

In Silico Classification of hERG Channel Blockers: a Knowledge-Based Strategy

Elodie Dubus,* Ismail Ijjaali, François Petitet, and André Michel^[a]

The blockage of the hERG potassium channel by a wide number of diverse compounds has become a major pharmacological safety concern as it can lead to sudden cardiac death. In silico models can be potent tools to screen out potential hERG blockers as early as possible during the drug-discovery process. In this study, predictive models developed using the recursive partitioning method and created using diverse datasets from 203 molecules tested on the hERG channel are described. The first model was built with hERG compounds grouped into two classes, with a separation limit set at an IC_{50} value of $1 \mu\text{M}$, and reaches an

overall accuracy of 81%. The misclassification of molecules having a range of activity between 1 and $10 \mu\text{M}$ led to the generation of a tri-class model able to correctly classify high, moderate, and weak hERG blockers with an overall accuracy of 90%. Another model, constructed with the high and weak hERG-blocker categories, successfully increases the accuracy to 96%. The results reported herein indicate that a combination of precise, knowledge management resources and powerful modeling tools are invaluable to assessing potential cardiotoxic side effects related to hERG blockage.

Introduction

Prolonged QT intervals, which correspond to the action potential duration of the surface electrocardiogram (ECG), can lead to ventricular tachyarrhythmias generally referred to as *torsades de pointes* (TdP), that can degenerate into ventricular fibrillation and sudden death.^[1] Consequently, several non-antiarrhythmic drugs such as sertindole (antipsychotic), terfenadine (antihistaminic), and cisapride (prokinetic agent) have recently received "black box" warnings or have been withdrawn from the market following reports of QT interval prolongation. Moreover, the number of reports of TdP and QT interval prolongation continues to increase, leading to a major pharmacological safety concern for the pharmaceutical industry and health regulatory agencies.^[2] Novel chemical compounds with this cardiotoxic side effect should therefore be identified as early as possible in the drug-discovery process to limit cost and save time in development.


The hERG K^+ channel ($K_{v11.1}$), which gives rise to the rapid component of the delayed rectifier K^+ -channel current (I_{Kr}), is involved in normal cardiac repolarization and has become the focus of many studies, as most of the QT-prolonging drugs have been shown to inhibit I_{Kr} .^[2,3] In addition, other drug characteristics can cause QT interval prolongation such as bioavailability at the target organ, metabolic issues, or the involvement of other ion channels.^[4] Although diverse functional in vitro assays have been developed to identify potential hERG inhibitors, patch-clamp electrophysiology is the most highly sensitive technique available but it is throughput-limited and expensive.

The possibility of using in silico screening to predict hERG affinity in the early phase of the drug-discovery process would provide cost-effective screening tools that can be used in association with other in vitro assays. Recently, several predictive models have been proposed, and some have been successful

in gaining insight into the molecular basis of hERG channel blockage, and in the prediction of the QT-prolonging potential for a large number of compounds.^[2,3,5] Computational approaches include both structure- and ligand-based methods.^[6–14] The latter methods include traditional and hologram QSAR^[11] studies, pharmacophore modeling,^[12] and neural network systems.^[13] Moreover, predictive models for hERG blockers based on the support vector machine method^[14] using 73 drugs from the literature have been described recently. Overall accuracies of 90 and 95% have been achieved according to a cutoff for separating hERG-active and -inactive compounds of 1 and $40 \mu\text{M}$, respectively. Herein, using a large validated dataset (203 molecules), we propose a novel in silico model based upon the recursive partitioning (RP) method to rapidly screen out potential hERG blockers. This classification technique is based on a decision tree algorithm, which divides compounds into a hierarchy of smaller and more homogeneous subgroups using the statistically most significant descriptors. More precisely, RP involves the creation of a decision tree composed of binary split nodes that divide the initial training set into smaller sets of higher purity, such as sets containing a majority of blockers or a majority of nonblockers as reported herein. Each split node can be compared to a binary question (yes or no) regarding the value of a particular molecular descriptor. The model can be used to classify any other new compound for which the descriptors used in the split nodes have been com-

[a] Dr. E. Dubus,[†] Dr. I. Ijjaali,[†] Dr. F. Petitet, Prof. A. Michel
Aureus Pharma, 174 quai de Jemmapes, 75010 Paris, France
Fax: (+33) 1-4018-5758
E-mail: elodie.dubus@aureus-pharma.com

[[†]] These authors contributed equally to this work.

 Supporting Information for this article is available on the WWW under <http://www.chemmedchem.org> or from the author.

puted. This rapid computational method has shown efficiency in a variety of classification exercises and with other statistical techniques.^[15–17] A decision tree model^[18] was previously reported based only on three calculated physicochemical descriptors to distinguish hERG binding. Equipped with a larger amount of data and diverse two-dimensional (2D) molecular descriptors, we report herein efficient predictive models for classifying hERG channel blockers with a higher degree of accuracy than previously described.

Results

Selection of the data sets

To generate RP models, 194 drugs were separated into two classes according to their activities on the hERG channel, high and weak blockage classes, with the cutoff being 1 μM . Furthermore, to increase the weak-blocker set, nine compounds which inhibited hERG channel activity by less than 20% at concentrations between 10 μM and 50 μM were added. In total there were 96 and 107 molecules in the high and weak class, respectively (Table 1, Model 1). The second model was based

	Model 1		Model 2	Model 3	
	Training	Test	Training	Training	Test
High ($\text{IC}_{50} \leq 1 \mu\text{M}$)	80	16	96	50	46
Moderate ($1 \mu\text{M} < \text{IC}_{50} < 10 \mu\text{M}$)			48		
Weak ($\text{IC}_{50} \geq 10 \mu\text{M}$ %Inh < 20%)	80	27	59	50	9

on a multi-class approach by dividing the dataset into three classes of high, moderate, and weak blockers, which relate to $\text{IC}_{50} \leq 1 \mu\text{M}$ (96 molecules), IC_{50} between 1 and 10 μM (48 molecules), and $\text{IC}_{50} \geq 10 \mu\text{M}$ (59 molecules), respectively (Table 1, Model 2). Finally, we selected the extreme classes for the third model, that is, high and weak hERG blockers (Table 1, Model 3). It should be noted that for models 1 and 3, sufficient numbers of molecules in each category (>50) allowed us to split the compounds in each class into a training set and a test set. The training set allowed us to generate the model with an internal validation, assessed by a cross-validation approach explained in the Computational Methods section. The test set consisted of molecules not present during any phase of model development, and therefore represented an external prediction set. Concerning Model 2, only the internal validation was assessed.

