Evaluation of machine learning approaches for building models to select quality candidate molecules in drug discovery.

Joseph T. Scavetta

Department of Computer Science, Rowan University, Glassboro, NJ, 08028, USA.

Abstract

Drug discovery can become a time-consuming and expensive process, with average development time exceeding a decade. Many methods have been developed to optimize drug candidate selection, such as rule-based methods for selecting compounds with values within set bounds on select properties. Machine learning techniques have also been employed in drug discovery, from predicting the values of drug properties, to classifying a compound as a drug or as not a drug. In this study, multiple machine learning methods are evaluated on their ability to classify approved drugs from non-approved drugs using physiochemical properties available on the public ChEMBL database. Further, these approaches are evaluated on their ability to classify approved drugs still on the market from drugs that have been withdrawn. The approaches evaluated and compared include logistic regression, random forest, support vector machine, and neural network. Drug-like properties, from those available in the ChEMBL database, are also evaluated to determine which are most important in explaining drug variance.

Introduction

1. Drug Discovery and Why Drugs Fail

Drug discovery has historically relied on many iterations of compound synthesis, *in vitro* and *in vivo* assays, and manual evaluation of the results. This overall process typically becomes time-consuming and expensive, with average development time exceeding a decade. Because of this, decisions into which drugs the development team should evaluate further, drop, or rework is a critical part of drug development. Making these decisions relies on many conditions, notably, a drug's performance on many assays. Often, those making these decisions will screen for a drug's absorption, distribution, metabolism, excretion, and toxicity data (ADMET) [1].

For a drug to become FDA approved, it must enter and succeed clinical trials. In the Phase I trials, most drugs fail due toxicity. In Phase II and Phase III trials, drugs often fail due to efficacy problems, though, toxicity still plays a large role in drug failure within these phases as well [2]. Because of these two major components of failure, chemical properties that relate to a compound's toxicity and absorption are important in determining a successful drug. These chemical features can be referred to as drug-like properties.

2. Rule-Based Drug Optimization

While the desired characteristics of drug-like properties can vary depending on the drug's target and route of transmission, there has been much research into defining common generalizations. These act as general guidelines as to which physiochemical properties, and their values, we should expect from a successful drug, compared to a less effective drug candidate. The well-known rule set, Lipinski's rule of 5, is often noted when discussion which properties make a compound drug-like. This rule set declares upper bounds on molecular properties such as hydrogen bound donor and acceptors, molecular mass, and lipophilicity [3]. While drugs that satisfy Lipinski's rule of 5 are often more successful than those that do not, many exceptions can occur where the model would disqualify a successful drug and vice-versa [4]. This is because the rule-set focuses on the absorption properties of a drug, only one of many important components. Further, they are primarily descriptive of permeability potential; solubility and dosage may also play a role in absorption. Also, the bounds apply to oral drugs that do not act as substrates for naturally occurring transporters. While there are limitations, Lipinski's rule of 5 acts as a useful starting point in selecting important drug-like properties to act as features in future drug-prediction models, specifically absorption models.

3. Extensions of the Rule-Based Methods

Further rule-based methods have been created in extension to Lipinski's rule of 5, which further include rotatable bonds, polar surface area, and the number of sp^3 hybridized carbon bonds in the considered bounds. While the rule-based methods provide straightforward methods for determining drug success, they are limited in substantial ways. Having a strict cutoff implies these properties are discrete in their effects, rather than continuous. Misjudging this assumption can result in many missed opportunities. Further, these rules are often generated only from properties that successful drugs have in common, however, if the distribution of these properties is the same for non-drugs as well, then these properties have no value in drug determination. Without comparing drugs and non-drugs, differences in property distribution cannot be known [5].

To replace cutoffs with a continuous scale, Bickerton et al. created the quantitative estimate of drug-likeness (QED) [6]. This performs well, however, it still does not consider whether a property is truly predictive, i.e. has a different distribution from non-drugs. To consider both issues, the relative drug likelihood (RDL) can be computed, which relies on a positive set and a negative set of data. These data sets are dependent on the question at hand. RDL performs better than QED, as it uses distributions from the positive and negative set to determine which properties have a high relative likelihood of distinguishing a drug [5].

4. Applying Machine Learning to Drug Discovery

Many models and commercial software have been developed to predict various ADMET properties for drug candidates [1]. This, including many other medicinal chemistry topics have begun to improve with the use of machine learning techniques [7]. For example, drugs

typically have a solubility, denoted as log S, ranging from -1 to -5. This is important for achieving adequate solubility in aqueous solution while staying partially hydrophobic to pass through cell membranes. However, finding the solubility of a compound is difficult. Rather than direct measurements, machine learning methods have been used, which can achieve predictions within a rms error of 0.8 log unit, performing as well as the uncertainty from experimental log S measures, averaging 0.6 log unit [8].

Machine learning techniques have also been applied to ligand-based visual screening. In other words, machine learning algorithms have been used to determine which drugs match some query the most out of a database of potential compounds. Support vector machines (SVMs) have been used in various ligand-based visual screening tasks. Notably, they have had success in distinguishing drugs from non-drugs. This is achieved by projecting compounds into high dimensional space so that they may become linearly separable. Random forest (RF) is another technique often employed in ligand-based visual screening. RF uses an ensemble of decision trees, where each tree is created from a random sample of training data. One advantage is that RF can handle imbalanced classes, which can be common in the chemical space [9].

Deep learning techniques have also been explored more recently, notably, convolutional neural networks, recurrent neural networks, and fully connected feed-forward networks. Deep neural networks have been shown to perform better than other machine learning techniques, such as random forest, at compound activity prediction [10]. However, deep learning often needs sample sizes in the millions. Unfortunately, drug discovery is limited for available data as drug development and testing is long and costly. Recent work has been done on employing one-shot learning in the drug candidate space, which requires only a few samples [11].

