Understanding hERG inhibition with QSAR models based on a one-dimensional molecular representation

David J. Diller · Doug W. Hobbs

Received: 15 February 2007 / Accepted: 29 April 2007 / Published online: 5 June 2007

Abstract Blockage of the potassium channel encoded by the human ether-a-go-go related gene (hERG) is well understood to be the root cause of the cardio-toxicity of numerous approved and investigational drugs. As such, a cascade of in vitro and in vivo assays have been developed to filter compounds with hERG inhibitory activity. Quantitative structure activity relationship (QSAR) models are used at the very earliest part of this cascade to eliminate compounds that are likely to have this undesirable activity prior to synthesis. Here a new QSAR technique based on the one-dimensional representation is described in the context of the development of a model to predict hERG inhibition. The model is shown to perform close to the limits of the quality of the data used for model building. In order to make optimal use of the available data, a general robust mathematical scheme was developed and is described to simultaneously incorporate quantitative data, such as IC50 = 50 nM, and qualitative data, such as inactive or IC50 > 30 μM into QSAR models without discarding any experimental information.

Keywords hERG · 1D-QSAR · Robust regression

D. J. Diller (✉)
Department of Molecular Modeling, Pharmacopeia Inc, CN5350, Princeton, NJ 08543-5350, USA
e-mail: ddiller@pharmacop.com

D. W. Hobbs
Department of Biochemistry and Molecular Biophysics, Washington University, St. Louis, MO 63110, USA

Introduction

Approved drugs such as Terfenadine, Astemizole, and Grepafloxin, have been withdrawn from the market while others such as Thioridazine and Pimozide have had their sales restricted because they induce a rare form of cardiac arrhythmia called torsades de pointes (TdP) [1–3]. Subsequently, all of these compounds have been shown to be potent inhibitors of the potassium channel encoded by the human ether-a-go-go related gene (hERG). The hERG potassium channel is the molecular determinant for the rapid component of the delayed rectifier current which makes it a key factor in the repolarization of cardiac cells. Thus hERG is a key determinant in the cardiac action potential duration and the QT interval of the electrocardiogram [4]. Blockage of the hERG potassium channel is now accepted as the most common cause of drug induced TdP.

Drug induced TdP most commonly occurs in conjunction with a variety of risk factors such as impaired hepatic/renal function, heart disease, electrolyte imbalance, drug–drug interactions, or genetic susceptibility [2]. As a result, the incidence of drug induced TdP is very rare often between 1 and 10 instances in 1,000,000 uses even with a potent hERG blocker [5]. The infrequency of occurrence makes drug induced TdP extremely difficult to detect in clinical trials and increases the importance of detecting and eliminating hERG blocking activity early in the drug development process.

The typical hERG screening paradigm begins with a high throughput assay such as rubidium efflux [6–9] or a competitive binding assay [10–13]. The next level assay, often considered the gold standard in vitro assay, is a patch clamp assay performed in an artificial cell line [14]. Next the prolongation of action potential duration is measured in mammalian purkinje fibers [15, 16]. Each assay is more resource and time consuming than the previous but is believed to be a better predictor of potential cardiotoxicity in man. Thus each assay acts as a filter to ensure the compounds with the greatest likelihood of exhibiting little or no cardiotoxicity are actually tested and ultimately that only those that are very likely to be safe enter clinical trials [5].

Computational models can play and in most places are playing a role [17–28] in the cascade of tests to ensure that new compounds are devoid of the hERG blocking activity. While not as predictive as their in vitro counterparts, computational models can still play a valuable role as they can help prioritize compounds prior to synthesis thereby decreasing the number of compounds that make it to development and still have hERG blocking liabilities. The impetus of the current work was to develop a computational model to predict hERG inhibition in patch clamp assays in artificial cell lines using data extracted from literature sources.

In general, data sets taken from multiple laboratories present many challenges that put them outside the scope of many classical QSAR methods. Three-dimensional methods, such as 3D-QSAR [17], pharmacophore modeling [18, 23, 26] or homology modeling [21, 27, 28], suffer from a conformational explosion as the flexibility of the molecules and the size of the data set increases. Determining the bioactive conformation remains a difficult problem for any single molecule with more than a small number of rotatable bonds. The uncertainty in the bioactive conformation particularly when extended over many molecules with different chemotypes adds a degree of uncertainty to the resulting models. Additional uncertainty is introduced when an explicit three-dimensional alignment is required for model building, such as with 3D-QSAR. The uncertainty with the three-dimensional alignment becomes particularly unwieldy for data sets having many chemotypes, like the data set described here. For this reason, hERG models built with three-dimensional methods have typically used relatively small data sets.

Traditional two-dimensional QSAR methods [19, 20, 22] do not suffer from the same problems as the three-dimensional methods. By two-dimensional QSAR we refer to those pattern recognition and regression techniques that rely on descriptors that can be calculated solely from the molecular topology and do not explicitly depend on any choice of three-dimensional coordinates. In particular, because these methods rely only on the topological structure of the molecules involved, the two-dimensional QSAR methods do not introduced any ambiguity due to a particular choice of conformation or three-dimensional alignment. Since they lack explicit structural information, however, two-dimensional QSAR techniques often have difficulty with modeling specific interactions with proteins particularly when multiple chemotypes are involved.

Here we describe a QSAR method that offers a balance between the two-dimensional and three-dimensional QSAR methods. The method relies on the one-dimensional representation developed by Dixon and Merz [29]. The one-dimensional representation is a projection of the molecule, using multi-dimensional scaling, into one dimension so as to maintain as much of the two-dimensional topological distance information as possible. The one-dimensional representation does not depend on explicit three-dimensional coordinates and has been shown to retain the majority of the topological information present in the structure [30].

Recently, the one-dimensional representation was used [30] to create multiple alignments of molecules with a common biological activity much like multiple sequence alignment is used for analyzing related protein or DNA sequences. The QSAR method described here, referred to as 1D-QSAR, builds on this effort. A 1D-QSAR model correlates computed properties of the atoms with their location in a multiple one-dimensional alignment much like 3D-QSAR correlates the steric and electrostatic aspects of each atom with its position within the three-dimensional alignment. Since the search space is much simpler than those encountered in three-dimensional problems, the alignment for a 1D-QSAR model is refined during model building thereby resulting in an alignment that is data driven and free from the bias of the expectation of the user.

A second common complication with data sets derived from literature sources is that we expect them to have high outlier content due to differences in experimental protocols. While differences in experimental protocols is one means for outliers to enter a data set there are many other mechanisms that can introduce outliers into a data set. Compounds that achieve their activity via a mechanism in a manner that is fundamentally different from that of the majority of the compounds will be outliers. As an example the majority of p38 inhibitors discovered to date are directly ATP competitive [31]. There are, however, examples of true allosteric ATP competitive p38 inhibitors [32]. If two types of compounds that achieve their biological activity in fundamentally different ways are used to train a model and one type of compound dominates the training data then in the best case the second type would show up as outliers and in the worst case would decrease the predictive power of the model on the first type of compound. As they can adopt multiple distinct conformations, such as open and closed forms, and often have multiple binding sites, the possibility for dual modes of inhibition of cation channels in general and hERG in particular is of genuine concern. A final type of outlier occurs in cases where a compound's representation is not consistent with its biological activity. Examples of this include a molecule that is incorrectly aligned in a 3D-QSAR model or important conformational or stereochemical effects that cannot be captured by two-dimensional descriptors.

A third complication with the data set used in this work is that it contains both qualitative and quantitative data. By a quantitative data point we mean a molecule having a measured IC50, and by qualitative data point we typically mean a molecule that has been shown to be inactive. As with most instances with specific binding, chemistry space contains far more inactive compounds than active compounds. Thus in order to develop a general model, information concerning inactive compounds is critical. Mathematically consistent techniques to simultaneously handle both qualitative and quantitative data without discarding information from either class are necessary to maximize the use of the typically limited available data. Unfortunately, such techniques are absent from QSAR methodology.

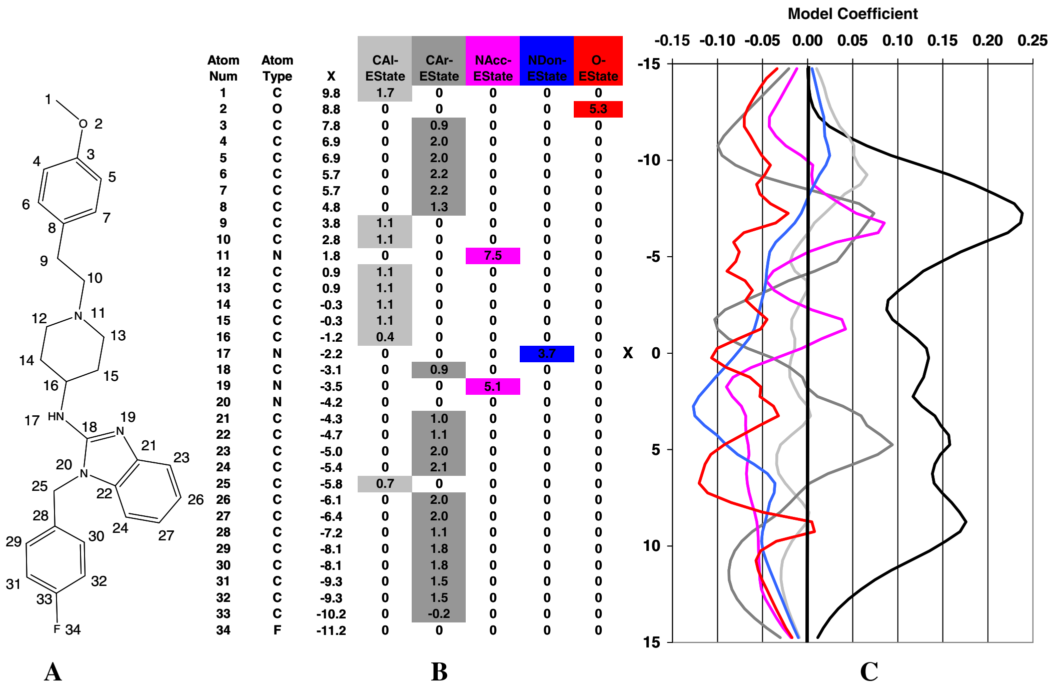
Since the data set used here consists of qualitative and quantitative data and is expected to have high outlier content, a robust mathematically consistent scheme to handle both types of data without discarding information is critical. While this mathematical scheme is presented in the context of the 1D-QSAR model building process, it is generally applicable to other regression methods.