Chemical diversity of data sets

The chemical diversity of the three classes was assessed using nearest-neighbor searching algorithm as implemented in

ChemAxon application Compr.^[20] Table 2 summarizes estimations of the average self-dissimilarity and inter-classes dissimilarity. The weak class showed the highest self-dissimilarity

Table 2. Average dissimilarity statistics between the three classes based on the nearest-neighbor searching algorithm (ChemAxon application Compr).

Class	High	Moderate	Low
High	57.2%		
Moderate	62.9%	63.4%	
Low	66.4%	67.0%	66.8%

(66.8%). The corresponding values for moderate and high classes (63.4 and 57.2%, respectively) still indicated a reasonable chemical diversity within these classes. Inter-class dissimilarities range from 62.9 to 67.0%, which suggests that these datasets were sufficiently diverse. In Figure 1, the whole hERG dataset of 203 molecules is widely spread in the plane defined by the two main Principal Component Analysis (PCA) axes calculated from the 184 molecular descriptors. The first two principal components explain 41.2% of the variance. The screening collection of the Specs^[25] database was used to illustrate the chemical diversity of the hERG dataset.

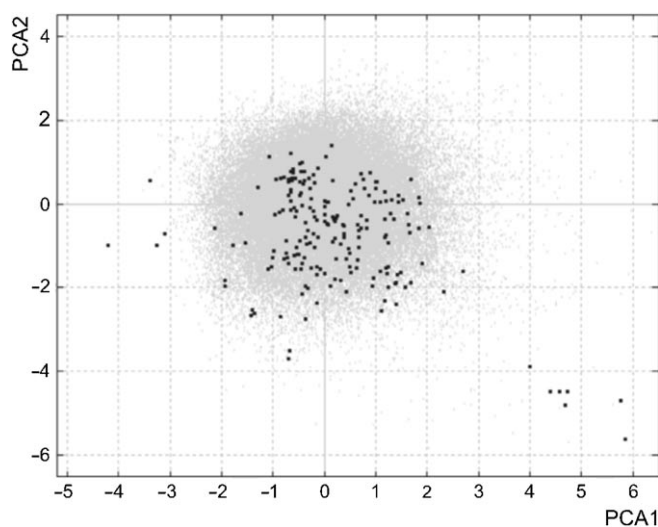


Figure 1. Representation of hERG compounds (203) within the chemical space of the Specs database (December 2005, 77438 compounds) according to the first two PCA axes computed using 184 2D molecular descriptors. Grey and black dots correspond to the Specs database and hERG compounds, respectively.

Model 1: cutoff of 1 μM

A support vector machine model and a QSAR approach have set the active/inactive boundary for hERG compounds at $\text{IC}_{50} = 1 \mu\text{M}$.^[11,14] To evaluate our approach, RP models have been generated using the same separation value. The RP method is known to be sensitive to unbalanced training sets for constructing models. Therefore, an equal number of compounds

(80 per class) has been chosen to create the training set. In addition, the molecules have been selected using Diverse Subset as described in the methods section, to obtain the 80 most various compounds in each class. The aim of this approach, was to obtain the most global model with a large coverage of the chemical space. To limit the over-fitting issue that can result by using a pool with too many descriptors, selection of the most relevant descriptors involved in decision tree-based models among all 184 2D descriptors was performed using CFS algorithm (Table 3). 23 uncorrelated relevant descriptors were identified which led to the best performance. Interestingly, several selected descriptors belong to the subdivided surface areas P_VSA parameter type, which express structural information based on molecular surfaces. Therefore, two predictive models were constructed using either the set of 23 relevant descriptors or the set of 32 P_VSA descriptors. Considering the training set, both models built using relevant or P_VSA descriptors have high classification ability, with 96 and 97.5% accuracy respectively (Table 4, Model 1). The overall classification accuracy of the models decreased to 74 and 81% when

the test set was applied to validate the models (Table 5, Model 1). Although 94% of high blockers were correctly predicted, the models had difficulty in classifying weak hERG blockers, with a precision of 63 and 74% for relevant and P_VSA descriptors, respectively. Among the ten false negatives misclassified in the model based on relevant descriptors, nine molecules had an IC₅₀ value between 1 and 10 μM. Similarly, five compounds of the seven false negatives in the P_VSA model possessed IC₅₀ values between 1 and 10 μM.

Model 2: tri-class models

Based on the observations of misclassification with molecules possessing biological activities between 1 and 10 μM, a new RP model was generated using three classes of compounds according to their IC₅₀ values, measured on the hERG channel (Table 1, Model 2). The former weak class was split creating an intermediate category composed of molecules having an IC₅₀ value between 1 and 10 μM (moderate blockage). The new weak hERG blockers possessed IC₅₀ ≥ 10 μM. Roche et al.^[13]

Table 3. List of most relevant molecular descriptors used in the decision-tree-based models, selected by the Correlation-based Feature Selection algorithm.