5. Aims and Goals

I propose to evaluate the usefulness of various drug-like properties in the drug candidate space, by employing principal component analysis (PCA) and a selection of supervised machine learning techniques in the task of classifying approved drugs from unapproved candidate drugs. The algorithms I will evaluate include logistic regression, random forest, support vector machine, and neural network. I will evaluate and compare the best models from each to determine which algorithm performs the best in this classification task. Secondly, I propose to create further models to classify between drugs that have remained on the market, from drugs that have been discontinued. These models would act as a starting point for creating tools that would support decision making in various drug development stages.

Methods

6. Data Collection

The ChEMBL database supplies publicly available data on drugs and drug-candidates, such as binding, functional and ADMET data [12]. ChEMBL currently has 1.8 million distinct compounds, including about 11 thousand approved drugs. Among the data available, they provide calculated molecular properties, including those used in Lipinski's rule of five and more.

To generate a dataset of approved drugs and drug-candidates that have not been approved, I filtered for compounds using their maximum approved phase in clinical trials. I classified approved drugs as those that passed phase 4 in at least one indication, and have not been withdrawn from the market; a total of 3,092 samples were retrieved. I classified unapproved candidate-drugs as those that have not passed phase 2 in any indication; a total of 1,696,443 samples were retrieved.

I also generated a dataset of withdrawn drugs by filtering for those that have passed phase 4 clinical trials in an indication, but was withdrawn from the market; a total of 233 samples was retrieved.

7. Data Preprocessing and Dimension Reduction

All non-numeric features were removed from analysis to simplify algorithm comparisons, resulting in 15 predictors. Samples were analyzed for missing data; those with missing data in more than 5 features were excluded from the dataset, reducing total samples in

approved drugs, non-approved drugs, and withdrawn drugs to 2,702, 1,668,411, and 230, respectively. I replaced missing data using generalized mean imputation on each feature, including all classes (approved, non-approved, and withdrawn) in the mean calculations.

Two datasets where created: approved vs. non-approved and approved vs. withdrawn. Each dataset was then balanced using random under-sampling, where the majority class was sub-sampled to match the number of samples in the minority class. They were then split using a random stratified selection of samples to achieve a 1:4 test to training data split, each balanced by class.

Using the training set, features were mean centered and normalized in variance. Following normalization, features were projected to less dimensions using PCA; the approved vs. non-approved dataset was reduced to 7 principal components to cover 92.8% of variance while the approved vs. withdrawn dataset was reduced to 6 principal components to cover 91.8% of variance. All testing set transformations were applied to the test set.

8. Hyperparameter and Model Validation

To determine the best hyperparameters for each approach, models with varying hyperparameter combinations were created and evaluated using 10-fold cross-validation on the testing sets. To determine the best hyperparameter selection, the average F1 score of each combination was compared, and the greatest was selected.

Models were then trained on the complete training sets using the best hyperparameter combination and evaluated on the held-out test sets. Final models were evaluated and compared using precision, recall, F1, and Cohen's kappa scores. All computational methods were completed in python using the python 3.7 interpreter. All models were created using scikit-learn 0.20.3.

Results

9. ChEMBL Data Analysis

Most samples in the ChEMBL dataset have complete data in all features evaluated features, however, there are some samples missing 1 – 4 features and some present with 9 or more of the features missing data (Fig. 1). Amongst the features missing data, ACD ApKa and ACD BpKa are the highest amongst all data sets, though, in all features more than half of all samples have data (Fig. 1).

10. Algorithm Performance

Logistic Regression.

For the approved vs. non-approved dataset, the best model used an inverse regularization strength of 0.84, though, changing the regularization parameter did not affect the model performance (Fig. 3). For the approved vs. withdrawn dataset, the best model used an inverse regularization strength of 3.68, though again, changing the regularization parameter did not affect the model performance (Fig. 3).

For the approved vs. non-approved dataset, logistic regression selected approved drugs from the test set with a precision of 0.81, recall of 0.76, F1 of 0.78, and a kappa of 0.58. For the approved vs. withdrawn dataset, logistic regression selected approved drugs on the test set with a precision of 0.60, recall of 0.46, F1 of 0.52, and a kappa of 0.15.

Support Vector Machine.

For the approved vs. non-approved dataset, the best model used an inverse regularization strength of 12.61 and a radial basis function kernel; gamma was automatically determined. For the approved vs. withdrawn dataset, the best model used an inverse regularization strength of 100.35 and a radial basis function kernel; again, gamma was automatically determined. For both datasets, changing the regularization parameter did not affect the model performance, however, there was a marginal improvement with the radial basis function kernel over the linear kernel, although, the radial basis function kernel had more presence of overfitting as well (Fig. 4).

For the approved vs. non-approved dataset, support vector machine selected approved drugs from the test set with a precision of 0.83, recall of 0.80, F1 of 0.82, and a kappa of 0.64. For the approved vs. withdrawn dataset, support vector machine selected approved drugs on the

test set with a precision of 0.68, recall of 0.59, F1 of 0.63, and a kappa of 0.30.

Random Forest.

For the approved vs. non-approved dataset, the best model used 377 estimators. For the approved vs. withdrawn dataset, the best model used 167 estimators. For both datasets, increasing the number of estimators beyond 50 did not dramatically increase model performance (Fig. 5).

For the approved vs. non-approved dataset, random forest selected approved drugs from the test set with a precision of 0.83, recall of 0.82, F1 of 0.83, and a kappa of 0.65. For the approved vs. withdrawn dataset, random forest selected approved drugs on the test set with a precision of 0.68, recall of 0.59, F1 of 0.63, and a kappa of 0.30.

Multilayer Perceptron (Neural Network).

For the approved vs. non-approved dataset, the best model used an alpha of 1.89, and 1 hidden layer of 100 nodes. For the approved vs. withdrawn dataset, the best model used an alpha of 178.92, and 3 hidden layers of 100 nodes each. For both datasets, altering alpha and the node structure wildly varied performance with no obvious trends present. (Fig. 6).

