Supporting Information

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, Noel T. Southall, Anton Simeonov 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

S1. Summary of autoencoder (AE) model.

Batch size: 256

Epochs: 5

Average loss (training): 0.075

Reconstruction rate: 80.2% (based on 1000 compounds)

S2. Summary of adversarial autoencoder (AAE) model.

Batch size: 256

Epochs: 5

Average loss (training): 0.078

Reconstruction rate: 94.0% (based on 1000 compounds)

S3. Results of hyperparameter optimization for DNN model based on training data.

Hyperparameter optimization (or grid search) was performed in two steps. The parameters investigated in Round 1 include: activation function, batch size, number of epochs and the learning rate of the optimizer. In Round 2, different dense layer architectures (i.e. dense candidates) were tested. The optimal hyperparameters that were employed in cross-validation and external validation for different descriptors are provided below:

RDKit:

```
    Round 1 {'activation': relu,
                'batch_size': 32,
                'dense_layer_sizes': [200, 100],
                'epochs': 20,
                'learn_rate': 0.0005}
```

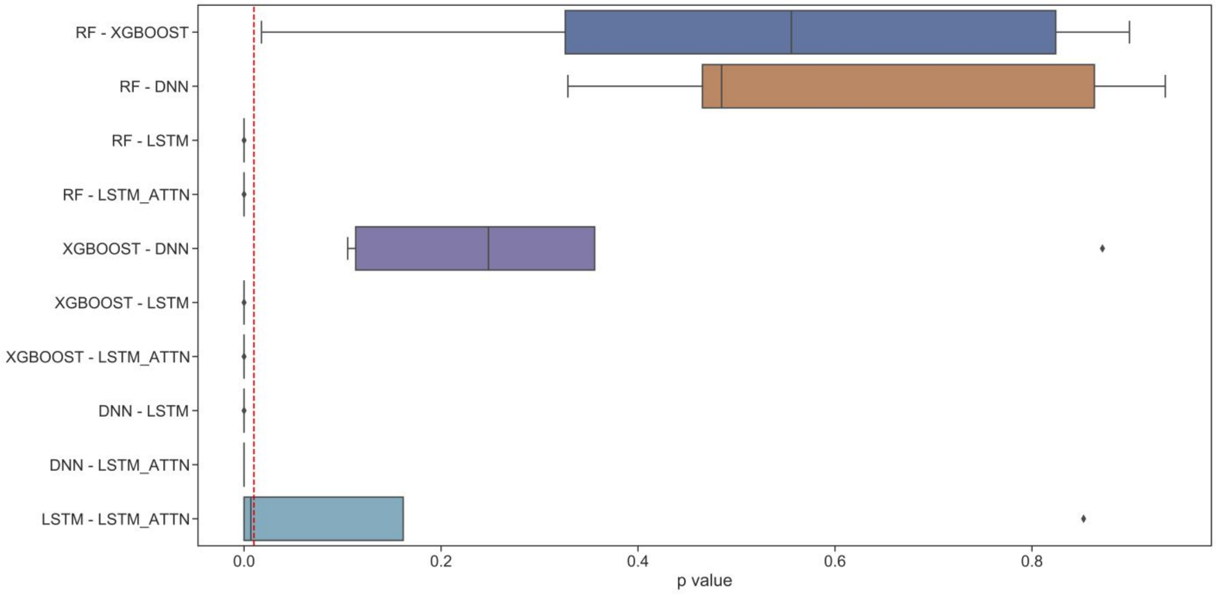
```
     Round 2 {dense_candidates = [300, 200, 100, 50, 1]}

MorganFP:

     Round 1 {'activation': relu,
                     'batch_size': 128,
                     'dense_layer_sizes': [700, 500],
                     'epochs': 30,
                     'learn_rate': 0.00001}

     Round 2 {dense_candidates = [2000, 2000, 1000, 700, 1]}

Latent1:

     Round 1 {'activation': relu,
                     'batch_size': 32,
                     'dense_layer_sizes': [700, 500],
                     'epochs': 30,
                     'learn_rate': 0.00005}

     Round 2 {dense_candidates = [1000, 700, 500, 300, 1]}

Latent2:

     Round 1 {'activation': relu,
                     'batch_size': 32,
                     'dense_layer_sizes': [700, 500],
                     'epochs': 30,
                     'learn_rate': 0.00005}

     Round 2 {dense_candidates = [1000, 700, 500, 1]}
```

S4. Five-fold cross-validation results for training data partitioned using scaffold split.

