

Chemistry Capstone Project: *An exploration of model complexity in the drug discovery field*

Authors: Phil Michaels and David Dunstan

University of Michigan: Masters of Applied Data Science Program

August 2022

Abstract

Model selection is an important aspect to any data science project. With new model frameworks being published regularly, there have been increasing calls for standardized datasets across industries to help compare the results. We herein describe an evaluation of different models of varying complexity in the drug-discovery context. We rely on cheminformatics to process and featurize 6 datasets that cover a range of sizes and drug discovery related tasks. We trained several common classification models as well as a directed message passing neural network from the Chemprop package for each dataset. We then compared the models using a range of metrics and found roc-auc to be the most useful for our examples. From our results, while Chemprop was amongst the highest scoring in all cases, it did not always outperform the simpler models. Therefore, we find that exploring a wide range of models to be prudent in establishing baselines and providing comparative insight.

Introduction

It is March of 2020, and a highly contagious virus is spreading rapidly across the world and killing without any known treatments. Scientists race to first understand the biology of the virus, decipher its DNA and protein structures and then identify opportunities for small molecules to interrupt or inhibit them. This is the exact scenario that many drug discovery scientists find themselves in everyday, instead of COVID-19 they are looking to treat HIV, Alzheimers, or cancer. The needs of patients drive the drug discovery industry to find these novel treatments with the utmost speed. Machine learning holds massive promise to help better predict which molecules can be used to treat these diseases and get safer drugs to patients faster.

The field of drug discovery is complex due to the interplay between biological mechanisms of disease and drug candidates. The search often begins with an efficient high-throughput screen to identify potential hit compounds from large library collections. Next, these hits are expanded and chemically iterated upon to optimize their potency as well as their properties so they can be absorbed into the body, reduce the risk of toxic side-effects and allow them to be produced efficiently. These optimized hits, called leads, are then moved into further testing where animal models are used to confirm the results before moving them towards the clinic.

The opportunity to apply modeling at all of these stages offers the potential to increase the speed, reduce the costs, and minimize the need for animal studies. Advances in high-throughput

experimentation and automation are enabling access to larger and more complete datasets than ever before. A typical drug-discovery workflow involves virtually designing a set of potential drug-like compounds and then applying machine learning models based on both historical and target-specific data. These models help to predict which compounds are most likely to be successful while minimizing the risk of poor properties. Scientists in the lab then produce the selected compounds and profile them in a range of assays that empirically test the predictions. The process is iterative in that the new data fuels re-training of the models and subsequent selections until the desired outcome is achieved or sufficient data is produced to convince the scientists that alternative molecules are needed due to an adverse finding.

Project Goal

Given the increasing role that modeling is playing in drug discovery, we wanted to assess whether the trend towards more complex modeling approaches is contributing to increased performance when compared with more simple baselines. While there are many ways to evaluate performance of a model, we took inspiration from a recent Nature paper in which the authors suggest that the use of standardized data sets would be a way to help better compare new modeling techniques. [1] We selected a subset of these datasets from Moleculenet [2] that would be typically relevant to an early drug discovery space. We then compared the results of these various models to each other using several scoring metrics.

While there are a multitude of complex models that have been published, we chose to explore a message passing neural network package that has been developed by MIT, called Chemprop. [3] In the initial paper, the authors describe and compare the Chemprop approach to a series of other models, however they looked primarily at a single metric and did not factor in other challenges that arise from utilization of complex models. [4] Furthermore, we feel that there is opportunity to dive deeper into the predictions between the models to try and understand where performance is breaking down. For the simple models, we chose to explore a range of scikit-learn model deployments. We also utilized scikit-learn based metrics for evaluation.

Datasets:

For the comparison of the various model types, we selected several datasets that would be relevant to the early drug discovery space. Considering the conclusions of Bender et al. [1], we selected a subset of datasets which were suggested for evaluation. As a result, the datasets that we have are all publicly available and have been utilized in multiple machine learning evaluations and projects. Therefore, there are few ethical concerns for the utilization of these datasets within our project. Any ethical considerations for this project are then the result of the misapplication of the resulting models, or from erroneous results, where our models may not be performing as expected. We therefore caution that all results here are made in an attempt to compare different model types to one another and not to utilize the resulting models for prediction of physicochemical or activity related predictions in a drug-discovery context.

We did diverge from the prior works in that we also converted all of the datasets to a classification target in order to better compare the resulting models to one another, and to limit the scope of the project. For the datasets that were sourced as regression targets, we utilized generally accepted literature guidance to assign a binary target value. While there are a multitude of regression modeling problems that would be beneficial for drug discovery, we felt that classification metrics would provide a better degree of interpretability and homogeneity. Overall, we feel that these datasets represent a good range of different size, class balance, and difficulty.

BACE Dataset: The BACE dataset is a set of compounds with their BACE (an enzyme involved in the formation of neurodegenerative amyloid plaques) activity. In order to convert this problem to a classification task, a generally accepted biological activity cut-off of pIC50 values > 6 was employed to separate “active” compounds, which inhibit the target, from “inactive” compounds, which do not show any effect. [5]

Dataset Size: Small (1513 molecules)

Class Balance: Balanced (1012 positive class : 501 negative class)

TOX21 Dataset: The TOX21 dataset is a collection of compounds and their activity across a range of different toxicity relevant assays. [6] While there are several different assays in the dataset, we decided to focus the dataset on the ‘NR-AhR’ assay which is a liver toxicity assay, due to the data completeness, relevance to early drug discovery, and broad applicability to a number of different disease areas.

Dataset Size: Medium (7831 molecules)

Class Balance: Unbalanced (768 positive class : 5781 negative class)

Clintox Dataset: The Clintox dataset is a list from the FDA of drugs and whether or not they have shown toxicity in patients. This dataset, while small, represents some of the most relevant data for human patients from a drug safety perspective.