For the approved vs. non-approved dataset, the multilayer perceptron selected approved drugs from the test set with a precision of 0.79, recall of 0.72, F1 of 0.75, and a kappa of 0.52. For the approved vs. withdrawn dataset, the multilayer perceptron selected approved drugs on the test set with a precision of 0.50, recall of 1, F1 of 0.67, and a kappa of 0.

Overall.

For both the approved vs. non-approved dataset and the approved vs. withdrawn dataset, random forest performed the best on the test set (Table 1) and achieved the best cross-validation scores (Fig. 7).

Discussion

11. Important Drug-like Properties

The ChEMBL database provides many samples of drugs and candidate drug compounds with each being connected to many features and properties. The set of features I used from the database were molecular properties that were readily available, though, more data can be retrieved from looking at many assay results connected to a compound. With most compounds having data in most features (Fig. 1), this database is a good starting point for assessing the usage of simple drug-like properties in machine learning methods.

After transforming the data using PCA, we can see some features are correlated in their importance towards the top two principal components (Fig. 2). A clear trend appears as three clusters, general polarity of a compound (polar surface area and hydrogen bond donor and acceptors), the structure of a compound (rotatable bonds, molecular weight, number of heavy atoms), and the aromaticity of a compound. The number of bioactivities, number of targets, ACD BpKa, and ACD ApKa do not appear to contribute much towards the top two principal components.

The presence of these correlated structures may provide a starting point in searching for important drug-like properties that can be used to distinguish compounds from one-another. For example, the polarity of a compound may play a role in its solubility, which has an important impact as oral drugs need to have enough polarity to become soluble in aqueous solution, but not too much that they cannot pass the membrane [8].

12. Approved Drug Prediction

In the task of predicting an approved drug from a non-approved drug, random forest performed the best, followed closely by support vector machine (Table 1). However, the best random forest model only had a precision of 83% and a recall of 82% on the test set. While this is better than guessing (50% due to a balanced dataset), it leaves much room for error which can be expensive and dangerous in drug discovery.

There was a presence of overfitting in the cross-validation results (Fig. 4; Fig. 5). Random forest and support vector machine could perform better on this task with more samples, though, on the other approaches, there was no overfitting present, indicating that more predictive features may be needed or

alterations to the approaches may be required (Fig. 3; Fig. 6).

While there is some error in the models, they all perform better than simply using Lipinski's rule of 5 to determine an approved drug from a non-approved drug. Using the test set and selecting approved drugs as those with less than two violations [3] results in a precision of 0.51% and a recall of 0.46%, significantly worse than the machine learning models' predictions (Table 1).

13. Withdrawn Drug Prediction

In the task of predicting an approved drug from a withdrawn drug, random forest and support vector machine performed the best on the test set (Table 1). The best random forest model only had a precision of 68% and a recall of 59% on the test set, which is only marginally better than guessing (50% due to a balanced dataset).

Like the earlier task, there was a presence of overfitting in the cross-validation results (Fig. 4; Fig. 5). As mentioned above, support vector machine and random forest could perform better on this task with more samples, though, on the other approaches, there was no overfitting present, showing that more predictive features may be needed or alterations to the approaches may be needed (Fig. 3; Fig. 6).

While there is some error in the models, they all perform better than simply using Lipinski's rule of 5 to determine an approved drug from a withdrawn drug. Using the test set and selecting approved drugs as those with less than two violations [3] results in a precision of 0.51% and a recall of 0.48%, worse than the machine learning models' predictions (Table 1).

14. Case Studies: Model Selections

Approved vs. non-approved

Using the top performing random forest model, the top five compounds from the test set selected as having the highest probability of being an approved drug were sulfanilamide, norepinephrine, idoxuridine, pralidoxime chloride, and cimetidine. All these compounds have passed phase IV trials, acting as an approved treatment for some conditions. Some common observations between these top five compounds are that they all have one phenyl or phenyl derivative group, they have only a few if any atoms that are not the typical atoms in an organic molecule such as carbon, hydrogen, nitrogen, or oxygen, and they all have low molecular weights (Fig. 8).

The bottom five compounds with the lowest probability of being an approved drug in the test set were all unnamed compounds that have not been tested in any clinical trials. Some common observations between these bottom five compounds are that they all have many ring structures, they have one or more halogen atoms, and they all have significantly higher molecular weights than the top five (Fig. 9).

Approved vs. withdrawn

Using the top performing random forest model, the top five compounds from the test set selected as having the highest probability of being an approved drug (that has not been withdrawn) were pamidronic acid, carbenicillin phenyl, meprednisone, levocabastine hydrochloride, and edoxaban tosylate. None of these compounds were withdrawn.

The bottom five compounds with the highest probability of being a withdrawn drug in the test set were sulfadoxine, fenfluramine hydrochloride, levomethadyl acetate, oxyphenbutazone, and fluorescein. Sulfadoxine and fluorescein were not withdrawn, however, the other three were withdrawn in at least one market for toxicity.

15. Next Steps

While these molecular properties, with the sample sizes available, did not supply enough information to achieve high performing separation, they do shed some light on the viability of machine learning in drug discovery using publicly accessible databases. The models perform better than simply using Lipinski's rule of five, so they may be a step in the right direction. To create more reliable models, more features will need to be evaluated, and new algorithms which perform better on smaller sample sizes should be explored.

The creation of reliable models in drug prediction could speed up the drug discovery process significantly, reduce any unnecessary costs attached to continuing development with a poor drug candidate, and most importantly, reduce the probability that poor or harmful compounds reach late phase trials or the public market.

References

[1] H. van de Waterbeemd and E. Gifford, “ADMET in silico modelling: Towards prediction paradise?,” *Nat. Rev. Drug Discov.*, vol. 2, no. 3, pp. 192–204, 2003.