Descriptor	Definition	Type	Models
PEOE_VSA + 5	Sum of v_i where PEOE- q_i is in the range [0.25,0.30].	Subdivided surface areas	1, 2
PEOE_VSA - 5	Sum of v_i where PEOE- q_i is in the range [-0.30,-0.25].	Subdivided surface areas	1, 2, 3
SlogP_VSA2	Sum of v_i such that SlogP is in (-0.2,0].	Subdivided surface areas	1, 2, 3
SlogP_VSA7	Sum of v_i such that SlogP is in (0.25,0.30].	Subdivided surface areas	1, 2, 3
SlogP_VSA8	Sum of v_i such that SlogP is in (0.30,0.40].	Subdivided surface areas	2, 3
SMR_VSA1	Sum of v_i such that SMR is in (0.11,0.26].	Subdivided surface areas	1, 2, 3
SMR_VSA2	Sum of v_i such that SMR is in (0.26,0.35].	Subdivided surface areas	1, 3
SMR_VSA5	Sum of v_i such that SMR is in (0.44,0.485].	Subdivided surface areas	1, 2, 3
SMR_VSA6	Sum of v_i such that SMR is in (0.485,0.56].	Subdivided surface areas	1, 2, 3
VdistEq	If m is the sum of the distance matrix entries, then VdistEq is defined to be the sum of $\log_2 m - p_i \log_2 p_i / m$ where p_i is the number of distance matrix entries equal to i .	Adjacency and distance matrix descriptors	1, 2, 3
BCUT_SLOGP_2	BCUT descriptors using atomic contribution to logP (using the Wildman and Crippen SlogP method) instead of partial charge; SlogP values in (-0.2,0].	Adjacency and distance matrix descriptors	2, 3
BCUT_SMR_2	The BCUT descriptors using atomic contribution to molar refractivity. SMR values in (-0.2,0].	Adjacency and distance matrix descriptors	1
GCUT_PEOE_1	GCUT descriptors are calculated from the eigenvalues of a modified graph distance adjacency matrix. PEOE- q_i is in the range [0.05,0.10].	Adjacency and distance matrix descriptors	2
GCUT_SLOGP_0	GCUT descriptors using atomic contribution to logP (using the Wildman and Crippen SlogP method). SlogP values ≤ -0.4.	Adjacency and distance matrix descriptors	1
GCUT_SLOGP_2	GCUT descriptors using atomic contribution to logP (using the Wildman and Crippen SlogP method). SlogP values in (-0.2,0].	Adjacency and distance matrix descriptors	1, 2, 3
GCUT_SMR_2	GCUT descriptors using atomic contribution to molar refractivity. SMR values in (-0.2,0].	Adjacency and distance matrix descriptors	1
balabanJ	Balaban's connectivity topological index.	Adjacency and distance matrix descriptors	1, 2, 3
Zagreb	Zagreb index.	Connectivity indices	1
PEOE_VSA_FHYD	Fractional hydrophobic van der Waals surface area.	Partial charge descriptors	1, 2, 3
PEOE_VSA_FPPOS	Fractional positive polar van der Waals surface area.	Partial charge descriptors	1, 2, 3
PEOE_VSA_FPNNEG	Fractional negative polar van der Waals surface area.	Partial charge descriptors	3
PEOE_VSA_HYD	Total hydrophobic van der Waals surface area.	Partial charge descriptors	1, 2
PEOE_VSA_FPOL	Fractional polar van der Waals surface area.	Partial charge descriptors	1
lip_acc	The number of O and N atoms.	Atom counts and bond counts	2
LogS	Log of the aqueous solubility.	Physical properties	1, 2, 3
vs_a_pol	Approximation to the sum of VDW surface areas of polar atoms (atoms that are both hydrogen bond donors and acceptors), such as -OH.	Pharmacophore feature descriptors	1, 2, 3
SlogP	Log of the octanol/water partition coefficient (including implicit hydrogen atoms).	Physical properties	1, 2, 3
TPSA	Total polar surface area.	Physical properties	1, 2, 3

Table 4. Correct classification determined by cross-validation (training set) for each model generated with relevant or P_VSA descriptors.			
		Descriptors	
		Relevant	P_VSA
Model 1	High	78/80 (97.5%)	78/80 (97.5%)
	Weak	75/80 (94%)	78/80 (97.5%)
	All	96%	97.5%
Model 2	High	89/96 (93%)	92/96 (96%)
	Moderate	33/48 (69%)	31/48 (65%)
	Weak	53/59 (90%)	59/59 (100%)
	All	86%	90%
Model 3	High	48/50 (96%)	47/50 (94%)
	Weak	49/50 (98%)	49/50 (98%)
	All	97%	96%

Table 5. Correct classification determined by external validation (test set) for each model generated with relevant or P_VSA descriptors.			
		Descriptors	
		Relevant	P_VSA
Model 1	High	15/16 (94%)	15/16 (94%)
	Weak	17/27 (63%)	20/27 (74%)
	All	74%	81%
Model 3	High	45/46 (98%)	42/46 (94%)
	Weak	8/9 (89%)	9/9 (100%)
	All	96%	93%

split their data set in a similar way, and failed to produce a predictive model because of the error range of the experimental results. Using relevant molecular descriptors to construct the first model, an overall classification accuracy of 86% has been obtained after a 5-fold cross-validation (Table 4, Model 2). Furthermore, all compounds from the weak class have been positively predicted in the second model based only on the P_VSA parameters with an improved accuracy of 90% (decision trees generated for Models 1 and 2 are described in the Supporting Information).

Although the precision of both models for the new weak class improved compared to Model 1, these new models still had difficulty in classifying the moderate blockers, with a precision of 69 and 65% for relevant and P_VSA descriptors, respectively. 12 of 15 and 9 of 17 misclassified moderate blockers were predicted as high blockers for relevant and P_VSA descriptors, respectively. As shown in Figure 2, we selected two relevant and conventional descriptors, TPSA and SlogP, involved in the decision tree generation to represent the distribution of compounds of each class. Interestingly, high and weak compounds seemed to be located in quite distinguishable zones on the graph. High blockers had generally, higher SlogP values and lower TPSA values than weak blockers. In contrast, compounds belonging to the moderate category hold great variability which might explain the low capacity of our models to correctly discriminate intermediate compounds.