Dataset Size: Small (1484 molecules)

Class Balance: Unbalanced (112 positive class : 1372 negative class)

Solubility Delaney Dataset: The solubility dataset is a measure of how much of a given compound can be dissolved in water, which is an important metric to optimize for any drug to ensure it is able to be properly absorbed into the bloodstream. In order to convert this to a classification problem, a solubility measure of 0.1M was utilized to determine if a compound should be classified as soluble or insoluble. [7]

Dataset Size: Small (1128 molecules)

Class Balance: Unbalanced (185 positive class : 943 negative class)

Deepchem Lipophilicity Dataset: The lipophilicity (how strongly a molecule is attracted to grease or fat compared with water) is also an important property for drug discovery. In order to be properly absorbed across the gut lumen and into the bloodstream, and to avoid potential clearance or toxicity issues, a molecule should have a lipophilicity between -0.4 and 5.6. [8] Therefore, we applied these thresholds to convert this initial dataset into a classification one. After classification, this dataset was imbalanced, but in favor of the positive class.

Dataset Size: Medium (4,200 molecules)

Class Balance: Unbalanced (4,055 positive class : 145 negative class)

HIV Dataset: The HIV dataset contains a set of molecules and their reported activity to inhibit HIV replication in a biochemical assay. The dataset was already a classification problem, but represents a relatively large dataset for our purposes.

Dataset Size: Very Large (41,127 molecules)

Class Balance: Unbalanced (1443 positive class : 39684 negative class)

For the comparison of chemical modeling approaches, we opted to utilize a cross-validation approach to model training and evaluation. Therefore, each dataset was only split into a training set and a withheld validation set for final performance evaluation. A key challenge in the area of chemical modeling is the idea of a chemical scaffold. When training a machine learning model, the model attempts to learn the various features of the molecules, however if there molecules within a dataset are very homogeneous in terms of a chemical structure, with only a few minor modifications, we risk the model learning to search for a specific sub-structure, instead of learning generalizable features. While not true data-leakage, it can limit the applicability of a model to new chemical structures. Typically within a drug discovery project there is a desire to have models that generalize well in order to facilitate exploration and testing of new scaffold types, and therefore limit the risk of unexpected later findings, such as toxicity.

In order to address the problem of scaffold balance we used K-nearest neighbors (KNN) clustering to help split the dataset such that the same chemical scaffolds are not fully represented in both the training and validation data. We also performed a more traditional random split for comparison. It should be noted that there are also more elaborated approaches, such as the scaffold based split available in Chemprop [9] and while potentially effective, we found that the KNN approach was straightforward to execute and effective. One risk that we were aware of was that the KNN approach might alter the class balance of the various datasets relative to the random split. However, from Figure 1, we can observe that the class balance is roughly the same between approaches and the validation sets share the same class balance as the training set. In order to assign the number of clusters, we

explored some metrics such as silhouette score, but found that dividing the total data set size by 30 produced clusters that were small enough to enable roughly even class balance, while still producing clusters that were grouped by chemical substructure.

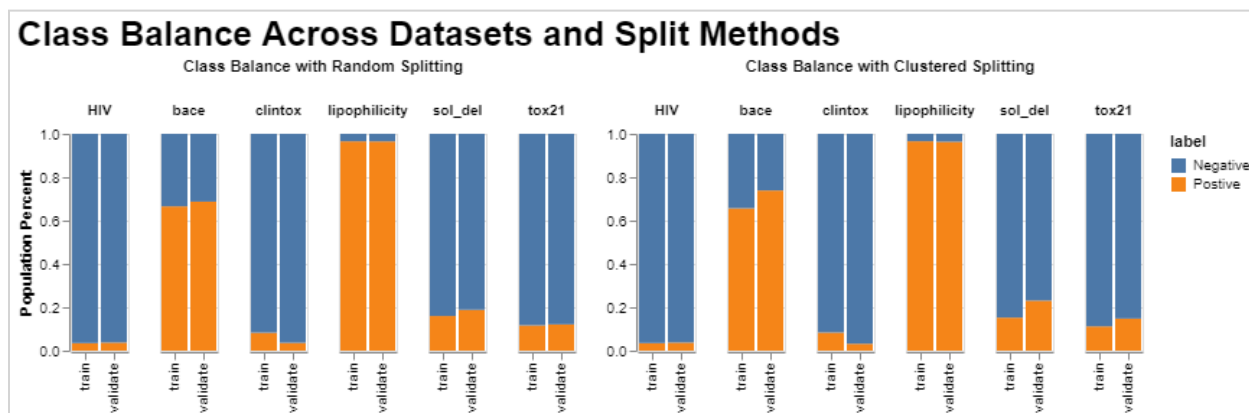


Figure 1. The final class balance metrics for the various data splits by KNN clustered and random splitting methods

In comparing the various clusters we also wanted to evaluate how well the clustering algorithm was doing at identifying groups of active molecules. In Figure 2, we can observe the red line indicating the overall dataset fraction of active compounds compared with the relative fraction for each cluster for the BACE dataset. The presence of over 38 of 50 total clusters above or below this line indicates that it is doing a good job in grouping like-compounds together and therefore helping to more realistically between training and validation sets.

