Development and Evaluation of an in Silico Model for hERG Binding

Minghu Song and Matthew Clark*

Locus Pharmaceuticals, 512 Township Line Rd., Blue Bell, Pennsylvania 19422

Received August 1, 2005

It has been recognized that drug-induced QT prolongation is related to blockage of the human ether-a-go-go-related gene (hERG) ion channel. Therefore, it is prudent to evaluate the hERG binding of active compounds in early stages of drug discovery. In silico approaches provide an economic and quick method to screen for potential hERG liability. A diverse set of 90 compounds with hERG IC_{50} inhibition data was collected from literature references. Fragment-based QSAR descriptors and three different statistical methods, support vector regression, partial least squares, and random forests, were employed to construct QSAR models for hERG binding affinity. Important fragment descriptors relevant to hERG binding affinity were identified through an efficient feature selection method based on sparse linear support vector regression. The support vector regression predictive model built upon selected fragment descriptors outperforms the other two statistical methods in this study, resulting in an r^2 of 0.912 and 0.848 for the training and testing data sets, respectively. The support vector regression model was applied to predict hERG binding affinities of 20 in-house compounds belonging to three different series. The model predicted the relative binding affinity well for two out of three compound series. The hierarchical clustering and dendrogram results show that the compound series with the best prediction has much higher structural similarity and more neighbors of training compounds than the other two compound series, demonstrating the predictive scope of the model. The combination of a QSAR model and postprocessing analysis, such as clustering and visualization, provides a way to assess the confidence level of QSAR prediction results on the basis of similarity to the training set.

INTRODUCTION

Blockage of the human ether-a-go-go-related gene (hERG) potassium ion channel is believed to be the major cause of drug-induced QT syndrome, which can lead to sudden death.[1] Several drugs including terfenadine, astemizole, grepafloxicin, terodiline, droperidol, lidoflazine, sertindole, levomethadyl, and cisapride have been withdrawn from the market or severely restricted in availability as a result of their association with drug-induced long QT syndrome. It is, therefore, prudent to screen drug candidates for hERG blockage liability at the early stages of drug discovery.

Current in vitro assays for hERG binding are technically demanding, costly, and labor-intensive.[2] Several in silico models derived from available in vitro data have been developed using 3D and 2D approaches to provide a supplement assessment for hERG liability. These in silico models are listed in Table 1. A structure-based study for hERG channel blockage, described by Mitcheson and co-workers, was based on alanine-scanning mutagenesis and a homology model built from the closed form of the K^+ channel structure.[3] Ekins and co-workers developed a pharmacophore model containing four hydrophobic and one positively charged feature using in vitro data of 15 drugs.[4] Their pharmacophore model was further validated using an additional 22 testing compounds. Cavalli and co-workers developed another pharmacophore model, along with a 3D QSAR (CoMFA) predictive model.[5] Their model included three aromatic moieties and one central positively charged tertiary amine. Another 3D QSAR (CoMSIA) model has recently been constructed by Pearlstein and co-workers at Aventis, and the derived structure−activity relationship was further interpreted by a homology model based on the crystal structure of an open MthK potassium channel.[6] Although these structure-based and 3D QSAR models provide impressive insight for the interaction between drugs and the hERG ion channel, their application is limited by the lack of a hERG crystal structure, effective techniques for the sampling of active conformations, and the need for 3D molecular alignment of diverse structures. Therefore, a variety of 2D QSAR models have been used as screening tools to estimate the hERG binding affinity of drug candidates. For example, Qikprop calculates topological and physicochemical whole molecular properties derived from Monte Carlo simulations and describes the relationship between hERG binding affinity and computed descriptors using multiple linear regression (MLR).[7] Roche et al. used various machine learning techniques, ranging from MLR to modern multivariate analysis techniques, such as self-organizing maps, principal component analysis, partial least squares (PLS), and supervised neural networks, to find appropriate molecular descriptors and then built predictive models.[8] In that study, the most accurate model, based on neural networks, produced a classification accuracy of 93% and 71% for nonblocking and blocking agents, respectively. Table 1 lists recent publications of in silico models for hERG binding affinity.

In this study, we developed 2D QSAR models for the quick estimation of hERG binding affinities. We computed fragment fingerprints as QSAR descriptors for a set of training compounds and generated in silico predictive models based on three different statistical algorithms: support vector

* Corresponding author e-mail: mclark@locuspharma.com.

10.1021/ci050308f CCC: $33.50
Published on Web 12/03/2005

Table 1. Recent Publications of in Silico Modeling of hERG Blockers

references	methods employed
	Homology Modeling
Mitcheson et al., 2000[3]	alanine-scanning mutagenesis, homology modeling using a closed K^+ channel structure (Kcsa), and docking analysis
Pearlstein et al., 2003[6]	homology model based on an open K^+ channel structure (MthK) and a CoMSIA model
	Pharmacophore Modeling
Cavalli et al., 2002[5]	conformational search and clustering using MacroModel and 3D-QSAR (CoMFA) models
Ekins et al., 2002[4]	pharmacophore model using Catalyst
Zolotoy, et al. 2003[9]	conformational analysis and ab initio calculation to understand the physicochemical determinants for hERG blockers
	2D QSAR (Classification and Regression)
Qikprop[7]	QikProp descriptors and multiple linear regression
Roche et al. 2002[8]	a variety of QSAR descriptors and statistical classification methods
Keseru et al. 2003[10]	hologram QSAR descriptors and partial least squares
Aronov et al. 2004[11]	classification model based on a 2D topological similarity filter and 3D pharmacophore ensemble
Bains et al. 2004[12]	fragment-based and experimental descriptors and an evolutionary algorithm

regression (SVR), PLS, and random forest (RF). A two-step modeling procedure was employed in this study: first, important fragment descriptors were identified using feature selection methods based on sparse linear SVR; then, linear SVR, PLS, and random forest models were constructed based upon these selected descriptors. The performance of these QSAR models was subsequently evaluated by leave-one-out cross validation and two testing data sets: a testing set containing 19 additional drugs not used in the model construction and a set of 20 proprietary compounds.

