Published in final edited form as:
Mol Pharm. 2017 December 04; 14(12): 4462–4475. doi:10.1021/acs.molpharmaceut.7b00578.

Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Datasets

Alexandru Korotcov[†], **Valery Tkachenko**[†,*], **Daniel P Russo**[‡,$], and **Sean Ekins**[‡,*]

[†]Science Data Software, LLC, 14914 Bradwill Court, Rockville, MD 20850, USA

[‡]Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Lab 3510, Raleigh, NC 27606, USA

[$]The Rutgers Center for Computational and Integrative Biology, Camden, NJ, 08102, USA

Abstract

Machine learning methods have been applied to many datasets in pharmaceutical research for several decades. The relative ease and availability of fingerprint type molecular descriptors paired with Bayesian methods resulted in the widespread use of this approach for a diverse array of endpoints relevant to drug discovery. Deep learning is the latest machine learning algorithm attracting attention for many of pharmaceutical applications from docking to virtual screening. Deep learning is based on an artificial neural network with multiple hidden layers and has found considerable traction for many artificial intelligence applications. We have previously suggested the need for a comparison of different machine learning methods with deep learning across an array of varying datasets that is applicable to pharmaceutical research. Endpoints relevant to pharmaceutical research include absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) properties, as well as activity against pathogens and drug discovery datasets. In this study, we have used datasets for solubility, probe-likeness, hERG, KCNQ1, bubonic plague, Chagas, tuberculosis and malaria to compare different machine learning methods using FCFP6 fingerprints. These datasets represent whole cell screens, individual proteins, physicochemical properties as well as a dataset with a complex endpoint. Our aim was to assess whether deep learning offered any improvement in testing when assessed using an array of metrics including AUC, F1 score, Cohen's kappa, Matthews correlation coefficient and others. Based on ranked normalized scores for the metrics or datasets Deep Neural Networks (DNN) ranked higher than SVM, which in turn was ranked higher than all the other machine learning methods. Visualizing these properties for training and test sets using radar type plots indicates when models are inferior or perhaps over trained. These results also suggest the need for assessing deep learning further

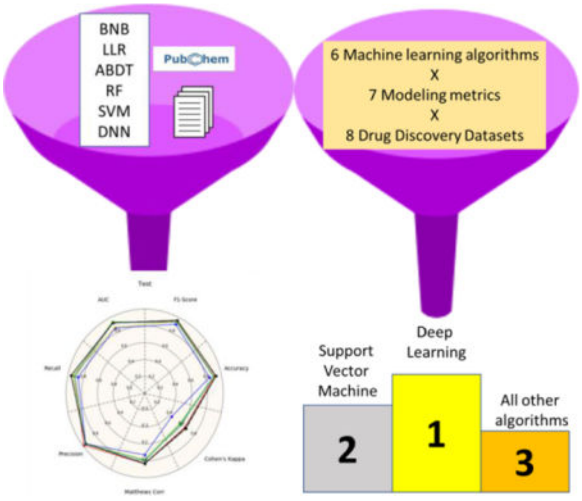
[*]Authors to whom correspondence should be addressed: Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Lab 3510, Raleigh, NC 27606, USA. E-mail address: collaborationspharma@gmail.com, Phone: +1 215-687-1320, Twitter: @collabchem; Science Data Software, LLC, 14914 Bradwill Court, Rockville, MD 20850, USA, tkachenko.valery@gmail.com.

Competing interests:
S.E. works for Collaborations Pharmaceuticals, Inc. and Phoenix Nest, Inc.

SUPPORTING INFORMATION
Supporting further details on the models, model metrics and an example of a Jupyter notebook are available https://github.com/scidatasoft/paper_ml_cmp. The training sets used in the models have been previously described by us and are available elsewhere[62]. This material is available free of charge via the Internet at http://pubs.acs.org.

using multiple metrics with much larger scale comparisons, prospective testing as well as assessment of different fingerprints and DNN architectures beyond those used.

Graphical abstract



Keywords

Deep Learning; Drug Discovery; Machine learning; Pharmaceutics; Support Vector Machine

INTRODUCTION

Drug discovery is now at the point where the increasing amount of public data in PubChem[1, 2], ChEMBL[3], and a growing list of other databases created from high throughput screens and high throughput biology in general, (including whole cell phenotypic screens, enzymes, receptors etc.) has placed it squarely in the realm of 'Big data'[4]. We are faced with a significant challenge. No longer can we be limited to a small number of molecules and their properties, we now have thousands of molecules and scores of properties to consider. How do we mine this data, use it and hope to learn from it such that we can make drug discovery more efficient and successful?

One approach which we propose is to use machine learning which can deal with this big data used in cheminformatics with methods such as support vector machines (SVM)[5–11], k-Nearest Neighbors (kNN)[12], Naïve Bayesian[13–17], Decision Trees[18] and others[19] which have been increasingly used[4, 20–22]. These methods can be used for binary classification, multiple classes classification or values prediction.

In recent years, deep artificial neural networks (including convolutional and recurrent networks) have won numerous contests in pattern recognition and machine learning[23–25]. Deep learning solves the central problem in representation learning by introducing representations that are expressed in terms of other, simpler representations. N-layer neural networks are shown in Figure 1. It is worth noting that a single-layer neural network describes a network with no hidden layers where the input is directly mapped to the output layer. In that sense, the logistic regression or SVMs are simply a special case of single-layer Neural Networks. In our work for simplicity of Deep Neural Networks (DNN) representation, we will be counting hidden layers only. Quite often 1–2 hidden layers NN are called shallow neural networks and 3 or more hidden layers NN, are called deep neural networks. A recent review addressed the development and application of deep learning in pharmaceutical research[26], a method that has proven very successful in learning images and languages elsewhere[27]. Previously deep learning has been used mostly for unsupervised learning and noisy data[28–34]. Limited efforts to use deep learning for pharmaceutical applications suggest a need for further exploration to access its utility for cheminformatics compared with other methods[35]. Deep Learning has been relatively widely used for bioinformatics[36] and computational biology[37]. Deep Learning has also been used to predict properties such as aqueous solubility using four published datasets and was shown to compare favorably to other machine learning methods using 10-fold cross validation[33]. Merck have performed a comparison of deep neural networks to date and have in turn compared them to random forests for use with large quantitative structure activity relationships (QSAR) datasets. They showed they out-performed random forests for 11 out of 15 datasets and 13 of 15 datasets in a second evaluation using time-split test sets[38]. Merck did not however look at other machine learning methods. One of the largest examples of validation of Deep learning models alongside other machine learning approaches is in the case of the Tox21 Challenge. Deep learning with multitask learning[39] slightly outperformed the closest consensus ANN method[40] across nuclear receptor and stress response datasets. Most recently, one group has suggested some datasets for molecular machine learning and used these for comparison with selected machine learning methods[41]. A second group has assessed several machine learning methods with 7 ChEMBL datasets but only focused on a single metric to assess performance[42]. Very frequently deep learning is applied to a single dataset in isolation and not compared to many of the available alternative methods. There are likely many more datasets that could benefit from deep learning even though they may be smaller[43].

These machine learning methods are increasingly used for virtual screening of compounds which can enable more efficient utilization of high throughput screening (HTS) resources, by enriching the set of compounds screened with compounds that are active[44–47]. In addition, such machine learning approaches can also be used for absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) properties as these factors can impact the success of drug discovery processes and their early assessment can later prevent failure[48–54]. Big pharma companies have access to these computational technologies but it is generally expensive to produce the data and models in house using commercial software as well as run the models and screening on project teams. Ideally, what if we could augment these resources such that a single program manager or chemist could use the models to prioritize the compounds they made and tested in a combined workflow to find “hits” that can then be reconfirmed and optimized[55]. Past collaborations[56, 57] suggested these computational approaches could greatly impact drug discovery efficiency.

Over the last decade we and others[15–17, 58–61] have increasingly focused on Bayesian approaches because of their ease of use and general applicability[56, 62–78] using molecular function class fingerprints[79, 80] of maximum diameter 6 and several other simple descriptors[81, 82]. Much of this work was centered on models for *Mycobacterium tuberculosis*[83–85] taking account of cytotoxicity and prospectively evaluating them to show high hit rates compared to random screening[85–87]. We have since followed this with datasets for Chagas disease[88] and Ebola[89] to repurpose approved drugs as well as model ADME properties such as aqueous solubility, mouse liver microsomal stability[90], Caco-2 cell permeability[62], toxicology datasets[91] and transporters[66, 92–97]. By making the fingerprints[98], and Bayesian model building algorithm open source[21, 62] there is the potential to further expand on this work.

The main aim of this study was to evaluate whether deep learning offered any improvement in testing when assessed using an array of metrics over other computational methods for drug discovery and ADME/Tox datasets. In the process, we have developed an approach to making Deep learning models more accessible.

EXPERIMENTAL SECTION

Computing

All computing was done on a single dual-processor, quad-core (Intel E5640) server running CentOS 7 with 96GB memory and two Tesla K20c GPU. The following software modules were installed: nltk 3.2.2, scikit-learn 0.18.1, Python 3.5.2, Anaconda 4.2.0 (64-bit), Keras 1.2.1, Tensorflow 0.12.1, Jupyter Notebook 4.3.1.

Datasets and Descriptors

Diverse drug discovery datasets that are publicly available for different type of activity prediction were used to develop prediction pipelines (Table 1). The same datasets have been used in Clark *et al.*,[62] for exploring applicability of an array of Bayesian models for ADME/Tox and other physicochemical properties prediction. In the current study the FCFP6 fingerprints, 1024 bins datasets were computed from SDF files using RDKit (http://www.rdkit.org/). A typical frequency of fingerprints occurrence in the 1024 bins compounds representation in a dataset shown in Figure 2.

