsurf_HL1$ is the hydrophilic–lipophilic balance, a ratio of the hydrophilic regions and hydrophobic regions of a molecule. $vsurf_HL1$ measures which of these two features are dominate in a molecule or if they are approximately the same. Increasing the hydrophilic region, based upon eq 16, decreases hERG binding, which is in agreement with many literature reports.^{17,18,26,32} The cross-correlation coefficient between these two descriptors for the training set is 0.47, a moderate, but not a high correlation. While these two descriptors measure *related* physical properties of a molecule, they are not very similar features in terms of their control over hERG activity.

Thermodynamic Descriptors. E_{sol} , the aqueous solvation energy, suggests that as the aqueous solubility of the compound increases, there is a decrease in predicted pIC_{50} . The other thermodynamic descriptor, $PEOE_VSA_POL$, is the total van der Waals surface area of a compound's polar atoms. Atoms are considered polar if the absolute value of the assigned Gasteiger atomic partial charge⁴⁸ is greater than 0.2, and the surface area is calculated using an approximate connection table. As the van der Waals polar surface area of a compound increases, pIC_{50} is predicted by eq 16 to strongly decrease. The $vsurf_HL1$ and $PEOE_VSA_POL$ molecular descriptors in eq 16 are consistent with the general finding that higher molecular lipophilicity promotes hERG activity.

Electronic Descriptors. a_{acc} is the only electronic descriptor in eq 16, and it is the number of hydrogen bond acceptor atoms of a compound. On the basis of eq 16, pIC_{50} is predicted to decrease as the number of hydrogen bond acceptors increases.

Model Validation Using the PubChem Bioassay Test Set and Additional Literature Set. To evaluate and validate the binary QSAR hERG blockage model given by eq 16, a large test set of hERG bioassay data was assembled from PubChem as described above. Currently, the hERG assays given in the PubChem repository only provide the percentage of hERG blockade and *not* an actual IC_{50} value. To construct the test set, compounds from PubChem were classified as active, or inactive, using a cutoff value of 20% hERG blockage. That is, compounds with a hERG blockage percentage of 20% or greater were considered active. The predicted value from the continuous QSAR model (eq 16) is a pIC_{50} value that cannot be directly converted into the percentage of hERG blockade. Thus, it was necessary to

develop a bridging relationship between the hERG blockage percentage cutoff from PubChem and the IC_{50} cutoff used in the binary form of the QSAR model (eq 16). In the PubChem hERG assay protocol, a $10\ \mu\text{M}$ dosing of terfenadine produced a 100% hERG blockade and was adopted to represent the positive control in test set validation screening. Correspondingly, a cutoff value less than the binary model cutoff IC_{50} value of $40\ \mu\text{M}$, successfully used for the training set in Table S1, must be identified to match the cutoff value of 20% in the PubChem classification protocol. To find an appropriate cutoff value to use with eq 16 for validation screening of the PubChem test set, the following was done:

- (a) First, eq 16, along with the “standard” cutoff value of $40\ \mu\text{M}$ ($pIC_{50} = 4.39$) was used to predict the active or inactive hERG blockage of all 1668 PubChem compounds. For the active molecules, 141 of 163 compounds were correctly predicted, and 724 compounds of 1505 inactive (nonblockers) were correctly estimated. The overall accuracy was 52% and, obviously, the model performs poorly.
- (b) Next, the binary active/inactive cutoff value was lowered to $10\ \mu\text{M}$ ($pIC_{50} = 5.0$) and, as expected, better hERG blockage predictions were obtained with an overall accuracy of 65% (74% for sensitivity and 64% for specificity). This lower cutoff value ($IC_{50} = 10\ \mu\text{M}$) is in much better agreement with the categorization of PubChem data than the $40\ \mu\text{M}$ cutoff.
- (c) Then the Lipinski’s rule-of-five and logP constraints, presented in the Materials and Methods section, were applied to the PubChem data set to take advantage of information learned from previous hERG studies and to recognize that the high-throughput screening data from PubChem might contain non drug-like and/or non-hERG-related compounds.
 - (i) When only the Lipinski’s rule-of-five constraint was applied, 1315 compounds were retained in the test set, which partition into 105 blockers (actives) and 1210 nonblockers (inactives). The overall accuracy is 69% with the cutoff value of $10\ \mu\text{M}$. Among the predicted 105 blockers, 72 compounds were correctly predicted, and for the nonblockers, 847 of the 1210 compounds were successfully classified. Overall, this initial constraining strategy provided marginal improvement.
 - (ii) Application of the logP constraint, as well as the Lipinski rule-of-five constraint, resulted in a filtered test set of 876 compounds from the PubChem repository database (29 active and 847 inactive compounds). When the IC_{50} cutoff value of $10\ \mu\text{M}$ was applied to this 876 compound test set, the overall accuracy realized is 82% with sensitivity and specificity values of 97% and 82%, respectively. Thus, these findings demonstrate that the binary form of the QSAR model, eq 16, is reliable and robust in classifying hERG toxicity for drug-like compounds.