Methods

A one-dimensional representation for a molecule is a single coordinate, x_i, for each atom in the molecule. Thus a molecule could have many one-dimensional representations. Each molecule is assigned its canonical one-dimensional representation according to the previously published scheme [29]. In essence, the canonical one-dimensional representation for a molecule is derived via multi-dimensional scaling from the two-dimensional topological distances so that the resulting one-dimensional atomic distances are as close as possible to their corresponding two-dimensional topological atomic distances. The one-dimensional representation for a molecule is unique up to translation of its one-dimensional coordinates or to flipping its coordinates, i.e., multiplying the coordinates by –1. Mathematically, if $x_1,..., x_n$ is the canonical one-dimensional representation then $px_1 + y,..., px_n + y$, where $p = \pm 1$ and y is any real number, is an equivalent one-dimensional representation.

In addition to its one-dimensional coordinate, each atom is assigned a finite number (J) of descriptors. For the development of the hERG model describe here, we settled on 6 atom descriptors (Fig. 1). The first descriptor is simply 1 for each non-hydrogen atom. While this might seem like adding a constant term to the model, it is not because the one-dimensional representations have varying densities of atoms along the one-dimensional axis. This descriptor captures the density of atoms along the one-dimensional axis and allows for the model to create regions of favored or disfavored atom occupancy. The remainder of the atom descriptors are based on the E-state keys [33–35] separated according to five atom types. The E-state keys were selected because they can be calculated solely from molecular topology, they capture valence and steric information about each atom, and they have been successfully used for a wide range of pharmaceutically related QSAR problems. The second atom descriptor is the E-state key for aliphatic (non-aromatic) carbons. The third descriptor is the E-state key for aromatic carbons. The fourth atom descriptor is the E-state key for nitrogen atoms that can act as hydrogen bond acceptors. The fifth atom descriptor is the E-state key for nitrogen atoms that are capable of acting as hydrogen bond donors. To classify nitrogen as an acceptor, donor or both the neutral form of the molecule was used. A nitrogen donor is any nitrogen with a bonded hydrogen. A nitrogen acceptor is any nitrogen with an available lone pair. For example, nitrogen acceptors include amine, pyridine and cyano nitrogens but not aniline, amide or urea nitrogens as the lone pair is treated as being conjugated with the adjoining atoms. As examples, the nitrogens of N-methyl piperidine, pyridine, and cyano have E-state keys of 2.4, 3.8 and 7.3 respectively. In order that it correlate positively with pK_a, 10 minus the E-state key was used as the descriptor for a nitrogen acceptor. The sixth atom descriptor is the E-state key for oxygen atoms. For all five E-state based descriptors, atoms that are not of the particular type are assigned a value of 0 for that particular descriptor.

The particular delineation of atom types used in conjunction with the E-state keys was based on hypotheses put forth in the literature regarding important elements for modulating hERG binding. Lipophilicity has consistently been indicated as a contributing factor in hERG blockers. For this reason, carbon was singled out as a descriptor. The aliphatic carbons and aromatic carbons were separated into different types because models that further separated the carbons into aromatic and aliphatic were clearly superior, data not shown, to those that treated all carbons as the same type. The second key factor in potent hERG blockers is a basic amine. Since a basic center is typically expected to be a key contributor to hERG activity, the nitrogen acceptor was made a separate type. Other than a basic center, polar functionality is generally expected to diminish hERG activity within a series of compounds. Thus the nitrogen donor and oxygen atoms were further distinguished. Separating oxygen into donor and acceptor types did not



Atom Num	Atom Type	X	CAl-EState	CAr-EState	NAcc-EState	NDon-EState	O-EState
1	C	9.8	1.7	0	0	0	0
2	O	8.8	0	0	0	0	5.3
3	C	7.8	0	0.9	0	0	0
4	C	6.9	0	2.0	0	0	0
5	C	6.9	0	2.0	0	0	0
6	C	5.7	0	2.2	0	0	0
7	C	5.7	0	2.2	0	0	0
8	C	4.8	0	1.3	0	0	0
9	C	3.8	1.1	0	0	0	0
10	C	2.8	1.1	0	0	0	0
11	N	1.8	0	0	7.5	0	0
12	C	0.9	1.1	0	0	0	0
13	C	0.9	1.1	0	0	0	0
14	C	-0.3	1.1	0	0	0	0
15	C	-0.3	1.1	0	0	0	0
16	C	-1.2	0.4	0	0	0	0
17	N	-2.2	0	0	0	3.7	0
18	C	-3.1	0	0.9	0	0	0
19	N	-3.5	0	0	5.1	0	0
20	N	-4.2	0	0	0	0	0
21	C	-4.3	0	1.0	0	0	0
22	C	-4.7	0	1.1	0	0	0
23	C	-5.0	0	2.0	0	0	0
24	C	-5.4	0	2.1	0	0	0
25	C	-5.8	0.7	0	0	0	0
26	C	-6.1	0	2.0	0	0	0
27	C	-6.4	0	2.0	0	0	0
28	C	-7.2	0	1.1	0	0	0
29	C	-8.1	0	1.8	0	0	0
30	C	-8.1	0	1.8	0	0	0
31	C	-9.3	0	1.5	0	0	0
32	C	-9.3	0	1.5	0	0	0
33	C	-10.2	0	-0.2	0	0	0
34	F	-11.2	0	0	0	0	0

Fig. 1 An overview of a one-dimensional QSAR model. (**A**) Astemizole. (**B**) The one-dimensional representation of Astemizole. The atom numbering (column 1) refers to the atom numbering in (**A**). The ''X'' column is the one-dimensional coordinate. The remaining five columns are the five atom descriptors based on the E-state keys. (**C**) An example of a one-dimensional QSAR model. The *black curve* corresponds to the const atom property which is not shown in (**B**) because it is one for every non-hydrogen atom. The *light gray curve* corresponds to the aliphatic carbon E-state property. The *dark gray curve* corresponds to the aromatic carbon E-state property. The *purple curve corresponds* to the nitrogen acceptor E-state property. The *blue curve* corresponds to the nitrogen donor E-state property. The *red curve* corresponds to the oxygen E-state curve. To calculate the score for a particular atom the value of each atom descriptor is multiplied by the value of the model coefficient (horizontal axis) at the corresponding one-dimensional coordinate (vertical axis) of the given atom. For example, for atom 11 the only non-zero descriptors are the constant property which has a value of 1.0 and the nitrogen acceptor E-state property which has a value of 7.5. Since the one-dimensional coordinate of this atom is 1.8, the nitrogen acceptor E-state property is multiplied by the value of the purple curve (–.09) at $X = 1.8$ and the constant property is multiplied by the value of the *black curve* (~0.12) at the same position. Thus the score for atom 11 is $1.0 \times 0.12 - 7.5 \times 0.09 \approx -0.56$, i.e., in this alignment to the model atom 11 will decrease the calculated activity of Astemizole by a little over half a log unit. If, however, the one-dimensional representation were translated such that the coordinate of this atom were –1.5, i.e., subtract 3.3 from each one-dimensional coordinate, the nitrogen acceptor E-state property is multiplied by the value of the *purple curve* (~0.04) at $X = -1.5$ and the constant property is multiplied by the value of the *black curve* (~0.1) at the same position. Thus the score for atom 11 is $7.5 \times 0.04 + 1.0 \times 0.1 \approx 0.4$. With this model and this particular alignment of Astemizole to the model, atom 11 contributes 0.4 log units to the predicted hERG IC50 for Astemizole. To get the calculated pIC50 for Astemizole in any alignment to the model simply repeat this procedure for each atom and sum the resulting atomic contributions

improve the models in any discernable way, and thus all oxygens where treated as a single type. No special type was created for sulfur atoms because these were fairly uncommon in the data set. In this data set sulfurs most commonly occurred in sulfonamides and thiophenes. In these cases the sulfur atom affects the E-state keys of neighboring atoms, and thus their effects are still felt even though they are not explicitly handled. Similarly, no type for halogen atoms was included though they are still important as they also significantly affect the E-state keys of neighboring atoms.

Finally, we noticed on initial use of these models that occasionally an unusual value for an E-state key would lead to a very surprising calculated pIC50—particularly many false positives. For example, in most cases the E-state key for an aliphatic carbon is positive. So a negative coefficient in the model would mean that a typical aliphatic carbon will decrease the calculated activity. But certain carbons, such as that in a CF3, will have a very large negative E-state key. Thus rather than decreasing the calculated activity where the aliphatic carbon coefficient is negative it increases the activity. To prevent these types of artifacts, only the positive portion of each atom descriptor was retained, i.e., if an atom descriptor was less than 0, it was set to 0. This has the added benefit that positive coefficients always indicate that the atom type increases the calculated activity and negative coefficients decrease the calculated activity.

