

Predictive Toxicology Modeling: Protocols for Exploring hERG Classification and *Tetrahymena pyriformis* End Point Predictions

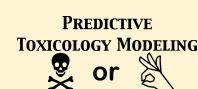
Bo-Han Su,^{†,‡} Yi-shu Tu,^{‡,§} Emilio Xavier Esposito,^{§,‡} and Yufeng J. Tseng^{*,†,‡,§}

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

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

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

ABSTRACT: The inclusion and accessibility of different methodologies to explore chemical data sets has been beneficial to the field of predictive modeling, specifically in the chemical sciences in the field of Quantitative Structure–Activity Relationship (QSAR) modeling. This study discusses using contemporary protocols and QSAR modeling methods to properly model two biomolecular systems that have historically not performed well using traditional and three-dimensional QSAR methodologies. Herein, we explore, analyze, and discuss the creation of a classification human Ether-a-go-go Related Gene (hERG) potassium channel model and a continuous *Tetrahymena pyriformis* (*T. pyriformis*) model using Support Vector Machine (SVM) and Support Vector Regression (SVR), respectively. The models are constructed with three types of molecular descriptors that capture the gross physicochemical features of the compounds: (i) 2D, 2 1/2D, and 3D physical features, (ii) VolSurf-like molecular interaction fields, and (iii) 4D-Fingerprints. The best hERG SVM model achieved 89% accuracy and the three-best SVM models were able to screen a Pubchem data set with an accuracy of 97%. The best *T. pyriformis* model had an R^2 value of 0.924 for the training set and was able to predict the continuous end points for two test sets with R^2 values of 0.832 and 0.620, respectively. The studies presented within demonstrate the predictive ability (classification and continuous end points) of QSAR models constructed from curated data sets, biologically relevant molecular descriptors, and Support Vector Machines and Support Vector Regression. The ability of these protocols and methodologies to accommodate large data sets (several thousands compounds) that are chemically diverse – and in the case of classification modeling unbalanced (one experimental outcome dominates the data set) – allows scientists to further explore a remarkable amount of biological and chemical information.



■ INTRODUCTION

The field of toxicology modeling, and in general predictive modeling, borrows heavily from nonphysical science based research areas such as statistics, psychology, sociology, and computer science. Thus, the field of predictive modeling benefits from a cornucopia of methodologies and protocols that is adept at managing data that is unbalanced, diverse, and correlated due to hidden factors. Like most areas of chemistry and biology, or any active research area, there are some systems (data sets) that have yet to be adequately described or analyzed because satisfactory methods and protocols have not been discovered.

The recently published article by Huang and Fan¹ explores two notoriously difficult systems to adequately model: the inhibition of the human Ether-a-go-go Related Gene (hERG) potassium channel and the toxicity and mutagenesis of *Tetrahymena pyriformis* (*T. pyriformis*). A genetic algorithm² was used to select molecular descriptors that were then provided to a support vector machine (SVM)^{3–5} to construct the models; the importance of each molecular descriptor was determined by the SVM. Unfortunately the work presented by Huang and Fan¹ did not take advantage of protocols or methodologies that would have improved the predictive abilities of their models. Using the hERG and *T. pyriformis* data sets and a universal protocol that is similar to the one

implemented in the Huang and Fan's article,¹ the predictive models discussed herein are markedly better at discerning if a compound is hERG active or toxic to *T. pyriformis* for their respective data set. The presented protocols take advantage of known physical properties to analyze and focus the data sets, while commonly employed protocols are used to construct sound models. The protocols are presented in a straightforward manner that can be easily adapted and used to construct predictive models for other data sets.

The human Ether-a-go-go Related Gene potassium channel is considered a critical and major component 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^{6–9} occurs. 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. Many *in silico* hERG models, using Quantitative Structure–Activity Relationship (QSAR) approaches, have been published to predict if a drug candidate has the inclination to block the hERG channel.^{10–13} Among the applied classification methodologies^{14,15} are Bayesian,¹⁶ decision

Received: January 31, 2012

Published: May 29, 2012

tree,¹⁷ neural network,¹⁵ support vector machine,^{18–21} and partial least-squares (PLS).²² A PLS classification hERG model built by Keseru²² had an accuracy of 85% for a training set of 55 compounds and an accuracy of 83% for a test set of 95 compounds. Sun¹⁶ published a Bayesian-based classification model using a training set of 1979 in-house compounds and a test set of 66 compounds. Sun's hERG classification model had a receiver operating characteristic (ROC) accuracy of 87% for the training set and a ROC accuracy of 88% for the test set. Gepp and Hutter¹⁷ reported a decision tree hERG classification model with an accuracy of 92% for a training set of 264 compounds and an accuracy of 76–80% for a test set of 75 compounds. Roche et al.¹⁵ implemented supervised neural networks to construct a hERG classifying model with an accuracy of 93% for a 244 compound training set and an accuracy of 82% for a test set comprised of 72 compounds. Li et al.²³ published a hERG classification model that employed a SVM and achieved an overall classification accuracy of 74% for the training set of 495 compounds and an accuracy of 73% for a test set of 1877 compounds from a PubChem data set (AID 376).²⁴ Overall, a sampling of successful hERG models from the literature has mostly used machine-learning methods to achieve high accuracy for the training set compounds. Among the studies presented above, only the model built by Li et al. resulted in a lower accuracy for the training and test sets than the other studies, but they used a considerably larger training set of 495 samples (compounds) and a test set with close to 1800 more compounds than the other models (72 to 95 compounds). Moreover, the previously presented model constructing methods – with the exception of Li et al.'s protocol – lacked sufficient model validation because they were only applied to small test sets containing between 72 and 95 compounds. Huang and Fan¹ used Li et al.²³ hERG training set of 495 compounds to construct SVM classification models whose descriptors were selected by a genetic algorithm^{2,25,26} (GA). The classification model was applied to an external test set of 1948 compounds from the PubChem bioassay database (AID 376). The best resulting model had an accuracy near 87% for the training set and 82% for the test set.¹

For better predictive accuracy and mechanistic understanding of the system of interest, it is preferred that only a single chemical mechanism is being captured within the QSAR model. In the field of toxicology predictive modeling, however, it is hard to confirm the detailed chemical mechanism being measured in the toxicological experiment. Moreover, many toxicological experiments are comprised of complicated chemical reactions, but effective predictive models are still needed to represent these toxicological experiments.²⁷ Additionally, the risk of overfitting or construction of models by chance is reduced due to the structural diversity of the training set and the use of a multiclass trial descriptor pool. Thus these models can be incorporated into a toxicology screen and have the potential to reduce the need for bioassays.

The *Tetrahymena pyriformis* assay was developed for predicting the aqueous toxicity of a compound as it relates to fish lethality.²⁸ Schulz and co-workers have analyzed more than 2,400 compounds and constructed the TETRATOX database²⁹ that is not publicly available. The Schulz and Cronin's group^{30–34} have published many *T. pyriformis* QSAR studies and models that explore subsets of the compounds based on functional groups. For these sets of QSAR models, the authors applied linear regression to a set of physicochemical molecular descriptors, such as HOMO and LUMO energies and logP (the

octanol–water partition coefficient), to construct the QSAR models. Although these QSAR models had regression coefficient of determination values that indicate statistical significance ($R^2 > 0.8$), the predictive abilities of these QSAR models was tuned for compounds that are similar to those used to build the model, i.e. the training set compounds.

Zhu and colleagues³⁵ built a general QSAR model to predict the *T. pyriformis* toxicity of compounds, collected 983 compounds from Schultz's publication, and randomly divided the compounds into a training set (644 compounds) and a test set (339 compounds). To validate the training set models, Zhu et al. collected an additional 110 compounds that were published more recently by Schultz et al.,³⁴ as a secondary test set. In the study by Zhu and co-workers, they applied several commonly used model construction methods to construct discrete QSAR models. Specifically, individual QSAR models were constructed using *k*-nearest neighbors (*k*NN), support vector machine (SVM), machine-learned ranking (MLR), ordinary least-squares (OLS), partial least-squares, associative neural network (ASNN), and artificial neural network (ANN); the optimal model from each method was combined to form a consensus model. The consensus model returned an R^2 value of 0.94 for the training set, 0.88 for test set 1, and 0.77 for test set 2. The individual models comprising the consensus model possessed a mean Q^2 value of 0.85 for the training set, and these values ranged from 0.72 to 0.95 for the individual models. Additionally, the mean R^2 value for the individual models applied to test set 1 was 0.75 (ranging from 0.49 to 0.85) and 0.51 for test set 2 (ranging from 0.37 to 0.66). The best performing individual model not included in the consensus model was constructed using ASNN; this model had an R^2 value of 0.85 and a Q^2 value of 0.83 for the training set and R^2 values of 0.85 and 0.66 for test sets 1 and 2, respectively.

In the study by Huang and Fan,¹ the authors used a GA^{2,25,26} to select the molecular descriptors and a SVM to construct a continuous model for the *T. pyriformis* data set. However, the GA descriptor selecting protocol made it difficult to find important descriptors that influence the model and was most likely the rational for Huang and Fan to suggest that molecular descriptors are not important when building a QSAR model. The best *T. pyriformis* SVM-GA model in the Huang and Fan study had a Q^2 of 0.86 and coefficient of determination (R^2) values of 0.88 and 0.60 for test sets 1 and 2, respectively. Additionally, the data sets of Zhu and co-workers have been used to explain factors that might influence a QSAR model. In the study of Tetko et al.,³⁶ the *T. pyriformis* data set was used to demonstrate that the applicability domain of a QSAR model is determined by the distance (molecular similarity) between the compound whose end point is being predicted (commonly the test set compound) and the training set.

