JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

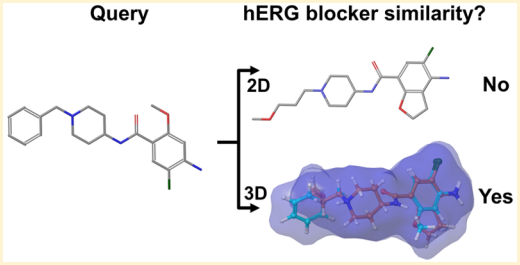
General Purpose 2D and 3D Similarity Approach to Identify hERG Blockers

Patric Schyman,* Ruifeng Liu, and Anders Wallqvist*

DoD Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, 2405 Whittier Drive, Frederick, Maryland 21702, United States

Supporting Information

ABSTRACT: Screening compounds for human ether-à-go-go-related gene (hERG) channel inhibition is an important component of early stage drug development and assessment. In this study, we developed a high-confidence (p-value < 0.01) hERG prediction model based on a combined two-dimensional (2D) and three-dimensional (3D) modeling approach. We developed a 3D similarity conformation approach (SCA) based on examining a limited fixed number of pairwise 3D similarity scores between a query molecule and a set of known hERG blockers. By combining 3D SCA with 2D similarity ensemble approach (SEA) methods, we achieved a maximum sensitivity in hERG inhibition prediction with an accuracy not achieved by either method separately. The combined model achieved 69% sensitivity and 95% specificity on an independent external data set. Further validation showed that the model correctly picked up documented hERG inhibition or interactions among the Food and Drug Administration- approved drugs with the highest similarity scores—with 18 of 20 correctly identified. The combination of ascertaining 2D and 3D similarity of compounds allowed us to synergistically use 2D fingerprint matching with 3D shape and chemical complementarity matching.



INTRODUCTION

Early assessment of absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties[1,2] is an important stepping stone in drug development. One of the toxicology screens that compounds must pass during early preclinical studies involves the human ether-à-go-go-related gene (hERG), a potassium ion channel involved in the normal cardiac repolarization activity of the heart.[3] Drug-induced blockade of the hERG function can cause long QT syndrome and can lead to arrhythmia and death.[4] Several experimental techniques are available to assess the hERG function; the most reliable being patch-clamp electrophysiological recording of the K^+ current.[5] The use of these methods is not routinely implemented in the early stages of drug design, and hence, computational methods that can assess and accurately predict hERG liabilities are of considerable interest.

Although the crystal structure of the hERG channel is not yet known, it shares structural similarities with other voltage-gated potassium (K_v) ion channel family members. Mutation experiments have identified crucial residues for binding, e.g., two polar residues (Thr623 and Ser624) located in the base of the channel and two hydrophobic residues (Tyr652 and Phe656) in the S6 region.[6] From the variety of different molecules that interact with hERG and the different response to site-directed mutation, it is clear that not all molecules bind in the same region of the channel. Although structure-based methods, e.g., molecular dynamics (MD) and docking calculations using homology modeling of the hERG ion channel[7−9] have been used, it remains a challenge to develop general and accurate prediction models based on detailed molecular structure interactions.

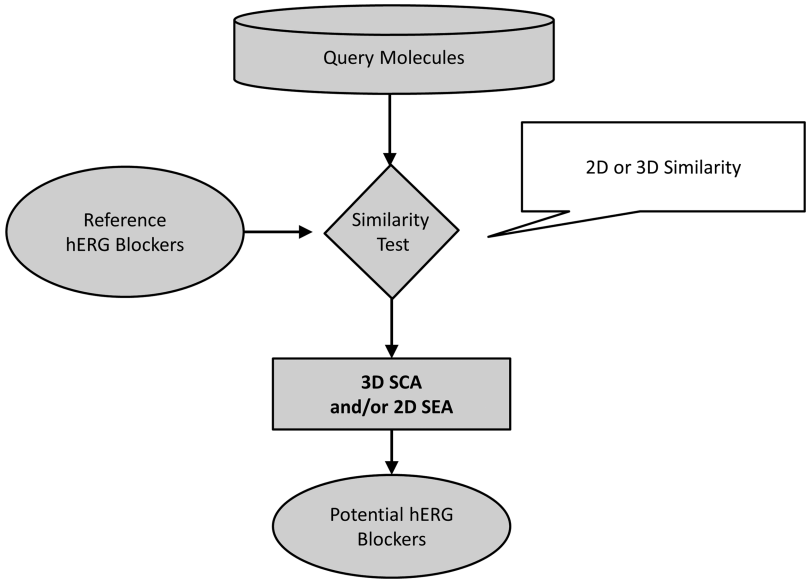
Nonetheless, hERG blockers share common molecular features, e.g., they often have (1) a positive ionizable center, (2) a hydrophobic group, (3) a V-shaped geometry, (4) a molecular weight of >250, and (5) a ClogP value of >1.[10,11] This has resulted in the development of general purpose ligand-based computational methods to predict hERG inhibition, e.g., quantitative structure activity relationships (QSARs), *k* nearest neighbor (*k*-NN), support vector machine (SVM), random forest (RF), and naive Bayesian classification (NBC).[10,12,13] Although the best methods give accuracies in the range of 70% to 90%, most of these studies use either small data sets or in-house data sets that are not publicly available for training and/or testing, which makes a fair comparison of the performance of different methods impossible.

In this study, we developed a hERG prediction tool based on 3D similarity scores between multiple molecular conformations, which we named the similarity conformation approach (SCA). The SCA shares some resemblance to the 2D similarity ensemble approach (SEA) introduced by Shoichet and co-workers,[14] which they originally used to relate proteins using setwise 2D chemical similarity of their ligands. Here, we compared 3D SCA and 2D SEA and combined them for

Received: October 9, 2015
Published: December 30, 2015

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222

Scheme 1. Flowchart of the 2D SEA and 3D SCA Procedure



optimal performance to identify hERG blockers. Although we focused on hERG for this report, the method can easily be applied to any protein with known inhibitors/substrates. The aim was to create a robust computational tool that will predict possible hERG blockers with high confidence and low false-positive rates to ensure that we do not dismiss otherwise perfectly acceptable drug candidates. It should be noted that several drugs currently on the market do inhibit hERG as a known side effect. The 3D SCA method predicted ~60% of our data set of hERG blockers with high confidence (p-value <0.01), and a specificity of 99%, which in comparison with other methods in the literature, is one of the highest specificity values.[15−21] The 3D SCA has advantages over 2D SEA as 3D methods can identify molecules with different scaffolds with similar shape and bioactivity.[22] However, a complete enumeration and coverage of conformational space is not practical, and, hence, we combined the 3D SCA with 2D SEA to achieve a sensitivity of 65% on the set of hERG blockers used in model-building and 69% sensitivity on an external test set.[15] Furthermore, retraining of the SCA component is simple and straightforward and does not include a molecular size-dependent distribution analysis to determine parameters, as is required by the 2D SEA method.