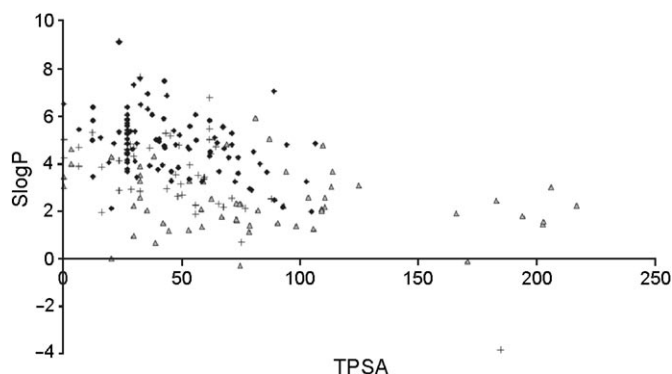


Figure 2. Score plot of two relevant descriptors TPSA and SlogP for hERG compounds from each class; High (◆), Moderate (+), Weak (▲).

Model 3: prediction of high and weak hERG blockers

The results obtained with the previous models indicate the difficulty in classifying the moderate class with a limited number of compounds. Therefore, new RP models using the same set of descriptors have been constructed omitting the moderate class, leaving the extreme classes (high and weak blockers). The molecules were split into training (see Supporting Information) and test sets following the same procedure used for Model 1, that is, a diverse balanced training set composed of 50 molecules per class (Table 1, Model 3). Using relevant descriptors, the tree built was able to successfully classify the compounds from the training set with an overall accuracy of 97% (Table 4, Model 3). In addition, validation of this model using the test set, correctly predicted 98 and 89% of high and weak hERG blockers (Table 5, Model 3). Moreover, another efficient model was obtained using the P_VSA descriptors, with a very low misclassification rate R(T) of 0.04, the model was able to predict correctly 94% of high and 98% of weak hERG blockers from the training set. To further validate this model, the predictions were performed using the validation set. The prediction is still of good quality with 94% of high and 100% of weak hERG blockers classified correctly. The decision trees are shown in Figure 3, correlation matrices of the main molecular descriptors involved in the classification of the hERG channel blockers are given in the Supporting Information and indicate a low degree of correlation among them. For both models, descriptors included logP-based descriptors related to the hydrophobic character of molecules, like SlogP, SlogP_VSA2, SlogP_VSA7, and refractivity descriptors such as SMR_VSA1, SMR_VSA4, SMR_VSA5, and SMR_VSA6 which take into account the size and the polarizability of molecules. The approximation to the sum of van der Waals surface areas of polar atoms (atoms that are both hydrogen bond donors and acceptors) was also found to be a discriminating feature. Other descriptors used were PEOE_VSA_FHYD, PEOE_VSA+0, and PEOE_VSA+1. These atomic partial charge descriptors reflect the positive charge that most of the molecules bear in the present molecular system. A difference between high and weak classes might be noted by observing the descriptor values. For both classes, Table 6 contains averages of the molecular descriptors (P_VSA and relevant) which were used to

