Critical Assessment of Artificial Intelligence Methods for Prediction of hERG Channel Inhibition in the 'Big Data' Era

Vishal B. Siramshetty, Dac-Trung Nguyen, Natalia J. Martinez, Anton Simeonov, Noel T. Southall and Alexey V. Zakharov*

National Center for Advancing Translational Sciences (NCATS), 9800 Medical Center Drive, Rockville, Maryland 20850, United States

*Corresponding author

Alexey V. Zakharov

Email: alexey.zakharov@nih.gov

ABSTRACT. The rise of novel artificial intelligence methods necessitates a comparison of this wave of new approaches with classical machine learning for a typical drug discovery project. Inhibition of the potassium ion channel, whose alpha subunit is encoded by **h**uman ***E**ther-à-go-go*-**R**elated **G**ene (hERG), leads to prolonged QT interval of the cardiac action potential and is a significant safety pharmacology target for the development of new medicines. Several computational approaches have been employed to develop prediction models for assessment of hERG liabilities of small molecules including recent work using deep learning methods. Here we perform a comprehensive comparison of prediction models based on classical (random forests and gradient boosting) and modern (deep neural networks and recurrent neural networks) artificial intelligence methods. The training set (~9000 compounds) was compiled by integrating hERG bioactivity data from ChEMBL database with experimental data generated from an in-house, high-throughput thallium flux assay. We utilized different molecular descriptors including the latent descriptors, which are real-valued continuous vectors derived from chemical autoencoders trained on a large chemical space (> 1.5 million compounds). The models were prospectively validated on ~840 in-house compounds screened in the same thallium flux assay. The deep neural networks performed significantly better than the classical methods with the latent descriptors. The recurrent neural networks that operate on SMILES provided highest model sensitivity. The best models were merged into a consensus model that offered superior performance compared to reference models from academic and commercial domains. Further, we shed light on the potential of artificial intelligence methods to exploit the chemistry big data and generate novel chemical representations useful in predictive modeling and tailoring new chemical space.

KEY WORDS. QT interval, cardiac arrhythmia, hERG, random forests, deep neural networks and recurrent neural networks.

INTRODUCTION. The *human ether-a-go-go-related gene* (hERG) encodes for the pore-forming alpha-subunit of voltage-gated potassium ion channels. The hERG channel regulates the efflux of potassium ions in cardiac myocytes and thereby plays a key role in coordination of heartbeat.[1, 2] Literature indicates that blockade of hERG channel leads to prolonged QT interval of the action potential which can result in fatal cardiac arrhythmia (*Torsade de pointes*).[3, 4] Several marketed drugs, including antiarrhythmic agents, were withdrawn after being reported to trigger cardiac arrhythmia that sometimes led to sudden death.[5, 6] Consequently, hERG channel emerged as an important off-target, marking early assessment of hERG liability an essential step in drug discovery.[7-9] The gold-standard *in vitro* and *in vivo* assays that facilitate screening of hERG channel inhibition are expensive and provide low throughput.[10-12] Meanwhile, *in silico* methods emerged as an alternative for early assessment of pharmacological and toxicological effects of chemical substances.[13-15]

Multiple studies reported *in silico* models for predicting hERG channel inhibition over the past several years. Tropsha et al. provided an overview of quantitative structure-activity relationship (QSAR) studies from the literature that were reported before 2014.[16] More recently, several other studies reported models based on simple methods like read across[17] and machine learning (ML) methods[18-21], including deep neural networks (DNNs).[22-25] Many models are based on proprietary or in-house datasets which restricts the use of this data to build newer models for academic drug discovery.[21] It is suspected that hERG channel can bind a wide variety of chemotypes and a major limitation to developing robust prediction models using publicly-domain hERG activity data is the fairly limited chemical diversity of available training data.[26] Although different combinations of ML algorithms and molecular descriptors have been tested, there is no combination of choice that performs well on unseen data. For instance, a consensus of Support Vector Machines, Random Forests, Gradient Boosting Model and Tree Bagging provided better performance in comparison to individual

models that performed very similar to each other when used with different descriptors.[16] Recent studies suggest that neural networks based on learnable representations offer broadly a better performance than classical ML algorithms.[27, 28] Latent descriptors that are derived directly from the neural network architecture are gaining popularity in molecular property prediction.[29] Furthermore, descriptor-free QSAR models that are based on recurrent neural networks (RNNs) and molecular representations like SMILES were reported to demonstrate superior generalization capabilities on out-of-domain test data.[30]

Despite this body of literature, no one has attempted to compare these methods against each other in a prospective validation study. In this study, we enriched the public domain hERG data with a dataset comprising bioactive compounds screened in a homogeneous high-throughput assay of hERG channel inhibition to provide the community with a large reference dataset of high integrity. The primary goals of the study are to develop prediction models representing both classical and novel AI developments, validate the best models on prospectively screened compounds, as well as on recently approved FDA drugs with hERG liability data.

MATERIALS AND METHODS.

Public Domain hERG Bioactivity Data. ChEMBL has been a major repository in the public domain for compound activity data extracted from scientific literature.[31] The most recently updated ChEMBL database (version 25, accessed 28 October, 2019) provides more than 20000 activity records for hERG channel (UniProt accession: Q12809). The activity records were preprocessed, as previously described in literature, to generate a high-confidence bioactivity dataset.[16, 21, 32] Briefly, only the potency and affinity values reported as IC_{50}, EC_{50} or K_i endpoints were retained, chemical structures were normalized and multiple

measurements from different assays were analyzed and treated. The post-processed dataset comprised of 6233 compounds.

Thallium Flux Assay Data. A high-throughput ion channel screen was developed and validated at National Center for Advancing Translational Sciences - NCATS (formerly known as the NIH Chemical Genomics Center) as a modified version of the FluxOR™ thallium flux assay that detects inhibition of the hERG channel by measuring flow of a surrogate ion, thallium.[33] The study compared the activities of 10 common hERG inhibitors measured in thallium flux and patch-clamp experiments and concluded that the homogeneous high-throughput assay can be used as a cost-effective alternative to patch-clamp technique.[33] More recently, NCATS reported a collection of 4,323 compounds screened in thallium flux assay used for generating support vector classification models of hERG channel inhibition.[20] In the present study, we aimed to merge this dataset with the high-confidence activity data obtained from ChEMBL database for development and critical assessment of modern machine learning approaches. First, we analyzed the correlation between flux assay data and ChEMBL data that originated from multiple assay types (e.g. electrophysiology, ion flux, radioligand binding, fluorescence, etc.). A subset of 86 approved drugs with activity data from both sources was identified for this purpose. However, since ChEMBL data spans multiple assay types, the correlation analysis was performed twice, first with patch-clamp data alone (Figure 1a) and next with data obtained after merging activities from different assay types (Figure 1b). We noticed that the outcomes from different assay types correlated well with the patch-clamp data, suggesting that the data could be used together. Furthermore, given the concordance between ChEMBL data and flux assay data, we decided to merge the two datasets to generate a combined dataset after removal of duplicates.