A 1D-QSAR model is a set of J functions, $f_1,..., f_J$, defined on a one-dimensional interval (L, R) (see Fig. 1c). Note that each of the J functions corresponds to one of the J atom descriptors described in the preceding paragraphs. With these functions defined, the predicted biological activity for a molecule, η, is given by

$$P(\eta) = C + \sum_{k=1}^{K}\sum_{j=1}^{J} d_{kj} f_j(x_k) \tag{1}$$

where C is a constant, K is the number of atoms in the molecule η, d_{kj} is the jth descriptor of the kth atom of this molecule, and x_k is the one-dimensional coordinate of the kth atom.

The functions f_j for $j = 1,..., J$, are determined as follows. The one-dimensional interval (L, R) is split into N even sub-intervals of width $\mathrm{d}x = (R - L)/N$. Then the values of the f_j at the points $x_\mathrm{n} = L + n \times \mathrm{d}x$ for $n = 0,..., N$ are treated as free variables. The values for the f_j at points in between two successive x_n are determined by linear interpolation. The values of $f_j(x_n)$ for $j = 1,..., J$ and $n = 0,..., N$ are determined by minimizing the following:

$$G(F) = \sum_{m=1}^{M} E(P(\eta_m) - a_m) + \gamma \sum_{j=1}^{J}\left[(f_j(x_0))^2 + (f_j(x_N))^2\right] + \gamma \sum_{j=1}^{J} \int_{L}^{R} \left(\frac{\mathrm{d}f_j}{\mathrm{d}x}\right)^2 \mathrm{d}x \tag{2}$$

where F refers collectively to the free variables $f_j(x_n)$ for $j = 1,..., J$ and $n = 0,..., N$, M is the number of molecules in the training set, E is the error function described below, η_m is the mth molecule, a_m is the experimentally measured biological activity of the molecule η_m, and γ is a positive constant referred to as the ridge coefficient. The first term in Eq. 2 simply penalizes for differences between the calculated activity and the measured activity. The second term penalizes the functions for being different from 0 at either endpoint. Typically, the aligned molecules are well within the left and right limits of the entire interval, and thus the functions at each end are undetermined. The second term forces the f_j to 0 at the ends of the interval unless the data suggest otherwise. The third term penalizes for excessive variation in the f_j. The third term greatly reduces the number of effective free variables and guarantees that the f_j vary only when there is significant data supporting the variation. The third term greatly reduces the risk of over fitting.

For linear regression based QSAR applications, the error function, E, in Eq. 3, is quadratic, i.e., $E(z) = z^2$. The chief advantage of using the quadratic penalty function is that it allows the minimum of Eq. 2 to be determined analytically using standard matrix manipulations. This is not a significant advantage because the minimum of Eq. 2 can be found numerically to high levels of precision using standard gradient-based minimization techniques.

A second advantage of linear regression is that it allows a mathematical analysis of the effective number of free parameters for constrained regression problems [36] such as this one. The information on the effective number of free parameters is critical for understanding when over-fitting is likely to occur. Figure 2 shows the effective number of free variables as a function of the ridge coefficient (γ in Eq. 2) when using linear regression. For the hERG data set described in this work, without the constraints there are 361 free variables: 6 functions with 60 free variables each and the constant term. With a ridge coefficient of 10 there are effectively 66 free variables, at 100 there are effectively 36 free variables and at 1,000 there are effectively 16 free variables. Given the size of the training sets (a few hundred compounds) and the information on the resulting number of free variables we should anticipate that the optimal value for the ridge coefficient would be in the range 10–1,000.

The shortcoming of using linear regression is that it causes undue bias as a result of the outliers in the data set, i.e., it is not robust. Since we expect that our data set contains numerous outliers, linear regression is not ideal for building models with this particular data set. To work with this data set, $E(z)$, where z is the difference between the measured and calculated activity, should possess the following properties:

1. $E(0) = 0$ and $E(z) \geq 0$ for all z, i.e., there is no penalty if the calculated activity agrees with the measured activity.
2. E is convex.

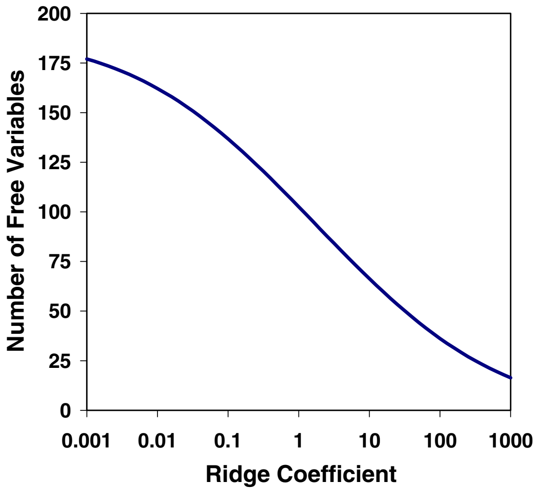


Fig. 2 The theoretical number of free variables as a function of the ridge coefficient

3. E is continuously differentiable.
4. $E(z)$ should increase roughly linearly as z gets large, i.e., $dE/dz \sim 1$ when $z >> 0$ and $dE/dz \sim -1$ when $z << 0$.

Property 2 guarantees that Eq. 2 has a unique local minimum. Strictly speaking property 3 could be replaced by the weaker requirement that E simply be continuous, but as written property 3 allows for the minimum of Eq. 2 to be found using simple gradient based minimization techniques such as conjugate gradient [37]. Property 4 prevents outliers from excessively influencing the final models, i.e., it makes the technique robust.

A second complication with the error function occurs when qualitative data are included. For the hERG data set described here, there are a significant number of inactive compounds. In most cases of specific binding to macromolecules, the majority of compounds in chemistry space are inactive. For developing a general model the inactive compounds contain valuable information that should not be discarded during model development. The negative data are particularly important in this case because the goal is to design compounds that show no hERG activity.

To satisfy properties 1–4 above and to handle the mixed data types, we use approximations to the error functions developed for support vector machines (SVM) [36, 38, 39]. For compounds with quantitative data (for example, $pK_i = 7.5$) we use the following penalty function:

$$E(z) = \frac{1}{\alpha}\ln\left(\frac{1+e^{-\alpha(\varepsilon+z)}+e^{-\alpha(\varepsilon-z)}}{1+2e^{-\alpha\varepsilon}}\right) \tag{3}$$

where α and ε are positive constants and z is the difference between the calculated activity and the observed biological activity (see Fig. 3a). For any positive value of α and ε one can show that $E(z)$ as defined by Eq. 3 satisfies the properties 1–4. The constant α controls the extent to which E approximates the functions used in support vector machines. As $\alpha \to \infty$, $E(z)$ converges exactly to the SVM penalty functions:

$$E(z) = \begin{cases} |z| - \varepsilon & \text{if } |z| > \varepsilon \\ 0 & \text{if } |z| \le \varepsilon \end{cases} \tag{4}$$

Typically, we take $\alpha = 4.0$. Loosely, ε is the expected intrinsic variability of the biological data. For large values of α if the calculated activity is within ε of the measured activity then the penalty will be essentially 0. Typically, we take $\varepsilon = 0.5$ which allows for an intrinsic variability of approximately a factor of 3 in the measured IC50s.

Minor modifications of Eq. 3 make the error function adequate for qualitative data. In cases where we have an upper bound for the pIC50 we use (see Fig. 3b)

$$E(z) = \frac{1}{\alpha}\ln(1+e^{\alpha z}) \tag{5}$$

and in cases for compounds with a lower bound on the pIC50 we use

$$E(z) = \frac{1}{\alpha}\ln(1+e^{-\alpha z}) \tag{6}$$

where again α is a positive constant and z is the calculated activity minus the measured bound for the biological

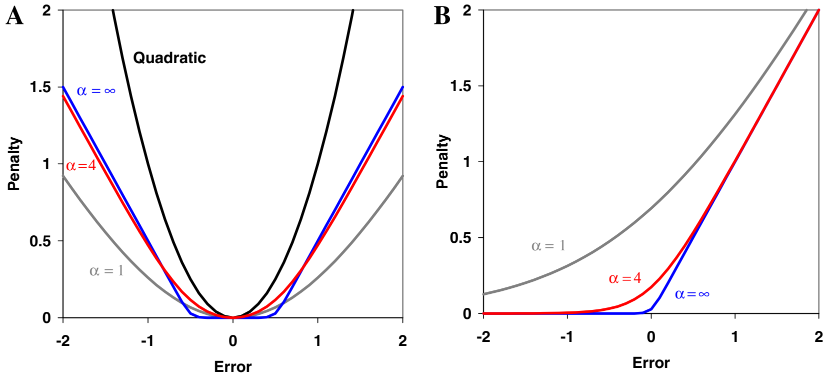


Fig. 3 The penalty functions used for model building. (**A**) The penalty function used for the quantitative data. In this case the error is simply the difference between the measured and calculated pIC50s. The quadratic penalty function (*black*) and the SVM penalty function (*blue*) are shown for comparison. For all the work done in this paper $\alpha = 4.0$ was used for both penalty functions and a value of $\varepsilon = 0.5$ was used for the second penalty function. (**B**) The penalty functions used for the qualitative data. For this purpose the qualitative data consist of inactive compounds in which case we assume the IC50 ≥ 30 μM, i.e., pIC50 ≤ 4.5. Thus the error (x-axis) is the calculate pIC50 minus 4.5. The choice of 30 μM as a cutoff for inactive compounds was based on the observation that compounds were often reported as inactive when they showed little or no activity at 10 μM. No activity at 10 μM should safely translate to IC50 ≥ 30 μM. No attempt was made to optimize the cutoff. The SVM penalty function ($\alpha = \infty$, *blue*) is shown for comparison purposes

activity of the compound. For the work presented here all the qualitative data are for compounds that are deemed inactive in which case we take pIC50 ≤ 4.5, i.e., IC50 ≥ 30 µM. Again as $\alpha \rightarrow \infty$, $E(z)$ converges exactly to the SVM functions (see Fig. 3b).

