JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 1224−1233

pubs.acs.org/jcim

The Catch-22 of Predicting hERG Blockade Using Publicly Accessible Bioactivity Data

Vishal B. Siramshetty,[†,‡] Qiaofeng Chen,[†,§] Prashanth Devarakonda,[†] and Robert Preissner*[,†,‡]

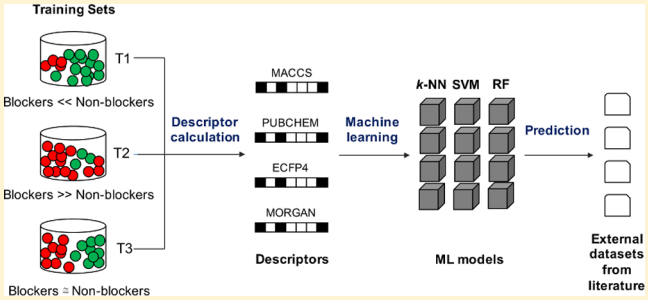
[†]Structural Bioinformatics Group, Charité - University Medicine Berlin, 10115 Berlin, Germany

[‡]BB3R − Berlin Brandenburg 3R Graduate School, Freie Universität Berlin, 14195 Berlin, Germany

[§]China Scholarship Council (CSC), Beijing 100044, China

Supporting Information

ABSTRACT: Drug-induced inhibition of the human ether-à-go-go-related gene (hERG)-encoded potassium ion channels can lead to fatal cardiotoxicity. Several marketed drugs and promising drug candidates were recalled because of this concern. Diverse modeling methods ranging from molecular similarity assessment to quantitative structure−activity relationship analysis employing machine learning techniques have been applied to data sets of varying size and composition (number of blockers and nonblockers). In this study, we highlight the challenges involved in the development of a robust classifier for predicting the hERG end point using bioactivity data extracted from the public domain. To this end, three different modeling methods, nearest neighbors, random forests, and support vector machines, were employed to develop predictive models using different molecular descriptors, activity thresholds, and training set compositions. Our models demonstrated superior performance in external validations in comparison with those reported in the previous studies from which the data sets were extracted. The choice of descriptors had little influence on the model performance, with minor exceptions. The criteria used to filter bioactivity data, the activity threshold settings used to separate blockers from nonblockers, and the structural diversity of blockers in training data set were found to be the crucial indicators of model performance. Training sets based on a binary threshold of 1 μM/10 μM to separate blockers ($IC_{50}/K_i \leq 1$ μM) from nonblockers ($IC_{50}/K_i > 10$ μM) provided superior performance in comparison with those defined using a single threshold (1 μM or 10 μM). A major limitation in using the public domain hERG activity data is the abundance of blockers in comparison with nonblockers at usual activity thresholds, since not many studies report the latter.



INTRODUCTION

In recent years, adverse events leading to toxicities accounted for a majority of postmarketing drug withdrawals.[1] Cardiotoxicity, along with hepatotoxicity and immune reactions, was reported as a predominant concern in the recall of several drugs.[2] Inhibition of the voltage-gated potassium ion channel, whose α subunit is encoded by the human ether-à-go-go-related gene (hERG), prolongs the QT interval of the cardiac action potential and subsequently results in fatal arrhythmias.[3] A number of cardiovascular and non-cardiovascular drugs were reportedly withdrawn because of the incidence of sudden cardiac death, while the use of several other drugs has been restricted because of their potential to cause severe cardiac arrhythmias, marking hERG as a major antitarget during drug development.[4,5] The U.S. Food and Drug Administration (FDA) now requires every new chemical entity to be screened against hERG.[6−8] This garnered serious attention from both the pharmaceutical industry and academia to develop predictive models that can detect hERG liability. Although multiple in silico models have been proposed to date, the conventional patch-clamp electrophysiological assay is the most widely accepted method to confirm hERG liabilities.[9−11] Although the overall throughput of semiautomated versions of the patch-clamp assay has improved, the huge costs involved and scope of experimental errors are considered as major drawbacks.[12−14] Furthermore, significant levels of variability were reported with high-throughput ion-channel electrophysiology screens.[15] Thus, in silico models that can reliably predict hERG blockade would still be attractive for medicinal chemists during optimization of lead molecules.

A recent study[16] provided a detailed overview of the quantitative structure−activity relationship (QSAR) studies that employed a wide range of methods, including recursive partitioning, partial least-squares, naïve Bayes, support vector machine (SVM), random forest (RF), neural networks, and nearest neighbors (k-NN). The activity thresholds used to distinguish blockers and nonblockers ranged from 1 to 40 μM,[17−20] suggesting a high variation in the training set compositions. While some considered data extracted from single or multiple assay/data sources,[21−25] most studies used in-house data or proprietary data from the pharma industry that

Received: March 16, 2018
Published: May 17, 2018

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233

are not publicly accessible.[19,26−31] A limited number of studies[16,32,33] reported classification models based on hERG data extracted from publicly accessible bioactivity databases such as ChEMBL and PubChem. The heterogeneous activity data obtained from such databases were shown to possess a considerable level of experimental uncertainty, and recommendations were made regarding how such data must be curated before model development.[34−36] Additional limitations such as small numbers of compounds used in modeling (often a few hundred), narrow or unreported applicability domains, and lack of proof of validation (e.g., Y-randomization tests) restrict the use of most previously published models. Although more recent studies cautiously curated activity data and came up with better-performing models in comparison with older studies, the amount of data considered by these studies is very small in comparison with the current wealth of hERG bioactivity data in the public domain. Especially, a few studies, although they evaluated multiple thresholds, did not evaluate the effects of training set composition (ratio of actives to inactives) or activity threshold settings on the model performance. Bajorath and co-workers[37] recently reported the influence of varying training set size and composition on the performance of a machine-learning method in predicting active compounds by employing data sets belonging to 10 different activity classes (or targets). While the performance increased with increasing number of negative-class compounds, it was cautioned that this was possible only if a required threshold of positive-class compounds was achieved. It is unclear whether this would be applicable to the currently available hERG data sets, suggesting the need to evaluate the influence of varying training set sizes and compositions. Moreover, the two commonly used thresholds (1 μM and 10 μM) were reported to result in training sets with completely contrasting compositions of hERG data from ChEMBL,[32] which highlights the lack of a large number of negative-class examples when considering a higher activity threshold.

In this article, we highlight the challenges involved in the development of a robust in silico model for predicting hERG blockade using publicly accessible bioactivity data. Multiple activity threshold settings were employed in order to evaluate their influence on the model performance. The modeling approaches *k*-NN, SVM, and RF were employed with four different molecular fingerprints as features. Model validation was performed using multiple data sets extracted from the literature, both independently and in combination with ChEMBL data, to demonstrate the importance of appropriate training set size and composition in the development of robust models.

MATERIALS AND METHODS

Training Data Sets. Bioactivity data for hERG were extracted from ChEMBL (version 23).[38] To ensure that we employed only data of high confidence, the data were preprocessed using filter criteria for compound selection originally proposed by Bajorath et al.[39] The filtering steps involved and corresponding changes to the data are presented in Table S1 in the Supporting Information. Chemical structures were standardized using the JChem Suite (http://www.chemaxon.com) in order to remove duplicate compound entries. Although a 10- to 30-fold difference between the IC_{50} and peak plasma concentration is considered as the industry guidance,[40] it is impractical to obtain such data for large-scale modeling, and no widely accepted activity threshold is recommended in the literature to distinguish hERG blockers from the nonblockers. Therefore, we used multiple threshold settings (see Table 1) to obtain three training data sets (T1, T2, and T3). Each training set was further split (in a stratified fashion using class labels) into an internal training set (90%) and an internal test set (10%) for 10-fold cross-validation (CV).