The training set is the cornerstone of a sound and robust predictive or QSAR model. Constructing a training set comprised of diverse and unique compounds that are within known physicochemical limits aids the predictive ability of models. To improve the accuracy of the SVM-based hERG model, compounds for the training set were carefully selected in an effort to filter out redundant information. In a previous study, we reported a hERG binary classification QSAR model³⁷ constructed using the genetic function approximation³⁸ (GFA) methodology, and this model was better at predicting a compound's propensity to be a hERG channel blocker than other published classification models.^{15,22,23,39–42} The training

set was constructed from a set of 250 structurally diverse compounds obtained from the literature with known hERG activity and a condensed version of the PubChem bioassay (AID 376) containing 876 compounds. This hERG classification model achieves 91% accuracy for the training set and 83% accuracy for the test set. Furthering our work in the area of hERG blocking classification modeling, we published another study⁴³ addressing the active-versus-inactive imbalance typically seen in high-throughput screening results. The PubChem hERG Bioassay data set (AID 376; 163 active and 1505 inactive compounds) was used as the training set after it was pruned of compounds violating Lipinski's Rule-of-Five and those that did not fall within the specified logP range.⁴³ To avoid overfitting the SVM model, the linear SVM modeling and deletion procedure was applied to reduce the size of the training set descriptor pool used to construct the model and then judiciously selected molecular features from the reduced descriptor pool. This is the preferred approach and maximizes the correct classification of compounds for hERG toxicity. An external data set (the test set) consisted of 356 compounds collected from available literature data and was comprised of 287 actives and 69 inactives; this collection of compounds was used to validate the models. The accuracy, sensitivity, and specificity of our best model determined from 10-fold cross-validation were 95%, 90%, and 96%, respectively; the overall accuracy was near 87% for the external test set. The knowledge gained from our previous hERG studies is applied to the data set used by Huang and Fan¹ along with the descriptor selection and model construction methods presented here to construct robust predictive models.

MATERIAL AND METHODS

hERG Data Sets. The hERG training set used in this study was derived from the one used in the study by Huang and Fan.¹ Huang and Fan combined the 495 training set compounds and the 66 external compounds collected by Li and co-workers²³ and designated this as their training set. In this study, Huang and Fan's initial training set of 561 compounds was cleaned to remove compounds complexed with metal ions along with those that sampled unstable conformations during the molecular dynamics simulation for the calculation of 4D-Fingerprints (4D-FPs). A total of 546 compounds remained and our training set consisted of 210 active and 336 inactive compounds. Initially, the hERG test set contained 1948 compounds obtained from the PubChem BioAssay AID 376.²⁴ The hERG test set was reduced to 1795 compounds (220 active and 1575 inactive) in the study by Huang and Fan,¹ because the molecular descriptors for 153 compounds could not be calculated by the chosen molecular descriptor application; DRAGON.² It is not uncommon for several compounds of a large data set to have difficulties when calculating their molecular descriptors. We used the same cleaning protocol for the hERG test set—as for the training set—and compounds with metal ions, structurally ambiguous compounds (an SDfile entry with two or more compounds), hERG activators, and compounds also present in our training set were removed. The cleaned test set contained a total of 1668 PubChem compounds comprised of 163 active and 1505 inactive compounds. The hERG data set compounds from PubChem AID 376 and the literature compounds were obtained as 2D molecular structures, converted into 3D structures using HyperChem 7.0⁴⁴ and geometry optimized

using HyperChem 7.0s MM+ force field (based on the Allinger MM2 force field⁴⁵).

The *Tetrahymena pyriformis* Toxicity Data Set. The *Tetrahymena pyriformis* toxicity data set was retrieved from the article published by Zhu et al.³⁵ This data set provided the compounds as SMILES strings along with each compound's biological end point as the negative logarithm of the concentration required to inhibit growth by 50% (pIGC50). The data set was separated by Zhu et al.³⁵ into a training set of 644 compounds and two external validation sets (test sets) with 339 and 110 compounds, respectively. The Zhu et al. *T. pyriformis* data set has also been used to explore the domain applicability of QSAR models³⁶ and factors influencing the reliability of QSAR models.¹ The 3D conformation of the compounds in the *T. pyriformis* data set were constructed and energy minimized using the MMFF94x force field and atomic partial charges⁴⁶ with Born solvation (MOE 2010.10 software⁴⁶).

Molecule Descriptors. The molecular descriptors used to construct the QSAR models for the hERG and the *T. pyriformis* data sets were obtained from two sources. The 2D, 21/2D, 3D, VolSurf-like, and Semi-Empirical molecular descriptors of each compound were calculated with Molecular Operating Environment (MOE) 2010.10 software⁴⁶ using MMFF94x force field with the Born solvation model. To capture high-dimensional dynamic molecular information, 4D-Fingerprints⁴⁷ of each compound were calculated and used as molecular descriptors.

Semi-Empirical Molecular Descriptors. Seven AM1 Semi-Empirical molecular descriptors, calculated with MOE 2010.10, are part of the trial descriptor pool. These molecular descriptors values capture the electronic physicochemical properties of the compounds, specifically the dipole moment, total SCF energy, electronic energy, heat of formation, the HOMO and LUMO energies, and the ionization potential.

2D and 21/2D Molecular Descriptors. Adding to the trial descriptor pool are 228 MOE 2D and 21/2D molecular descriptors. 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 & Hall connectivity and kappa 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 and included measures of the conformational potential energy and its components, molecular surfaces, volumes and shapes, and conformation dependent charge descriptors. These descriptors are dependent on the conformation of the molecule.

VolSurf Molecular Interaction Fields. The VolSurf^{48,49} descriptor set contains 76 molecular features based on molecular interaction fields. These descriptors are alignment independent and are not strongly dependent on each compound's molecular conformation; the 3D molecular interaction fields are represented as a single numerical value. The compound is placed in a grid (with the exception of four VolSurf descriptors that measure the molecular volume, surface area, globularity, and rugosity), 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 combinations of interaction energies and molecular volumes are used as molecular descriptors.

Universal 4D Fingerprints. The 4D-Fingerprints (4D-FP) were developed to model⁴⁷ and classify⁵⁰ compounds using the conformation information of a compound from a molecular dynamic (MD) simulation. With a compound sampling multiple conformations during a MD simulation, the problem that too few conformations are analyzed in traditional 3D fingerprints is alleviated. The descriptor size of 4D-FP varies related to the number of atoms within a molecule. A total of 613 4D-FPs were calculated and used for *T. pyriformis* data set, and 5271 4D-FP descriptors were calculated for the hERG data set. The difference in the number of 4D-FPs between the hERG and *T. pyriformis* data set is due to the physical size of the compounds in each data set. The largest compound in the hERG data set contains more atoms than the largest compound in the *T. pyriformis* data set, thus the hERG data set required more 4D-FPs to capture the molecular information contained in the hERG compounds. The time to compute the 4D-FPs features of hERG data set and *T. pyriformis* data set are approximately 850 and 490 s (or 14 and 8 min), respectively. These durations are acceptable for the number of compounds of interest, and inclusion of these descriptors with those of standard molecular calculation packages is reasonable.

Support Vector Machine (SVM) and Support Vector Regression (SVR). A support vector machine (SVM)^{4,5} is a supervised machine-learning technique that applies a hyper-plane within the descriptor space in an attempt to separate (classify) the samples, whereas support vector regression (SVR)³ applies cost functions to the support vectors and is suitable for regression model building when the samples have continuous end points. The end points for each compound (sample) of the hERG data set are binary – the compounds are classified as active (1s) or inactive (0s) – while the end points for the *T. pyriformis* data set are continuous (the logarithm of 50% growth inhibitory concentration, pIGC₅₀, values ranging from –2.67 to 3.05). Thus, the hERG classification models were constructed using a SVM, and the *T. pyriformis* continuous models were constructed using a SVR. Models were constructed and validated using the LIBSVM v2.88⁵¹ application and interfaced to R v2.12.2⁵² using the e1071 package v1.6.⁵³

Partial Least Squares (PLS). To identify important molecular descriptors that accurately represent the pIGC₅₀ end points of the *T. pyriformis* data set, the PLS loading matrix was used to assist in the selection of descriptors. PLS applies principal component analysis and regression methods to highlight important features and suggests approximate models using the descriptors and activity values. The pls package v2.2⁵⁴ in R v2.12.2⁵² was used to calculate the PLS loading matrix.

Evaluation of the Classification Model's Predictive Ability. To provide a broader understanding of the classification model's performance, the *accuracy* (*Acc*; correctly predicted active and nonactive compounds; eq 1), *sensitivity* (*Sen*; correctly predicted active compounds; also referred to as *Recall*; eq 2), *specificity* (*Spe*; correctly predicted nonactive compounds; eq 3), and the *Geometric-Mean* (*G-mean*; the square-root of the *sensitivity* multiplied by the *specificity*; eq 4) are important model evaluation criteria that should be considered when examining models. When *accuracy* is the only evaluation measure of a classification model reported, the true predictive nature of the model for the active and inactive entities can be biased because all of the compounds (entities)

are included in the *accuracy* measure and inaccuracies in the prediction of the actives (positives) or the inactives (negatives) are diminished. Calculating the *G-mean* value – taking into consideration the *sensitivity* and *specificity* – provides a single value (like *accuracy*) that provides a realistic predictive ability value of the model (unlike *accuracy*)

$$accuracy = \frac{tp + tn}{tp + fn + tn + fp} \quad (1)$$

$$sensitivity = \frac{tp}{tp + fn} \quad (2)$$

$$specificity = \frac{tn}{tn + fp} \quad (3)$$

$$Geometric - Mean = \sqrt{sensitivity \times specificity} \quad (4)$$

where *tp* is the number of correctly predicted positives (true positives; actives), *tn* is the number of correctly predicted negatives (true negatives; inactives), *fp* is the number of incorrectly predicted positives (false positives; negatives incorrectly predicted to be positives; inactives predicted to be actives), and *fn* is the number of incorrectly predicted negatives (false negatives; positives incorrectly predicted to be negatives; actives predicted to be inactives). A fifth method to analyze classification models is Cohen's kappa (κ), and it measures the agreement between classification models or predicted and known classifications.⁵⁵ It is defined as

$$Cohen's \kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (5)$$

where $Pr(a)$ is relative observed agreement between the predicted classification of the model and the known classification, and $Pr(e)$ is the hypothetical probability of chance agreement. The $Pr(a)$ and $Pr(e)$ values are calculated from a confusion matrix. Cohen's *kappa* analysis returns values between –1 (no agreement) and 1 (complete agreement). Predictive models, when compared to the known classification of the data set, with Cohen's *kappa* values between –1.0 and 0.4 indicate that the model is a poor predictor, values between 0.4 and 0.6 indicate that the model is average, values between 0.6 and 0.8 imply that the model is acceptable, and values between 0.8 and 1.0 denote that the model is highly predictive. While *G-mean* is the primary model evaluation method for this study, this quintet of classification model evaluation measures are included to determine the full abilities of the classification model.

