Bias-Correction of Regression Models: A Case Study on hERG Inhibition

Katja Hansen,[†] Fabian Rathke,[†] Timon Schroeter,[†] Georg Rast,[‡] Thomas Fox,[§] Jan M. Kriegl,[*,§] and Sebastian Mika[*,||]

University of Technology, Berlin, Germany, Departments of Drug Discovery Support and Lead Discovery, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach a.d. Riss, Germany, and idalab GmbH, Berlin, Germany

Received February 27, 2009

In the present work we develop a predictive QSAR model for the blockade of the hERG channel. Additionally, this specific end point is used as a test scenario to develop and evaluate several techniques for fusing predictions from multiple regression models. hERG inhibition models which are presented here are based on a combined data set of roughly 550 proprietary and 110 public domain compounds. Models are built using various statistical learning techniques and different sets of molecular descriptors. Single Support Vector Regression, Gaussian Process, or Random Forest models achieve root mean-squared errors of roughly 0.6 log units as determined from leave-group-out cross-validation. An analysis of the evaluation strategy on the performance estimates shows that standard leave-group-out cross-validation yields overly optimistic results. As an alternative, a clustered cross-validation scheme is introduced to obtain a more realistic estimate of the model performance. The evaluation of several techniques to combine multiple prediction models shows that the root mean squared error as determined from clustered cross-validation can be reduced from 0.73 ± 0.01 to 0.57 ± 0.01 using a local bias correction strategy.

INTRODUCTION

In recent years, avoiding drug induced cardiac arrhythmia has become an important optimization parameter during the discovery and development of new drugs.[1,2] One of the most common issues is the prolongation of the QT interval by blocking the human ether-a-go-go related gene-encoded potassium channel (hERG channel) expressed in cardiac muscle cells.[3,4] QT prolongation enhances the risk of potentially fatal torsades de pointes. A number of drugs that were withdrawn from the market due to QT prolongation such as terfenadine or cisapride were shown to cause an unwanted blockade of the hERG channel. Following the "fail fast – fail cheap" paradigm of drug discovery it is highly desirable to identify compounds which exhibit hERG inhibition early in the discovery process.[5] In this context, in-silico methods have been developed and established to either cope with limited capacities for in vitro testing or to assess virtual compounds. For a survey on computational efforts toward a model for hERG blockade, comprising homology models, pharmacophore approaches, and QSAR models, we refer to recent reviews.[6−11]

In the present paper we describe the development of predictive QSAR models using machine learning techniques. Various modern machine learning methods which relate molecular descriptors with biological activities are available, and techniques like support vector machines (SVMs[12]), artificial neural networks,[13] or, more recently, Gaussian Processes[14] (GPs) have been applied to address drug absorption, distribution, metabolism, excretion, or toxicity, and hERG inhibition.[15−20] Regression models based on public domain hERG data sets in general exhibit predictive powers between $r^2 = 0.5$ and $r^2 = 0.7$ (estimated from cross-validation experiments or predictions for independent test set molecules).

Although the usage of different machine learning techniques and descriptor sets has led to a series of more or less equally performing models for hERG inhibition, less effort has been made to investigate how the predictions of individual regression models can be fused to obtain more robust and/or more accurate models. For categorical models, ensemble or consensus approaches often outperform individual models.[21] In this study, we compare different ways to fuse regression models. Following the concept of consensus modeling, the most straightforward approach to combine regression models would be to calculate an average predicted value from all models.[22] Alternatively, one can select the model that will most likely exhibit the lowest prediction error for each individual compound. The model is selected according to the similarity to the closest member of a set of reference compounds with known experimental values. This approach was proposed by Kühne and coworkers and applied to water solubility.[23] The similarity to a set of reference compounds can be exploited in an alternative approach to correct each individual prediction by a local bias estimate. This refers to the idea of associative or correction libraries that has been recently introduced and applied to different ADMET end points.[24−27] We extend the concept of model selection to the development of a biased model in which the expected prediction error of the selected

[*] Corresponding author e-mail: jan.kriegl@boehringer-ingelheim.com (J.M.K.), mika@idalab.de(S.M.).

[†] University of Technology.

[‡] Department of Drug Discovery Support, Boehringer Ingelheim Pharma GmbH & Co. KG.

[§] Department of Lead Discovery, Boehringer Ingelheim Pharma GmbH & Co. KG.

[||] idalab GmbH.

10.1021/ci9000794 CCC: $40.75 © 2009 American Chemical Society
Published on Web 05/12/2009

model is used as a local correction term. Additionally, we improve the estimate of the predictivity of our models by performing clustered cross-validation with multiple random partitions in addition to standard cross-validation experiments.

DATA

Functional Data on hERG Inhibition. To assess the inhibition of the hERG channel, we considered patch clamp data recorded in mammalian cell lines. Literature data were combined with data for proprietary compounds to increase the size and chemical diversity of the data set. Only compounds that were measured in patch clamp settings using either HEK293 or Chinese hamster ovary (CHO) cells were taken from public domain sources. For details on the selection procedure, we refer to Kramer et al.[28] In-house measurements were carried out as follows:

HEK293 cells stably transfected with hERG cDNA were cultivated on glass coverslips for up to 5 days. These coverslips were placed in a 2 mL recording chamber perfused with buffer containing (mM): NaCl (137), KCl (4.0), $MgCl_2$ (1.0), $CaCl_2$ (1.8), Glucose (10), HEPES (10), pH 7.4 with NaOH. Patch pipettes (2−5 MΩ) were filled with pipet solution (mM): K-aspartate (130), $MgCl_2$ (5.0), EGTA (5.0), K_2ATP (4.0), HEPES (10.0), pH 7.2 with KOH. Membrane currents were recorded using an EPC-10 patch clamp amplifier and PatchMaster software (HEKA Electronics), using the whole-cell configuration. Cells were held at −60 mV, and the following pulse pattern was applied: 40 mV for 2000 ms; −120 mV for 2 ms; ramp to 40 mV in 2 ms; 40 mV for 50 ms repeated at 15 s intervals. During each interpulse interval 4 pulses scaled down by a factor of 0.2 were recorded for a P/n leak subtraction procedure. Rs compensation was employed up to a level that safely allowed recording devoid of ringing. Peak current amplitudes were measured 3 ms after the ramp to +40 mV. Residual currents were calculated for each cell and compound concentration. A logistic concentration−response curve was fitted to the residual current data, and the resulting inhibitory constants were converted to pIC_{50} values.

The intra- and interlaboratory variability of pIC_{50} values collected using the patch clamp technique under comparable experimental conditions ranges between a factor of 2 and three (see[29] and BI unpublished data).

A comparison of the published and in-house determined pIC_{50} values for ten reference compounds spanning an activity range between 3.5 and 9.0 yielded an overall correlation of $r^2 = 0.9$. The pIC_{50} values for seven out of ten compounds agreed within 0.5 log units, and for only one compound, the deviation was close to one log unit. This comparison, together with the results in the following section, indicates that in this case literature and in-house data can safely be pooled in one data set.

Data Sets. The data set which entered the preprocessing included 563 compounds with pIC_{50} in-house measurements and 113 measurements of hERG inhibition for druglike compounds that were gathered from the literature. The literature set and the proprietary set span a similar range of pIC_{50} values, but they are centered in different regions of the chemical space. The PCA plot in Figure 1 illustrates this observation. Initial experiments showed that separate machine learning models for the two sets do not perform significantly better than a single model build using data from both sets. To this end, it was evaluated whether it is possible to use in-house data to predict literature data and vice versa. This fails almost completely, i.e. such predictors are worse than predicting just an average value. Further it was tested whether the performance of models built and evaluated exclusively on in-house or literature data, respectively, perform better than those built and evaluated on all data. In a standard cross-validation setting (*vide infra*), the performance is slightly better on the in-house data alone (presumably since they are more consistent) than on the literature data alone. However, in total the single set performance is very close to the performance of models build using all data. Therefore, in-house and literature data were combined into a single set of data.