Only one of the 29 active compounds, NSC7814, was misclassified in the PubChem test set that consists of 876 compounds. NSC7814, shown in Figure 3, is arguably only weakly active and has structural and property features that have been suggested should *not* lead to active hERG blockages. Some hydrophilic groups, such as benzenesulfonyl, hydroxide groups, and amine tend to abolish hERG blocking potency as shown in the Song and Clark study.¹⁷ The misclassified active compound, NSC7814, contains

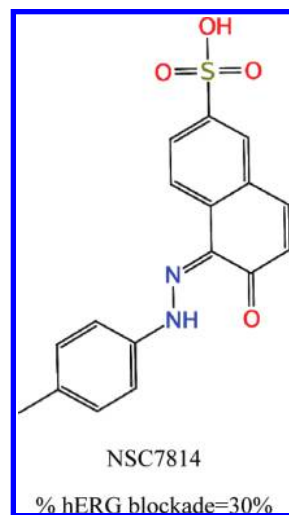


Figure 3. Chemical structure of the one weak hERG binder from the PubChem test set that were misclassified by eq 16 when using a $10\ \mu\text{M}$ cutoff.

benzenesulfonyl, hydroxide, and amine functional groups that are chemical fragments associated with compounds that are not hERG blockers.

For the nonblockers, 154 out of 847 compounds were misclassified as hERG binders. Among the 154 misclassified compounds, the average number of nitrogen atoms is approximately 3, and only 7 of these compounds do not have any nitrogen atoms. It has been established that a basic nitrogen atom in a compound is a molecular characteristic that is often associated with hERG affinity.²⁶ Thus, many of these 154 outliers containing basic nitrogens and exhibiting low hERG blockade in the current PubChem assay, might actually be weak hERG binders. Furthermore, many of these 154 outliers also contain *N*-methyl-piperidine, tertiary amino, benzyl, secondary aliphatic amine, sulfonamide, and hydrophobic groups all of which have been suggested as important features for hERG binding. Four examples of outlier compounds containing these pro-active hERG groups are shown in Figure 4.

Two compounds predicted by eq 16 as false positives are given in Figure 5. These two outlier compounds both have multiple polar-negative groups located 6–8 Å from hydrogen bond donors which, according to eq 16, increases the propensity for hERG binding. Additionally, these compounds each contain a large number of polar groups that increase the *vsurf_CW2* term of eq 16 and, thus, hERG potency. Hence, these two compounds are predicted to be hERG blockers using eq 16 with the cutoff value set to $10\ \mu\text{M}$.

Table 4b lists the test set accuracy (Acc), sensitivity (Sen), and specificity (Spe) for the PubChem data set (all compounds and the filtered data set of 876 compounds) for the ten 80–20 models. The PubChem external test sets demonstrate the robustness of the models through (a) high individual accuracy values with the average accuracy of the ten models (for the filtered PubChem data set) the same as the best model (eq 16) and (b) the sensitivity for the best model is higher than the models based on the randomly selected training set.

Thus, it is reasonable to suggest eq 16 is highly reliable and functionally useful for the prediction of hERG toxicity across a wide range of chemistry. Overall, eq 16 could actually be more robust in correctly making hERG activity predictions than is realized from evaluation and validation

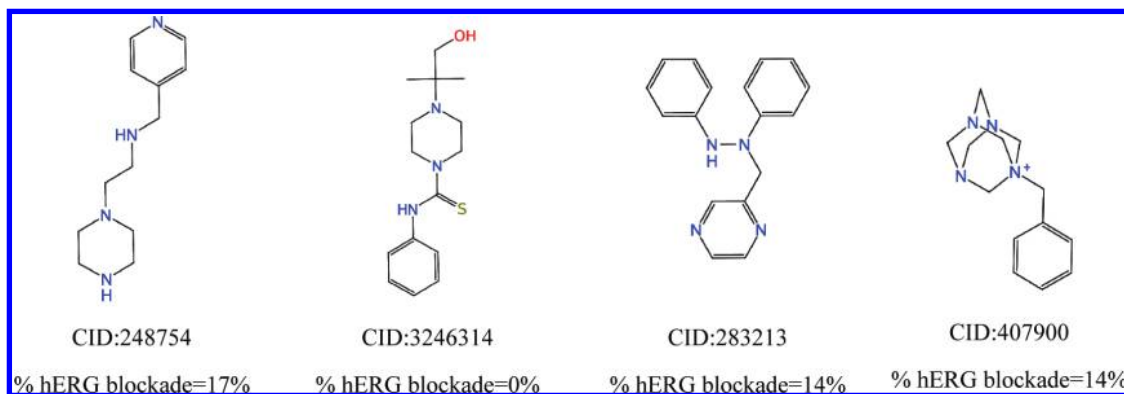


Figure 4. Four examples of hERG nonblockers in the PubChem test set that are misclassified by eq 16 using the 10 μ M cutoff yet contain significant hERG binding features.