[2] D. Schuster, C. Laggner, and T. Langer, “Why Drugs Fail - A Study on Side Effects in New Chemical Entities,” *Curr. Pharm. Des.*, vol. 11, no. 27, pp. 3545–3559, Oct. 2005.

[3] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings,” *Adv. Drug Deliv. Rev.*, vol. 64, no. SUPPL., pp. 4–17, Dec. 2012.

[4] P. D. Leeson and B. Springthorpe, “The influence of drug-like concepts on decision-making in medicinal chemistry,” *Nat. Rev. Drug Discov.*, vol. 6, no. 11, pp. 881–890, 2007.

[5] I. Yusof and M. D. Segall, “Considering the impact drug-like properties have on the chance of success,” *Drug Discov. Today*, vol. 18, no. 13–14, pp. 659–666, 2013.

[6] G. R. Bickerton, G. V Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, “Quantifying the chemical beauty of drugs,” *Nat. Chem.*, vol. 4, no. 2, pp. 90–98, Feb. 2012.

[7] J. Panteleev, H. Gao, and L. Jia, “Recent applications of machine learning in medicinal chemistry,” *Bioorganic Med. Chem. Lett.*, vol. 28, no. 17, pp. 2807–2815, 2018.

[8] W. L. Jorgensen and E. M. Duffy, “Prediction of drug solubility from structure,” *Adv. Drug Deliv. Rev.*, vol. 54, no. 3, pp. 355–366, Mar. 2002.

[9] A. Lavecchia, “Machine-learning approaches in drug discovery: Methods and applications,” *Drug Discov. Today*, vol. 20, no. 3, pp. 318–331, 2015.

[10] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, “The rise of deep learning in drug discovery,” *Drug Discov. Today*, vol. 23, no. 6, pp. 1241–1250, 2018.

[11] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, “Low Data Drug Discovery with One-Shot Learning,” *ACS Cent. Sci.*, vol. 3, no. 4, pp. 283–293, Apr. 2017.

[12] A. Gaulton *et al.*, “ChEMBL: A large-scale bioactivity database for drug discovery,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. 1100–1107, 2012.

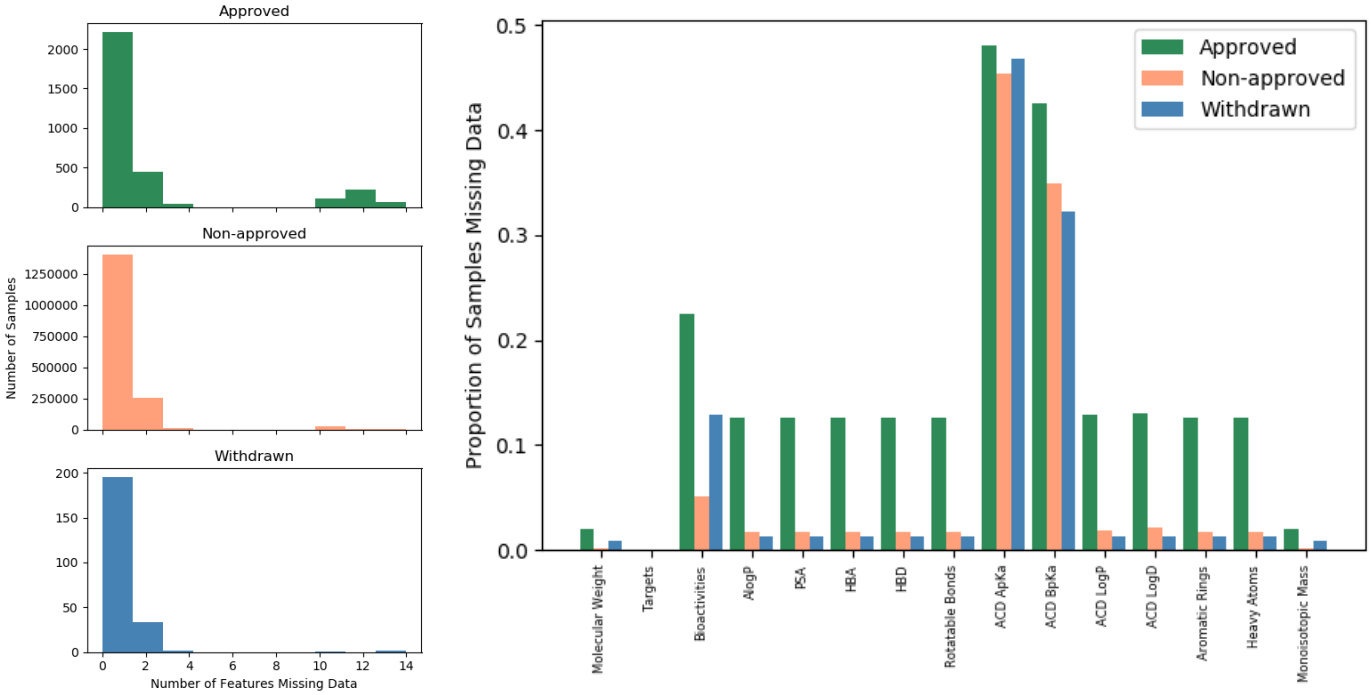

Figure 1: Analysis of the retrieved data from the ChEMBL database. Data were analyzed separately by their dataset: approved drugs, non-approved drugs, and withdrawn drugs.