Table 1. Summary of Training Sets and Activity (Ac) Threshold Settings Used to Separate Blockers from Nonblockers

data set	threshold settings	no. of compounds	no. of blockers	no. of nonblockers
T1	blockers: Ac $\leq$ 1 μM nonblockers: Ac > 1 μM	5804	1406	4398
T2	blockers: Ac $\leq$ 10 μM nonblockers: Ac > 10 μM	5804	4096	1708
T3	blockers: Ac $\leq$ 1 μM nonblockers: Ac $\geq$ 10 μM	3223	1406	1817

External Data Sets. We collected data for external validation from four independent studies that previously reported predictive hERG models based on different classification approaches. We separately extracted the training and test data sets (see Table 2), as they were used in each of these studies, to be able to compare the performance of our models with those reported. Although we considered a larger pool of studies from which to extract the data sets, not many studies provided the data sets, and multiple studies used the very same data sets. Furthermore, the data sets we chose represent both patch-clamp (% inhibition) and radioligand binding assay (IC_{50}) types.

Table 2. Overview of the External Data Sets Employed for Evaluation of Our Models

data set	no. of compounds	no. of blockers	no. of nonblockers	assay type; source
E1 (training)[a]	3024	483	2541	thallium flux (training) and patch-clamp (test) assays; Sun et al.[41]
E1 (test)[a]	66	53	13	
E2 (training)	2389	1004	1385	multiple assay types; Doddareddy et al.[42]
E2 (test)	255	108	147	
E3 (training)	476	110	366	multiple assay types; Li et al.[25]
E3 (test)	66	44	22	
E4 (training)	368	199	169	multiple assay types; Marchese Robinson et al.[20]
E4 (test)	313	197	116	

[a]The activity type of the assays was % inhibition, and the thresholds used were described in the original study.[41] For all of the other data sets, the activity type was IC_{50}, and the threshold was 10 μM.

Molecular Descriptors. We employed four types of molecular fingerprints: Molecular Accession System (MACCS) keys; PubChem fingerprints; extended connectivity fingerprints (ECFP); and Morgan fingerprints. The first two fingerprints belong to the category of substructure-based molecular fingerprints, while the other two are circular fingerprints based on a layered atom environment representation. MACCS keys and PubChem fingerprints vary in terms of length (166 and 881 bits, respectively), while ECFP and Morgan fingerprints are of the same length (1024 bits) but vary in the amount of information encoded, as we chose a diameter of 4 for ECFP (i.e., ECFP4), which is the radius in the case of

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233

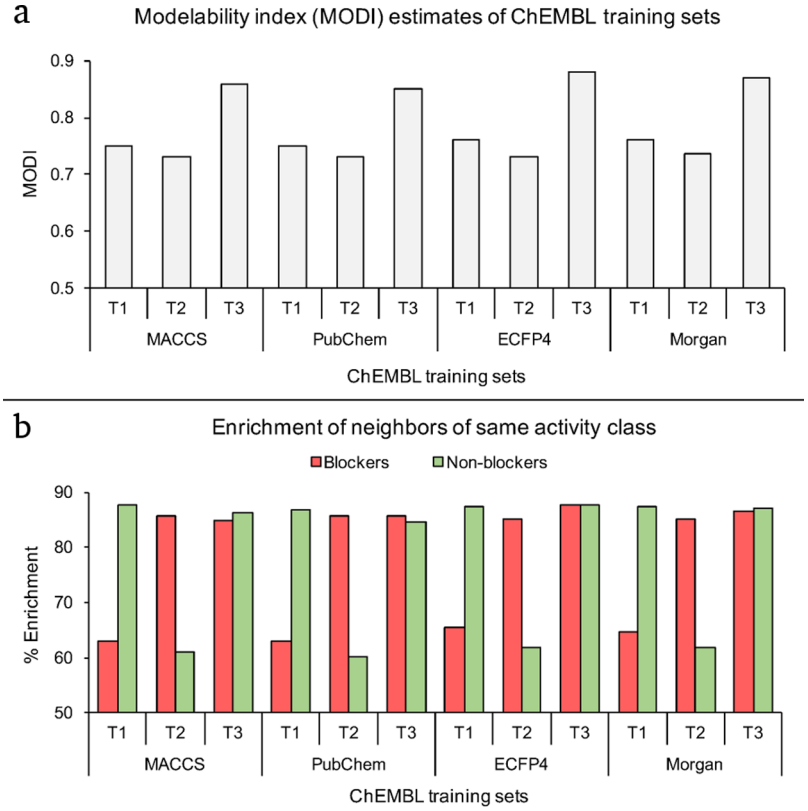


Figure 1. (a) Comparison of MODI estimates for training sets generated from ChEMBL using different molecular fingerprints. (b) Enrichment of nearest neighbors of the same activity class for blockers and nonblockers in the ChEMBL training sets T1, T2, and T3 using different molecular fingerprints.

Morgan fingerprints. Heat map representations of the chemical similarity of the three training data sets were generated to study the distribution of the similarity of the structures. Furthermore, the individual bits were used as features for a principal component analysis (PCA) to visually inspect the chemical space of hERG blockers and nonblockers. All of the fingerprints were calculated using the Chemistry Development Kit (CDK) nodes in KNIME (version 3.2.1).

Model Development and Validation. In a recent study, deep learning was evaluated in comparison with traditional machine learning approaches to model pharmaceutically relevant end points, including hERG, using diverse drug discovery data sets.[43] While the method offered improvement across different metrics used for performance evaluation, little impact was noticed for the hERG end point, especially over the test set in comparison to other classifiers studied.[43] Therefore, we stick to the classical approaches *k*-NN, SVM, and RF for modeling in this study. *k*-NN is a simple, nonparametric classification and regression approach that considers *k* closest training instances in the feature space. While the classification approach outputs the class membership, the regression approach outputs the property value, which in this case is the probability of a molecule to be active against hERG. In this study, we developed our own NN model employing multiple *k* values (k = 1, 3, 5, and 10), which outputs a class label and a probability for each test set compound to be a hERG blocker. SVM aims to maximize the margin hyperplane that differentiates the two classes. It was used here because of its effectiveness in high-dimensional spaces and its excellent generalization capabilities.[41] In this study, we employed C-SVC with a linear kernel function and default SVM parameters (the penalty parameter was C = 1.0). RF is an ensemble classifier that builds many decision tree classifiers and outputs the most predicted class, while using averaging to improve accuracy and limit overfitting. It was used here because it does not need any hyperparameter optimization.[44] The Gini index was employed as the splitting criterion, and a static random seed was chosen. The number of estimators (or the number of trees in the forest) used was 100, and no bagging was employed. All of the models were developed in the Python programming language using the machine learning library Scikit-learn[45] and the open-source cheminformatics tool kit RDKit (http://www.rdkit.org). The availability of data sets and models is detailed in section S10 in the Supporting Information. The test sets from the literature were used to test the models that were built using our three data sets, the corresponding training set from the literature, and combined sets obtained from the first two data sets. The performance of each classification model was assessed on the basis of the sensitivity, specificity, AUC, and balanced accuracy. In this context, the sensitivity (or true positive rate) is the proportion of blockers correctly predicted as blockers (eq 1):