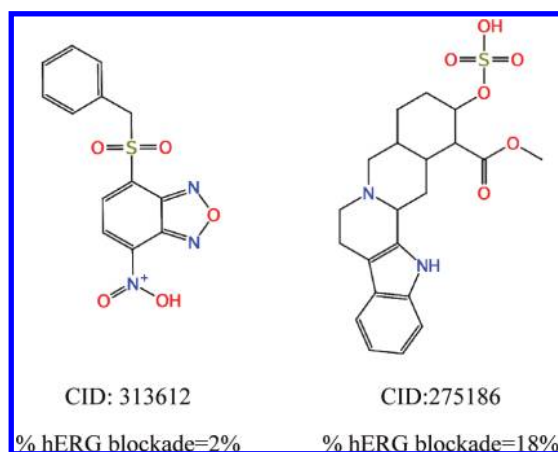


Figure 5. Examples of hERG nonblockers in the PubChem test set that are misclassified by eq 16 using the 10 μ M cutoff yet contain a relatively large number of hydrogen bond donor or acceptor atoms.

studies using the PubChem test set because of limitations in the fidelity and comparability of the hERG activity measures using this test set.

The second external test set of 106 compounds was used to further validate eq 16 as a reliable hERG screening model. When the IC_{50} cutoff value of 40 μ M was applied to the second external test set, the overall accuracy was 77% with corresponding sensitivity and specificity values of 92% and 38%, respectively. Even though the calculated root-mean-square deviation (rmsd) between the observed and predicted pIC_{50} was 1.39, eq 16 has a high sensitivity score that is also found in the external PubChem data set (sensitivity of 74% for all compounds, and 97% for filtered compounds). Again, this illustrates that eq 16 is very reliable in being able to predict active hERG compounds. Additional data on this validation analysis are given in the Supporting Information Table S2.

Model Comparisons. To further compare the quality and predictive power of the continuous and binary QSAR models constructed in this study, five published hERG studies employing different methods and models were compared to the findings of this study and are summarized in Table 7. The first column of Table 7 describes each of the modeling methodologies, the second column contains the training set accuracy and the number of training set compounds in parentheses, and the last column has the test set accuracy with the number of test set compounds also given in parentheses. The summary findings from applying each

method, except for those from this investigation, are sorted (in ascending order) by the number of compounds of the training set. The studies listed in the second through fifth rows in Table 7 have good training set accuracy, but these models are based upon relatively small training sets. Conversely, the support vector machine study was applied to a large training set, but, correspondingly, the training set accuracy is reduced. The size of the supervised neural network study and the molecular similarity-based study are near to the number of compounds in our training set. Among these three studies, the training set predictions from the models of this study are slightly less significant than the models from the other two studies. However, it is quite reasonable to ascertain that the neural network study involves a degree of nonlinear fitting, and the use of more descriptors than presented in eq 16. It is hard to determine how many descriptors were used, what the fitting protocols were used for the molecular similarity-based study and the accuracy of the similarity-based strategy is limited to the chemical space of the applied training set. The overall accuracy values of the test sets (internal and external), when predicted with eq 16, are comparable to the other referenced hERG blocker prediction studies.

Two other evaluation analyses, based upon using the entire PubChem hERG data set as a test set, and each analysis using a different model development method, namely, SVM classification²⁶ and random forest (RF) decision tree analysis,⁴⁹ were carried out. A summary of the findings from these two studies is as follows: (a) The SVM model achieved 73% accuracy (57% sensitivity). (b) The RF model achieved about 69% accuracy (47% sensitivity). Overall, the predictive performance of eq 16 in this comparative study is superior to both the SVM and the RF models.

DISCUSSION

The continuous QSAR models derived from the three molecular descriptor pools used in this study have (a) r^2 and q^2 values that are nearly the same, (b) accuracies, as binary classification models, that are at least 90%, while their r^2 values to predict pIC_{50} only explain around 60% of the of the training set variance, and (c) their performance as binary classification models is not particularly sensitive to the cutoff value used for classifying compounds as active or inactive.