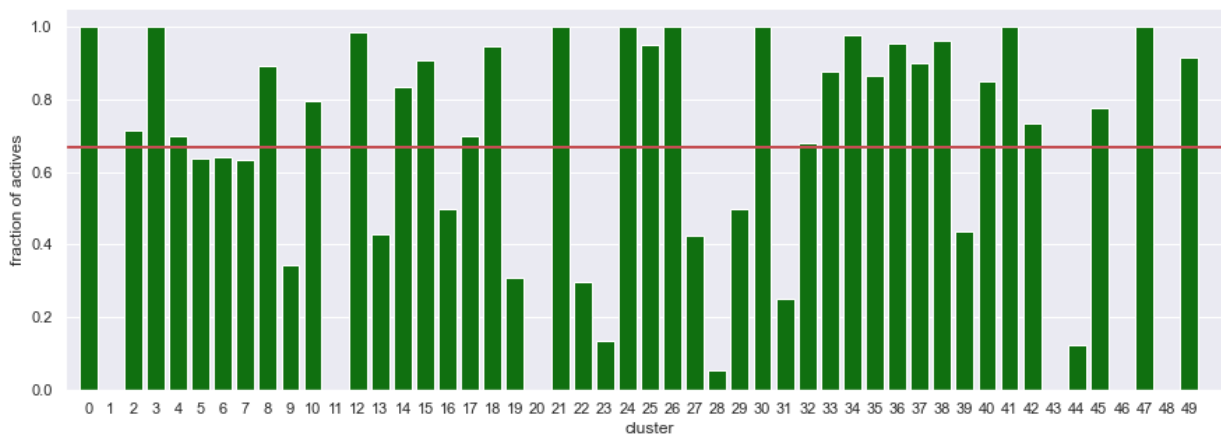


Figure 2. The fraction of active molecules in each cluster, with the red-line indicating the overall class ratio in the dataset.

The other evaluation we can perform on initial datasets is to look at how the molecular fingerprints might cluster with respect to the relevant activity data. In order to assess this in a generic way, we generated the fingerprints and then employed the Uniform Manifold Approximation (Umap) to reduce

the bit vectors to two dimensions for visualization. We decided to utilize Umap over PCA since it does a better job at preserving local structure and is more interpretable than TSNE. [10] From the resulting visualization in Figure 3, we can get a rough gauge of how difficult the various classification tasks might be. For instance, we can observe for the BACE dataset that many of the local clusters are separated based on activity, while for the HIV dataset, the active compounds appear to be randomly spread in the chemical space. Furthermore, BACE, TOX21, Clintox and Solubility all appear to have clear clusters and structure within the data, whereas HIV and Lipophilicity both appear to have mainly a single cluster. Therefore, we make the broad generalization that the HIV and Lipophilicity datasets are more difficult than the others.

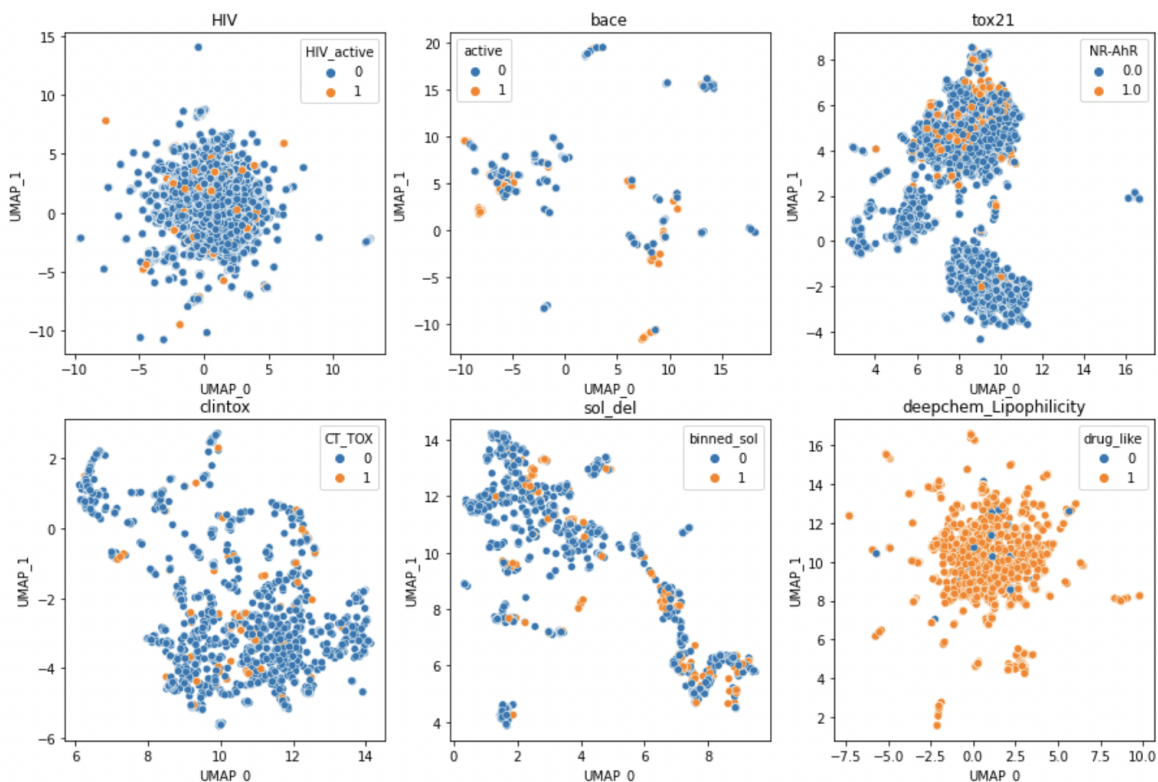


Figure 3. The Umap dimensionality reduction visualizations for each of the datasets.

Modeling:

To explore the question of how much value does model complexity add in the context of early drug-discovery programs, we needed to select the models to compare. There is not necessarily a concrete definition of what separates simple and complex models, and instead they operate on more of a gradient. We applied general heuristics in that if the model is too complex it might overfit the data, and fail to generalize. While, if a model is too simple, it might underfit the data and fail to give optimal performance. Interpretability is also something which can often favor simpler models, however there are also techniques for doing this in more complex modeling space. Lastly, the computational costs should also be considered, with simpler models typically being faster to train and easier to optimize

hyperparameters for than more complex ones. [11] For this project we evaluated several of the readily available Scikit-learn models against the MIT Chemprop model.