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

where TP is the number of true positives and FN is the number of false negatives. The specificity (or true negative rate) is the proportion of nonblockers correctly predicted as nonblockers (eq 2):

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (2)$$

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233

where TN is the number of true negatives and FP is the number of false positives. The AUC is the area under the receiver operating characteristic curve, which is obtained by plotting the true positive rate against the false positive rate at different threshold settings. The balanced accuracy (BACC) is the average of the proportions correctly predicted for each class individually.[46] It is calculated as the average of the sensitivity and specificity values (eq 3):

$$\mathrm{BACC} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{NP}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right) \quad (3)$$

RESULTS AND DISCUSSION

Before evaluation of the predictive power of the models, it is important to understand the underlying data and their potential to be used in the generation of prediction models. Therefore, we first provide an overview of the data sets and the feasibility of using them, followed by the cross-validation and external validation results and the related discussion based on these considerations.

Overview of the Data Sets. For each training data set, the feasibility of obtaining a prediction model was estimated by calculating the modelability index (MODI), as proposed by Tropsha and co-workers,[36] who define it as the "activity class-weighted ratio of the number of nearest-neighbor pairs of compounds with the same activity class versus the total number of pairs". MODI provides a prior estimate of the feasibility of obtaining externally predictive models using a data set of bioactive compounds, which emerged in the context of evaluating the influence of activity cliffs on the overall performance of QSAR models.[47] In this context, we evaluated how varying the training set size and composition would influence the data set modelability. The third training set (T3), which was generated using a binary activity threshold, was associated with significantly higher MODI values (Figure 1a) in comparison with the other two training sets. In particular, the trend remained same with four different molecular fingerprints used to calculate the MODI values. This is in agreement with the findings of Tropsha and co-workers that the choice of descriptors had a weak influence on MODI values when evaluated on multiple data sets. However, a general limitation could be that instances of the closest nearest neighbors with the same Euclidean distance value belonging to different activity classes could influence the final MODI value in a data set containing highly similar molecules. However, closely investigating the enrichment of nearest neighbors belonging to same activity class with different training sets and different fingerprints confirmed that training set T3 showed better enrichment for both blockers and nonblockers (Figure 1b). Taken together, these observations indicate that employing training set T3 would result in better-performing models. A discussion of how the data set modelability correlates with the performances of different models will be presented in the next sections.

The activity effects were analyzed for the three different training sets in order to understand the impact of activity cliffs on model performances. Activity cliffs (using MMPs) and matched molecular pairs (MMPs)[48] were generated from the three training sets using activity annotations (1 or 0) as the criteria for cliff and MMP formation rather than defined activity values.[49] To ensure high data quality, subsets of the training sets belonging to two different activity types (IC_{50} and K_i) were treated separately. Activity cliffs and MMPs were generated in Discovery Studio using the settings described in section S3 in the Supporting Information.

In line with the previous observations, the training set T3 resulted in a relatively low number of activity cliffs and MMPs (Table 3). At a higher activity threshold (T2), the number of cliffs significantly increased compared with the number formed at a lower activity threshold (T1) in the case of IC_{50} data. On the other hand, in the case of K_i data, the number of cliffs formed at the higher threshold (T2) was less than that formed at the lower threshold (T1), suggesting the importance of data confidence in QSAR studies. Furthermore, closely examining the chemical space representations (similarity heat maps and PCA plots) of the training sets revealed a greater diversity among the compounds in the training set T3 compared with the other two data sets (section S4 in the Supporting Information). Though these observations indicate a higher potential for the training set designed using a binary activity threshold, this would be confirmed only after comparing the cross-validation results.

Table 3. Activity Cliffs and MMPs Obtained with Different Training Sets and Activity Types

	IC_{50} data			K_i data		
data set	no. of compounds	no. of cliffs	no. of MMPs	no. of compounds	no. of cliffs	no. of MMPs
T1	4780	882	62	1024	255	11
T2	4780	1203	62	1024	111	11
T3	2629	171	30	594	32	2

Cross-Validation Results. In all of the cross-validation runs, employing four fingerprint types (MACCS, PubChem, ECFP4, and Morgan) and three classification methods, training set T3 provided superior performance in terms of AUC values compared with training sets T1 and T2 (Figure 2). While the AUC values from the best-performing models (on training set T3) varied slightly with the SVM classifier for different fingerprints, the NN and RF classifiers tended to provide the same or comparable performance with different fingerprints. In particular, the RF classifier appeared to be unbiased, if not completely, to the fingerprint of choice. For instance, SVM provided an AUC of 0.56 and a BACC of 0.53 with MACCS fingerprints on training set T1, which are much lower than the values achieved with other fingerprints. On the other hand, RF and NN provided comparable AUC values for all of the fingerprints, including MACCS, with RF being the most consistent performer. When compared in terms of BACC, the training sets based on a single activity threshold (T1 and T2) performed inferior to the training set based on a binary activity threshold (T3). This could be the case because the first two data sets (T1 and T2) are highly imbalanced (i.e., they have higher numbers of training examples belonging to one class in comparison with those belonging to the other class). In the case of NN, the best results were obtained with $k = 1$ among all of the k values considered ($k = 1$, 3, 5 and 10). While PubChem and ECFP4 performed the best with the SVM and RF classifiers, PubChem and Morgan worked well for k-NN. Detailed cross-validation results, including results for all fingerprint types and standard deviations for AUC values, are provided in Table S5.

In data set T1, as a result of the lower activity threshold (1 μM), the number of blockers is significantly lower than the number of nonblockers, which could explain the lower

