# Similarity-Based Classifier Using Topomers to Provide a Knowledge Base for hERG Channel Inhibition

Britta Nisius and Andreas H. Göller*

Bayer HealthCare AG, Global Drug Discovery, Lead Generation and Optimization, Aprather Weg 18a,
D-42096 Wuppertal, Germany

Received August 29, 2008

In a proof of principle study we show that the similarity property principle can be applied to predict ADMET properties, exemplified on the case of hERG $K^+$ channel inhibition. Early prediction of a drug candidate's hERG channel activity is becoming increasingly important in the drug discovery process because blockade of the hERG channel may lead to life-threatening cardiac arrhythmias. Using the Tripos Topomer Search technology as compound similarity measure, we query molecules with unknown hERG activity for similar molecules in a set of compounds with known hERG activity. The hERG activity of the query molecule is then predicted based on the hERG activity of its Topomer Search neighbors and their distances to the query molecule. The similarity property principle can be applied with promising performance to predict hERG inhibition as long as there is a high structural overlap between the chemical spaces of the query compounds and the reference data set. We show that this is achievable for database sizes of about 10000 structurally diverse molecules. In this case topoHERG is a similarity-based hERG classifier, which also acts as a knowledge base for hERG channel inhibition.

## INTRODUCTION

The hERG channel (human Ether-a-go-go Related Gene[1]) is a voltage-gated potassium channel expressed in the heart and the nervous system, which is responsible for the repolarization during the action potential.[2] Blockage of the hERG channel extends the repolarization phase leading to a prolonged QT interval of the electrocardiogram. The QT interval is defined as the time required for ventricular repolarization during a single cardiac cycle.[3] This prolongation of the QT interval is the desired effect for class III antiarrhythmic drugs, but it is an undesired side effect for noncardiovascular drugs which may lead to life-threatening cardiac arrhythmias such as Torsades de Pointes (TdP).[4] Due to an undesired blockade of the hERG channel at least eight noncardiovascular drugs among those cisapride, terfenadine, and astemizole were withdrawn from the market, and several other drugs have been linked to arrhythmias and sudden death.[2] Therefore, the hERG channel is a general antitarget in the pharmaceutical industry, and early prediction of a drug candidate's hERG channel activity is becoming increasingly important in the drug discovery process.[5]

Drugs from various indications showing diverse chemotypes were found to be hERG channel blockers.[6] hERG channel binding is very complex, rather unspecific, and inhomogeneous, and it is very difficult to understand the underlying SAR(s).

Alanine scanning mutagenesis studies were performed to provide clues to the mechanism of drug interference with hERG channel functioning. An initial investigation on the binding site of dofetilide indicated that Ser620 is important for hERG binding. Further experimental mutagenesis studies suggested two aromatic residues (Tyr652, Phe656) involved in hERG binding of many drugs (cisapride, terfenadine quinidine, chloroquine). Additionally, mutagenesis of Ser631 which is located outside the pore region modulates hERG channel binding of fluvoxamine, indicating the existence of an additional binding site besides the prominent binding site at the inner end of the large channel pore.[7] Therefore, these studies provide possible explanations why the hERG channel is blocked by such a broad spectrum of drugs although site-directed mutagenesis studies themselves are no direct experimental proof of the mechanism of hERG channel binding.

The current gold standard method for obtaining high-quality data of the functional effects of compounds at ion channels is the patch clamp technique,[8] but its low throughput limits its application to drug discovery and safety testing.[9] Automated approaches for higher throughput patch clamping are available and show reasonable correlation to the conventional voltage clamp assay.[10,11] However, the achievable throughput is far away from what is considered high throughput.

There are other methodologies such as radioligand binding, fluorescence, and ion flux measurements which are higher in throughput, but these methods lack the sensitivity of the patch clamp technique.[8] For instance, high throughput hERG assays using the displacement of high-affinity, radioactively labeled blockers are often nonfunctional because they focus only on the binding site of the replaced ligand. Therefore, at the present state *in silico* hERG models based on selected patch clamp data as experimental basis display a promising, fast, and inexpensive alternative for the early prediction of a drug candidate's hERG inhibition potency.

Several review articles describe in detail the approaches used to build in silico hERG models.[7,11,12] The applied approaches

* Corresponding author phone: +49-202-36-5442; fax: +49-202-365461; e-mail: andreas.goeller@bayerhealthcare.com.

10.1021/ci800304t CCC: $40.75
Published on Web 01/15/2009

can generally be divided into structure-based and ligand-based approaches. Structure-based approaches make use of homology models or mutagenesis data because a hERG X-ray structure is not yet available, whereas ligand-based approaches include both pharmacophore and QSAR methods. Especially in QSAR modeling the entire spectrum of available methodologies, including 3D QSAR models like CoMFa,[13] Catalyst,[14] and CoMSiA[15] or machine learning methods like Support Vector Machines,[16,17] Neural Networks,[18,19] Decision Trees,[20,21] Bayesian Classifiers,[22,23] or binary QSAR[24] were applied. A recent paper describes a more sophisticated approach to predict hERG activity: Kramer and co-workers differentiate between nonspecific hERG binders and two types of specific hERG binders by using pharmacophore scanning as a primary step.[25] Three different partial least-squares models are available for molecules hitting one of two pharmacophores (specific binders) or for molecules matching none of the two pharmacophores (nonspecific binders).

Herein we describe the results of a proof of principle study which differs from all approaches so far. The basis of our approach is the similarity property principle proposed by Johnson and Maggiora saying that overall similar molecules are likely to have similar biological properties.[26] The principle is the work horse of medicinal chemistry from lead finding to lead optimization and the basis for the development of new drugs, though it is common knowledge that sometimes small chemical changes in an active molecule, the so-called activity cliffs,[27−31] can have a significant influence on activity. There are examples in the literature showing that chemical similarity is not always obviously equivalent to biological similarity and that even small structural changes can cause significant changes in binding to biological macromolecules.[32,33] However, the similarity property principle has successfully been applied e.g. in ligand-based virtual screening, and therefore we decided to investigate in a proof of principle study whether this similarity property principle can also be applied to predict ADMET properties.

As a measure for similarity we use the Tripos Topomer Search technology,[34] which is designed to screen large reference data sets for molecules which are similar to the query compound and returns a pharmacophore and shape-based distance between the query molecule and its neighbors in the reference data set. For a new molecule with unknown hERG activity our approach searches for similar compounds in a large reference data set containing molecules with known hERG activity using Topomer Search. The hERG activity of the query molecule is then predicted based on the found Topomer Search neighbors, their distances, and their hERG activities. We chose to name our approach topoHERG because we use Topomer Search to capture similarity.