Complex Models

In order to evaluate the complex modeling space, we opted to explore a publicly available package instead of trying to develop a novel framework on our own. Due to the time of the project, it did not seem feasible to build something which would function better than the work resulting from collaborations spanning several years between academia and industry. While there are many packages available such as Few-Shot Learning for Molecules ([FS-MOL](#)) [12] from a Microsoft Collaboration, to DeepChem a scientific learning package from Stanford ([DeepChem](#)) [13], we opted to go with the MIT package of Chemprop. For this project we felt that Chemprop provided a good framework that balanced both ease of use as well as having in-built features such as data splitting, early stopping and cross validation. Furthermore, evaluations of Chemprop against other modeling methods are present in the literature, which help give further comparison and insight into our results. [4]

The Chemprop package offers many different approaches to training the models, however the core is based around a directed message passing neural network (D-MPNN). This network works by passing a directed message across a molecular “network” made up of bonds as edges and atoms as nodes. The resulting representation is then utilized to create a learned representation of the molecule which the model uses to train against the property prediction task. In addition, the Chemprop package also allows for a fixed feature representation, such as a Morgan Fingerprint, or generic physicochemical properties and descriptors to be utilized in conjunction. For this project we utilized a Morgan Fingerprint representation in addition to the D-MPNN for the model to enable a better comparison to the simpler models. The Chemprop models were then trained over 50 epochs using ROC-AUC as the optimization metric, a 5-fold cross validation. The remaining parameters were set as default.

Simple Models

There are a vast array of different models that can be applied to classification problems, and many packages that offer versions of them. For this project, we chose to explore the Scikit-learn implementations since they are generally accepted in the machine learning industry and also offer many helpful functions such as cross validation and data splitting. Furthermore, the Scikit-learn package has implementations of many common model types which helped provide a range of models for us to compare. [14] For this project we decided to utilize Logistic regression and K-nearest neighbors (KNN) as some of the most simple models. We also trained a Dummy Classifier that would only predict the most frequent class in order to provide a very simple baseline. We looked at both Random Forest and Gradient Boosted Decision Trees as different decision tree type classifiers, that are utilized often within the chemistry community and are less sensitive to data normalization, while still remaining interpretable. Lastly, we also fit a Support Vector Machine Classifier (SVC) as these were quite popular in many chemistry applications during the early 2000’s, before Neural networks surpassed them in popularity. [15, 16]

For the simple models, we featurized the molecules using the Morgan Fingerprint generation algorithm available in the RDkit package. While we explored different radii and bit numbers initially, we found that a radius of 2 and a bit length of 1024 produced reasonable results without generating excessively large vectors. While it is possible that these parameters could be further tuned in future work, we were unable to thoroughly evaluate the fingerprint parameters across all of the datasets due to the limited time. For the majority of the simple models we kept the parameters at their default values, with the exception of logistic regression where we increased the maximum iterations to 1000. For the KNN model we specified 5 neighbors. For both the Gradient Boosted and Random Forest models we used 100 estimators and for the learning rate we used 0.1 in the case of gradient boosting. The SVC model utilized the Sklearn algorithm to produce the predicted probabilities as these are not available by default.

Modeling Methods

While the simple models could be trained in a reasonable amount of time on our local computers, the Chemprop model did take a significant amount of computational time. We therefore opted to utilize a GPU to help speed this up. From an initial test on the BACE dataset, which was one of our smaller ones, training the chemprop model locally on a MacBook Air took over 1.5 hours, while on the University of Michigan Great Lakes Cluster it took less than 5 minutes. When factoring in the cross validation and multiple training attempts, the Chemprop training would have been prohibitive to do locally. While many tools exist to access GPU resources, this highlights one key challenge to utilizing more complex models. For this project we opted to use the University of Michigan Great Lakes cluster and a Nvidia Tesla V100 GPU to train our Chemprop models.

In order to get a sense of how well each model was generalizing, while leveraging the entire dataset, we opted to use cross-validation. Cross-validation is where the dataset is divided into several equal parts and one part is held out as a test set during a round of training. Training and this testing is then repeated with each split until all have been used once. The resulting difference between each test score, can give an empirical measure of how well the model is generalizing to new data. For this project, we opted to perform 5-fold cross validation to keep the training times reasonable. This was performed either using the Scikit-learn function, or as a command line feature in the Chemprop package. After performing cross-validation, we then trained the various models on the entire training set and made predictions on both the training and validation sets. The resulting model was then used to predict on both validation and training splits and the metrics were aggregated for evaluation. All the code utilized for this project is available in the git repository (<https://github.com/PJMichaels/Chemistry-Capstone/tree/final>)

Evaluation and Results

Split Method Comparison

In order to determine which data splitting method would be best to carry through for evaluation, we compared a random split to the cluster based split for each of the different model types. In Figure 4, we can see that, with the understandable exception of the dummy model, the performance of all model classes is lower in the case of the clustered split. While this does mean that the models may appear to be lower performing, we feel that this is more representative of what a real-world drug discovery team would face where they are trying to evaluate novel targets based on the limited data they have available and predict into new chemical spaces that the models have not seen. We therefore utilized the cluster based method for further evaluation.