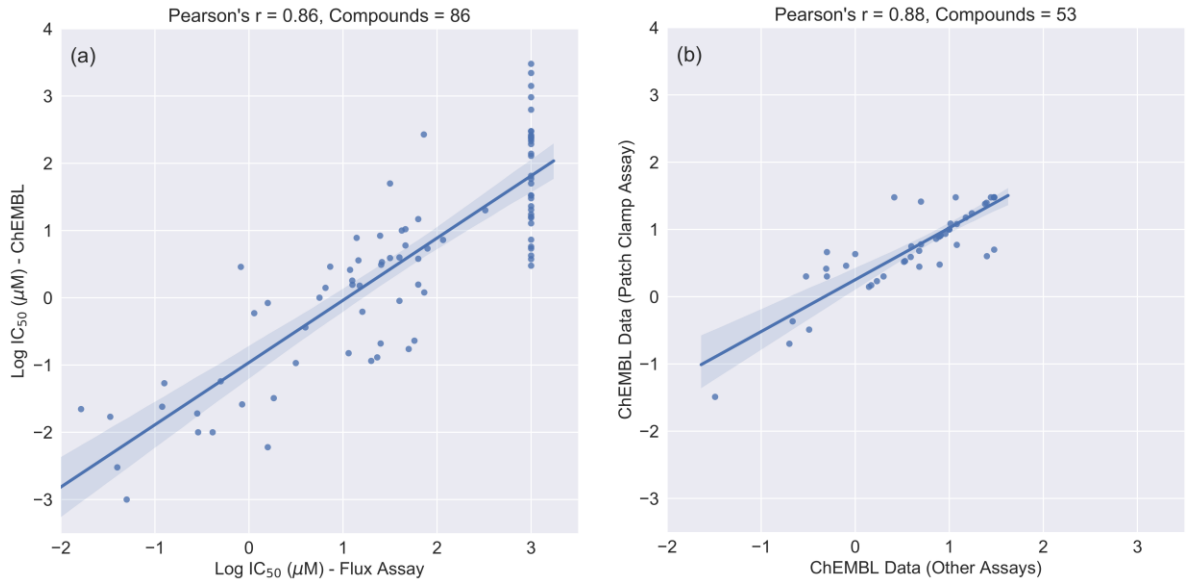

Figure 1. Correlation between hERG activity data from: (a) flux assay and ChEMBL database; (b) patch-clamp assays and other assays in ChEMBL database.

Modeling Datasets. The dataset generated by merging ChEMBL and flux assay data was used for training the classification models. Additionally, a collection of 840 compounds was selected from an in-house library to generate a test dataset for prospective validation of the classification models. These compounds were measured in the same thallium flux assay to generate IC_{50} data. We did not use a single activity threshold to discriminate blockers from non-blockers in ChEMBL whole-cell patch clamp data. Previous studies using a binary threshold (1μM and 10 μM) provided superior performance as compared to using a single threshold (1μM or 10 μM).[21] For the thallium-flux assay, a threshold of 30μM was used considering the average fold difference in the activity for compounds with data available from both sources. Finally, whole-cell patch clamp hERG data was manually extracted for 177 FDA approved drugs (2012 to 2018) from their pharmacological and safety reviews.[34] Many previously published studies do not include such data into their training or validation sets. We believe that validation on this data offers a more realistic evaluation setting for our models. Since it is hard to determine the hERG liability of a drug based on the IC_{50} value

alone,[35] without the knowledge of the peak serum concentration unbound to plasma proteins, we use both activity thresholds (1μM and 10 μM) to classify the drugs as blockers and non-blockers. Table 1 summarizes the datasets used for modeling and validation.

Table 1. An overview of the datasets used in this study.

Dataset	Sources (Assay Type)	Activity Type	Cpds	Blockers	Non-blockers
Training Set	ChEMBL (Multiple); NCATS (Flux Assay)	IC_{50}, K_i, EC_{50}	8154	2164	5990
Prospective Validation Set	NCATS (Flux Assay)	IC_{50}	839	53	786
FDA-Approved Drugs	FDA Pharmacological and Safety Reviews (Patch Clamp)	IC_{50}	177	15 (1μM) 46 (10μM)	162 (1μM) 131 (10μM)

Molecular Descriptors. Molecular fingerprints and physicochemical properties are widely used for the development of QSAR models.[36, 37] Five different sets of molecular descriptors were calculated and used in combination with different methods in this study. Morgan fingerprints (1024 bits) and RDKit descriptors were calculated using the RDKit toolkit.[38] Recent progress in deep learning facilitates the development of the different molecular representations such as latent vectors that are used as descriptors for modeling molecular properties.[29, 39, 40] In this study we utilized several different approaches to generate fixed-length latent vector representations for our modeling sets. The detailed methodology involved in generation of latent vectors is described in the next section. RNNs that learn on sequential data such as SMILES as input have been employed for molecular property prediction.[30, 41, 42] Three variations of SMILES including the commonly known canonical SMILES, randomized SMILES[43] and DeepSMILES[44] were employed as input for RNN models in this study. While

the RDKit toolkit was used to generate canonical SMILES, we relied on their original implementations for randomized SMILES and DeepSMILES.

Autoencoder and Adversarial Autoencoder Models. RNNs that encode SMILES and decode them back to SMILES constitute the most recent generation of methods for ligand-based *de novo* design.[45] In order to generate new structures, an RNN learns to predict the probability of the next character in a SMILES string, given the previous characters.[45] An autoencoder (AE) constitutes a special architecture that generates a compressed representation of the provided input that could be used to reconstruct the chemical structures. The code layer (i.e., the compressed representation) produced by an autoencoder is a fixed-length vector of descriptors increasingly referred to as 'latent descriptors' that have been used for molecular property prediction.[40, 46] Variations of the original autoencoder architecture have been proposed that translate between different string representations of molecules (e.g., canonical SMILES, InChI, etc.). One such variation focused on the ability of AEs to generate new samples which resulted in variational autoencoders (VAEs).[39] Adversarial autoencoders (AAEs) are modifications of VAEs in which an AE is combined with a generative adversarial network (GAN). GANs facilitate generation of novel structures given a prior distribution of the training set examples without the explicit need to define and manipulate a probability distribution. While AEs have been previously reported to be used for modeling molecular properties, this is the first study to use AAEs to generate latent descriptors for QSAR. Both AE and AAE models were built using data from ChEMBL database (version 25).

The latest release of ChEMBL (version 25) contains a total of 1,870,462 molecules. The reason behind choosing ChEMBL is that most small molecules from the database have been synthesized and tested against biological targets which indicates that there are good

chances to learn representations of synthesizable bioactive molecules. This set was first preprocessed using the criteria suggested by Brown et al.[47] after which we ended up with 1,641,316 molecules. Briefly, salts were removed, charges were neutralized, length of SMILES was restricted to 100 characters and lastly, omitted molecules with metals and other unusual atoms. After partitioning, the training and test sets contained 1,313,054 molecules and 328,262 molecules, respectively. The training set was used to build AE and AAE models that were validated on the test set for reconstruction ability. Training of models were performed in batches of 256 compounds for a total of five epochs with the length of latent descriptor predefined as 512. Model summaries and the reconstruction performances of the AE and AAE architectures are provided in the supporting information (S1 and S2). The smiles2latent models from the AE and AAE architectures were used to generated latent descriptors Latent1 and Latent2, respectively.

QSAR Models. Two different modeling pipelines were developed. The first pipeline consisted of the classical ML methods, based on four descriptors (RDKit, MorganFP, Latent1 and Latent2), that serve as baseline models. *Scikit-learn*,[48] an open-source ML library was used to train and validate the models. The second pipeline consisted of two types of neural networks: feed-forward neural networks based on the four descriptors and RNNs based on SMILES. The models were built using *Keras*[49] deep learning library with *TensorFlow*[50] as background. We implemented two types of data splits: random split and scaffold split, both adapted from the *DeepChem*[51] library. The training set was partitioned into internal training (80%) and test (20%) sets. Parameter optimization was performed on these partitions in a five-fold cross-validation format. For this purpose, the split was performed five times each for both split types. Finally, models based on the best parameters were built using the unpartitioned

training set and evaluated on the prospective validation set and FDA approved drugs. All learning methods are briefly explained below.

Random Forests

Random forest (RF)[52] is an ensemble of decision trees that are fitted on various subsamples of the data and uses averaging to restrict overfitting and improve prediction accuracy. The 'RandomForestClassifier' method from *Scikit-learn* was used to build the model. The number of estimators was set to 300 and random state was set to an integer. The rest of the parameters were set to default values.

Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is an ML method that allows both regression and classification. It is based on the Gradient Boosting Decision Tree technique and has been widely applied in the field of data mining.[53] Due to its recently gained popularity over RF in cheminformatics, we used XGBoost as the second baseline method. Similar to RF, XGBoost models were used with a total of 300 estimators and random state set to an integer. The remaining parameters were set to default values.

Feedforward Neural Networks

Artificial neural networks (ANNs) have been applied for a wide range of QSAR tasks.[54-57] Increase in the use of RF and Support Vector Machines for classification and regression in cheminformatics led to a decline in the use of ANNs. Eventually, the ANNs have evolved into DNNs. Unlike ANNs, a DNN consists of multiple fully connected layers with two or more hidden (or intermediate) layers between the input and output layers. In a feedforward neural network (referred simply as DNN in the rest of the study), the information passed through the input layer flows in forward direction through the hidden layers to the output

layer.[58] A number of parameters are available for tuning a DNN such as the number of hidden layers, number of epochs, activation function, optimizer and its learning rate. Hyperparameter optimization is essential to improve the performance of DNN and avoid overfitting on training data. This is detailed in the *Model Optimization* section of the results.

RNNs based on SMILES

Long Short-Term Memory (LSTM) networks are RNNs that can be used to model sequence data such as natural language.[41] Previous studies reported the use of LSTMs to learn directly from SMILES which led the community towards descriptor-free QSAR models.[30] The LSTM networks built in this study were fed with canonical SMILES that are first encoded into one-hot vectors and then passed to the computing cell which performs as many computations as the length of the input SMILES in a loop. At each step, one character of SMILES is taken as input and the computed activation value is passed to the next step which takes the next character as input. In this way, the information from previous characters is persisted while the next characters are being processed. Finally, the network produces a prediction probability between 0 and 1. These values can be used to obtain the binary classification labels. Furthermore, we investigated attention-based modeling in which the neural network architecture is extended to search for parts of the input sequences that are relevant to the target variable.[59, 60] In case of LSTM networks, the attention mechanism gives importance to certain parts of the sequence (i.e., SMILES) rather than considering the whole sequence as important. For this purpose, we implemented Multiplicative Attention from *Keras Self-Attention* library[61] with regularization and without any attention bias.

Performance Assessment. The performance of the models was mainly accessed using the area under the curve (AUC) from the receiver operating characteristic (ROC) curves. A ROC

curve plots the true positive rate against the false positive rate and thus provides an estimate of the performance of a binary classifier. In addition to AUC, the following metrics were estimated:

Sensitivity and Specificity

The sensitivity (or the true positive rate) of a model is the proportion of hERG blockers correctly predicted as blockers. Specificity (or the true negative rate) is the proportion of non-blockers correctly predicted as non-blockers.

$$\text{Sensitivity } (Sens) = TP/(TP + FN)$$

$$\text{Specificity } (Spec) = TN/(TN + FP)$$

Here, *TP* = number of true positives; *FN* = number of false negatives; *TN* = number of true negatives; and *FP* = number of false positives.

Balanced Accuracy

The balanced accuracy (BACC) of a model is an average of the proportions correctly predicted for each class (i.e., *Sensitivity* and *Specificity*).

$$\text{Balanced accuracy } (BACC) = (Sensitivity + Specificity)/2$$

Reference Models. We compared our results with state-of-the-art models from Pred-hERG and StarDrop.[62, 63] Pred-hERG provides consensus predictions based on different machine learning models and molecular fingerprints.[16] Their models outperformed several existing models and were made publicly accessible while constantly being updated with hERG

bioactivity data from ChEMBL database.[63, 64] StarDrop's hERG model is available from the ADME QSAR module of the software. Gold standard patch-clamp IC_{50} data was used to build a regression model based on a non-linear Gaussian Processes technique. These two models represent two different domains (academic and commercial) and we believed that it would be appropriate to use them for a comprehensive comparison.

Availability of Models and Datasets. All datasets and model implementations are available in our GitHub repository (https://github.com/ncats/herg-ml). The preprocessed ChEMBL data and scripts for building AE and AAE models are also provided. In addition to the training data, structures and hERG activity annotations for the prospective validation set compounds and FDA approved drugs subset are made publicly available.

RESULTS AND DISCUSSION

Model Optimization. The baseline methods Random Forests and XGBoost are robust and do not require extensive parameter optimization.[65] The quality of deep learning models is more dependent on the number of descriptors, hyperparameters and computational capabilities (e.g., use of GPU). We first report parameter optimization performed using one of the five-fold cross-validation training sets. For the DNN, a series of 260 models were built using different combinations of optimizer learning rate, activation function, number of epochs and batch size for each descriptor type. The same dense layer architecture was maintained for the first round of grid search and once the best parameters were obtained, we tried to find an optimal dense layer structure for each descriptor. The DNNs in this study typically consisted of three to five layers with decreasing number of neurons as it moves forward that resulted in a pyramidal network structure, previously reported as an optimal setting for DNNs.[66, 67] The number of units in the input layer was defined based on the shape of the incoming descriptors and was reduced

in the hidden layers and finally, the output layer consists of only one unit which uses a *sigmoid* activation function to return the output. Different combinations of the number of hidden layers and the number of neurons per hidden layer were examined and the best performing architecture was retained for both five-fold cross-validation and final validation. In the case of LSTM, along with the parameters considered for DNN, we also investigated the number of LSTM units. The grid search results for DNN are provided in the supporting information (S3). While the best performing parameters varied with the descriptors for DNN, *relu* activation function always provided superior accuracy. The LSTM models provided best performance with Adam optimizer (learning rate = 0.01), *tanh* activation function, 64 LSTM units, 64 neurons in the first dense layer, 10 epochs and batch size of 128. These settings were retained for the LSTM model with attention mechanism in which different attention widths (2, 4, 6, 8, 16) were evaluated. In general, smaller attention widths provided superior performance.

Table 2. Overview of the hyperparameter settings used following a grid search optimization.

Method	**Descriptors (Length)**	**Hyperparameters**
DNN	RDKit (119) MorganFP (1024) Latent1 (512) Latent2 (512)	*activation* = [relu, selu] *epochs* = [10, 20, 30] *batch_size* = [32, 64, 128] *learn_rate* = [0.0001, 0.0005, 0.00001, 0.00005] *dense_layers* = 3 to 5
LSTM	Canonical SMILES (100)	*activation* = [relu, tanh] *epochs* = [5, 10, 15] *batch_size* = [64, 128] *learn_rate* = [0.01, 0.001, 0.0001] *dense_layers* = 2 *lstm_dim* = [64, 128]

Classical AI versus Modern AI Methods. Based on five-fold cross-validation performed on the training set, we expected that the modern methods would provide relatively better performance in comparison to the classical methods. The baseline models based on RF and XGBoost provided highest AUC (Figure 2) and BACC (Table 3) with RDKit descriptors. However, these two methods provided the worst performance (Sensitivity < 0.6) when used together with the latent descriptors. In contrast, the DNN models provided better performance (BACC > 0.8) with latent descriptors. Though the DNN models demonstrated fairly similar performance with different descriptor types, MorganFP provided a better tradeoff between Sensitivity and Specificity. The LSTM models based on SMILES performed on par with the DNN and baseline models with BACC as high as 0.81. RF and DNN were the best performers in cross-validation. Overall, for different methods and descriptors, 'random split' provided superior performance in comparison to scaffold split. Results for cross-validation using scaffold split are provided in supporting information (S4).