## MATERIALS AND METHODS

**Data Sets.** HERG in vitro inhibition $IC_{50}$ values were gathered from the literature. From the whole data set we extracted only patch clamp measurements on mammalian cell lines (HEK, CHO, and COS) to ensure consistency of our data set because measurements on different tissues are not always comparable. Mammalian cell lines display the ideal cell line because they allow the use of physiological temperatures of human species.[7] $IC_{50}$ values obtained in a nonmamalian cell line e.g. *Xenopus* Oocytes should be used with caution because the highly lipophilic environment in the cell line limits the access of the drug to its site of action, which may lead to a significant underestimation of the drug's potency as a hERG channel blocker.[35]

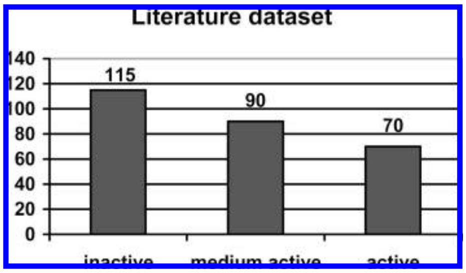


**Figure 1.** Distribution of the three $pIC_{50}$ classes in the literature data set. Molecules having a $pIC_{50}$ value lower than 5 are classified as hERG inactive, molecules with a $pIC_{50}$ between 5 and 6 are classified as medium hERG active, and molecules having a $pIC_{50}$ larger than 6 are classified as hERG active.

When using hERG $IC_{50}$ inhibition measurements as the data basis of our in silico hERG model it has to be considered that although there is a strong correlation between hERG inhibition and QT prolongation or cause of TdP, there is no conclusive evidence that the blocking of hERG leads to TdP. Therefore our model displays a classifier for hERG channel inhibition and was not designed to truly predict the torsadesogenic potential of a compound.

The collected $IC_{50}$ values were transformed to $pIC_{50}$ values for further analysis. Based upon these $pIC_{50}$ values the data set was grouped into three different $pIC_{50}$ classes according to Roche et al.:[19] compounds with $pIC_{50} \leq 5$ were considered as inactive, compounds with $pIC_{50} > 6$ were considered as active, and compounds with $5 < pIC_{50} \leq 6$ were considered as medium active.

We used these $pIC_{50}$ classes to join multiple $pIC_{50}$ values for the same molecule resulting from different measurements. If all available $pIC_{50}$ values for one molecule belong to the same $pIC_{50}$ class, the mean $pIC_{50}$ value was calculated. Multiple measurements for one compound with inconsistent $pIC_{50}$ classes were inspected further. The mean $pIC_{50}$ was determined if the difference between the $pIC_{50}$ values was lower than one $pIC_{50}$ unit. If this was not the case, but just one outlier was responsible for the difference in the $pIC_{50}$ classes, the outlier was deleted, and the mean $pIC_{50}$ value for the remaining measurements was calculated. If the $pIC_{50}$ values differed too much and multiple measurements were responsible for the inconsistent activity classes, all measurements for the special molecule were discarded.

All in all, we obtained a literature data set containing 275 compounds having $pIC_{50}$ values between 2.0 and 9.0. According to the three $pIC_{50}$ classes 115 of the compounds are inactive, 90 compounds are medium active, and 70 compounds are active. The distribution of the three $pIC_{50}$ classes in the literature data set is shown in Figure 1. An SD-File containing all molecules of our literature data set and their corresponding $pIC_{50}$ classes is given in the Supporting Information.

Earlier in the development process of the topoHERG model we used a subset of this literature data set containing only 232 molecules. These 232 molecules are those molecules we collected at the beginning of our proof of principle study, and the remaining 43 molecules of the literature data

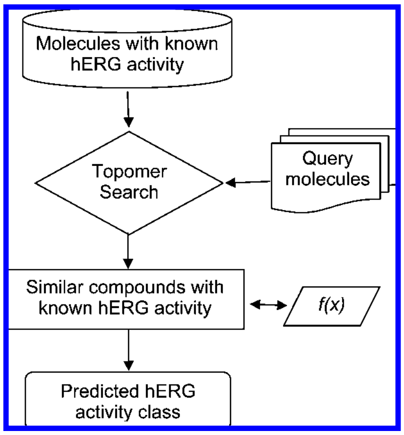

Molecules with known hERG activity
Topomer Search
Query molecules
Similar compounds with known hERG activity
f(x)
Predicted hERG activity class

**Figure 2.** Flowchart of the topoHERG workflow: for each query molecule Topomer Search is applied onto a reference data set containing molecules with known hERG activity to find similar molecules. The hERG activity of the query molecule is then predicted based on these Topomer Search neighbors, their distances to the query molecule, and their hERG activity using the activity prediction function *f(x)*.

set were collected during the development process of the topoHERG model. This subset of the literature data set used in some of the presented results contains 102 inactive, 67 medium active, and 63 hERG active molecules.

**Topomer Search.** To capture similarity between the query molecules and the compounds in the reference data set, we use the Tripos Topomer Search technology, which is based on the Topomer approach. A Topomer is defined to be "an invariant 3D representation of a molecular fragment, generated from its 2D topology by deterministic rules that produce absolute configuration, conformation and orientation".[36] Topomer Search screens large databases for molecules that are similar to the query molecule and calculates distances between them.

The similarity between a query molecule and a molecule from the reference data set is determined by comparing five pharmacophoric features and the similarity in shape of the associated Topomers. The pharmacophoric features used for Topomer comparison include aromatic features, hydrogen bond donors and acceptors, and positive or negative charges. The similarity in shape of two Topomers is measured based on a CoMFa-like steric field. All these similarities in shape or pharmacophoric features contribute to the overall distance between two Topomers, which is defined as the weighted Euclidean sum of all those distance contributions. Each pharmacophoric or shape contribution to the Topomer distance has its own weight, defining the maximum distance contribution for this special feature. The distance between the two molecules is then defined as the Euclidean sum of all fragment or Topomer similarities.

We use a special test version of Topomer Search provided by Tripos in 2006. To our knowledge and as confirmed by Tripos there are no substantial algorithmic differences between this test version and the current commercial Topomer Search version. All modifications by Tripos were done for the sake of implementation into the graphical user interface or to remove experimental features.

**Development of the topoHERG Model.** A flowchart displaying the overall procedure of our topoHERG approach is shown in Figure 2. A Topomer Search for each query molecule is performed in a reference data set containing molecules with known hERG activity. The output of Topomer Search is a list of all neighbors found and the corresponding distances to the query molecule. To determine the hERG $pIC_{50}$ class of a query compound we use these distances and the known hERG $pIC_{50}$ classes of all Topomer Search neighbors found as input for the activity prediction function *f(x)*, which is a central element of the topoHERG model. The output of this activity prediction function is a "hERG active" or "hERG inactive" prediction, if Topomer Search is able to find neighbors in the reference data set. The activity prediction function returns "unknown", if no Topomer Search neighbors are found in the reference data set.