To build a 1D-QSAR model, the one-dimensional representations of the molecules must initially be put into a common frame of reference, i.e., they must all be aligned similar to a multiple sequence alignment of protein sequences. To develop a multiple alignment of the molecules we typically align a small diverse subset (~10–20) of the most potent molecules from the training set using the methods described earlier [30]. The remaining molecules are then aligned one by one to the multiple alignment of this subset.

Initial attempts to build a 1D-QSAR model using the above fitting procedure resulted in models that performed to within the limits of the data quality (mean absolute errors of ~0.6) on test and training sets. Conceptually, we imagine this process as building an image of the hERG binding site, in so far as is possible, in one dimension. Once an image of the binding site is available the most rational way to make predictions on new molecules is to position them in the image where maximal activity is achieved. For example, when docking compounds their predicted activity is based on where they achieve their optimal score. Thus we decided that the most reasonable method to put new molecules into the frame of reference of the model was to align each new molecule so as to maximize its calculated biological activity. We emphasize that the realignment was not to improve the fit to data only to ensure that each molecule was aligned where it achieved its greatest calculated activity. Since this is a one-dimensional search problem, the global maximum calculated activity for a molecule can be found very rapidly using essentially systematic search techniques [37].

Unfortunately, realigning each member of the training set to the initial 1D-QSAR model so as to maximize each calculated pIC50 resulted in calculated pIC50s that showed little or no correlation with their measured pIC50s. A close examination showed that while the most potent members of the training set did not significantly change position within the multiple alignment, the calculated pIC50s of many of the weakly active and inactive molecules increased significantly because they could find another way to align to the model to improve their calculated pIC50. This observation combined with the realization that realigning each molecule to the model so as to maximize its calculated pIC50 results in a new one-dimensional alignment led to the final protocol we used for building a 1D-QSAR model (see Fig. 4):

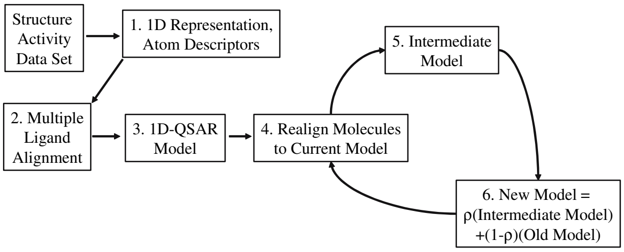

Structure Activity Data Set
1. 1D Representation, Atom Descriptors
2. Multiple Ligand Alignment
3. 1D-QSAR Model
4. Realign Molecules to Current Model
5. Intermediate Model
6. New Model = ρ(Intermediate Model) +(1-ρ)(Old Model)

Fig. 4 The process used to build a 1D-QSAR model

1. From an initial 1D alignment of the training set compounds, build a 1D-QSAR model according to the procedure described above. Proceed to step 2 with this model as the current model.
2. Realign all molecules in the training set to the current model so as to maximize each molecule's calculated activity. This results in a new multiple one-dimensional alignment.
3. Using the new multiple alignment from Step 2, build another 1D-QSAR model according to the procedure described above. Refer to this model as the intermediate model.
4. Construct a new model as the weighted average of the current model and the intermediate model from step 3, i.e., each of the f_j is $f_j^{\text{new}} = \rho f_j^{\text{current}} + (1-\rho) f_j^{\text{intermediate}}$ where $0 < \rho < 1$.
5. Return to step 2 using the new model constructed in step 4 as the current model.

Steps 2–5 are repeated until the 1D-QSAR model converges. Ultimately, this procedure builds a model and refines the multiple alignment so as to minimize Eq. 2 subject to the constraint that each molecule is aligned to the model at its global maximum calculated potency. Though the outlined procedure does not guarantee that it solves this global optimization problem, it has in practice provided near optimal solutions. No attempt was made to pick an optimal value for the parameter ρ used for averaging of models from step to step. A few different values were tested, and they all resulted in essentially identical final models. For all the work presented here in the nth iteration through steps 2–5 above we set $\rho = 0.7/\sqrt{n}$. Other aspects of the process, such as the ridge coefficient and the initial alignment, were rigorously examined and are discussed below.

Results

The data set

The data set was assembled from literature sources. A total of 190 compounds with at least one measured hERG IC50

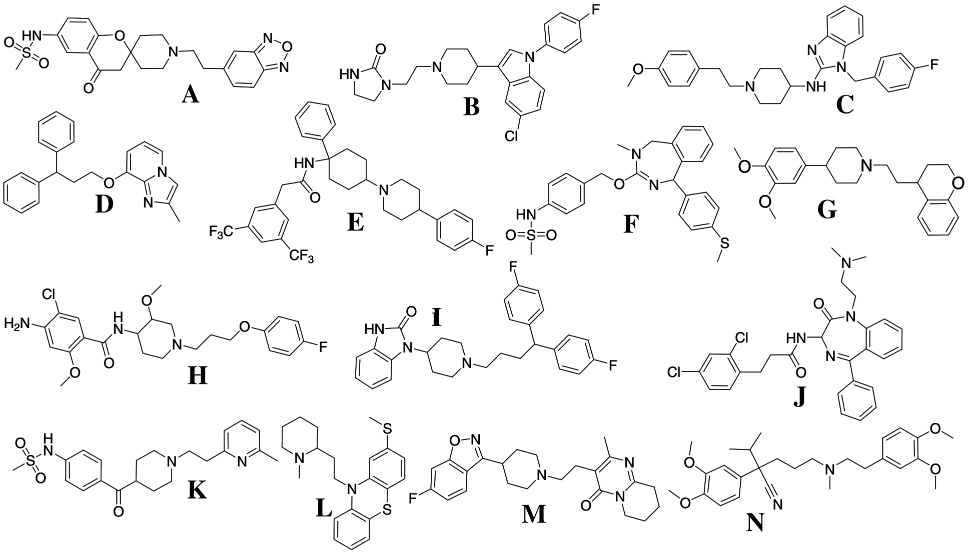


Fig. 5 Examples of compounds from the active data set. All the data given below is averaged after the correction procedure over what could be found in the literature for each compound. (**A**) IC50 = 1.4 nM [40]. (**B**) Sertindole—IC50 = 3 nM [18, 41–43]. (**C**) Astemizole—IC50 = 3.7 nM. (**D**) IC50 = 4.0 nM [44]. (**E**) IC50 = 6.8 nM [45]. (**F**) IC50 = 6.8 nM [46]. (**G**) RP58866—IC50 = 7.1 nM [47, 48]. (**H**) Cisapride—IC50 = 11 nM [21, 49–53]. (**I**) Pimozide—19 nM [11, 18, 21, 54]. (**J**) IC50 = 32 nM [55]. (**K**) E-4031—IC50 = 45 nM [11, 18, 56–59]. (**L**) Thioridazine—IC50 = 51 nM [18]. (**M**) Risperidone—IC50 = 100 nM [18, 41, 60]. (**N**) Verapamil—IC50 = 190 nM [61, 62]

from a patch clamp assay were extracted from the literature: see Fig. 5 for examples of the most potent compounds. Since the data set was constructed from measurements taken from many different laboratories under many different assay types and conditions there should be genuine concern about the internal consistency of the data. In order to ascertain the consistency and limits of the data set, compounds for which data could be found measured independently in multiple laboratories were closely examined. When two laboratories use the same assay format, particularly when they express the hERG channel in the same cell line, the inter-laboratory variation is usually minimal. Table 1 lists the compounds we found with measured hERG IC50s in different laboratories but with the same cell line. When measured in the same cell line, the inter-laboratory variation in hERG IC50s is typically a factor of 3–5, i.e., the pIC50s vary by about 0.5–0.7 log units. The biggest exception in our data set is with Sildenafil in which case there are two reported hERG IC50s in HEK293 cells: 100 μM [71] and 3 μM [18]. Thus even with essentially identical assays there is the potential for large differences in measured binding.

Table 1 Inter-laboratory variation in hERG IC50s

Compound	IC50s (nM)	Assay system	References
Amiodarone	3,800, 9,800	XO	[63, 64]
Astemizole	48, 69, 480	XO	[65–67]
Cisapride	7, 15	HEK	[49, 50]
	16, 5	CHO	[21, 51]
Dofetilide	10, 12, 15, 59	HEK	[11, 42, 50, 68]
	125, 160, 320	XO	[59, 69, 70]
E-4031	18, 77	HEK	[18, 56]
Sertindole	14, 15	HEK	[18, 42]
Sildenafil	3,300, 100,000	HEK	[18, 71]
Terfenadine	56, 204, 213	HEK	[18, 50, 72]
	250, 300, 330, 350, 431	XO	[65–67, 73, 74]

Examples of the inter-laboratory variation in hERG pIC50 when using the same expression system. The Assay System refers to the cell type used to express the hERG potassium channel. XO = *Xenopus Oocytes*, HEK = Human Embryo Kidney cells, and CHO = Chinese Hamster Ovary cells. Only measurements from different laboratories were compared to create this table

The inter-laboratory variation is much greater when measurements are made with assays using different cell lines. As a first example consider the data for Dofetilide given in Table 1: there is approximately a 10-fold difference between the hERG IC50s when measured in HEK cells compared to *Xenopus Oocytes*. As a second example, E-4031 (compound **K** of Fig. 5) has a large range of measured and published hERG IC50s: 7.7 nM [56] and 18 nM [18] in HEK293 cells, 397 nM [57] in guinea pig ventricular myocytes, 910 nM [58] in rabbit ventricular muscle, and 3,500 nM [59] in *Xenopus Oocytes*. In the worst case the measured IC50s vary by as much as a factor of 500.