Distance to Model Calculation. Tetko and colleagues suggested that the descriptor distances between the test compounds and the training set influence the predictability of QSAR models.³⁶ To remove potential outliers from test set 2 of the *T. pyriformis* data set, the overall and 3-nearest-neighbor (3-NN) mean Euclidean distance between the training set and each test set compound was calculated using the scaled descriptors from the 204-term SVR model.

RESULTS AND DISCUSSION

hERG Data Set. SVM Model Building Using the Raw hERG Training Set. The complete hERG training set (the raw training set) contains 210 hERG blocking compounds and 336 hERG inactive compounds resulting in a ratio of 5:8, actives to inactives. All 4D-FPs and MOE descriptors were used to build

Table 1. Prediction Performance Measures for the hERG Data Set from the Raw Training and Test Set SVM Models Determined Using All MOE and 4D-FPs Descriptors (Name *raw* in the Third Row), 1289 Selected Features (Name *select1289* in the Fourth Row), 1000 Selected Features (Name *select1000* in the Fifth Row), and 900 Selected Features (Name *select900* in the Sixth Row)

model	training set percentage (number correct/total number)				test set percentage (number correct/total number)			
	accuracy	sensitivity	specificity	G-means	accuracy	sensitivity	specificity	G-means
<i>raw</i>	81 (440/546)	85 (179/210)	78 (261/336)	81.4	61 (1015/1668)	64 (105/163)	60 (910/1505)	62.0
<i>select1289</i>	81 (440/546)	76 (159/210)	84 (281/336)	79.9	68 (1136/1668)	53 (87/163)	70 (1049/1505)	60.9
<i>select1000</i>	81 (444/546)	85 (179/210)	79 (265/336)	81.9	74 (1231/1668)	43 (71/163)	77 (1160/1505)	57.5
<i>select900</i>	79 (430/546)	90 (188/210)	72 (242/336)	80.5	82 (1365/1668)	41 (67/163)	86 (1298/1505)	59.4

the classification SVM model. The best raw training set SVM model was selected based on the *G-mean* metric (eq 4), and the training and test set performances are listed in the third row of Table 1. The *accuracy*, *sensitivity*, *specificity*, and *G-mean* values for the training set are listed in the first four columns, while the last four columns provide these same measures for the test set. The *G-mean* value for the raw training set SVM with 10-fold cross-validation was 81% along with an *accuracy* of 81%. The *G-mean* when applying the model to the external test set was 62% with an *accuracy* of 61%. This illustrates that the “best” classification SVM model built from a raw training set without limitations applied to the selection of descriptors produces an overfit model with poor performance when classifying the hERG test set. Constructing a SVM model using 5577 descriptors (5271 4D-FPs and 306 MOE descriptors) could reduce the performance of the model’s predictive power since there is the possibility that a majority of the molecular descriptors are intercorrelated, and therefore it is harder to interpret the molecular features associated with a compound that exhibits hERG blocking behavior.

SVM Model Building Using “F-Score” Filter for Feature Selection. A simple and effective “F-score” technique for feature selection was adopted to identify the influential model features (molecular descriptors) and reduce the need to use a large descriptor set during SVM model construction.⁵⁶ F-score is a simple method that evaluates the discriminative nature of positive and negative instances. The larger the F-score, the more likely the molecular feature is especially discriminative. Used as a criterion to identify if a feature is important for a SVM model, the F-score was used to explore the impact of limiting the number of molecular descriptors (features) used to construct a classification SVM for the hERG data set. The values of calculated F-scores of the 5577 molecular descriptors ranged from 0.0 to 0.27. Most of the F-score values for the molecular descriptors were 0.0 (no significant influence), and only 200 descriptors had F-score values greater than 0.1. This indicates that most of the descriptors have small F-scores, and none of the descriptors have especially discriminative characteristics based on the F-scores evaluation. Therefore, three F-score threshold values of 0.0, 0.001, and 0.01 were applied to retain molecular descriptors with F-score values greater than the defined thresholds. Trial descriptor pools containing 1289, 1000, and 900 descriptors were used to construct and evaluate SVM models. The performance of the best SVM model built using these three feature sets are listed in the fourth to sixth rows of Table 1. The *accuracy*, *sensitivity*, *specificity*, and *G-mean* values for the training set are listed in the first four columns, while the last four columns provide these same classification model evaluation measures for the test set.

The 1289 most influential features based on the largest F-score values, from the set of 5577 molecular features, were

selected. The best training set SVM model built with 1289 features and subjected to 10-fold cross-validation has a *G-mean* value of 80% along with an *accuracy* of 81%. Applying this model to the hERG test set resulted in a *G-mean* value of 61% and an *accuracy* of 68%. Compared to the SVM model constructed for the raw training set with all of the molecular features, the 1289-feature classification model had approximately the same *accuracy* for the training set and a pronounced improvement with respect to the *accuracy* for the test set. The 1000-feature SVM model for the training set – again evaluated with 10-fold cross-validation – returned a *G-mean* value of 82% with an *accuracy* of 81%. The *G-mean* value of the test set when evaluated with the 1000-feature model was 58% with an increased *accuracy* of 74%. The best 900-feature SVM classification model for the training set with 10-fold cross-validation had a *G-mean* value of 81% along with an *accuracy* of 79%. The *G-mean* value for the classification of the test set using the 900-feature SVM model was 59% with an *accuracy* of 82%. By constructing classification SVM models using the most prominent molecular descriptors (features) the *accuracy* of the predictions for the test sets is improved (the training set *accuracy* remains constant), whereas the *G-mean* score is steady for the training set while decreasing for the test set. The *G-mean* score for the test set is reduced in concert with the reduction of the number of features used to construct the SVM models. This is most likely due to the reduction in the number of molecular features – removal of descriptors that add noise to the model – while preserving the important information contained within the subset of molecular descriptors used to construct the model.

SVM Model Building by Adopting the logP Filter and Reduced 4D-FPs. Implementing a logP filter to focus the training set, as carried out in our previous hERG blocking GFA modeling study,³⁷ combined with the Lipinski’s Rule-of-Five⁵⁷ filter reduced the size of the training set to 206 compounds (37 active and 169 inactive compounds) and a test set of 876 compounds (29 active and 847 inactive compounds). The Lipinski’s Rule-of-Five is a well-known filter used to classify compounds as drug-like or not and can be used as a high-throughput screening protocol to partition the compounds of interest. Compounds considered to be nondrug-like, based on Lipinski’s Rule-of-Five, might have unreliable experimental end points and were considered noisy data, thus they discarded. It is well-known that there is a direct correlation between a compound’s hydrophobic nature and its propensity to induce hERG blockage; increasing the hydrophobic nature of a compound increases the hERG blocking effect and vice versa.^{20,23,58,59} Thus, compounds were removed from the training set based on calculated logP values. Specifically, active compounds with a logP value less than 4.1 and inactive compounds with a logP value greater than 2.8 were removed to

Table 2. Prediction Performance Measures from the Top Three Training Set SVM Models Determined Using All MOE and 4D-FPs Descriptors and Adopting the logP and Lipinski's Rule-of-Five Constraints for the hERG Data Set^a

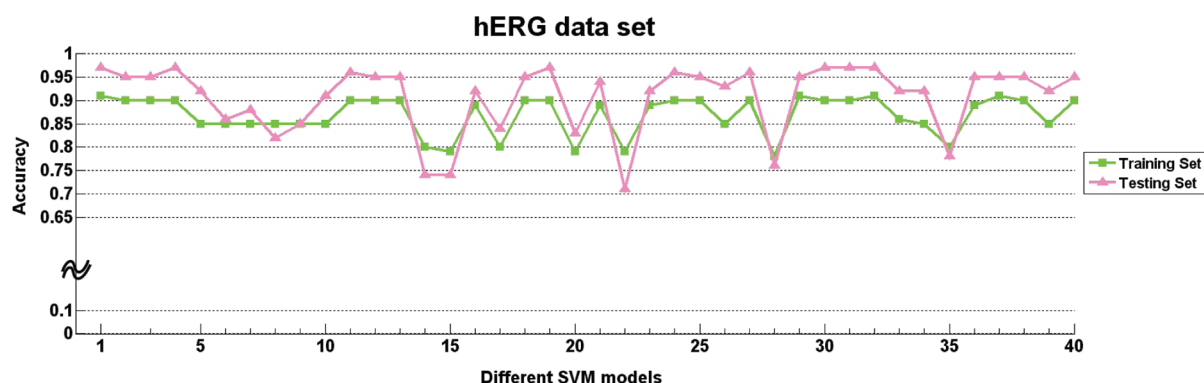
training set percentage (number correct/total number)				test set percentage (number correct/total number)			
accuracy	sensitivity	specificity	G-mean	accuracy	sensitivity	specificity	G-mean
89(184/206)	65(24/37)	95(160/169)	78.6	97(848/876)	66(19/29)	98(829/847)	80.4
84(173/206)	65(24/37)	88(149/169)	75.6	90(788/876)	83(24/29)	90(764/847)	86.4
85(175/206)	62(23/37)	90(152/169)	74.7	91(798/876)	79(23/29)	92(775/847)	85.3

^aThe models are ranked using the *G-mean* metric for the evaluation of the training set.