**Table 1.** Four Most Important Parameters of Topomer Search Optimized Using a Genetic Algorithm

| parameter | description |
|---|---|
| max_dist | maximum overall distance between two similar molecules |
| max_hb | maximum distance contribution for all hydrogen bond donor or acceptor atoms |
| max_pn | maximum distance contribution for all positive and negative centers |
| max_arom | maximum distance contribution for all aromatic rings |

During the development of the topoHERG approach one important task was to find the best activity prediction function. A further important aspect was to determine the best parameters for Topomer Search.

**The Activity Prediction Function.** The activity prediction function is a central element of the topoHERG model. A general requirement to this function is that close Topomer Search neighbors should contribute disproportionately in comparison with Topomer Search neighbors having large distances. Additionally, the overall Topomer Search distance between two molecules is nonlinear because it is a weighted Euclidean sum. To capture this nonlinearity and to allow a much larger contribution of close Topomer Search neighbors, we use the inverse of the Topomer Search distance as a weight for each Topomer Search neighbor. The hERG $pIC_{50}$ class of a query molecule is determined based on eq 1:

$$S = \sum_{neighbors} \frac{1}{dist} * \begin{cases} 1 & active \\ -1 & inactive \end{cases}$$
$$Activity = \begin{cases} active & S > 0 \\ inactive & S \leq 0 \end{cases} \quad (1)$$

To differentiate between hERG active and hERG inactive neighbors, the inverse of the Topomer Search distance is multiplied by 1 if the Topomer Search neighbor is hERG active, and it is multiplied by −1 if the Topomer Search neighbor is not a hERG channel blocker. A query compound is classified hERG active, if the overall sum *S* is larger than zero, and it is classified to be hERG inactive otherwise.

**Find the Optimal Parameters for Topomer Search.** A further important aspect during the development process of topoHERG was to investigate which Topomer Search parameters influence the outcome of our topoHERG model and which are the best values for these parameters. Table 1

**Table 2.** Results of the Genetic Algorithm Using Two Different Fitness Functions

| fitness function | max_dist | tp | tn | fn | fp | up | un | accuracy | sensitivity | specificity |
|---|---|---|---|---|---|---|---|---|---|---|
| Matthews Score | 180 | 27 | 66 | 4 | 4 | 32 | 99 | 0.92 | 0.87 | 0.94 |
| 3tp+tn-3fn-fp | 270 | 36 | 132 | 24 | 32 | 3 | 24 | 0.75 | 0.60 | 0.80 |

gives a description of the four Topomer Search parameters we optimized using a Genetic Algorithm (GA). GAs are used in computing to find exact or approximate solutions to optimization or search problems. They are categorized as global search heuristics. GAs display a particular class of evolutionary algorithms using techniques inspired by evolutionary biology such as inheritance, mutation, selection, and recombination.[37] The application of GAs in molecular modeling is described in detail in a book by Devillers.[38]

We applied the GA onto the subset of our literature data set containing 232 molecules. This data set is used as the reference data set and as the set of query molecules simultaneously in a leave-one-out like way: When using molecule A as the query molecule, the remaining 231 molecules of the data set besides molecule A are used as the reference data set.

An important aspect of a GA is the fitness function: a measure used to determine the quality of a given parameter set. We compared different fitness functions, two of them yielding reasonable minima. The first fitness function resulting in a reasonable minimum is the Matthews Correlation Coefficient[39]

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tn+fn)\cdot(tp+fn)\cdot(tn+fp)\cdot(tp+fp)}} \qquad (2)$$

with *tp* defining the number of true positives (hERG active molecules, classified correctly), *tn* defining the number of true negatives (hERG inactive molecules, classified correctly), *fn* defining the number of false negatives (hERG active molecules predicted to be hERG inactive), and *fp* defining the number of false positives (hERG inactive molecules predicted to be hERG active).

The second minimum was found using a fitness function which we defined reflecting the different numbers of active and inactive molecules in the data set. Using $\mathrm{pIC}_{50} > 6$ as a threshold for hERG activity, the data set contains about three times more inactive than active molecules, and therefore we multiplied the number of true positives and the number of false negatives by three, resulting in the following fitness function:

$$fitness = 3tp + tn - 3fn - fp \qquad (3)$$

## RESULTS AND DISCUSSION

**The Optimal Topomer Search Parameter Sets As Determined by the GA.** While analyzing the results of the GA, we detect that the quality of our topoHERG model is almost only influenced by the parameter max_dist, defining the threshold for the overall distance between two similar molecules. The remaining parameters max_hb, max_pn, and max_arom influence the quality of a parameter set only marginally.

Table 2 lists the optimal values for the parameter max_dist found using the GA with the two different fitness functions and the resulting statistics for the 232 molecules. The Matthews Correlation Coefficient fitness function yields the minimum max_dist = 180 resulting in 92% correct classifications, a sensitivity (fraction of hERG active molecules correctly predicted to be hERG active) of 87% and a specificity (fraction of hERG inactive molecules correctly predicted to be hERG inactive) of 94%. Therefore, within the classified compounds, the accuracy for this parameter set is very promising. However, this minimum has a drawback because 56% of all molecules remain unclassified. This means that due to the low maximum distance between similar molecules, Topomer Search is for many query molecules not able to find Topomer Search neighbors, but if close Topomer Search neighbors are found the similarity property principle holds in almost all cases resulting in the excellent accuracy.

Using eq 3 as a fitness function for the GA yields a second minimum with max_dist = 270 resulting in a significantly lower fraction of unclassified molecules, due to the enlarged maximum distance between similar molecules. Only 12% of all molecules remain unclassified. But the increase of the parameter max_dist results also in a loss of similarity between a query molecule and its Topomer Search neighbors which leads to a decrease of accuracy: 82% of all predictions are correct, including a sensitivity of 60% and a specificity of 85%.

Based on the value for the parameter max_dist we appoint the two models topoHERG_180 and topoHERG_270. Since the quality of our topoHERG model is influenced only marginally by the values for the parameters max_hb, max_pn, and max_arom, the application of the GA to find the best parameter set was no real training phase, and a retraining of our topoHERG models is not necessary when new molecules are added to our data set.

Due to the high number of unclassified molecules it is obvious that topoHERG_180 cannot be used as a single global model when only the molecules in the literature data set are used as the reference data set for Topomer Search. However, on the basis of its excellent accuracy within the classified compounds, we propose that topoHERG_180 can act very well as a primary classifier in a two-stage hERG model. The overall idea behind this two-stage approach is that each query molecule is classified first using our topoHERG_180 model. The remaining unknowns of the topoHERG_180 model are then classified using a second classifier.