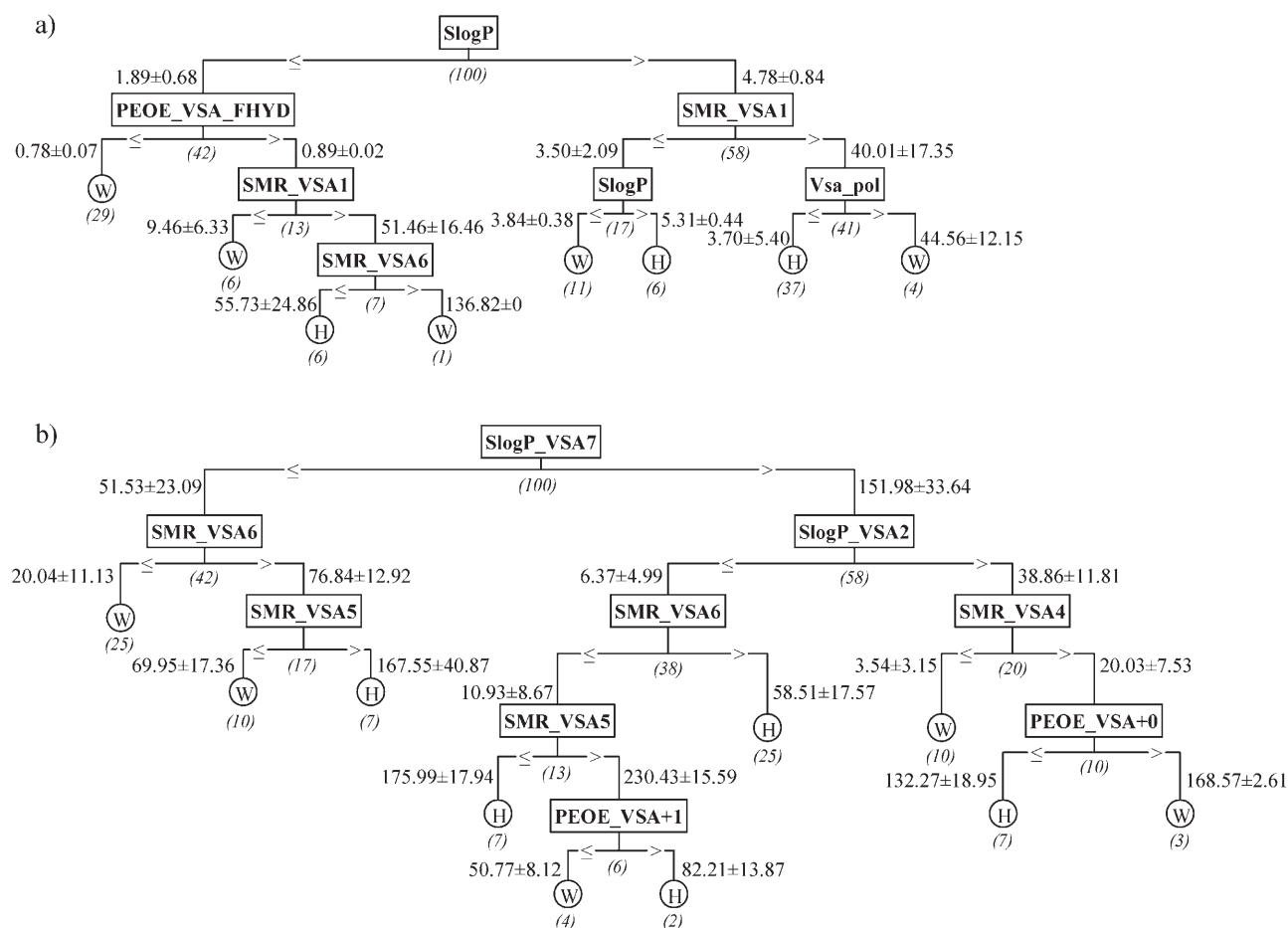


Figure 3. Decision trees for the model 3 training set using either a) relevant descriptors or b) P_VSA descriptors. W and H define compounds classified as weak or high blockers, respectively. For each node, descriptor average value \pm SEM is indicated and distribution of compounds is represented in brackets.

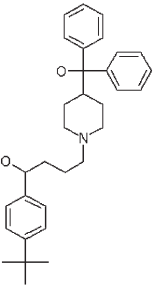
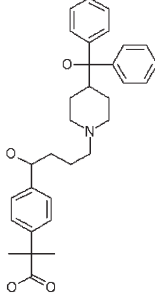
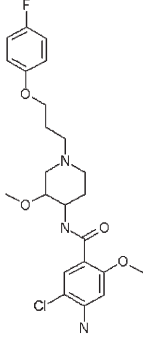
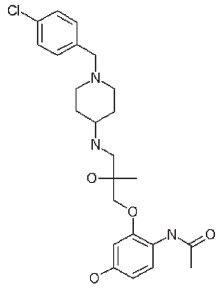
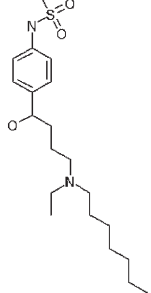
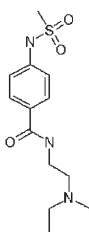
build Model 3. All these descriptors, except *vsa_pol* and *SlogP_VSA2*, were found to be larger for hERG channel blockers. Some relevant selected descriptors are exemplified for molecules from both classes (Table 7). For example, Terfenadine and Fexofenadine belong to different classes with obvious distinct descriptors, and yet have other relevant descriptors which are very close. A simple substitution of *tert*-butyl by isobutyric acid group modifies the molecule classification. This can be used as a local predictor, by changing the substituents around the chemical scaffold to make descriptor values fall into the high- or weak-blocker range.

Interestingly, passing the 48 compounds belonging to the moderate class into these models led to a balanced split of the molecules, for the P_VSA model 52% were classified as high and 48% classified as weak hERG blockers, and using the relevant descriptors set, 69% of the compounds were classified as high and 31% were classified as weak. No significant correlation was found between the IC_{50} values and the classification (data not shown).

Table 6. Differences in the values of selected descriptors in the training set (100 molecules) of Model 3 using P_VSA and relevant descriptors for blockade classification.

Relevant descriptors			P_VSA descriptors		
Descriptor	Average value		Descriptor	Average value	
	High	Weak		High	Weak
<i>SlogP</i>	4.68	2.45	<i>SlogP_VSA2</i>	11.68	45.15
<i>PEOE_VSA_FHYD</i>	0.92	0.84	<i>SlogP_VSA7</i>	135.72	83.86
<i>SMR_VSA1</i>	33.68	31.72	<i>SMR_VSA4</i>	13.08	9.76
<i>SMR_VSA6</i>	52.76	44.22	<i>SMR_VSA5</i>	188.83	122.36
<i>vsa_pol</i>	4.71	17.49	<i>SMR_VSA6</i>	52.76	44.22
			<i>PEOE_VSA + 0</i>	110.59	107.97
			<i>PEOE_VSA + 1</i>	64.82	53.48

Table 7. Descriptor comparison of six compounds with their observed class and their predicted class.

	Molecules ^[a]					
						
Class/Prediction	Terfenadine High/High	Fexofenadine Weak/Weak	Cisapride High/High	C ₂₄ H ₃₂ ClN ₃ O ₄ Weak/Weak	Ibutilide High/High	Sematilide Weak/Weak
SlogP	6.85	5.92	3.66	3.68	4.26	1.13
PEOE_VSA_FHYD	0.93	0.84	0.88	0.86	0.88	0.83
SMR_VSA1	53.89	53.89	40.08	64.89	28.50	3.12
SMR_VSA6	55.32	63.06	81.50	81.50	55.32	79.01
vsa_pol	27.13	54.27	0	32.82	13.57	0

[a] Additional information regarding these molecules can be found in the Supporting Information.

Discussion

In this study, we have developed efficient predictive filters to screen out potential hERG blockers using high quality data sets and a quite powerful statistical method, recursive partitioning classification. One major element in the success of a predictive model is the precision and quality of the biological data from which the model is calculated. From the knowledge base developed at Aureus Pharma, high quality datasets which have a large range of chemical diversity and detailed biological information can easily be generated. Furthermore, although data came from diverse sources with wide experimental variations, a precise selection of data sets allowed us to generate robust binary classification trees with overall accuracies consistently above 80% and as high as 96% for the extreme classes model.

RP is gaining popularity as a method for analyzing large drug discovery screening sets.^[15,17,26] Binary classification trees are simple to understand and interpret, and can be computed very quickly and efficiently. This approach has been successfully used in the field of ADMET. Indeed, Susnow et al.^[15] using binary classifications for the prediction of inhibitors of the cytochrome P450 isoform 2D6, reported 80% overall accuracy using an external set of 51 compounds. Recently a test set of 88 COX-2 inhibitors were correctly classified with an accuracy of 82%.^[16] In this study, we developed three different RP models based on a biological activity cutoff for hERG channel blockage prediction.