DATA SET

In vitro hERG inhibition data (IC_{50}) for 90 structurally diverse drugs were collected from the literature. Among this set, 19 drugs were randomly selected as testing compounds, and the remaining 71 drugs were used as training data for constructing the QSAR models. Most IC_{50} values were measured in mammalian cells (HEK, CHO, and COS); IC_{50} measurements from nonmammalian cell lines were used in five cases when mammalian cell data were not available. These binding values (IC_{50}) were transformed to pIC_{50} ($-\log IC_{50}$, concentration in molarity) for statistical analysis. The binding data of these compounds as well as their original references are listed in the Supporting Information.

Twenty Locus proprietary compounds with measured hERG binding affinities were used as an additional test of the predictive performance of the models. Their experimental binding affinities were expressed as the percentage of current inhibition determined in voltage-clamped HEK293 cells, measured as I_{kr} current reduction after a steady-state effect had been reached in the presence of the drug relative to the control current before the drug was introduced. Since the binding of in-house compounds was expressed in percentage of inhibition, we converted the percentage binding data for these compounds into a binding affinity constant unit that is compatible with pIC_{50}. This relative binding constant was calculated from the percent binding measurement using eq 1, which has been traditionally used to fit the concentration−response curve and to calculate IC_{50}.

$$\log k_{\text{approximate}} = -\log D + \log \frac{B\%}{1 - B\%} \qquad (1)$$

$\log k_{\text{approximate}}$ is the binding affinity constant approximated from the percentage of binding data, and D is the concentration (1 μM) of the drug used in the assay. In a recent study,[13] the same formula was used to convert the percentage binding data to the equivalent binding affinity constant for plasma protein binding prediction. Since this approximation is quite crude, these converted data were not used for constructing models. Instead, they were only used as additional data to qualitatively assess the ability of the in silico HERG models to differentiate active and inactive in-house compounds.

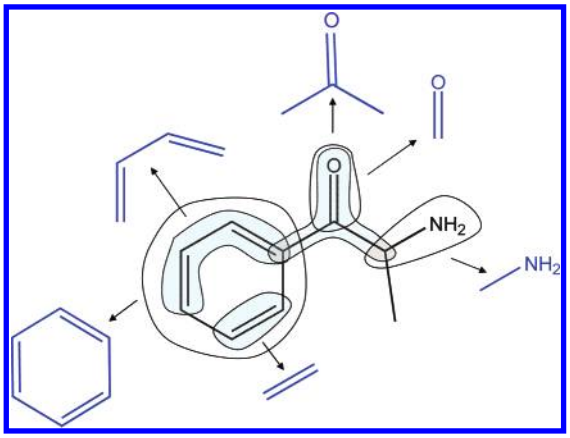


Figure 1. Simplified example of overlapping fragments found in cathinone.

METHODS

Structure Preparation and QSAR Fragment Descriptor Calculation. An initial 2D fragment library was constructed using 261 structurally diverse fragments that frequently appear in bioactive compounds. The algorithms used to compute 2D fragment-based descriptors for the collected compounds with the structures in their neutral states have been recently reported.[14] A simplified example of the overlapping fragments found in cathinone is shown in Figure 1. During substructure searching, if a particular fragment is found in the query molecule, the substructure score is set to the number of occurrences of this fragment in the molecule. This substructure search algorithm has recently been applied to create several ADME (absorption, distribution, metabolism, and excretion) predictive models, such as solubility and log P models.[14]

Feature Selection and Predictive Model Construction. Prior to any statistical modeling, a preprocessing step was employed to discard descriptors having the same value for

all compounds. After preprocessing, 141 fragment descriptors remained. Since the number of remaining descriptors, 141, was still much larger than the number of training compounds, 71, a feature selection procedure was employed to select descriptors relevant to variations in hERG binding affinity. In this study, we used a linear programming formulation of the ν-SVR algorithm, called sparse ν-SVR,[15] to perform feature selection and simultaneously construct the linear SVR predictive models based upon the 45 selected descriptors. Support vector machines (SVMs) are a class of supervised learning algorithms for pattern recognition problems initially proposed by Vapnik in statistical learning theory[16] and further extended asSVR[17] to solve regression problems. Traditional regression algorithms construct models by minimizing the training error, so their predictive performance cannot be guaranteed for novel compounds, especially when the training data itself is noisy. This problem is known as "overfitting". To overcome this problem, instead of minimizing the training error, SVR minimizes a regularized error that controls both the training error and model complexity. This method varies the coefficients for each fragment and removes fragments with low contributions to the model. A brief overview of SVR and the linear formulation of ν-SVR is provided in the Appendix. In addition, the selected descriptors were used to construct QSAR models using two other well-known statistical algorithms, PLS and RF.[18,19] To further test the performance of the QSAR models, scramble testing[20] was performed. In scramble testing, the bioactivity responses, hERG pIC_{50} in this case, of the training data are randomly shuffled and, then, a new training model is constructed using shuffled responses. The predictive ability of these new models is then tested on intact test data. The correlation is computed as 1 − (sum of squares of errors)/(sum of squares deviations from the mean) and, thus, ranges from 1.0 to negative values if the variance of the error is larger than the data variance. The hypothesis behind the scramble testing is that models constructed using scrambled data should not be predictive for the test set. If the test compounds are still well-predicted (i.e., a predictive $r^2 > 0.5$), we can conclude that the model does not make meaningful predictions.