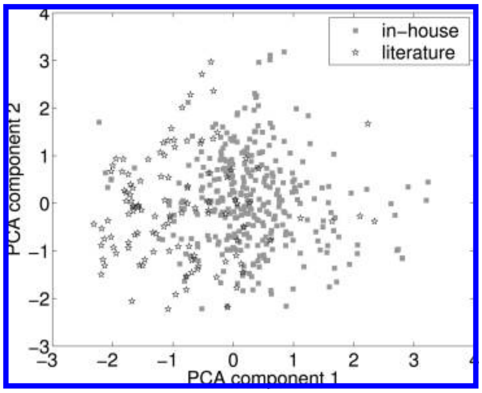

Figure 1. Projection of the data set on the first and second component of a PCA model using all descriptors. The in-house and literature set are marked to illustrate the different distributions.

Molecular Descriptors, Preprocessing, and Feature Selection. A number of descriptor sets was chosen to cover different aspects of chemical information about a compound. Properties solely derived from the 2D structure of the molecule as well as a 3D characterization of the interaction of the molecule with its surroundings were included. Descriptors were generated as follows: All compounds were ionized at pH 7.4 according to ChemAxon's pK_a predictor (JChem pK_a plugin, ChemAxon Kft, Budapest, Hungary). To calculate descriptors based on the 3D structure of a molecule, a single 3D conformation was generated with Corina (Version 3.4, Molecular Networks GmbH, Erlangen, Germany). The conformational energy was then minimized in the MMFF94x force field available in MOE (MOE 2007.09, Chemical Computing Group, Montreal, Canada). Chemical properties based on the 1D and 2D representation of the molecule such as size, shape, lipophilicity, atom and ring counts, surface areas, and topological features were taken from the 2D subset of the QSAR descriptors available in MOE. The 2D topological arrangement of pharmacophoric interaction points was characterized by ChemAxon pharmacophoric fingerprints (ChemAxon Kft, Budapest, Hungary) and CATS descriptors.[30] To assess the interaction energy of the molecule with its environment, we employed the VolSurf package (vsplus 0.4.5a, Molecular Discovery Ltd., U.K.), using four standard chemical probes (water, hydrophobic probe, carbonyl oxygen, and amide nitrogen[31,32]).

All descriptors were preprocessed as follows: Constant features, i.e. those that do not change over all compounds, were removed. Counts in MOE 2D descriptors and

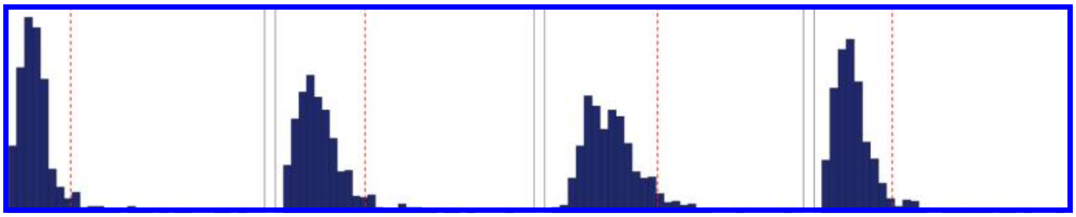

Figure 2. Histograms of δ outlier scores for each descriptor set (from left to right: ChemAxon, CATS, MOE, and Volsurf) using $m = 7$ neighbors. Dashed vertical lines indicate the respective cutoff corresponding to treating the 50 compounds with the highest scores as outliers.

ChemAxon pharmacophoric fingerprints were log scaled, i.e. $x = log(abs(x) + 1)*sign(x)$. Finally, all features were normalized as follows: From each feature the median of this feature over all data was subtracted, and the feature was normalized such that the largest absolute value is smaller than four. In contrast to subtracting the mean and scaling to standard deviation one, this approach is in our experience more robust when the distribution of the descriptors is skewed. All three descriptor sets were concatenated to one vector for each compound.

In cross-validation runs and for evaluation purposes the preprocessing was done on the training part of the data only. The results from the training set were then applied to the respective test data (i.e., remove those features that were constant in the training data and normalize using the parameters calculated on the training data).

The benefits of feature selection in statistical modeling have been studied extensively (see refs 33 and 34 and references therein). If controlling the number of input features is considered as a means of regularization, feature selection can, in principle, help to reduce the risk of overfitting. On the other hand, it is also possible to over fit by selecting a small number of features which are biased toward the training set.[35,36] In previous studies, we applied modern learning algorithms like Gaussian Processes and Support Vector Machines (SVMs) to several prediction tasks in drug discovery and design.[18,37−42] We observed that kernel based learning methods (GP, SVM, etc.) can benefit from using many (>1000) features, even if some of these features contain only a little information, provided that a proper regularization is applied during parameter tuning. Regularization is introduced for instance by optimizing a likelihood function in GP learning or by optimizing prediction performance on held out data in inner CV loops for SVR. Therefore we decided to leave the feature selection to the learning algorithm itself (apart from removing constant features).

Analysis of Outliers. Visual inspection of the raw descriptors and different PCA visualizations indicated that several percent of all compounds in the data set might be outliers. Therefore, we decided to analyze the data considering the κ, *gamma*, and δ indices introduced by Harmeling et al.[43] The first two indices, κ and *gamma*, are variants of heuristics that have been previously used (see refs 39 and 41 and references therein): κ is simply the distance to the mth nearest neighbor, and *gamma* is the mean distance to m nearest neighbors. The last index, δ, corresponds to the length of the mean vector to all m nearest neighbors. Since κ and *gamma* are only based on distances, they do not explicitly indicate whether interpolation or extrapolation is going to be necessary to make a prediction. δ allows for making this decision and indicates exactly how much extrapolation is necessary.

A first evaluation of the impact of outliers on modeling was performed as follows: δ indices were computed for each individual set of descriptors and for the concatenated set of all descriptors (see Figure 2) using $m = 7$ neighbors. After a visual inspection of these histograms for the four sets of δ indices, the number of outliers was set to 50, i.e. by this working definition, a compound is an outlier if its δ-index is in the top 50 of δ-indices for any of the four sets of descriptors. The complete evaluation of single models (see Evaluation Strategy) was then performed twice, once including and once excluding the outliers.

MACHINE LEARNING METHODOLOGY

The hERG pIC_{50} inhibition value was measured for a set of n chemical compounds. Based on these measurements we aim to predict the pIC_{50} inhibition value of new chemical compounds. More precisely, we look for a regression function f which can predict the pIC_{50} inhibition value for any compound represented as descriptor vector x.

When measuring the quality of f, contradictory aspects have to be considered: On the one hand, the complexity of the function f must be sufficient to express the relation between the given experimental pIC_{50} measurements $(y_1,y_2,.........,y_n)$ and the corresponding descriptor vectors $(x_1,x_2,.........,x_n)$ accurately. On the other hand, f should not be too complex (e.g., too closely adapted to the training data) to allow for reliable pIC_{50} predictions of unknown compounds. This trade-off is captured mathematically in the minimization of the *regularized empirical loss function*[44]

$$\min R_{emp}^{reg}(f) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(f(\mathbf{x}_i),y_i)}_{quality\ of\ fit} + \underbrace{\lambda\cdot r(f)}_{regularizer} \quad . \qquad (1)$$

where ℓ refers to a loss function, r to a regularization function, and λ to a positive balance parameter. The first term in eq 1 measures the quality of the fit of the model on the training data, and the second term penalizes the complexity of the function f to prevent overfitting. The parameter λ is used to adjust the influence of the regularization function r.