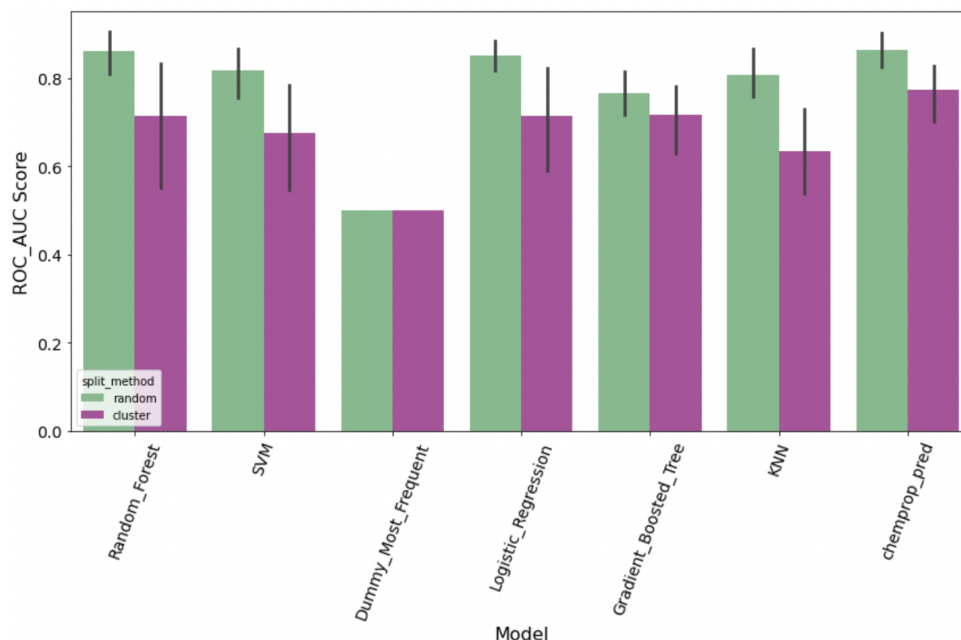


Figure 4. This figure shows a comparison between the random and clustered split ROC-AUC scores for the different models on the validation dataset. The results are aggregated across all 6 datasets to provide a generalized view.

Cross Validation Results

To explore the cross validation results, we chose to compare the ROC-AUC metric because it was not affected by the discrimination threshold and was the metric we utilized to optimize the Chemprop model during training. From Figure 5, we can see that generally Chemprop scored highest, even when factoring in the standard deviations. Interestingly, for the ClintTox dataset, Chemprop scored about the same as SVC, logistic regression and the gradient boosted tree. With the exception of the BACE data set, the others did have models that were close to the score of the Chemprop model from a cross validation perspective, suggesting that some of the simple models can still generalize well and they might be performant in many circumstances. Factoring in the standard deviations on some of the scores as well might make it difficult to differentiate them by ROC-AUC alone. The variation in the cross validation scores for the simple models, with values as high as 0.12 in some cases suggests that the data set splits might be more important in those cases than for Chemprop where these were <0.04. The one note here is that we were not able to perform scaffold or cluster based splitting during the

cross validation process, so the random presence of the scaffolds in the train/test splits might make these results overly optimistic. To account for this we utilize the results of the withheld validation data below. A future adaptation could work to incorporate a scaffold based splitting approach to the cross validation process to improve this moving forward.

	Gradient Boosted Tree	Dummy Most Frequent	Random Forest	Logistic Regression	KNN	SVC	Chemprop
Solubility	0.78 ± 0.1	0.5 ± 0.0	0.79 ± 0.09	0.85 ± 0.08	0.7 ± 0.07	0.83 ± 0.07	0.93 ± 0.04
HIV	0.69 ± 0.05	0.5 ± 0.0	0.74 ± 0.03	0.69 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.8 ± 0.02
Lipophilicity	0.72 ± 0.12	0.5 ± 0.0	0.68 ± 0.16	0.77 ± 0.1	0.63 ± 0.11	0.73 ± 0.09	0.86 ± 0.04
TOX21	0.82 ± 0.03	0.5 ± 0.0	0.84 ± 0.04	0.81 ± 0.03	0.72 ± 0.06	0.76 ± 0.04	0.87 ± 0.01
Clintox	0.78 ± 0.01	0.5 ± 0.0	0.74 ± 0.07	0.8 ± 0.07	0.58 ± 0.07	0.78 ± 0.07	0.79 ± 0.03
BACE	0.74 ± 0.09	0.5 ± 0.0	0.73 ± 0.11	0.74 ± 0.12	0.74 ± 0.08	0.7 ± 0.11	0.91 ± 0.02

Figure 5. The ROC-AUC scores resulting from the cross validation of the various models on their respective datasets.

Modeling Results on Withheld Validation Data

A major consideration for an experiment of this complexity, with multiple datasets and multiple models, is which scoring metric is most appropriate for model optimization on a given dataset. Scikit-learn offers over twenty different classification metrics to choose from. We decided to look at 3 of the more common metrics, accuracy, f1 score, and roc-auc. Each of these have their own pros and cons. Drug-like molecules, and especially the subset that have the specific therapeutic effect for a specific disease, are often an underrepresented population, as is the case for the majority of the datasets in this study. With this in mind, it is important to consider the information in Figure 1 (above) which highlights class proportionality when selecting metrics to measure model success. For this project, we explored how well each of these metrics differentiated the models from a dummy classifier that simply picked the majority class. We evaluated each of these metrics against the withheld validation data from the cluster based splitting results.