TopoHERG_270 can be used as a single model or as a secondary classifier in combination with topoHERG_180 as a primary classifier accepting a certain rate for unknowns. While using topoHERG_270 one aspect has to be kept in mind: the larger the distances between a query molecule and its Topomer Search neighbors the lower the similarity between them. Due to the nonlinearity of the distance a change of Topomer Search distance from 120 to 180 represents a much larger change in structure than a change from 0 to 60 would. The Topomer Search User Manual states that Topomer Search distances above 220 should be used with caution, especially distances higher than 240.[40]

To validate that it is nevertheless reasonable to use neighbors with a Topomer Search distance up to 270 to predict hERG activity, we compared each query molecule with its topoHERG_180 neighbors, with its topoHERG_270 neighbors, and with all the molecules in the reference data set, respectively, using some standard descriptors in SciTegic's Pipeline Pilot[41] (listed in the Supporting Information). Based on the descriptor values, which were scaled between 0 and 1, we calculated the mean Euclidean Distance for each query molecule. The resulting distance distribution histograms are shown in Figure 3. The mean Euclidean distance for each query molecule to the whole reference data set is determined to be 0.05, which is substantially higher than the mean Euclidean distance for each query molecule to its topoHERG_270 neighbors which is 0.03, which again is higher than the mean Euclidean distance for each query molecule to its topoHERG_180 neighbors (0.01). Therefore, topoHERG_270 neighbors are less similar than topoHERG_180 neighbors, but they show considerable more similarity to the query molecule than all compounds from the reference data set. That is why we conclude that a similarity grouping as implemented in topoHERG_270 is acceptable, and therefore we decided to use topoHERG_270 as one secondary classifier in combination with topoHERG_180 as a primary classifier.

**Comparison of Different Two-Stage Methods.** Besides using topoHERG_180 in combination with topoHERG_270, we also investigated whether topoHERG_180 can be used in combination with further hERG *in silico* models to improve their performances. The first model we combined topoHERG_180 with is a Decision Tree model published by Buyck et al.[42] and reconfirmed by Aronov et al.[43] This model classifies a molecule as hERG active if it contains a strongly basic group, if it is large, but not too large ($11.075 \leq \mathrm{CMR} < 17.64$), and if it is lipophilic enough ($\mathrm{CLOGP} \geq 3.666$). Additionally we investigated whether combining topoHERG_180 with the Schroedinger hERG QikProp[44] improves the performance of the QikProp model.

For all three combinations of two-stage models we compared the performance for two different thresholds for hERG channel activity. The first threshold of $\mathrm{pIC}_{50} > 6$ classifies only highly potent inhibitors as hERG active. This threshold is appropriate for building hERG classifiers which are used to screen large databases for high-potent hERG inhibitors. Based on our three $\mathrm{pIC}_{50}$ classes this threshold implies that hERG inactive and hERG medium active molecules are grouped together to form the hERG inactive class. We also used the contrary approach while grouping together hERG medium active and hERG active compounds to form the hERG active group, resulting in a threshold $\mathrm{pIC}_{50} > 5$ for hERG activity. This threshold is appropriate for building hERG classifiers which are applied in the lead-optimization phase.

The results for $\mathrm{pIC}_{50} > 5$ and $\mathrm{pIC}_{50} > 6$ are shown in Figure 4(a)−(f). In both figures the light gray bars show the performances of using topoHERG_270, the Decision Tree model, and the QikProp model as single models. The dark gray bars show the performances of the two-stage models: first, all molecules are classified using topoHERG_180, and afterwards the remaining unknowns are classified using topoHERG_270, the Decision Tree model, or the QikProp model.

The Decision Tree model shows very good specificity and rather weak sensitivity for both activity thresholds when used as a single model. QikProp as a single model gives very good sensitivity and rather weak specificity for both activity thresholds. Therefore, these two models perform oppositional with regard to hERG active and hERG inactive molecules: the Decision Tree model shows high accuracy in the prediction of hERG inactive molecules, whereas the QikProp model shows high accuracy in predicting hERG binders.

When combining these two contrary models with topoHERG_180 it becomes apparent that using topoHERG_180 in a two-stage approach adjusts the weaknesses of the associated secondary classifiers: combining topoHERG_180

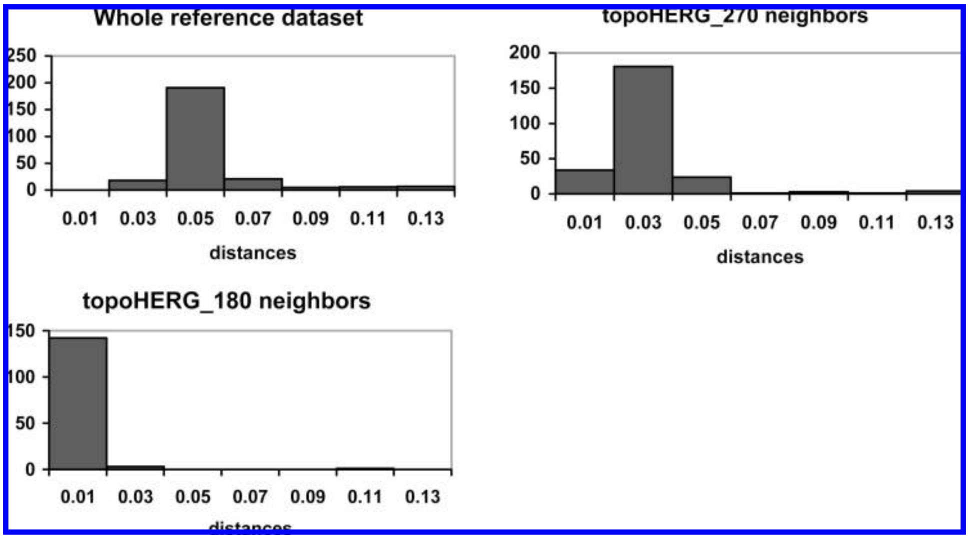


**Figure 3.** Histograms of the mean Euclidean distances based on common 2D descriptors (refer to Table S1, Supporting Information) of each query compound to all reference data set molecules, to all neighbors with a maximum Topomer Search distance of 180, and to all neighbors with a distance of 270, respectively. TopoHERG_270 neighbors show substantial more similarity to the query molecule than all compounds from the reference data set. Therefore, we conclude that a similarity grouping as implemented in topoHERG_270 is acceptable.