Machine learning

Two general prediction pipelines were developed. The first pipeline solely built using only classic Machine Learning (CML) methods, such as Bernoulli Naive Bayes, Linear Logistic Regression, AdaBoost Decision Tree, Random Forest, and Support Vector Machine. Open source Scikit-learn (http://scikit-learn.org/stable/, CPU for training and prediction) ML python library was used for building, tuning, and validating all CML models included in this pipeline. The second pipeline was built using Deep Neural Networks (DNN) learning models of different complexity using Keras (https://keras.io/), a deep learning library, and Tensorflow (www.tensorflow.org, GPU training and CPU for prediction) as a backend. The developed pipeline consists of a random splitting of the input dataset into training (80%) and test (20%) datasets, while maintaining equal proportions of active to inactive class ratios in each split (stratified splitting). Thus, all the models' tuning and hyper parameters search were conducted solely through 4-fold cross validation on training data for better model generalization. An example Jupyter notebook is provided (Supplemental information 1).

Bernoulli Naive Bayes

The Naive Bayes method is a supervised learning algorithm based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Bernoulli Naive Bayes implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable.

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods and have been widely used[102]. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one-dimensional distribution. On the other hand, although naive Bayes is known as a decent classifier, it is known to be not a very good estimator, so the class probability outputs are not very accurate.

Our Bernoulli Naive Bayes (BNB) models were tuned and trained using *BernoulliNB()* method from the Naïve Bayes module of Scikit-learn. The 4-fold stratified cross-validation with a non-parametric approach based on isotonic regression for balancing classes (most of datasets are heavily imbalanced) have been used. The cross-validation generator estimates the model parameter on the train portions of cross-validation split for each split, and the calibration is done on the test cross-validation split of the train dataset, then the probabilities predicted for the folds are then averaged. AUC, F1-score and other metrics listed in data analysis section were computed using those probabilities.

Linear Logistic Regression with regularization

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution, thus predicting the probability of particular outcomes. The L2 binominal regularized logistic regression method was used to classify the activities. A stochastic average gradient optimizer was used in the *LogisticRegressionCV()* method from the Linear Module of Scikit-learn. A 4-fold stratified cross-validation method was used in grid search of the best regularization parameter (L2 penalties were in logarithmic scale between $1e^{-5}$ and $1e^{-1}$). The AUC of ROC was used for scoring the classification (maximizing AUC) performance for each fold of balanced classes' classification task.

AdaBoost Decision Tree

AdaBoost is a type of "Ensemble Learning" where multiple learners are employed to build a stronger learning algorithm by conjugating many week classifiers. The Decision Tree (DT) was chosen as a base algorithm in our implementation of the AdaBoost method (ABDT). *AdaBoostClassifier()* method with 100 estimators and 0.9 learning rate from Scikit-learn ensemble methods was used. Similarly to Naïve Bayes, the ABDT models were tuned using isotonic calibration for the imbalanced classes with 4-fold stratified cross-validation method.

Random Forest

Random forest (RF) method is another ensemble method, which fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The *RandomForestClassifier()* method with maximum depth of tree 5 and balanced classes weights were was used to build the model. The 4-fold stratified cross-validation grid search was done using 5, 10, 25, and 50 estimators with the AUC of ROC as a scoring function of the estimator.

Support Vector Machine

Support Vector Machine (SVM) is one of the most popular supervised machine learning algorithms mostly used in classification problems and it is quite effective in high dimensional spaces[7]. The learning of the hyperplane in the SVM algorithm can be done using different kernel functions for the decision function. C SVM classification with libsvm implementation method from Scikit-learn was used (*svm.SVC()*). The 4-fold stratified cross-validation grid search using weighted classes was done for 2 kernels (linear, rbf), C (1, 10, 100), and gamma values (1e−2, 1e−3, 1e−4). The parameter C, common to all SVM kernels, trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly. Gamma defines how much influence a single training example has. The larger gamma is, the closer other examples must be to be affected. Our implementation of SVM automatically finds the best parameters and save the best SVM model for activity predictions.

Deep Neural Networks

Two basic approaches to avoid DNN model overfitting are used in training including L2 norm and drop out regularization for all hidden layers. The following hyperparameters optimization was performed using a 3 hidden layers DNN (Keras with Tensorflow backend) and the grid search method from Scikit-learn. The following parameters were optimized prior to final model training: optimization algorithm (*SGD*, *Adam*, *Nadam)*, learning rate (0.05, 0.025, 0.01, 0.001), network weight initialization (*uniform*, *lecun_uniform*, *normal*, *glorot_normal*, *he_normal*, *he_normal)*, hidden layers activation function (*relu*, *tanh*, *LeakyReLU*, *SReLU)*, output function (*softmax*, *softplus*, *sigmoid)*, L2 regularization (0.05, 0.01, 0.005, 0.001, 0.0001), dropout regularization (0.2, 0.3, 0.5, 0.8) and number of nodes all hidden layers (512, 1024, 2048, 4096).

The following hyperparameters were used for further DNN training: *SGD*, learning rate 0.01 (automatically 10% reduced on plateu of 50 epochs), weight initialization he_normal, hidden layers activation SReLU, output layer function *sigmoid*, L2 regularization 0.001, dropout is 0.5. The *binary crossentropy* was used as a loss function. In order to save training time, an early training termination were implemented by stopping training if no change in loss were observed after 200 epochs. The number of hidden nodes in all hidden layers were set equal to number of input features (number of bins in fingerprints). The DNN model performance was evaluated on up to 8 hidden layers DNNs.

Data analysis—In this study, several traditional measurements of model performance were used, including recall, precision, F1-Score, accuracy, ROC curve and the area under it (AUC), Cohen's Kappa[103, 104], and the Matthews correlation[105]. For the metric definitions, we will use the following abbreviations: the number of true positives (TP), the number of false positives (FP), the number of true negatives (TN), and the number of false negatives (FN). Model Recall (also known as the True Positive Rate or Sensitivity) can be thought of percentage of *true* class labels correctly identified by the model as *true* and is defined: $Recall = \frac{TP}{TP+FN}$. Similarly, model precision (also known as the Positive Predictive Value) is the probability a predicted *true* label is indeed *true* and is defined: $Precision = \frac{TP}{TP+FP}$. The F1-Score is simply the harmonic mean of the Recall and Precision: $F_1 Score = 2\frac{Precision \cdot Recall}{Precision + Recall}$. Accuracy is another measure of all-around model robustness, and is the percentage of correctly identified labels out of the entire population. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. The ROC curve can be computed by plotting the Recall vs the False Positive Rate (FPR) at various decision thresholds T, where $FPR = \frac{FP}{FP+TP}$. In this study, all constructed models are capable of assigning a probability estimate of a sample belonging to the *true* class. Thus, we can construct an ROC curve by measuring the Recall and FPR performance when we considered a sample with a probability estimate > T as being True for various intervals between 0 and 1. The AUC can be constructed from this plot and can be thought of as the ability of the model to separate classes, where 1 denotes perfect separation and 0.5 is random classification. Another measure of overall model classification performance is the Matthew's Correlation Coefficient (MCC), which is not subject to heavily imbalanced classes and is defined as $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$. As a measure of correlation, its value can be between −1 and 1. Cohen's Kappa (CK), another metric estimating overall model performance, attempts to leverage the Accuracy by normalizing it to the probability that the classification would agree by chance (p_e) and is calculated by $CK = \frac{Accuracy - p_e}{1 - p_e}$, where $p_e = p_{True} + p_{False}$, $p_{True} = \frac{TP+FN}{TP+TN+FP+FN} \cdot \frac{TP+FP}{TP+TN+FP+FN}$, $p_{False} = \frac{TN+FN}{TP+TN+FP+FN} \cdot \frac{TN+FP}{TP+TN+FP+FN}$.

Experimental solubility determination—For one of our drug discovery projects the measurement of the aqueous solubility of 3 test compounds in pH 7.4 PBS was performed by Bioduro (San Diego, CA) utilizing the published modified shake flask method[106, 107]. These molecules were used as an external test set for the solubility models developed in this study.

RESULTS

The AUC, F1-score, accuracy, Kappa, Matthews correlation, precision and recall values of the all trained models for compounds represented as FCFP6 fingerprints in 1024 bins have been summarized (Table 2 and Supplemental Tables 1–16) while individual model files are also provided (Supplemental Information 2). For clarity, we have grouped all the metrics by each dataset training and test set and presented them as radar plots. A perfect score on all metrics would be represented by a circle the size of the complete plot. Based on the ROC score alone, the Naïve Bayes models are comparable to those described previously by Clark *et al.*[62] (Table 2). In this case however, we are using a different source of descriptors and modeling algorithms FCFP6 vs ECFP6 and the RDKit vs CDK[108]. In many cases the test sets for the SVM models were generally better than any other models in the same pipeline.

When the radar plots are analyzed it is readily apparent which models are likely over-trained. In this case the model training sets have large scores across all the metrics for the training set and have much lower scores for the test set. The shape of the plots can also be indicative of the quality of the models. The larger the circle for the test set, the better the model. The solubility model (Figure 3) is a good example of a model that is well balanced. The training and test sets are virtually both represented by similar circular plots and it is clear that the BNB method performs worst across most of the metrics. The Probe-like model (Figure 4) has an irregular arrangement of scores for the test set which shows it performs very poorly for the Cohen's Kappa across all methods. The test set for the hERG model (Figure 5) shows that most methods are comparable across the metrics and ABDT stands out as performing the worst with the test set. Overall, the Cohen's Kappa is the most sensitive metric for this dataset. The KCNQ1 model (Figure 6) shows that DNN and SVM outperform other methods for training and testing and the Matthews correlation and Cohen's Kappa scores are dramatically lower than all other metrics. The Bubonic plague model (Figure 7) is a difficult example, DNN easily out performs all methods in training and testing (AUC, Matthews correlation and accuracy metric perform best). The Chagas disease dataset (Figure 8) again shows that the DNN has preferable training and testing performance with Cohen's kappa as the most sensitive metric. The tuberculosis dataset (Figure 9) is another example where DNN outperforms all methods on training by a large margin and by a small margin on the test set, except for the recall statistics. Precision, F1-score, and Cohen's Kappa are poor for all methods with the test set. The Malaria dataset (Figure 10) shows the impact of DNN with big improvements in precision, F1-score and Cohen's Kappa for the training and test set when compared to the other machine learning methods.