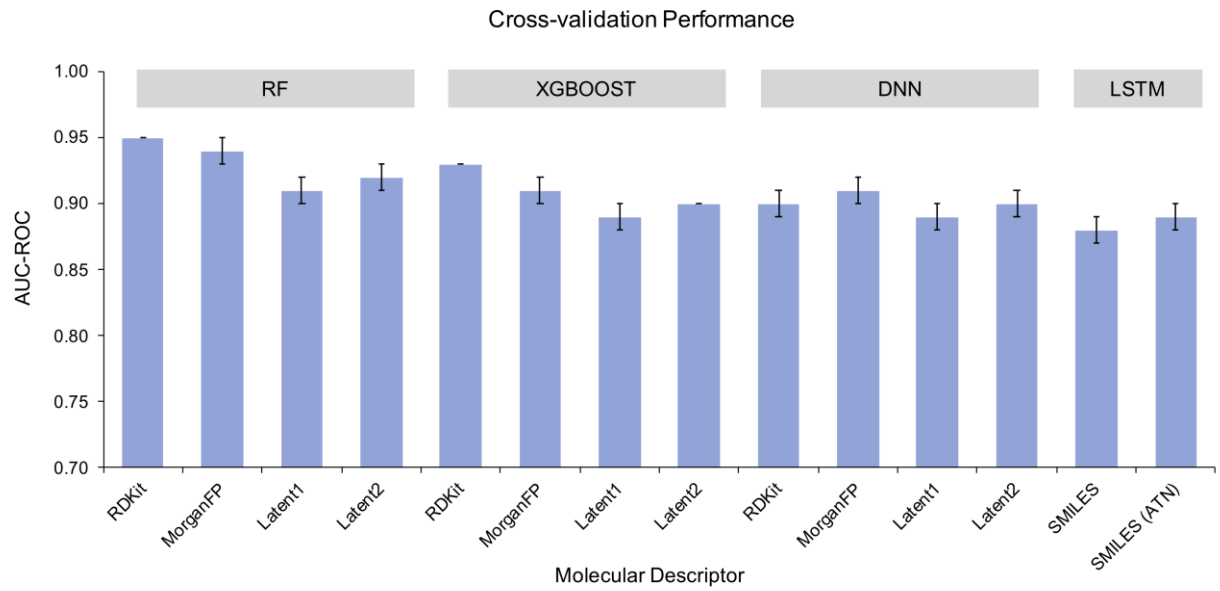

Figure 2. Model performance on training data generated using the random splitting scheme. For each method-descriptor combination, the standard deviation of the average of performance for different folds (*N*=5) is presented as an error bar.

The prospective validation set containing 839 compounds was used to evaluate the models. A nearest-neighbor analysis with the training set revealed that a majority (>80%) of these compounds are below a Tanimoto similarity threshold of 0.6 (supporting information, S5). The optimal settings from cross-validation were retained for DNN and LSTM models. A performance trend similar to cross-validation was observed (Table 4), except that the XGBoost model based on RDKit descriptors provided the best performance. The DNN model performed well with all descriptor types while RF performed the best using RDKit descriptors. The LSTM model based on attention mechanism provided better AUC and BACC in comparison to the model without attention. Overall, recent developments such as DNN and LSTM provide robust predictions using different descriptors and simple sequence-based descriptor such as SMILES. In particular, the latent descriptors derived from encoder-decoder architectures performed very well on validation set and emphasize their applicability in prediction of molecular properties and biological activity. However, classical ML methods such as XGBoost and RF are still in the league of best performing models, in agreement with previous studies.[16, 21]

Table 3. Other cross-validation performance metrics. For each model, the average of different folds (*N*=5) and corresponding standard deviation are listed.

Method	Descriptor	BACC	Sensitivity	Specificity
RF	RDKit	0.84 +/- 0.02	0.73 +/- 0.04	0.95 +/- 0.01
	MorganFP	0.83 +/- 0.02	0.68 +/- 0.03	0.97 +/- 0.01
	Latent1	0.73 +/- 0.01	0.48 +/- 0.03	0.98 +/- 0.01
	Latent2	0.74 +/- 0.01	0.50 +/- 0.02	0.98 +/- 0.01
XGBoost	RDKit	0.82 +/- 0.01	0.72 +/- 0.02	0.93 +/- 0.01
	MorganFP	0.79 +/- 0.02	0.62 +/- 0.03	0.96 +/- 0.01
	Latent1	0.75 +/- 0.01	0.56 +/- 0.02	0.94 +/- 0.01
	Latent2	0.76 +/- 0.01	0.58 +/- 0.01	0.94 +/- 0.01

DNN	RDKit	0.81 +/- 0.09	0.86 +/- 0.04	0.76 +/- 0.03
	MorganFP	0.84 +/- 0.02	0.75 +/- 0.04	0.93 +/- 0.01
	Latent1	0.82 +/- 0.01	0.73 +/- 0.04	0.91 +/- 0.03
	Latent2	0.82 +/- 0.01	0.72 +/- 0.04	0.91 +/- 0.02
LSTM	SMILES	0.80 +/- 0.01	0.85 +/- 0.03	0.75 +/- 0.02
	SMILES-ATN	0.81 +/- 0.01	0.78 +/- 0.04	0.83 +/- 0.02

Similar to cross-validation, XGBoost and DNN obtained better Sensitivity over the RF models when using the latent descriptors. Furthermore, two different latent descriptors provided similar results. In few cases, descriptors based on AAE model (Latent2) provided slightly better results. Recently, Gomez-Bombarelli, *et al.* proposed a variational autoencoder (VAE), which is an autoencoder with generative ability to propose new compounds with desired properties.[39] The original implementation of this VAE model, built on a subset of ZINC database,[68] was used to generate latent descriptors of length 192 bits for both training and validation set compounds. DNNs were used to evaluate the utility of these descriptors for prediction of hERG channel blockade. Hyperparameter optimization was performed similar to other descriptors. This model did not perform as well as the DNN models based on our latent descriptors (Table 5). This could be due to the fact that the latent space of the VAE model was originally shaped for predicting specific molecular properties such as the water-octanol partition coefficient. We also noticed that the reconstruction rate of the encoder-decoder model can influence the QSAR model performance. An inverse correlation was observed between the reconstruction rate of the encoder-decoder models and the improvement in performance of QSAR models using the latent descriptors. Considering this into account, we trained our AE and AAE models in a small number of epochs to limit the reconstruction rate and obtain optimal performance using the latent descriptors. However, a detailed investigation to arrive at the best

reconstruction rate could not be performed due to the huge computational costs involved development of these models.

Table 4. Performance of the models on prospective validation set.

Method	Descriptor	AUC-ROC	BACC	Sensitivity	Specificity
RF	RDKit	0.84	0.77	0.66	0.88
	MorganFP	0.81	0.68	0.40	0.97
	Latent1	0.76	0.64	0.32	0.97
	Latent2	0.75	0.63	0.30	0.97
XGBoost	RDKit	0.84	0.80	0.77	0.83
	MorganFP	0.83	0.65	0.38	0.92
	Latent1	0.78	0.69	0.49	0.88
	Latent2	0.76	0.62	0.38	0.87
DNN	RDKit	0.82	0.74	0.75	0.72
	MorganFP	0.83	0.76	0.66	0.85
	Latent1	0.78	0.73	0.74	0.72
	Latent2	0.80	0.74	0.72	0.75
LSTM	SMILES	0.76	0.72	0.77	0.67
	SMILES-ATN	0.79	0.75	0.75	0.74

Table 5. Performance of DNN models based on different latent descriptors.