The regularization function r not only prevents overfitting but also is often used to ensure that the problem in eq 1 is not illposed which is required by various optimization methods.

The loss function ℓ determines the loss resulting from the inaccuracy of the predictions given by f. Most of the methods in this work use the *squared error loss function*

$$\ell(f(\mathbf{x}_i), y_i) = (f(\mathbf{x}_i) - y_i)^2 \tag{2}$$

Most inductive machine learning methods minimize the empirical risk function with respect to different regularization terms r and loss functions ℓ.

Standard Machine Learning Approaches. The following algorithms were used to build the regression models in this paper:

Ridge Regression. Ridge Regression[45] is a technique to regularize a standard linear regression model. The underlying loss function is the squared-error loss function as given in eq 2. The function f describes a hyperplane, and the regularization term r can be written as

$$f(\mathbf{x}) = x'\mathbf{w} + b \text{ and } r(f) = \|\mathbf{w}\|^2 \text{ with } \mathbf{w} \in \mathbb{R}^d,\ b \in \mathbb{R} \tag{3}$$

The entries of the vector $\mathbf{w}$ define the incline of the hyperplane and are called (regression) weights whereas the real number b is denoted as offset.

The regularization function r is especially important when dealing with correlated inputs as in the present case. In a typical linear model, a widely large positive weight can be canceled by a similarly negative weight in its correlated counterpart. The regularizer which is applied in Ridge Regression imposes a penalty on the sum of squares of the regression weights.

Gaussian Processes. Gaussian Processes (GPs) are techniques from the field of Bayesian statistics. The idea of GP modeling is to assume a prior probability distribution for the model underlying the data and to update this probability distribution in the light of the observed data to finally obtain a posterior distribution.[14] Each GP prediction is a Gaussian distribution where the mean can be interpreted as the predicted value and the variance as a confidence estimate or uncertainty.

GPs can be understood as a 'Ridge Regression' model using the following functional class f and regularizer r:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b \text{ and } r(f) = \sum_{i,j} \alpha_i k(\mathbf{x}_i, \mathbf{x}_j)\alpha_j \tag{4}$$

The loss function is again given by eq 2.

GPs require a covariance or kernel function k that needs to be specified. This function determines how the trade-off between the smoothness of f and the quality-of-fit is modeled. In the experimental part we use a combination of the "radial basis" kernel function and the "rational quadratic" kernel function[14]

$$k(\mathbf{x}, \mathbf{x}') = \eta \cdot \exp\left(\frac{-\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2}\right) + (1 - \eta)\left(1 + \sum_{i=1}^{d} w_i (x_i - x_i')^2\right)^{-\upsilon}$$

Support Vector Regression. The goal of Support Vector Regression (SVR)[46−48] is to estimate a function f of the form described in eq 4. Contrary to the GP approach SVR is based on an ε-intensive loss function

$$\ell(f(\mathbf{x}_i), y_i) = \left| f(\mathbf{x}_i) - y_i \right|_{\varepsilon} = \begin{cases} 0 & \text{if } |f(\mathbf{x}_i) - y_i| \le \varepsilon \\ |f(\mathbf{x}_i) - y_i| - \varepsilon & \text{else} \end{cases} \tag{5}$$

Absolute deviations up to ε are tolerated, and larger differences are penalized linearly. For SVR we also have to choose a kernel function which represents the class of functions from which the solution is taken from. Here we limited ourselves to radial basis function kernels, which are in our experience well suited for QSAR problems (e.g. ref 18):

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2}\right)$$

Random Forests. A Random Forest is essentially a collection of tree predictors where each tree depends on the value of a randomly sampled parameter vector.[49] The tree predictor itself recursively splits the training data into subsets and fits a constant model, i.e. the mean hERG inhibition value of the included compounds, on each subset. In each step, one subset of compounds X is divided into two subsets X_L, X_R guided by a least-squares error criterion. The loss of a set X with n_X compounds is defined as

$$\ell_{tree}(X) = \frac{1}{n_X} \sum_{x_i \in X} (y_i - \bar{y}_X)^2 \tag{6}$$

where $\bar{y}_X$ refers to the mean inhibition value of the compounds in X.[50] The best split of a subset X is the split that maximizes

$$\ell_{tree}(X) - \ell_{tree}(X_L) - \ell_{tree}(X_R) \text{ with } X_L \cup X_R = X \tag{7}$$

Since a different optimization problem is solved in each step, this approach cannot be considered as a global risk minimization problem (eq 1) and results in a discontinuous prediction function.

Baseline Model. All modeling approaches are compared to a baseline model. This model predicts identical values for each compound, namely the average target value seen in the training data.

ENSEMBLE MODELING APPROACHES

The individual models which were generated by the methods described in the previous section can be combined by different approaches to obtain ensemble models. Following the work by Kühne et al.[23] we assume the following general setting: Several *single models* f_i $i = 1,\ldots\ldots\ldots,l$ are generated and evaluated on disjoint training and test sets. For all compounds in the test set, the true value is known. Now an unknown compound t is added to the set. The central idea is to derive a prediction $f^*(x_t)$ for the unknown compound t by considering the performance of the models on the neighboring compounds within the test set. Thus, the test set can be considered as correction set. When validating such an approach we employ cross-validation, i.e. the test set of each cross-validation run is considered as correction set. Each compound of the correction set is left out once, and its predicted value is deduced from the neighboring compounds of the correction set. For the determination of the neighboring compounds a measure of molecular similarity is required. In this work we define the distance of two chemical compounds a and b as the Euclidean distance of the corresponding descriptor vectors taking a reduced set of normalized features into account:

$$\|\mathbf{x}_a - \mathbf{x}_b\|_{red} \tag{8}$$

Due to the curse of dimensionality, measuring the Euclidean distance in the whole descriptor space of more than 400 dimensions would result in unspecific distances. To diminish this effect the descriptor vector is reduced to a small selection of descriptors which are most relevant in the context of hERG inhibition. Initial experiments showed that a set of 100 features with the highest weighting factors as determined in the GP model is an appropriate set of descriptors. They are given in the Supporting Information.

Ensemble Models. The following ensemble modeling approaches were evaluated:

Selection by MAE (MAE Model). The single model with the **lowest mean absolute error on the neighboring compounds** is selected to predict the desired value for the unknown compound.[23] The k nearest neighbors are selected based on the distance measure introduced in eq 8, and the mean absolute error (MAE) is calculated as

$$\mathrm{MAE}(f_i) = \frac{1}{k}\sum_{j=1}^{k} |f_i(\mathbf{x}_j) - y_j| \tag{9}$$

Here f_i refers to one of the l trained single models. The predicted value $f^*(x_t)$ of this ensemble model is given by

$$f^*(\mathbf{x}_t) = f_{\mathrm{minMAE}}(\mathbf{x}_t) \text{ with } f_{\mathrm{minMAE}} = \underset{f_i\, i=1,..,l}{\operatorname{argmin}}(\mathrm{MAE}(f_i)) \tag{10}$$

If not stated otherwise, we set the number of nearest neighbors k that are considered in this or any other of the following ensemble approaches to $k = 10$.