Sparse ν-SVR was implemented in the R programming language.[21] The linear optimization solution was obtained using the linear programming solver, lp_solve 5.0.[22] PLS and RF calculations were carried out using two statistical packages in R, pls.pcr and randomForests. For all three statistical methods, both the full set of descriptors and the descriptors selected by linear SVR feature selection were used to construct models. For random forests, the variable m_{try} was set to one-third of the number of descriptors for regression and 500 trees were used to construct the forest ensemble. For PLS, predictive models were constructed using the optimal number of latent variables, as determined by cross validation. For scramble testing, the random shuffle was carried out 100 times.

RESULTS AND DISCUSSION

The statistical results obtained from support vector regression, partial least squares, and random forest models are listed in Table 2. The SVR method gave the best results for both the training ($r^2 = 0.912$, RMSE = 0.440) and the test sets ($r^2 = 0.849$, RMSE = 0.597), with 45 fragments selected (RMSE = root-mean-square error). Cross-validation results are improved for PLS and RF when selected descriptors are used. Since descriptor selection is intrinsic in the random forest algorithm, the performance of the random forest method is generally not sensitive to the presence of irrelevant descriptors, and additional feature selection does not markedly improve the performance.[18] Among the three statistical methods, the SVR method slightly outperforms the other two methods in both training and testing predictions, although two other algorithms also gave satisfactory predictions on the basis of the testing set predictions.

Table 2. Comparison of Statistical Results of Different Predictive Models Based on Support Vector Regression, Partial Least Squares, and Random Forest

	SVR (FS)[a]	PLS (FS)[a]	PLS (ALL)[b]	RF (FS)[a]	RF (ALL)[b]
training r^2	0.912	0.882	0.820	0.884	0.889
training RMSE	0.440	0.510	0.623	0.505	0.493
cross-validated r^2	0.636	0.678	0.430	0.495	0.383
scramble r^2	−0.281	−0.430	−0.429	−0.109	−0.136
testing r^2	0.849	0.781	0.753	0.785	0.823
testing RMSE	0.597	0.719	0.763	0.712	0.645

[a] FS: only descriptors selected by linear SVR were used for model construction. [b] ALL: the full descriptor set before feature selection was used to construct the predictive models.

The scramble test, after feature selection, results in an average predictive r^2 of −0.281 for SVR. The negative coefficients observed for all methods indicate that the variances of the errors of prediction are larger than the variance of the data, and thus, the models of the scrambled data have no predictive ability. This confirms that the predictive models are not chance correlations. The scatter plot of experimental versus predicted pIC_{50} values by the SVR model for training compounds after feature selection and the corresponding leave-5-out cross-validation result after feature selection are shown in parts a and b of Figure 2, respectively. The scatter plot of the SVR testing results is shown in Figure 3. The predicted pIC_{50} values of 71 training and 19 testing drugs are listed in the Supporting Information. All of the results support the effectiveness of the linear SVR algorithm for both feature selection and constructing predictive models; therefore, the SVR model was chosen to predict the hERG binding affinity of Locus proprietary compounds.

Linear SVR models not only provide predictions, but they also imply relationships between the hERG binding affinity and 2D structural patterns in molecular structures through the fragment descriptor coefficients. The fragments identified as important by the linear SVR model as well as their coefficient weights in the linear model are shown in Table 3. The presence of fragments with positive coefficients increases the predicted hERG pIC_{50} value and, therefore, the potential liability, while fragments bearing negative coefficients diminish the hERG binding affinity. The magnitudes of the coefficients suggest how much the fragments will affect hERG binding. In Table 3, lipophilic fragments, such as benzyl and chloronaphthyl, bear positive coefficients in the linear model, and hydrophilic groups, such as primary amine, acetamide, and carbonyl groups, generally have negative coefficients. Lipophilic groups have been shown to increase hERG channel binding, and some hydrophilic

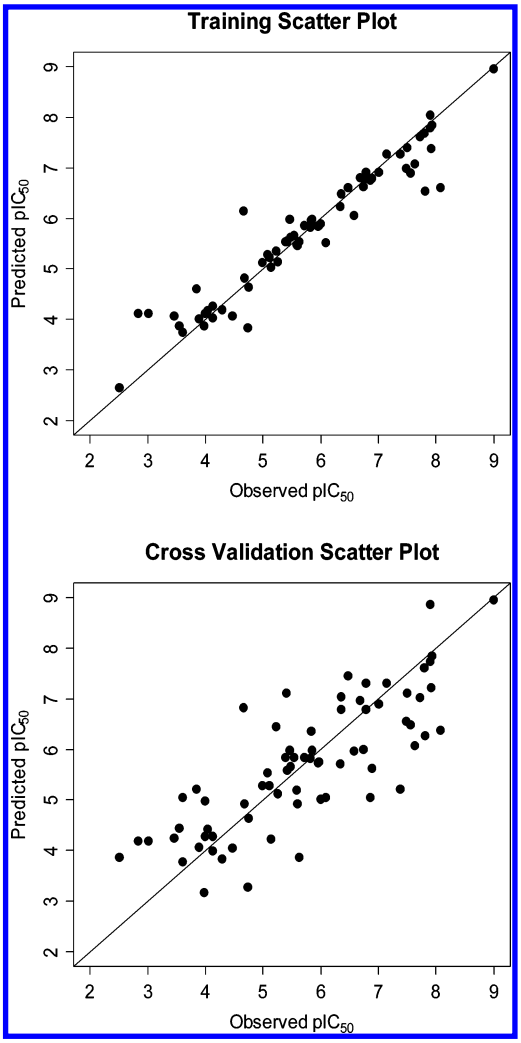


Figure 2. Training (left) and cross-validation (right) performance of hERG pIC50 predictions using the SVR model with feature selection.