Overall, the QSAR models developed in this study are not particularly good in precisely explaining the continuous pIC_{50} values of the training set compounds but are quite good

Table 7. Comparisons of the Performances of Several hERG Models Including eq 16 of this Paper

modeling methodology	accuracy of training set predictions (number of compounds)	accuracy of testing set predictions (number of compounds)
binary QSAR, this study (40 μM cutoff)	91% (250)	83% (876)
PLS (traditional and hologram QSAR) ⁵⁰	83–87 (55)	83% (95)
shape signatures ³⁰	69%~73% (83)	85–95% (21)
fragment-based–evolutionary algorithm ⁵¹	87–89% (70–100)	85–90% (22–24)
recursive partition ⁵²	96% (100)	93–96% (55)
binary QSAR model ³²	83–87% (150–223)	78–86% (64)
supervised neural network ⁴⁷	93% (244)	82% (72)
similarity-based method ³⁸	76% (275)	80% (500 ^a)
supported vector machine, SVM ²⁶	70–86% ^a (495)	73% (1877)

^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.

Table 8. Relative Importance of the Descriptor Terms of Eq 16 when *PEOE_VSA_POL*, *E_SOL*, *vsurf_CW2*, and *vsurf_HLI* are Grouped Together and Considered as a Single Descriptor

descriptor name	relative importance of descriptors over the training set for eq 16
$\epsilon 1(\text{hs}, \text{hs})$	−4.77
<i>PEOE_VSA_POL</i> + <i>vsurf_CW2</i> + <i>E_SOL</i> + <i>vsurf_HLI</i>	−4.12
<i>a_acc</i>	−1.58
$\epsilon 6(\text{all}, \text{hbd})$	−1.31
$\epsilon 7(\text{pn}, \text{hbd})$	1.16

in classifying the lower half of the training set pIC_{50} values from the upper half. Of the various QSAR models developed in this study, eq 16 performs slightly better for two-state classification than the others.

The most definitive 3D structure–activity interpretation from eqs 15 and 16 is that all IPE pairs, other than *pn* and *hbd*, which are separated by 6 to 8 Å, will decrease pIC_{50} . But (*pn*, *hbd*) IPE pairs 6 to 8 Å apart in a compound will increase pIC_{50} . The implication of this finding to minimize hERG blockage is to avoid having negatively partially charged atoms or groups (including hydrogen bond acceptors) in a compound that are 6–8 Å from hydrogen bond donors. The other 4D-FP structural finding, which has the highest significance in the QSAR but is not readily translated into definitive chemical structures, is that polar atoms, like nitrogen, oxygen, and sulfur atoms, which are grouped close together in space are a good feature for avoiding hERG activity.

It also remains unclear, in terms of specific chemical structures, how to take direct advantage of the other descriptors in eq 16 to decrease hERG blockage. However, *PEOE_VSA_POL*, *vsurf_CW2*, *E_SOL*, and *vsurf_HLI* can be conceptually organized into one feature class related to the solvent accessible polar surface of a compound. With this in mind, the relative importance of the descriptor terms listed in Table 5 have been recomputed with the *PEOE_VSA_POL*, *vsurf_CW2*, *E_SOL*, and *vsurf_HLI* descriptors considered, effectively, as a single descriptor and are given in Table 8. The combined contribution of these four descriptors is −4.12, and thus, this new cluster feature becomes the second most significant molecular property. This cluster feature suggests that as the gross polar solvent

accessible surface area of a compound increases, its hERG affinity will decrease.

Perhaps the most novel and useful part of this QSAR analysis and the corresponding models is that pharmacophore and molecular property information can be directly extracted and then used as a guideline in designing and refining drug candidates. As an example to demonstrate this capability, a potent hERG blocker and a quite inactive compound have been selected from the training set. The compound desmethyldastemizole, shown in Figure 6, has an IC_{50} value of 0.001 μM and is a strong inhibitor of the hERG channel. Figure 6 displays the chemical structure and a 3D ball-stick model of desmethyldastemizole in its lowest-energy conformation. Superimposed on the ball-stick model of desmethyldastemizole are representations of the important structural features, based upon the descriptors of eq 16, which increase and decrease the hERG blockage potency of the compound. A structural feature colored red indicates a positive descriptor term in eq 16 (an increase in hERG activity) and features in different shades of blue represent the negative descriptors. A darker blue color indicates a higher weight of the descriptor term to decreasing hERG activity. The distances between an oxygen atom (#1) and two nitrogen atoms (#2 and #3) are within 6–8 Å, and both distances delineate the one important feature, $\epsilon-7(\text{pn}, \text{hbd})$, that increases hERG affinity based on eq 16. In Figure 6 the blue dots depict the partial polar surface areas of the molecule. These rather sparse polar surface area regions do not significantly decrease hERG affinity. In addition, desmethyldastemizole has three hydrogen bonding acceptor atoms (#2, #3, and #4) contributing to the least significant negative descriptor term, *a_acc* of eq 16. Desmethyldastemizole has only two heavy atoms (#2 and #4), which are in close proximity to one another and thus contribute to satisfy the $\epsilon 1(\text{hs}, \text{hs})$ 4D-FP descriptor term of eq 16. Thus, desmethyldastemizole is predicted by eq 16, and shown in Figure 6, to have potent hERG activity because it lacks significant contributions from the most two important features, total solvent accessible polar-surface area and $\epsilon 1(\text{hs}, \text{hs})$, that can decrease the hERG affinity. Desmethyldastemizole is not rich in those structural features of eq 16 that explicitly increase hERG blockage.