Weighted Model. This model is based on the idea that a **weighted sum of all predictions** of the different single models may result in a greater improvement than selecting the prediction of only one model. In the simplest way, all individual predictions can be combined with equal weighting, i.e. the average predicted value is calculated from all models. Here we determine the weight of each model, v_{f_i}, according to the mean absolute error of the model on the neighboring compounds

$$v_{f_i} = \frac{1}{\mathrm{MAE}(f_i)}\left(\sum_{j=1}^{l}\frac{1}{\mathrm{MAE}(f_j)}\right)^{-1} \text{ with } \sum_{i=1}^{l} v_{f_i} = 1 \tag{11}$$

The prediction of the weighted model is then given by

$$f^*(\mathbf{x}_t) = \sum_{i=1}^{l} v_{f_i} f_i(\mathbf{x}_t) \tag{12}$$

The higher the accuracy of a single model f_i in the neighborhood of the unknown compound t, the greater the impact of the model f_i on the prediction of the weighted model.

Bias Corrected Model. In this approach one single model is selected according to the minimum mean absolute error on the neighboring compounds — similarly to the MAE model. Then the prediction of the selected model is **corrected by the mean error** on the neighbors. To incorporate the distance between the unknown compound t and its neighbors we define a *distance weight* d_j for each of the k nearest neighbors as

$$d_j = \frac{1}{\|\mathbf{x}_t - \mathbf{x}_j\|_{red}}\left(\sum_{i=1}^{k}\frac{1}{\|\mathbf{x}_t - \mathbf{x}_i\|_{red}}\right)^{-1} j = 1, ..., k \text{ with } \sum_{j=1}^{k} d_j = 1 \tag{13}$$

This way close compounds receive high distance weights. The selected model is now given as

$$f_{\mathrm{weightedDist}} = \underset{f_i\, i=1,..,l}{\operatorname{argmin}}\left(\frac{1}{k}\sum_{j=1}^{k}\frac{\|f_i(\mathbf{x}_j) - y_j\|_{red}}{d_j}\right) \tag{14}$$

The prediction is given as the prediction of $f_{\mathrm{weightedDist}}$ reduced by the prediction error on the neighborhood:

$$f^*(\mathbf{x}_t) = f_{\mathrm{weightedDist}}(\mathbf{x}_t) - \frac{1}{k}\sum_{j=1}^{k}\frac{f_{\mathrm{weightedDist}}(\mathbf{x}_j) - y_j}{d_j} \tag{15}$$

This approach is closely related to the concept of correction libraries.[26,27]

Average KNN Model and Random Choice Model. These reference models quantify the improvement achieved by applying ensemble models. In the Random Choice model the prediction of one single model is chosen **randomly** as the predicted value.

Unlike all other models, the Average model predicts the value for the unknown compound without considering the prediction of the single models. The prediction is the **average of the true values** of the neighboring compounds

$$f^*(\mathbf{x}_t) = \frac{1}{k}\sum_{j=1}^{k} y_j \tag{16}$$

EVALUATION STRATEGY

In order to test and compare the performance of all modeling approaches that have been introduced so far, a meaningful validation scheme is required.

Evaluation of Single Models. We evaluate all models in a **standard 3-fold cross-validation** setting: The data set is randomly divided into three disjoint sets. All five models are trained using two folds and evaluated on the third (test) fold. After three iterations with different test folds, each compound of the data set was once part of a test set, and exactly one prediction for each compound in the whole set was generated. We perform the whole process for 50 different random partitions and evaluate the performance of the different models on each test set. In addition, the standard error and the variance of different performance measures over all 50 trials is calculated.

Additionally we evaluate the models in a **clustered cross-validation** setting. The data set is grouped into 15 equally sized clusters using the geo-clust algorithm.[51] The similarity of two compounds in the descriptor-space is defined by the distance measure introduced in eq 8. Each cluster is then randomly allocated into three folds, each composed of five clusters and processed as described above for the standard cross-validation setting. This form of cross-validation helps to prevent too optimistic performance estimates by avoiding to have similar compounds in both the training and the test set.

For the Ridge Regression model and the Support Vector Regression we implement a "nested cross-validation" as follows: To determine the hyperparameters that result in the best possible generalization to unseen data, each training set is used for a three-times 5-fold "inner cross-validation". Afterward, the model is trained using the whole training set and the parameters that were determined in the inner cross-validation. In the case of Ridge Regression, the parameter λ is optimized, while for Support Vector Regression, the hyperparameters λ, σ, and ε need to be determined (see Machine Learning Methodology). By performing inner cross-validation, the test set of the individual outer cross-validation loops remains always truly unseen and allows for the estimation of the overall generalization performance. For the Random Forest a variant of the algorithm introduced by Breiman[49] is used. We train each tree using the full training set and keep the parameters constant. Contrary to the other learning techniques, all parameters in the Gaussian Process are estimated on the fly and need not to be specified a priori or be chosen by an inner cross-validation.

Evaluation of Ensemble Models. We evaluate **two different settings**: In the first setting we combine a Support Vector Regression, a Random Forest, and a Gaussian Process model that were trained on the same set of compounds. In the second setting we fuse several models that were obtained by applying the same learning algorithm but trained on different sets of compounds. This way we can distinguish between the performance improvement which results from the variety of machine learning methods and the improvement which originates from differences in the training sets. We restrict ourselves to Random Forests; the evaluation of our single models shows that similar results can be expected for SVR or GP models (*vide infra*).

For the evaluation of the ensemble models in both settings, we focus on clustered cross-validation. In each cross-validation loop, all five left-out clusters compile the correction set.

In the second ensemble setting we use the training set (composed of 10 clusters) to create 20 different subsets using a bagging approach, i.e. we sample with replacement. Each of the 20 different subsets is then used to train a regression model by using the same machine learning method. In this study we have chosen Random Forests; however, due to the similar performance of all nonlinear modeling techniques that were investigated here we expect similar results also for GPs and SVMs. The ensemble of 20 bagged Random Forests is now evaluated on the left out correction set as described in the first case. As in the evaluation of the single models, the whole process is repeated 50 times, and the performance of the different models is evaluated on each correction set.

Performance Measures. Models are evaluated with respect to the following performance criteria:

- **RMSE:** The root mean squared error (rmse) is defined as

$$\sqrt{\frac{1}{N}\sum_{i}(y_i - f(\mathbf{x}_i))^2} \tag{17}$$

- **Correlation:** The r^2 value or correlation coefficient is defined as

$$r^2 = \frac{(\sum_{i}((y_i - \bar{y})(f(\mathbf{x}_i) - \bar{f}(\mathbf{x}))))^2}{\sum_{i}(y_i - \bar{y})^2 \sum_{i}(f(\mathbf{x}_i)\bar{f}(\mathbf{x})^2)} \tag{18}$$

where y_i are the labels, $f(x_i)$ are the predictions, and $\bar{x}$ denotes the respective mean values.

Table 1. Evaluation of Single Model Approaches: Different Error Measures Applied in the Standard Cross-Validation Setting and the Clustered Cross-Validation Setting[a]

method	RMSE	r^2	LOG05	LOG1
Standard Cross-Validation				
Baseline Model	0.86	-	0.50	0.77
Ridge Regression	0.91	0.25	0.51	0.79
Gaussian Process	0.62	0.49	0.66	0.92
Support Vector Regression	0.62	0.48	0.66	0.91
Random Forest	0.63	0.48	0.64	0.91
Clustered Cross-Validation				
Baseline Model	0.87	-	0.49	0.77
Ridge Regression	1.15	0.11	0.38	0.65
Gaussian Process	0.73	0.30	0.54	0.85
Support Vector Regression	0.73	0.29	0.55	0.84
Random Forest	0.73	0.31	0.55	0.85

[a] RMSE denotes the root mean squared error, r^2 denotes the correlation coefficient, and LOG05 and LOG1 denote the fraction of predictions falling within 0.5 and 1 (log) units of the true value, respectively. The corresponding standard errors across all 50 repetitions are all below 0.01. See text for details.