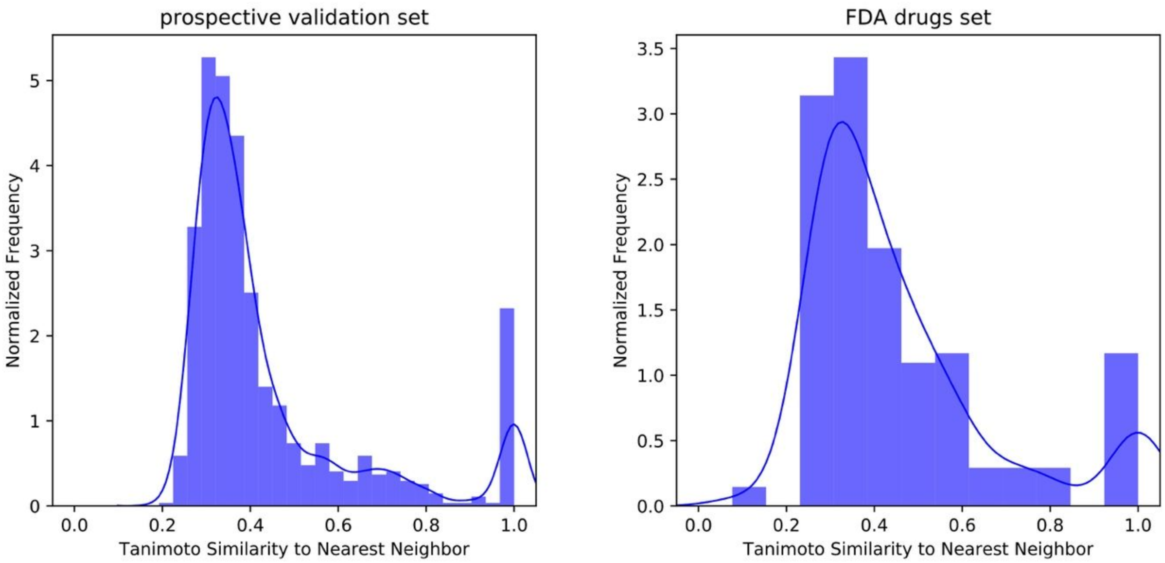
Model	Descriptor	AUC-ROC	BACC	Specificity	Sensitivity
RF	RDKit	0.90 +/- 0.01	0.80 +/- 0.02	0.66 +/- 0.04	0.94 +/- 0.01
	Morgan FP	0.90 +/- 0.01	0.78 +/- 0.02	0.58 +/- 0.02	0.97 +/- 0.01
	Latent AE	0.86 +/- 0.03	0.68 +/- 0.02	0.39 +/- 0.04	0.97 +/- 0.01
	Latent AAE	0.86 +/- 0.02	0.70 +/- 0.02	0.42 +/- 0.04	0.97 +/- 0.01
XGBoost	RDKit	0.89 +/- 0.01	0.79 +/- 0.01	0.67 +/- 0.02	0.92 +/- 0.01
	Morgan FP	0.87 +/- 0.01	0.75 +/- 0.01	0.56 +/- 0.03	0.94 +/- 0.01
	Latent AE	0.83 +/- 0.02	0.71 +/- 0.02	0.51 +/- 0.04	0.91 +/- 0.01
	Latent AAE	0.85 +/- 0.01	0.72 +/- 0.02	0.53 +/- 0.04	0.92 +/- 0.01
FF-DNN	RDKit	0.87 +/- 0.01	0.77 +/- 0.01	0.81 +/- 0.09	0.72 +/- 0.10
	Morgan FP	0.88 +/- 0.01	0.79 +/- 0.01	0.70 +/- 0.03	0.89 +/- 0.02
	Latent AE	0.86 +/- 0.02	0.77 +/- 0.02	0.69 +/- 0.09	0.86 +/- 0.07
	Latent AAE	0.87 +/- 0.01	0.77 +/- 0.01	0.64 +/- 0.04	0.89 +/- 0.03
LSTM	SMILES	0.83 +/- 0.01	0.75 +/- 0.01	0.82 +/- 0.04	0.69 +/- 0.03
LSTM-ATN	SMILES	0.84 +/- 0.01	0.76 +/- 0.01	0.79 +/- 0.05	0.73 +/- 0.05

S5. Statistical analysis for comparing the individual models developed as part of cross-validation. A total of five models (RF-RDKIT, XGBOOST-RDKIT, DNN-MORGANFP, LSTM-SMILES and LSTM_ATTN-SMILES) were selected for statistical analysis since these were the best developed individual models for each method using different descriptors. In order to compare a given pair of models, we resorted to McNemar's Test which acts as a pairwise version of