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233

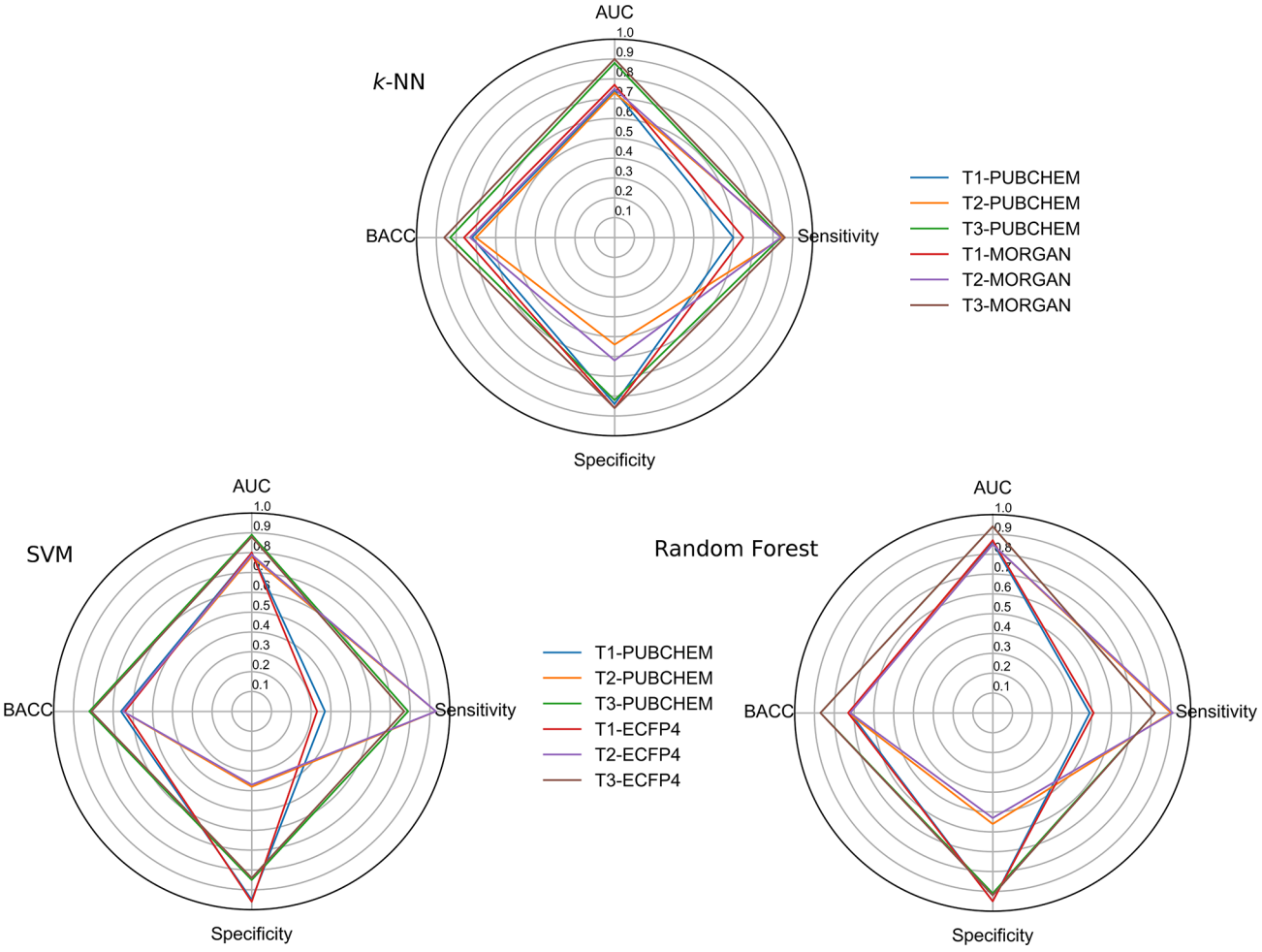


Figure 2. 10-fold cross-validation performance (AUC) on the three training sets T1, T2, and T3 employing the two best-performing molecular fingerprints and three classifiers: (top) *k*-NN; (bottom left) SVM; (bottom right) RF.

sensitivity (true positive rate) values achieved with this data set. In contrast, data set T2 provided a lower specificity (true negative rate) because of the lack of a sufficient number of nonblockers at the higher activity threshold (10 μM). However, training set T3, which is more balanced in terms of the number of blockers and nonblockers, provided higher sensitivity and specificity values and hence provided higher balanced accuracies. However, it was found that NN performed better than SVM and RF in terms of sensitivity and specificity values with training sets T1 and T2. The SVM and NN classifiers provided varying BACC values for different fingerprints, while RF provided comparable or nearly the same values. These findings confirm that training set T3 provides a better trade-off between sensitivity and specificity, which is evident from the AUC and BACC values. SVM failed to achieve better results especially when using fingerprints that describe the negative-class and positive-class examples in a highly similar chemical space (e.g., MACCS), which is clear from the similarity heat map and 2D scatter plot representations (section S4 in the Supporting Information). On the basis of these findings, only training set T3 was used for external validation in combination with PubChem and ECFP4 fingerprints in the case of the RF and SVM classifiers and with PubChem and Morgan fingerprints in the case of NN, which provided better performance in cross-validation.

Data Quality as an Indicator of Performance. In order to evaluate the influence and importance of data quality, particularly in the context of availability of a huge bioactivity data set containing more than 18 000 bioactivities from ChEMBL, which aggregates them from multiple sources, an additional training data set containing 5176 molecules was used for cross-validation employing all three classifiers with the two best-performing fingerprints from their corresponding cross-validation outcomes. This data set was generated by combining training set T3 with additional compounds from ChEMBL's data set that were annotated as "active" and "inactive" in the column "ACTIVITY COMMENT". Since these compounds were directly chosen on the basis of annotations, the resulting data set is denoted as a "low-confidence" data set. A 10-fold CV indicated that the performance decreased with this data set. Both the AUC and BACC values (Table S6) indicate superior performance with training set T3. Taken together, these results show that the quality of the bioactivity data set in hand influences the model performance.

Influence of Structural Diversity of hERG Blockers. To further identify other factors that could significantly influence model performance, we evaluated the effect of diversity of the hERG blockers in the training data set. In order to achieve this, three data sets were generated that contained increasing numbers of diverse blockers while the number of nonblockers was kept the same as in training set T3. The three data sets

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233

$T3_{d1}$, $T3_{d2}$, and $T3_{d3}$ contained 300, 400, and 500 structurally diverse blockers, respectively, and 1817 nonblockers each. All three data sets were generated using the RDKit Diversity Picker node in KNIME, which picks diverse compounds on the basis of the Tanimoto distance between fingerprints (here we employed ECFP4 fingerprints) using a MaxMin algorithm.[50] A 10-fold cross-validation was performed using these data sets, employing all three classifiers with the two best-performing fingerprints from the initial cross-validation. The results confirmed an increase in performance (true positive rate or sensitivity) with increasing number of structurally diverse blockers in the training set (Table 4). While RF remained the best performer, all of the classifiers achieved AUC values that were either the same as (RF) or comparable to (NN and SVM) those obtained using the original training set T3, which contains more than 1400 blockers compared with 500 blockers in data set $T3_{d3}$. However, the BACC values were slightly lower than those obtained with T3, with NN being an exception, which demonstrated robust performance in terms of BACC too. These findings suggest the use of a structurally diverse set of blockers in the training set to model hERG channel blockade, although it is hard to estimate how many such diverse compounds are required to achieve a consistent performance. At the same time, the influence of the minimum number of negative-class examples required must not be ignored.[37]

Table 4. 10-Fold Cross-Validation Performance (AUC, Sensitivity, Specificity, and BACC) on the Three Training Sets Containing Increasing Numbers of Structurally Diverse hERG Blockers

classifier	fingerprint	training set	AUC	Sensitivity	Specificity	BACC
NN	PubChem	$T3_{d1}$	0.78	0.19	0.93	0.56
		$T3_{d2}$	0.82	0.54	0.91	0.73
		$T3_{d3}$	**0.87**	**0.76**	**0.91**	**0.84**
	Morgan	$T3_{d1}$	0.79	0.27	0.93	0.60
		$T3_{d2}$	0.86	0.68	0.93	0.81
		$T3_{d3}$	**0.89**	**0.81**	**0.93**	**0.87**
SVM	PubChem	$T3_{d1}$	0.72	0.03	0.99	0.51
		$T3_{d2}$	0.85	0.36	0.96	0.66
		$T3_{d3}$	**0.90**	**0.61**	**0.95**	**0.78**
	ECFP4	$T3_{d1}$	0.65	0.01	1.00	0.50
		$T3_{d2}$	0.84	0.36	0.97	0.66
		$T3_{d3}$	**0.90**	**0.61**	**0.95**	**0.78**
RF	PubChem	$T3_{d1}$	0.80	0.07	0.99	0.53
		$T3_{d2}$	0.90	0.37	0.98	0.68
		$T3_{d3}$	**0.94**	**0.65**	**0.98**	**0.81**
	ECFP4	$T3_{d1}$	0.79	0.01	1.00	0.50
		$T3_{d2}$	0.91	0.35	0.99	0.67
		$T3_{d3}$	**0.94**	**0.65**	**0.98**	**0.82**