- **LOG** ε: The fraction of predictions within a specific interval ε around the true value.

RESULTS AND DISCUSSION

Single Models. Table 1 shows the averaged performance measures (RMSE, r^2, LOG05, and LOG1) after 50 repetitions of standard cross-validation as well as for clustered cross-validation. The distribution of RMSE values is illustrated in Figure 3, and the relation between prediction and measured pIC_{50} value is visualized in Figure 4.

The *Ridge Regression* model does not perform very well, moreover, it is outperformed by the baseline model. In contrast, all other models yield results significantly improving upon the baseline prediction. Hence, a linear model seems be too "simple", i.e. it has not enough flexibility to capture the complex molecular mechanisms which determine the inhibition of the hERG channel. Although the data set is relatively small, and taking the complexity of the biological mechanism into account, the nonlinear methods yield models where up to two-thirds of all predictions are within 0.5 log units of the experimental value.

The results across all 50 repetitions of our cross-validation experiments are very consistent, showing only a small "within-method" and "between-method" variance. From this we conclude that all models are close to the performance which is achievable on this data set. Notably, this holds especially true when comparing the "kernel-based" learners SVR and GP and the "density-based" Random Forest. Moreover, the model performances when using trainings sets with and without the outliers identified by the δ-indices did not differ significantly for GP, SVR, and RF models. We conclude that these learning algorithms are robust enough to deal with the outliers included in the present set of data. Hence, we will present and discuss results that were obtained with the whole data set in the following sections. The *clustered cross-validation* and the *standard cross-validation* show the same tendencies, but the latter one yields more optimistic performance estimates — especially for Ridge Regression. This is not surprising since the most informative neighboring points of each test compound are taken away when leaving out whole clusters of compounds

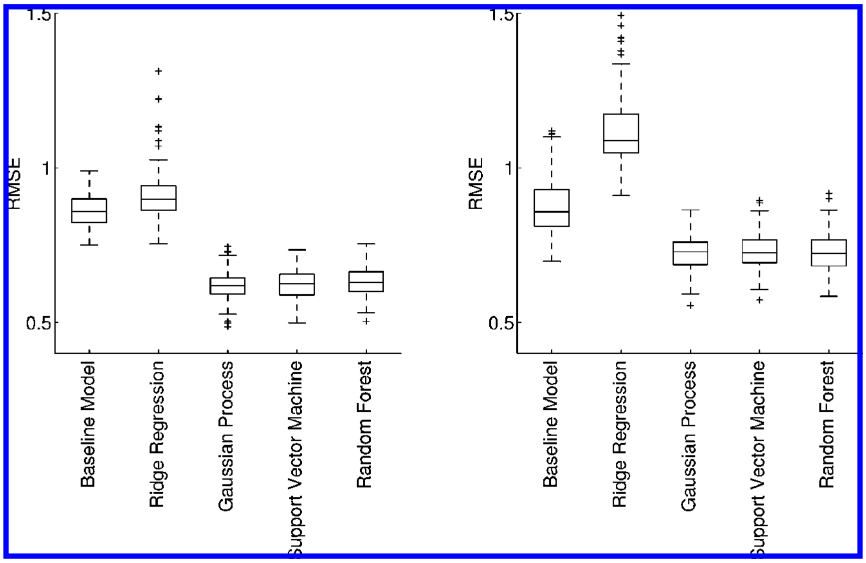

Figure 3. Box-plot depiction of the root mean squared error (RMSE) in the **standard** (left) and **clustered** (right) cross-validation setting over 50 repetitions. The box covers 50% of the actual data, the box height being the interquartile range, the horizontal line denotes the median. The whiskers are at most 1.5 the interquartile range. Points outside this range are marked as outliers.

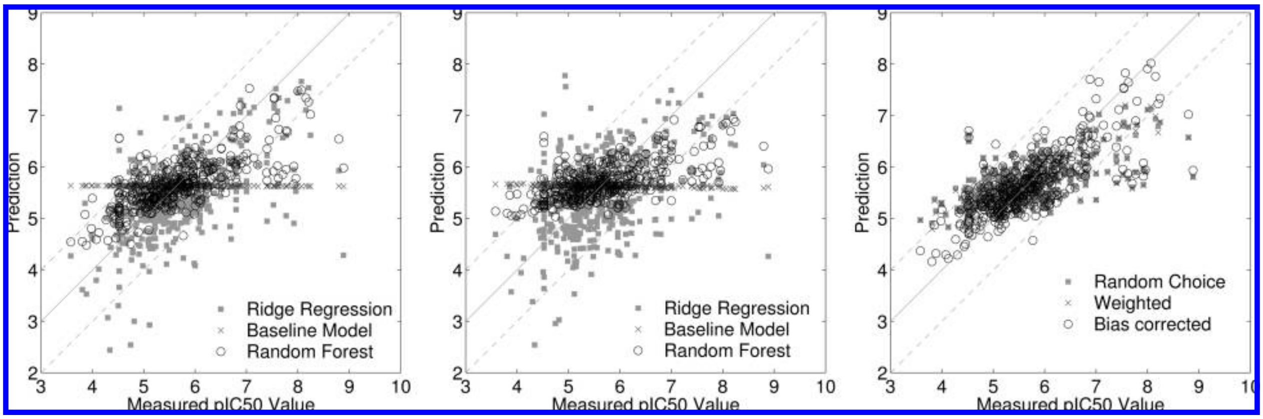

Figure 4. Relation between predictions and true values for the Baseline model, the Ridge Regression model and the Random Forest model evaluated in the standard cross-validation setting (left),and the clustered cross-validation setting (middle) over 50 repetitions. The plots for the Gaussian Process and the Support Vector Regression model (not shown) nearly equal the one of the Random Forest model. On the right we show the relation between predictions and true values for the Random Choice model, the Bias corrected model, and the Weighted model evaluated in the clustered cross-validation setting over 50 repetitions.

from the training set. Especially, as can be seen from Figure 4 the spread of the predicted *y*-values is significantly smaller in the clustered cross-validation. This is an effect that we observe in many practical applications of QSAR models: During application, contrary to the validation results, such models tend to miss out on the tails, i.e. the high and low values, of the target distribution. This underpins once more that a clustered cross-validation yields more realistic performance estimates for a real world application of the model. In fact, in a realistic application scenario in drug research, these models often have to be applied to new chemical series which might be significantly dissimilar to the compounds that have entered the model training process. The following observation fits into these considerations: the performance of the prediction models differs significantly between the individual clusters. Figure 5 shows the squared correlation coefficient between predictions and measured pIC_{50} values calculated on each cluster separately. It can be seen that there is a large spread in the performance depending on the cluster. Clusters 7 and 11 can be predicted with acceptable performance, whereas other clusters (e.g., 4, 8, 13) are very difficult to predict.