METHODS AND MATERIALS

hERG Data Sets. We retrieved 282 hERG blockers from the work of Wang et al.,[15] which includes data collected by Li et al.[23] and the Wombat-PK data set. Although other hERG data are available, primarily high-throughput binding and functional assay data in ChEMBL, we built our model based on hERG inhibition IC_{50} values measured in primarily patch-clamp electrophysiology assay using Chinese hamster ovary (CHO) and human embryonic kidney (HEK) mammalian cell lines. We chose an IC_{50} value of $\leq$10 μM as a cutoff value for hERG blockers. This represents a practical and convenient choice for early stage screening and assessment. In later stages of drug discovery, one can use the therapeutic window to better select a tolerable level of hERG inhibition. If a drug candidate is potent at a low therapeutic dose, hERG inhibition may not be an issue even if an in vitro assay classifies the active compound as a hERG channel inhibitor. However, in early stages of drug discovery, compound potency is typically not optimized, and no reliable estimate of the therapeutic window exists. Most published literature studies have adopted the 10 μM IC_{50} value as a practical cutoff to use. All data sets are provided as Supporting Information. In the absence of a large and experimentally verified set of known non-hERG blockers, we used 25 000 diverse and randomly chosen molecules obtained from the National Cancer Institute (NCI). We selected the diverse set of molecules from the NCI library using the following procedure: (1) filtering the library contents to only contain organic compounds, (2) stripping salts and standardizing charges and stereo representations, (3) protonating and deprotonating acids and bases, respectively, (4) removing duplicate structures, (5) removing structures with >10 rotatable bonds and a molecular weight of <150 Da, and (6) selecting 25 000 organic molecules by structure dissimilarity based on extended connectivity footprints (ECFP_4)[24] and Tanimoto similarity. Although this set will contain a small fraction of hERG blockers, it avoids comingling data from, for example, congeneric series that are typically provided in the literature data, which tend to be associated with residual hERG activity due to structural similarity to the active hERG blockers. If we restricted the nonblocking data to these smaller data sets, we could derive an accurate model but with a severely limited applicability domain.[25] Absence of hERG activity was assessed from 400 known non-hERG blockers obtained from functional assay data deposited in the ChEMBL database[26] with an IC_{50} value of >10 μM.[21]

Additional Data Sets. An external test set of 120 molecules from Wang et al.[15] was used for evaluation. At an IC_{50} delimiter value of 10 μM, the data set consists of 54 hERG blockers and 66 non-hERG blockers. The 120 hERG activity data are collected from available IC_{50} measurements in the literature using primary mammalian cell line data from HEK and CHO. However, when mammalian data are not available, nonmammalian cell lines such as *Xenopus laevis* oocytes are included. As a further external validation, we used the known drugs from Drug Bank,[27] which, after removing hERG

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222

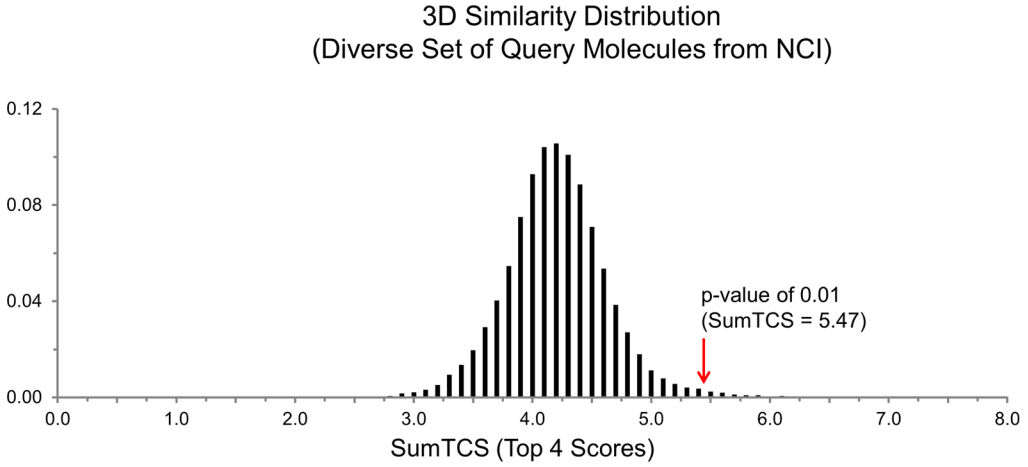


Figure 1. 3D similarity score distribution. The distribution of scores was derived from analyzing a large and diverse set of 25000 query molecules based on similarity to 282 known reference hERG blockers. The p-value 0.01 is marked in the figure where 99% of the queries have a SumTCS value of < 5.47. NCI, National Cancer Institute; SumTCS, sum of Tanimoto combo score.

molecules included in our training set, yielded 1072 drug molecules suitable for computational testing.

2D SEA Method. The SEA method is described elsewhere[14] and is only briefly outlined here. The 2D SEA method relates a query molecule with a set of known hERG blockers via 2D Tanimoto similarity (Schematic 1). A *raw score* is calculated by summing all the Tanimoto coefficients (Tcs) between the set of query molecules and the set of hERG blockers. The raw score is strongly dependent on the number of molecules in the two sets. The size-independent *z*-score is calculated from the raw score with the mean and standard deviation values that are modeled as functions of the product of the ligand sets sizes. To emphasize ligands with high similarity, a similarity cutoff value is applied below which Tc values do not contribute to the raw score. Varying the cutoff value and finding which *z*-score distribution best fits to an extreme value distribution determines the optimal cutoff value. In our study, we summed all Tc values above the optimal cutoff value of 0.57 between a query molecule and any hERG reference molecules to give a total Tc (TTc). The higher the TTc value, the more likely it is that the query is a hERG blocker. As molecular descriptors for identifying structurally similar compounds, we used Accelrys extended connectivity fingerprint within a diameter of four chemical bonds (ECFP_4)[24] and a bit size of 2048. The TTc values are therefore dependent on the data set and the chosen fingerprints.

3D SCA Method. Our 3D SCA method used the Tanimoto combo score (TCS), which includes two terms: 3D structure similarity (shape) and pharmacophore property similarity (color) to calculate the 3D similarity and a summation of an optimal fixed number of top TCSs, which gave a sum of TCS (SumTCS).