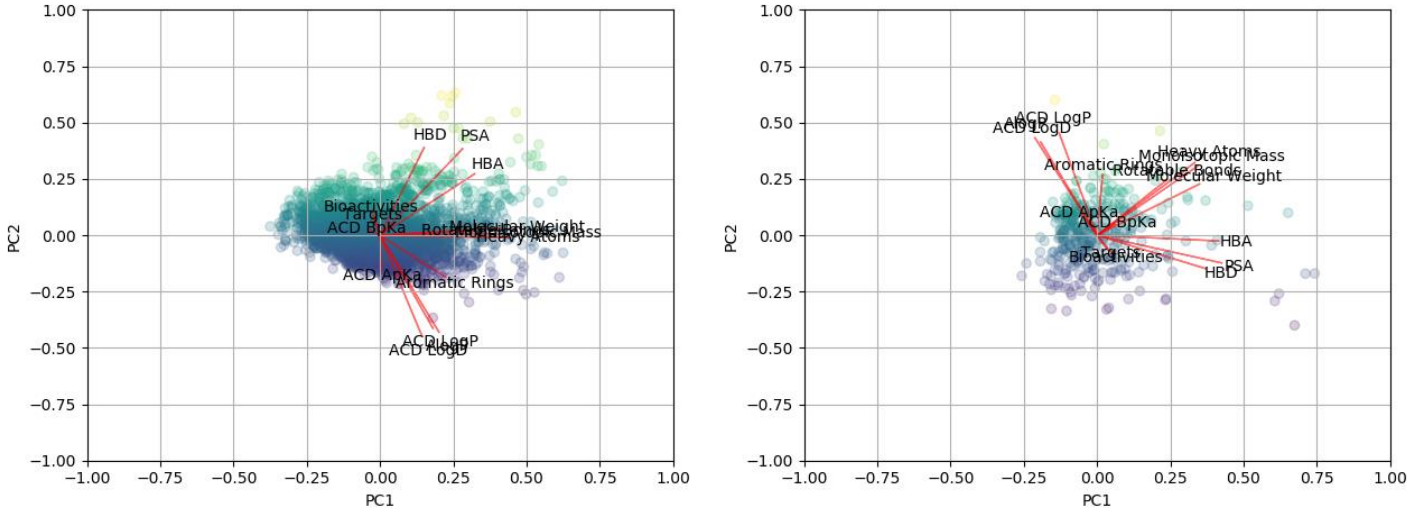

Figure 2: PCA biplot of two joint datasets: approved vs. non-approved (left) and approved vs. withdrawn (right). Direction of arrows show the features' contribution to that principal component. Lengths of arrows show the amount of the features' contribution to the principal components.

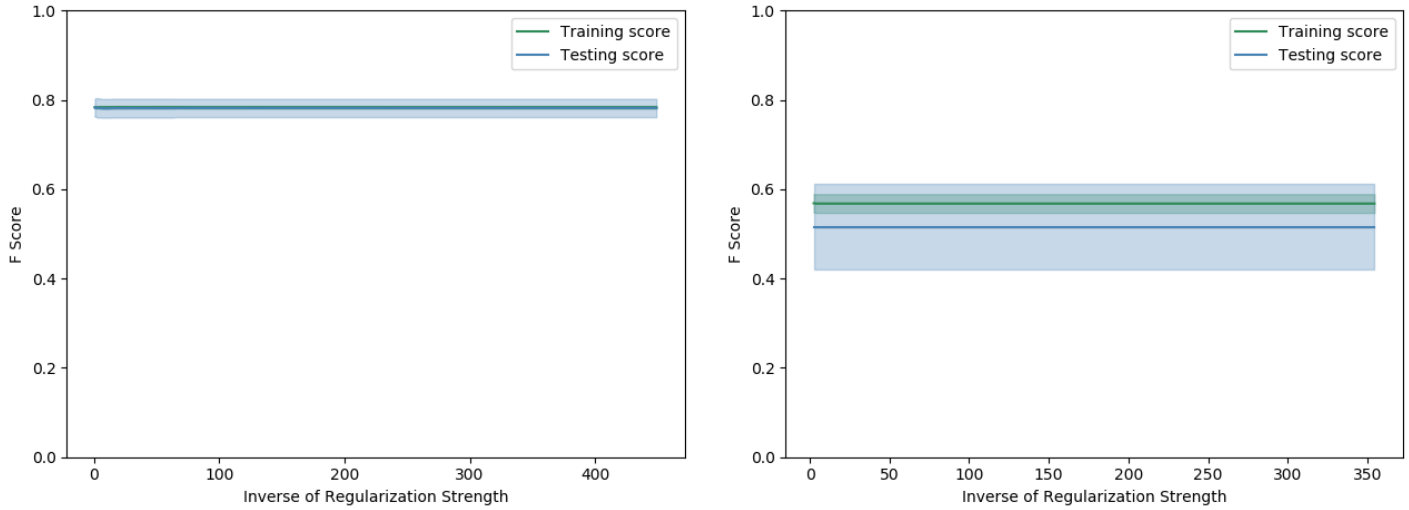

Figure 3: Logistic regression 10-fold cross validation results of two joint datasets: approved vs. non-approved (left) and approved vs. withdrawn (right). Models with varying inverse regularization strengths were evaluated on their F1 score.