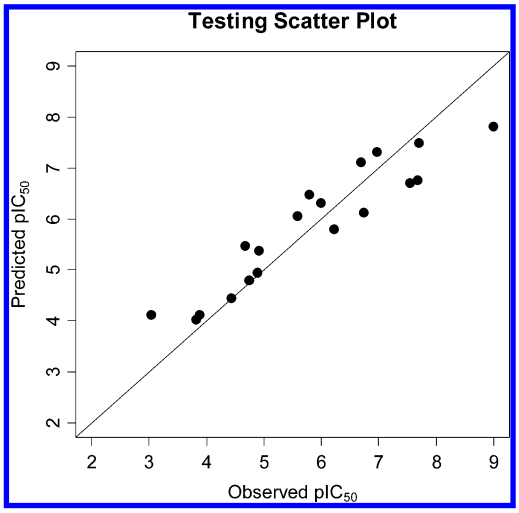


Figure 3. Scatter plot of observed vs predicted pIC_{50} of 19 test compounds using the SVR model with feature selection. The r^2 is 0.849; the RMS error is 0.597.

groups, such as carboxylic acid, ketones, hydroxide groups, and amine or primary aliphatic amines decrease hERG binding affinity. We also observe that compounds with substitutions of fluorine or methane sulfonamide groups increase hERG binding. In contrast to other hydrophilic fragments, the positive coefficient associated with secondary aliphatic amines indicate that this group will increase a drug's binding to the hERG channel, as suggested in Bain et al.'s recent work.[12] Tertiary amines have been identified as significant contributors for hERG binding in published pharmacophore models[4−6] and a structure-based modeling study.[3] In our study, *N*-methyl piperidine shows strong contribution to hERG binding, as the third most significant fragment, which is in agreement with the significance of tertiary amines in those 3D models. However, tertiary amines also present in weak hERG binders (e.g., $IC_{50} \gg 1\ \mu M$) indicate that the effect of tertiary amines is modulated by other structural features of the ligands. Several representatives of weak hERG binders containing other forms of tertiary amines are shown in Figure 4. Recently, Zolotoy et al.[9] proposed a structural model to address the duality of effects of tertiary amines in terms of charge and charge shielding. They observed that potent hERG inhibitors often bear a charged amine center that is surrounded by bulky hydrophobic groups. On the other hand, if the charged center is less sterically shielded, it is more easily deprotonated and has less potency. Moreover, we observe in our model that the presence of benzenesulfonamide diminishes hERG binding affinity, which is consistent with a previous observation that the benzenesulfonyl group is present in 20% of the channel nonblockers but in only 2% of the blockers.[8]

The predicted pIC_{50} values of 20 Locus proprietary compounds are listed in Table 4. The original experimental hERG binding data, percent binding, and the relative binding constant values converted from single-point percent binding data are also shown. These relative binding constants were compared with experimental pIC_{50} values to qualitatively judge whether the model can differentiate strong from weak hERG binders. The proprietary compounds belong to three series: series A (16 compounds), series B (2 compounds), and series C (2 compounds). The RMS error for all 20 compounds is 1.26 log units, with an r^2 of 0.29. The in silico model provides a qualitatively good estimation of hERG binding for series A compounds, as shown in Figure 5. For this 16 compound subset, the RMS error is 0.67 and the r^2 is 0.76. Only two compounds in series A and B, A_5 and A_{13}, had prediction errors greater than one log unit of experimental pIC_{50} values. To readily distinguish strong hERG binders from their weaker analogues, a threshold of predicted $-\log IC_{50}$ is introduced by observing the predictions for drugs that have been reported as potential hERG blockers. The experimental and predicted $-\log IC_{50}$ values of these 18 drugs are listed in Table 3 of the Supporting Information. The experimental IC_{50} values of these compounds are lower than 1 μM, and most of their predicted $-\log IC_{50}$ values are much larger than 6.5. Therefore, this value (6.5) is used as the cutoff of predicted pIC_{50} to differentiate potential hERG blockers from nonblockers in this study. When this cutoff is used, strong hERG binders with greater than 80% inhibition can be readily differentiated from other weaker binders below 40% inhibition, with only two false positives, compounds A_6 and A_{13}. Since the determination of this threshold is somewhat arbitrary, more studies will be needed in order to determine a reliable threshold. In addition, we observed that, when a substitution was changed from methanesulfonamide to acetamide in compounds A_3 and A_4, hERG binding affinities at a 1 μM concentration dropped from 84.0% (A_3)