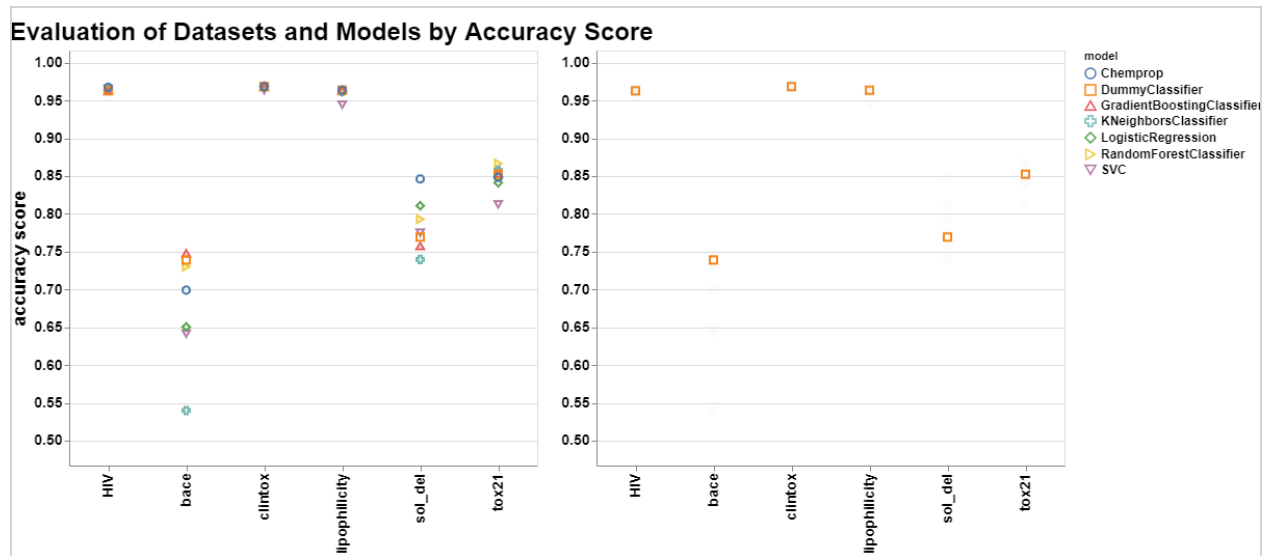


Figure 6: Comparison of dataset and model combinations by means of accuracy scores and with reference to a most frequent class prediction dummy classifier

The first metric we looked at was accuracy score (Figure 6), which represents the number of correct predictions over the total number. Accuracy score is easy to communicate to stakeholders, but in cases where there is significant class imbalance it over represents the majority class and does not properly represent model success. Figure 6 displays accuracy scores for different models and different dataset combinations and compares this against our sanity check with the dummy model predicting. Considering the class imbalance in all but the BACE dataset, it is unsurprising that accuracy scores do not significantly differentiate model success from the dummy classifier, and therefore is not an effective scoring metric for our datasets.

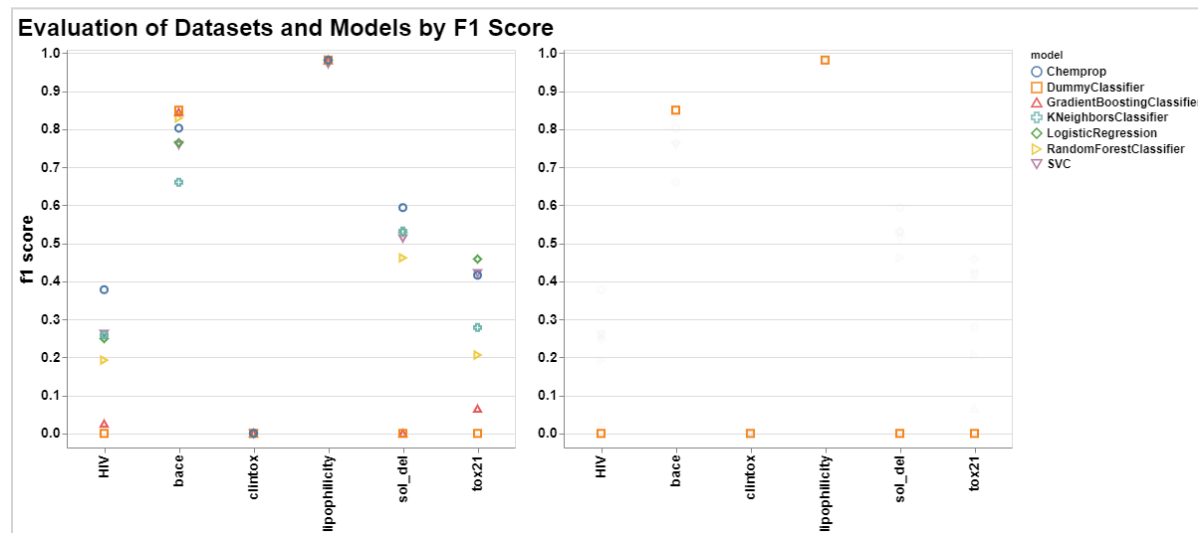


Figure 7: Comparison of dataset and model combinations by means of f1 scores and with reference to a most frequent class prediction dummy classifier

Figure 7 shows model results for different datasets represented by the F1 score, which is also known as the harmonic mean of recall and precision. This metric does a good job of representing model success as long as the positive class is the under-represented class. In the case of our Lipophilicity dataset, the negative class is the one that is significantly under-represented, and using this metric does not allow for differentiation of models trained on this dataset. In all other cases we can see more significant model differentiations. It is interesting that with the BACE dataset, all models performed equal to or worse than the dummy classifier. Chemprop differentiated itself from other models for the HIV and solubility datasets, and performed in the top cluster of models for TOX21, but was not the top model. In general, these models are all scoring rather low, which highlights the difficulty in these chemistry prediction problems.