The examples in the preceding paragraph offer a rather dismal view of the inter-assay variation and call into

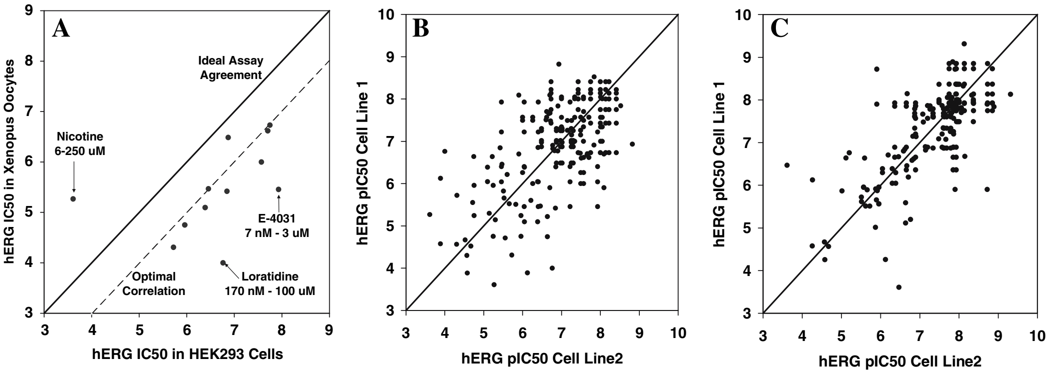


Fig. 6 The assessment of inter-assay variation in the data set. (**A**) This shows that while there are significant differences between measurements in HEK293 cells and *Xenopus Oocytes* the differences are for the most part a consistent shift from one assay to the other. The hERG IC50s measured in HEK cells are on average 1 log unit more potent than when measured in *Xenopus Oocytes*. (**B**) This shows the overall inter laboratory variation with no correction. To create this plot a set of compounds that were measured in at least two different assays were assembled. By different assays we mean different cell types were used for expression of the hERG channel. The compounds in this set were then plotted twice once with measurement 1 as the *x*-axis and measurement 2 as the *y*-axis and once with measurement 2 as the *x*-axis and measurement 1 as the *y*-axis. Plotting each molecule twice avoids an arbitrary designation as to which measurement should be which axis. In total this set consisted of 103 pairs of measurements. With no correction, 12 of the pairs of measurements differed by more than 2 log units, 17 differed by between 1 and 2 log units, 30 differed between 0.5 and 1 log units and the remaining 44 differed by less than 0.5 log units. (**C**) This shows the inter-laboratory variation after the correction. With the correction, only two of the pairs of measurements differed by more than 2 log units, 14 differed by between 1 and 2 log units, 25 differed by between 0.5 and 1 log units, and the remaining 62 differed by less than 0.5 log units. Thus the correction diminishes the inter-laboratory variation, but we still expect a fair number of outliers in this data set. For this reason we consistently report mean absolute errors rather than the more commonly reported root mean square errors. Finally, for compounds with multiple measurements, the mean pIC50 value over all correct measurements was used. Any compound that had a standard deviation of greater than 1 log unit was eliminated from the data set. There were only two compounds that were eliminated via this criterion

question whether data from different assays can be compared. A close examination of the data, however, showed that the variation was not entirely random scatter and to a certain extent could be corrected. Figure 6a shows those compounds for which measured hERG inhibition has been published both in an assay using HEK cells and in an assay using *Xenopus Oocytes*. For the majority of the data the correlation across the two assays is reasonable but the compounds generally prove to be 10-fold more potent when HEK cells are used than when *Xenopus Oocytes* are used. This consistent difference in measured potency is easily corrected. Thus the data were corrected by creating a constant adjustment factor for each cell type so that it best agreed with the corresponding data in HEK cells. After this correction, the mean absolute error for the hERG IC50's measured in different laboratories and using different cell lines was 0.6 log units (see Fig. 6b). This number should be taken as the best any model should fit this data. Any model that fits this data with a mean absolute error less than 0.6 log units is at least in part fitting noise and consequently is over fit. After the correction, the variation in the data was less than 1 log unit 84% of the time and greater than 2 log units 2% of the time. Thus even with the correction there are still instances with large differences in measured binding between laboratories, and thus we should expect that this data set has a fair number of outliers. This highlights the need during model development to use robust techniques. Note that in general we report mean absolute errors rather than the more common root mean square errors because the latter is overly effected by outliers.

In addition to the 190 compounds with reported hERG IC50s from patch clamp assays, a second set of 317 compounds was assembled. This data set consists of compounds with a high likelihood of showing little or no hERG inhibition but for which no measured IC50 was available. These inactive compounds were those that had been shown to be inactive in a patch clamp assay as well as drugs that have been on the market for some time and have not shown cardiotoxicity. For these compounds, we assumed that the hERG IC50 was greater than 30 μM, i.e., pIC50 less than 4.5. While this is certainly true for the majority of this second set of compounds, there is a good chance that a small number of these compounds have significant hERG activity, particularly in the 1–10 μM range. This again highlights the need for robust model building procedures.

Model building

The data set and the iterative procedure described above (Fig. 4) were used to build a number of models to predict hERG inhibition. To test the dependence of the final models on the initial alignment 100 models were created with different initial alignments. To create initial one-dimensional alignments, a subset of 8–20 compounds of the training set with IC50 < 200 nM were randomly selected. Subset alignments were created using the methods described previously [30]. The full alignment was then generated by aligning all members of the training set to the subset alignments again using the methods described previously [30]. While minor differences were seen in the final models depending on the initial alignments, the predictions on the test set typically showed high correlation, $R^2 > 0.9$, between models. Since these differences were small compared to the estimated error in the data itself, we believe the procedure is essentially independent of the initial alignment. All subsequent models were built based on a subset alignment of ziprasidone [18, 41], verapamil [61, 62], thioridazine [18, 41], sertindole [18, 41, 42], risperidone [18, 41], pimozide [18, 41], loratadine [65, 72], haloperidol [18, 21, 75], E-4031 [11, 18], and cisapride [21, 49–51].

Next, to thoroughly determine the optimal value of the ridge coefficient (γ of Eq. 2) and to test the potential for over fitting due to the refinement of the alignment during the cyclic model building process, models were generated for a range of values of the ridge coefficient from 10^{-3} to 10^5 using the cyclic procedure and using just the starting alignment. Figure 7 compares the fit to the training data (black-static, red-cyclic) and the quality of the predictions on the test set (purple-static, blue-cyclic) for both the static model building process and the cyclic model building process. For the static model building process we see that we get ‘‘near perfect’’ fits to the data for low values of the ridge coefficient, i.e., high numbers of free parameters. As is always the case, near perfect fits to data result in nearly useless models. This is indicated by the fact that the performance on the test set is very poor for low values of the ridge coefficient. As the ridge coefficient increases, the fit to data becomes worse but the predictions on the test set become much better. For values above 500 the performance on the training and test set are close though there remains a noticeable gap between the two.

For the cyclic model building procedure with low values of the ridge coefficient, there is a large gap between the models' predictive capability and ability to fit the data indicating that the procedure is overfitting the data. Unlike the static model building procedure, however, the fit to data does not become ‘‘near perfect’’ as the ridge coefficient becomes very small. We believe that this occurs because with too many free variables the cyclic procedure is unable to converge on the optimal model. For intermediate values of the ridge coefficient (10–100) with the cyclic procedure the performance on the test and training set is nearly identical (mean absolute error ~0.6) whereas with the static model building procedure there is a wide discrepancy between the performance on the test and training sets: for a ridge coefficient of 10 the mean absolute error on the training set is 0.4 whereas it is 1.1 on the test set. Thus while it may seem that the refinement of the alignment during model building increases the potential for overfitting, it in fact diminishes the extent of overfitting.

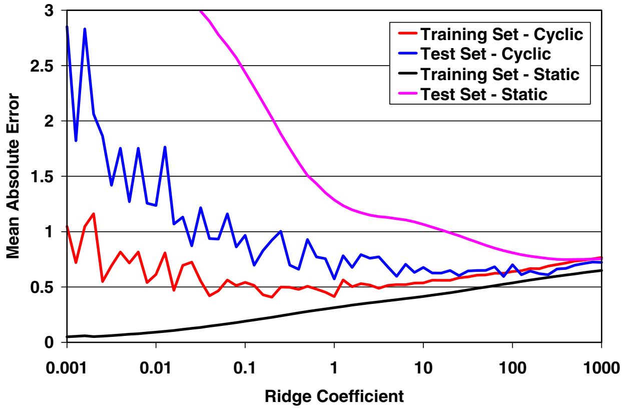


Fig. 7 The mean absolute error in the training and test set as a function of the ridge coefficient. The *purple curve* corresponds to the test set for the static model building procedure. The *blue curve* corresponds to the test set for the dynamic model building procedure. The *red curve* corresponds to the training set for the dynamic model building procedure. The *black curve* corresponds to the training set for the static model building procedure. For low values of the ridge coefficient (<1) there is a wide discrepancy between the mean absolute error on the test and training sets with both the static and cyclic modeling building procedures though the effect is much more pronounced with the static model building procedure. This indicates that the procedure is overfitting the training set. We believe the oscillation in the errors in the test and training sets with the cyclic procedure arise at low values of the ridge coefficient because there are many more local minima in the fitting procedure making it more difficult to converge on the optimal model. We believe this is an additional indication of over fitting. Note that for this and other reported statistics we use mean absolute error rather than the more traditional rms deviation or R^2 because the mean absolute error is a robust statistic whereas neither the rms deviation nor R^2 is

Based on the analysis shown in Fig. 7, the optimal value of the ridge coefficient is near 100 where both the training and test set exhibit a mean absolute error of 0.62. This value of 0.62 for a mean absolute error compares favorably to the optimal value determined by examining the inter-laboratory variation in the measured data (see Fig. 6). For the value of 100 for the ridge coefficient the calculated versus measured activities of the compounds in the test and training sets are shown in Fig. 8. The scatter shown in Fig. 8 is comparable to that shown in Fig. 6c where different assay systems were compared. For different assay systems, after the correction, we estimated a mean absolute error of 0.6. On both the training and test sets the mean absolute error is approximately 0.62. Thus statistically the model is fit to the limits of the data. For all subsequent discussion, the model built with a ridge coefficient of 100 is used.