By using a fixed number of similarity scores, we could simplify the procedure to train our model compared with the more extreme retraining required by the SEA. The primary reason we did not implement a 3D SEA method is the requirement of a large enough data set to be able to accurately fit an extreme value distribution to the size dependent *z*-score mean and standard deviation values. The currently available hERG inhibition data did not allow us to make a reliable fit. In order to overcome this issue, we instead developed the 3D SCA method with a fixed number of similarity hits for each query. We calculated the SCA conformational and similarity calculations in the following manner: The query molecules were represented by a set of multiple low-energy (<8 kcal/mol) conformations and a root-mean-square distance of >0.8 Å to avoid degenerate conformations. We calculated the similarity score for all pairs between a set of query molecules and the set of hERG reference molecules, including their conformations (Schematic 1). We used OMEGA[28,29] (Open Eye Scientific Software, Santa Fe, New Mexico) to generate molecular confirmations for both query and reference molecules with the following parameters: number of maximal conformations (*maxconfs*) 200, root-mean-square distance (*rms*) 0.8 Å, maximum allowed conformational energy (*ewindow*) 8 kcal/mol, maximal rotational bonds (*maxrot*) 10, and maximum number of stereo centers (*flipper_maxcenter*) 4. This resulted in large conformational data sets, e.g., the total number of conformations for the 282 hERG blockers were 10 377, and roughly one million for the 25 000 random molecules. Although we used multiple conformations to represent each molecule, the number of conformations that we can practically include imposes a limitation of 3D structure similarity approaches. As any discretization of conformational space is limited, we could potentially classify two similar molecules as dissimilar if their conformations do not match within the chosen discretization scheme. It should be noted that we did considered chirality in the 3D SCA method. If chirality of an input structure was given, only conformers of the specific chirality were generated. If chirality of an input structure was not given, conformers of all possible chirality were generated.

We used Rapid Overlay of Chemical Structures (ROCS)[30,31] software to perform the 3D alignment based on shape and chemical similarity. ROCS compares the set of queries with the hERG reference molecules and returns a TCS value between 0 and 2. If the score is 2, both the molecular shape and the chemical property match; if the score is 0, there is no shape or chemical property match. These calculations are very fast for a few molecules, but for our exhaustive conformational calculations the number of overlay calculations are substantial, e.g., approximately 10 billion overlay calculations are needed for evaluating the diverse set of 25 000 molecules for potential hERG blockers.

We performed two steps to find the optimal sum. The first step was to run a 3D similarity test using a diverse set of 25 000 molecules from NCI as query molecules to determine the

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222

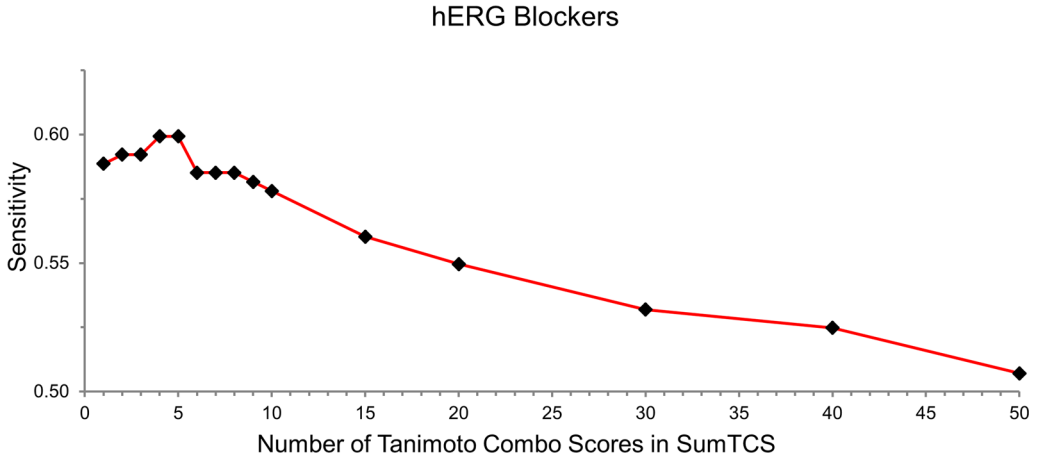


Figure 2. Optimizing the sensitivity. The true positive ratio (sensitivity) of hERG blockers based on including increasing numbers of highest Tanimoto combo scores.

SumTCS value that corresponds to a p-value of 0.01. Figure 1 shows the distribution for a diverse set of randomly selected molecules with the SumTCS of the four highest similarity scores. In the second step, we used the SumTCS value to determine how many correctly identified hERG blockers (sensitivity) we found using a 5-fold cross-validation on our set of hERG blockers. Figure 2 shows the sensitivity when we varied the number of top scores summed ranging from 1 to the top 50 scores. The corresponding specificity values were 99% if we considered that all random molecules are non-hERG blockers, which is an exaggeration, since most likely a small fraction of the randomly selected molecules are in fact hERG blockers. The best performance occurred when adding four or five top similarity scores. In this study, we chose the sum of four, which corresponds to a SumTCS cutoff value of 5.47. It is noteworthy that by just including the best similarity score the sensitivity was still good (0.59), and that after using >5 values in the summation, the sensitivity declined (<0.59). The requirement of adding four scores implies that we need a minimum of four high-scoring conformations for a positive hit. The procedure maximized the true positive hits by reducing the average pairwise similarity score required to maintain a p-value of <0.01.

Model Performance Measures. We used the following metrics to measure the quality of the classification models:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{2}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

$$\text{kappa} = \frac{\text{accuracy} - \Pr(\text{e})}{1 - \Pr(\text{e})} \tag{4}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Kappa is a metric for assessing the quality of binary classifiers,[32] and Pr(e) is an estimate of the probability of a correct prediction by chance. It is calculated as

$$\Pr(\text{e}) = \frac{(\text{TP} + \text{FN})(\text{TP} + \text{FP}) + (\text{FP} + \text{TN})(\text{TN} + \text{FN})}{(\text{TP} + \text{TN} + \text{FP} + \text{TN})^2} \tag{5}$$

The sensitivity is a measure of the model's ability to correctly detect true positive and the specificity measures the model's ability to detect true negative. Kappa compares the probability of correct predictions to the probability of correct predictions by chance. Its values range from +1 (perfect agreement between model prediction and experiment) to −1 (complete disagreement), with 0 indicating no agreement beyond that expected by chance.

RESULTS AND DISCUSSION

2D Similarity Ensemble Approach Applied to hERG. In the 2D SEA method, we summed all similarity coefficients >0.57 for each query that is similar to any of the hERG reference molecules (the ensemble). The basic idea is that a query that shows similarity with several hERG reference molecules is more likely to be a hERG blocker than a query with just one similar reference molecule, however, that is not always necessarily true.

Table 1 shows the 2D results from the data set containing 282 known hERG inhibitors with IC_{50} values ≤10 μM and a test set of 400 molecules with IC_{50} values >10 μM. It should be noted that several of the inactive molecules are structurally similar to some of the potent hERG inhibitors, and hence, they tend to have some IC_{50} activity and could be classified as weak hERG inhibitors. As discussed in the Methods and Materials section, use of the existing small data sets for non-hERG blockers would severely limit the applicable chemical space for predicting possible nonblockers.