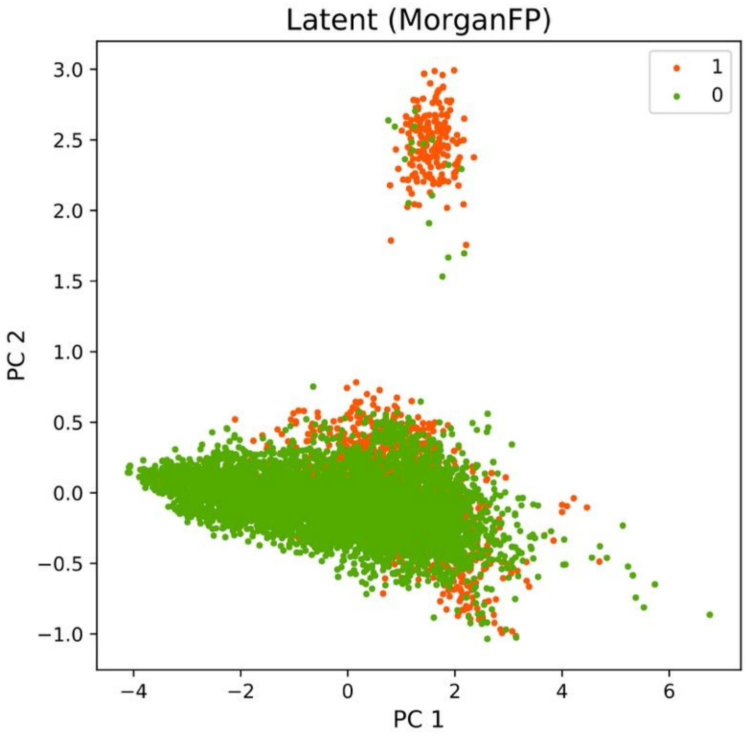
chi-squared test. A chi-square statistic is calculated that is transformed into a p-value. We performed a pairwise analysis for all five models which resulted in 10 model pairs. For each model pair, the distribution of p-values for the five folds of cross-validation is presented in the box plot. The threshold for significance was adjusted by employing Bonferroni correction (significance threshold = 0.01) and is shown in the box plot as the red dashed line.



S6. Distribution of similarity of validation set and approved drugs set towards training set. A majority of compounds from both sets are below a Tanimoto (T_c) threshold of 0.6. Those compounds that were found to be identical (T_c = 1.0) were closely examined and it was found that a majority of these are either stereo analogues or have opposite stereo configurations which could not be accounted in the 2D descriptors used to measure similarity.



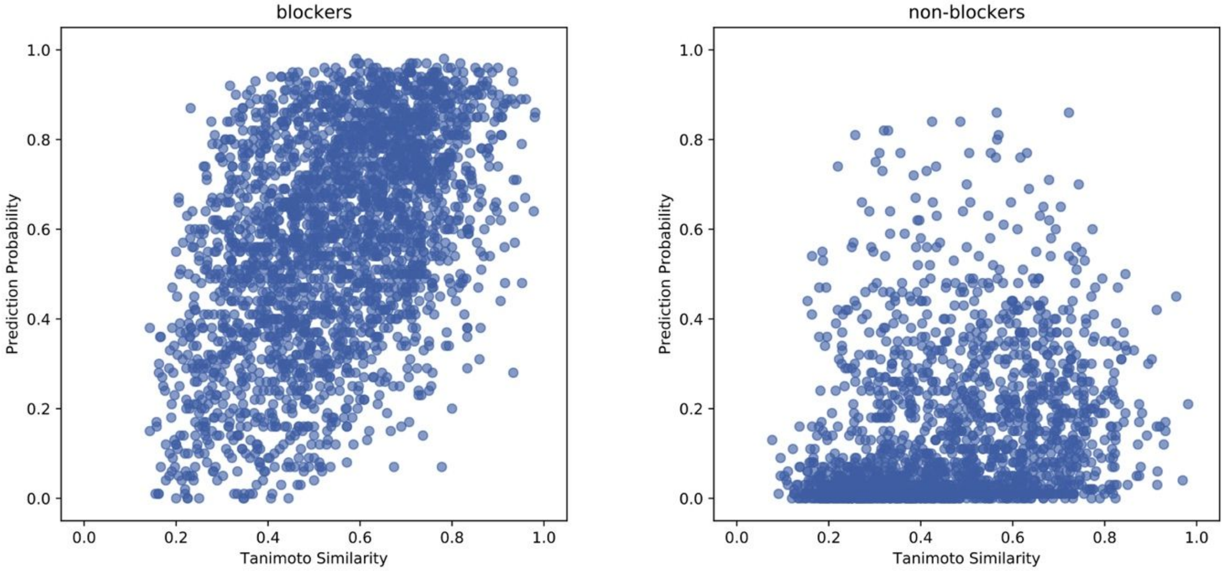
S7. PCA plot using the latent descriptors derived from AE model based on MorganFP.



S8. Performance of autoencoder (AE) derived latent descriptors from different sources (Canonical SMILES and molecular fingerprints) in external validation.

Classifier	Latent Descriptor Source	AUC-ROC	BACC	Sensitivity	Specificity
RF	Canonical SMILES	0.76	0.64	0.32	0.97
RF	MorganFP	0.70	0.64	0.28	0.99
XGBoost	Canonical SMILES	0.78	0.69	0.49	0.88
XGBoost	MorganFP	0.76	0.67	0.43	0.91
FF-DNN	Canonical SMILES	0.78	0.73	0.74	0.72
FF-DNN	MorganFP	0.78	0.70	0.60	0.79

S9. Correlation of hERG activity and similarity towards the training set for the newly generated compounds.


blockers
non-blockers
Prediction Probability
Tanimoto Similarity