Nifedipine can be considered an inactive compound, with respect to hERG blockage, having an IC_{50} value of 275 (μM). The equivalent graphical representations used for desmethyldastemizole in Figure 6 are given in Figure 7 for nifedipine.

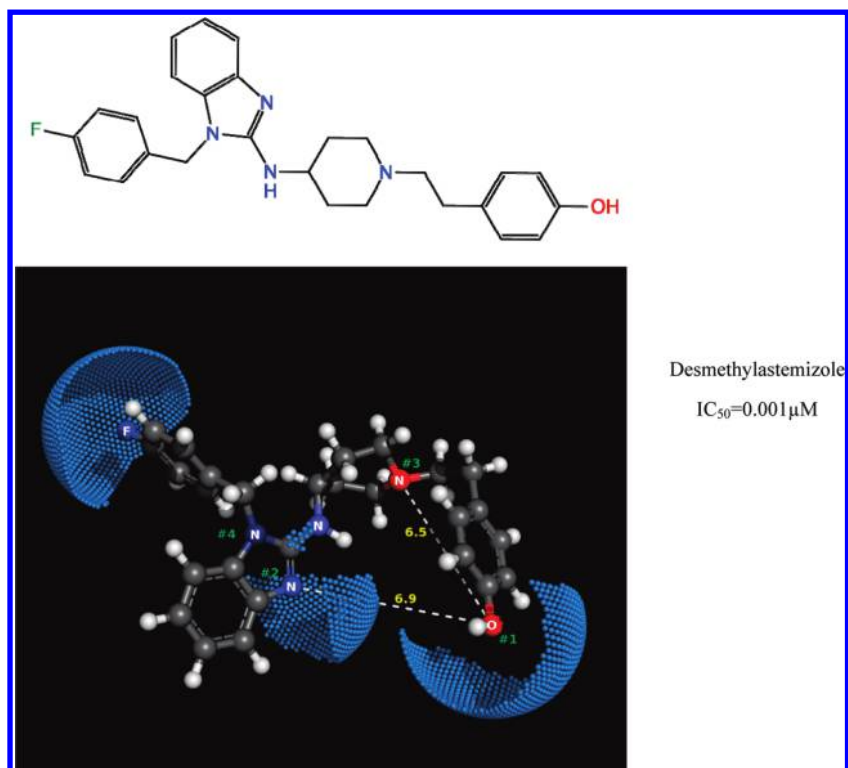


Figure 6. Desmethylastemizole (chemical structure shown at the top) in its lowest-energy conformation. The structural features, based upon the descriptors of eq 16, which contribute increasing hERG blockage are shown in red. Features that decrease hERG activity are portrayed in different shades of blue to reflect their relative importance with dark blue representing the most highly weighted features. The blue dots depict the polar surface regions and the numbers along the white dotted line indicates the distances between the specified pairs of atoms in Ångstroms. Atoms of interest are highlighted with green numbers.