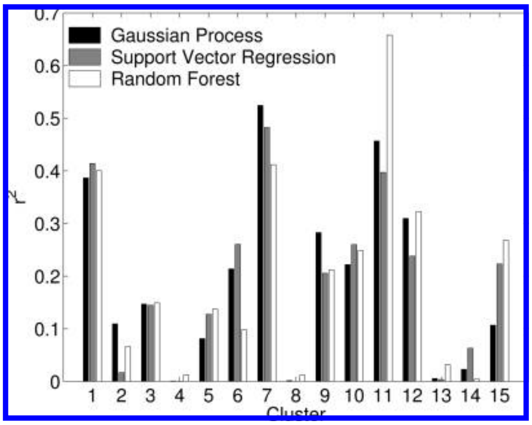

Figure 5. Calculation of the squared correlation coefficient on the clustered cross-validation results for each cluster separately.

Assuming that the clusters group structurally similar compounds, this finding resembles some of our experiences when applying QSAR models of this type: the interaction between members of certain structural classes and the hERG channel is in some cases only partially covered by the

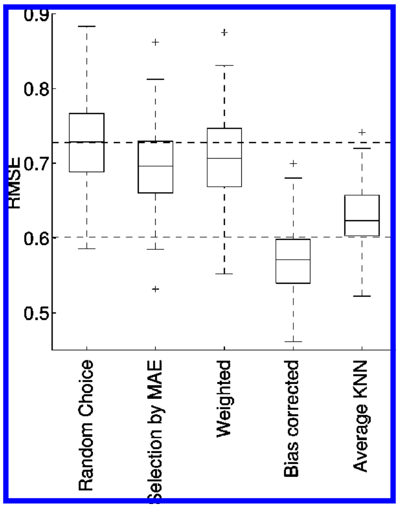

Figure 6. Combination of a Random Forest, a Gaussian Process, and a SVR model trained on equal sets: Box-plot depiction of the root mean squared error (RMSE) of the different ensemble methods in the clustered cross-validation setting over 50 repetitions. For details, see also Figure 3.

Table 2. Evaluation of Ensemble Model Approaches[a]

method	RMSE	r^2	LOG05	LOG1
Combination of GP, SVR, and Random Forest				
Random Choice	0.73	0.30	0.55	0.85
selection by MAE	0.7	0.35	0.58	0.87
weighted	0.71	0.33	0.56	0.86
bias corrected	0.57	0.54	0.71	0.93
average KNN	0.63	0.44	0.62	0.9
single Random Forest	0.73	0.31	0.55	0.85
Leave-One-Out Random Forest	0.6	0.726	0.66	0.92
Combination of Random Forest Models Trained on Different Sets				
Random Choice	0.76	0.26	0.52	0.84
selection by MAE	0.7	0.35	0.56	0.86
weighted	0.74	0.31	0.53	0.85
bias corrected	0.57	0.55	0.69	0.93
average KNN	0.63	0.46	0.63	0.91
(bagging) single Random Forest	0.76	0.26	0.52	0.83
Leave-One-Out Random Forest	0.6	0.726	0.66	0.92

[a] RMSE denotes the root mean squared error, r^2 denotes the correlation coefficient, and LOG05 and LOG1 denote the fraction of predictions falling within 0.5 and 1 (log) units of the true value, respectively. The corresponding standard errors across all 50 repetitions are all below 0.02. See text for details.

molecular descriptors and our models, whereas the structure–activity relationships revealed by other structural classes are much better reproduced. It is interesting to note that there is no direct correlation between the cluster composition into proprietary and public domain data and the corresponding model performance (data not shown). However, a more thorough analysis of the structural classes represented by each cluster and their putative mode of interaction with the hERG channel is beyond the scope of this paper.

Ensemble Models. *Benchmarks for Ensemble Models.* To allow for more insights into the performance gain achieved by different ensemble strategies we compare them not only to both baseline models (Average KNN and Random Choice) but also to the following two quantities:

- The RMSE of a *single Random Forest model* is taken as an upper benchmark: The ensembles are expected to achieve a RMSE that is smaller than this upper bound.

- The RMSE of a *leave-one-out cross-validated Random Forest model* is taken as a lower benchmark. This model is trained on all compounds in the training set and all (but one) compounds in the correction set. Since only one test compound is left out in each iteration, the model has almost full knowledge (contrary to the ensembles which are validated in clustered cross-validation only).

Combination of Different Single Models Trained on Equal Training Sets. Figure 6 visualizes the distribution of the RMSE over 50 repetitions for the different ensemble approaches when training three single models on identical data (all based on ten nearest neighbors). Table 2 and Figure 6 illustrate that the *Random Choice* model and the single Random Forest model yield similar results. This could be expected, because all single models (GP, SVR, and Random Forest) perform about equally well. The *Weighted* as well as the *Selection by MAE* approach do not improve the performance significantly compared to a single Random Forest model (dashed line). The reason for this observation is illustrated in Figure 7: The prediction errors of each individual model for the compounds are highly correlated. If one particular model yields an inaccurate prediction, the other single models show similar prediction errors. A compensation of prediction errors by combining single models is very unlikely. Therefore, a simple, unweighted averaging of the predictions of all individual models would in general also not lead to a significant improvement. The local bias correction in the *Average KNN* and the *Bias corrected* model show a large improvement. For the latter one the mean RMSE is even smaller than the RMSE of a Random Forest model evaluated in leave-one-out cross-validation (lower dashed line). This result indicates that a local bias correction is more important than the choice of the prediction method. Considering the fact that in the *Bias corrected* consensus model the single models are only trained on two-thirds of the data set and not on almost the entire data set like in the leave-one-out Random Forest model this works surprisingly well. Interestingly, incorporating additional information about just the ten nearest compounds allows for the reaching of this small RMSE.

So far the evaluation was focused on ensemble models which incorporate a neighborhood of ten compounds. To determine the influence of the neighborhood size on the quality of the model we performed the same evaluation varying the number of neighbors between one and ten. The results are summarized in Figure 8: The *Average KNN* and the *Bias corrected* model are strongly dependent on the number of neighbors, where an optimal number of neighbors seems to be 5. However, the RMSE does not significantly decrease when more neighbors are taken into account. The MAE and the Weighted model do not improve with the size of the neighborhood. This may again be caused by highly correlated prediction errors (cf. Figure 7).

Combination of Single RF Models Trained on Different Training Sets. In this section we evaluate the performance of ensemble models which combine the predictions of 20 Random Forests trained on different parts of the training set. Using a bagging approach we constructed a different training set for each Random Forest and combined them using the same ensemble models and

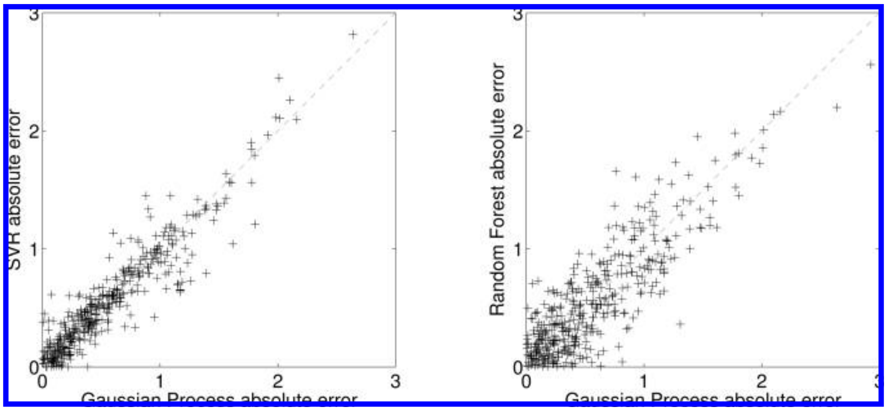

Figure 7. Visualization of the strong correlation between the absolute error of the GP and SVR model (left) and the GP and the Random Forest model (right). The corresponding correlation coefficients amount to 0.96 (GP versus SVR), 0.86 (GP versus Random Forest), and 0.82 (SVR versus Random Forest).