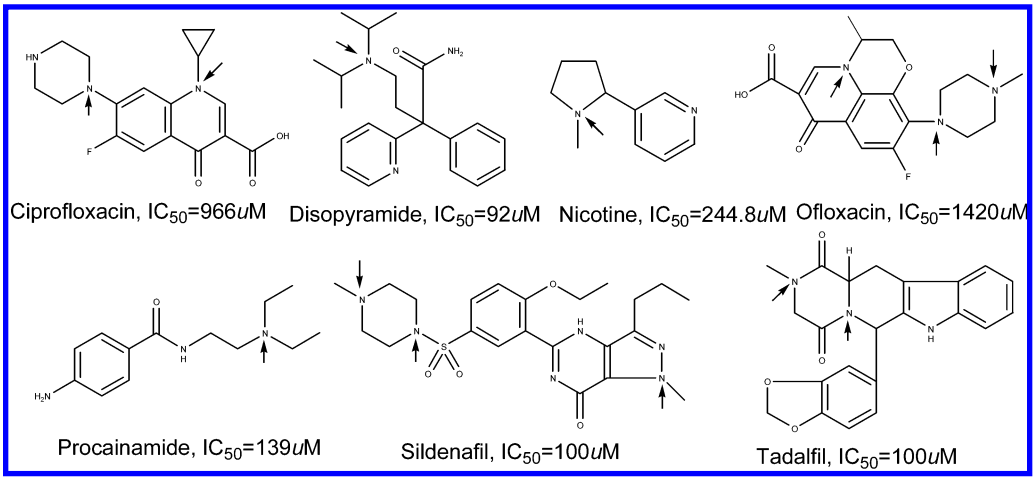
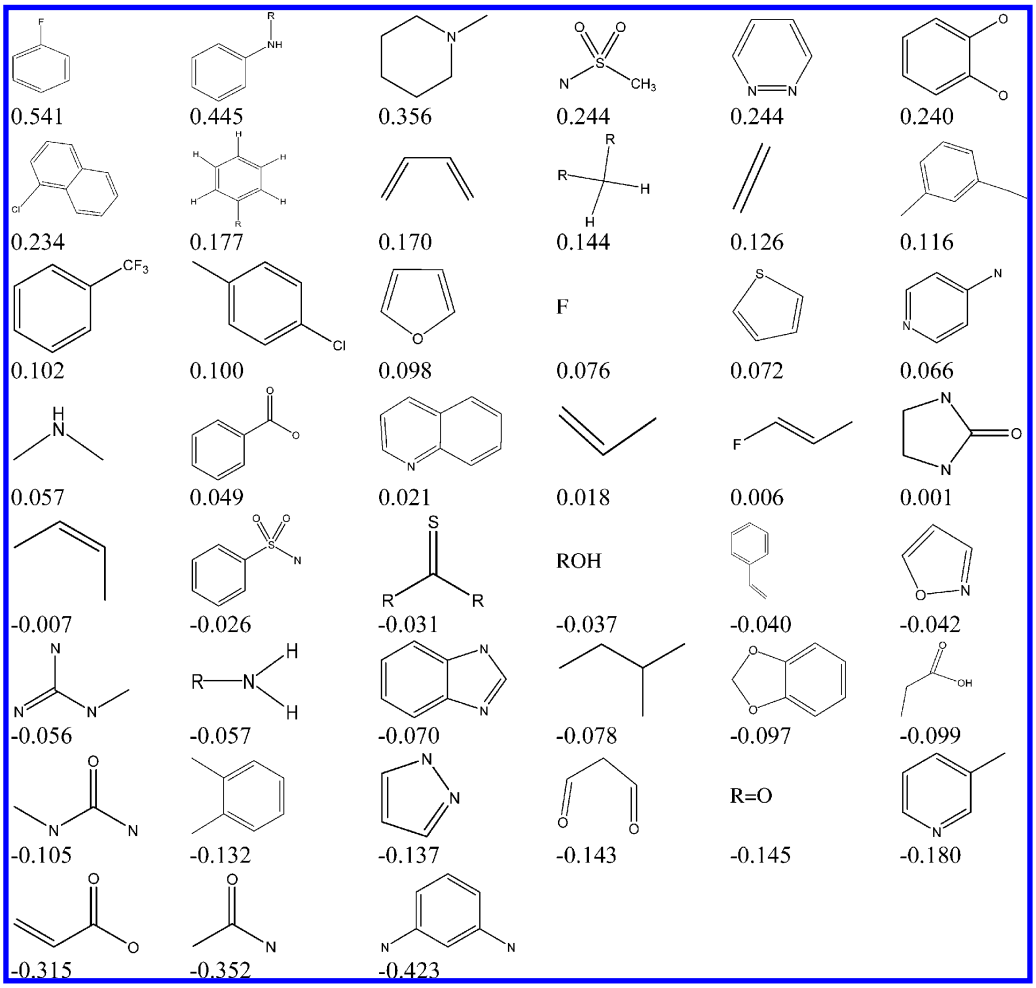


Figure 4. Weak hERG binders containing tertiary amines (highlighted with arrows) and their hERG binding affinities. Although tertiary amines are associated with high binding, they occur in both strong and weak binding compounds.

Table 3. Molecule Fragments Identified as the Most Relevant Descriptors for hERG Binding Affinity by Linear Support Vector Regression and Their Corresponding Weights in the Linear SVR Equation



to 7.0% (A_{12}) and from 82.5% (A_4) to 4.5% (A_{15}), respectively. This structure−activity relationship is consistent with the fragment coefficients that suggest that the presence of a methanesulfonamide group increases hERG binding, while the acetamide group reduces it.

In terms of absolute value of the binding affinity, series C compounds were poorly predicted, with errors greater than three log units, which may be due to either the estimation of the binding constant from the percentage binding or the diversity of the training data. Since the predicted pIC_{50} values are close to the approximated binding constants in series A and B, the effect from the diversity of the training data may be the major contribution to the discrepancy for series C. Since the training data only covers 71 compounds, the chemical space of the training data is limited and may not cover the chemical space for series C compounds. Generally,

Table 4. Prediction of hERG pIC_{50} Values for 20 In-House Compounds by the Linear SVR Model and Their Approximated Binding Affinity Constants Converted from Original Percentage Inhibition Data

compounds	% inhibition[a]	approximate binding affinity constants[b]	pred. pIC_{50}[c]
series A_1	95	7.28	7.38
series A_2	94.3	7.22	7.10
series A_3	84.0	6.72	6.92
series A_4	82.5	6.67	7.16
series A_5	47.6	5.96	7.02
series A_6	38	5.79	6.57
series A_7	35	5.73	6.15
series A_8	34	5.71	6.13
series A_9	27.6	5.58	6.16
series A_{10}	12.5	5.15	5.35
series A_{11}	12	5.13	5.27
series A_{12}	7.0	4.88	5.25
series A_{13}	6	4.81	6.53
series A_{14}	5	4.72	5.36
series A_{15}	5	4.72	5.36
series A_{16}	4.5	4.67	5.48
series B_1	6.0	4.81	5.11
series B_2	0.3	3.48	4.09
series C_1	7.0	4.88	8.28
series C_2	0.4	3.60	7.12

[a] Percentage inhibition data measured in HEK293 cells at the drug concentration of 1 μM. [b] Approximated relative binding constants calculated from percentage binding data. [c] Predicted pIC_{50} values for 20 in-house compounds from a constructed support vector regression model.