Table 3. Prediction Performance Measures from the Top Three Training Set SVM Models Determined Using All MOE and selected 4D-FPs Descriptors and Adopting the logP and Lipinski's Rule-of-Five Constraints for the hERG Data Set^a

training set percentage (number correct/total number)					test set percentage (number correct/total number)				
accuracy	sensitivity	specificity	G-mean	kappa	accuracy	sensitivity	specificity	G-mean	kappa
90 (185/206)	65 (24/37)	95 (161/169)	78.6	0.94	97 (850/876)	62 (18/29)	98 (832/847)	78.0	0.98
85 (174/206)	65 (24/37)	89 (150/169)	75.9	0.90	91 (798/876)	79 (23/29)	92 (775/847)	85.2	0.95
84 (173/206)	65 (24/37)	88 (149/169)	75.6	0.90	89 (782/876)	83 (24/29)	90 (758/847)	86.4	0.94

^aThe models are ranked using the *G-mean* and kappa metric for the evaluation of the training set.

**Figure 1.** Prediction performance for the top forty training set SVM models determined using all descriptors and adopting the logP and Lipinski's Rule-of-Five constraints of the hERG data set. The X-axis denotes the SVM models constructed using different parameters, and the Y-axis indicates the corresponding accuracies for the training set (green) and the test set (purple).

match the mean logP value of active and inactive compounds from our previous binary classification GFA-QSAR model for hERG blockage prediction.³⁷ Applying the logP constraint resulted in training sets that were focused toward the physicochemical requirements of the hERG receptor. The three-best performing classification SVM models derived from this data set are listed in Table 2 and ranked by their *G-mean* values. The training set values for *accuracy*, *sensitivity*, *specificity*, and *G-mean* are on the left, and the test set values are on the right, respectively. The best SVM model constructed by this protocol and validated with 10-fold cross-validation achieved 89% *accuracy* and a *G-mean* value of 79%. The three-best SVM models were used to evaluate the pruned test set (constructed from the PubChem data set; AID 376). The resulting *accuracy* and *G-mean* values for applying the best SVM model to the test set are 97% and 80%, respectively. This finding demonstrates that the best SVM model built from this data set is reliable and robust for classifying the hERG toxicity of drug-like compounds. The removal of "noisy compounds" from the training set, by employing the physical property logP filter in addition to the Lipinski's Rule-of-Five filter, effectively "cleans" the data sets and correspondingly increases the classification performance of the models, especially when employing SVM strategies.

Based on the results and success of our previous study,⁴³ combining the descriptors pool for the training set to consist of all of the MOE descriptors and a selection of the 4D-FP descriptors results in the construction of a well-performing hERG classification SVM model. The "select" set of 4D-FPs contains the hydrogen bond acceptor (HBA), polar-positive (PP), and polar-negative (PN) interaction pharmacophore elements (IPEs) that have been shown to contain important molecular information needed for hERG blockage classification. To confirm that a specific IPE is significant for any hERG data set, a SVM model based on the filtered hERG data set is constructed using the selected 4D-FPs minus the IPE of interest and all the MOE descriptors; the results are listed in Table 3. The top three hERG classification models from the training set are ranked by their *G-mean* values, and the *accuracy*, *sensitivity*, *specificity*, *G-means*, and *kappa*⁶⁰ values are listed in the first five columns while the last five columns provide these same measures for the test set. For the best SVM model, the overall *accuracy* is near 90%, with a *sensitivity* value of 65%, a *specificity* value of 95%, and a *G-mean* value of 79%. Moreover, the *G-mean* value for the top three models constructed from all the MOE and the selected 4D-FPs descriptors are all near 78%. The predicted *accuracy*, *sensitivity*, and *specificity* values for the test set using the top three models range from 89% to 97%,

62% to 83%, and 90% to 98%, respectively, with the *G-mean* values between 78% and 86%. In Table 3, Cohen's *kappa* values are also provided as another measurement of the classification accuracy. Similar to the *G-mean* values calculated for our hERG classification models, the *kappa* analysis also highly ranked the top three models with values between 0.90 and 0.98. This indicates that these models are highly effective for predicting if a compound is hERG active. These findings illustrate that the SVM models constructed using the *selected 4D-FPs* and *all of the MOE descriptors* improves the predictive ability of the models for the training and test sets.

Comparison of Model Building Methods. Since Huang and Fan¹ only provide *accuracy* values, we will also focus on the *accuracy* values within our discussion. In Figure 1, the performance of our top 40 SVM models using the logP and Lipinski's Rule-of-Five constraints and constructed along with the classification SVM models constructed with all of the molecular descriptors for the hERG data set are plotted (the X-axis denotes the top 40 SVM models and the Y-axis denotes the *accuracy* for the corresponding SVM models). The green line with squares shows the *accuracy* values for the training set, while the purple line with the triangles denotes the *accuracy* values for the test set. The training set *accuracy* ranges from 78% to 90% and from 73% to 97% for the test set. The top SVM model based on training set *accuracy*, with a value of 90%, has an *accuracy* of 97% for the test set. In the study by Huang and Fan,¹ the best SVM model had a training set *accuracy* near 100%, yet the test set had an *accuracy* of 77%.¹ Selecting the best hERG classification SVM model from the Huang and Fan study based on the *accuracy* of the test set highlights a model with an *accuracy* of 83% for the test set and an *accuracy* near 87% for the training set. The performance of our models obviously outperforms the models of Huang and Fan (and other hERG classification studies), and this can be credited to the use of Lipinski's Rule-of-Five and logP constraints to focus and balance the numbers of active and inactive hERG compounds comprising the training set.

Interpretation of Key Descriptors. In addition to improving the screening ability, sound predictive models provide the ability to interpret key molecular descriptors related to hERG cardiotoxicity. These descriptors can then be visualized on compounds known to participate – and those known not to participate – in hERG blockage, providing scientists with insight to the structural and physicochemical aspects of molecules that are hERG active and inactive. Since the same protocols from our previous SVM models⁴³ were applied to Huang and Fan's data set to construct improved hERG blockage predictive models, the important molecular descriptors are the same as those identified in our previous hERG blockage model. The seven structural features identified included five positive terms that increase the predicted value for hERG blockage, logP(o/w), *b_{ar}*, *a_{nCl}*, SlogP_VSA6, and $\epsilon^*(np,np)$, and two negative terms that reduce the predicted value for hERG blockage, $\epsilon_1(np,hba)$ and $\epsilon_4(np,hbd)$. To visually investigate the contribution of these seven descriptors to hERG blockage, projecting the structural characteristics onto molecules from the Huang and Fan data set is helpful. A strong hERG blocker, *Wombat_64*, has been selected and depicted in its 2D chemical structure and 3D ball-and-stick rendering (lowest energy conformation), Figure 2. All structural features colored red represent constructive descriptor terms that result in an increase in predicted hERG activity; the darker the shading the more influence the descriptor and thus structural

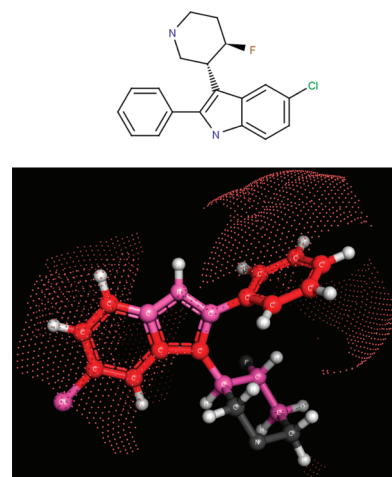


Figure 2. *Wombat_64* (2D depiction and lowest energy conformation). The structural features that contribute to the compound's increased hERG blockage are shown in red. The descriptors that increase hERG activity are depicted in varying shades of red to reflect their relative importance; dark red represents the most significant molecular descriptor. The red dots illustrate the surface regions of the nonpolar atoms. The biological end point for *Wombat_64* is $IC_{50} = 7 \mu M$.

feature imparts. *Wombat_64* does not have any structural features that contribute to the reduction of a compound's predicted tendency for hERG blockage. The physical characteristics (properties) of logP(o/w) and SlogP_VSA6 are combined and correspond to the effective solvent accessible surface area for the compound's nonpolar atoms. The red dots represent the nonpolar solvent accessible surface area, and these regions significantly increase hERG affinity and blockage. Additionally, *Wombat_64* has many nonpolar atoms that are in close proximity to one another and satisfy the significant constructive (increase the likelihood of hERG blockage) descriptor $\epsilon^*(np,np)$ along with containing many aromatic atoms to satisfy another positive descriptor, *b_{ar}* (number of aromatic bonds). The atoms associated with these molecular descriptors are colored red, while atoms that represent one hERG blocking descriptor – to increase hERG blockage – are colored pink. A molecular characteristic that increases the potential for a compound to block the hERG channel is the number of chlorine atoms, *a_{nCl}*, and *Wombat_64* also contains a single chlorine atom. Thus, *Wombat_64* – containing the five molecular descriptors [logP(o/w), *b_{ar}*, *a_{nCl}*, SlogP_VSA6, and $\epsilon^*(np,np)$] that are instrumental to a compound's ability to block the hERG channel – is an excellent example of the molecular descriptors that contribute to a strong hERG channel blocker.