Table 1. Number of Predicted hERG Blockers (Sensitivity) Compared with a Set of Nonblockers and a Set of 25 000 Diverse Molecules

data set	2D SEA	3D SCA	2D SEA + 3D SCA
blockers	143 (50.7%)	165 (58.5%)	182 (64.5%)
nonblockers	18 (4.5%)	38 (9.5%)	40 (10.0%)
diverse molecules	28 (0.1%)	226 (1.0%)	230 (1.0%)

To evaluate the 2D SEA model performance on the data set, we used 5-fold cross-validation by randomly dividing the data

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222

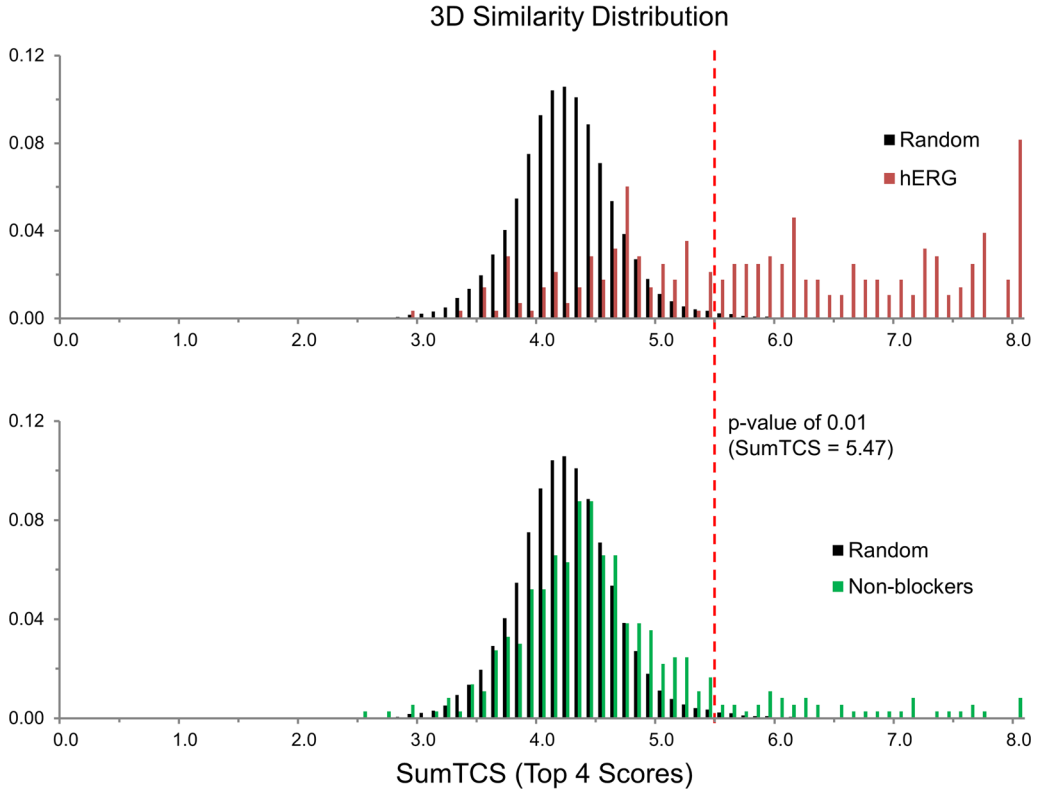


Figure 3. 3D similarity distribution of a set of random molecules compared with hERG blockers (top) and non-hERG blockers (bottom).

set into five equal-sized groups, where four groups were used as a training set and the fifth group as a test set of hERG blockers. The process was repeated five times. The goal of our hERG model was to predict potential hERG blockers with high confidence.

The 2D SEA method correctly identified 51% of the hERG blockers (true positives), while only 5% of the nonblockers were incorrectly identified as blockers (false positives). We also examined a diverse set of 25 000 molecules, and the 2D SEA method identified <1% as potential hERG blockers, reaffirming the high specificity of the model.

3D Similarity Conformation Approach Applied to hERG. The developed 3D SCA is a variety of the similarity ensemble approach developed by Shoichet and co-workers.[14] While the method shares similarities with other approaches that use multiple conformations to represent reference and query molecules, here we combined it with an optimized sum of similarity scores (Figure 2). Each hERG blocker in the reference data set was represented by multiple low-energy (<8 kcal/mol) conformations. In the same manner, a set of multiple conformations was generated for each query molecule. The top four TCSs were summed to yield a SumTCS value for each query and all reference molecules. If the sum exceeded 5.47, the query was classified as a hERG blocker, which corresponds to a p-value <0.01 (see Methods and Materials).

To test the 3D SCA performance, we used a similar 5-fold cross-validation procedure on the same data set that we used for the 2D SEA (Table 1). Figure 3 shows the SumTCS distributions for a set of hERG blockers (<10 μM) and a set of 400 non-hERG blockers (>10 μM) compared with a set of random molecules. The hERG distribution can be distinguished from the random distribution with 59% of the hERG blockers having a p-value of <0.01. The set of nonblockers overlap more with the diverse set but have a right tail indicating that there are some molecules that share similarity with hERG blockers. This is due to the set of 400 nonblockers containing roughly 10% of compounds that are structurally similar to the set of hERG blockers but do not meet the experimental criteria of <10 μM for being an inhibitor.

To compare the 3D SCA method with the 2D SEA, we used the set of 400 non-hERG blockers as negatives and the set of 282 hERG blockers as positives. Table 2 shows the overall accuracy, sensitivity, and kappa value for the 2D and 3D methods. The accuracy and kappa values are similar for both methods, showing that 2D SEA and 3D SCA could accurately predict potential hERG blockers. However, there was a noticeable difference in the increased sensitivity (true positive) to 59% for the 3D method, compared with 51% for the 2D SEA.

Table 2. Method Performance on a Test Set of 282 hERG Blockers and 400 Non-hERG Blockers

method	accuracy	sensitivity	specificity	kappa
2D SEA	77%	51%	96%	0.49
3D SCA	77%	59%	91%	0.51
2D SEA + 3D SCA	79%	65%	90%	0.56

Understanding and Analyzing 2D and 3D Differences. At first glance, it seems as though the 2D and 3D performances are very similar, but the overall performance hides important detailed differences. To highlight these differences, we first created one cluster of all hERG blockers based on maximal dissimilarity where molecules near the center of the cluster have high similarity and molecules far away from the center have low similarity. We then selected 50% of the hERG blockers that are

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222

structurally most similar to the cluster center as our high-similarity compounds. We then removed the next 25% most structurally similar hERG blockers and kept the remaining 25% as a low-similarity test set. The latter set contained the structurally most dissimilar molecules compared with the cluster center. Figure 4 shows the different regions with

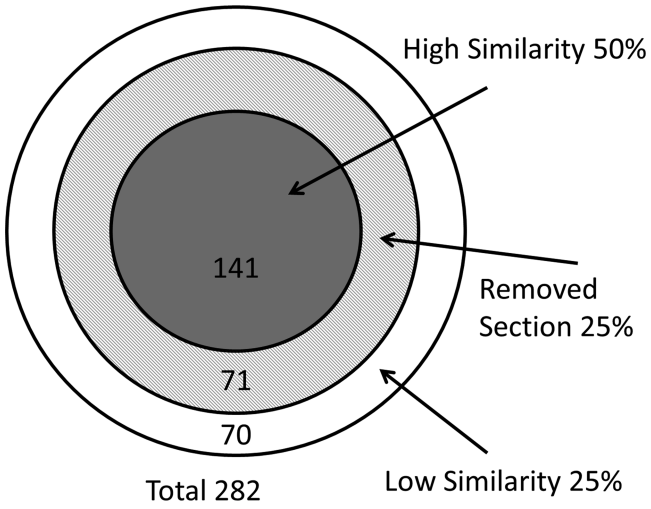


Figure 4. Schematic of the 2D similarity distribution relative to the cluster center (high similarity) and to the outer region with most dissimilar compounds (low similarity).

molecules with high similarity in the center and low similarity in the outer region. We then used the two data sets of the center and outer region to study differences of the 2D versus 3D similarity approaches, as summarized in Table 3.