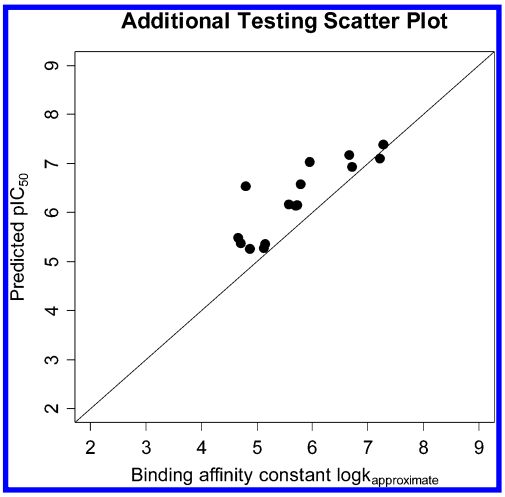


Figure 5. Scatter plot of predicted pIC_{50} values of 16 series A compounds listed in Table 4 vs their equivalent binding affinity constants converted from percentage binding data.

the predictive ability of a QSAR model is related to two factors: the similarity of the predicted molecules to the nearest training compounds and the number of such training neighbors in chemical space.[23] The molecules with more training neighbors or a high similarity with neighbors will generally be better predicted. To explore the similarity relationship between our proprietary compounds and the training compounds, hierarchical clustering was performed on the combination set of training compounds and 20 Locus compounds from the three series. The dissimilarity matrix of these compounds was derived from their fragment fingerprint vector distances, and their normalized distances between compounds were used as the dissimilarity score for clustering. The resulting clustering tree structure is demonstrated via a hierarchical dendrogram shown in Figure 6. The dendrogram is a binary tree in which each leaf represents one compound. It starts with a cluster for each compound and recursively joins compounds at each node into clusters based on structural dissimilarity scores. All compounds merge into one cluster when the tree reaches the root. The horizontal scale in the dendrogram corresponds to the dissimilarity score used to join different cluster in the hierarchical clustering process. Compounds with a high structural similarity are generally located within the same or adjacent clusters. For example, 16 series A compounds are tightly clustered together in the regions highlighted by the large arrow. This cluster is mainly composed of two groups: one contains those hERG binders (A_1, A_2, A_3, A_4, and A_5) with relatively high affinities (e.g., larger than 45% inhibition) and the other contains compounds (from A_6 to A_{16}, except A_{11}) with lower affinities below 40% inhibition. Compounds in the first group are further separated into two subgroups. Compounds A_1 and A_2 with affinities above 90% inhibition are clustered together, while the other three compounds form the other subgroup. In the second group, most compounds with affinities lower than 20% inhibition (A_{10}, A_{12}, A_{14}, A_{15}, and A_{16}) form a subgroup, which can be differentiated from another subgroup including compounds (A_7, A_8, and A_9) with affinities between 20% and 40% inhibition. In Figure 6, the dendrogram was plotted so that the nodes or branches with structurally similar compounds are assigned to the right side of figure. The cluster-joining process merges branches from right to left on the basis of their dissimilarity scores. In this figure, the compounds on the right side have a higher structural similarity than those compounds on the left side. We observe that two series C compounds are in branches on the left side and away from most of the training compounds, while series A compounds have more training neighbors around them. These results are consistent with the observation that the SVR model predictions were less accurate for these compounds than for series A compounds and explains the predictive scope of the QSAR models for the three different compound series. Two series B compounds also spread away from most training compounds; however, they cluster close to the training compound olanzapine. This may explain why they are well-predicted by the SVR model, although their predictions are less reliable than the predictions for series A compounds, since they have fewer training neighbors.

While the fragment-based QSAR method is a powerful approach, it shares the limitations of other QSAR approaches, among them the requirement of a high-quality, diverse training data set. In addition, the variety of different cells used to measured hERG binding introduces noise into the collected training data. Even a large set of training compounds cannot be guaranteed to cover the chemical space of novel compounds, and therefore, QSARs have uneven performance on various series of novel compounds. In that sense, postprocessing methods, such as visualization and model interpretation, are valuable to analyze the prediction in question and further guide the design of relevant training sets for relevant compounds. In our study, the hierarchical dendrogram is presented as a method to further investigate the predictive scope of QSAR models for novel compounds. Besides QSAR modeling, the visualization and interpretation of other computational models has recently drawn the attention of the computational chemistry community.

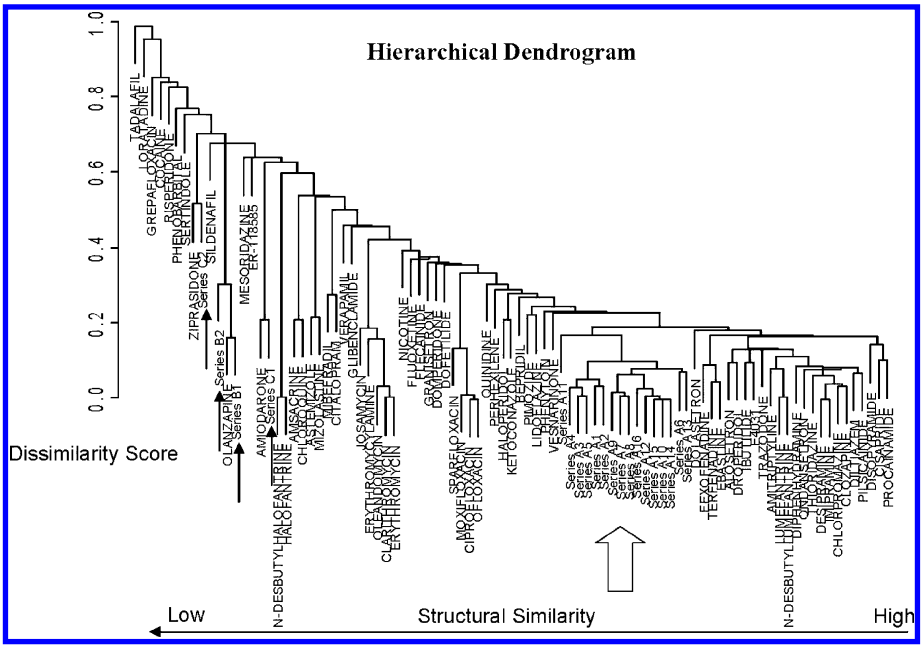


Figure 6. Hierarchical clustering dendrogram of the data set including 71 training drugs with 20 in-house compounds. The dendrogram is sorted by the dissimilarity score among all compounds from left to right, with less similar compounds on the left. The large arrow points to series A compounds and the smaller arrows to series B and C compounds.