Descriptor	Length	AUC-ROC	BACC	Sensitivity	Specificity
Latent2 (best model)	512 bits	0.80	0.74	0.72	0.75
Latent VAE	196 bits	0.75	0.69	0.64	0.74

O'Boyle & Dalke recently proposed an adaption of the original SMILES known as DeepSMILES that could be used instead of the conventional SMILES representations in building generative neural networks.[44] They tried to address the syntactical limitations of

SMILES that could be a reason behind the poor validity of the newly generated structures. In another benchmark, canonical SMILES and DeepSMILES were compared to 'Randomized SMILES' for the development of generative RNN models.[43] Randomized SMILES were earlier proposed as a data augmentation technique to improve the performance of QSAR models.[69] Further, they were also shown to improve the relevance of latent descriptors for QSAR when used in generation of autoencoder models.[29] In this study, we developed LSTM models using these two SMILES adaptations and compared the performance with our best LSTM model based on canonical SMILES. In the case of Randomized SMILES, different enumeration factors (e = 2, 3, 4, 5) were considered i.e. in case of e = 5, five unique randomized SMILES were generated for each molecule in the training set. In all cases, the LSTM model started to provide higher Sensitivity although the overall performance declined. Similarly, DeepSMILES did not perform as well as the canonical SMILES (see Table 6). In order to evaluate DeepSMILES on a larger dataset, the AE model developed in this study was rebuilt using the same ChEMBL data but this time using DeepSMILES. Again, the AE model based on canonical SMILES resulted in a higher reconstruction performance and the latent descriptors derived from the same model provided better QSAR performance.

Table 6. LSTM models based on different SMILES representation. For canonical SMILES, the performance values reported are from LSTM model with attention mechanism.

SMILES Type	AUC-ROC	BACC	Sensitivity	Specificity
Canonical SMILES	0.79	0.75	0.75	0.74
DeepSMILES	0.76	0.71	0.73	0.69
Randomized SMILES	0.76	0.69	0.79	0.59

Comparison with available hERG models. Reference models from Pred-hERG webserver and StarDrop software were used to obtained predictions for the validation set. The StarDrop model predicts a pIC_{50} value for each compound. An activity threshold of $pIC_{50} = 6$ was used

to assign the final prediction labels. Predictions from Pred-hERG model are based on a consensus of four different machine learning methods (RF, Gradient Boosting, TreeBag and Support Vector Machines) that were individually built using different molecular descriptors (pharmacophoric fingerprints, featMorgan fingerprints, PubChem fingerprints and MACCS fingerprints). Similarly, we generated a consensus model based on four different methods, each in combination with the best performing molecular descriptor when tested individually (RF-RDKit; XGBoost-RDKit; DNN-MorganFP; LSTM-ATN-SMILES). Our consensus model outperformed both reference models in predicting the validation set (see Table 7).

Table 7. Performance of our consensus model in comparison to the reference hERG models.

hERG Model	**BACC**	**Sensitivity**	**Specificity**
StarDrop 6.4.0	0.69	0.58	0.79
Pred-hERG 4.2	0.77	0.74	0.81
Our Consensus	0.80	0.74	0.86

Performance on recently approved drugs. The consensus model was used to predict the hERG liabilities of the FDA approved drugs. Since two activity thresholds are used, predictions were validated twice taking into account different thresholds and the results are presented in Table 8. With more stringent activity criteria, the consensus model achieved a BACC of 0.79. Similar to the validation set, a majority (>80%) of these drugs were found to be below a Tanimoto similarity threshold of 0.6 (supporting information, S5). Thus, we demonstrated the ability of our models to provide robust predictions on unseen chemical space. At the same time, it is clear that the activity threshold used to separate blockers from non-blockers can result in a completely different dataset and model performance. While 10μM is the generally accepted threshold, in the case of this dataset, we believe that 1μM offers a realistic composition with more non-blockers than blockers. Furthermore, no clear

trend was noticed in the evaluated time period (2012 to 2018) for the potential of newly approved drugs to inhibit hERG (see Figure 3), while the expectation was to observe a decrease in the inhibitory potential over the time. This should draw the attention of the community to the question - *is hERG blockade still a concern for drug discovery?*

Table 8. Performance of the consensus model on FDA approved drugs.

Activity Threshold	AUC-ROC	BACC	Sensitivity	Specificity
1μM	0.79	0.75	0.71	0.78
10μM	0.77	0.67	0.44	0.89

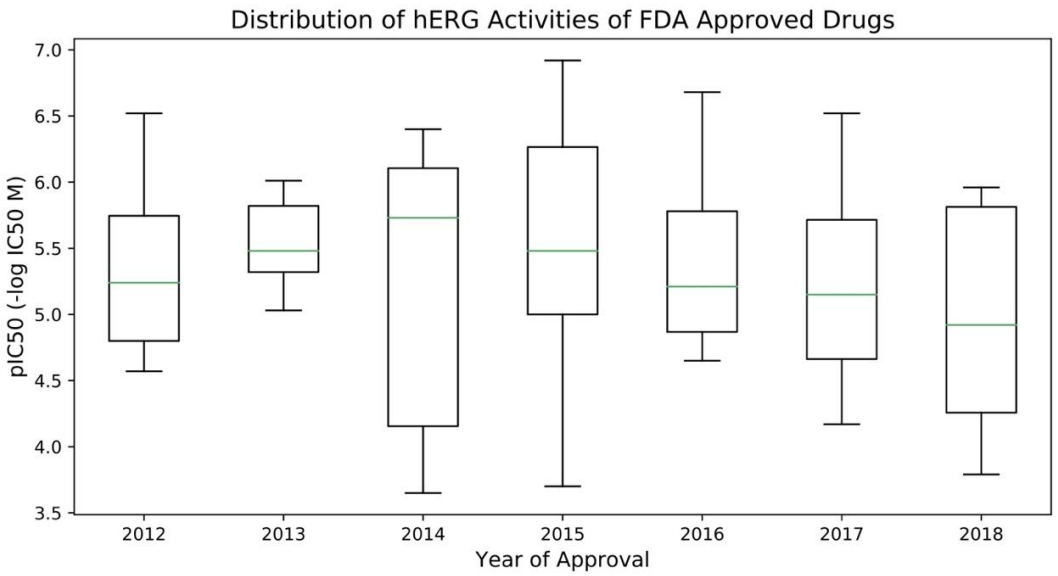

Figure 3. Trend of hERG activities of 72 drugs approved by FDA between 2012 and 2018.

Traditional descriptors versus Latent Descriptors. The choice of descriptors is a key factor for the development of a robust predictive model. Except SMILES, all descriptors used in this study are numerical descriptors: either binary (MorganFP) or real-valued (RDKit, Latent1 and Latent2). The performance obtained using latent descriptors from AE and AAE models was comparable to that obtained using fingerprints and other descriptors only when employed

with DNNs. The poor performance of RF and XGBoost models with latent descriptors could be attributed to the continuous distribution of the compounds in the low-dimensional space. Overall, MorganFP performed the best among all numerical descriptors. The PCA plots in Figure 4 indicates that the blockers could be better discriminated from the non-blockers by MorganFP. The continuous distribution of compounds in the latent space explains the poor ability of simple classifiers such as RF and XGBoost to distinguish blockers from non-blockers. Previously, these representations have been shown to provide improvements over baseline models based on molecular fingerprints.[29] It is also worth noting that these representations are not only useful in reconstruction of molecules but also in capturing properties of molecules that include biological activity.