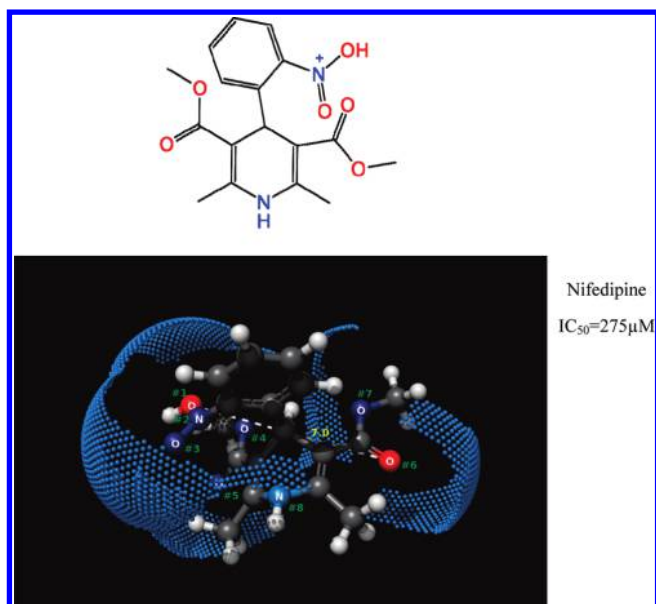


Figure 7. Nifedipine (chemical structure shown at the top) in its lowest-energy conformation. The structural features, based upon the descriptors of eq 16, which contribute decreasing hERG activity are portrayed in different shades of blue to reflect their relative importance with dark blue being most highly weighted. The blue dots depict the polar surface regions, and the numbers along the white dotted line indicate the distances between the specified pairs of atoms in Ångstroms. Atoms of interest are highlighted with green numbers. The distance between atoms #1 and #6 is the only feature that modestly contributes to the increasing hERG binding affinity, and for this reason, these atoms are colored red.

This compound has six hydrogen bond acceptor atoms (a_{acc}) (#2, #3, #4, #5, #6, and #7), which is twice the number as desmethylastemizole. It is also seen in Figure 7

that the polar surface area for nifedipine is a much larger fraction of the total surface area than is the case for desmethylastemizole. The large amounts of these two features contribute to eliminate nifedipine's hERG activity. In addition, there are three clusters of polar atoms (the first consisting of atoms #1, #2, and #3, the second being atoms #4 and #5, and the third composed of atoms #6 and #7) that lead to major contribution from the most significant molecular feature, $\epsilon 1(hs,hs)$, that diminishes the hERG binding affinity of the molecule. The only structural feature of nifedipine that increases pIC_{50} is the spatial distance of a hydrogen bond donor (hbd) atom, #1, and a negative-polar (np) atom, #6. These atoms are the required 6–8 Å apart, based upon the $\epsilon-7(pn,hbd)$ fingerprint, to enhance a compound's hERG blocking ability. Two levels of structural feature significance that decrease hERG binding are correspondingly colored by light and dark blue representations in Figure 7. The contributions to the polar surface area are colored light blue (#8, least significant). Five atoms are colored dark blue (most significant), which contribute to the $\epsilon 1(hs,hs)$ descriptor and also to the hydrogen bond acceptor atom count. The structural feature, $\epsilon-7(pn,hbd)$ formed by atoms #1 and #6, which contributes to increasing hERG activity, has, overall, significantly less weight when compared to the structural features that decrease hERG activity. The structural features which decrease hERG activity dominate the structure of nifedipine, as can be seen in Figure 7, and therefore, the compound is an inactive hERG blocker.

CONCLUSION

The QSAR models developed and presented here are applicable to both virtual screening and limited chemical

modification studies based on interpretation of the models. The approach taken herein also permits the inclusion of 3D and 4D information without introducing the alignment and pose complications that typically arise in structure-based studies. Having a set of descriptors that can be readily interpreted in terms of chemical structure is a useful feature of this modeling approach. However, it does remain problematic in deciding how much weight to assign to the inclusion or exclusion of molecular features derived from any hERG QSAR model, as compared to the need for that feature to realize therapeutic potency. Structurally modifying a compound based upon any of the descriptors in eq 16 to reduce hERG activity will, at least initially in the design phase of the drug discovery process, likely be trumped by the requirement of the pharmacophore needed for therapeutic activity. As such, the findings from eq 16 may only be implemented as guides or constraints to decrease hERG pIC₅₀ that are secondary to realizing therapeutic potency.

ACKNOWLEDGMENT

This work was funded by the Taiwan National Science Council, Grants 96-2218-E-002-019 and 96-2627-B-002-012, and National Institutes of Health through the NIH Roadmap for Medical Research, Grant R21 GM075775-01. Resources of the Laboratory of Computational Molecular Design and Detection, Department of Computer Science and Information Engineering, and Graduate Institute of Biomedical Engineering and Bioinformatics of National Taiwan University were used in performing these studies. The authors would like to thank the reviewers for their insightful comments and suggestions that helped make this a more-well rounded study.