CONCLUSIONS

This study described a computational method that employs fragment fingerprint and statistical approaches to predict hERG binding affinity. The most predictive model was generated with support vector regression, although PLS and random forests also produced satisfactory statistical results. Our method avoids the extensive efforts required for molecular 3D alignment and conformation sampling of the active configurations, making it efficient to create and use predictive models. The results show that the predictive model allows accurate predictions of compounds similar to the training set with an error of 0.6 log units and the prediction of more diverse proprietary compounds with an RMS error of 1.3 log units. The SVR algorithm was developed using linear programming so that it is quite conveniently implemented or integrated into other cheminformatics packages for virtual screening. The promising prediction results for the proprietary compounds further demonstrate the ability of the model to quickly screen structurally diverse compounds for hERG binding. Fragments identified as important descriptors in the feature selection step can provide insight to medicinal chemists for lead optimization. Moreover, the hierarchical clustering and dendrogram tree provides an additional postprocessing approach to determine the prediction confidence for different series of compounds.

ACKNOWLEDGMENT

The authors gratefully acknowledge Dr. Suobao Rong for his valuable discussion and critical comments, as well as the Scientific Computation, Computational Chemistry, Biochemistry, and Biology groups at Locus Pharmaceuticals.

APPENDIX: SPARSE N-SUPPORT VECTOR REGRESSION

In this appendix, we will explain how sparse ν-SVR works for regression. Given a collection of training data points $\{(x_1,y_1)\cdots(x_i,y_i)\cdots(x_M,y_M)\}$, where x_i is a vector in N-dimensional descriptor space and y_i is the bioactivity value, the goal of a regression problem is to find an optimal mathematical function $f(x)$ with the least deviation between predicted and observed responses. If linear functions $f(x_1\cdots\cdots x_N) = \sum_{j=1}^{N} w_j x_j + b$ in the input space are considered, w and b are parameters that need to be optimized in the regression problem, where w is the weight coefficient in the linear function. In SVR, w and b are optimized by minimizing a regularized error that controls both the training error and model complexity:

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{M}|y_i - f(x_i)|_\epsilon \qquad (2)$$

where $^1/_2||w||^2$ is the regularization factor which controls the model complexity ($||w||^2$ denotes the dot product of weight vector $w \times w$) and the second term represents the training error, in which C is a non-negative parameter that plays the tradeoff between the above two terms. In SVR, the training error is usually defined by the so-called ϵ-insensitive loss

$$L_g[y - f(x)]: = |y - f(x)|_\epsilon = \min[0, |y - f(x)| - \epsilon] \quad (3)$$

where ϵ is a non-negative parameter that determines the tolerance to error, and only deviations larger than ϵ are considered as errors during the optimization process. Figure 7 illustrates the basic concept of ϵ-insensitive loss in the linear case for one-dimensional input. The geometry space between the two lines $f(x) = (wx + b) \pm \epsilon$ defines an ϵ tube. The data points inside the ϵ tube will not be considered as error points. The use of the ϵ-insensitive loss introduces tolerance for noisy data, thus reducing overfitting.

However, in ϵ-SVR, it is not straightforward to determine a proper value for parameter $\nu \in (0,1]$. Hence, ν-SVR was proposed to use a new input parameter $\nu \in (0,1]$ to automatically select ϵ. In ν-SVR, parameter ν provides a

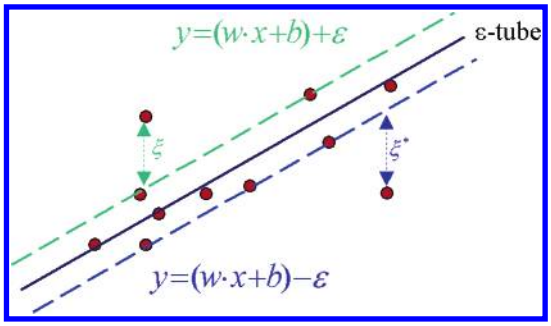


Figure 7. Graphical illustration of the ϵ-insensitive loss and the ϵ tube.[24] Only the two data points outside the ϵ tube are considered in the calculation of training errors, and the ϵ-insensitive loss for these two points are denoted by ξ and ξ^*.

lower bound on the fraction of support vectors and an upper bound on the fraction of error points outside the ϵ-insensitive tube. Since ν has a narrow selection range $\{\nu \in (0,1]\}$ and represents the accuracy level of the optimization solution, it is easier and more intuitive to adjust ν to find the proper ϵ for the model optimization than to directly tune ϵ. The ν-SVR is formulated as a convex optimization problem as follows:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + \frac{C}{M}\sum_{i=1}^{M}(\xi_i + \xi_i^*) + C\nu\epsilon \\ \text{subject to } & y_i - \langle wx_i\rangle - b \le \epsilon + \xi_i, \quad \xi_i \ge 0 \\ & \langle wx_i\rangle + b - y_i \le \epsilon + \xi_i^*, \quad \xi_i^* \ge 0 \\ & \epsilon \ge 0 \end{aligned} \tag{4}$$