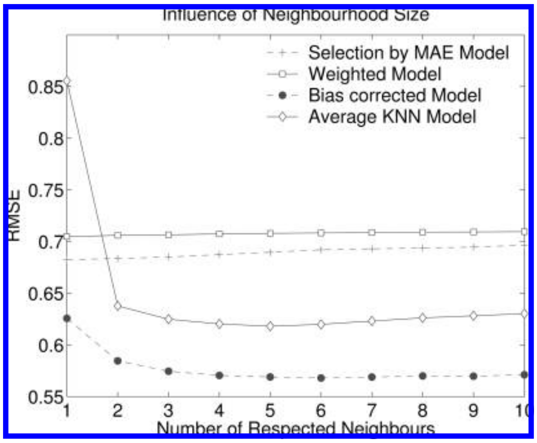

Figure 8. Influence of the number of considered neighbors on the ensemble model performance.

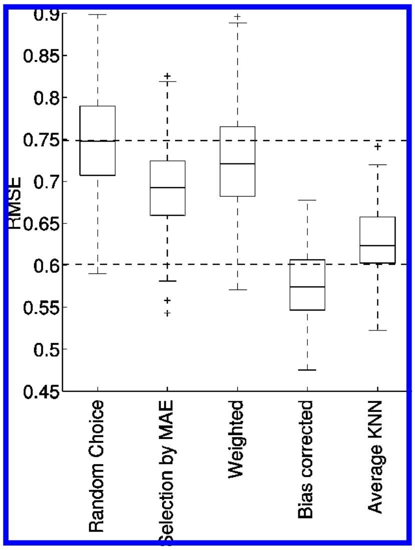

Figure 9. Combination of 20 classical bagging Random Forest models trained on different sets: Box-plot depiction of the root mean squared error (RMSE) of the different ensemble methods in the clustered cross-validation setting over 50 repetitions. For details, see also Figure 3.

cross-validation setting as described before. The main results of this evaluation are summarized in Table 2 and Figure 9. The underlying single models are now trained on data sets with more variety. Due to the different training sets, the differences between each pair of single models now occur on different clusters of compounds. Hence, some errors of the single models are compensated in the ensemble model. However, the distribution of RMSE values shows similar tendencies as in the previous setting: For the *Random Choice* model we observe a worse performance than for a single model (upper dashed line). The *Weighted* model again only achieves a small improvement. In contrast to the previous observation, the *Selection by MAE* model now achieves a somewhat larger improvement with respect to the single model. The *Bias corrected* model again achieves the largest improvement of all ensemble methods.

CONCLUSION

In this study we investigated the performance of several machine learning algorithms in single and ensemble model settings to address hERG inhibition. To broaden the basis for model training, literature and in-house manual patch clamp data were carefully combined. Single Gaussian Process, Support Vector Regression, and Random Forest models which were trained on the combined data set gave RMSE values of roughly 0.6 in standard cross-validation and about 0.7 in clustered cross-validation, whereas the linear Ridge Regression model was not able to discover a relationship between the molecular descriptors and hERG inhibition. Although all three nonlinear models are based on different ideas and algorithms, their prediction errors are highly correlated. The performance of a consensus model based on these three models thus only slightly exceeds the single model performance when building models by applying different learning algorithms to *identical training sets* in a standard cross-validation setting. In the more realistic clustered cross-validation setting, combining different models improves the performance of the final model. This can be observed for an ensemble whose individual models are trained with the same data set as well as for an ensemble based on different training subsets. In both cases, a local bias correction yields the best results. These findings are encouraging in two aspects: first they indicate that local bias correction may be a good way to cope with the influence of subtle structural modifications on the interaction with the hERG channel, while global trends such as the overall influence of compound lipophilicity[8,52,53] are still covered. Second, the calculation of a simple local bias correction from new measurements can substitute retraining of a model using the expanded data set, as proposed also in earlier studies.[26] Further investigations are necessary to evaluate in which cases the bias corrected approach is adequate and in which cases retraining should be preferred. Also, in the absence of a controlled test bed like in the present work it is mandatory to employ a suitable form of applicability test (see e.g. ref 41 and references therein) to check whether the compounds to be predicted are similar enough to the available correction set.

All QSAR models that have been discussed in this study only hardly give practical hints which molecular features should be altered during compound optimization to overcome hERG interaction. They are rather intended to provide a fast and reliable method for assessing large compound sets which originate from HTS and virtual screening campaigns or combinatorial libraries. Of course, a variety of experimental high-throughput methods such as competitive binding, rubidium efflux, or high throughput automated patch clamp assays are available for these tasks.[54,55] However, well-tuned *in-silico* models which were trained on high-quality experimental data can be of use especially in an early stage of a drug discovery project. They can be applied for virtual compounds before they are synthesized or purchased from external vendor catalogues. Moreover, *in-silico* methods are faster and cheaper to run. Since the accuracy of the experimental values may suffer from sample impurities, poor solubility, poor chemical stability, and a tendency to stick to surfaces or other properties, a predictive *in-silico* model may be a valuable alternative to support decisions such as the prioritization of HTS clusters, selection of compounds from vendor databases, or even the assistance to medicinal chemists in prioritizing synthesis plans.

ACKNOWLEDGMENT

We gratefully acknowledge helpful discussions with Klaus-Robert Müller (TU Berlin). We would like to thank Christian Kramer (University of Erlangen-Nuremberg) for compiling the literature data set, Dr. Paul Czodrowski (BI Austria) for calculating the descriptors, and Dr. Brian Guth (BI Pharma) for helpful comments on the manuscript. This work was supported in part by the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886 and DFG grant MU 987/4-1.

Supporting Information Available: List of the 50 descriptors selected for the calculation of the Euclidean descriptor distance of molecules. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Fenichel, R.; Malik, M.; Antzelevitch, C.; Sanguinetti, M.; Roden, D.; Priori, S.; Ruskin, J.; Lipicky, R.; Cantilena, L. Drug-induced torsades de pointes and implications for drug development. *J. Cardiovasc. Electrophysiol.* **2004**, *15*, 475–495.

(2) Recanatini, M.; Poluzzi, E.; Masetti, M.; Cavalli, A.; Ponti, F. D. QT prolongation through hERG K(+) channel blockade: Current knowledge and strategies for the early prediction during drug development. *Med. Res. Rev.* **2005**, *25*, 133–166.

(3) Fermini, B.; Fossa, A. The impact of drug-induced QT interval prolongation on drug discovery and development. *Nat. Rev. Drug Discovery* **2003**, *2*, 439–447.

(4) Sanguinetti, M.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **2006**, *440*, 463–469.

(5) Stansfeld, P.; Sutcliffe, M.; Mitcheson, J. Molecular mechanisms for drug interactions with hERG that cause long QT syndrome. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 81–94.

(6) Aronov, A. Predictive in silico modeling for hERG channel blockers. *Drug Discovery Today* **2005**, *10*, 149–155.

(7) Aronov, A. Tuning out of hERG. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 128–140.

(8) Jamieson, C.; Moir, E.; Rankovic, Z.; Wishart, G. Medicinal chemistry of hERG optimizations: Highlights and hang-ups. *J. Med. Chem.* **2006**, *49*, 5029–5046.

(9) Hutter, M. In silico prediction of drug properties. *Curr. Med. Chem.* **2009**, *16*, 189–202.