External Validation Results. External test data sets from four different studies were used to evaluate the model performance. In addition to using ChEMBL training set T3, we used the actual training data sets employed in the corresponding studies in order to be able to compare the performance of our models with the reported ones. Apart from these two sets, we generated combination data sets (by combining each external training set with ChEMBL training set T3) to evaluate the potential of public domain bioactivity data to improve the model performance on independent external test data sets. All of the overlapping compounds in the external training sets were removed to obtain unique combination sets. Altogether, for each study, we have three different training sets and one test set. In the case of two test sets (Sun et al. and Li et al.) out of four, the training set from ChEMBL (T3) clearly outperformed the external training sets used in the original studies (Figure 3a,c). In the other two cases (test sets from Doddareddy et al. and Robinson et al.), the external training sets performed slightly better (Figure 3b,d). However, in almost all cases (data sets and fingerprint types), the combination set performed better than the individual training sets. The test data set from Robinson et al. was predicted well with the training set from the original study. The huge overlap (41%; 151 of 368 compounds) of the external training set from Robinson et al. with ChEMBL training set T3 could be the reason behind the modest influence of the T3 set. Interestingly, the external training set from Li et al. showed poor performance with all three classifiers. In fact, this is the only case where SVM performed much better than the other two classifiers. The performance of the RF model highly varied when used with two fingerprint types, PubChem and ECFP4, while the difference was either small or negligible in cross-validation and other external validations. In the case of Li et al., the PubChem fingerprint always provided inferior performance compared with ECFP4, possibly because of the high diversity of the training set compounds. In comparison with the other external training sets, this data set showed the least number of overlapping compounds with ChEMBL training set T3. Overall, the RF classifier provided the best performance. Detailed external validation results are provided in Table S7. Furthermore, it was found that external validation performance using either ChEMBL training set T3 alone or in combination with literature-extracted training sets was always better than the performance of each of these training sets in their original studies (details are provided in Table S8).

Catch-22 of Using Public Domain Bioactivity Data. With more than 18 000 activity records (for more than 14 000 unique compounds), ChEMBL provides the largest publicly available bioactivity data set to model hERG blockade. It must be noted that ChEMBL collects bioactivity data from multiple sources, including published studies. These studies measure different assay end points for hERG under varying assay conditions. Using such heterogeneous bioactivity data for predictive modeling was already shown to introduce significant amounts of uncertainty.[35] Therefore, preprocessing was essential to ensure that only high-confidence data were used in this study. However, the number of compounds in the final data set was significantly low. Furthermore, the activity threshold that separates blockers from nonblockers heavily influences the number of compounds in the positive and negative classes. While a strict threshold ($\leq$1 μM) limits the number of blockers, higher thresholds ($\geq$10 μM or above) result in low numbers of nonblockers (negative-class examples) in the training set that insufficiently span the chemical space of corresponding compounds in the test data sets. A threshold that omits compounds in the dubious activity value range (10 μM > activity value > 1 μM) provided a more balanced data set. This became clear from the cross-validation performance of training set T3 and the MODI values obtained alone and in combination with external test data sets. The observation remained the same with all three classifiers (k-NN, SVM, and RF). However, the performance based on training set T3 with different fingerprint types varied only with the SVM classifier. The fact that there was no difference in performance with the

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233

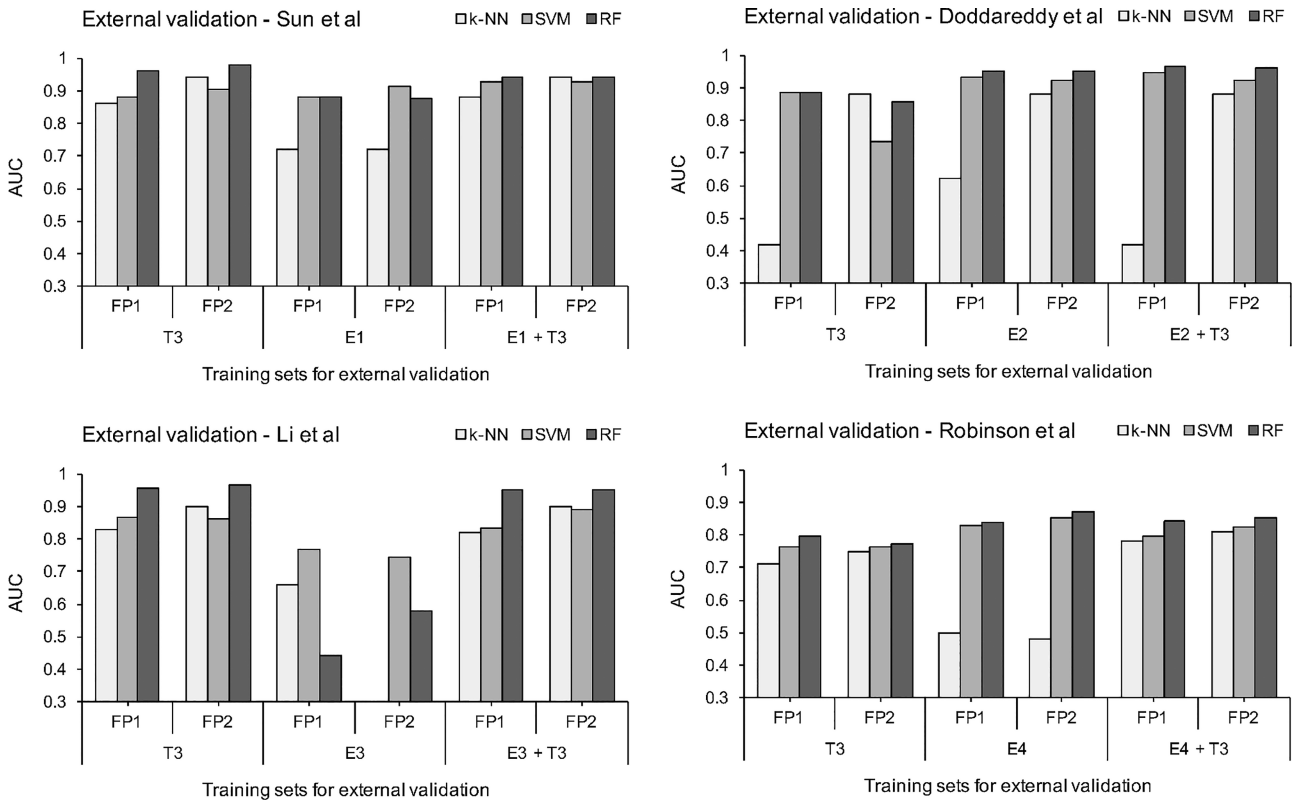


Figure 3. External validation performance (AUC) of the NN, SVM, and RF classifiers on test sets extracted from four different studies from the literature (Sun et al.,[41] Doddareddy et al.,[42] Li et al.,[25] and Robinson et al.[20]). Two best-performing fingerprints (FP1 and FP2) from the cross-validation were used for each classifier. FP1 is PubChem for all the classifiers, while FP2 is MORGAN for *k*-NN and ECFP4 for SVM and RF.