Instead of minimizing the Euclidean norm (l_2-norm) $||w||^2$, sparse ν-SVR regularizes the l_1-norm $||w||_1 = \sum_{j=1}^{N}|w_j|$ of weights w in the linear model. To form a linear program, the l_1-norm $||w||_1$ is expressed as $\sum_{j=1}^{N}(\alpha_j + \alpha_j^*)$, where we define $\alpha_j = \alpha_j - \alpha_j^*$ subject to $\alpha_j \ge 0$ and $\alpha_j^* \ge 0$. Then, the sparse ν-SVR is formulated as follows:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\sum_{j=1}^{N}(\alpha_j + \alpha_j^*) + \frac{C}{M}\sum_{i=1}^{M}(\xi_i + \xi_i^*) + C\nu\epsilon \\ \text{subject to } & y_i - \sum_{j=1}^{N}(\alpha_j - \alpha_j^*)x_{ij} - b \le \epsilon + \xi_i, \\ & \qquad\qquad i = 1, 2, ..., M \\ & \sum_{j=1}^{N}(\alpha_j - \alpha_j^*)x_{ij} + b - y_i \le \epsilon + \xi_i^*, \\ & \qquad\qquad i = 1, 2, ..., M \\ & \alpha_j, \alpha_j^*, \xi_i, \xi_i^*, \epsilon \ge 0, i = 1, ..., M, j = \\ & \qquad\qquad 1, ..., N \end{aligned} \tag{5}$$

The jth weight w_j or $\alpha_j - \alpha_j^*$ in the resulting linear model from eq 5 determines how the jth descriptor affects the biological response. The sign of w_j indicates if the associated jth descriptor increases or decreases the biological activity, and its magnitude determines the significance of its impact on the response in the linear model. Moreover, sparse ν-SVR produces weight vectors where most fragment descriptors are diminished or driven to zero during the model optimization process as a result of the effect of 1-norm regularization and the convex property of linear programming. Only a few fragments will "survive" with a nonzero weight after the model construction. Therefore, these fragments with nonzero weights are considered as the descriptors most relevant to biological properties and are used to construct the final predictive model.

Supporting Information Available: Data tables describing training set compounds with the experimental and predicted bindings and the source references for the data. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Keating, M. T.; Sanguinetti, M. C. Molecular genetic insights into cardiovascular disease. *Science* **1996**, *272*, 681−685.
(2) Witchel, H. J.; Milnes, J. T.; Mitcheson, J. S.; Hancox, J. C. Troubleshooting problems with in vitro screening of drugs for QT interval prolongation using HERG K+ channels expressed in mammalian cell lines and Xenopus oocytes. *J. Pharmacol. Toxicol. Methods* **2002**, *48*, 65−80.
(3) Mitcheson, J. S.; Chen, J.; Lin, M.; Culberson, C.; Sanguinetti, M. C. A structural basis for drug-induced long QT syndrome. *PNAS* **2000**, *97*, 12329−12333.
(4) Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. Three-Dimensional Quantitative Structure−Activity Relationship for Inhibition of Human Ether-a-Go-Go-Related Gene Potassium Channel. *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427−434.
(5) Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K(+) channel blockers. *J. Med. Chem.* **2002**, *45*, 3844−3853.
(6) Pearlstein, R. A.; Vaz, R. J.; Kang, J.; Chen, X.-L.; Preobrazhenskaya, M.; Shchekotikhin, A. E.; Korolev, A. M.; Lysenkova, L. N.; Miroshnikova, O. V.; Hendrix, J.; Rampe, D. Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829−1835.
(7) Duffy, E. M.; Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000**, *122*, 2878−2888.
(8) Roche, O.; Trube, G. J. Z.; Pflimlin, P.; Alanine, A.; Schneider, G. A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *ChemBioChem.* **2002**, *3*, 455−459.
(9) Zolotoy, A. B.; Plouvier, B. P.; Beatch, G. B.; Hayes, E. S.; Wall, R. A.; Walker, M. J. Physicochemical determinants for drug induced blockade of HERG potassium channels: Effect of charge and charge shielding. *Curr. Med. Chem.: Cardiovasc. Hematol. Agents* **2003**, *1*, 225−241.
(10) Keseru, G. M. Prediction of hERG potassium channel affinity by traditional and hologram qSAR methods. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773−2775.
(11) Aronov, A. M.; Goldman, B. B. A model for identifying HERG K+ channel blockers. *Bioorg. Med. Chem.* **2004**, *12*, 2307−2315.
(12) Bains, W.; Basman, A.; White, C. HERG binding specificity and binding site structure: evidence from a fragment-based evolutionary computing SAR study. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 205−233.
(13) Kratochwil, N. A.; Huber, W.; Muller, F.; Kansy, M.; Gerber, P. R. Predicting plasma protein binding of drugs: a new approach. *Biochem. Pharmacol.* **2002**, *64*, 1355−1374.
(14) Clark, M. Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **2005**, *45*, 30−38.
(15) Bennett, K. P. B. J.; Embrechts, M.; Breneman, C.; Song, M Dimensionality Reduction via Sparse Support Vector Machines. *J. Machine Learning Res.* **2003**, *3*, 1229−1243.
(16) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.
(17) Smola, A. J. S. B. *A Tutorial on Support Vector Regression*; NeuroCOLT2 Technical Report NC2-TR-1998-030; 1998.
(18) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(19) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2003**, *43*, 1947−1958.
(20) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Model.* **1994**, *34*, 854−866.
(21) *R*, 2.2.0; The R Foundation: Vienna, Austria, 2005. http://www.r-project.org/.

(22) Berkelaar, M. *lp_solve*; Eindhoven Institute of Technology: Eindhoven, Netherlands, 2005. http://lpsolve.sourceforge.net/.
(23) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Model.* **2004**, *44*, 1912−1928.
(24) Bi, J.; Bennett, K. P. Duality, Geometry and Support Vector Regression. In *Advances in Neural Information Processing Systems 14*; Dietterich, T., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, 2002; pp 593−600.

CI050308F