Table 3. Calculated Sensitivities in Center and Outer Region for Different Methods

predicting hERG	2D SEA	3D SCA	2D SEA + 3D SCA
high-similarity compounds (center region)	50%	52%	60%
low-similarity compounds (outer region)	10%	23%	24%

The predicted sensitivities with the 2D and 3D methods for the center region were similar, 50% and 52%, respectively, when using 5-fold cross-validation. However, the 2D and 3D methods identified different compounds, and by combining the two results, the sensitivity increased to 60%. Approximately 70% of the high-similarity compounds were predicted by both methods, indicating that ~30% of the hERG blockers are uniquely identified in either the 2D or 3D method.

The difference between the 2D and 3D approaches becomes more noticeable when predicting the molecules in the outer region, using the compounds in the center high-similarity region as reference molecules. The 2D SEA correctly identified 10%, while the 3D SCA identified 23% of the low-similarity compounds. Eighty-six percent of the compounds detected by the 2D method were also found by our 3D SCA method. This shows the strength of using 3D over 2D similarity matching, in that it can detect compounds that do not necessarily have similar molecular scaffolds but have similar overall chemical structures.

Figures 5 and 6 show examples of query molecules that have high 3D similarity scores but low 2D similarity scores. The

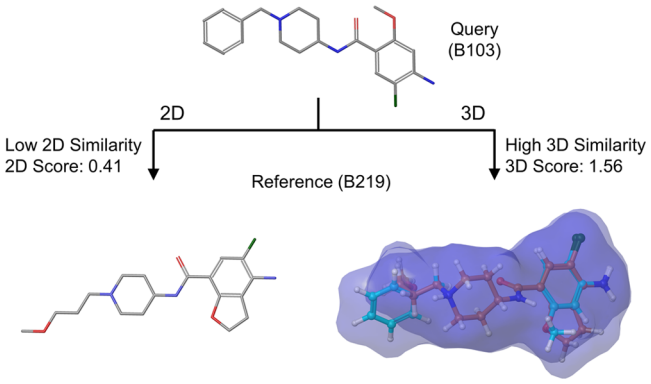


Figure 5. Example of 2D versus 3D similarity matching. In spite of low 2D fingerprint similarity between the query (B103) and the reference molecule (B219) they have high 3D chemical similarity in space and atom type overlap for certain conformations, as seen from the overlay picture.

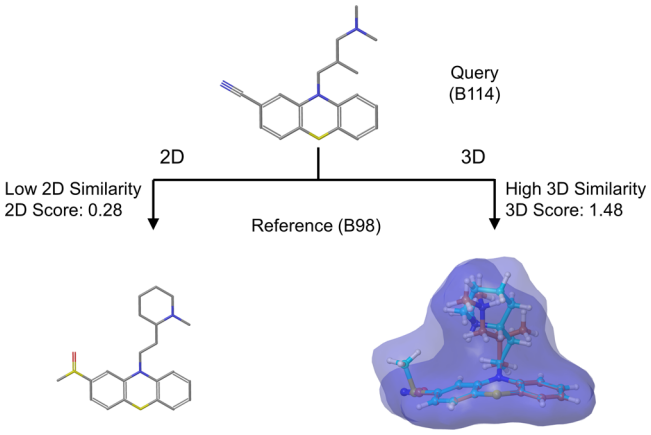


Figure 6. Example of 2D versus 3D similarity matching. The 2D fingerprint similarity between the query (B114) molecule and reference molecule (B98) is low with the major difference being that the B114 had a 1-methylpiperidine group where the B98 had a trimethylamine. This difference still gave a high 3D similarity score, as the volumes of these two compounds were almost the same.

query molecule B103 in Figure 5 had a low 2D fingerprint similarity score with the best-matching hERG reference molecule (B219) as their scaffold. Although these were quite different, the 3D shape matched and gave a high 3D similarity score, as can be seen in the 3D molecular overlay. Figure 6 shows another example, where the B114 query molecule was best matched with the reference molecule B98. The major difference was that B114 had a 1-methylpiperidine group, whereas the B98 had a trimethylamine. This change resulted in a low 2D similarity score as the atom types differ, but as can be seen in the 3D overlay picture the volumes of these two compounds were almost the same. In these two examples, the 2D similarity score failed to pass the 0.57 Tc cutoff, but the 3D SCA method identified these as positive hits, which used the sum of four conformational overlap scores that met the SumTCS cutoff of 5.47. These examples highlighted differences between 2D and 3D similarity matching and pointed toward the possibility of combining the methods.

Combined 2D SEA and 3D SCA Model Applied on hERG. Based on these results, we combined our 2D and 3D methods to increase the sensitivity without sacrificing the overall performance of the 3D SCA method. Tables 1, 2, and 3

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222

show the results from the combined 2D and 3D methods for identifying hERG blockers and the overall improvement of the predictions. We observed the greatest improvement in the calculated sensitivity when combining 2D SEA and 3D SCA, which increased from 59% to 65%. At the same time, we did not significantly increase the false positives for the data set of non-hERG blockers or for the diverse data set. The overall accuracy, sensitivity, and kappa value for the combined 2D SEA and 3D SCA method were 79%, 65%, and 0.56, respectively. These results showed that combining 2D SEA and 3D SCA approaches improved the prediction of hERG blockers.

External Test Set. It is almost impossible to make an accurate and impartial comparison between methods as there are many variables involved, such as which IC_{50} value was used to select hERG blockers and nonblockers for the training set. In an effort to qualitatively compare our 3D SCA method with another method, we chose a test set from Wang et al.,[15] which they use to assess their Bayesian classifier model based on molecular properties and the ECFP_8 fingerprints. We removed any molecules from our set of hERG blockers that also appeared in this test set. Table 4 shows the comparison of specificity and sensitivity between the different methods with an IC_{50} threshold value for hERG blockers at 10 μM.