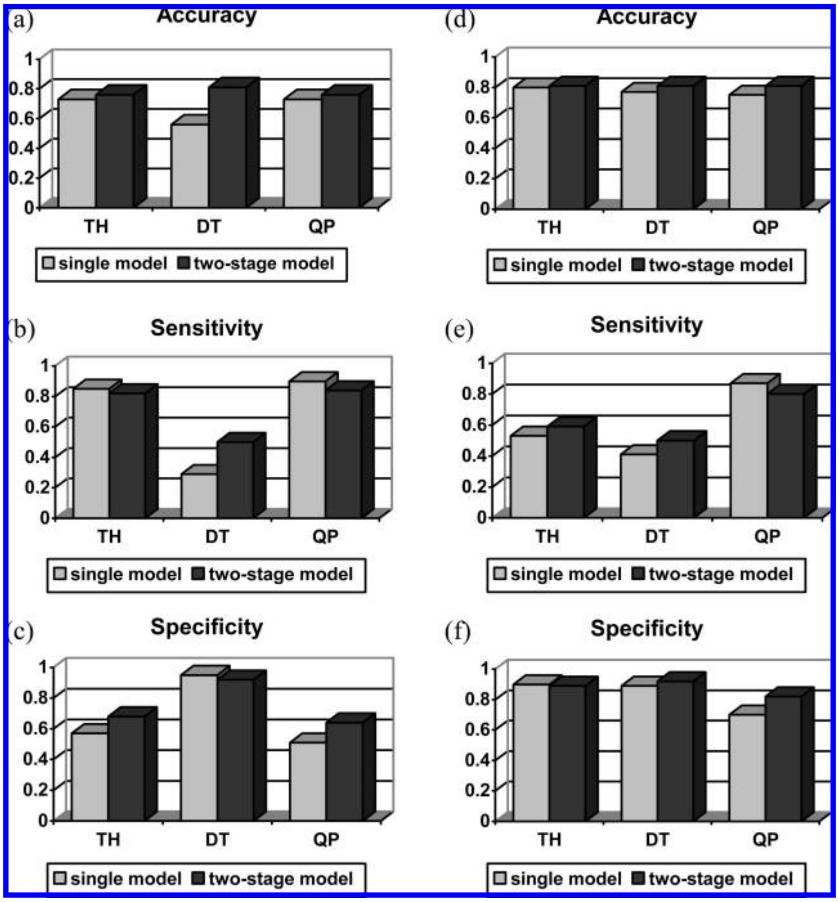


**Figure 4.** Comparison of three different two-stage models for two different hERG channel activity thresholds (pIC50 > 5 and pIC50 > 6). Performances of topoHERG_270 (TH), Decision Tree (DT), and QikProp (QK) models when used as a single model are shown in light gray. The results of two-stage approaches first applying topoHERG_180 and then either topoHERG_270, the Decision Tree, or QikProp to classify the remaining unknowns are shown in dark gray: (a) accuracy for $pIC_{50} > 5$, (b) sensitivity for $pIC_{50} > 5$, (c) specificity for $pIC_{50} > 5$, (d) accuracy for $pIC_{50} > 6$, (e) sensitivity for $pIC_{50} > 6$, and (f) specificity for $pIC_{50} > 6$.

and QikProp significantly improves the weak specificity of the Qikprop model, while combining topoHERG_180 and the Decision Tree model significantly improves the weak sensitivity of the Decision Tree model.

A crucial result for both activity thresholds and all three secondary classifiers is that using topoHERG_180 as a primary classifier always improves the accuracies of the single models. The largest increase in accuracy is achieved when combining topoHERG_180 and the Decision Tree model using $pIC_{50} > 5$ as the activity threshold: the Decision Tree model yields 56% accuracy, whereas the corresponding two-stage approach yields 81% accuracy. The Decision Tree model performs weakest as a single model for the activity threshold $pIC_{50} > 5$, but the corresponding two-stage model outperforms the other two-stage approaches.

When using $pIC_{50} > 6$ as the activity threshold, the overall accuracies of all two-stage models are the same (81%), but the corresponding sensitivities and specificities differ. For this activity threshold the Decision Tree model and our topoHERG_270 model both show a rather bad sensitivity and a good specificity, whereas the QikProp model shows a good sensitivity and a medium specificity. With regard to sensitivity and specificity the combination of topoHERG_180 and QikProp performs best for this activity threshold. However, due to the low number of hERG binders in the data set for this activity threshold the overall accuracies for all three two-stage approaches are the same.

All in all, the combination of Decision Tree and topoHERG_180 model performs best for an activity threshold of $pIC_{50} > 5$, and the combination of QikProp and topoHERG_180 model performs best for $pIC_{50} > 6$. The combination of our two topoHERG models performs comparably well for both activity thresholds, but it has to be considered that about 10% of all molecules remain unclassified because there are no Topomer Search neighbors within a maximum distance of 270 for these molecules.

**topoHERG_180 Performance.** To further investigate the performance of using topoHERG_180 as a single model we developed different test cases for our approach. First of all, we split our literature data set into a training set and a test set and applied the topoHERG_180 model to all molecules in the test set using the training set as the reference data set for Topomer Search.

We used a hierarchical, agglomerative clustering approach based on unity fingerprints and a maximum Tanimoto Coefficient of 0.7 to divide our literature data set into two independent data sets. We selected all molecules in clusters to form the reference data set for Topomer Search and all remaining singletons to form the set of query molecules resulting in a reference data set containing 174 molecules and a set of 101 query structures, whereof 50 are hERG inactive, 19 are hERG active, and 32 molecules show medium hERG activity. Using this approach we chose the reference data set structurally as dissimilar as possible from the set of query structures based on the unity fingerprints similarity measure.

Applying topoHERG_180 to each molecule in the set of query structures results in 5 classified compounds. Four of these 5 classified molecules are predicted correctly to be inactive. The remaining classified molecule shows medium hERG channel activity. Using $pIC_{50} > 6$ as the threshold for hERG channel activity results in a correct classification of this molecule. When using $pIC_{50} > 5$ as the activity threshold, this medium hERG active molecule is misleadingly predicted to be hERG inactive.

Thus, within the classified compounds the accuracy is quite well because depending on the selected activity threshold 80% or 100% of all molecules are classified correctly. The reason for the low number of classified compounds is the introduced maximized diversity between the set of query molecules and the reference data set. Therefore it is not surprising that the similarity measure Topomer Search is able to detect similarities between the reference data set and a query molecule in only very few cases and thus give a prediction based on the detected similar compounds.

Additionally, we used a data set containing in-house Bayer Schering Pharma (BSP) hERG inhibition measurements as an independent test set in two different manners: On the one hand we used the literature data set as the reference data set and the BSP data set as the set of query molecules. On the other hand we used the BSP data set as the set of query structures and as the reference data set.