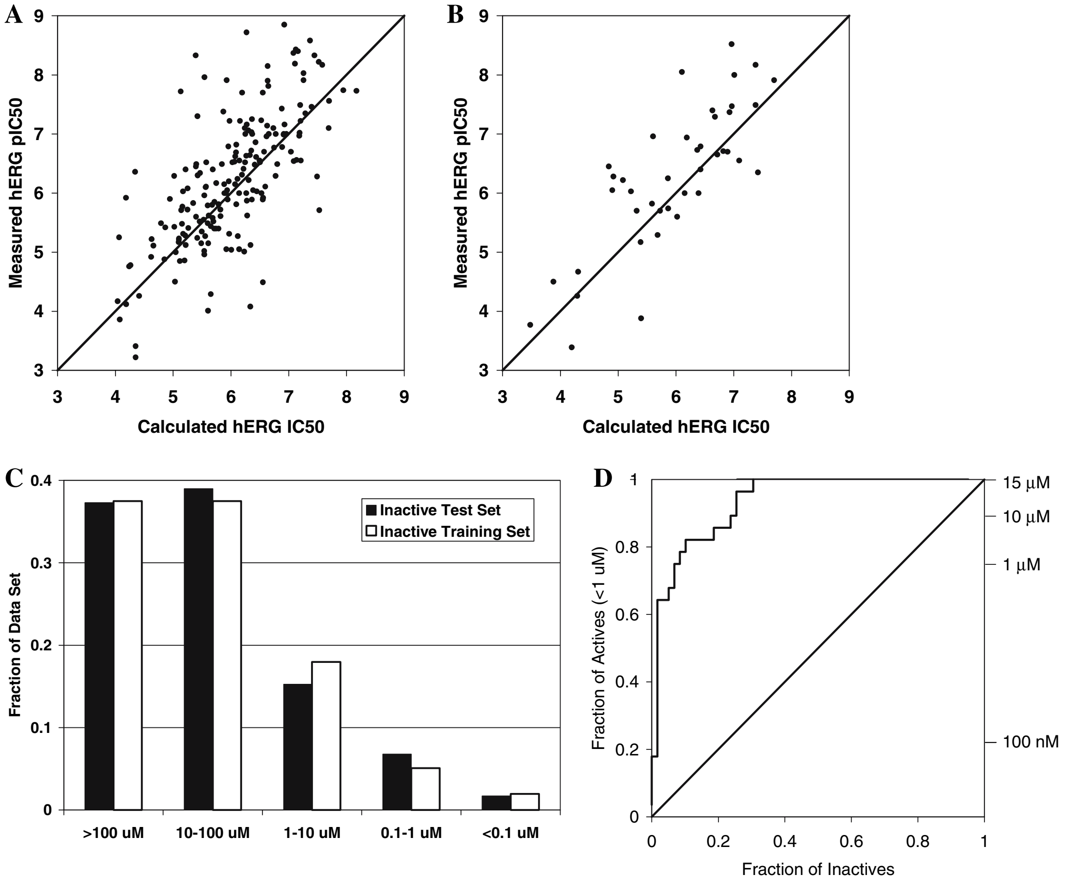


Fig. 8 Model performance on the training and test set. (**A**) The performance on the training set (189 compounds). The mean absolute error on the training set is 0.62. About 52% of the compounds have an error of less than 0.5 log units, 80% have an error of less than 1 log unit, while 4% have an error greater than 2 log units. (**B**) The performance on the test set (41 compounds). The mean absolute error on the test set is 0.62. 54% of the compounds in the test set have a prediction error of less than 0.5 log units, 78% have an error of less than 1 log unit, while no compound of the test set has an error greater than 2 log units. (**C**) The performance on the negative data set. The training set (256 compounds) is in black and the test set (59 compounds) is in white. For the inactive compounds we regressed to IC50 ≥ 30 μM, i.e., pIC50 ≤ 4.5. (**D**) The receiver operating characteristic (ROC) curve using the members of the test set with IC50 ≤ 1 μM as the actives (28 compounds) and inactive members of the test set (59 compounds) as the inactives. The area under the ROC curve is 0.93

An examination of the final model

There are two ways to more closely examine the final hERG model. The first is simply to analyze the curves for each of the atomic descriptors (see Fig. 9). For this it is critical to recall that variation in the curves is penalized to decrease the number of effective free variables. Thus a curve will vary only when supported by sufficient data. Also, it is important that all the atomic descriptors are positive and thus positive/negative values for the curves indicate positive/negative contributions to the calculated activity when atoms of the corresponding type are positioned at the particular location in the model. A second approach to examining the final model is derived from the fact that the model assigns to each atom a contribution to the activity. This information can be displayed on a molecule's structure to indicate which portions of a particular molecule contribute significantly to the activity (see Fig. 10).

The model building procedure isolates two of the atom descriptors as being key positive contributors to the hERG activity of small molecules (see Fig. 9). The first is the aromatic carbon E-state. For aromatic carbons there are two clear regions that contribute positively to the hERG activity (see Fig. 9b). The first region occurs between $X = -5$ and –2.5 peaking at –3.5 and the second region occurs between 4.5 and 9.5 with a peak at 6.0. As the units for X are bonds, the model indicates that the optimal separation for the two phenyl rings is 9–10 bonds. A typical aromatic carbon in one of these two regions will contribute 0.2–0.3 log units to the molecule's calculated hERG activity. Thus an optimally placed phenyl ring can

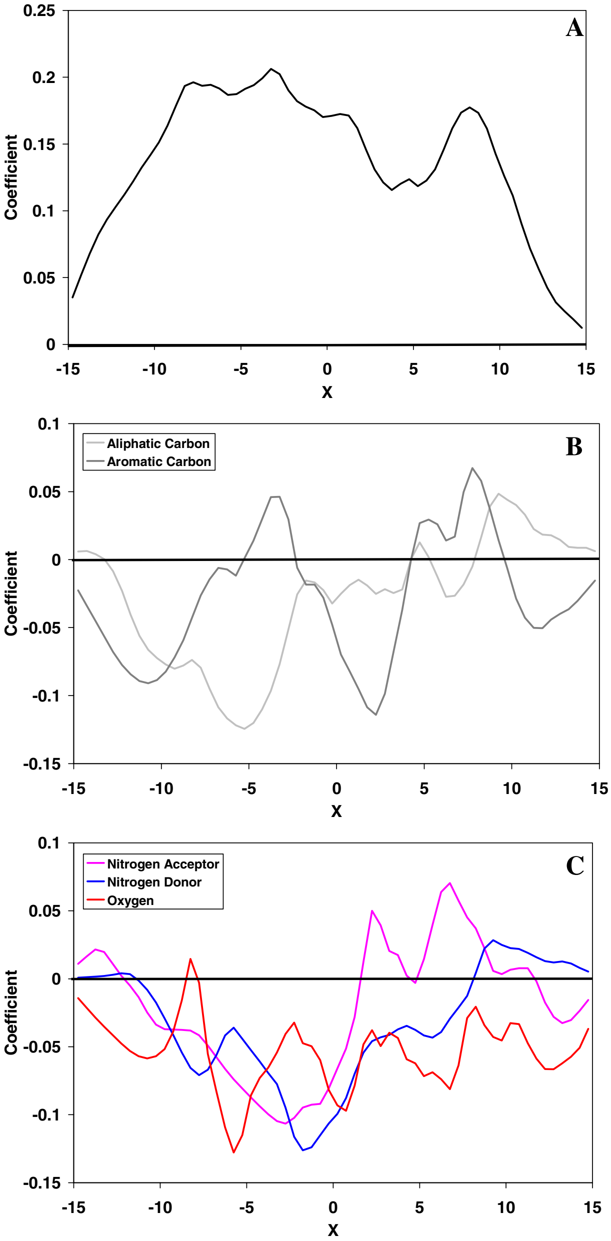


Fig. 9 The final hERG model. The units on the x-axis are essentially bond units. (**A**) The contribution of size to hERG activity. This curve indicates that there is a large contribution to hERG activity that arises simply from a molecule's size. From this analysis, on average an atom contributes 0.1 log units to hERG activity. (**B**) The contribution of hydrophobic groups to hERG activity. *Dark gray* corresponds to aromatic carbons, *light gray* to aliphatic carbons. (**C**) The contribution of polar functionality to hERG activity. *Purple* corresponds to nitrogen acceptors, *blue* to nitrogen donors and *red* to oxygen

contribute as much as 1.5 log units. The presence of two large areas favorable for aromatic rings is consistent with alanine-scanning mutagenesis which has shown that residues Y652 and F656 of the S6 domain make critical interactions with high affinity hERG binders [76, 77]. Furthermore, homology modeling and docking efforts suggest that these two aromatic residues form critical PI–PI stacking interactions with pendant aromatic rings of hERG blockers [21, 27, 28].