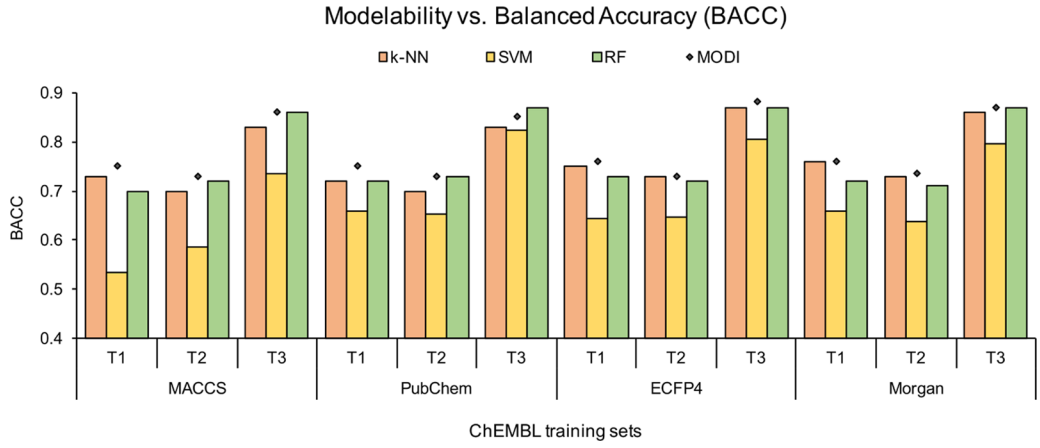


Figure 4. Cross-validation performance (BACC) on ChEMBL training sets T1, T2, and T3 in comparison with MODI estimates for the data sets.

RF classifier could be due to the classifier's ability to naturally handle correlation between features or descriptors without the need for a separate descriptor selection procedure. However, the similarity heat maps based on two different fingerprints (MACCS and ECFP4) look completely different. SVM classifiers based on the Tanimoto kernel may show a significant difference in performance when these two fingerprints are used as features. Overall, the performances on external test sets generally improved using T3 either alone or in combination with an external training set. These findings indicate that in general there would be an improvement in the performance of a classifier with a well-balanced hERG data set extracted from public bioactivity resources like ChEMBL. Although the extent of overlap of chemical space with the external training set also plays a key role, as a general rule it seems to be that the bigger and more diverse the training set, the better is the performance of the model, which is in agreement with previous reports.[51]

Modelability versus Applicability Domain. Another important factor considered in this study is the modelability of data sets for hERG blockade modeling. The quality of the data set in hand defines its modelability. Although certain other filter criteria were applied to arrive at a high-confidence data set, factors such as the size of the data set and diversity of the chemical structures also influence the quality of data set. The concept of the modelability index was introduced to estimate the suitability of data sets for QSAR modeling. Although it was validated on a large number of data sets, it is not clear whether a MODI estimate can be an absolute indicator of model

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233

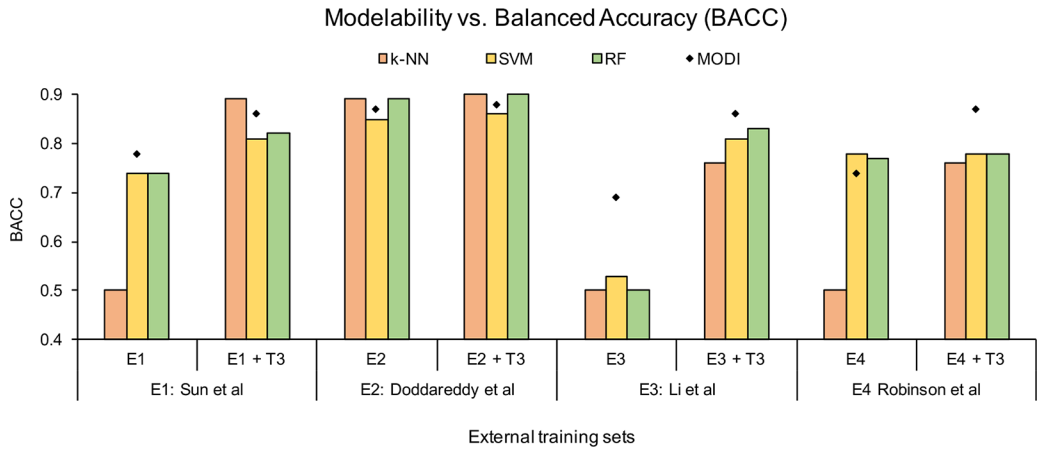


Figure 5. Comparison of MODI estimates with the external validation performance (BACC) on external training sets alone and in combination with T3.

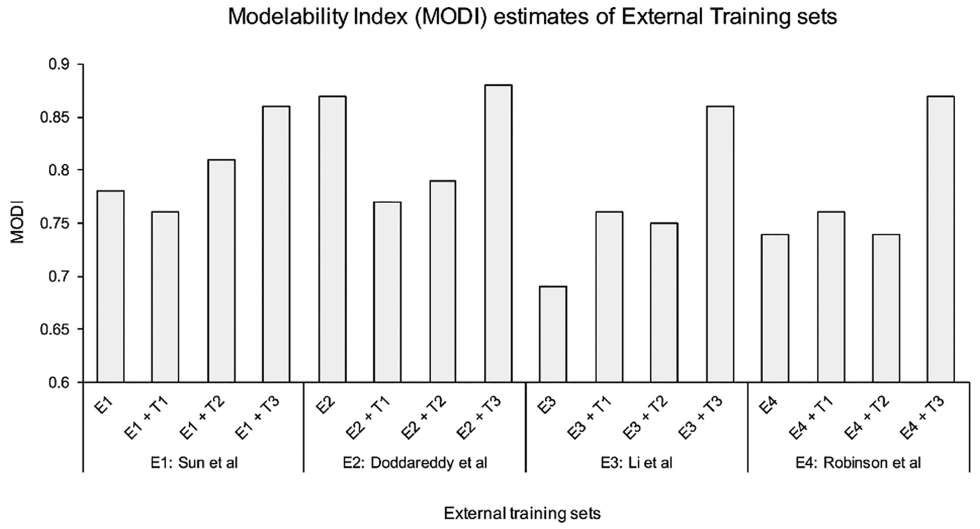


Figure 6. Comparison of MODI estimates of all external training sets alone and in combination with ChEMBL training sets T1, T2, and T3.

performance. In this study, training set T3 was associated with both the highest MODI estimates and performance values in comparison with the other two training sets, T1 and T2. This trend (MODI vs BACC) remained the same with all fingerprint types and classifiers in cross-validation (Figure 4). Similarly, in external validation, with increasing MODI value the performance (BACC) improved in all cases except for the training set of Robinson et al. (Figure 5). The MODI values improved in all cases when the external training sets were combined with ChEMBL training set T3 (Figure 6). This is due to the enrichment of the chemical space of blockers and nonblockers within the external training sets. However, the performance on external test sets did not show the same trend with two different fingerprints and classifiers because this depends on the nature of the chemical structures in the external test sets. This cannot be estimated from the MODI value, which is only a training set characteristic. The applicability domain (APD), estimated on the basis of Euclidean distances, provides an estimate of the coverage of chemical space of test set molecules within the training set.[52] In other words, the domain of model applicability must be defined in order to flag the compounds in the test set that could have unreliable predictions. These estimates were calculated using the APD node in KNIME.[53,54] Analogous to the MODI index, APD calculation estimates the distance of a test set compound to its nearest neighbor in the training set compared with a predefined APD threshold. Except for the training set from Sun et al., all of the external training sets showed 100% reliable predictions when combined with ChEMBL training set T3 (Table S9). The greatest improvement was observed with the training set from Li et al., which originally provided about 82% reliable predictions when used alone. The APD values are in agreement with the MODI estimates, confirming that either metric can be used in estimating the suitability or applicability of a data set to model an end point.