T. pyriformis Data Set. The PLS Descriptor Selection of T. pyriformis Data Set. To focus the molecular descriptors and retain those that are closely correlated with the continuous biological end points, we used the PLS loading matrix to rank the descriptors. Analyzing the loadings of the first three components (Figure 3), we highlight six descriptors (Table 4) that are highly related to the pIGC50 end points. In the loadings comparison of components 1 and 2 (Figure 3a), the indicated descriptors were related to the Semi-Empirical calculated total energy (AM1_E) and electrostatic energy (AM1_Eele). Principal Moments of Inertia (PMI; pmi1, pmi2, and pmi3) and hydrophobic surface descriptors (vsurf_D1; a VolSurf-like molecular interaction field 3D molecular descriptor

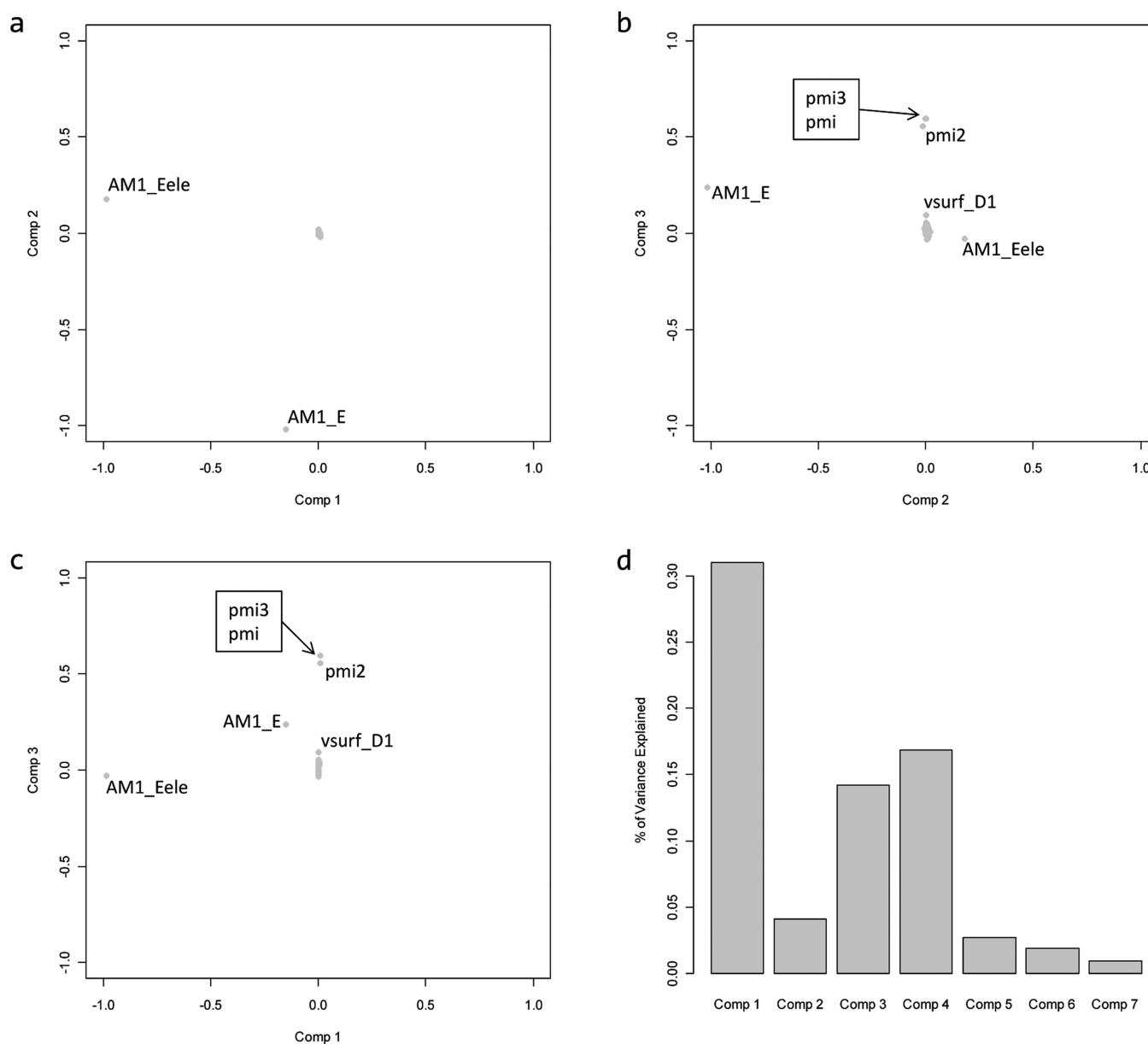


Figure 3. The PLS loadings for the first three components of the PLS preliminary *T. pyriformis* data set model. (a) Component 1 versus Component 2. (b) Component 2 versus Component 3. (c) Component 1 versus Component 3. (d) Percent variance explained by each PLS component. Important descriptors are labeled.

that describes the hydrophobic volume around a molecule) are indicated as molecular descriptors that are strongly correlated with the biological activity of the compounds (Figure 3b and c). These results may imply that *T. pyriformis* toxicity of these compounds is related to electrostatic energy, the ability of the

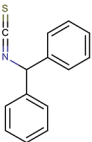
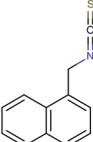


Table 4. Six Descriptors Distinguished by Loadings of the 3 PLS Components of the *T. pyriformis* Data Set

symbols	description
AM1_E	total energy applying AM1 semi-empirical parametrization
AM1_Eele	electrostatic energy applying AM1 semi-empirical parametrization
pmi	principal movement of inertia
pmi2	second component of principal movement of inertia
pmi3	third component of principal movement of inertia
vsurf_D1	descriptor related to the hydrophobic surface volume

compound to rotate around an axis, and the hydrophobic nature of the compound.

To investigate the relationship between the biological end points (pIGC₅₀) and the six most influential descriptors, the two most and least active compounds and their associated descriptors of interest were selected for analysis, Table 5. The descriptor values of the toxic compounds (most active) are between four- and eight-times greater in numerical value compared to the nontoxic compounds (least active). The values for the pmi-related descriptors of active compounds are approximately 100-times greater than for the inactive compounds. Although the molecular similarity and molecular size are significantly different between the most and least toxic compounds displayed in Table 5, the existence of a relationship between the activity and these descriptors is still apparent from the constructed predictive models.

Table 5. *T. pyriformis* Compounds with Highest and Lowest pIGC₅₀ Values (Most and Least Active) in the Training Set and Their Corresponding Descriptor Values for the 6-Term Model

Structure				
pIGC ₅₀	3.05	3.05	-2.67	-2.45
AM1_Eele	-324998	-264926	-24674.6	-59591.1
AM1_E	-54042.2	-47511.4	-11622.6	-19679.6
pmi	2194.41	1938.504	22.749	135.387
pmi2	1366.247	1523.165	20.32429	74.91601
pmi3	2009.849	1913.504	21.13052	123.1368
vsurf_D1	822.75	764.25	97.625	172.625

We built a SVR regression model using these six molecular descriptors (Figure 4), and the coefficient of determination

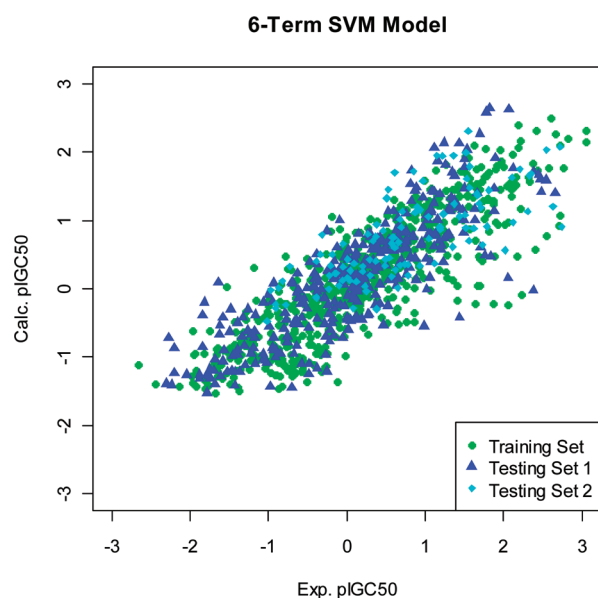


Figure 4. The predicted pIGC₅₀ value (X-axis), compared with experimental pIGC₅₀ value (Y-axis), of the SVR Model using descriptors selected by the loadings of the first three components of the PLS model constructed from the *T. pyriformis* data set. The training set and the test sets 1 and 2 are represented respectively as green circles, blue triangles, and cyan diamonds.

(R^2) for the training set and two test sets were 0.731, 0.695, and 0.552, respectively (Table 6). A reason for the low R^2 values might be the insufficient amount of information contained within the initial six descriptors, thus we selected more descriptors by considering the significant descriptors in the first-seven loading matrices, PLS components 1 through 7.