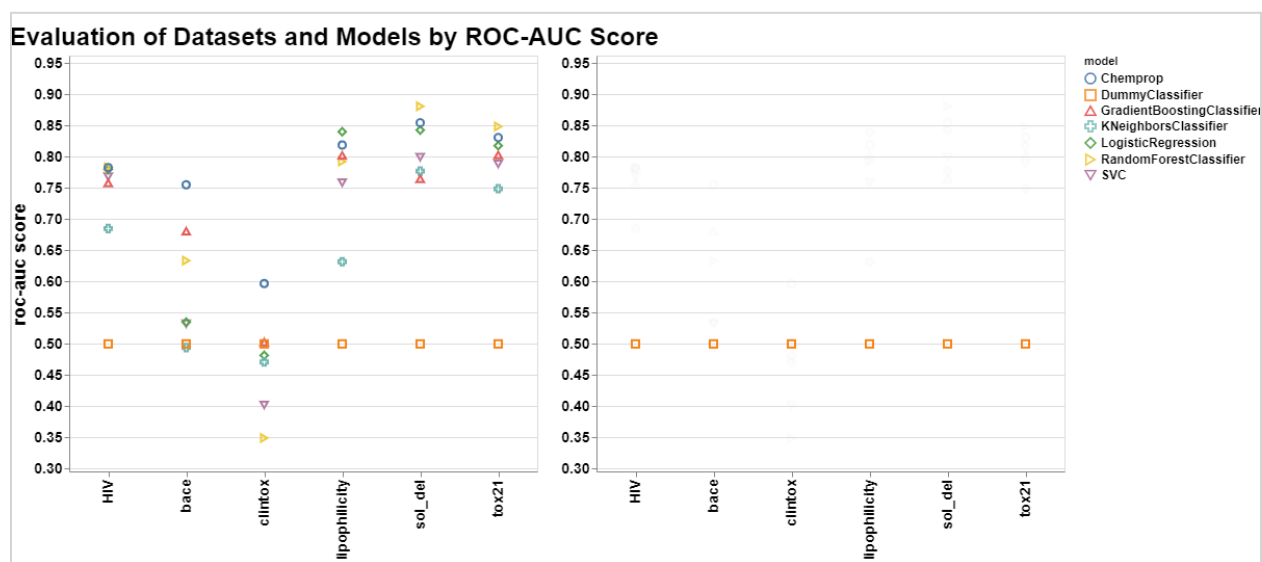


Figure 8: Comparison of dataset and model combinations by means of roc-auc scores and with reference to a most frequent class prediction dummy classifier

The metric we found most informative was roc-auc score (Figure 8), which seems to do a good job of differentiating model success from the dummy models. Models that score above 0.5 show positive differentiation of predictions from random guessing, with a roc-auc value of 1.0 being ideal. From the roc-auc scores, there was no clear model that was best across the datasets. Chemprop achieved considerably better scores for the BACE and clintox datasets, and competitive scores for all other models. Random Forest performed best for the Solubility and TOX21 datasets. Logistic Regression achieved the top roc-auc score for the lipophilicity dataset. In general, the KNN models seem to be low performing for these datasets. Clintox is an interesting dataset because for the majority of the models they performed worse than the dummy classifier suggesting they were not able to learn a signal from the noise. This is one case where Chemprop did something that no other model could. Overall, these model results suggest that there is no single best model for these datasets, but instead different models may still offer advantages in certain circumstances.



Figure 9: This figure compares model training and validation ROC-AUC scores across models for all datasets.

Another factor to consider in regards to model performance is overfitting. Figure 9 compares training and validation set prediction results by ROC-AUC score across all datasets, which if drastically different can be a sign that a model is overfit and therefore not generalizing well onto new data. Figure 9 reveals two major insights. First, the models where nearly every training score is almost perfect are the ones that are most prone to overfitting, and users should be extra cautious when optimizing these for real world data. Chemprop and Random Forest in particular stand out in this regard as there is almost no variance in the training scores, and they are very high scoring on training data and lower performance on the validation data. The GradientBoosted classifier appears to be the most guarded against overfitting, even after minimal parameter optimization. This may make it one of the easier models to get up and running.

While we did not explore hyperparameter optimization extensively, this would be a good place for future efforts. From Figure 9, we can see that there is still room for model improvement through tuning. Models that show a wide gap between their train and validation data can often be further generalized in terms of their learnings to increase the overall model output. Something that would be interesting to evaluate in future work would be a relative effort and time spent on hyperparameter tuning compared with relative increase in performance for each of these models. A “simple” model might no longer qualify for that category if to achieve maximum performance required significant effort tuning. KNN for instance stands out as a model where this could be the case given the large gap between training and validation scores.

BACE Results In-depth Review

In order to explore the differences between modeling techniques further, we chose to look closely at the BACE dataset due to its manageable size compared to the HIV dataset and the prediction of on-target potency is often the most critical task for many drug-discovery programs. In order to maximize performance, the hyperparameter optimization was performed using the Chemprop

algorithm that involves a Bayesian optimization approach to search over the various neural network parameter space. The hyperparameter optimization suggested that increasing the drop-out from 0 to 0.3 helped performance, suggesting that a reduction in the unoptimized model complexity might help improve generalizability. Interestingly, the other parameters of hidden-size and network depth both increased from 300 to 900 and 3 to 6 layers respectively. This suggests that adding additional model complexity could help with the performance. It should be noted that the hyperparameter optimization was performed to improve the roc-auc metric, so further exploration might yield other parameters that improve other metrics.

Following the updated training, if we explore the predictions for each of the models relative to the actual labels, shown in Figure 10, it is clear that many of the models are not performing particularly well. The Dummy Most Frequent classifier appears in dark green for most samples because it simply assigns with complete confidence the positive label for this dataset. The simpler models of Logistic Regression, Random Forest and Gradient Boosted Trees all show a lighter color, indicating that the models are not particularly confident in their decisions. The KNN classifier is clearly not a good choice in this situation as it has large sets that are not correctly predicted. Interestingly, the complex Chemprop models appear to do a better overall job based on the heatmap, however there are some clear groupings of molecules where all of the models appear to perform poorly.