(10) Inanobe, A.; Kamiya, N.; Murakami, S.; Fukunishi, Y.; Nakamura, H.; Kurachi, Y. In Silico Prediction of the Chemical Block of Human Ether-a-Go-Go-Related Gene (hERG) K(+) Current. *J. Physiol. Sci.* **2008**, *58*, 459–470.

(11) Nisius, B.; Göller, A. H. Similarity-Based Classifier Using Topomers to Provide a Knowledge Base for hERG Channel Inhibition. *J. Chem. Inf. Model.* **2009**, *49*, 247–256.

(12) Müller, K.-R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. An Introduction to Kernelbased Learning Algorithms. *IEEE Neural Networks* **2001**, *12*, 181–201.

(13) Bishop, C. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, U.K., 1995.

(14) Rasmussen, C. E.; Williams, C. K. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, 2005.

(15) Fox, T.; Kriegl, J. Machine learning techniques for in silico modeling of drug metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1579–1591.

(16) Obrezanova, O.; Csanyi, G.; Gola, J.; Segall, M. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847–1857.

(17) Thai, K.; Ecker, G. Predictive models for HERG channel blockers: ligand-based and structure-based approaches. *Curr. Med. Chem.* **2007**, *14*, 3003–3026.

(18) Schwaighofer, A.; Schroeter, T.; Mika, S.; Hansen, K.; ter Laak, A.; Lienau, P.; Reichel, A.; Heinrich, N.; Müller, K.-R. A Probabilistic Approach to Classifying Metabolic Stability. *J. Chem. Inf. Model.* **2008**, *48*, 785–796.

(19) Nisius, B.; Göller, A. H.; Bajorath, J. Combining Cluster Analysis, Feature Selection and Multiple Support Vector Machine Models for the Identification of Human Ether-a-go-go Related Gene Channel Blocking Compounds. *Chem. Biol. Drug Des.* **2009**, *73*, 17–25.

(20) Li, Q.; Joergensen, F.; Oprea, T.; Brunak, S.; Taboureau, O. hERG classification Model Based on a Combination of Support Vector Machine Method and GRIND Descriptors. *Mol. Pharm.* **2008**, 117–127.

(21) O'Brien, S.; de Groot, M. Greater than the sum of its parts: combining models for useful ADMET prediction. *J. Med. Chem.* **2005**, *48*, 1287–1291.

(22) Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **2006**, *6*, 21–44.

(23) Kühne, R.; Ebert, R.-U.; Schürmann, G. Model Selection Based on Structural Similarity-Method Description and Application to Water Solubility Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 636–641.

(24) Tetko, I. Neural Network Studies. 4. Intoduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717–728.

(25) Tetko, I. Associative neural network. *Methods Mol. Biol.* **2008**, *458*, 185–202.

(26) Rodgers, S.; Davis, A.; Tomkinson, N.; van de Waterbeemd, H. QSAR modeling using automatically updating correction libraries: application to a human plasma protein binding model. *J. Chem. Inf. Model.* **2007**, *47*, 2401–2407.

(27) Bruneau, P.; McElroy, N. logD(7.4) Modeling Using Bayesian Regularized Neural Networks. Assessment and Correction of the Errors of Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 1379–1387.

(28) Kramer, C.; Beck, B.; Kriegl, J.; Clark, T. A composite model for HERG blockade. *J. Chem. Med. Chem.* **2008**, *3*, 254–265.

(29) Kirsch, G.; Trepakova, E.; Brimecombe, J.; Sidach, S.; Erickson, H.; Kochan, M.; Shyjka, L.; Lacerda, A.; Brown, A. Variability in the measurement of hERG potassium channel inhibition: Effects of temperature and stimulus pattern. *J. Pharmacol. Toxicol. Meth.* **2004**, *50*, 93–101.

(30) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894–2896.

(31) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, 29–39.

(32) Fortuna, C. G.; Barresi, V.; Berellini, G.; Musumarra, G. Design and synthesis of trans 2-(furan-2-yl)vinyl heteroaromatic iodise with antitumor activity. *Bioorg. Med. Chem.* **2008**, *16*, 4150–4159.

(33) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077–1084.

(34) Demel, M.; Janecek, A.; Thai, K.-M.; Ecker, G.; Gansterer, W. Predictive QSAR models for polyspecific drug targets: The importance of feature selection. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 91–110.

(35) Qi, Y. A.; Minka, T. P.; Picard, R. W.; Ghahramani, Z. Predictive automatic relevance determination by expectation propagation. *ICML '04: Proceedings of the twenty-first international conference on Machine learning*; New York, NY, 2004; p 85.

(36) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.

(37) Müller, K.-R.; Rätsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'Drug-likeness' with Kernel-Based Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 249–253.

(38) Schroeter, T.; Schwaighofer, A.; Mika, S.; ter Laak, A.; Sülzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Predicting Lipophilicity of Drug Discovery Molecules using Gaussian Process Models. *J. Chem. Med. Chem.* **2007**, *2*, 1265–1267.

(39) Schroeter, T.; Schwaighofer, A.; Mika, S.; Laak, A. T.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Estimating the Domain of Applicability for Machine Learning QSAR RModels: A Study on Aqueous Solubility of Drug Discovery Molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 485–498.

(40) Schroeter, T.; Schwaighofer, A.; Mika, S.; Laak, A. T.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Estimating the Domain of Applicability for Machine Learning QSAR RModels: A Study on Aqueous Solubility of Drug Discovery Molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 651–664.

(41) Schroeter, T.; Schwaighofer, A.; Mika, S.; Laak, A. T.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Machine Learning Models for Lipophilicity and their Domain of Applicability. *Mol. Pharm.* **2007**, *4*, 524–538.

(42) Schwaighofer, A.; Schroeter, T.; Mika, S.; Laub, J.; ter Laak, A.; Sülzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Accurate Solubility Prediction with Error Bars for Electrolytes: A Machine Learning Approach. *J. Chem. Inf. Model.* **2007**, *47*, 407–424.

(43) Harmeling, S.; Dornhege, G.; Tax, D.; Meinecke, F. C.; Müller, K.-R. From outliers to prototypes: ordering data. *Neurocomputing* **2006**, *69*, 1608–1618.

(44) Smola, A. J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.

(45) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Verlag: Berlin, Germany, 2001.

(46) Vapnik, V. *Statistical Learning Theory*; Wiley: New York, NY, 1998.

(47) Cristianini, N.; Shawe-Taylor, J. *Support Vector Machines*; Cambridge University Press: Cambridge, MA, 2000.

(48) Schölkopf, B.; Smola, A. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.

(49) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(50) Torgo, L. Functional Models for Regression Tree Leaves. *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*; San Francisco, CA, 1997; pp 385−393.

(51) Wen, Y. M. B. L. L.; Zhao, H. Equal clustering makes min-max modular support vector machine more efficient. *Proceedings of the 12th International Conference on Neural Information Processing (ICONIP 2005)*; Taipei, 2005; pp 77−82.

(52) Waring, M. J.; Johnstone, C. A quantitative assessment of hERG liability as a function of lipophilicity. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1759–1764.

(53) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51*, 817–834.

(54) Sorota, S.; Zhang, X.-S.; Margulis, M.; Tucker, K.; Priestly, T. Characterizing of a hERG Screen Using the IonWorks HT: Comparison to a hERG Rubidium Efflux Screen. *Assay Drug Dev. Technol.* **2005**, *3*, 47–57.

(55) Bridgland-Taylor, M. Optimisation and validation of a medium-throughput electrophysiology-based hERG assay using IonWorks HT. *J. Pharmacol. Toxicol. Meth.* **2006**, *54*, 189–199.

CI9000794