Using the literature data set as the reference data set and the BSP data set as the set of query structures results in 80% correct classifications within the classified compounds. However, topoHERG_180 gives a prediction only for about 5% of the molecules, highlighting the low structural overlap between the literature and the BSP data set.

When applying topoHERG_180 to the BSP data set in the leave-one-out like manner described above, 84% of all molecules are classified correctly, including a sensitivity of 76% and a specificity of 88%. However, the fraction of unclassified molecules is again high (37%). Applying the two-stage approach combining topoHERG_180 and topoHERG_270 increases this fraction of remaining unknowns to only 13%, whereas the accuracy decreases only slightly: 81% of all molecules are correctly classified, including a sensitivity of 69% and a specificity of 86%.

Finally, we used the 43 structures that were added to our initial data set containing 232 compounds as validation set of query structures and the 232 data set as the set of reference structures. We get predictions with topoHERG_180 for 8 molecules (19%). For the threshold pIC50 > 5, all compounds are predicted correctly, and for pIC50 > 6 only one compound is misclassified. Therefore, although the fraction of classified compounds is again rather low, the prediction accuracy for this test case is excellent because only one medium active compound is misclassified to be a strong hERG binder.

**Comparison with Other Similarity-Based Methods.** To underline the quality of our topoHERG_180 model, we compared our approach with other similarity-based models. Since there are no other published similarity-based hERG models, we decided to build similarity-based models using fingerprints. We used MACCS structural keys[45,46] and ECFP_4[47] fingerprints to capture similarities between the query molecule and the molecules in the reference data set using the Tanimoto Coefficient as the similarity measure. In analogy to the development process of our topoHERG models we compared different activity prediction functions and different thresholds for the maximum Tanimoto Coefficient between similar molecules.

**Table 3.** Comparison of topoHERG_180 with Further Similarity-Based hERG Classifiers Using MACCS and ECFP_4 Fingerprints

| | topoHERG_180 | MACCS fingerprints | ECFP_4 fingerprints |
|---|---|---|---|
| accuracy | 0.90 | 0.85 | 0.84 |
| sensitivity | 0.83 | 0.73 | 0.71 |
| specificity | 0.93 | 0.90 | 0.89 |
| unknowns | 0.54 | 0.54 | 0.54 |

We tested two different activity prediction functions: (i) a simple nearest neighbor approach (assign the $pIC_{50}$ class of the nearest neighbor to the query molecule) and (ii) an activity prediction function analogously to the topoHERG function, which uses all neighbors within a certain maximum distance to determine the activity of the query molecule:

$$S = \sum_{neighbors} TC * \begin{cases} 1 & active \\ -1 & inactive \end{cases}$$

$$Activity = \begin{cases} active & S \geq 0 \\ inactive & S < 0 \end{cases} \qquad (4)$$

A weighted sum $S$ considering all neighbors with a Tanimoto Coefficient lower than a certain threshold is determined. The weight of each neighbor is the Tanimoto Coefficient between the query molecule and the current neighbor, this weight is multiplied by 1 if the neighbor is hERG active, and it is multiplied by −1 if the neighbor is hERG inactive. If the overall sum $S$ is larger than zero, then the query molecule is classified to be hERG active; otherwise it is predicted to be hERG inactive. If there are no neighbors within a certain threshold both activity prediction functions return "unknown".

Additionally, we varied the threshold for the maximum Tanimoto Coefficient of similar molecules between 0.4 and 0.9. For ECFP_4 fingerprints and for MACCS fingerprints we compared the resulting accuracies for both activity prediction functions applied with the different maximum Tanimoto Coefficients. From the resulting set of different models we selected the best ECFP_4 model and the best MACCS model showing the identical number of unknowns as our topoHERG_180 model. The best MACCS fingerprint model is defined via the nearest neighbor activity prediction function with a maximum Tanimoto Coefficient of 0.8. The best ECFP_4 fingerprint model is based on eq 4 and a maximum distance between neighbored compounds of 0.5.

The comparison of these two fingerprint models with the topoHERG_180 model is shown in Table 3. For the identical number of unknowns our topoHERG_180 model is superior to both fingerprint-based models showing the power of the pharmacophore and shape-based Topomer Search approach. Of course all three similarity-based models have the problem of a large number of unclassified molecules, if the structural overlap between the query molecules and the molecules in the reference data set is too low. But the comparison of topoHERG_180 with fingerprint-based methods revealed that

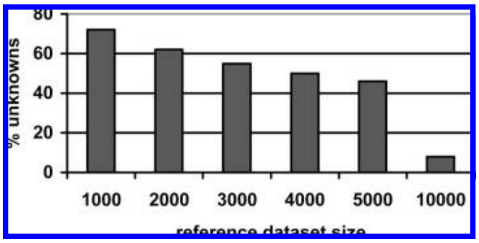


**Figure 5.** Percentages of unclassified molecules for differently sized structurally diverse data sets. We applied topoHERG_180 to random subsets of the BSP compound library containing 1000, 2000, 3000, 4000, 5000, and 10000 molecules and report the corresponding fractions of remaining unknowns. Since the probability to find Topomer Search neighbors increases with an increase in size and structural diversity of the reference data set, a reference data set containing about 10000 structurally diverse molecules is necessary to use topoHERG_180 as a global model.

topoHERG_180 outperforms the fingerprint-based models in detecting and handling the existing similarities.

**Requirements for a Global topoHERG Model.** The structural overlap between the query compounds and the reference data set is crucial and of course the limiting factor of all similarity-based models. Therefore, a large and structurally diverse reference data set is necessary for all similarity-based models to ensure that there are similar reference molecules for the majority of all query molecules. Since every new molecule enhances our knowledge and therefore the probability that a similar molecule for a query molecule can be found, topoHERG will improve with every new molecule added to the reference data set. A technological advantage of our approach is that retraining is not necessary when new molecules are added to the reference data set.

The fact that a large structurally diverse reference data set is necessary for a global similarity-based hERG model raises of course the question how large a reference data set for topoHERG_180 has to be to achieve an acceptable low percentage of unknowns. To answer this question we selected a random subset of the entire BSP compound library containing 10000 structurally diverse molecules. From this data set we selected again random subsets containing 1000, 2000, 3000, 4000, and 5000 molecules. Afterward we applied topoHERG_180 to these data sets in the leave-one-out like way described earlier. From this procedure we obtain the remaining percentages of unclassified molecules shown in Figure 5.

Since the probability to find Topomer Search neighbors increases with an increase in size and structural diversity of the reference data set, a reference data set containing about 10000 structurally diverse molecules is necessary to achieve a rate of unknowns lower than 10%. With a structurally diverse reference data set of about 10000 molecules, topoHERG_180 will act as a global hERG model. In this case, topoHERG acts as a set of models for different binding modes. Each subset of molecules in the reference data set showing the same SAR can be considered as the basis for a submodel for this special binding mode. Gavaghan et al.[48] presented a hERG model based on almost 9000 molecules, and the model published by O'Brien and de Groot[49] utilizes a data set containing about 60000 molecules, showing that the allocation of a large and structurally diverse data basis for a global topoHERG_180 model is feasible.