Mutagenesis and homology modeling studies have shown that Tyr652 and Phe656 are the key residues primarily responsible for the high affinity of hERG channel for several known ligands.^[6,7,27] These residues are located on the S6 domain, or near the pore helix. Two important physicochemical interactions were identified. The first is a hydrophobic interaction

with Phe656, and the second is an electrostatic interaction (cation- π interaction) between the basic N of the ligand and Tyr652.

3D pharmacophore modeling analysis corroborated these assumptions.^[7-9] In the present study, we employed 2D molecular descriptors to build classification models. Considering the best model, model 3, both P_VSA and relevant descriptors used to build the model encode principal chemical features involved in the blockage of the hERG channel. Hydrophobic interaction is expressed by logP-based descriptors, whereas atomic partial charge descriptors PEOE_VSA_FHYD, PEOE_VSA + 0, and PEOE_VSA + 1, are related to the electrostatic interaction. The size effect of the inhibitor ligand was revealed by the molecular refractivity-based descriptor SMR_VSA. In their SVM study, Tobita et al.^[14] used 57 2D descriptors computed by MOE, and 51 molecular fragment-count descriptors. For a threshold of 1 μ M, they found three 2D descriptors and five molecular fragment-count descriptors to be important for classification accuracy. These descriptors include VSA_Base, PEOE_VSA + 0, SMR_VSA0, and four other fragment-count descriptors. These descriptors are equivalent to the descriptors selected to build our models. These findings confirm the robustness of the recursive partitioning method as used in our present work.

Our first model, separated compounds into high and weak blockers at IC₅₀ = 1 μ M. The trained model reached an overall classification accuracy of 97.5% after a 5-fold cross-validation for the P_VSA descriptor set. Similarly, Tobita and co-workers^[14] reported an accuracy of 90% with an SVM-based model using the same cutoff after a 10-fold cross-validation on 73 drugs. Their model correctly predicted 86% of high, and 93% of weak blockers. Using a larger amount of data (203 molecules), our P_VSA model correctly classified 97.5% of both high and weak blockers. To further validate this model, we applied an

external data set of 43 compounds. Keserü et al.^[11] using a QSAR approach, generated a model (cutoff of 1 μM) able to classify 83% of actives and 87% of inactives correctly, from a test set of 13 compounds. Although we obtained a better prediction for the high blockers (94 vs. 83%), our model had less capacity to classify the weak compounds (74 vs. 87%). Interestingly, among the false negatives a majority had IC_{50} values between 1 and 10 μM , suggesting the involvement of different chemical features for discriminating this category of molecules. Therefore, we generated a second approach using three classes of compounds, including a new moderate class of molecules with this range of activity (1–10 μM). The results reported herein, showed a better positive prediction of high (96%) and weak blockers (100%) achieved with the P_VSA descriptors but a difficulty to predict moderate blockers (overall accuracy of 90%). In order to gain some insight, the distribution of compounds belonging to the moderate class were represented, and compared to high and weak blockers using two relevant descriptors from the decision tree. As shown in Figure 2, the discrimination of moderate blockers was difficult to assess using this set of descriptors. Furthermore, as there were a limited number of compounds in the moderate class, it was not possible to create an external test set. Collection of additional data is currently in process, as is assessment of new statistical approaches to further update and improve our models. In a previous study, Roche et al.^[13], described a model in a supervised neural network system, which considered only the extreme classes of blockage, that is, compounds with a high IC_{50} value under 1 μM and compounds with a low IC_{50} value above 10 μM . 71% of the high blockers and 93% of the non-blockers were correctly predicted in a validation set of 95 proprietary compounds. To critically evaluate our approach, a third model was generated with the extreme classes and using the same descriptors sets. Regarding our test set, the model achieved a prediction accuracy of 96 and 93% for relevant and P_VSA descriptor sets, respectively. The relevant descriptors seem the more interesting model as the depth tree is only 4 and needs less descriptors than the P_VSA model to classify the molecules.

In conclusion, we have developed computer-based models which allow the effective classification of hERG channel blockers. In silico models are gaining interest as the cost of cardiotoxicity failures is recognized. However, it should be noted that the prediction accuracy of these models is strongly affected by the diversity of compounds used as the training set. For this purpose, the Aureus Pharma hERG knowledge database allowed us to extract a biologically filtered and diverse molecular set. The fast RP method is suitable for the early discovery process to screen out new compounds and can be used as a general filter for cardiotoxic side effects related to hERG channel blockage.

Computational Methods

Data set extraction: A series of compounds tested in electrophysiology experiments on the wild type hERG channel have been selected from our hERG knowledge database. The Aureus Pharma

hERG knowledge database is comprised of comprehensive knowledge from literature on the hERG channel, as well as other channels involved in cardiotoxicity.^[19] This first selection retrieved 510 molecules associated with 1970 biological activities, coming from 187 publications and 11 patents (June 2005 release). We then selected 217 molecules having at least, a measured IC_{50} value. As shown in Figure 4, a wide range of biological activities were covered with this set of compounds.

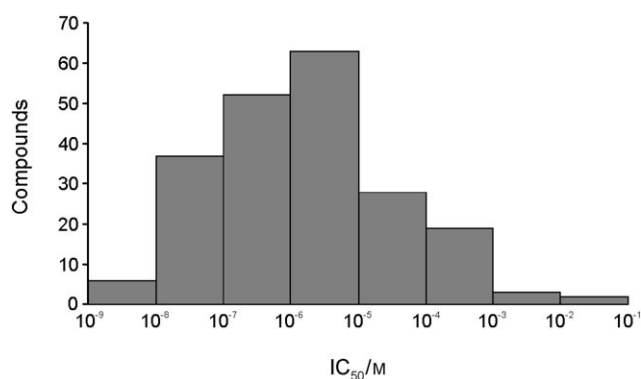


Figure 4. Distribution of IC_{50} values for compounds studied on hERG channel. IC_{50} values from electrophysiology experiments on hERG channel have been retrieved from the hERG database^[19]. The ordinate axis represents the number of compounds and the abscissa represents the range of IC_{50} values.