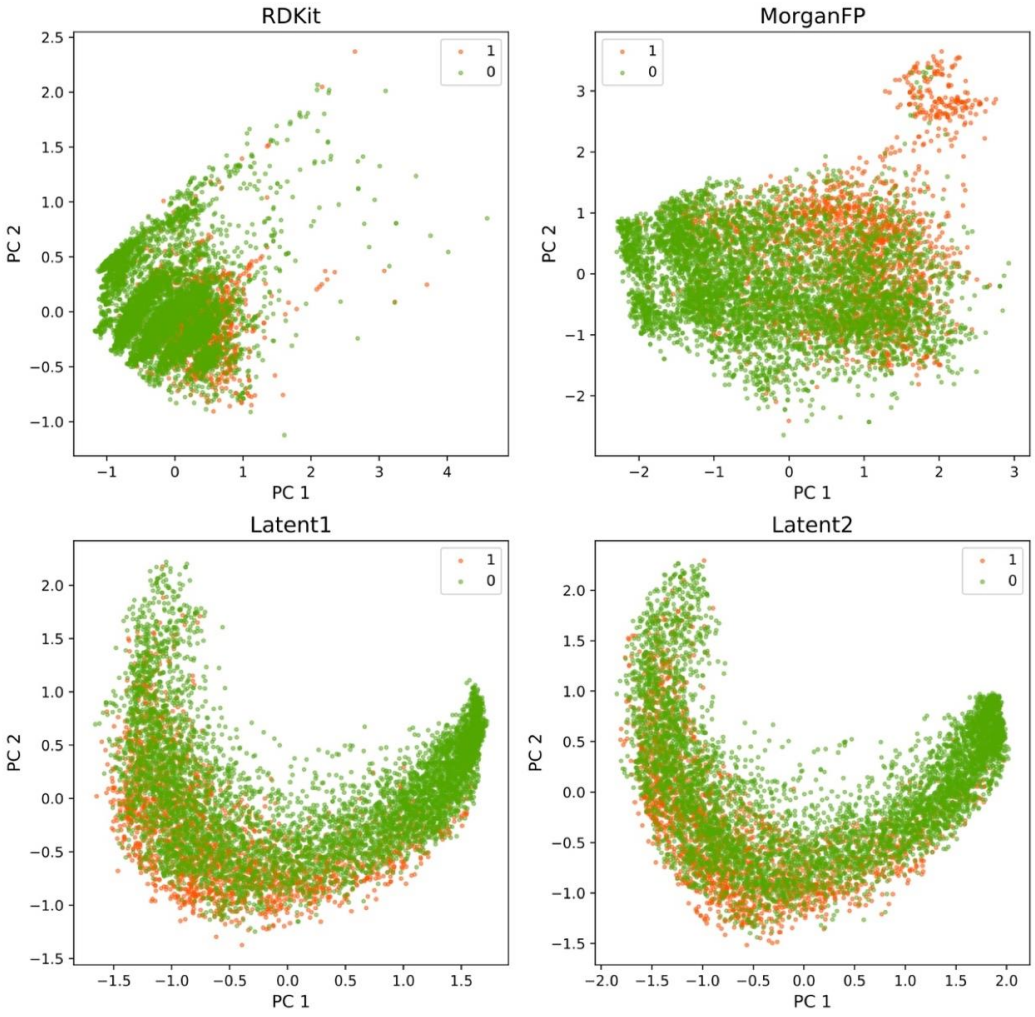

Figure 4. Two-dimensional PCA plots based on different descriptors for the training data: blockers (1) and non-blockers (0) are represented in red and green colors, respectively.

The choice of chemical representation has a great influence on the properties of the latent space. Thus, we investigated whether using molecular fingerprints as input strings for the autoencoder model could improve the QSAR performance of the latent descriptors as the fingerprint length is fixed by default and the vocabulary is essentially simple. As anticipated, the latent descriptors derived from fingerprints provided a completely different distribution of the training data (supporting information, S6). Performance of the RF, XGBoost and DNN models trained on these descriptors was competitive although the latent descriptors that originated from canonical SMILES performed relatively better (supporting information, S7).

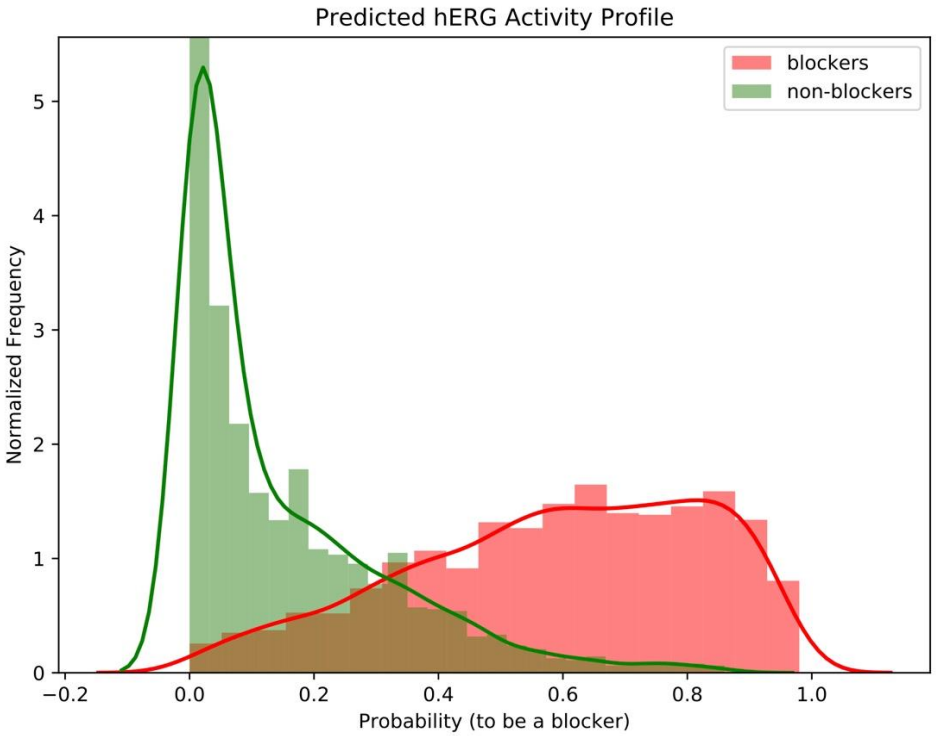

Figure 5. Distribution of the sampled compounds based on their probabilities to be hERG blockers. The solid lines represent the shapes of distribution of the corresponding subsets.

Generating compounds without hERG liabilities. The recently introduced sequence-to-sequence based models that rely on the encoded representation (i.e. latent space) of molecules facilitates exploration of new chemical space. Apart from its novel applicability in QSAR modeling[29] and virtual screening[40], the encoded representation has been explored to generate focused chemical libraries with molecular properties of interest.[39, 70-74] A number of key factors such as validity, novelty, diversity and synthetic feasibility of the sampled molecules have been addressed. Such models have been recently reported to identify promising drug candidates.[75] In this context, the AAE architecture was used to sample new compounds using hERG blockers and non-blockers as separate starting points. Distribution of the prediction probabilities for the newly generated compounds (using the consensus model) revealed that most compounds generated around non-blockers were predicted as non-blockers by the consensus model (Figure 5). Similarly, a majority of new structures sampled from the blockers were predicted as blockers (Figure 5 and Figure 6).

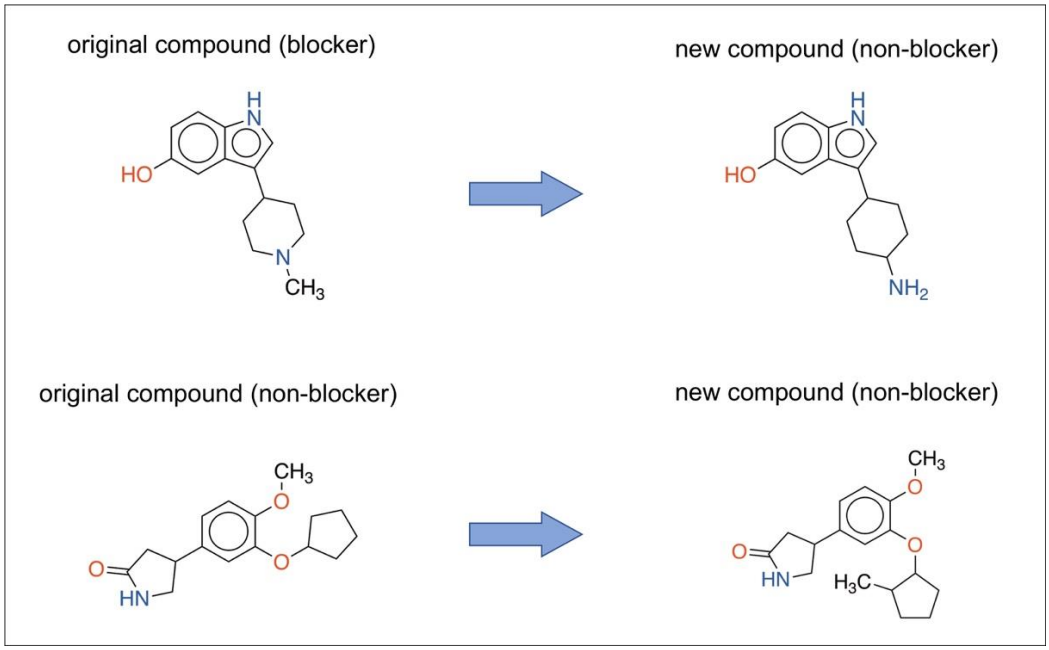

Figure 6. Exemplary compounds sampled from blockers and non-blockers in the training set. The activities of new compounds were predicted using the consensus model.