Table 4. Calculated Sensitivities and Specificities on an External Test Set

method	sensitivity	specificity
2D SEA	61%	98%
3D SCA	64%	97%
2D SEA + 3D SCA	69%	95%
Bayesian classifier[15]	70%	88%

The 2D SEA and 3D SCA methods gave lower sensitivity values of 61% and 64%, respectively, compared to 70% for the Bayesian classifier model. On the contrary, 2D SEA and 3D SCA methods gave much higher specificity values of 98% and 97%, respectively, than the corresponding specificity of 88% reported for the Bayesian classifier model. By combining the 2D SEA and 3D SCA methods, the sensitivity increased in the same manner as we previously observed—to 69% while maintaining a high of 95% specificity. The combined 2D SEA and 3D SCA sensitivity was on par with the result of Bayesian classifier model, but with an improved specificity.

Test on Known Drugs. In one of our tests, we reviewed 1150 drug molecules that are currently on the market. Although the majority of drug molecules do not inhibit hERG, there are several drugs that are known potent hERG blockers, e.g., Verapamil and Quinidine, and some of them were included in our data set as hERG blockers. We therefore removed any duplicates from the drug set, leaving 1072 drugs. We used the drug data set as our query molecules to test for potential hERG blockers. From the 1072 drugs, 73 (6.8%) were identified as hERG blockers with the 2D SEA, 134 (12.5%) with the 3D SCA, and 142 (13.2%) for the combined 2D SEA and 3D SCA method were identified as potential hERG blockers. Figure 7 shows the similarity distribution for the drug molecules compared with the random distribution. The distribution is similar to the distribution for the nonblockers with a tail to the right indicating that there are some molecules that are similar to the hERG blockers. In fact, many of the 142 drugs predicted to be hERG blockers are known in the literature to interact with the hERG ion channel, but with limited toxicity. We performed a literature search for references regarding hERG inhibition to evaluate our method and to test if the drugs predicted to be hERG blockers were correctly identified. Table 5 shows the 20 drugs that are most likely to interact with hERG based on their 2D and 3D similarity scores, for which we found literature support for 18 of 20 (90%) of them being hERG blockers. Among the 20 drugs, there are 5 groups with similar structures with up to 4 members.

Reasons why hERG inhibition might be acceptable in a drug include the following: (1) the potency of the drug to the primary target is much higher and therefore will not severely interfere with hERG at a low level of concentration, (2) there is evidence that if a drug is interacting with multiple ion channels the toxic effect of hERG inhibition becomes less pronounced,[50] and (3) the benefit of the drug is greater than the potential risk of incurring long QT syndrome.

It is worthwhile noting that in some cases it is the metabolites of a drug that are responsible for in vivo hERG inhibition. Our prediction models were developed using in vitro cell-based assays that do not include appreciable amounts of metabolizing enzymes to allow for drug metabolism. For studies of pro-drugs, i.e., drugs that needs to be metabolized to become active, or drugs that are extensively metabolized in vivo, the actual structure of the metabolites are required as an input to our model to make any assessment of potential hERG

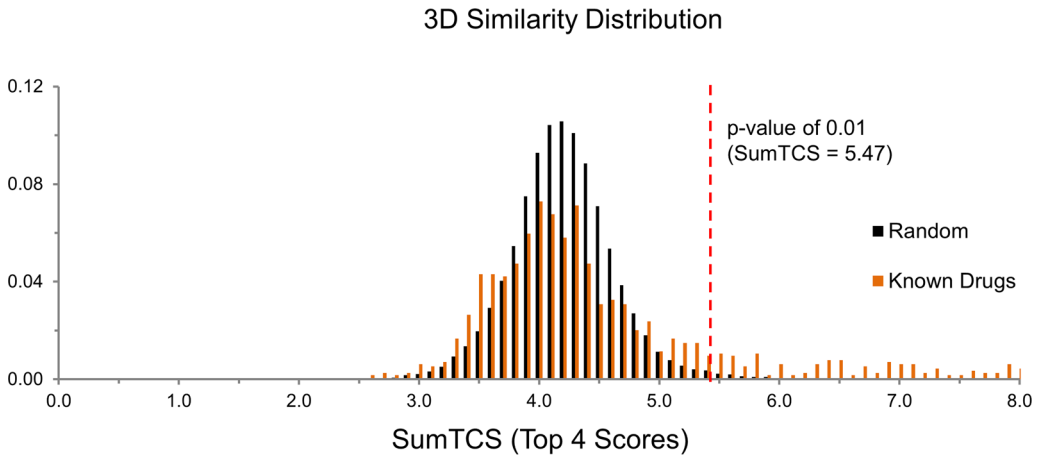
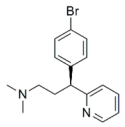
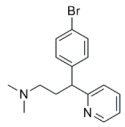
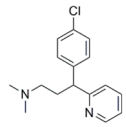
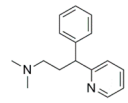
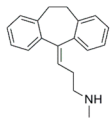
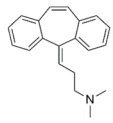
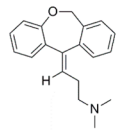
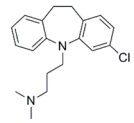
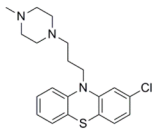
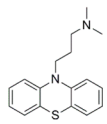
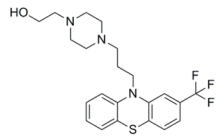
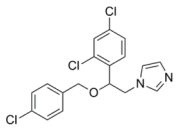
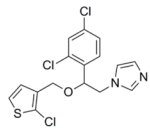
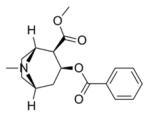
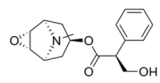
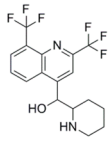
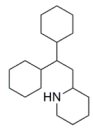
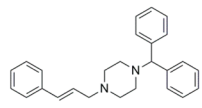
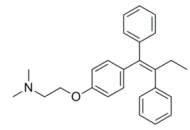
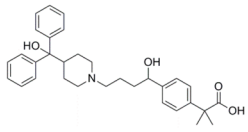


Figure 7. 3D similarity distribution of a set of random molecules compared with FDA-approved drug molecules.

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222

Table 5. Top 20 Drugs Predicted to Be hERG Blockers Based on 2D SEA and 3D SCA

Group				
1	 Dexbrompheniramine (B[a])[33]	 Brompheniramine (B)[33]	 Chlorphenamine (B)[34]	 Pheniramine (B)
2	 Nortriptyline (B)[35]	 Cyclobenzaprine (B)[36]	 Doxepin (B)[37]	 Clomipramine (B)[38]
3	 Prochlorperazine (B)[39]	 Promazine (B)[40]	 Fluphenazine (B)[41]	
4	 Econazole (B)[42]	 Tioconazole (B)		
5	 Cocaine (B)[43]	 Scopolamine (N[b])[44]		
Singles				
	 Mefloquine (B)[45]	 Perhexiline (B)[46]	 Cinnarizine (B)[47]	 Tamoxifen (B)[48]
	 Fexofenadine (N)[49]			

[a]B = blocker. [b]N = nonblocker.

liabilities. This is not a limitation of our method per se and if a user has inhibition data for any metabolites these can easily be added to the reference data set.