**Exemplary Output of the topoHERG Model.** The topoHERG output is shown in Figure 6. It gives columns for the 2D structure of the query molecule, the 2D structures of the corresponding Topomer Search neighbors, their distances, and their hERG activities for two different activity thresholds $pIC_{50} > 5$ and $pIC_{50} > 6$. Finally the query molecule's predicted hERG activities for these two activity thresholds are shown.

Due to this special output format, our topoHERG model can act as a knowledge base for hERG channel activity. Via the 2D structures of each Topomer Search neighbor and the corresponding hERG activity classes for the two different activity thresholds the user can deduce the underlying hERG SAR. If e.g. two neighbors are found and one is hERG active and one is hERG inactive, the user can visually inspect these neighbors to detect features that might reduce hERG channel activity. Additionally, our output format enables the user to evaluate the predicted hERG activities of the query molecules. The predictions of the topoHERG model are based on the neighbor's Topomer Search distances and their hERG channel activities. These values are displayed in the output, and thereby the user can rationalize and evaluate the prediction.

## SUMMARY AND CONCLUSIONS

Our proof of principle study revealed that the similarity property principle can be applied to predict hERG activity. Based on a Genetic Algorithm we developed two versions of our topoHERG model: topoHERG_180 and topoHERG_270. The topoHERG_180 model allows only very close neighbors to influence the predicted activity of the corresponding query molecule, resulting in an excellent accuracy due to the high similarity between the query molecule and the neighbors found via Topomer Search. A drawback of this strict similarity threshold is that a large and structurally diverse reference data set is required for a global model. In our literature data set topoHERG_180 is predictive in only about 50% of all cases. If similar compounds are found, the predicted hERG activity is highly accurate. The topoHERG_270 model also allows more distant neighbors to participate in the prediction of the query molecule's hERG activity. Therefore the remaining number of unclassified molecules decreases significantly, but the prediction accuracy also decreases. Since topoHERG_180 shows a significantly better accuracy than topoHERG_270, we propose to use these two models in a two-stage approach. Applying the combined topoHERG_180 and topoHERG_270 to the public data set results in 76% correctly classified molecules, and an application to an independent test set results in 81% correct classifications.

A comparison of topoHERG with other similarity-based approaches using fingerprints revealed that topoHERG outperforms the fingerprint-based models showing the power of the pharmacophore and shape-based Topomer Search approach.

The topoHERG approach offers several appealing advantages over conventional QSAR models:

The topoHERG model output format provides the user with a knowledge-base for hERG channel activity. The reference data set contains the available knowledge about

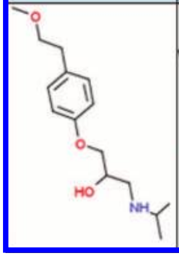
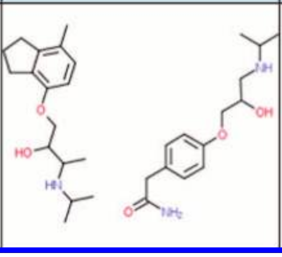
| Query molecule | Topomer Search neighbors | Distances | Activities (10 μM) | Activities (1 μM) | Predicted activity (10 μM) | Predicted activity (1 μM) |
|---|---|---|---|---|---|---|
|  |  | 174<br>93 | active<br>inactive | inactive<br>inactive | inactive | inactive |

**Figure 6.** Exemplary output of the topoHERG_180 model using metoprolol with a $pIC_{50}$ of 3.84 as query molecule. The output includes the 2D structure of the query molecule and the 2D structures of the corresponding Topomer Search neighbors in columns 1 and 2. The Topomer Search distances and the hERG activities for two different activity thresholds ($pIC_{50} > 5$ and $pIC_{50} > 6$) of these Topomer Search neighbors are shown in columns 3, 4, and 5. Finally the query molecule's predicted hERG activity for the two activity thresholds are shown in columns 6 and 7.

hERG binders and hERG nonbinders. The model predictions are reported in a way to allow the user to learn from the neighbors which structural features prevent hERG channel binding and also evaluate the predicted activity.

The second major advantage is that topoHERG provides a strict measure of the domain of applicability, in that it classifies only molecules with close similarity to the known cases.

Third, topoHERG provides intrinsic submodels for the different SARs in hERG channel binding. Each subset of molecules in the reference data set showing its own binding mode displays the basis for the submodel for this special binding mode.

Finally, every new molecule which is added to the reference data set, without retraining, will enhance the knowledge about hERG activity and therefore the probability that topoHERG finds similar molecules for a query compound.

The downside of our proof of principle study based on the available literature data is the low number of experimental values in our knowledge base. When topoHERG_180 is applied to the BSP test set using the literature data set as the reference data set, a prediction is given for only 5% of all compounds due to the low structural overlap between the BSP and the literature data set. Our model is only capable to interpolate within the chemical space of the reference data set but cannot be extended beyond this. However, this is the classical failure of all SAR approaches, and it is questionable whether the accuracy and prediction confidence of other QSAR approaches for chemicals beyond the chemical space of the training set requiring extrapolation can be specified.[50] However, we were able to show in a benchmark study that our topoHERG_180 model can be used as a global model for hERG channel activity when a reference data set containing about 10000 carefully selected structurally diverse molecules is available. The selection process of these compounds should be based on Topomer Search dissimilarities.

We expect that the approach is applicable for a variety of other ADMET end points for off-targets with unspecific binding and multiple SARs. Among these rather unspecific end points are CYP inhibition, HSA binding, or blood-brain barrier penetration.

Our proof of principle study reveals that applying the similarity property principle using Topomer Search as a similarity measure displays a promising alternative to classical QSAR and pharmacophore modeling of hERG channel binding.

## ACKNOWLEDGMENT

The authors thank Richard Cramer and Bernd Wendt from Tripos for the fruitful discussions and support with Topomer Search and Andreas Sutter, Mario Lobell, and Herbert Himmel for the their support on the preparation of the data sets.

**Supporting Information Available:** Descriptors used to validate the use of neighbors with a Topomer Search distance up to 270 (Table S1), data set containing 681 hERG patch-clamp measurements on mammalian cell lines (HEK, CHO, COS) and corresponding references (Table S2), and SD-File containing 275 molecules from a literature data set. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Warmke, J. W.; Ganetzky, B. A family of potassium channel genes related to eag in Drosophila and mammals. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 3438–3442.