The most significant descriptors from the first-seven PLS components were selected using the maximum-absolute loading value as a method to rank each descriptor, and the thresholds of the loading values were set to 0.01 and 0.001, respectively. An absolute loading threshold value greater than 0.01 resulted in the identification of 102 descriptors, while an absolute loading

Table 6. Leave-One-Out R^2 Validation of SVM Models with Different Number of Terms Selected by the Preliminary PLS Model for the *T. pyriformis* Data Set

	training set	testing set #1	testing set #2
6 terms	0.731	0.695	0.552
102 terms max(abs(loadings)) > 0.01	0.912	0.817	0.613
204 terms max(abs(loadings)) > 0.001	0.928	0.832	0.620

value of 0.001 resulted in 204 descriptors being highlighted. The results of these two SVR models are shown in Figure 5 and Table 6, compared with the original 6-term SVR model. Increasing the number of terms in the SVR model from six to 102 increased the R^2 of the training set from 0.731 to 0.912 and

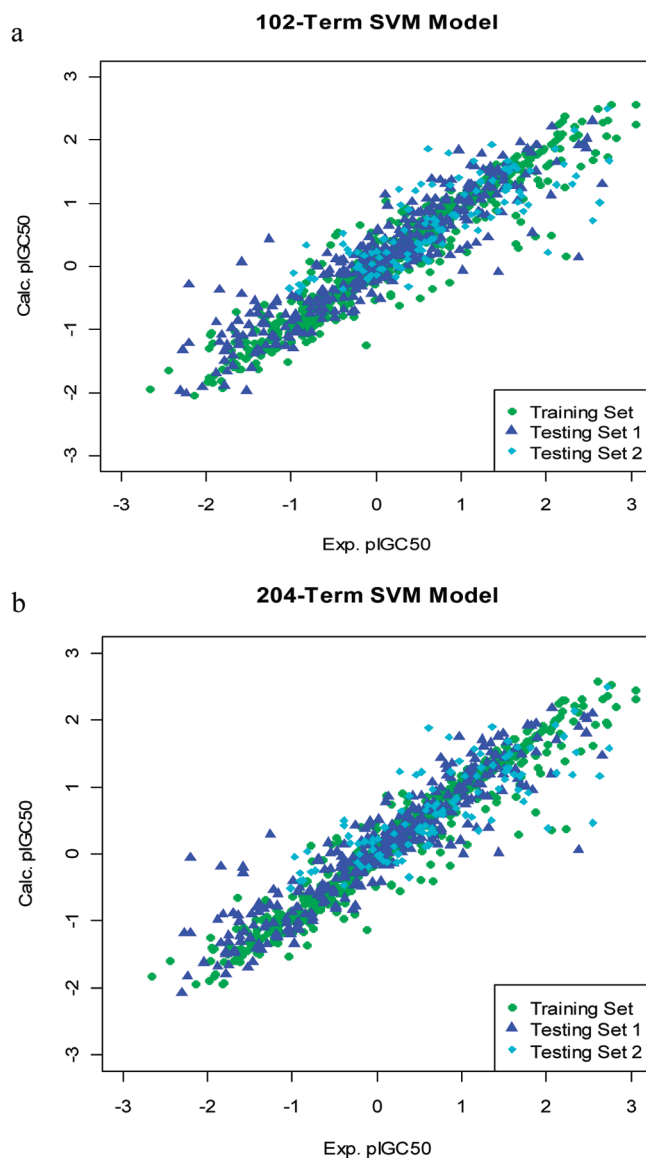


Figure 5. The predicted pIGC₅₀ value (X-axis) of the *T. pyriformis* data set, compared with experimental pIGC₅₀ value (Y-axis), of the SVR Models of (a) the 102-term model (maximum abs(loadings) > 0.01) and (b) the 204-term model (maximum abs(loadings) > 0.001). The training set and the test sets 1 and 2 are represented respectively as green circles, blue triangles, and cyan diamonds.

resulted in test sets 1's and 2's R^2 increasing from 0.695 to 0.817 and 0.552 to 0.613, respectively. Increasing the number of descriptors available to 204, and thus constructing a 204-term SVR model, increased the R^2 of the training set and test sets 1 and 2 to 0.924, 0.832, and 0.620, respectively. The test set validation results for the 204-term SVR model is approximate to the models reported by Huang and Fan¹ and Zhu and co-workers.³⁵

To determine if the prediction performance would improve with additional descriptors, SVR models were built by adding descriptors based on the order of their maximum absolute PLS loading values (score). In an iterative fashion, a SVR model was built with the most important molecular descriptor followed by the creation of another SVR model built with the two most important descriptors. This process was continued until all 719 molecular descriptors were used to construct a SVR model. The coefficient of determination (R^2) is plotted in Figure 6 for the

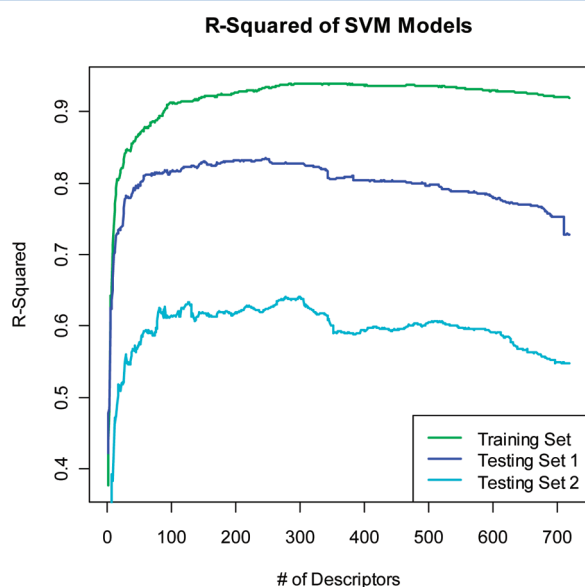


Figure 6. The R^2 of SVM models built by descriptors added to the model by sorting of the maximum of absolute value of loadings of the *T. pyriformis* data set. The training set and the test sets 1 and 2 are represented respectively as green, blue, and cyan lines.

training and test sets. As illustrated in Figure 6, the R^2 values of test sets 1 and 2 start to decrease once the SVR model is constructed with more than approximately 300 of the most important molecular descriptors – based on PLS loading values. A reason the predictive ability of the models' decreased with each additional descriptor (for models with greater than approximately 300 molecular descriptors) for each of the test sets is most likely that the models are exhibiting signs of overfitting. Therefore, the 204-term SVR model was selected for the following outlier removal experiments because of its predictive abilities for the training and test sets.

Outlier Removal Using "Distance to Model" Methods. Tetko et al. suggested that within a QSAR model the distance between the descriptor values of the training set and the test set, or the compound(s) of interest, influences the error in the predicted activity.³⁶ The descriptor space – the numerical range that each descriptor in the model (training set) covers – directly impacts how well the model will predict compounds that were not used to construct the model, commonly referred to as the "Applicability Domain" of a model. In the studies

relating to the *T. pyriformis* data set,^{1,35,36} the R^2 values of test set 2 were below 0.70. Typically, R^2 values below this value indicate that a QSAR model was not very adept at predicting the bioactivity for a series of compounds, and, in this example, the model is not able to adequately predict the experimental bioactivity values for test set 2 of the *T. pyriformis* data set. A possible reason for the low predictive performance of the models when applied to test set 2 is that there were some compounds outside of the "Applicability Domain" as defined by the training set compounds. To explore the applicability domain of our 204-term SVR model, the descriptor distance between each test set compounds and the training set was calculated; this is commonly referred to as the "Distance to Model".

We compared the "Distance to Model" to the prediction error (Figure 7) and calculated the Pearson's correlation coefficient (R ; Table 7) between the two values. The training set's correlation coefficient between the error of prediction and the "Distance to Model" was 0.100, while the correlations coefficients for test sets 1 and 2 were 0.273 and 0.215, respectively. A test set compound might only need a small portion of the compounds in the training set to adequately describe its activity. To explore this concept the distance between the nearest neighbor distances were explored. Accordingly, the pairwise distances between all of the training set and test set compounds were calculated to determine the 3-nearest-neighbor (3-NN) distances. Using the 3-NN method, the training set R increased to 0.118, and the correlation coefficients of the two test sets increased to 0.368 and 0.316, respectively. Reducing the number of neighboring compounds from "all compounds" to "three compounds" for the correlation between the error of prediction and the "Distance to Model" evaluation resulted in a "stronger" correlation. These results can be interpreted in several ways; the first is that more than three compounds are needed to sufficiently predict the end point for the compounds in the training and test sets, the second is that the prediction error cannot be estimated by calculating the 3-NN distance, and the third is that it can be determined if a test set compound fits the model and thus can aid in removing outliers.

Using the 3-NN criteria to filter outliers from the test set was explored by restricting the "distance" between the test set compound and its three nearest-neighbors of the training set; the 3-NN distance was bound to less than or equal to 20, 15, and 10. Investigating the effects of outlier removal was focused on test set 2 due to the poor performance of the QSAR model and to provide insight regarding the removal of test set compounds that are not similar to training set compounds (outliers). The R^2 (predicted versus experimentally determine end points) for the complete version of test set 2 was 0.620 and is not as impressive as test set 1's R^2 of 0.832. Using the 3-NN distance as a filter, test set 2 compounds with a distance greater than 20, 15, and 10 were removed, and the R^2 was recalculated for each remaining allotment of test set 2 compounds. Including only test set 2 compounds with a "Distance to Model" less than 20 from the training set slightly improved the R^2 to 0.621. Reducing the distance threshold to 15 increases the R^2 value to 0.675, while further reducing the threshold to 10 caused the R^2 value to decrease to 0.645. The decrease in the R^2 value when reducing the threshold value from 15 to 10 might be due to compounds that fit within the applicability domain – and helps to explain the important physicochemical properties of the test set 2 compounds – were filtered out. While the R^2