Supporting Information Available: Listing of the 250 compound training set used in this study that includes (a) compound name, (b) the observed and predicted (using eq 16) pIC₅₀ values, and (c) the corresponding observed and predicted hERG active and (d) inactive classifications based on two different classification cutoff values (40 and 10 μ M) (Table S1). In addition, a listing of the additional literature validation set is provided that includes (a) compound name, (b) the observed and predicted (using eq 16) pIC₅₀ values, and (c) the corresponding observed and predicted hERG active and inactive classifications based upon the cutoff values of 40 μ M (Table S2). This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- Brown, A. M. Drugs, hERG and sudden death. *Cell. Physiol. Biochem.* **2004**, *35*, 543–547.
- Perlstein, R.; Vaz, R. J.; Rampe, D. Understanding the structure–activity relationship of the human ether-a-go-go-related gene cardiac K⁺ channel. A model for bad behaviour. *J. Med. Chem.* **2003**, *46*, 2017–2022.
- Sanguinetti, M. C.; Jiang, C.; Curran, M. E.; Keating, M. T. A mechanistic link between an inherited and an acquired cardiac arrhythmia: hERG-encoded the I_{Kr} potassium channel. *Cell. Physiol. Biochem.* **1995**, *81*, 299–307.
- Recanatini, M.; Poluzzi, E.; Masetti, M.; Cavalli, A.; De Ponti, F. QT prolongation through hERG K(+) channel blockade: Current knowledge and strategies for the early prediction during drug development. *Med. Res. Rev.* **2005**, *25*, 133–166.
- Doyle, D. A.; Morais Cabral, J.; Pfuetzner, R. A.; Kuo, A.; Gulbis, J. M.; Cohen, S. L.; Chait, B. T.; MacKinnon, R. The structure of the potassium channel: Molecular basis of K⁺ conduction and selectivity. *Science* **1998**, *280*, 69–77.
- Zhou, Y.; Morais-Cabral, J. H.; Kaufman, A.; MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K⁺ channel-Fab complex at 2.0 Å resolution. *Nature* **2001**, *414*, 43–8.
- Jiang, Y.; Lee, A.; Chen, J.; Cadene, M.; Chait, B. T.; MacKinnon, R. Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **2002**, *417*, 515–22.
- Jiang, Y.; Lee, A.; Chen, J.; Ruta, V.; Cadene, M.; Chait, B. T.; MacKinnon, R. X-ray structure of a voltage-dependent K⁺ channel. *Nature* **2003**, *423*, 33–41.
- Long, S. B.; Campbell, E. B.; MacKinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K⁺ channel. *Science* **2005**, *309*, 897–903.
- Farid, R.; Day, T.; Friesner, R. A.; Pearlstein, R. A. New insights about hERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorg. Med. Chem. Lett.* **2006**, *14*, 3160–73.
- Osterberg, F.; Aqvist, J. Exploring blocker binding to a homology model of the open hERG K⁺ channel using docking and molecular dynamics methods. *FEBS Lett.* **2005**, *579*, 2939–44.
- Rajamani, R.; Tounge, B. A.; Li, J.; Reynolds, C. H. A two-state homology model of the hERG K⁺ channel: Application to ligand binding. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1737–41.
- Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, universal QSAR and QSPR descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1526–1539.
- MOE (Molecular Operating Environment), 2008.10; Chemical Computing Group, Inc.: Montreal, Canada, 2008.
- Cruciani, G.; Pastor, M.; Guba, W. VolSurf: A new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29–S39.
- Talete, S. R. L. *DRAGON for Windows (software for molecular descriptor calculations)*, version 5.4; 2006.
- Stuper, A. J.; Jurs, P. C. ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques. *J. Chem. Inf. Comput. Sci.* **1976**, *16* (2), 99–105.
- MOLCONN-Z. <http://www.eslc.vabiotech.com/molconn/>. (accessed February 9, 2010).
- Song, M.; Clark, M. Development and evaluation of an in silico model for hERG binding. *J. Chem. Inf. Model* **2006**, *46*, 392–400.
- Yoshida, K.; Niwa, T. Quantitative structure–activity relationship studies on inhibition of hERG potassium channels. *J. Chem. Inf. Model* **2006**, *46*, 1371–8.
- Seierstad, M.; Agrafiotis, D. K. A QSAR model of hERG binding using a large, diverse, and internally consistent training set. *Chem. Biol. Drug. Des.* **2006**, *67*, 284–96.
- Gavaghan, C. L.; Arnby, C. H.; Blomberg, N.; Strandlund, G.; Boyer, S. J. Development, interpretation, and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 189–206.
- Hall, L. H.; Kier, L. B.; Lipkowitz, K.; Boyd, D. The molecular connectivity χ indices and κ shape indices in structure–property modeling. *Rev. Comput. Chem.* **1991**, *2*, 367–422.
- Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.
- Bruneau, P. Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1605–1616.