In general, the DNN models performed well for external test set predictions except for the AUC performance of the probe-like dataset. For AUC DNN-3 outperforms BNB on 6 of 8 datasets (Table 2). For F1 score the DNN-3 model outperforms BNB on 6 of 8 datasets (Supplemental Table 1). For Accuracy DNN-3 outperforms BNB on 8 datasets (Supplemental Table 2). For Kappa it outperforms BNB on 7 datasets (Supplemental Table 3). For Matthews correlation DNN-3 outperforms or equals BNB on 6 of 8 datasets (Supplemental Table 4). For precision, DNN-3 outperforms BNB on 6 of 8 datasets (Supplemental Table 5). For recall it outperforms BNB on 5 of 8 datasets (Supplemental Table 6). 4 of 8 datasets had AUC values higher for DNN-3 than SVM while 7 of the 8 datasets showed the same of greater F1 values for DNN-3. For accuracy DNN-3 was equal or better than SVM in all 8 datasets. For Kappa, Matthews correlation and recall 5 datasets were equal or better than SVM. For precision 6 datasets were equal or better than SVM.

In order to further understand which models performed the best we have used ranked normalized scores for each machine learning algorithm ranked by metric (Table 3) and by dataset (Table 4). This approach has been previously used by others[109] to compare multiple machine learning methods and performance criteria. When the models are ranked by both metric or dataset, Deep learning DNN-5 and DNN-4) ranks above SVM and all other methods are below this (Table 3 and 4).

The solubility of 3 compounds from one of our drug discovery projects was assessed using all the different solubility machine learning models developed in this study. The cut off for a soluble molecule was $LogS = -5$ (10 μM/L). The experimental solubility for the 3 compounds evaluated ranged from 80.8 μM to 465 μM. In virtually all cases the molecules were correctly predicted as soluble (Table 5).

DISCUSSION

To date there have been relatively few studies that have made comparisons of deep learning to the wide array of classical machine learning methods or have discussed this methods application in pharmaceutical research[41, 110, 111] or even used the models for actual predictions for ongoing projects. This study therefore fills a void related to drug discovery applications of these methods. For all the computational modeling approaches which we may want to consider they are dependent upon the model applicability domain[112] and are affected by the quality of the underlying data,[113, 114] which may in turn determine the utility and relevance of any model and prediction[115]. Comparisons of deep learning with other machine learning algorithms across a range of applications suggests this method frequently improves when using predominantly internal cross-validation as the form of evaluation[26]. In this study, we used an external test set in all cases for comparison of the different algorithms. In addition, we compared several metrics for assessing performance, which is unlike most published studies in this field which rely on one or a narrow range. Our approach for comparison uses a rank normalized score approach which has been used by others to compare machine learning algorithms and performance metrics in other areas[109].

It is likely that the lack of a commercial machine learning platform that is accessible for users to build, validate and test machine learning models based on their own data, is holding back smaller organizations and academic groups from using these approaches. Making these machine learning models something that can be created and used without the need for an expert in cheminformatics would be a considerable achievement. We are far from there yet but are using the Jupyter Notebook to seamlessly integrate chemical operations, dataset manipulation, and machine learning models (e.g., DL, as well as Bayesian, Trees, etc.) within one framework. Deep learning methods have not been widely assessed using prospective validation. Making such models accessible will allow us to take previously published and novel data, enable building of models, and evaluate them for internal quality, before validating them using predictions on vendor libraries, purchase and testing.

The results of this study show there is still work to be done to improve the deep learning models overall due to the differences across all the metrics for training and test set evaluation. Also it is apparent that some of the metrics are less sensitive than others. For example, AUC is far less sensitive than Cohen's Kappa. This might therefore represent a more useful test set metric than AUC. However, using AUC alone we can discern differences in the models with external testing (Table 2). At least 7 out of 8 datasets show an improvement from Bayesian to SVM and 5 out of the 8 datasets show an improvement in the Test set ROC from Bayesian to DNN-3. This represents results from a true external test set. In 8 of 8 cases the training ROC increases from Bayesian to DNN3. The rank normalized score approach indicates that DNN outperforms SVM, which in turn then outperforms all other algorithms (Table 3 and 4). These results suggest SVM and DNN should perhaps be further embraced by scientists and more rigorously evaluated in different scenarios for drug discovery.

Regardless of the individual models' shortcomings, we investigated the 'real world' applicability of one of the models, namely solubility, to a recent drug discovery project. Solubility is a physicochemical property which is usually determined for many compounds in the process of a drug discovery project and any efforts that could help accurately predict this property might ultimately help in the drug design process. Here, we use classification models and regression models for solubility to identify the predicted solubility for three in-house compounds. 8 of the 9 models (including all DNN models) correctly identified all three compounds as being soluble (Solubility > 10 μM/L). The only misclassifications came from the AdaBoost Decision Tree model, labeling the two less soluble compounds as insoluble. Notably, the DNN models all labeled the compounds as having > 99% probability of being soluble. We then evaluated regression performance on predicting solubility using a linear classifier (Elastic Net) and the DNN models. All models achieved accuracy within approximately 1 log unit of solubility, with a tendency to under-predict solubility (Supplemental Table 7). This is admittedly a very small test set but represents compounds which are advanced leads and therefore of considerable interest. Further prospective assessment of these solubility machine learning models requires much larger numbers of molecules, but this represents a starting point.

While we did not evaluate different descriptors in this study, FCFP6 does quite well with the datasets in this study. However future studies may evaluate additional descriptors such as other non-fingerprint descriptors with deep learning. A recent paper described molecular graph convolutions which represents a simpler encoding of molecules as undirected graphs of atoms for machine learning applications[116]. The development of additional descriptors and their assessment with different machine learning methods would go some way towards finding the best combination of descriptors and machine learning algorithm. We did not observe any obvious effect of dataset size or balance and this could be due to the limited number of datasets assessed.

Increasingly there are efforts to develop open platforms for connecting scientists and sharing data for many types of projects relevant to drug discovery, these will often include various prediction algorithms e.g. qsardb.org and ochem.eu[99, 100] or private sharing and development tools such as Chembench[101] and CDD Vault[62]. It should be noted that up until recently[43] the open source deep learning toolkits for cheminformatics have been relatively inaccessible to the average scientist to use for either building models or generating predictions. Facilitating making machine learning accessible to non-expert users will help in increasing the potential impact of these methods for drug discovery. Ideally these and other machine learning models need to be simple to use and transparent.

As our previous prospective testing of Bayesian models for various drug discovery projects (built with commercial software) has led to several promising leads and candidates for *in vivo* animal testing[85–87, 117, 118], a meaningful improvement in model statistics could lead to improved hit rates in prospective testing. We also propose the need for actually applying the models to real world drug discovery tasks like assessing solubility (Table 5). A key finding of this study is that a wider array of model metrics is likely essential for model comparison and reliance on AUC may not be ideal. In future, it is likely we will also use these and other metrics to evaluate how Deep Learning performs for prospective virtual screening as well as assess a much larger array of datasets. Such case studies are ongoing and will be of value to understand the scope and limitations of deep learning and other machine learning methods for drug discovery.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Alex Clark is kindly acknowledged for collaboration on the previous Bayesian studies described herein and used for benchmarking. Dr. Mary Lingerfelt is acknowledged for assistance with obtaining the solubility data.

Grant information

We acknowledge NIH funding R43GM122196 from NIGMS. The solubility data was obtained as part of R21TR001718 from NCATS.

ABBREVIATIONS

ADME/Tox	Absorption, Distribution, Metabolism, Excretion/Toxicology
ABDT	AdaBoost
ANN	artificial neural networks
AUC	area under the curve
BNB	Bernoulli Naive Bayes
CML	classic Machine Learning
DT	Decision Tree
DNN	Deep Neural Networks
hERG	human ether a-go-go related gene

HTS	high throughput screening
kNN	k-Nearest Neighbors
QSAR	quantitative structure activity relationships
RF	Random forest
ROC	receiver operating characteristic
SVM	support vector machines

References

1. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. 2016; 44(D1):D1202–13. [PubMed: 26400175]
2. Anon The PubChem Database. http://pubchem.ncbi.nlm.nih.gov/
3. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012; 40:D1100–7. Database issue. [PubMed: 21948594]
4. Ekins S, Clark AM, Swamidass SJ, Litterman N, Williams AJ. Bigger data, collaborative tools and the future of predictive drug discovery. J Comput Aided Mol Des. 2014; 28(10):997–1008. [PubMed: 24943138]
5. Bennet KP, Campbell C. Support vector machines: Hype or hallelujah? SIGKDD Explorations. 2000; 2:1–13.
6. Christianini, N., Shawe-Taylor, J. Support vector machines and other kernel-based learning methods. Cambridge University Press; Cambridge, MA: 2000.
7. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2001
8. Lei T, Chen F, Liu H, Sun H, Kang Y, Li D, Li Y, Hou T. ADMET Evaluation in Drug Discovery. Part 17: Development of Quantitative and Qualitative Prediction Models for Chemical-Induced Respiratory Toxicity. Mol Pharm. 2017; 14(7):2407–2421. [PubMed: 28595388]
9. Kriegl JM, Arnhold T, Beck B, Fox T. A support vector machine approach to classify human cytochrome P450 3A4 inhibitors. J Comput Aided Mol Des. 2005; 19(3):189–201. [PubMed: 16059671]
10. Guangli M, Yiyu C. Predicting Caco-2 permeability using support vector machine and chemistry development kit. J Pharm Pharm Sci. 2006; 9(2):210–21. [PubMed: 16959190]
11. Kortagere S, Chekmarev D, Welsh WJ, Ekins S. Hybrid scoring and classification approaches to predict human pregnane X receptor activators. Pharm Res. 2009; 26(4):1001–11. [PubMed: 19115096]
12. Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A. Development and validation of k-nearest neighbour QSPR models of metabolic stability of drug candidates. J Med Chem. 2003; 46:3013–3020. [PubMed: 12825940]
13. 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(8):2855–66. [PubMed: 27379394]
14. Li D, Chen L, Li Y, Tian S, Sun H, Hou T. ADMET evaluation in drug discovery. 13. Development of in silico prediction models for P-glycoprotein substrates. Mol Pharm. 2014; 11(3):716–26. [PubMed: 24499501]
15. Nidhi, Glick M, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model. 2006; 46(3):1124–33. [PubMed: 16711732]

16. Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, Bender A, Jenkins JL, Urban L. Modeling promiscuity based on in vitro safety pharmacology profiling data. ChemMedChem. 2007; 2(6): 874–80. [PubMed: 17492703]

17. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. ChemMedChem. 2007; 2(6):861–873. [PubMed: 17477341]

18. Susnow RG, Dixon SL. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. J Chem Inf Comput Sci. 2003; 43(4):1308–15. [PubMed: 12870924]

19. Mitchell JB. Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci. 2014; 4(5):468–481. [PubMed: 25285160]

20. Zhu H, Zhang J, Kim MT, Boison A, Sedykh A, Moran K. Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. Chem Res Toxicol. 2014; 27(10):1643–51. [PubMed: 25195622]

21. Clark AM, Ekins S. Open Source Bayesian Models: 2 Mining A "big dataset" to create and validate models with ChEMBL. J Chem Inf Model. 2015; 55:1246–1260. [PubMed: 25995041]

22. Ekins S, Freundlich JS, Reynolds RC. Are Bigger Data Sets Better for Machine Learning? Fusing Single-Point and Dual-Event Dose Response Data for Mycobacterium tuberculosis. J Chem Inf Model. 2014; 54:2157–65. [PubMed: 24968215]

23. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015; 61:85–117. [PubMed: 25462637]

24. Capuzzi SJ, Politi R, Isayev O, Farag S, Tropsha A. QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. Frontiers in Environmental Science. 2016; 4(3)

25. Russakovsky, O., deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, AC., Fei-Fei, L. ImageNet large scale visual recognition challenge. https://arxiv.org/pdf/1409.0575.pdf

26. Ekins S. The next era: Deep learning in pharmaceutical research. Pharm Res. 2016; 33:2594–603. [PubMed: 27599991]

27. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–44. [PubMed: 26017442]

28. Jha A, Gazzara MR, Barash Y. Integrative deep models for alternative splicing. Bioinformatics. 2017; 33(14):i274–i282. [PubMed: 28882000]

29. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adam M, Gertych A, Tan RS. A deep convolutional neural network model to classify heartbeats. Comput Biol Med. 2017; 89:389–396. [PubMed: 28869899]

30. Cang Z, Wei GW. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLoS Comput Biol. 2017; 13(7):e1005690. [PubMed: 28749969]

31. Javed K, Gouriveau R, Zerhouni N. A New Multivariate Approach for Prognostics Based on Extreme Learning Machine and Fuzzy Clustering. IEEE Trans Cybern. 2015; 45(12):2626–39. [PubMed: 25643420]

32. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One. 2013; 8(6):e66341. [PubMed: 23826094]

33. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inf Model. 2013; 53(7):1563–75. [PubMed: 23795551]

34. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. IEEE Trans Med Imaging. 2016; 35(5):1313–21. [PubMed: 26891484]

35. Baskin II, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery. Expert Opin Drug Discov. 2016; 11:785–795. [PubMed: 27295548]

36. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2016

37. Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016; 12(7):878. [PubMed: 27474269]

38. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. J Chem Inf Model. 2015; 55(2):263–74. [PubMed: 25635324]

39. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. Front Environ Sci. 2016; 3:80.

40. Abdelaziz A, Spahn-Langguth H, Schramm K-W, Tetko IV. Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge. Front Environ Sci. 2016; 4:2.

41. Wu, Z., Ramsundar, B., Feinberg, EN., Gomes, J., Geniesse, C., Pappu, AS., Leswing, K., Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. https://arxiv.org/ftp/arxiv/papers/1703/1703.00564.pdf

42. Koutsoukas A, Monaghan KJ, Li X, Huan J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J Cheminform. 2017; 9:42. [PubMed: 29086090]

43. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low Data Drug Discovery with One-Shot Learning. ACS Cent Sci. 2017; 3(4):283–293. [PubMed: 28470045]

44. Oprea TI, Matter H. Integrating virtual screening in lead discovery. Curr Opin Chem Biol. 2004; 8(4):349–58. [PubMed: 15288243]

45. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: applications to targets and beyond. Br J Pharmacol. 2007; 152:21–37. [PubMed: 17549046]

46. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol. 2007; 152:9–20. [PubMed: 17549047]

47. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD. Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model. 2007; 47(4):1504–19. [PubMed: 17591764]

48. Lombardo F, Obach RS, Dicapua FM, Bakken GA, Lu J, Potter DM, Gao F, Miller MD, Zhang Y. A hybrid mixture discriminant analysis-random forest computational model for the prediction of volume of distribution of drugs in human. J Med Chem. 2006; 49(7):2262–7. [PubMed: 16570922]

49. Lombardo F, Obach RS, Shalaeva MY, Gao F. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. J Med Chem. 2004; 47(5):1242–50. [PubMed: 14971904]

50. Lombardo F, Obach RS, Shalaeva MY, Gao F. Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding. J Med Chem. 2002; 45:2867–2876. [PubMed: 12061889]

51. Lombardo F, Shalaeva MY, Tupper KA, Gao F. ElogDoct: A tool for lipophilicity determination in drug discovery. 2 Basic and neutral compounds. J Med Chem. 2001; 44:2490–2497. [PubMed: 11448232]

52. Lombardo F, Blake JF, Curatolo WJ. Computation of brain-blood partitioning of organic solutes via free energy calculations. J Med Chem. 1996; 39:4750–4755. [PubMed: 8941388]

53. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Del Rev. 1997; 23:3–25.

54. Ekins S, Ring BJ, Grace J, McRobie-Belle DJ, Wrighton SA. Present and future in vitro approaches for drug metabolism. J Pharm Tox Methods. 2000; 44:313–324.

55. Tanrikulu Y, Kruger B, Proschak E. The holistic integration of virtual screening in drug discovery. Drug Discov Today. 2013; 18(7–8):358–64. [PubMed: 23340112]

56. Gupta RR, Gifford EM, Liston T, Waller CL, Bunin B, Ekins S. Using open source computational tools for predicting human metabolic stability and additional ADME/TOX properties. Drug Metab Dispos. 2010; 38:2083–2090. [PubMed: 20693417]

57. Ekins S, Gupta RR, Gifford E, Bunin BA, Waller CL. Chemical space: missing pieces in cheminformatics. Pharm Res. 2010; 27(10):2035–9. [PubMed: 20683645]

58. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, Shoichet BK, Urban L. Large-scale prediction and testing of drug activity on side-effect targets. Nature. 2012; 486(7403):361–7. [PubMed: 22722194]

59. Crisman TJ, Parker CN, Jenkins JL, Scheiber J, Thoma M, Kang ZB, Kim R, Bender A, Nettles JH, Davies JW, Glick M. Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. J Chem Inf Model. 2007; 47(4): 1319–27. [PubMed: 17608469]

60. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. J Med Chem. 2006; 49(23):6802–10. [PubMed: 17154510]

61. Jenkins JL, Glick M, Davies JW. A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. J Med Chem. 2004; 47(25):6144–59. [PubMed: 15566286]

62. Clark AM, Dole K, Coulon-Spector A, McNutt A, Grass G, Freundlich JS, Reynolds RC, Ekins S. Open source bayesian models: 1. Application to ADME/Tox and drug discovery datasets. J Chem Inf Model. 2015; 55:1231–1245. [PubMed: 25994950]

63. Kortagere S, Ekins S. Troubleshooting computational methods in drug discovery. J Pharmacol Toxicol Methods. 2010; 61(2):67–75. [PubMed: 20176118]

64. Ekins S, Williams AJ. Precompetitive Preclinical ADME/Tox Data: Set It Free on The Web to Facilitate Computational Model Building to Assist Drug Development. Lab on a Chip. 2010; 10:13–22. [PubMed: 20024044]

65. Ekins S, Honeycutt JD, Metz JT. Evolving molecules using multi-objective optimization: applying to ADME. Drug Discov Today. 2010; 15:451–460. [PubMed: 20438859]

66. Bahadduri PM, Polli JE, Swaan PW, Ekins S. Targeting drug transporters - combining in silico and in vitro approaches to predict in vivo. Methods Mol Biol. 2010; 637:65–103. [PubMed: 20419430]

67. Ekins S, Bugrim A, Brovold L, Kirillov E, Nikolsky Y, Rakhmatulin E, Sorokina S, Ryabov A, Serebryiskaya T, Melnikov A, Metz J, Nikolskaya T. Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms. Xenobiotica. 2006; 36(10–11):877–901. [PubMed: 17118913]

68. Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, Sorokina S, Bugrim A, Nikolskaya T. A Combined Approach to Drug Metabolism and Toxicity Assessment. Drug Metab Dispos. 2006; 34:495–503. [PubMed: 16381662]

69. Ekins S. Systems-ADME/Tox: resources and network approaches. J Pharmacol Toxicol Methods. 2006; 53(1):38–66. [PubMed: 16054403]

70. Chang, C., Ekins, S. Pharmacophores for human ADME/Tox-related proteins. In: Langer, T., Hoffman, RD., editors. Pharmacophores and pharmacophore searches. Wiley-VCH; Weinheim: 2006. p. 299-324.

71. Ekins S, Nikolsky Y, Nikolskaya T. Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. Trends Pharmacol Sci. 2005; 26(4):202–9. [PubMed: 15808345]

72. Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, Bugrim A, Nikolskaya T. Computational prediction of human drug metabolism. Expert Opin Drug Metab Toxicol. 2005; 1(2):303–24. [PubMed: 16922645]

73. Balakin KV, Ivanenkov YA, Savchuk NP, Ivaschenko AA, Ekins S. Comprehensive computational assessment of ADME properties using mapping techniques. Curr Drug Disc Tech. 2005; 2:99–113.