Next, all molecules for which the IC_{50} had been measured in non-mammalian cells, such as *Xenopus laevis* oocytes (15 out of 217) were removed. We excluded these values based on a correlation study performed with values from mammalian cells and oocytes. It was shown that hERG blocker drug potency was underestimated (as much as 100-fold) in frog cells. This has also been noted by other studies.^[5] To improve our data set, 9 compounds which were described only as having the ability to inhibit hERG channel activity by less than 20% were added to the set. After withdrawing molecules which were stereo-chemical duplicates, the global dataset contained 203 unique compounds, all of which had a biological measure on the hERG channel (for information on the 203 compounds see Supporting Information).

Data set diversity: The degree of diversity of the datasets was evaluated using an algorithm that applies nearest-neighbor searching (ChemAxon application Compr).^[20] A weighted Euclidean distance calculation applied the Tanimoto (Jaccard) coefficient based on ChemAxon CF fingerprints.^[20] The dissimilarity between molecules was given by the following formula:

$$D(A,B) = 1 - T(A,B) = \{[1 - T(A,B)] + w_1[C_1(A) - C_1(B)]^2 + w_2[C_2(A) - C_2(B)]^2 + \dots\}^{1/2} \quad (1)$$

Where $w_1, w_2 \dots$ are weights, $T(A,B)$ is the Tanimoto coefficient for molecules A and B, and $C_i(A)$ is the value of descriptor i of molecule A. Diversity statistics of the library self-dissimilarity test were expressed by the average and maximum dissimilarity.

Two-dimensional molecular descriptors: 2D molecular descriptors were calculated by the QuaSAR-Descriptor module from Molecular Operating Environment (MOE) software.^[21] Two sets of descriptors were generated for all the compounds, 32 P_VSA descriptors and a set containing all 2D descriptors available including P_VSA (184 de-

scriptors). These descriptors are based on evaluating a descriptor for a specific range $[u, v]$ of the specified property values P . The descriptor value corresponds to the sum of the atomic van der Waals (VSA) contributions of each atom i with P_i in $[u, v]$ range. The quantity $P_VSA(u, v)$ is defined by Equation (2):

$$P_VSA(u, v) = \sum V_i \delta(P_i \in [u, v]) \quad (2)$$

where V_i is the atomic contribution of atom i to the VSA of the molecule. A set of n descriptors associated with the property P are defined as follows:

$$P_VSA(u, v) = \sum_1 V_i \delta(P_i \in [a_{k-1}, a_k]) \quad k = 1, 2, \dots, n \quad (3)$$

where $a_0 < a_k < a_n$ are interval boundaries such that $[a_0, a_n]$ includes all values of P_i in any molecule. Each P_VSA -type descriptor is characterized as the amount of surface area with P in a certain range. Thus, Labute defined three sets of P_VSA molecular descriptors: 10 $SlogP_VSA_k$ intended to capture hydrophobic and hydrophilic effects, 8 SMR_VSA_k intended to capture both the size and the polarizability of a molecule, and 14 $PEOP_VSA_k$ to reflect the electrostatic interactions.^[22]

The P_VSA descriptors were found to be weakly correlated with each other over a large collection of compounds and reasonably good QSAR/QSPR models were built using these descriptors.^[16, 22] Other molecular descriptors included: physical properties, atom and bond count descriptors subdivided according to various criteria, Kier and Hall connectivity and Kappa shape indices intended to capture different aspects of molecular shape, and adjacency and distance matrix descriptors including BCUT and GCUT descriptors. Pharmacophore feature descriptors considered only the heavy atoms of a molecule and assigned a pharmacophoric type to each atom. The feature set was donor, acceptor, polar (both donor and acceptor), positive (base), negative (acid), and hydrophobic. Other descriptors were evaluated on the basis of partial charges. MOE uses the Partial Equalization of Orbital Electronegativities (PEOE) method of Gasteiger to calculate partial charges.^[23]

Relevant descriptors: Commonly called feature selection, the process of selecting more relevant descriptors can have a positive effect on the performance of classification algorithms, and enhance their accuracy and speed. To accomplish this task, we used a Correlation-based Feature Selection (CFS) algorithm.^[24] CFS uses a search algorithm along with a function to evaluate the merits of feature subsets, where features are referred to by molecular descriptors. It takes into account the usefulness of individual features for predicting the class label along with the intercorrelation among them. The following Equation (4) formalizes the CFS algorithm:

$$Merit_S = \frac{k \bar{r}_{cf}}{\sqrt{k + k(k-1) \bar{r}_{ff}}} \quad (4)$$

where $Merit_S$ is the heuristic "merit" of a feature subset S containing k descriptors, \bar{r}_{cf} the average feature-class correlation, and \bar{r}_{ff} the average feature-feature intercorrelation. The numerator can be thought of as giving an indication of how predictive are a group of features, and the denominator of how much redundancy there is among them.

Diverse subset composing: Diverse Subset is used to rank entries in a database, based on their distance from each other. The ranking order is suitable for extracting a subset of the database comprising entries which are farthest from each other. To calculate the

distance between two entries, MACCS keys fingerprints were employed using Tanimoto similarity metric.