The second atom property that contributes heavily to the calculated hERG activity is the nitrogen acceptor. This descriptor includes any nitrogen that has an accessible lone pair in its neutral form. Thus this includes functionality such as an amine, pyridine nitrogen etc. There are two clear peaks in the curve corresponding to this descriptor. The first peak centered near $X = 2.5$ corresponds to the basic center frequently found in potent hERG inhibitors and central to most of the examples given in Fig. 5. A tertiary amine positioned here will contribute approximately 0.5 log units to the molecule's overall calculated hERG activity. Near this first peak there is a strong penalty for a nitrogen donor. The result is that tertiary amines are strongly preferred over secondary or primary amines. This peak in the nitrogen acceptor curve and the two aromatic regions are consistent with the classic hERG pharmacophore consisting of a tertiary amine between two phenyl rings. This pharmacophore is most clearly illustrated in Fig. 10a with verapamil but also is illustrated in Fig. 10b and c with pimozide and E-4031. It should be noted that functionality other than an amine can contribute in this position. The basic iso-urea nitrogen of compound **F** of Fig. 5 is aligned near $X = 2.5$ and contributes approximately 0.4 log units to its calculated hERG activity, see Fig. 10d. As with the importance of the aromatic rings, the presence of a site for a basic center approximately midway between the two aromatic sites is consistent with other models [17, 18, 27, 28]. The exact nature of the interaction between the basic center and the hERG channel is less clear but has been suggested to be involved in a PI-cation interaction [17] or simply that it is complimentary to the negative electrostatic field found in the pore of the hERG channel [28].

The second peak in the nitrogen acceptor curve (centered at $X = 6.0$) occurs in the second aromatic region. This second peak often corresponds to heterocyclic basic centers such as a pyridine. This putative interaction appears to be a feature unique to this model and is illustrated in both Fig. 10e with compound **D** of Figs. 5 and 10c with the pyridine nitrogen of E-4031.

Beyond the nitrogen acceptors, most polar functionality significantly decreases hERG activity (Fig. 9c). As an example the ketone oxygen of E-4031 (Fig. 10c) decreases its calculated activity by 0.5 log units. Including polar functionality is a well known technique to decrease hERG binding [78]. This could be due to the fact that the hERG pore is fairly hydrophobic and offers little opportunity for interactions with polar functionality on small molecules. The lone example of a positive contribution from polar functionality other than the nitrogen acceptors is a small

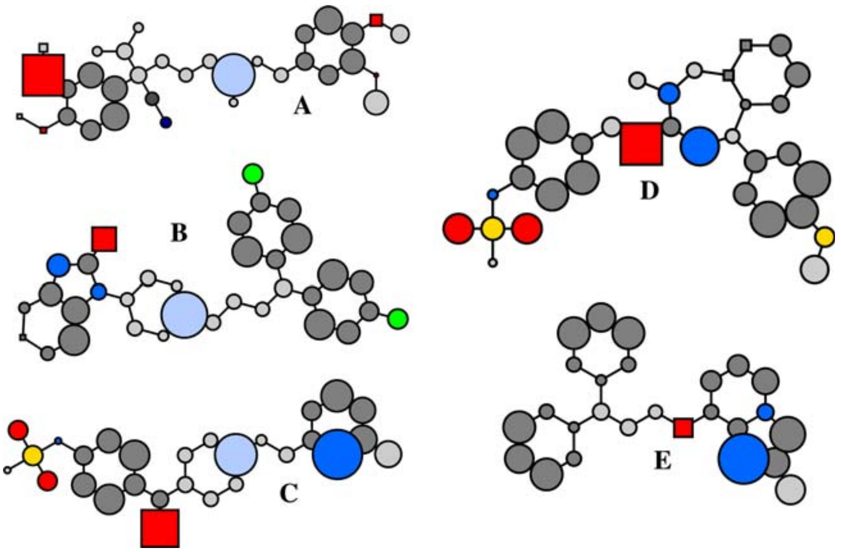


Fig. 10 The model depiction. The atoms are sized by the amount they contribute to the predicted hERG inhibition. The *circles* are atoms predicted to contribute positively, i.e., they interact favorably with the hERG channel. The *squares* are atoms that are predicted to contribute negatively, i.e., decrease the predicted hERG binding. The atoms are colored according to standard atom types: *gray*—carbon, *blue*—nitrogen, *red*—oxygen, *green*—fluorine. The *shade* indicates the hybridization. The *light colors* are sp^3 hybridized atoms. The *dark colors* are sp^2 or sp^1 hybridized atoms. In all cases, the molecules are oriented left to right as they are optimally positioned relative to the model. All structures are shown in the same configuration as in Fig. 5. (**A**) Verapmil—Compound **N** of Fig. 5. (**B**) Pimozide—Compound **I**. (**C**) E-4031—Compound **K**. (**D**) Compound **F**. (**E**) Compound **D**

preference for an oxygen near $X = -8.0$. This peak in the oxygen curve corresponds to the sulfonamide oxygens of compounds such as that found in E-4031 (Fig. 10c) and compound **F** (Fig. 10e). While this is a small peak and should not be over interpreted it could be consistent with the putative interactions proposed by Aronov [26] for neutral hERG blockers.

Discussion

The data set used here highlights the need to develop new approaches to handle the complex data sets generated in drug discovery and development programs. For this data set the inactive compounds played an important role in refinement. A close analysis of how the alignments of various compounds moved during the iterative model building process revealed that the position of the most potent compounds quickly stabilized whereas the less potent and inactive compounds continued to find new ways to align relative to the rest of the compounds. Ultimately, this led to the constraint that the inactives have calculated IC50s ≥ 30 μM in all possible alignments to the model rather than in a single pre-chosen alignment which meant that the information in the inactives had a much greater impact on the final model when compared to a model built with a static alignment. Since the goal in drug discovery programs is to design compounds devoid of hERG activity, the inactive compounds are particularly important.

Additionally, robust model building techniques should be the norm for QSAR applications rather than the exception. While the data set used here may seem to have a disproportionate number of outliers, nearly all data sets have the potential for outliers. Clearly, there can be large inter-laboratory differences in measured hERG inhibition, but there is potential for outliers of other types as well. The majority of the compounds in this data set inhibit the hERG potassium channel by binding directly in the channel pore. It is possible, however, that some of these compounds produce this activity by interfering with protein trafficking [79]. These compounds are also outliers. In addition, there is the potential that even within those compounds that directly block the hERG channel there are multiple binding sites or different states, such as open or close, of the hERG channel. Thus, outliers are likely to be present even in the absence of inter-laboratory variation.

Computational models can play and in most places are playing a role in the cascade of tests to ensure desirable ADME and toxicity properties. While not as predictive as their in vitro counterparts, computational models can still play a valuable role as they can help prioritize compounds prior to synthesis thereby decreasing the amount of effort put into synthesizing compounds that have liabilities, such as hERG inhibition. Further computational models can be useful in helping to solve experimentally uncovered liabilities by indicating which portions of the small molecule structure are contributing to the unwanted activity and thus how new molecules can be designed that are devoid of the activity. Since the predicted hERG pIC50 with this model is a sum of contributions from the atoms of the molecule it is straightforward to visualize the contributions of the individual atoms and substructures to the predicted IC50. Depicting the model's predictions, as in Fig. 10, can aid scientists in developing hypotheses as to how to solve a hERG liability within a chemical series thereby better complementing the intuition and experience of the individual scientist. Furthermore, as compounds are synthesized based on hypotheses developed via the model, the resulting data will be much more valuable for testing the model's limits and ultimately improving the model.

There are several clear ways that the 1D-QSAR approach could be extended. For example, while the E-state keys were a rather fortunate choice in atomic descriptors no investigation was done into other combinations of atomic descriptors. In particular, having a pK_a estimate for the basic nitrogens would be a logical avenue of investigation. In addition, electrostatic descriptors, such as charge for heteroatoms or heteroatom–hydrogen dipole for hydrogen bond donors would also be reasonable descriptors to test. Additional descriptors could broaden the applicability of this model or this technique to other data sets. A second extension of the 1D-QSAR approach would be to consider pattern recognition techniques other than regression. Possibilities include naïve Bayes models or neural networks which only require binary end points such as active/inactive. Doing so would expand the applicability of the method to a larger number of data sets such as those arising in high throughput screening. These will be active areas of future research.