Although synthesizability of the generated structures is a bottleneck for generative models, it was recently demonstrated that the fraction of synthesizable molecules is comparable to that of training set used to derive the new compounds.[76] Since our training set compounds originate from ChEMBL database that reports bioactivities for already synthesized compounds and in-house high-throughput assay, it is our expectation that the newly generated compounds have similar rate of synthesizability. Furthermore, the generated chemical structures are fairly diverse and not completely similar to the original training set subsets used for sampling (Figure 7; supporting information, S8). These findings emphasize the potential of generative models in designing new chemical libraries with desired properties (or poor toxic liabilities), particularly in combination with predictive models.

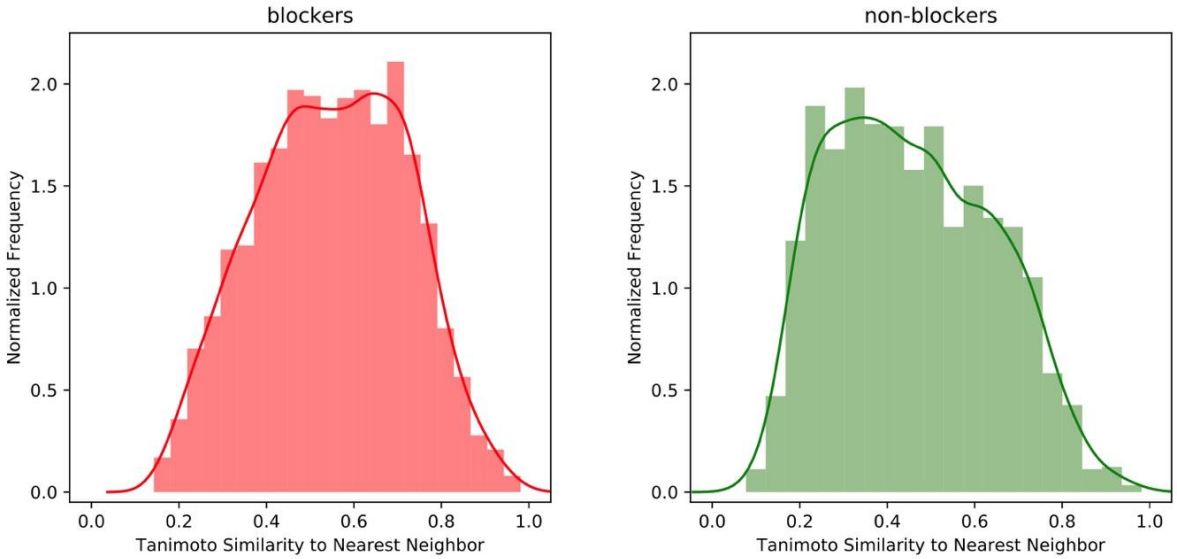

Figure 7. Distribution of newly sampled compounds based on Tanimoto similarity towards the nearest neighbor in the original blocker and non-blocker subsets. Solid lines represent the shapes of the distributions.

CONCLUSIONS. Modeling hERG channel inhibition has been important ever since the recall of marketed drugs due to fatal cardiac arrhythmias. To date, several computational modeling approaches have been proposed for early assessment of hERG liability and several *in silico* models have been reported in the recent years. In this study, both classical and modern learning approaches were evaluated and compared for their ability to predict hERG liabilities of small molecules. Both feed-forward neural networks (DNN models) and recurrent neural networks (LSTM models) performed on par with classical machine learning methods. It was also demonstrated that novel representations derived from the latent space of chemical autoencoders offer an alternative to traditional descriptors in structure-activity and structure-property modeling. Particularly, the DNNs provided a significantly better performance using these novel descriptors. Further, the utility of generative models to derive a new chemical space with a property of interest has been demonstrated. In addition, we also provide a high-quality reference dataset obtained by combining public domain hERG activity data with experimental data generated in a high-throughput thallium flux assay as well as hERG activity data for small molecules approved between 2012 and 2018. The validation data from this study can be used to evaluate hERG models proposed in future studies.

ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the National Institutes of Health, National Center for Advancing Translational Sciences.

SUPPORTING INFORMATION

S1, **S2**. Summaries of the autoencoder model and adversarial autoencoder models; **S3**. Results of hyperparameter optimization; **S4**. Cross-validation results based on scaffold split; **S5**. PCA plot for latent descriptors derived from fingerprints; **S6**. Comparison of performance of different latent descriptors; **S7**. Correlation of hERG activity and similarity of newly generated compounds towards training set.

REFERENCES

1. Smith, P. L.; Baukrowitz, T.; Yellen, G., The inward rectification mechanism of the HERG cardiac potassium channel. *Nature* **1996**, *379*, 833-836.
2. 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.
3. Sanguinetti, M. C.; Jiang, C.; Curran, M. E.; Keating, M. T., A mechanistic link between an inherited and an acquired cardiac arrhythmia: HERG encodes the IKr potassium channel. *Cell* **1995**, *81*, 299-307.
4. Sanguinetti, M. C.; Tristani-Firouzi, M., hERG potassium channels and cardiac arrhythmia. *Nature* **2006**, *440*, 463-469.
5. Haverkamp, W.; Breithardt, G.; Camm, A. J.; Janse, M. J.; Rosen, M. R.; Antzelevitch, C.; Escande, D.; Franz, M.; Malik, M.; Moss, A.; Shah, R., The potential for QT prolongation and proarrhythmia by non-antiarrhythmic drugs: clinical and regulatory implications. Report on a policy conference of the European Society of Cardiology. *Eur Heart J* **2000**, *21*, 1216-1231.
6. 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.
7. Witchel, H. J., The hERG potassium channel as a therapeutic target. *Expert Opin Ther Targets* **2007**, *11*, 321-336.
8. Raschi, E.; Vasina, V.; Poluzzi, E.; De Ponti, F., The hERG K+ channel: target and antitarget strategies in drug development. *Pharmacol Res* **2008**, *57*, 181-195.
9. Darpo, B.; Nebout, T.; Sager, P. T., Clinical evaluation of QT/QTc prolongation and proarrhythmic potential for nonantiarrhythmic drugs: the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use E14 guideline. *J Clin Pharmacol* **2006**, *46*, 498-507.
10. 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.
11. Wisniowska, B.; Polak, S., hERG in vitro interchange factors--development and verification. *Toxicol Mech Methods* **2009**, *19*, 278-284.
12. 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.
13. Jorgensen, W. L., The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813-1818.
14. Xiang, M.; Cao, Y.; Fan, W.; Chen, L.; Mo, Y., Computer-aided drug design: lead discovery and optimization. *Comb Chem High Throughput Screen* **2012**, *15*, 328-337.

15. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr., Computational methods in drug discovery. *Pharmacol Rev* **2014**, *66*, 334-395.
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. Alves, V. M.; Golbraikh, A.; Capuzzi, S. J.; Liu, K.; Lam, W. I.; Korn, D. R.; Pozefsky, D.; Andrade, C. H.; Muratov, E. N.; Tropsha, A., Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure-Activity Relationship Models. *J Chem Inf Model* **2018**, *58*, 1214-1223.
18. 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.
19. Wang, S.; Sun, H.; Liu, H.; Li, D.; Li, Y.; Hou, T., ADMET Evaluation in Drug Discovery. 16. Predicting hERG Blockers by Combining Multiple Pharmacophores and Machine Learning Approaches. *Mol Pharm* **2016**, *13*, 2855-2866.
20. Sun, H.; Huang, R.; Xia, M.; Shahane, S.; Southall, N.; Wang, Y., Prediction of hERG Liability - Using SVM Classification, Bootstrapping and Jackknifing. *Mol Inform* **2017**, *36*.
21. Siramshetty, V. B.; Chen, Q.; Devarakonda, P.; Preissner, R., The Catch-22 of Predicting hERG Blockade Using Publicly Accessible Bioactivity Data. *J Chem Inf Model* **2018**, *58*, 1224-1233.
22. Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V., Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model* **2015**, *55*, 263-274.
23. 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 Pharm* **2017**, *14*, 4462-4475.
24. Sharifi, M.; Buzatu, D.; Harris, S.; Wilkes, J., Development of models for predicting Torsade de Pointes cardiac arrhythmias using perceptron neural networks. *BMC Bioinformatics* **2017**, *18*, 497.
25. Cai, C.; Guo, P.; Zhou, Y.; Zhou, J.; Wang, Q.; Zhang, F.; Fang, J.; Cheng, F., Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. *J Chem Inf Model* **2019**, *59*, 1073-1084.
26. Vandenberg, J. I.; Perozo, E.; Allen, T. W., Towards a Structural View of Drug Binding to hERG K(+) Channels. *Trends Pharmacol Sci* **2017**, *38*, 899-907.
27. Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* **2018**, *9*, 513-530.
28. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R., Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model* **2019**, *59*, 3370-3388.
29. Bjerrum, E. J.; Sattarov, B., Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8*.
30. Chakravarti, S. K.; Alla, S. R. M., Descriptor Free QSAR Modeling Using Deep Learning With Long Short-Term Memory Neural Networks. *Frontiers in Artificial Intelligence* **2019**, *2*, 17.
31. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; 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.
32. Fourches, D.; Muratov, E.; Tropsha, A., Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model* **2016**, *56*, 1243-1252.
33. Titus, S. A.; Beacham, D.; Shahane, S. A.; Southall, N.; Xia, M.; Huang, R.; Hooten, E.; Zhao, Y.; Shou, L.; Austin, C. P.; Zheng, W., A new homogeneous high-throughput screening assay for profiling compound activity on the human ether-a-go-go-related gene channel. *Anal Biochem* **2009**, *394*, 30-38.
34. Drugs@FDA: FDA-Approved Drugs. https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm (accessed December 27).

35. Lehmann, D. F.; Eggleston, W. D.; Wang, D., Validation and Clinical Utility of the hERG IC50:Cmax Ratio to Determine the Risk of Drug-Induced Torsades de Pointes: A Meta-Analysis. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* **2018**, *38*, 341-348.
36. Danishuddin; Khan, A. U., Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today* **2016**, *21*, 1291-1302.
37. Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B., Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538-1546.
38. RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/.
39. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **2018**, *4*, 268-276.
40. Winter, R.; Montanari, F.; Noe, F.; Clevert, D. A., Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci* **2019**, *10*, 1692-1701.
41. Hochreiter, S.; Schmidhuber, J., Long short-term memory. *Neural Comput* **1997**, *9*, 1735-1780.
42. Greff, K.; Srivastava, R. K.; Koutník, J.; Steunebrink, B. R.; Schmidhuber, J., LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems* **2017**, *28*, 2222-2232.
43. Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O., Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics* **2019**, *11*, 71.
44. Noel, O. B.; Andrew, D., *DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures*. 2018.
45. Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P., Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Central Science* **2018**, *4*, 120-131.
46. Karpov, P. V.; Osolodkin, D. I.; Baskin, II; Palyulin, V. A.; Zefirov, N. S., One-class classification as a novel method of ligand-based virtual screening: the case of glycogen synthase kinase 3beta inhibitors. *Bioorg Med Chem Lett* **2011**, *21*, 6728-6731.
47. Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C., GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling* **2019**, *59*, 1096-1108.
48. scikit-learn: Machine Learning in Python. https://scikit-learn.org/stable/.
49. Keras: The Python Deep Learning library. https://keras.io/.
50. TensorFlow. https://www.tensorflow.org/.
51. DeepChem. https://deepchem.io/.
52. Tin Kam, H., The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1998**, *20*, 832-844.
53. Ji, X.; Tong, W.; Liu, Z.; Shi, T., Five-Feature Model for Developing the Classifier for Synergistic vs. Antagonistic Drug Combinations Built by XGBoost. *Frontiers in Genetics* **2019**, *10*, 600.
54. Huuskonen, J.; Salo, M.; Taskinen, J., Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J Chem Inf Comput Sci* **1998**, *38*, 450-456.
55. Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S., On the use of neural network ensembles in QSAR and QSPR. *J Chem Inf Comput Sci* **2002**, *42*, 903-911.
56. Mosier, P. D.; Jurs, P. C., QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. *J Chem Inf Comput Sci* **2002**, *42*, 1460-1470.
57. Winkler, D. A., Neural networks as robust tools in drug lead discovery and development. *Mol Biotechnol* **2004**, *27*, 139-168.
58. Basheer, I. A.; Hajmeer, M., Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* **2000**, *43*, 3-31.

59. Bahdanau, D.; Cho, K.; Bengio, Y. J. C., Neural Machine Translation by Jointly Learning to Align and Translate. **2014**, *abs/1409.0473*.
60. Luong, T.; Pham, H.; Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation. In EMNLP, 2015; 2015.
61. Keras-Self-Attention. https://github.com/CyberZHG/keras-self-attention (accessed December 4).
62. StarDrop. https://www.optibrium.com/stardrop/ (accessed December 17).
63. Braga, R. C.; Alves, V. M.; Silva, M. F.; Muratov, E.; Fourches, D.; Liao, L. M.; Tropsha, A.; Andrade, C. H., Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Mol Inform* **2015**, *34*, 698-701.
64. Pred-hERG 4.2. http://predherg.labmol.com.br/ (accessed December 17).
65. 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.
66. Angermueller, C.; Parnamaa, T.; Parts, L.; Stegle, O., Deep learning for computational biology. *Mol Syst Biol* **2016**, *12*, 878.
67. Wenzel, J.; Matter, H.; Schmidt, F., Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J Chem Inf Model* **2019**, *59*, 1253-1268.
68. Irwin, J. J.; Shoichet, B. K., ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **2005**, *45*, 177-182.
69. Bjerrum, E. J. J. A., SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. **2017**, *abs/1703.07076*.
70. Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P., Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent Sci* **2018**, *4*, 120-131.
71. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H., Molecular de-novo design through deep reinforcement learning.
72. Gupta, A.; Muller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G., Generative Recurrent Networks for De Novo Drug Design. *Mol Inform* **2018**, *37*.
73. Popova, M.; Isayev, O.; Tropsha, A., Deep reinforcement learning for de novo drug design. *Sci Adv* **2018**, *4*, eaap7885.
74. Sattarov, B.; Baskin, II; Horvath, D.; Marcou, G.; Bjerrum, E. J.; Varnek, A., De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J Chem Inf Model* **2019**, *59*, 1182-1196.
75. Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A., Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology* **2019**, *37*, 1038-1040.
76. Gao, W.; Coley, C. W., The Synthesizability of Molecules Proposed by Generative Models. *arXiv:2002.07007v1 [q-bio.QM]* **2020**.