CONCLUSIONS

We have developed a 3D SCA method that classifies an inhibitor or blocker based on the sum of an optimal number of the highest 3D similarity scores between conformations of a set of known reference ligands and a set of query ligands. We implemented the 3D SCA method for the potassium ion channel, hERG, as a drug target to predict potential hERG inhibitors. The benefit of the 3D SCA method over 2D similarity approaches was that it could make reliable predictions for compounds that lack 2D fingerprint similarity but share shape and chemical similarity with the training set, i.e., it can make predictions for novel chemical scaffolds. Although our 3D SCA method correctly predicted more hERG blockers than 2D SEA, combining the two approaches enhanced overall performance and, in particular, improved the sensitivity of the model. The 3D SCA method is easily implemented, requiring only a training set of known ligands that affect the protein target. Furthermore, the 3D SCA itself requires little training when

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222

new data become available, which is crucial for maintaining a model with high coverage.[51,52]

Complex molecular targets such as channels, receptors, and other multicomponent macromolecular assemblies typically have multiple binding sites that can affect their function. Given that we do not have detailed structural information about these structures, chemical similarity remains the key principle for developing high-throughput methods to assess chemicals against these target classes. Because the 3D SCA method captures both structural shape and chemical similarity, it has the capabilities to detect similarities not encoded in 2D approaches. Limitations with using rapid enumeration of conformations instead of comprehensive sampling techniques imply that combining 2D and 3D methods provides a practical means to develop both highly sensitive and specific models for complex ADMET end points.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00616.

Data sets used in this study; SMILES strings for the hERG blockers, non-hERG blockers, external test sets, drug molecules from Drug Bank, and our diverse set of molecules from NCI (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*Phone: 301-619-1941. Fax: 301-619-1983. E-mail: pschyman@bhsai.org (P.S.).

*E-mail: sven.a.wallqvist.civ@mail.mil (A.W.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors were supported by the U.S. Army Medical Research and Materiel Command (Ft. Detrick, MD) and the Defense Threat Reduction Agency grant CBCall14-CBS-05-2-0007. High-performance computing resources were made available by the U.S. Department of Defense High Performance Computing Modernization Program. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

■ REFERENCES

(1) Wang, J.; Urban, L. The Impact of Early ADME Profiling on Drug Discovery and Development Strategy. *Drug Discovery World* **2004**, *5*, 73−86.

(2) Mullard, A. New Drugs Cost US $2.6 Billion to Develop. *Nat. Rev. Drug Discovery* **2014**, *13*, 877.

(3) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG Potassium Channels and Cardiac Arrhythmia. *Nature* **2006**, *440*, 463−469.

(4) De Ponti, F.; Poluzzi, E.; Montanaro, N. Organising Evidence on QT Prolongation and Occurrence of Torsades de Pointes with Non-Antiarrhythmic Drugs: A Call for Consensus. *Eur. J. Clin. Pharmacol.* **2001**, *57*, 185−209.

(5) Polak, S.; Wisniowska, B.; Brandys, J. Collation, Assessment and Analysis of Literature in vitro Data on hERG Receptor Blocking Potency for Subsequent Modeling of Drugs' Cardiotoxic Properties. *J. Appl. Toxicol.* **2009**, *29*, 183−206.

(6) Mitcheson, J. S.; Chen, J.; Lin, M.; Culberson, C.; Sanguinetti, M. C. A Structural Basis for Drug-Induced Long QT Syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 12329−12333.

(7) Boukharta, L.; Keränen, H.; Stary-Weinzinger, A.; Wallin, G.; de Groot, B. L.; Åqvist, J. Computer Simulations of Structure-Activity Relationships for HERG Channel Blockers. *Biochemistry* **2011**, *50*, 6146−6156.

(8) Di Martino, G. P.; Masetti, M.; Ceccarini, L.; Cavalli, A.; Recanatini, M. An Automated Docking Protocol for hERG Channel Blockers. *J. Chem. Inf. Model.* **2013**, *53*, 159−175.

(9) Farid, R.; Day, T.; Friesner, R. A.; Pearlstein, R. A. New Insights About HERG Blockade Obtained from Protein Modeling, Potential Energy Mapping, and Docking Studies. *Bioorg. Med. Chem.* **2006**, *14*, 3160−3173.

(10) Wang, S.; Li, Y.; Xu, L.; Li, D.; Hou, T. Recent Developments in Computational Prediction of HERG Blockage. *Curr. Top. Med. Chem.* **2013**, *13*, 1317−1326.

(11) Jamieson, C.; Moir, E. M.; Rankovic, Z.; Wishart, G. Medicinal Chemistry of hERG Optimizations: Highlights and Hang-Ups. *J. Med. Chem.* **2006**, *49*, 5029−5046.

(12) Villoutreix, B. O.; Taboureau, O. Computational Investigations of hERG Channel Blockers: New Insights and Current Predictive Models. *Adv. Drug Delivery Rev.* **2015**, *86*, 72−82.

(13) Braga, R. C.; Alves, V. M.; Silva, M. F. V.; Muratov, E.; Fourches, D.; Tropsha, A.; Andrade, C. H. Tuning hERG Out: Antitarget QSAR Models for Drug Development. *Curr. Top. Med. Chem.* **2014**, *14*, 1399−1415.

(14) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.

(15) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET Evaluation in Drug Discovery. 12. Development of Binary Classification Models for Prediction of hERG Potassium Channel Blockage. *Mol. Pharmaceutics* **2012**, *9*, 996−1010.

(16) Broccatelli, F.; Mannhold, R.; Moriconi, A.; Giuli, S.; Carosati, E. QSAR Modeling and Data Mining Link Torsades de Pointes Risk to the Interplay of Extent of Metabolism, Active Transport, and HERG Liability. *Mol. Pharmaceutics* **2012**, *9*, 2290−2301.

(17) Du-Cuny, L.; Chen, L.; Zhang, S. A Critical Assessment of Combined Ligand- and Structure-Based Approaches to HERG Channel Blocker Modeling. *J. Chem. Inf. Model.* **2011**, *51*, 2948−2960.

(18) 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.

(19) Jia, L.; Sun, H. Support Vector Machines Classification of hERG Liabilities Based on Atom Types. *Bioorg. Med. Chem.* **2008**, *16*, 6252−6260.

(20) Thai, K. M.; Ecker, G. F. Similarity-Based SIBAR Descriptors for Classification of Chemically Diverse hERG Blockers. *Mol. Diversity* **2009**, *13*, 321−336.

(21) Czodrowski, P. hERG Me Out. *J. Chem. Inf. Model.* **2013**, *53*, 2240−2251.

(22) Yera, E. R.; Cleves, A. E.; Jain, A. N. Chemical Structural Novelty: On-Targets and Off-Targets. *J. Med. Chem.* **2011**, *54*, 6771−6785.