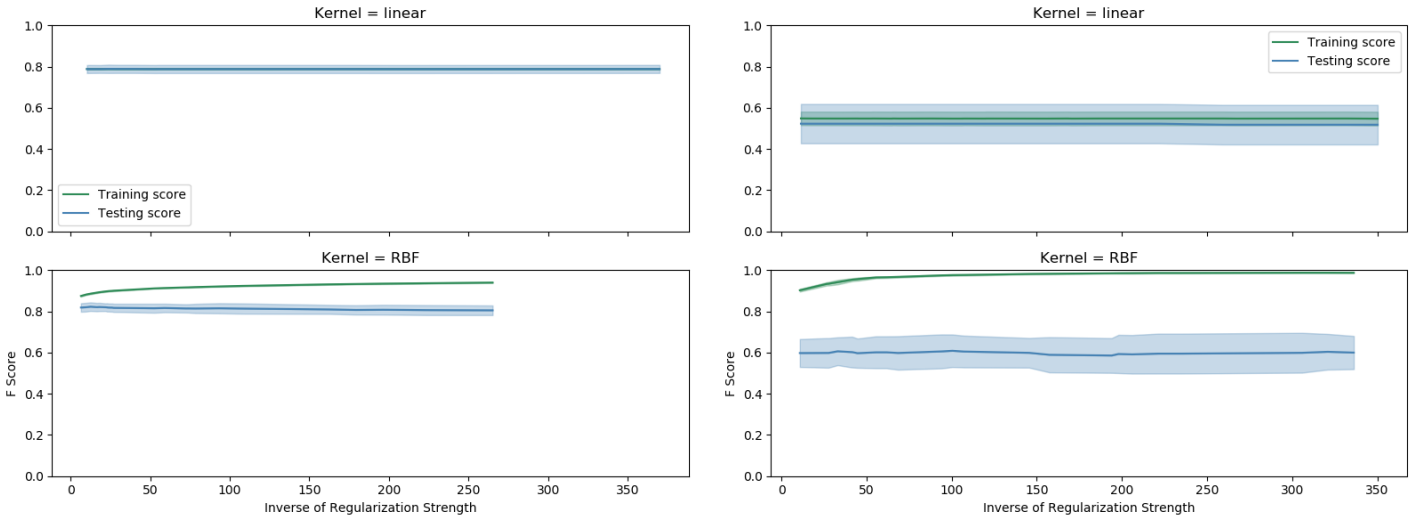

Figure 4: Support vector machine 10-fold cross validation results of two joint datasets: approved vs. non-approved (left) and approved vs. withdrawn (right). Models with varying inverse regularization strengths and kernels were evaluated on their F1 score.

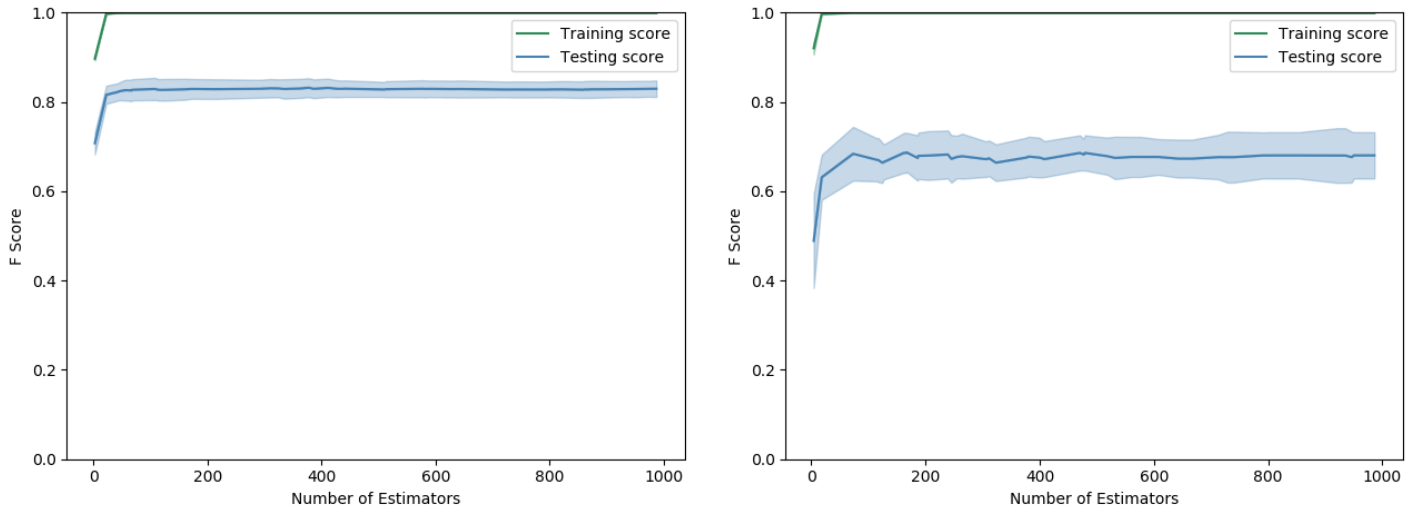

Figure 5: Random forest 10-fold cross validation results of two joint datasets: approved vs. non-approved (left) and approved vs. withdrawn (right). Models with varying number of estimators were evaluated on their F1 score.