74. Ekins S, Swaan PW. Computational models for enzymes, transporters, channels and receptors relevant to ADME/TOX. Rev Comp Chem. 2004; 20:333–415.

75. Ekins S, Boulanger B, Swaan PW, Hupcey MA. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. Mol Divers. 2002; 5(4):255–75. [PubMed: 12549676]

76. Ekins S, Wrighton SA. Application of in silico approaches to predicting drug–drug interactions. J Pharmacol Toxicol Methods. 2001; 45(1):65–9. [PubMed: 11489666]

77. Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA, Wikel JH. Progress in predicting human ADME parameters in silico. J Pharmacol Toxicol Methods. 2000; 44(1):251–72. [PubMed: 11274894]

78. Ekins S, Ring BJ, Grace J, McRobie-Belle DJ, Wrighton SA. Present and future in vitro approaches for drug metabolism. J Pharmacol Toxicol Methods. 2000; 44(1):313–24. [PubMed: 11274898]

79. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010; 50(5):742–54. [PubMed: 20426451]

80. Rogers D, Brown RD, Hahn M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. J Biomol Screen. 2005; 10(7):682–6. [PubMed: 16170046]

81. Ekins S, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Hohman M, Bunin B. A Collaborative Database And Computational Models For Tuberculosis Drug Discovery. Mol BioSystems. 2010; 6:840–851.

82. Ekins S, Kaneko T, Lipinksi CA, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Ernst S, Yang J, Goncharoff N, Hohman M, Bunin B. Analysis and hit filtering of a very large library of compounds screened against Mycobacterium tuberculosis. Mol BioSyst. 2010; 6:2316–2324. [PubMed: 20835433]

83. Ananthan S, Faaleolea ER, Goldman RC, Hobrath JV, Kwong CD, Laughon BE, Maddry JA, Mehta A, Rasmussen L, Reynolds RC, Secrist JA 3rd, Shindo N, Showe DN, Sosa MI, Suling WJ, White EL. High-throughput screening for inhibitors of Mycobacterium tuberculosis H37Rv. Tuberculosis. 2009; 89(5):334–53. [PubMed: 19758845]

84. Maddry JA, Ananthan S, Goldman RC, Hobrath JV, Kwong CD, Maddox C, Rasmussen L, Reynolds RC, Secrist JA 3rd, Sosa MI, White EL, Zhang W. Antituberculosis activity of the molecular libraries screening center network library. Tuberculosis. 2009; 89(5):354–63. [PubMed: 19783214]

85. Ekins S, Reynolds R, Kim H, Koo M-S, Ekonomidis M, Talaue M, Paget SD, Woolhiser LK, Lenaerts AJ, Bunin BA, Connell N, Freundlich JS. Bayesian Models Leveraging Bioactivity and Cytotoxicity Information for Drug Discovery. Chem Biol. 2013; 20:370–378. [PubMed: 23521795]

86. Ekins S, Reynolds RC, Franzblau SG, Wan B, Freundlich JS, Bunin B. A Enhancing Hit Identification in Mycobacterium tuberculosis Drug Discovery Using Validated Dual-Event Bayesian Models. PLOSONE. 2013; 8:e63240.

87. Ekins S, Casey AC, Roberts D, Parish T, Bunin BA. Bayesian models for screening and TB Mobile for target inference with Mycobacterium tuberculosis. Tuberculosis (Edinb). 2014; 94(2):162–9. [PubMed: 24440548]

88. Ekins S, Lage de Siqueira-Neto J, McCall LI, Sarker M, Yadav M, Ponder EL, Kallel EA, Kellar D, Chen S, Arkin M, Bunin BA, McKerrow JH, Talcott C. Machine Learning Models and Pathway Genome Data Base for Trypanosoma cruzi Drug Discovery. PLoS Negl Trop Dis. 2015; 9(6):e0003878. [PubMed: 26114876]

89. Ekins S, Freundlich J, Clark A, Anantpadma M, Davey R, Madrid P. Machine learning models identify molecules active against Ebola virus in vitro. F1000Res. 2015; 4:1091. [PubMed: 26834994]

90. Perryman AL, Stratton TP, Ekins S, Freundlich JS. Predicting mouse liver microsomal stability with “pruned’ machine learning models and public data. Pharm Res. 2015; 33:433–449. [PubMed: 26415647]

91. Ekins S. Progress in computational toxicology. J Pharmacol Toxicol Methods. 2014; 69(2):115–40. [PubMed: 24361690]

92. Dong Z, Ekins S, Polli JE. Quantitative NTCP pharmacophore and lack of association between DILI and NTCP Inhibition. Eur J Pharm Sci. 2014; 66C:1–9.

93. Dong Z, Ekins S, Polli JE. Structure-activity relationship for FDA approved drugs as inhibitors of the human sodium taurocholate cotransporting polypeptide (NTCP). Mol Pharm. 2013; 10(3): 1008–19. [PubMed: 23339484]

94. Ekins S, Diao L, Polli JE. A Substrate Pharmacophore for the Human Organic Cation/Carnitine Transporter Identifies Compounds Associated with Rhabdomyolysis. Mol Pharm. 2012; 9:905–913. [PubMed: 22339151]

95. Diao L, Ekins S, Polli JE. Quantitative Structure Activity Relationship for Inhibition of Human Organic Cation/Carnitine Transporter. Mol Pharm. 2010; 7:2120–2130. [PubMed: 20831193]

96. Zheng X, Ekins S, Raufman JP, Polli JE. Computational models for drug inhibition of the human apical sodium-dependent bile acid transporter. Mol Pharm. 2009; 6(5):1591–603. [PubMed: 19673539]

97. Diao L, Ekins S, Polli JE. Novel Inhibitors of Human Organic Cation/Carnitine Transporter (hOCTN2) via Computational Modeling and In Vitro Testing. Pharm Res. 2009; 26:1890–1900. [PubMed: 19437106]

98. Clark AM, Sarker M, Ekins S. New target predictions and visualization tools incorporating open source molecular fingerprints for TB Mobile 2.0. J Cheminform. 2014; 6:38. [PubMed: 25302078]

99. Aruoja V, Moosus M, Kahru A, Sihtmae M, Maran U. Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga Pseudokirchneriella subcapitata. Chemosphere. 2014; 96:23–32. [PubMed: 23895738]

100. Sushko I, Novotarskyi S, Korner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des. 2011; 25(6):533–54. [PubMed: 21660515]

101. Walker T, Grulke CM, Pozefsky D, Tropsha A. Chembench: a cheminformatics workbench. Bioinformatics. 2010; 26(23):3000–1. [PubMed: 20889496]

102. Xia X, Maliski EG, Gallant P, Rogers D. Classification of kinase inhibitors using a Bayesian model. J Med Chem. 2004; 47(18):4463–70. [PubMed: 15317458]

103. Carletta J. Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics. 1996; 22:249–254.

104. Cohen J. A coefficient of agreement for nominal scales. Education and Psychological Measurement. 1960; 20:37–46.

105. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405(2):442–51. [PubMed: 1180967]

106. Glomme A, Marz J, Dressman JB. Comparison of a miniaturized shake-flask solubility method with automated potentiometric acid/base titrations and calculated solubilities. J Pharm Sci. 2005; 94(1):1–16. [PubMed: 15761925]

107. Bergstrom CA, Norinder U, Luthman K, Artursson P. Experimental and computational screening models for prediction of aqueous drug solubility. Pharm Res. 2002; 19(2):182–8. [PubMed: 11885560]

108. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J Chem Inf Comput Sci. 2003; 43(2):493–500. [PubMed: 12653513]

109. Caruana, R., Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. 23rd International Conference on Machine Learning; Pittsburgh, PA. 2006; Pittsburgh, PA:

110. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. Mol Pharm. 2016; 13(5):1445–54. [PubMed: 27007977]

111. Gawehn E, Hiss JA, Schneider G. Deep Learning in Drug Discovery. Mol Inform. 2016; 35(1):3–14. [PubMed: 27491648]

112. Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI. Can we estimate the accuracy of ADME-Tox predictions? Drug Discov Today. 2006; 11(15–16):700–7. [PubMed: 16846797]

113. 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(7):1189–204. [PubMed: 20572635]

114. Williams AJ, Ekins S, Tkachenko V. Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation. Drug Disc Today. 2012; 17:685–701.

115. Vracko M, Bandelj V, Barbieri P, Benfenati E, Chaudhry Q, Cronin M, Devillers J, Gallegos A, Gini G, Gramatica P, Helma C, Mazzatorta P, Neagu D, Netzeva T, Pavan M, Patlewicz G, Randic M, Tsakovska I, Worth A. Validation of counter propagation neural network models for predictive toxicology according to the OECD principles: a case study. SAR QSAR Environ Res. 2006; 17(3):265–84. [PubMed: 16815767]

116. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. 2016; 30(8):595–608. [PubMed: 27558503]

117. Ekins S, Freundlich JS, Clark AM, Anantpadma M, Davey RA, P M. Machine learning models identify molecules active against the Ebola virus in vitro. F1000Res. 2016; 4:1091.