References

1. Haverkamp W, Breithardt G, Camm AJ, Janse MJ, Rosen MR, Antzelevitch C, Escande D, Franz M, Malik M, Moss A, Shah R (2000) Cardiovasc Res 47:219
2. De Ponti F, Poluzzi E, Montanaro N (2000) Eur J Clin Pharmacol 56:1
3. Redfern WS, Carlsson L, Davis AS, Lynch WG, MacKenzie I, Palethorpe S, Siegl PKS, Strang I, Sullivan AT, Wallis R, Camm AJ, Hammond TG (2003) Cardiovasc Res 58:32
4. Yap YG, Camm AJ (1999) Clin Exp Allergy 29:174
5. De Ponti F, Poluzzi E, Cavalli A, Recanatini M, Montanaro N (2002) Drug Saf 25:263
6. Tang W, Kang J, Wu X, Rampe D, Wang L, Shen H, Li Z, Dunnington D, Garyantes T (2001) J Biomol Screen 6:325
7. Rezazadeh S, Hesketh JC, Fedida D (2004) J Biomol Screen 9:588
8. Cheng CS, Alderman D, Kwash J, Dessaint J, Patel R, Lescoe MK, Kinrade MB, Yu W (2002) Drug Dev Ind Pharm 28:177
9. Chaudhary KW, O'Neal JM, Mo Z-L, Fermini B, Gallavan RH, Bahinski A (2006) Assay Drug Dev Technol 4:73
10. Finlayson K, Sharkey J (2004) In: Yan Z, Caldwell G (eds) Optimization in drug discovery. Humana Press, Totowa, NJ, pp353–368
11. Finlayson K, Turnbull L, January CT, Sharkey J, Kelly JS (2001) Eur J Pharmacol 430:147
12. Diaz GJ, Daniell K, Leitza ST, Martin RL, Su Z, McDermott JS, Cox BF, Gintant GA (2004) J Pharmacol Toxicol Methods 50:187
13. Chiu PJS, Marcoe KF, Bounds SE, Lin C-H, Feng J-J, Lin A, Cheng F-C, Crumb WJ, Mitchell R (2004) J Pharmacol Sci (Tokyo Japan) 95:311
14. Witchel HJ, Milnes JT, Mitcheson JS, Hancox JC (2002) J Pharmacol Toxicol Methods 48:65
15. Lu HR, Vlaminckx E, Van Ammel K, De Clerck F (2002) Eur J Pharmacol 452:183
16. Varro A, Balati B, Iost N, Takacs J, Virag L, Lathrop DA, Csaba L, Talosi L, Papp JG (2000) J Physiol 523:67
17. Cavalli A, Poluzzi E, De Ponti F, Recanatini M (2002) J Med Chem 45:3844
18. Ekins S, Crumb WJ, Sarazan RD, Wikel JH, Wrighton SA (2002) J Pharmacol Exp Ther 301:427
19. Roche O, Trube G, Zuegge J, Pflimlin P, Alanine A, Schneider G (2002) Chembiochem 3:455
20. Keseru GM (2003) Bioorg Med Chem Lett 13:2773
21. Pearlstein RA, Vaz RJ, Kang J, Chen X-L, Preobrazhenskaya M, Shchekotikhin AE, Korolev AM, Lysenkova LN, Miroshnikova OV, Hendrix J, Rampe D (2003) Bioorg Med Chem Lett 13:1829
22. Aronov AM, Goldman BB (2004) Bioorg Med Chem 12:2307
23. Du L-P, Tsai K-C, Li M-Y, You Q-D, Xia L (2004) Bioorg Med Chem Lett 14:4771
24. Zhang H, Hancox JC (2004) Biochem Biophys Res Commun 322:693
25. Aronov AM (2005) Drug Discov Today 10:149
26. Aronov AM (2006) J Med Chem 49:6917
27. Rajamani R, Tounge BA, Li J, Reynolds CH (2005) Bioorg Med Chem Lett 15:1737
28. Farid R, Day T, Friesner RA, Pearlstein RA (2006) Bioorg Med Chem 14:3160
29. Dixon SL, Merz KM Jr (2001) J Med Chem 44:3795
30. Wang N, DeLisle RK, Diller DJ (2005) J Med Chem 48:6980
31. Diller DJ, Lin TH, Metzger A (2005) Curr Top Med Chem 5:953
32. Pargellis C, Tong L, Churchill L, Cirillo PF, Gilmore T, Graham AG, Grob PM, Hickey ER, Moss N, Pav S, Regan J (2002) Nat Struct Biol 9:268
33. Hall LH, Mohney B, Kier LB (1991) Quant Struct Act Relat 10:43
34. Hall LH, Mohney B, Kier LB (1991) J Chem Inf Comput Sci 31:76
35. Kier LB, Hall LH (1990) Pharm Res 7:801
36. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Data mining, inference, and prediction. Springer-Verlag, New York
37. Press WH, Teulkolsky SA, Vetterling WT, Flannery BP (1997) Numerical recipes in C, 2 edn. Cambridge University Press, Cambridge, p 994
38. Vapnik VN, (1999) Support vector method for function estimation, US05950146
39. Vapnik VN, Golowhich SE (2001) Support vector method for function estimation, US06269323
40. Lynch JJ Jr, Wallace AA, Van der Gaag LH, Baskin EP, Bear CM, Gehret JR, Kothstein T, Stupienski RF, Appleby SD, Sanguinetti MC et al (1993) J Pharmacol Exp Ther 265:720
41. Kongsamut S, Kang J, Chen X, Roehr J, Rampe D (2002) Eur J Pharmacol 450:37

42. Rampe D, Murawsky MK, Grau J, Lewis EW (1998) J Pharmacol Exp Ther 286:788
43. Thomsen MB, Volders PGA, Stengl M, Spaetjens RLHMG, Beekman JDM, Bischoff U, Kall MA, Frederiksen K, Matz J, Vos MA (2003) J Pharmacol Exp Ther 307:776
44. Beatch GN, Liu Y, Plouvier BMC (2001) Preparation of imidazo [1,2-a]pyridine ether compounds as ion channel modulators. Cardiome Pharma Corp., Canada, p 111
45. Cooper LC, Carlson EJ, Castro JL, Chicchi GG, Dinnell K, Di Salvo J, Elliott JM, Hollingworth GJ, Kurtz MM, Ridgill MP, Rycroft W, Tsao K-L, Swain CJ (2002) Bioorg Med Chem Lett 12:1759
46. Johnson RE, Silver PJ, Becker R, Birsner NC, Bohnet EA, Briggs GM, Busacca CA, Canniff P, Carabateas PM, Chadwick CC et al (1995) J Med Chem 38:2551
47. Jurkiewicz NK, Wang J, Fermini B, Sanguinetti MC, Salata JJ (1996) Circulation 94:2938
48. Valentin J-P, Hoffmann P, De Clerck F, Hammond TG, Hondeghem L (2004) J Pharmacol Toxicol Methods 49:171
49. Mohammad S, Zhou Z, Gong Q, January CT (1997) Am J Physiol 273:H2534
50. Rampe D, Roy ML, Dennis A, Brown AM (1997) FEBS Lett 417:28
51. Walker BD, Singleton CB, Bursill JA, Wyse KR, Valenzuela SM, Qiu MR, Breit SN, Campbell TJ (1999) Br J Pharmacol 128:444
52. Carlsson L, Amos GJ, Andersson B, Drews L, Duker G, Wadstedt G (1997) J Pharmacol Exp Ther 282:220
53. Drolet B, Khalifa M, Daleau P, Hamelin BA, Turgeon J (1998) Circulation 97:204
54. Kang J, Wang L, Cai F, Rampe D (2000) Eur J Pharmacol 392:137
55. Selnick HG, Liverton NJ, Baldwin JJ, Butcher JW, Claremon DA, Elliott JM, Freidinger RM, King SA, Libby BE, McIntyre CJ, Pribush DA, Remy DC, Smith GR, Tebben AJ, Jurkiewicz NK, Lynch JJ, Salata JJ, Sanguinetti MC, Siegl PK, Slaughter DE, Vyas K (1997) J Med Chem 40:3865
56. Zhou Z, Gong Q, Ye B, Fan Z, Makielski JC, Robertson GA, January CT (1998) Biophys J 74:230
57. Sanguinetti MC, Jurkiewicz NK (1990) J Gen Physiol 96:195
58. Toyama J, Kamiya K, Cheng J, Lee JK, Suzuki R, Kodama I (1997) Circulation 96:3696
59. Ficker E, Jarolimek W, Kiehn J, Baumann A, Brown AM (1998) Circ Res 82:386
60. Magyar J, Banyasz T, Bagi Z, Pacher P, Szentandrassy N, Fueloep L, Kecskemeti V, Nanasi PP (2002) Naunyn Schmiedebergs Arch Pharmacol 366:350
61. Waldegger S, Niemeyer G, Morike K, Wagner CA, Suessbrich H, Busch AE, Lang F, Eichelbaum M (1999) Cell Physiol Biochem 9:81
62. Zhang S, Zhou Z, Gong Q, Makielski JC, January CT (1999) Circ Res 84:989
63. Kamiya K, Nishiyama A, Yasui K, Hojo M, Sanguinetti MC, Kodama I (2001) Circulation 103:1317
64. Kiehn J, Thomas D, Karle CA, Schols W, Kubler W (1999) Naunyn Schmiedebergs Arch Pharmacol 359:212
65. Taglialatela M, Pannaccione A, Castaldo P, Giorgio G, Zhou Z, January CT, Genovese A, Marone G, Annunziato L (1998) Mol Pharmacol 54:113
66. Chachin M, Katayama Y, Yamada M, Horio Y, Ohmura T, Kitagawa H, Uchida S, Kurachi Y (1999) Eur J Pharmacol 374:457
67. Suessbrich H, Waldegger S, Lang F, Busch AE (1996) FEBS Lett 385:77
68. Snyders DJ, Chaudhary A (1996) Mol Pharmacol 49:949
69. Weerapura M, Hebert TE, Nattel S (2002) Pflugers Arch 443:520
70. Lees-Miller JP, Duan Y, Teng GQ, Duff HJ (2000) Mol Pharmacol 57:367
71. Geelen P, Drolet B, Rail J, Berube J, Daleau P, Rousseau G, Cardinal R, O'Hara GE, Turgeon J (2000) Circulation 102:275
72. Crumb WJ Jr (2000) J Pharmacol Exp Ther 292:261
73. Ko CM, Ducic I, Fan J, Shuba YM, Morad M (1997) J Pharmacol Exp Ther 281:233
74. Roy M, Dumaine R, Brown AM (1996) Circulation 94:817
75. Suessbrich H, Schonherr R, Heinemann SH, Attali B, Lang F, Busch AE (1997) Br J Pharmacol 120:968
76. Chen J, Seebohm G, Sanguinetti MC (2002) Proc Natl Acad Sci USA 99:12461
77. Mitcheson J, Perry M, Stansfeld P, Sanguinetti MC, Witchel H, Hancox J (2005) Novartis Found Symp 266:136
78. Jamieson C, Moir EM, Rankovic Z, Wishart G (2006) J Med Chem 49:5029
79. Eckhardt LL, Rajamani S, January CT (2005) Br J Pharmacol 145:3