Computation: Computing molecular descriptors, preparing datasets, building the trees and predicting the classes were performed using MOE on a Windows platform computer (3 GHz CPU, 1 Gb RAM). Decision trees were constructed from the training data using the QuaSAR-Classify module, that implements a recursive partitioning algorithm. The prediction accuracy was evaluated by means of 5-fold cross-validation methodology. To avoid overtraining during the tree growing, QuaSAR Classify used a pruning process. A sequence of subtrees was constructed from the initial tree, and the internal test data set was used to choose the final output tree from this sequence. "Pruning" removed one or more branches of a tree. The roots of the branches removed remained part of the pruned tree becoming leaf nodes. Optimized parameters for Node Split Size (5) and Max Tree Depth (15) were used. A node sized less than or equal to the specified Node Split Size parameter will not be split further, and will become a leaf node of the tree. The Max Tree Depth parameter defines the limit number of splits between the root node and the lowest leaf node. It should be noted that the maximum tree depth observed in this study reached only seven. The final selection of the tree was made by comparison of the internal quality score; the misclassification rate $R(T)$. It measures the proportion of cases that are incorrectly classified by a tree. $R(T)$ can be defined as $N_{\text{misclassified}}/N_{\text{total}}$ where $N_{\text{misclassified}}$ is the total number of misclassified cases and N_{total} is the total number of cases in the training set. Additionally, the performance of obtained models was measured by the quantity of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) used to evaluate: 1) the overall classification accuracy of a prediction model where $\text{Accuracy} = TP + TN / (TP + TN + FP + FN) = 1/R(T)$ including both active and inactive compounds, 2) the accuracy of predicting an active class, $\text{Precision} = TP / (TP + TN)$, and 3) the ability of a predictive model to select instances of a certain class from a data set, $\text{Recall} = TP / (TP + FN)$.

Acknowledgements

The authors express their appreciation to the ChemAxon team for providing JChem tools and their helpful support. We thank Sophie Ollivier and the knowledge management team, as well as Dominique Neaud and the IT team for their valuable help during the preparation of this work. We gratefully thank Mary Donlan for her comments, and for assisting us with the proofreading of the manuscript.

Keywords: computer chemistry · hERG · molecular modeling · QT interval prolongation · recursive partitioning

- [1] a) Y. G. Yap, A. J. Camm, *Heart* **2003**, *89*, 1363–1372; b) I. Cavero, M. Mestre, J.-M. Guillon, W. Crumb, *Expert Opin. Pharmacother.* **2000**, *1*, 947–973.
- [2] K. Finlayson, H. J. Witchel, J. McCulloch, J. Sharkey, *Eur. J. Pharmacol.* **2004**, *500*, 129–142.
- [3] M. Recanatini, E. Poluzzi, M. Masetti, A. Cavalli, F. De Ponti, *Med. Res. Rev.* **2005**, *25*, 133–166.
- [4] J. Tamargo, *Jpn. J. Pharmacol.* **2000**, *83*, 1–19.
- [5] a) M. C. Sanguinetti, J. S. Mitcheson, *Trends Pharmacol. Sci.* **2005**, *26*, 119–124; b) A. M. Aronov, *Drug Discovery Today* **2005**, *10*, 149–155.
- [6] J. S. Mitcheson, J. Chen, M. Lin, C. Culbertson, M. C. Sanguinetti, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 12329–12333.

- [7] R. A. Pearlstein, R. J. Vaz, J. Kang, X. L. Chen, M. Preobrazhenskaya, A. E. Shchekotikhin, A. M. Korolev, L. N. Lysenkova, O. V. Miroshnikova, J. Hendrix, D. Rampe, *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829–1835.
- [8] A. Cavalli, E. Poluzzi, F. De Ponti, M. Recanatini, *J. Med. Chem.* **2002**, *45*, 3844–3853.
- [9] S. Ekins, W. J. Crumb, R. D. Sarazan, J. H. Wikel, S. A. Wrighton, *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427–434.
- [10] a) R. Rajamani, B. A. Tounge, J. Li, C. H. Reynolds, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1737–1741; b) G. Cianchetta, Y. Li, J. Kang, D. Rampe, A. Fravolini, G. Cruciani, R. J. Vaz, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3637–3642.
- [11] G. M. Keserü, *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773–2775.
- [12] A. M. Aronov, B. B. Goldman, *Bioorg. Med. Chem.* **2004**, *12*, 2307–2315.
- [13] O. Roche, G. Trube, J. Zuegge, P. Pflimlin, A. Alanine, G. Schneider, *ChemBioChem* **2002**, *3*, 455–459.
- [14] M. Tobita, T. Nishikawa, R. Nagashima, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886–2890.
- [15] R. G. Susnow, S. L. Dixon, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1308–1315.
- [16] N. Baurin, J.-C. Mozziconacci, E. Arnoult, P. Chavatte, C. Marot, L. Morin-Allory, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276–285.
- [17] A. Rusinko, M. W. Farmen, C. G. Lambert, P. L. Brown, S. S. Young, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017–1026.
- [18] C. Buyck in *EuroQSAR 2002. Designing Drugs and Crop Protectants: Processes, Problems, and Solutions* (Eds.: M. Ford, D. Livingstone, J. Dearden, H. Van der Waterbeemd), Blackwell, Oxford, **2003**, pp. 86–89.
- [19] Aureus Pharma, Paris, France, <http://www.aureus-pharma.com>.
- [20] JChem, version 3.0.10, ChemAxon, Budapest, Hungary, <http://www.chemaxon.com>.
- [21] MOE (Molecular Operating Environment), version 2004.03, Chemical Computing Group Inc., Montreal, Canada, <http://www.chemcomp.com>.
- [22] P. Labute, *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.
- [23] J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219–3228.
- [24] M. A. Hall, PhD thesis, Waikato University, New Zealand, **1999**.
- [25] Specs, Delft, Holland, <http://www.specs.net>.
- [26] P. E. Blower, K. P. Cross, M. A. Fligner, G. J. Myatt, J. S. Verducci, C. Yang, *Curr. Drug Discovery Technol.* **2004**, *1*, 37–47.
- [27] D. Fernandez, A. Ghanta, G. W. Kauffman, M. C. Sanguinetti, *J. Biol. Chem.* **2004**, *279*, 10120–10127.

Received: December 19, 2005
Published online on May 5, 2006