CONCLUSION

In this study, we emphasized the challenges involved in modeling of the pharmaceutically relevant hERG blockade end point. By far the biggest data set of hERG bioactivities, extracted from the ChEMBL database, was used at three different activity thresholds, which resulted in training sets of distinct size and composition. Three different machine learning classifier approaches, *k*-NN, SVM, and RF were used for model building by employing substructure-based and circular molecular fingerprints. We found that the quality of the data set, the activity thresholds used to design the training sets, and the structural diversity of the hERG blockers play key roles in

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233

achieving a robust model. In particular, application of certain activity data filtering criteria is mandatory when dealing with data from multiple assay types and bioactivity resources. Second, application of a single threshold (1 μM or 10 μM) resulted in highly imbalanced training sets with unusual numbers of blockers and nonblockers. A binary activity threshold provided a balanced training set that demonstrated higher modelability when used both alone and in combination with training sets extracted from previous studies. Overall, a well-balanced and diverse training set extracted by preprocessing of the hERG bioactivity data set from ChEMBL proves to be a valuable starting point in the development of in silico models to predict hERG blockade. By employing external training sets from previous studies alone and in combination with the ChEMBL training set, we demonstrated that our models performed better than the previously reported values in the corresponding studies. To this end, all of the data sets used and models constructed have been made publicly available.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00150.

Table S1: Criteria used to process the hERG bioactivity data from ChEMBL; Table S2: Overview of the extent of overlap of chemical space of external data sets with the training data set 3; S3: Discovery Studio software settings used for calculation of activity cliffs and MMPs; S4: Chemical space representations of the training data sets; Table S5: 10-fold CV performance of all three classifiers on ChEMBL training sets; Table S6: 10-fold CV performance of the low-confidence training set; Table S7: Detailed performance of external training sets alone and in combination with ChEMBL's training set T3; Table S8: Performance of external training sets in previous studies; Table S9: Applicability domains of the external training sets; S10: Availability of data sets and hERG models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: robert.preissner@charite.de.

ORCID

Vishal B. Siramshetty: 0000-0002-5980-8288

Robert Preissner: 0000-0002-2407-1087

Funding

Berlin-Brandenburg Research Platform BB3R, Federal Ministry of Education and Research (BMBF), Germany [031A262C]; DKTK.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Hongmao Sun and Dr. Rajarshi Guha from the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, for kindly providing their hERG data set based on the thallium flux assay.

■ ABBREVIATIONS

hERG, human ether-à-go-go-related gene; *k*-NN, *k*-nearest neighbors; SVM, support vector machine; RF, random forest; CV, cross-validation; FDA, Food and Drug Administration; CDK, Chemistry Development Kit; MACCS, Molecular Accession System; ECFP, extended connectivity fingerprint; PCA, principal component analysis; AUC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient; BACC, balanced accuracy; MODI, modelability index; MMP, matched molecular pair; APD, applicability domain

■ REFERENCES

(1) Siramshetty, V. B.; Nickel, J.; Omieczynski, C.; Gohlke, B. O.; Drwal, M. N.; Preissner, R. Withdrawn–a Resource for Withdrawn and Discontinued Drugs. *Nucleic Acids Res.* **2016**, *44*, D1080−1086.

(2) Onakpoya, I. J.; Heneghan, C. J.; Aronson, J. K. Post-Marketing Withdrawal of 462 Medicinal Products because of Adverse Drug Reactions: A Systematic Review of the World Literature. *BMC Med.* **2016**, *14*, 10.

(3) Vandenberg, J. I.; Perry, M. D.; Perrin, M. J.; Mann, S. A.; Ke, Y.; Hill, A. P. Herg K(+) Channels: Structure, Function, and Clinical Significance. *Physiol. Rev.* **2012**, *92*, 1393−1478.

(4) Brown, A. M. Drugs, Herg and Sudden Death. *Cell Calcium* **2004**, *35*, 543−547.

(5) Pearlstein, R.; Vaz, R.; Rampe, D. Understanding the Structure-Activity Relationship of the Human Ether-a-Go-Go-Related Gene Cardiac K+ Channel. A Model for Bad Behavior. *J. Med. Chem.* **2003**, *46*, 2017−2022.

(6) Kratz, J. M.; Schuster, D.; Edtbauer, M.; Saxena, P.; Mair, C. E.; Kirchebner, J.; Matuszczak, B.; Baburin, I.; Hering, S.; Rollinger, J. M. Experimentally Validated Herg Pharmacophore Models as Cardiotoxicity Prediction Tools. *J. Chem. Inf. Model.* **2014**, *54*, 2887−2901.

(7) *Guidance for Industry: S7b Nonclinical Evaluation of the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation) by Human Pharmaceuticals*; U.S. Food and Drug Administration: Rockville, MD, 2005.

(8) *Guidance for Industry: E14 Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-antiarrhythmic Drugs*; U.S. Food and Drug Administration: Rockville, MD, 2005.

(9) Polonchuk, L. Toward a New Gold Standard for Early Safety: Automated Temperature-Controlled hERG Test on the PatchLiner. *Front. Pharmacol.* **2012**, *3*, 3.

(10) Kiss, L.; Bennett, P. B.; Uebele, V. N.; Koblan, K. S.; Kane, S. A.; Neagle, B.; Schroeder, K. High Throughput Ion-Channel Pharmacology: Planar-Array-Based Voltage Clamp. *Assay Drug Dev. Technol.* **2003**, *1*, 127−135.

(11) Wen, D.; Liu, A.; Chen, F.; Yang, J.; Dai, R. Validation of Visualized Transgenic Zebrafish as a High Throughput Model to Assay Bradycardia Related Cardio Toxicity Risk Candidates. *J. Appl. Toxicol.* **2012**, *32*, 834−842.

(12) Hamill, O. P.; Marty, A.; Neher, E.; Sakmann, B.; Sigworth, F. J. Improved Patch-Clamp Techniques for High-Resolution Current Recording from Cells and Cell-Free Membrane Patches. *Pfluegers Arch.* **1981**, *391*, 85−100.

(13) Muster, W.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Muller, L.; Pahler, A. Computational Toxicology in Drug Development. *Drug Discovery Today* **2008**, *13*, 303−310.

(14) 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.