(23) 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. Pharmaceutics* **2008**, *5*, 117−127.

(24) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(25) Liu, R.; Wallqvist, A. Merging Applicability Domains for in silico Assessment of Chemical Mutagenicity. *J. Chem. Inf. Model.* **2014**, *54*, 793−800.

(26) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; et al. Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: A Large-Scale Bioactivity

Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.
(27) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−1097.
(28) Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(29) *OMEGA 2.5.1.4*; OpenEye Scientific Software, Santa Fe, NM.
(30) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(31) *ROCS 3.2.0.4*; OpenEye Scientific Software, Santa Fe, NM.
(32) Dunn, G.; Everitt, B. *Clinical Biostatistics: An Introduction to Evidence-Based Medicine*; Edward Arnold: London, 1995.
(33) Park, S.-J.; Kim, K.-S.; Kim, E.-J. Blockade of HERG K^+ Channel by an Antihistamine Drug Brompheniramine Requires the Channel Binding within the S6 Residue Y652 and F656. *J. Appl. Toxicol.* **2008**, *28*, 104−111.
(34) Hong, H. K.; Jo, S. H. Block of HERG K^+ Channel by Classic Histamine H_1 Receptor Antagonist Chlorpheniramine. *Korean J. Physiol. Pharmacol.* **2009**, *13*, 215−220.
(35) Jeon, S.-H.; Jaekal, J.; Lee, S. H.; Choi, B.-H.; Kim, K.-S.; Jeong, H.-S.; Han, S. Y.; Kim, E.-J. Effects of Nortriptyline on QT Prolongation: A Safety Pharmacology Study. *Hum. Exp. Toxicol.* **2011**, *30*, 1649−1656.
(36) Jo, S. H.; Youm, J. B.; Lee, C. O.; Earm, Y. E.; Ho, W. K. Blockade of the HERG Human Cardiac K^+ Channel by the Antidepressant Drug Amitriptyline. *Br. J. Pharmacol.* **2000**, *129*, 1474−1480.
(37) Duncan, R. S.; McPate, M. J.; Ridley, J. M.; Gao, Z.; James, A. F.; Leishman, D. J.; Leaney, J. L.; Witchel, H. J.; Hancox, J. C. Inhibition of the HERG Potassium Channel by the Tricyclic Antidepressant Doxepin. *Biochem. Pharmacol.* **2007**, *74*, 425−437.
(38) Jo, S. H.; Hong, H. K.; Chong, S. H.; Won, K. H.; Jung, S. J.; Choe, H. Clomipramine Block of the hERG K^+ Channel: Accessibility to F656 and Y652. *Eur. J. Pharmacol.* **2008**, *592*, 19−25.
(39) Kim, M. D.; Eun, S. Y.; Jo, S. H. Blockade of HERG Human K^+ Channel and I_{Kr} of Guinea Pig Cardiomyocytes by Prochlorperazine. *Eur. J. Pharmacol.* **2006**, *544*, 82−90.
(40) Thomas, D.; Wu, K.; Kathofer, S.; Katus, H. A.; Schoels, W.; Kiehn, J.; Karle, C. A. The Antipsychotic Drug Chlorpromazine Inhibits HERG Potassium Channels. *Br. J. Pharmacol.* **2003**, *139*, 567−574.
(41) Hong, H. K.; Lee, B. H.; Park, M. H.; Lee, S. H.; Chu, D.; Kim, W. J.; Choe, H.; Choi, B. H.; Jo, S. H. Block of hERG K^+ Channel and Prolongation of Action Potential Duration by Fluphenazine at Submicromolar Concentration. *Eur. J. Pharmacol.* **2013**, *702*, 165−173.
(42) Xia, M.; Shahane, S. A.; Huang, R.; Titus, S. A.; Shum, E.; Zhao, Y.; Southall, N.; Zheng, W.; Witt, K. L.; Tice, R. R.; Austin, C. P. Identification of Quaternary Ammonium Compounds as Potent Inhibitors of hERG Potassium Channels. *Toxicol. Appl. Pharmacol.* **2011**, *252*, 250−258.
(43) Guo, J.; Gang, H.; Zhang, S. Molecular Determinants of Cocaine Block of Human Ether-a-Go-Go-Related Gene Potassium Channels. *J. Pharmacol. Exp. Ther.* **2006**, *317*, 865−874.
(44) Van de Water, A.; Van der Linde, H.; De Clerck, F. *Cardiovascular, Electrophysiologic and Pulmonary Effects of High Intravenous Doses of Cisapride in Anesthetized Dogs*; Janssen Research Foundation, 1999.
(45) Kang, J.; Chen, X. L.; Wang, L.; Rampe, D. Interactions of the Antimalarial Drug Mefloquine with the Human Cardiac Potassium Channels KvLQT1/minK and HERG. *J. Pharmacol. Exp. Ther.* **2001**, *299*, 290−296.
(46) Walker, B. D.; Valenzuela, S. M.; Singleton, C. B.; Tie, H.; Bursill, J. A.; Wyse, K. R.; Qiu, M. R.; Breit, S. N.; Campbell, T. J. Inhibition of HERG Channels Stably Expressed in a Mammalian Cell Line by the Antianginal Agent Perhexiline Maleate. *Br. J. Pharmacol.* **1999**, *127*, 243−251.
(47) NIH Chemical Gnomics Center. PubChem BioAssay Database; AID = 588834; https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=588834 (accessed Aug 21, 2015).
(48) Thomas, D.; Gut, B.; Karsai, S.; Wimmer, A. B.; Wu, K.; Wendt-Nordahl, G.; Zhang, W.; Kathofer, S.; Schoels, W.; Katus, H. A.; Kiehn, J.; Karle, C. A. Inhibition of Cloned HERG Potassium Channels by the Antiestrogen Tamoxifen. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **2003**, *368*, 41−48.
(49) Scherer, C. R.; Lerche, C.; Decher, N.; Dennis, A. T.; Maier, P.; Ficker, E.; Busch, A. E.; Wollnik, B.; Steinmeyer, K. The Antihistamine Fexofenadine does not Affect I_{Kr} Currents in a Case Report of Drug-Induced Cardiac Arrhythmia. *Br. J. Pharmacol.* **2002**, *137*, 892−900.
(50) Kramer, J.; Obejero-Paz, C. A.; Myatt, G.; Kuryshev, Y. A.; Bruening-Wright, A.; Verducci, J. S.; Brown, A. M. MICE Models: Superior to the HERG Model in Predicting Torsade de Pointes. *Sci. Rep.* **2013**, *3*, 2100.
(51) Liu, R.; Schyman, P.; Wallqvist, A. Critically Assessing the Predictive Power of QSAR Models for Human Liver Microsomal Stability. *J. Chem. Inf. Model.* **2015**, *55*, 1566−1575.
(52) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.

DOI: 10.1021/acs.jcim.5b00616
J. Chem. Inf. Model. 2016, 56, 213−222