(2) Mitcheson, J. S.; Perry, M. D. Molecular determinants of high-affinity drug binding to hERG channels. *Curr. Opin. Drug. Discovery Dev.* **2003**, *6*, 667–674.

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

(4) Antzelevitch, C.; Shimizu, W. Cellular mechanisms underlying the long QT syndrome. *Curr. Opin. Cardiol.* **2002**, *17*, 43–51.

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

(6) De Ponti, F.; Poluzzi, E.; Montanaro, N. QT-interval prolongation by non-cardiac drugs: Lessons to be learned from recent experience. *Eur. J. Clin. Pharmacol.* **2000**, *56*, 1–18.

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

(8) Wood, C.; Williams, C.; Waldron, G. J. Patch clamping by numbers. *Drug Discovery Today* **2004**, *9*, 434–441.

(9) Brown, A. Drugs, hERG and sudden death. *Cell Calcium* **2004**, *35*, 534–546.

(10) Dubin, A. E.; Nasser, N.; Rohrbacher, J.; Hermans, A. N.; Marrannes, R.; Grantham, C.; Van Rossem, K.; Cik, M.; Chaplan, S. R.; Gallacher, D.; Xu, J.; Guia, A.; Byrne, N. G.; Mathes, C. Identifying modulators of hERG channel activity using the PatchXpress planar patch clamp. *J. Biomol. Screen.* **2005**, *10*, 168–181.

(11) Bridgland-Taylor, M. H.; Hargreaves, A. C.; Easter, A.; Orme, A.; Henthorn, D. C.; Ding, M.; Davis, A. M.; Small, B. G.; Heapy, C. G.; Abi-Gerges, N.; Persson, F.; Jacobson, I.; Sullivan, M.; Albertson,

N.; Hammond, T. G.; Sullivan, E.; Valentin, J.-P.; Pollard, C. E. Optimization and validation of a medium-throughput electrophysiology-based hERG assay using IonWorks HT. *J. Pharmacol. Toxicol. Methods* **2006**, *54*, 189–199.

(12) Thai, K.-T.; Ecker, G. F. Predictive models for hERG channel blockers: Ligand-based and structure-based approaches. *Curr. Med. Chem.* **2007**, *14*, 3003–3026.

(13) Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recantini, M. Towards 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.

(14) Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. Three-dimensional quantitative structure-activity relationships for inhibition of human ether-a-go-go-related gene potassium channel. *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427–434.

(15) Pearlstein, R. A.; Vaz, R. J.; Kang, J.; Chen, X.-L.; Preobrazhenskaya, M.; Shchekotikhin, A. E.; Korolev, A.; 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.

(16) Li, Q.; Jorgensen, F. S.; Oprea, T.; Brunak, S.; Taboureau, O. hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharm.* **2007**, *5*, 117–127.

(17) Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z. Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol. Sci.* **2004**, *79*, 170–177.

(18) Seierstad, M.; Agrafiotis, D. K. A QSAR model of hERG binding using a large, diverse, and internally consistent training set. *Chem. Biol. Drug. Des.* **2006**, *67*, 284–296.

(19) Roche, O.; Trube, G.; Zuegge, J.; Pfimlin, 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.

(20) Gepp, M. M.; Hutter, M. C. Determination of hERG channel blockers using a decision tree. *Bioorg. Med. Chem.* **2006**, *14*, 5325–5332.

(21) Dubus, E.; Ijjaali, I.; Petitet, F.; Michel, A. In silico classification of hERG channel blockers: A knowledge-based strategy. *ChemMedChem* **2006**, *1*, 622–630.

(22) Sun, H. An accurate and interpretable Bayesian classification model for prediction of hERG liability. *ChemMedChem* **2006**, *1*, 315–322.

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

(24) Thai, K.-M.; Ecker, G. F. A binary QSAR model for classification of hERG potassium channel blockers. *Biorg. Med. Chem.* **2008**, *16*, 4107–4119.

(25) Kramer, C.; Beck, B.; Kriegl, J. M.; Clark, T. A composite model for hERG blockade. *ChemMedChem* **2007**, *3*, 254–165.

(26) *Concepts and applications of molecular similarity*; Johnson, M., Maggiora, G. M., Eds.; John Wiley and Sons: New York, 1990.

(27) Maggiora, G. M. On outliers and activity cliffs - Why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(28) Peltason, L.; Bajorath, J. SAR Index: Quantifying the nature of structure-activity relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.

(29) Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

(30) Guha, R.; Van Drie, J. H. Assessing how well a modeling protocol captures a structure-activity landscape. *J. Chem. Inf. Model.* **2008**, *48*, 1716–1728.

(31) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.

(32) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.

(33) Kubinyi, H. Chemical similarity and biological activities. http://www.kubinyi.de//dd-06.pdf (accessed October 15, 2008).

(34) Cramer, R. D.; Jilek, R. J.; Andrews, K. M. Dbtop: Topomer similarity searching of conventional databases. *J. Mol. Graphics Modell.* **2002**, *20*, 447–462.

(35) 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.

(36) Jilek, R. J.; Cramer, R. D. Topomers: A validated protocol for their self-consistent generation. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1221–1227.

(37) Sivanandam, S. N.; Deepa, S. N. *Introduction to Genetic Algorithms*; Springer: Berlin, Heidelberg, 2007.

(38) Devillers, J. *Genetic Algorithms in Molecular Modeling*; Academic Press: New York, 1996.

(39) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(40) Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319, U.S.A. Dbtop documentation as provided for Beta-testing in 2006.

(41) *Pipeline Pilot, version 6.1.5*; Accelrys Inc.: 10188 Telesis Court, Suite 100, San Diego, CA 92121, U.S.A.

(42) Buyck, C.; Tollenaere, J.; Engels, M.; Clerck, F. D. An in silico model for detecting potential hERG blocking. Poster presentation. Euro-QSAR 2002, Bournemouth, Sep 8−13, 2002.

(43) Aronov, M.; Goldman, B. B. A model for identifying hERG $K^+$ channel blockers. *Bioorg. Med. Chem.* **2004**, *12*, 307–2315.

(44) 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.

(45) McGregor, G. B.; Pallai, P. Clustering of large databases of compounds: Using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.

(46) Durant, J.; Leland, B.; Henry, D.; Nourse, J. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(47) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity-fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682–686.

(48) Gavaghan, C. L.; Hasselgren Arnby, C.; Blomberg, N.; Strandlund, G.; Boyer, S. Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 189–206.

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

(50) Tong, W.; Hong, H.; Xie, Q.; Shi, L.; Fang, H.; Perkins, R. Assessing QSAR Limitations - A Regulatory Perspective. *Curr. Comput.-Aided Drug Des.* **2005**, *11*, 195–205.

CI800304T