(15) Elkins, R. C.; Davies, M. R.; Brough, S. J.; Gavaghan, D. J.; Cui, Y.; Abi-Gerges, N.; Mirams, G. R. Variability in High-Throughput Ion-Channel Screening Data and Consequences for Cardiac Safety Assessment. *J. Pharmacol. Toxicol. Methods* **2013**, *68*, 112−122.

(16) Braga, R. C.; Alves, V. M.; Silva, M. F.; 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.

(17) Aronov, A. M.; Goldman, B. B. A Model for Identifying Herg K + Channel Blockers. *Bioorg. Med. Chem.* **2004**, *12*, 2307−2315.
(18) Tobita, M.; Nishikawa, T.; Nagashima, R. A Discriminant Model Constructed by the Support Vector Machine Method for Herg Potassium Channel Inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886−2890.
(19) Dubus, E.; Ijjaali, I.; Petitet, F.; Michel, A. In Silico Classification of Herg Channel Blockers: A Knowledge-Based Strategy. *ChemMedChem* **2006**, *1*, 622−630.
(20) Marchese Robinson, R. L.; Glen, R. C.; Mitchell, J. B. Development and Comparison of Herg Blocker Classifiers: Assessment on Different Datasets Yields Markedly Different Results. *Mol. Inf.* **2011**, *30*, 443−458.
(21) Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. Toward a Pharmacophore for Drugs Inducing the Long QT Syndrome: Insights from a Comfa Study of Herg K(+) Channel Blockers. *J. Med. Chem.* **2002**, *45*, 3844−3853.
(22) Cianchetta, G.; Li, Y.; Kang, J.; Rampe, D.; Fravolini, A.; Cruciani, G.; Vaz, R. J. Predictive Models for Herg Potassium Channel Blockers. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3637−3642.
(23) Leong, M. K. A Novel Approach Using Pharmacophore Ensemble/Support Vector Machine (Phe/Svm) for Prediction of Herg Liability. *Chem. Res. Toxicol.* **2007**, *20*, 217−226.
(24) Thai, K. M.; Ecker, G. F. A Binary QSAR Model for Classification of Herg Potassium Channel Blockers. *Bioorg. Med. Chem.* **2008**, *16*, 4107−4119.
(25) 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.
(26) 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.
(27) 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.
(28) Sun, H. An Accurate and Interpretable Bayesian Classification Model for Prediction of Herg Liability. *ChemMedChem* **2006**, *1*, 315−322.
(29) Jia, L.; Sun, H. Support Vector Machines Classification of Herg Liabilities Based on Atom Types. *Bioorg. Med. Chem.* **2008**, *16*, 6252−6260.
(30) Fenu, L. A.; Teisman, A.; De Buck, S. S.; Sinha, V. K.; Gilissen, R. A.; Nijsen, M. J.; Mackie, C. E.; Sanderson, W. E. Cardio-Vascular Safety Beyond Herg: In Silico Modelling of a Guinea Pig Right Atrium Assay. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 883−895.
(31) Hansen, K.; Rathke, F.; Schroeter, T.; Rast, G.; Fox, T.; Kriegl, J. M.; Mika, S. Bias-Correction of Regression Models: A Case Study on Herg Inhibition. *J. Chem. Inf. Model.* **2009**, *49*, 1486−1496.
(32) Czodrowski, P. Herg Me Out. *J. Chem. Inf. Model.* **2013**, *53*, 2240−2251.
(33) Chavan, S.; Abdelaziz, A.; Wiklander, J. G.; Nicholls, I. A. A K-Nearest Neighbor Classification of Herg K(+) Channel Blockers. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 229−236.
(34) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.
(35) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K(I) Data. *J. Med. Chem.* **2012**, *55*, 5165−5173.
(36) Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data Set Modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 1−4.
(37) Rodriguez-Perez, R.; Vogt, M.; Bajorath, J. Influence of Varying Training Set Composition and Size on Support Vector Machine-Based Prediction of Active Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 710−716.
(38) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The Chembl Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−1090.
(39) Hu, Y.; Bajorath, J. Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics. *J. Chem. Inf. Model.* **2014**, *54*, 3056−3066.
(40) Redfern, W. S.; Carlsson, L.; Davis, A. S.; Lynch, W. G.; MacKenzie, I.; Palethorpe, S.; Siegl, P. K. S.; Strang, I.; Sullivan, A. T.; Wallis, R.; Camm, A. J.; Hammond, T. G. Relationships between Preclinical Cardiac Electrophysiology, Clinical QT Interval Prolongation and Torsade De Pointes for a Broad Range of Drugs: Evidence for a Provisional Safety Margin in Drug Development. *Cardiovasc. Res.* **2003**, *58*, 32−45.
(41) Sun, H.; Huang, R.; Xia, M.; Shahane, S.; Southall, N.; Wang, Y. Prediction of hERG Liability - Using SVM Classification, Bootstrapping and Jackknifing. *Mol. Inf.* **2017**, *36*, 1600126.
(42) Doddareddy, M. R.; Klaasse, E. C.; Shagufta; Ijzerman, A. P.; Bender, A. Prospective Validation of a Comprehensive in Silico Herg Model and Its Applications to Commercial Compound and Drug Databases. *ChemMedChem* **2010**, *5*, 716−729.
(43) Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning with Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharmaceutics* **2017**, *14*, 4462−4475.
(44) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(45) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(46) Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*; IEEE, 2010; pp 3121−3124.
(47) Maggiora, G. M. On Outliers and Activity Cliffs–Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
(48) Kenny, P. W. S. J. *Structure Modification in Chemical Databases*; Wiley-VCH: Weinheim, Germany, 2004.
(49) Hu, Y.; Maggiora, G. M.; Bajorath, J. Activity Cliffs in Pubchem Confirmatory Bioassays Taking Inactive Compounds into Account. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 115−124.
(50) Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of Diverse Database Subsets Using Property-Based and Fragment-Based Molecular Descriptions. *Quant. Struct.-Act. Relat.* **2002**, *21*, 598−604.
(51) Zhang, C.; Zhou, Y.; Gu, S.; Wu, Z.; Wu, W.; Liu, C.; Wang, K.; Liu, G.; Li, W.; Lee, P. W.; Tang, Y. In Silico Prediction of Herg Potassium Channel Blockage by Chemical Category Approaches. *Toxicol. Res. (Cambridge, U. K.)* **2016**, *5*, 570−582.
(52) Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A Novel Automated Lazy Learning QSAR (All-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated All-QSAR Models. *J. Chem. Inf. Model.* **2006**, *46*, 1984−1995.
(53) Melagraki, G.; Afantitis, A.; Sarimveis, H.; Igglessi-Markopoulou, O.; Koutentis, P. A.; Kollias, G. In Silico Exploration for Identifying Structure-Activity Relationship of Mek Inhibition and Oral Bioavailability for Isothiazole Derivatives. *Chem. Biol. Drug Des.* **2010**, *76*, 397−406.
(54) Afantitis, A.; Melagraki, G.; Koutentis, P. A.; Sarimveis, H.; Kollias, G. Ligand-Based Virtual Screening Procedure for the Prediction and the Identification of Novel Beta-Amyloid Aggregation Inhibitors Using Kohonen Maps and Counterpropagation Artificial Neural Networks. *Eur. J. Med. Chem.* **2011**, *46*, 497−508.

DOI: 10.1021/acs.jcim.8b00150
J. Chem. Inf. Model. 2018, 58, 1224−1233