118. Ekins S, de Siqueira-Neto JL, McCall LI, Sarker M, Yadav M, Ponder EL, Kallel EA, Kellar D, Chen S, Arkin M, Bunin BA, McKerrow JH, Talcott C. Machine Learning Models and Pathway Genome Data Base for Trypanosoma cruzi Drug Discovery. PLoS Negl Trop Dis. 2015; 9(6):e0003878. [PubMed: 26114876]

119. Huuskonen J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. J Chem Inf Comput Sci. 2000; 40(3):773–7. [PubMed: 10850781]

120. Litterman N, Lipinski CA, Bunin BA, Ekins S. Computational Prediction and Validation of an Expert's Evaluation of Chemical Probes. J Chem Inf Model. 2014; 54:2996–3004. [PubMed: 25244007]

121. Wang S, Li Y, Wang J, Chen L, Zhang L, Yu H, Hou T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. Mol Pharm. 2012; 9(4):996–1010. [PubMed: 22380484]

122. Du F, Yu H, Zou B, Babcock J, Long S, Li M. hERGCentral: a large database to store, retrieve, and analyze compound-human Ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. Assay Drug Dev Technol. 2011; 9(6):580–8. [PubMed: 22149888]

123. Guiguemde WA, Shelat AA, Bouck D, Duffy S, Crowther GJ, Davis PH, Smithson DC, Connelly M, Clark J, Zhu F, Jimenez-Diaz MB, Martinez MS, Wilson EB, Tripathi AK, Gut J, Sharlow ER, Bathurst I, El Mazouni F, Fowble JW, Forquer I, McGinley PL, Castro S, Angulo-Barturen I, Ferrer S, Rosenthal PJ, Derisi JL, Sullivan DJ, Lazo JS, Roos DS, Riscoe MK, Phillips MA, Rathod PK, Van Voorhis WC, Avery VM, Guy RK. Chemical genetics of Plasmodium falciparum. Nature. 2010; 465(7296):311–5. [PubMed: 20485428]

124. Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, Hasan S, Brown JR, Peishoff CE, Cardon LR, Garcia-Bustos JF. Thousands of chemical starting points for antimalarial lead identification. Nature. 2010; 465:305–310. [PubMed: 20485427]

125. Gagaring, K., Borboa, R., Francek, C., Chen, Z., Buenviaje, J., Plouffe, D., Winzeler, E., Brinker, A., Diagena, T., Taylor, J., Glynne, R., Chatterjee, A., Kuhen, K. Novartis-GNF Malaria Box. ChEMBL-NTD (www.ebi.ac.uk/chemblntd)

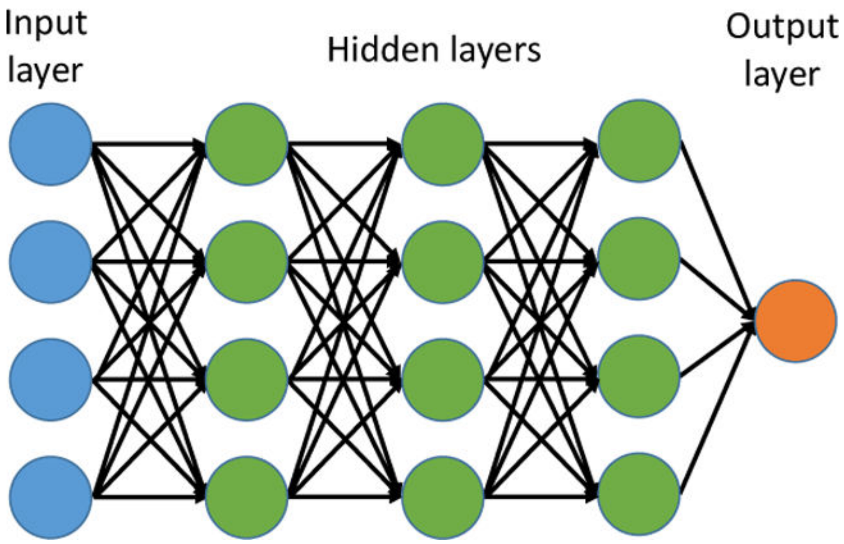


Figure 1.

A 4-layer neural network with four inputs, two hidden layers of 4 neurons each and one output layer. Notice that connections are between neurons across layers, but not within a layer.

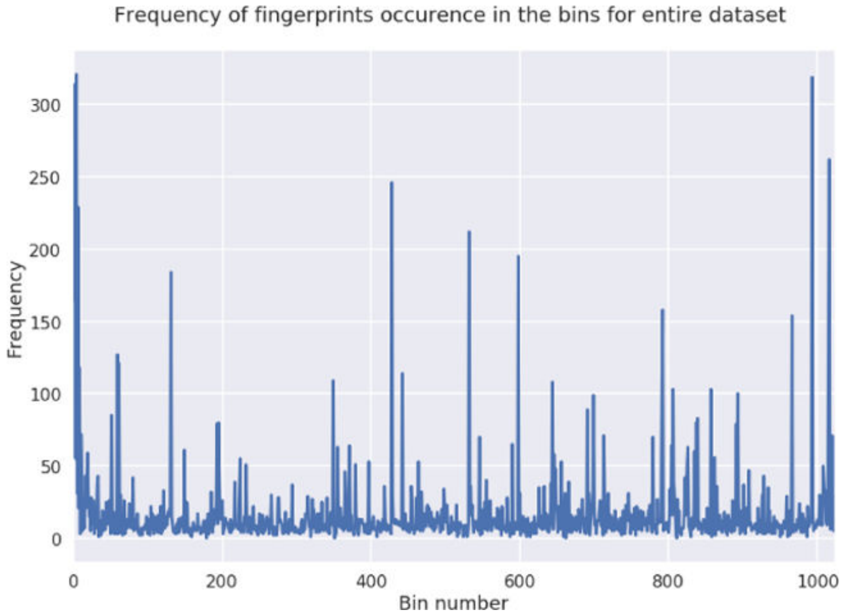


Figure 2.
Typical frequency of fingerprints occurrence in the 1024 bins compounds representation in a dataset.

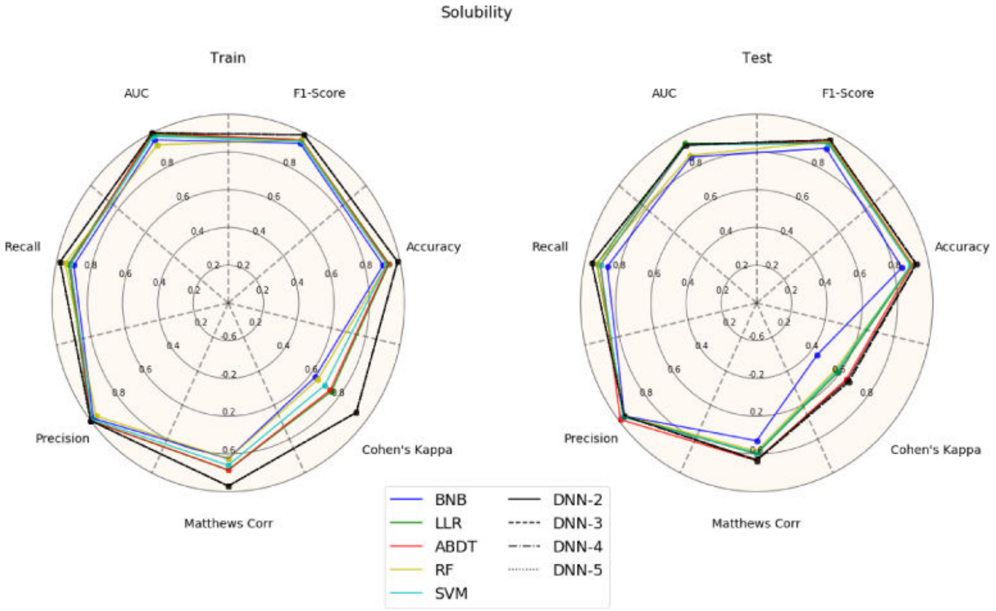


Figure 3.

Radar plot for the Solubility training and testing data.

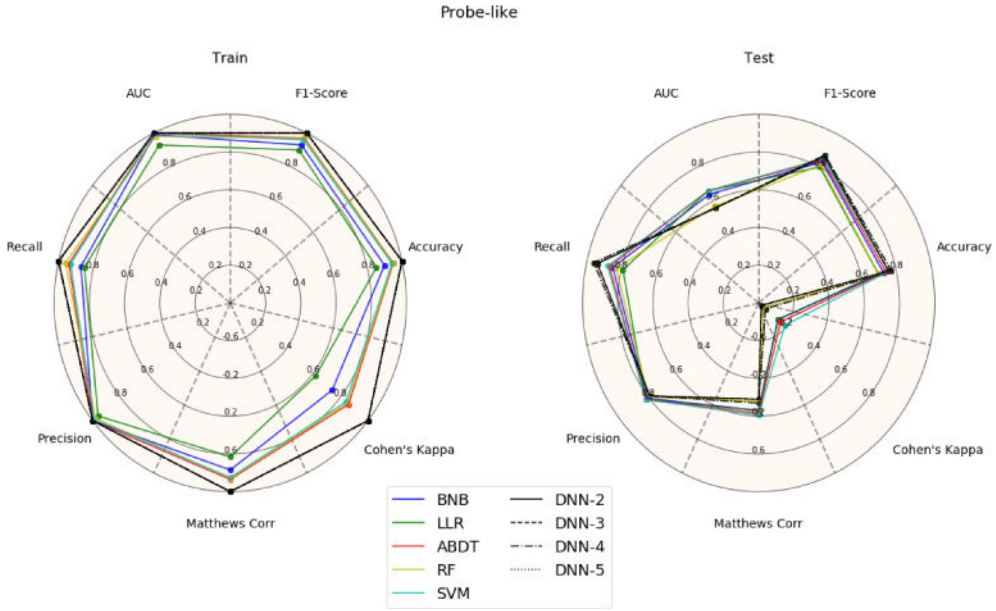


Figure 4.

Radar plot for the Probe-like training and testing data.