- Bridgland-Taylor, M. H.; Hargreaves, A. C.; Eastera, A.; Ormeb, A.; Henthorn, D. C.; Dings, M.; David, A. M.; Smalla, B. G.; Heapy, C. G.; Abi-Gerges, N.; Persson, F.; Jacobson, I.; Sullivan, M.; Albertson, N.; Hammond, T. G.; Sullivan, E.; Valentina, J.-P.; Pollard, C. E. Optimisation and validation of a medium-throughput electrophysiology-based hERG assay using IonWorks HT. *J. Pharmacol. Toxicol. Methods* **2006**, *54* (2), 189–99.
- Tagliatala, M.; Pannaccione, A.; Castaldo, P.; Giorgio, G.; Zhou, Z.; January, C.; Genovese, A.; Marone, G.; Annunziato, L. Molecular basis for the lack of hERG K⁺ channel block-related cardiotoxicity by the H1 receptor blocker cetirizine compared with other second-generation antihistamines. *Mol. Pharmacol.* **1998**, *54*, 113–121.
- Li, Q.; Jørgensen, F. S.; Oprea, T.; Brunak, S.; Taboureau, O. hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharmaceutics* **2008**, *5* (2), 117–127.
- Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. Three-dimensional quantitative structure–activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427–34.
- Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of hERG K(+) channel blockers. *J. Med. Chem.* **2002**, *45*, 3844–53.
- 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.; Rampe, D. Characterization of hERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829–35.
- (30) Chekmarev, D. S.; Kholodovych, V.; Balakin, K. V.; Ivanenkov, Y.; Ekins, S.; Welsh, W. J. Shape signatures: New descriptors for predicting cardiotoxicity in silico. *Chem. Res. Toxicol.* **2008**, *21* (6), 1304–1314.
- (31) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRIND-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43* (17), 3233–3243.
- (32) Thai, K. M.; Ecker, G. F. A binary QSAR model for classification of hERG potassium channel blockers. *Bioorg. Med. Chem.* **2008**, *16* (7), 4107–19.
- (33) Sun, H. An accurate and interpretable bayesian classification model for prediction of HERG liability. *ChemMedChem* **2006**, *1* (3), 315–322.
- (34) American Chemical Society. *SciFinder Scholar*. version 2007; Chemical Abstracts Service: Columbus, Ohio, 2007.
- (35) PubChem website. <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?cid=376>. (accessed November 18, 2008).
- (36) Weller, H. N. Purification of combinatorial libraries. *Mol. Diversity* **1999**, *4*, 47–52.
- (37) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- (38) Nisius, B.; Goller, A. H. Similarity-based classifier using topomers to provide a knowledge base for hERG channel inhibition. *J. Chem. Inf. Model.* **2008**, *49*, 247–256.
- (39) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure–permeation relationships: The VolSurf approach. *J. Mol. Struct. (THEOCHEM)* **2000**, *503*, 17–30.
- (40) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- (41) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
- (42) Friedman, J. *Multivariate Adaptive Regression Splines*, Technical Report No. 102; Laboratory for Computational Statistics, Department of Statistics, Stanford University; Stanford, CA, Nov 1988 (revised Aug 1990).
- (43) Aronov, A. M.; Goldman, B. B. A model for identifying hERG K⁺ channel blockers. *Bioorg. Med. Chem.* **2004**, *12*, 2307–2315.
- (44) O'Brien, S. E.; de Groot, M. J. Greater than the sum of its parts: combining models for useful ADMET prediction. *J. Med. Chem.* **2005**, *48*, 1287–1291.
- (45) Fioravanzo, E.; Cazzolla, N.; Durando, L.; Ferrari, C.; Mabilia, M.; Ombrato, R.; Parenti, M. D. General and Independent Approaches to Predict hERG Affinity Values. *Internet Electron. J. Mol. Des.* **2005**, *4*, 625–646.
- (46) Tobita, M.; Nishikawa, T.; Nagashima, R. A discriminant model constructed by the support vector machine method for hERG potassium channel inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886–2890.
- (47) Roche, O.; Trube, G.; Zuegge, J.; Pflimlin, P.; Alanine, A.; Schneider, G. A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. *ChemBioChem* **2002**, *3*, 455–9.
- (48) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegative—A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3288.
- (49) Aha, D. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *Int. J. Man–Mach. Stud.* **1992**, *36* (2), 267–287.
- (50) Keseru, G. M. Prediction of hERG potassium channel affinity by traditional and hologram qSAR methods. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773–5.
- (51) Bains, W.; Basman, A.; White, C. hERG binding specificity and binding site structure: Evidence from a fragment-based evolutionary computing SAR study. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 205–33.
- (52) Dubus, E.; Ijjaali, I.; Petitet, F.; Michel, A. In silico classification of hERG channel blockers: A knowledge-based strategy. *ChemMedChem* **2006**, *1* (6), 622–630.

CII00081J