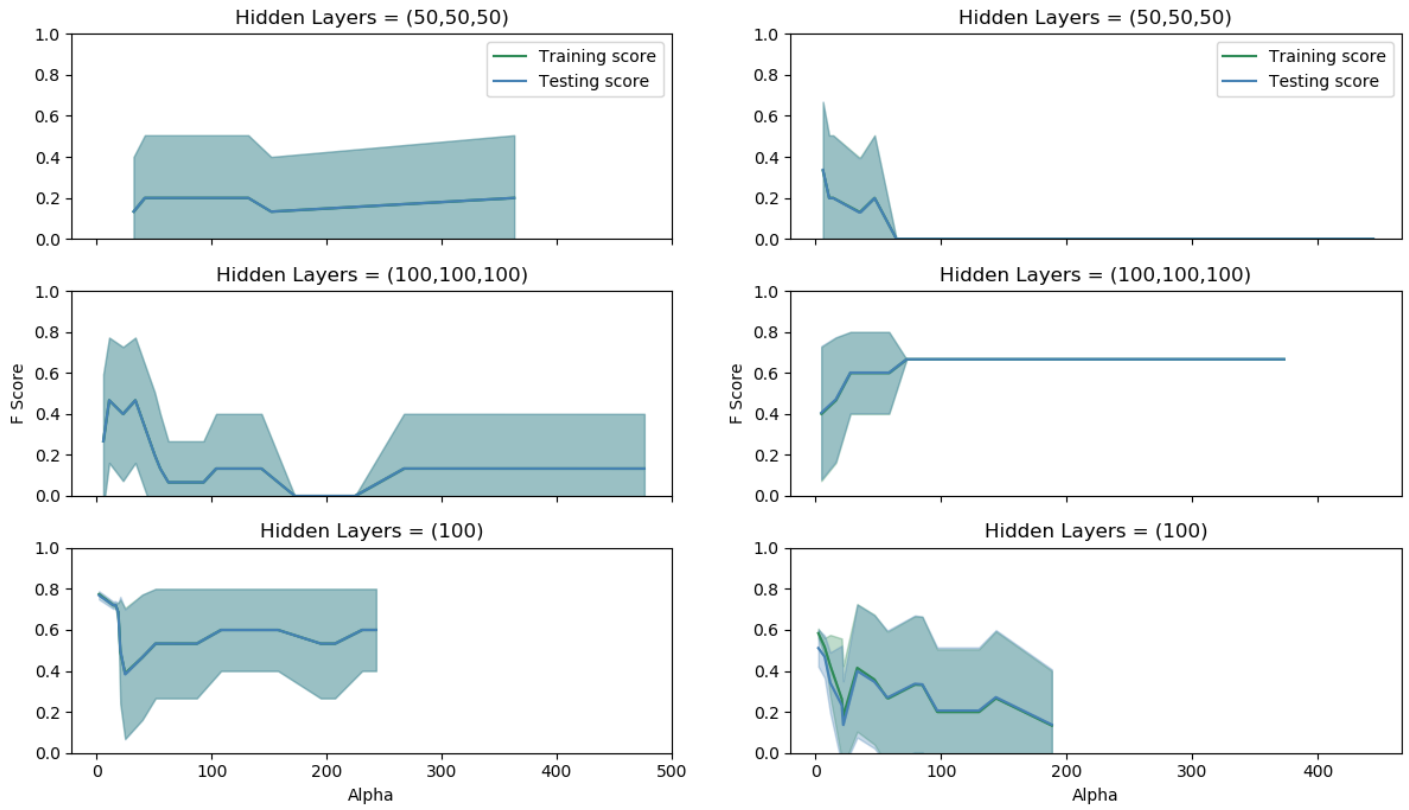

Figure 6: Multilayer Perceptron 10-fold cross validation results of two joint datasets: approved vs. non-approved (left) and approved vs. withdrawn (right). Models with varying alphas and hidden layer structures were evaluated on their F1 score.

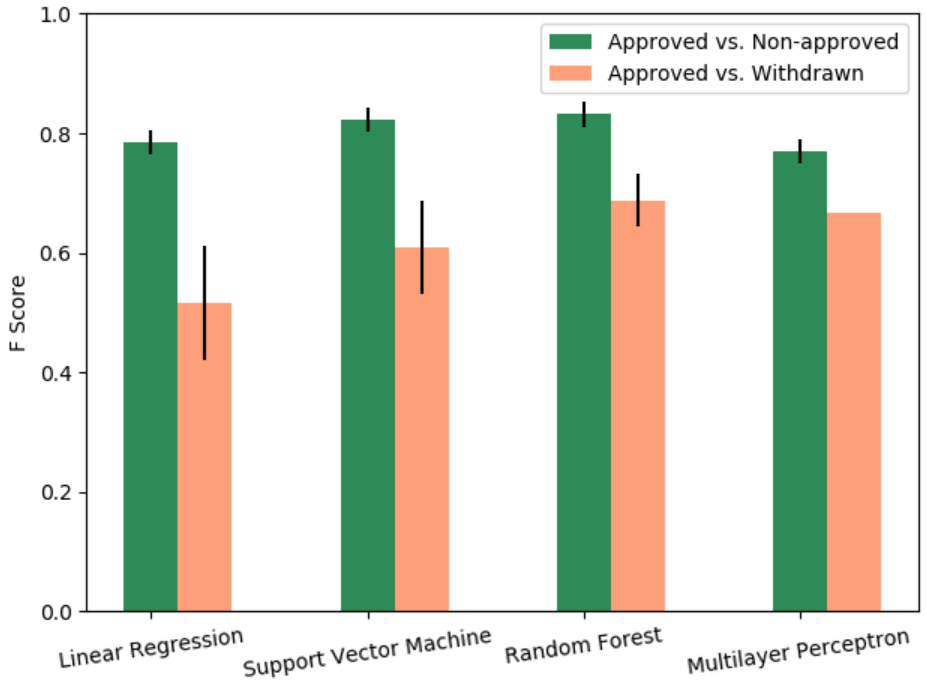

Figure 7: Cross-validation scores of the best models for each machine learning approach.