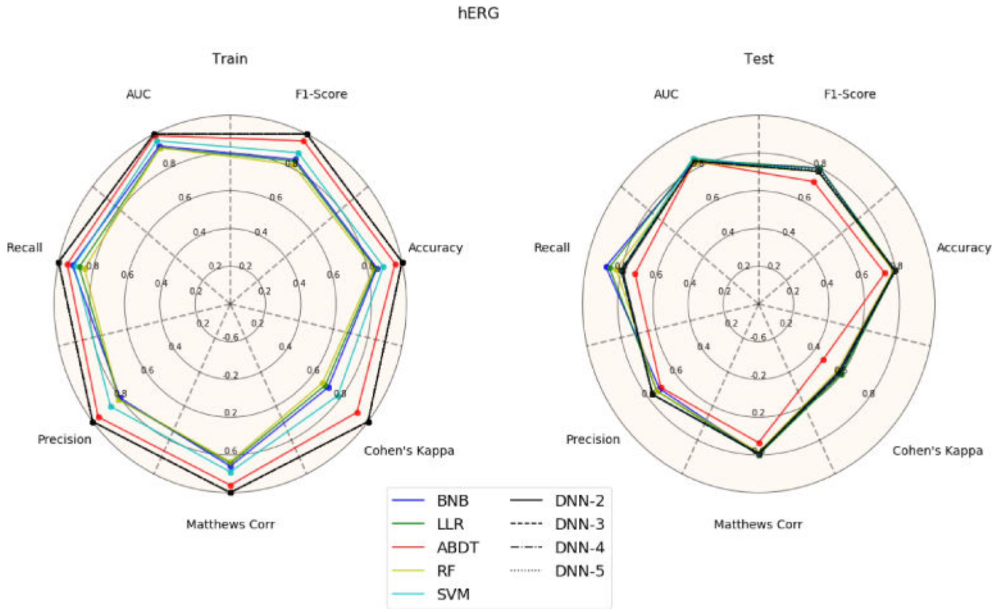


Figure 5.

Radar plot for hERG training and testing data.

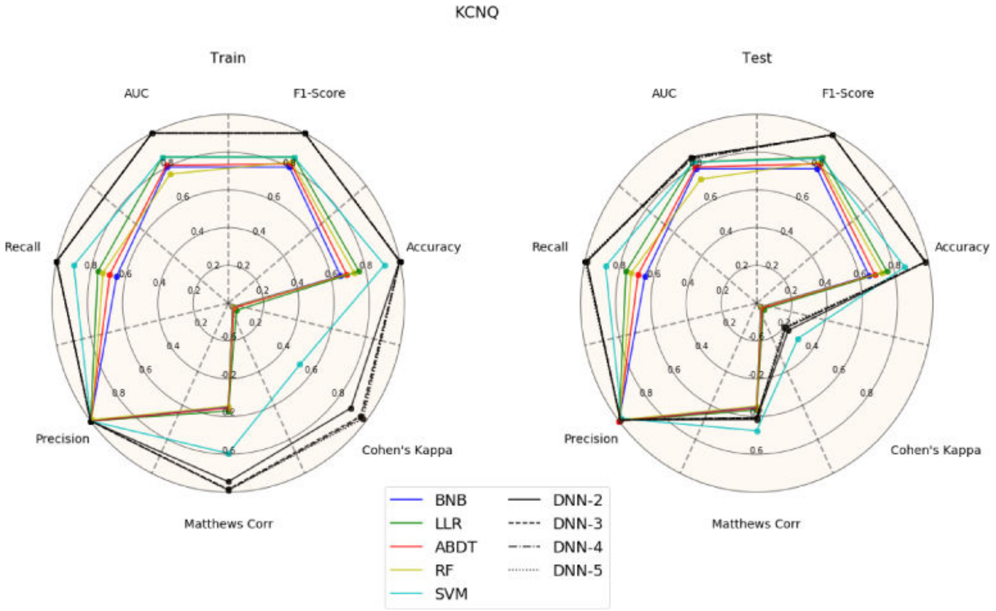


Figure 6.

Radar plot for KCNQ1 training and testing data.

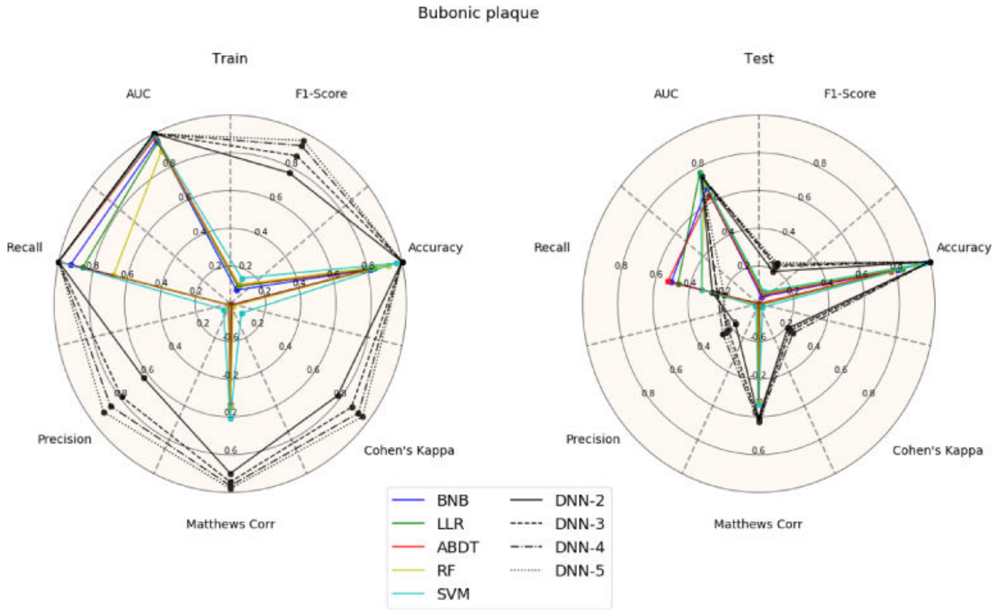


Figure 7.

Radar plot for Bubonic plague training and testing data.

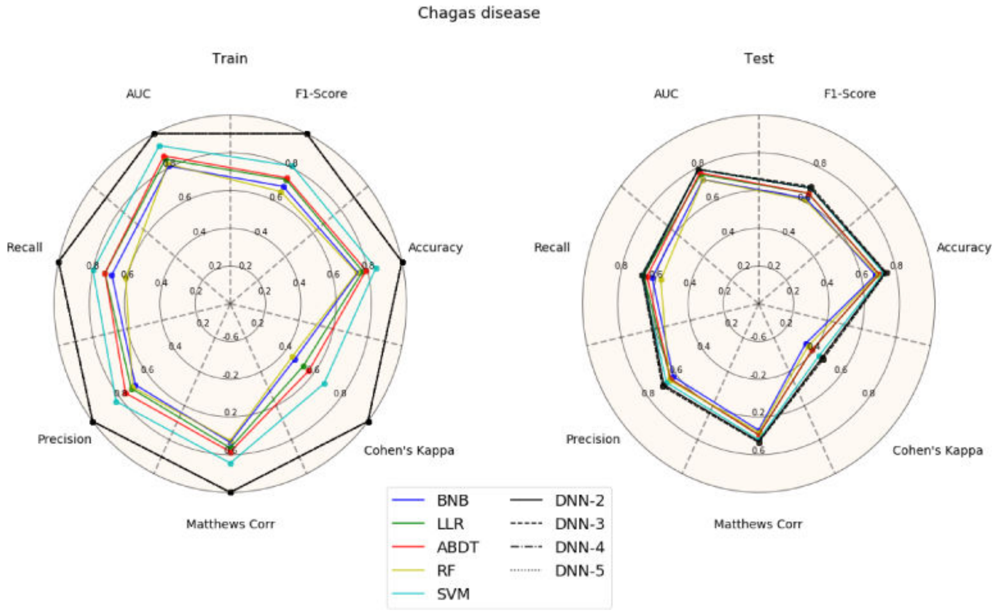


Figure 8.

Radar plot for Chagas training and testing data.

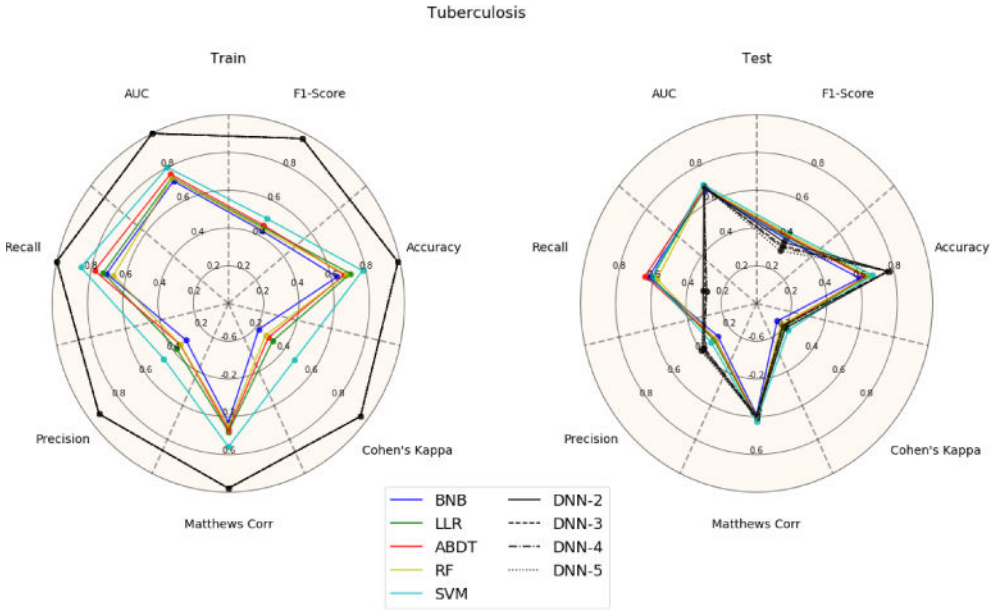


Figure 9.

Radar plot for tuberculosis training and testing data.