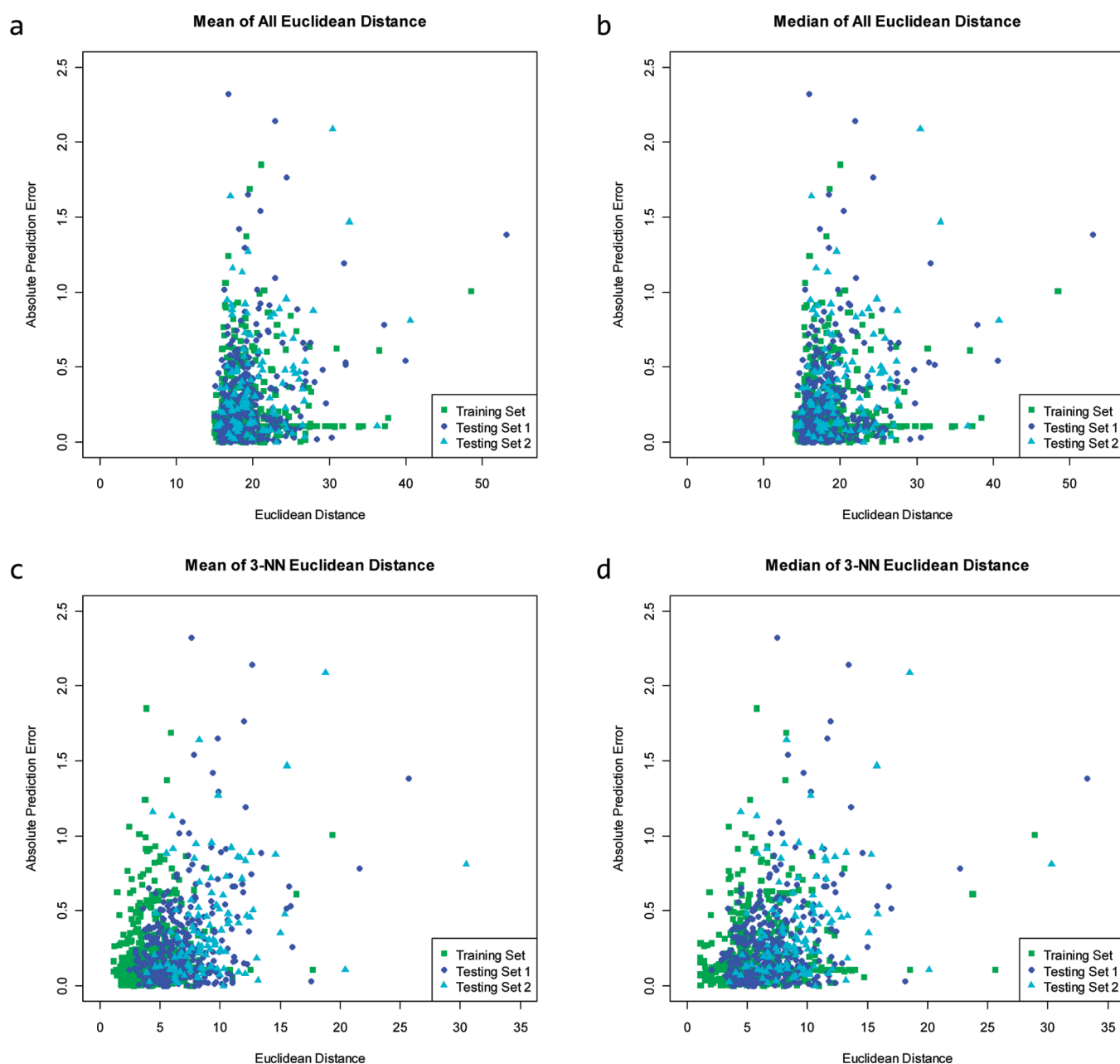


Figure 7. The distance to model calculations of the *T. pyriformis* data set using the mean and median of Euclidean distances between the test set and training set compounds, compared with the prediction error. (a) The mean distance to all training set compounds for each compound. (b) The median distance to all training set compounds for each compound. (c) The mean distance to the 3-nearest training set compounds for each compound. (d) The median distance to the 3-nearest training set compounds for each compound.

value for test set 2 only increases by 0.055 (0.620 to 0.675) when removing possible outliers, comparing the complete and filtered versions of test set 2 using the nearest-neighbor method of filtering compounds (“Distance to Model” threshold value of

15) is a sound method to determine if a test set compound fits within the domain of the model. The calculated R^2 values after filtering are presented in Table 8.

Comparison of Model Building Methods. For the *T. pyriformis* data set, the primary difference between the model construction methods in this study and those applied in Huang and Fan’s study¹ relates to the descriptor selection method.

Table 7. Pearson Correlation Coefficient Between “Distance to Model” and the Prediction Error of the *T. Pyriformis* Data Set

		training set	test set #1	test set #2
all	mean	0.100	0.273	0.215
	median	0.077	0.255	0.208
3-NN	mean	0.118	0.368	0.316
	median	0.117	0.378	0.308

Table 8. Validation Result of R^2 of Outlier Removal of the Test Set 2 of the *T. pyriformis* Data Set

	training set	test set #1	test set #2	test set #2 (D < 20)	test set #2 (D < 15)	test set #2 (D < 10)
R^2	0.928	0.832	0.620	0.621	0.675	0.645

The utilization of a genetic algorithm (GA)²⁵ is a familiar practice for descriptor selection and the construction of an ensemble of QSAR models. Where the GA protocol of Huang and Fan fails is the duration of the GA's evolutionary growth.¹ It is common to terminate a GA after several hundred generations (a thousand generations is considered standard) without a change to the surviving population (best models). The GA parameters, as described by Huang and Fan, constrained the number of descriptors in each model to between three and ten molecular descriptors – which is acceptable – but the low number of models within the population (50 models) and the extremely short evolutionary period (30 generations) did not allow the GA to adequately explore the descriptor space. The short period of model creation *might* highlight a few influential molecular descriptors, but it is definitely not enough evolutionary time to construct a sufficiently sound QSAR ensemble and reveal the truly meaningful descriptors. An efficient and sound protocol to determine the most relevant molecular descriptors is the analysis of a preliminary PLS model. Our final SVR model was built from approximately 300 molecular descriptors that were ranked using the PLS loading values; this protocol resulted in a suitable SVR model that did not exhibit overfitting.

■ DESIGNING THE PROTOCOLS TO IMPROVE THE MODELS

The protocols used to assemble the training set along with the selection of descriptors and the construction of the model have the most influence over the models' robustness and predictive ability. For the hERG data set, the training and test sets were comprised of compounds that passed the Lipinski's Rule-of-Five⁵⁷ and were within a specific relative lipophilicity (octanol/water partition coefficient, logP, values) range (training set compounds: active compounds with a calculated logP value *less than* 4.1 and inactive compounds with a logP value *greater than* 2.8).³⁷ Additionally, the hERG descriptor trial pool included a reduced set of the 4D-FP descriptors based on PLS loading values to ensure the SVM model's hyperplanes contained (molecular) information that was strongly relevant to the biological end points. These key features of the hERG data set protocols resulted in a SVM model with an improved predictive nature for the training set and, most importantly, the test set.

The *T. pyriformis* data set was defined by Zhu et al.,³⁵ thus the protocol to improve the predictive ability of the models was focused on the manner of selecting molecular descriptors for the SVR model and developing a better understanding of the relationship between the compounds in the training set and test sets. While it can be argued that the models presented herein are marginally better than Zhu and co-workers,³⁵ the overall knowledge gained by including a mixed-class and biologically relevant set of molecular descriptors provides a better molecular understanding of the overall toxicology for the data set. As noted above, the GA search for collections of important molecular descriptors in the Huang and Fan¹ study did not have adequate evolutionary periods, and thus the most important descriptors were not passed on to the SVR models. Instead of using a GA to select the descriptors, the work presented here used the PLS loadings to determine the set of descriptors that relate to the biological end points.

The two data sets – hERG blockage (a discrete data set) and *T. pyriformis* (a continuous data set) – were discussed and explored because they are commonly considered difficult with respect to constructing practical predictive models. Each data

set presented different hurdles to overcome and provided the opportunity to demonstrate the protocols and methodologies that are commonly used in building sound predictive models for discrete and continuous data sets. For example, selecting descriptors for the hERG data set required the use of the F-score, yet this methodology cannot be applied to continuous data sets like the *T. pyriformis* data set, while the PLS loading score method used to distinguish important molecular descriptors in the *T. pyriformis* data set is not suitable for discrete data sets. However, like all experiments, the protocols and methodologies must be adapted to suit the data set of interest.

■ CONCLUSION

The ability to construct, evaluate, and apply a QSAR model using contemporary methodologies and protocols is not a trivial process and requires attention to detail. The data sets, hERG and *T. pyriformis*, explored in this work have been extensively investigated, and each study provides additional information about the overall molecular system and/or insights to the modeling protocols and methods. The focus of this study was to provide insight to robust protocols, demonstrate current QSAR methodologies, and the implementation of these protocols and methodology to notoriously difficult data sets. Using various statistical methods, sound predictive models were constructed and analyzed. Selecting significant molecular descriptors using PLS loadings values to construct a SVR model reduced the number of weightless molecular descriptors. Harnessing the power of biologically relevant molecular descriptors across multiple descriptor class improves the overall interpretability of the models,⁶⁰ while the combination of sound protocols and methodologies helps to showcase the possibilities of QSAR modeling.

Overall, the protocols used for the hERG and *T. pyriformis* data sets have common threads that can be applied to any predictive modeling study: (i) construct the trial descriptor pool from relevant descriptors, (ii) remove the noise from the training set (whether this be molecules or other entities), (iii) select descriptors that are strongly related (correlated) to the end points of interest to construct the models, and (iv) analyze and understand the results for the training and test sets. While these protocols are somewhat apparent and universal, the manner they are applied can vary from system to system, and discovering the optimal combination is key to robust and insightful predictive models.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +886.2.3366.4888 #529. Fax: +886.2.23628167. E-mail: yjtseng@csie.ntu.edu.tw.