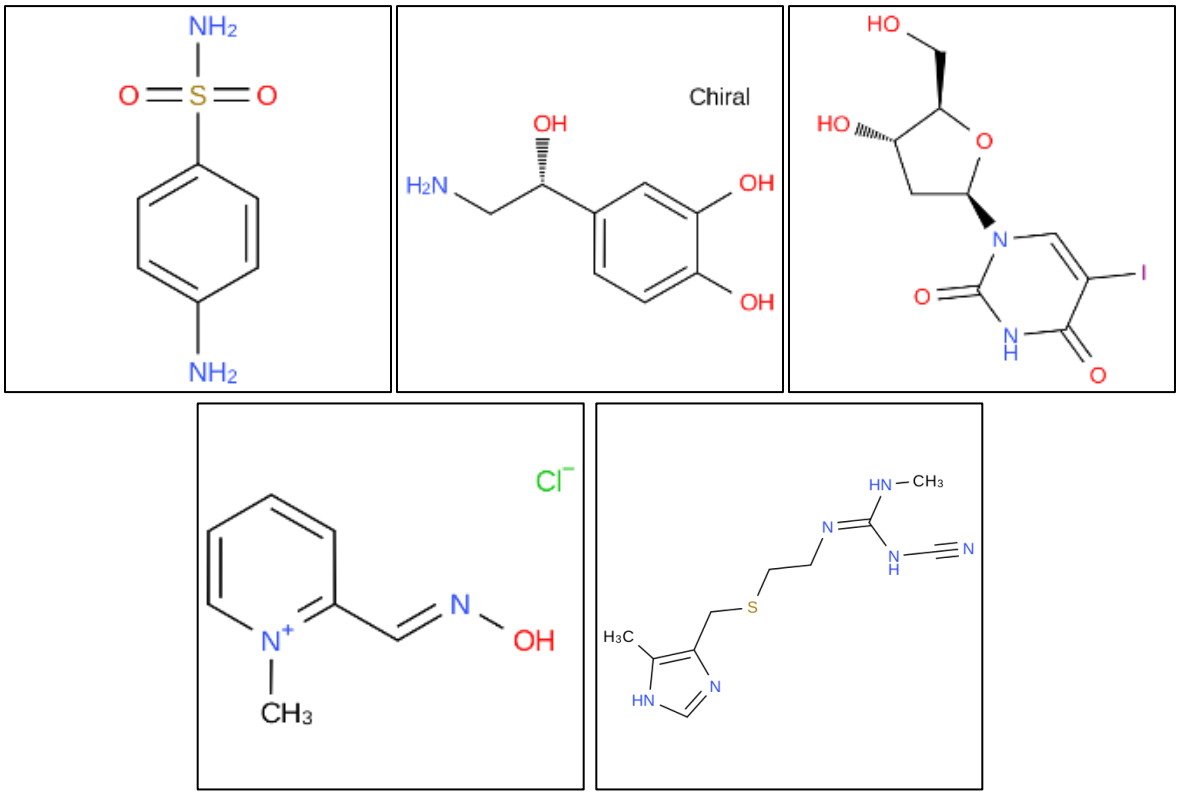

Figure 8: Chemical structure of the top five compounds likely to be approved drugs in the approved vs. non-approved test set. Pictures are sulfanilamide, norepinephrine, idoxuridine, pralidoxime chloride, and cimetidine (from left to right).

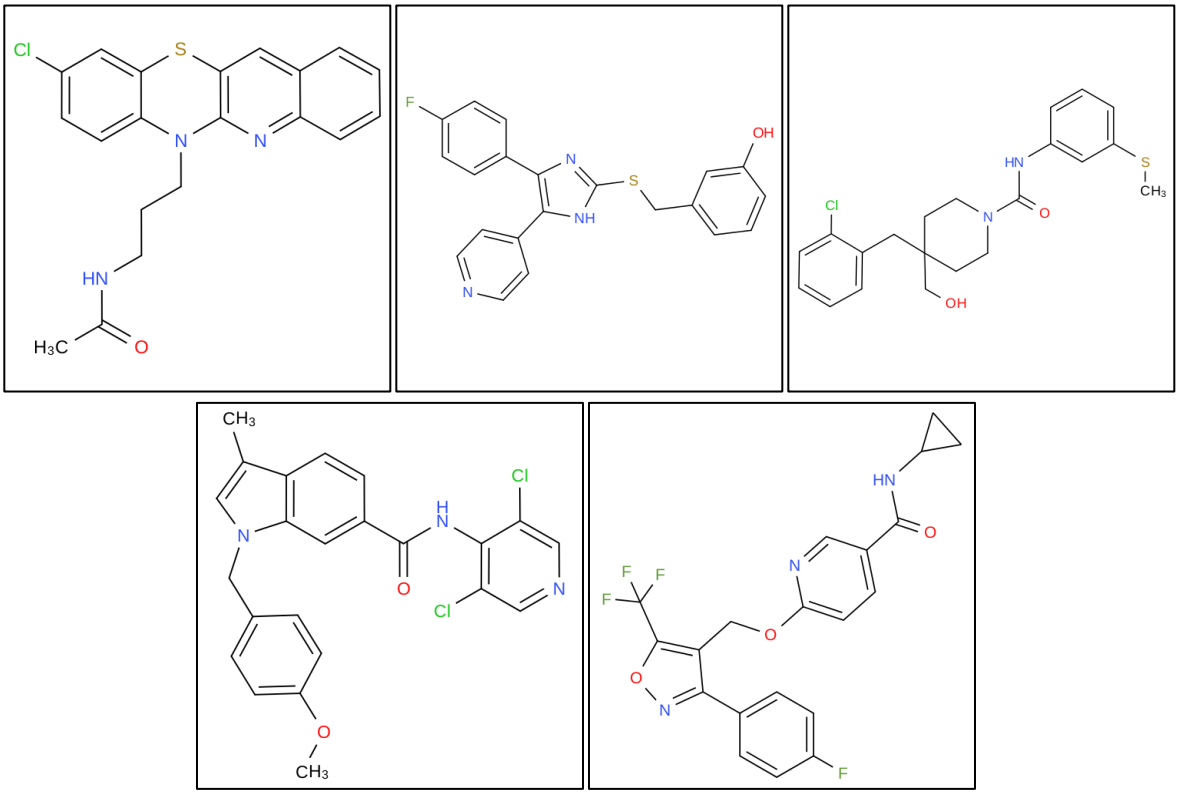

Figure 9: Chemical structure of the top five compounds likely to be non-approved drugs in the approved vs. non-approved test set. All compounds are unnamed.

Table 1: Performance on the Test Set using Multiple Evaluation Metrics.

APPROVED VS. NON-APPROVED	PRECISION	RECALL	F1	KAPPA
LOGISTIC REGRESSION	0.81	0.76	0.78	0.58
SUPPORT VECTOR MACHINE	**0.83**	0.80	0.82	0.64
RANDOM FOREST	**0.83**	**0.82**	**0.83**	**0.65**
MULTILAYER PERCEPTRON	0.79	0.72	0.75	0.52
RULE OF FIVE	0.51	0.46	0.48	NA

APPROVED VS. WITHDRAWN	PRECISION	RECALL	F1	KAPPA
LOGISTIC REGRESSION	0.60	0.46	0.52	0.15
SUPPORT VECTOR MACHINE	**0.68**	**0.59**	**0.63**	**0.30**
RANDOM FOREST	**0.68**	**0.59**	**0.63**	**0.30**
MULTILAYER PERCEPTRON	0.50	1.00	0.67	0.00
RULE OF FIVE	0.51	0.48	0.49	NA