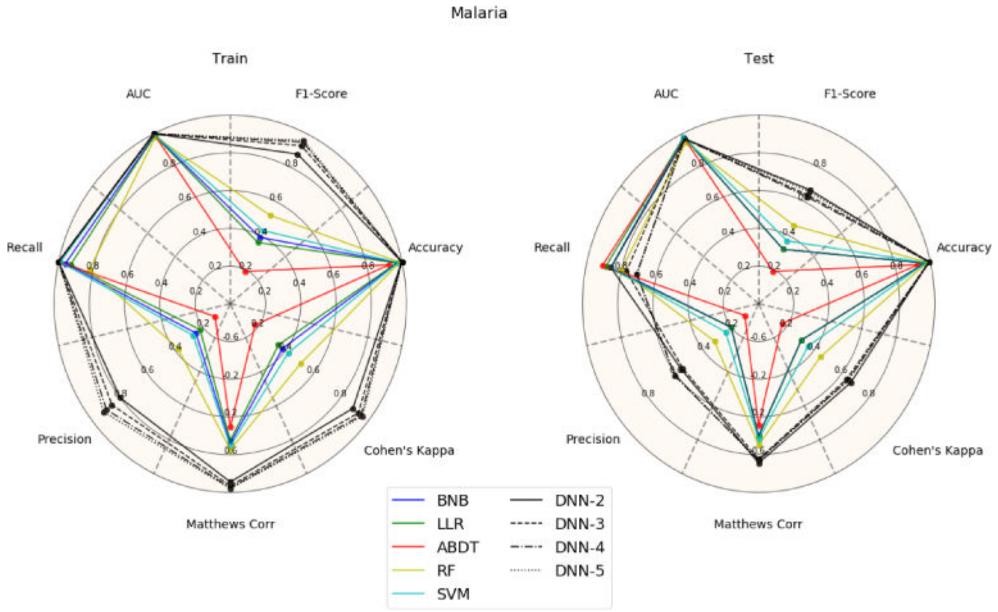


Figure 10.

Radar plot for Malaria training and testing data.

Table 1

Binary classification datasets used for evaluating multiple computational methods for activity prediction. Note, that some datasets have heavily imbalanced (This table was adapted from Clark et al, Table #2). Note the active / inactive ratios for hERG and KCNQ1 are reversed as we are trying to obtain compounds that are more desirable (active = non-inhibitors). Models for solubility and hERG are important for drug discovery because they represent an important physicochemical property and a toxicological target, respectively. The probe-like model represents small drug-like molecules that have been scored for probe-likeness by a medicinal chemist. KCNQ1 represents a very large dataset from a screen which could be relevant to a large target-based high throughput screening campaign. Currently this dataset is on the upper end of the size scale in the public domain for drug discovery. The remaining 4 datasets (bubonic plague, Chagas disease, Tuberculosis and malaria) represent whole cell phenotypic screens of various sizes and with differing ratios of active to inactive compounds.

Model	**Datasets used and references**	**Cutoff for active**	**Number of molecules and ratio**
solubility	[119]	Log solubility = −5	1144 active, 155 inactive, ratio 7.38
probe-like	[120]	described in[120]	253 active, 69 inactive, ratio 3.67
hERG	[121]	described in ref[121]	373 active, 433 inactive, ratio 0.86
KCNQ1	PubChem BioAssay: AID 2642[122]	using actives assigned in PubChem	301,737 active, 3878 inactive, ratio 77.81
Bubonic plague (*Yersina pestis*)	PubChem single-point screen BioAssay: AID 898	active when inhibition ≥50%	223 active, 139, 710 inactive, ratio 0.0016
Chagas disease (*Typanosoma cruzi*)	Pubchem BioAssay: AID 2044	with EC_{50} <1 μM, >10-fold difference in cytotoxicity as active as described in[88]	1692 active, 2363 inactive, ratio 0.72
TB (*Mycobacterium tuberculosis*)	*in vitro* bioactivity and cytotoxicity data from MLSMR, CB2, kinase, and ARRA datasets[22]	*Mtb* activity and acceptable Vero cell cytotoxicity selectivity index = (MIC or IC_{90})/CC_{50} ≥10	1434 active, 5789 inactive, ratio 0.25
malaria (*Plasmodium falciparum*)	CDD Public datasets (MMV, St. Jude, Novartis, and TCAMS)[123–125]	3D7 EC_{50} <10 nM	175 active, 19,604 inactive, ratio 0.0089

Table 2

AUC values for all machine learning models and tested compounds (ECFP6, bin=1024) implemented in our pipelines.

Model	BNB	LLR	ABDT	RF	SVM	DNN-2	DNN-3	DNN-4	DNN-5	Published Naïve Bayesian 5 fold cross validation with ECFP6[62]
solubility train	0.96	0.99	1.00	0.93	0.98	1.00	1.00	1.00	1.00	0.87
solubility test	0.86	0.94	0.93	0.87	0.93	0.93	0.93	0.93	0.93	
probe-like train	0.99	0.93	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.76
probe-like test	0.64	0.66	0.66	0.57	0.66	0.56	0.56	0.56	0.56	
hERG train	0.93	0.92	0.99	0.92	0.96	1.00	1.00	1.00	1.00	0.85
hERG test	0.84	0.85	0.84	0.83	0.86	0.84	0.84	0.84	0.84	
KCNQ1 train	0.80	0.86	0.81	0.76	0.86	1.00	1.00	1.00	1.00	0.84
KCNQ1 test	0.79	0.83	0.80	0.73	0.83	0.86	0.86	0.85	0.85	
Bubonic plague train	0.96	0.95	0.98	0.90	0.99	1.00	1.00	1.00	1.00	0.81
Bubonic plague test	0.68	0.77	0.64	0.71	0.76	0.75	0.75	0.75	0.75	
Chagas disease train	0.81	0.85	0.87	0.82	0.93	1.00	1.00	1.00	1.00	0.80
Chagas disease test	0.73	0.76	0.77	0.73	0.79	0.79	0.79	0.79	0.79	
Tuberculosis train	0.72	0.74	0.76	0.74	0.80	1.00	1.00	1.00	1.00	0.73
Tuberculosis test	0.67	0.68	0.68	0.68	0.70	0.69	0.68	0.69	0.69	
Malaria train	0.99	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.98
Malaria test	0.98	0.98	0.97	0.95	0.98	0.97	0.97	0.97	0.97	

BNB - Bernoulli Naive Bayes, LLR - Logistic linear regression, ABDT - AdaBoost Decision Trees, RF - Random Forest, SVM - Support Vector Machines, DNN-N - DNN with X hidden layers. For comparison, we added previously published 5-fold cross validation ROC data.

Table 3

Ranked normalized scores for each machine learning algorithm by metric (average over eight datasets for the test dataset).

MODEL	AUC	F1-score	ACC	Cohen-Kappa	Matthews	Precision	Recall	MEAN	RANK
DNN-5	0.798	0.688	**0.874**	0.382	0.387	**0.683**	0.700	**0.6448**	1
DNN-4	0.799	0.684	0.872	**0.383**	**0.387**	0.682	0.705	0.6446	2
DNN-2	0.800	0.685	0.868	0.378	0.382	0.664	0.723	0.6431	3
DNN-3	0.799	**0.689**	0.871	0.378	0.381	0.670	0.711	0.6428	4
SVM	**0.813**	0.620	0.832	0.341	0.366	0.599	0.744	0.6164	5
LLR	0.809	0.604	0.789	0.286	0.316	0.584	**0.746**	0.5907	6
RF	0.760	0.612	0.795	0.277	0.297	0.594	0.700	0.5765	7
ABDT	0.786	0.575	0.763	0.260	0.294	0.570	0.741	0.5698	8
BNB	0.774	0.587	0.760	0.251	0.282	0.574	0.733	0.5659	9

BNB - Bernoulli Naive Bayes, LLR - Logistic linear regression, ABDT - AdaBoost Decision Trees, RF - Random Forest, SVM - Support Vector Machines, DNN-N - DNN with X hidden layers. For comparison, we added previously published 5-fold cross validation ROC data.

Table 4

Ranked normalized scores of each machine learning algorithm by datasets (averaged over seven metrics for test datasets).

MODEL	Solubility	Probe-like	hERG	KCNQ	Bubonic	Chagas	Tuberculosis	Malaria	MEAN	RANK
DNN-5	0.870	0.582	0.740	0.746	0.408	0.649	0.412	0.752	**0.6448**	1
DNN-4	**0.870**	0.582	0.739	0.748	**0.428**	0.640	0.405	0.747	0.6446	2
DNN-2	0.865	0.564	0.731	**0.753**	0.401	0.640	0.418	**0.773**	0.6431	3
DNN-3	0.870	0.566	0.739	0.749	0.414	**0.649**	0.409	0.746	0.6428	4
SVM	0.831	**0.624**	0.742	0.721	0.320	0.633	**0.456**	0.604	0.6164	5
LLR	0.839	0.578	**0.750**	0.626	0.311	0.601	0.437	0.583	0.5907	6
RF	0.823	0.541	0.729	0.590	0.296	0.566	0.430	0.637	0.5765	7
ABDT	0.857	0.608	0.665	0.592	0.290	0.597	0.440	0.509	0.5698	8
BNB	0.763	0.590	0.747	0.574	0.298	0.564	0.413	0.578	0.5659	9

Table 5

Observed and predicted solubility for 3 novel external test compounds as part of a drug discovery project (Solubility class probabilities in parenthesis).

Compound	BNB	LLR	ABDT	RF	SVM	DNN-2	DNN-3	DNN-4	DNN-5	Experimental
1	Soluble (0.886)	Soluble (0.799)	Insoluble (0.348)	Soluble (0.622)	Soluble (0.930)	Soluble (0.999)	Soluble (0.999)	Soluble (0.999)	Soluble (0.999)	168 μM at pH 7.4
2	Soluble (0.799)	Soluble (0.709)	Insoluble (0.154)	Soluble (0.540)	Soluble (0.926)	Soluble (0.998)	Soluble (0.998)	Soluble (0.999)	Soluble (0.999)	80.8 μM at pH 7.4
3	Soluble (0.799)	Soluble (0.782)	Soluble (0.590)	Soluble (0.590)	Soluble (0.973)	Soluble (0.996)	Soluble (0.998)	Soluble (0.998)	Soluble (0.998)	465 μM at pH 7.4