Author Contributions

#Equal contribution to this work.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS:

T. pyriformis = *Tetrahymena pyriformis*

hERG = Ether-a-go-go Related Gene

QSAR = Quantitative Structure–Activity Relationship

SVM = Support Vector Machine

SVR = Support Vector Regression

GA = Genetic Algorithm

MOE = Molecular Operating Environment
4D-FPs = 4D-Fingerprints
TdP = Torsades de Pointes
PLS = partial least-squares
ROC = receiver operating characteristic
GF = A genetic function approximation
kNN = k-nearest neighbors
MLR = machine-learned ranking
OLS = ordinary least-squares
ASNN = associative neural network
ANN = artificial neural network
 R^2 = coefficient of determination
LOO = leave-one-out
 Q^2 = leave-one-out cross-correlation
MD = molecular dynamic
IPEs = interaction pharmacophore elements
HBA = hydrogen bond acceptor
PP = polar-positive
PN = polar-negative

REFERENCES

- (1) Huang, J.; Fan, X. Why QSAR Fails: An Empirical Evaluation Using Conventional Computational Approach. *Mol. Pharm.* **2011**, *8* (2), 600–608.
- (2) Holland, J. H. *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan, Ann Arbor, MI, 1975.
- (3) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. *Adv. Neural Inf. Process. Syst.* **1997**, 155–161.
- (4) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.
- (5) Vapnik, V. N. *The Nature of Statistical Learning Theory*; 2000.
- (6) Brown, A. M. Drugs, hERG and Sudden Death. *Cell. Physiol. Biochem.* **2004**, *35*, 543–547.
- (7) 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*, 1829–1835.
- (8) 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.
- (9) Sanguinetti, M. C.; Jiang, C.; Curran, M. E.; Keating, M. T. A Mechanistic Link Between an Inherited and an Acquired Cardiac Arrhythmia: hERG Encodes the IKr Potassium Channel. *Cell* **1995**, *81* (2), 299–307.
- (10) Aptula, A.; Cronin, M. Prediction of hERG K^+ Blocking Potency: Application of Structural Knowledge. *SAR QSAR Environ. Res.* **2004**, *15* (5–6), 399–411.
- (11) Cianchetta, G.; Li, Y.; Kang, J.; Rampe, D.; Fravolini, A.; Cruciani, G.; Vaz, R. Predictive Models for hERG Potassium Channel Blockers. *Bioorg. Med. Chem. Lett.* **2005**, *15* (15), 3637–3642.
- (12) Coi, A.; Massarelli, I.; Murgia, L.; Saraceno, M.; Calderone, V.; Bianucci, A. Prediction of hERG Potassium Channel Affinity by the CODESSA Approach. *Bioorg. Med. Chem.* **2006**, *14* (9), 3153–3159.
- (13) Obrezanova, O.; Csanyi, G.; Gola, J. M. R.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, *47* (5), 1847–1857.
- (14) Chen, X.; Li, H.; Yap, C.; Ung, C.; Jiang, L.; Cao, Z.; Li, Y.; Chen, Y. Computer Prediction of Cardiovascular and Hematological Agents by Statistical Learning Methods. *Cardiovasc. Hematol. Agents Med. Chem.* **2007**, *5* (1), 11–19.
- (15) 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.
- (16) Sun, H. An Accurate and Interpretable Bayesian Classification Model for Prediction of hERG Liability. *ChemMedChem* **2006**, *1* (3), 315–322.
- (17) Gepp, M.; Hutter, M. Determination of hERG Channel Blockers Using a Decision Tree. *Bioorg. Med. Chem.* **2006**, *14* (15), 5325–5332.
- (18) Jia, L.; Sun, H. Support Vector Machines Classification of hERG Liabilities Based on Atom Types. *Bioorg. Med. Chem.* **2008**, *16* (11), 6252–6260.
- (19) Leong, M. A Novel Approach Using Pharmacophore Ensemble/Support Vector Machine (PhE/SVM) for Prediction of hERG Liability. *Chem. Res. Toxicol.* **2007**, *20* (2), 217–216.
- (20) Song, M.; Clark, M. Development and Evaluation of an *in silico* model for hERG Binding. *J. Chem. Inf. Model.* **2006**, *46*, 392–400.
- (21) 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.
- (22) Keseru, G. M. Prediction of hERG Potassium Channel Affinity by Traditional and Hologram qSAR Methods. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773–5.
- (23) 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.
- (24) National Center for Biotechnology Information, hERG Channel Activity (AID: 376, Source: PDSP). In *The PubChem BioAssay Database*; National Center for Biotechnology Information: Bethesda, Maryland, USA, 2009.
- (25) Hopfinger, A. J.; Patel, H. C. Application of Genetic Algorithms to the General QSAR Problem and to Guiding Molecular Diversity Experiments. In *Genetic algorithms in molecular modeling*; Devillers, J., Ed.; Academic Press: 1996; pp 131–157.
- (26) Meffert, K.; Meseguer, J.; Martí, E. D.; Meskauskas, A.; Vos, J.; Rotstan, N.; Knowles, C.; Sangiorgi, U. B. JGAP - Java Genetic Algorithms and Genetic Programming Package.
- (27) Schultz, T. W. Structure–Toxicity Relationships for Benzenes Evaluated with *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **1999**, *12* (12), 1262–1267.
- (28) Schultz, T. W. TETRATOX: *Tetrahymena pyriformis* Population Growth Impairment Endpoint-A Surrogate for Fish Lethality. *Toxicol. Mech. Methods* **1997**, *7* (4), 289–309.
- (29) The TETRATOX Database. <http://www.vet.utk.edu/TETRATOX/index.php>.
- (30) Cronin, M. T. D.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Wayne Schultz, T. Comparative Assessment of Methods to Develop QSARs for the Prediction of the Toxicity of Phenols to *Tetrahymena pyriformis*. *Chemosphere* **2002**, *49* (10), 1201–1221.
- (31) Cronin, M. T. D.; Gregory, B. W.; Schultz, T. W. Quantitative Structure–Activity Analyses of Nitrobenzene Toxicity to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **1998**, *11* (8), 902–908.
- (32) Cronin, M. T. D.; Schultz, T. W. Development of Quantitative Structure–Activity Relationships for the Toxicity of Aromatic Compounds to *Tetrahymena pyriformis*: Comparative Assessment of the Methodologies. *Chem. Res. Toxicol.* **2001**, *14* (9), 1284–1295.
- (33) Dearden, J. C.; Cronin, M. T. D.; Schultz, T. W.; Lin, D. T. QSAR Study of the Toxicity of Nitrobenzenes to *Tetrahymena pyriformis*. *Quant. Struct.-Act. Relat.* **1995**, *14* (5), 427–432.
- (34) Schultz, T. W.; Hewitt, M.; Netzeva, T. I.; Cronin, M. T. D. Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action. *QSAR Comb. Sci.* **2007**, *26* (2), 238–254.
- (35) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48* (4), 766–784.

- (36) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against *Tetrahymena pyriformis*: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48* (9), 1733–1746.
- (37) Su, B.-H.; Shen, M.-Y.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. *In silico* Binary Classification QSAR Models Based on 4D-Fingerprints and MOE Descriptors for Prediction of hERG Blockage. *J. Chem. Inf. Model.* **2010**, *50* (7), 1304–1318.
- (38) 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.
- (39) 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.
- (40) Dubus, E.; Ijjaali, L.; Petitet, F.; Michel, A. *In silico* Classification of hERG Channel Blockers: a Knowledge-Based Strategy. *Chem-MedChem* **2006**, *1* (6), 622–630.
- (41) Nisius, B.; Goller, A. H. Similarity-Based Classifier Using Topomers to Provide a Knowledge Based for hERG Channel Inhibition. *J. Chem. Inf. Model.* **2008**, *49*, 247–256.
- (42) 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.
- (43) Shen, M.-y.; Su, B.-H.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. A Comprehensive Support Vector Machine Binary hERG Classification Model Based on Extensive but Biased End Point hERG Data Sets. *Chem. Res. Toxicol.* **2011**, *24* (6), 934–949.
- (44) Hypercube *Hyperchem Release 7.0*; 2008.
- (45) Allinger, N. L. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms. *J. Am. Chem. Soc.* **1977**, *99* (25), 8127–8134.
- (46) Chemical Computing Group Inc. *MOE (Molecular Operating Environment)*, 2010.10; Montreal, Canada, 2010.
- (47) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Model.* **2004**, *44* (5), 1526–1539.
- (48) Cruciani, G.; Crivori, P.; Carrupt, P.; Testa, B. Molecular Fields in Quantitative Structure–Permeation Relationships: the VolSurf Approach. *J. Mol. Struct. (Theochem)* **2000**, *503* (1–2), 17–30.
- (49) 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.
- (50) Iyer, M.; Hopfinger, A. J. Treating Chemical Diversity in QSAR Analysis: Modeling Diverse HIV-1 Integrase Inhibitors Using 4D Fingerprints. *J. Chem. Inf. Model.* **2007**, *47* (5), 1945–1960.
- (51) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1–27.
- (52) R Development Core Team, *R: A Language and Environment for Statistical Computing*; Vienna, Austria, 2011.
- (53) Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, a. A. *e1071: Misc Functions of the Department of Statistics (e1071)*, TU Wien; 2011.
- (54) Wehrens, R.; Mevik, B.-H. *pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*; 2007.
- (55) Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20* (1), 37–46.
- (56) Akay, M. F. Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis. *Expert Syst. Appl.* **2009**, *36*, 3240–3247.
- (57) 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.
- (58) 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.
- (59) Yoshida, K.; Niwa, T. Quantitative structure-activity relationship studies on inhibition of hERG potassium channels. *J. Chem. Inf. Model.* **2006**, *46*, 1371–8.
- (60) Tseng, Y. J.; Hopfinger, A. J.; Esposito, E. X. The Great Descriptor Melting Pot: Mixing Descriptors for the Common Good of QSAR Models. *J. Comput.-Aided Mol. Des.* **2012**, *26* (1), 39–43.