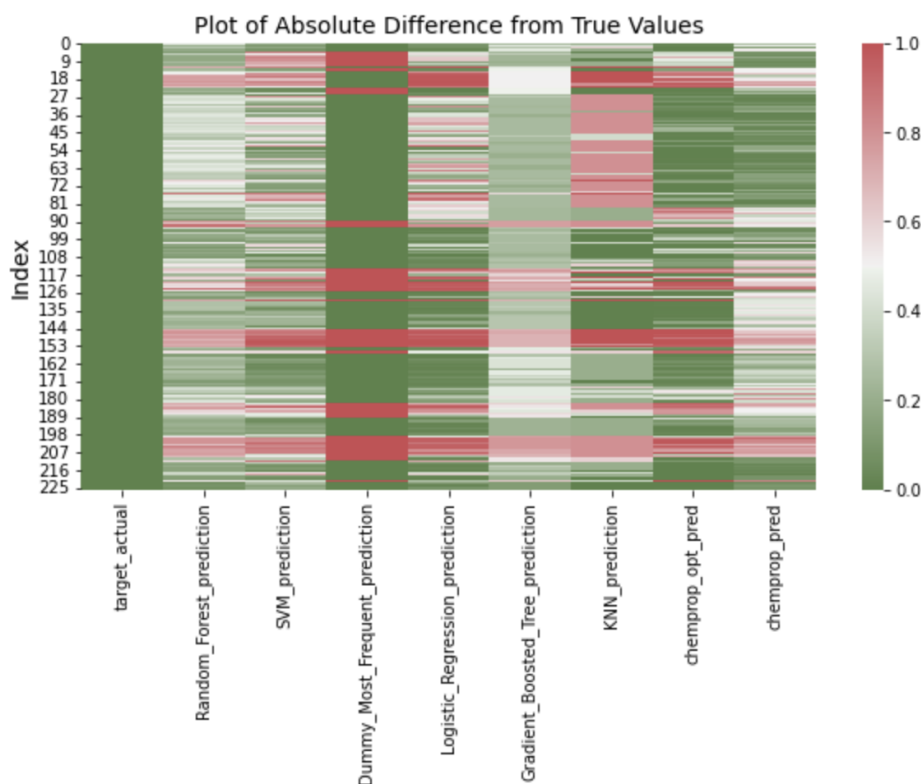


Figure 10. A heat map showing the absolute difference between the predicted probabilities and the actual target labels. Darker green colors indicate less difference from the true value, while red colors indicate a stronger in-correct prediction.

The other aspect of optimization that we then explored for the various different models, was how the decision threshold affected the true positive and false positive rate. To evaluate this across the different models, we plotted the ROC curve (Figure 11) for the different models. This curve can help determine how sensitive the prediction accuracy is to changing the threshold at which a prediction is considered to be positive or negative. The area under each curve is also the roc-auc metric that was discussed above. From Figure 11, it can be seen that both Chemprop models are relatively high performing, and from the simple models, the Gradient Boosted Tree is the best performing. Therefore, we selected the Optimized Chemprop model and the Gradient Boosted Tree as our models for further evaluation. This curve also helped us select a discrimination threshold for further scoring. To accomplish this we calculated the J-statistic for each threshold value and selected the maximum.^[17] However, these optimal thresholds were quite high for some of the models, such as the optimized chemprop which made it difficult to compare the models. Therefore, an intermediate discrimination threshold of 0.6 was selected, which seemed to strike a reasonable balance of true positives, which are of more interest in the drug discovery field to avoid missing potentially useful compounds.

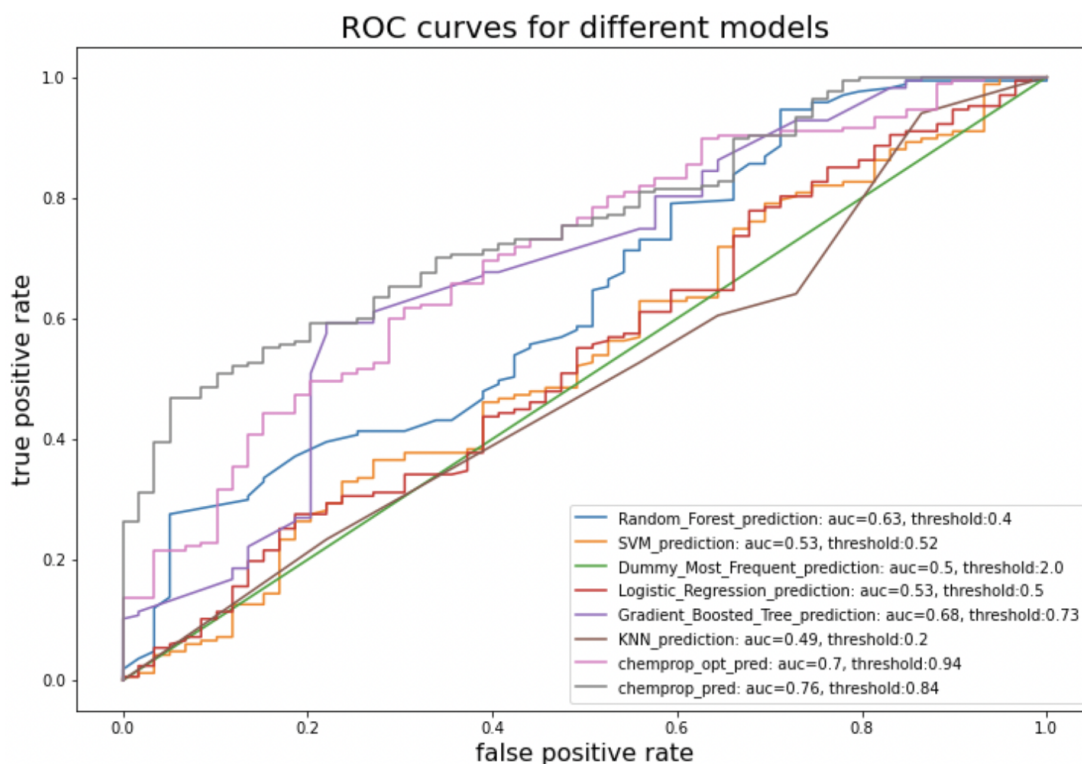


Figure 11. The Receiver Operating Characteristic curve for the models on the BACE dataset.

Once the discrimination threshold was determined, the series of scoring metrics were calculated in order to explore the model performance. From Figure 12, we can see that overall the scores are not particularly impressive. In part, since there is some class imbalance, the dummy prediction is relatively high scoring in terms of accuracy and f1. However, we can see that the Matthews Correlation Coefficient, similar to a Pearson's R, but for classification problems, is quite poor for the Dummy

model, as is the roc-auc. The optimized Chemprop model shows similar performance to the Dummy classifier, along with a higher Matthews correlation coefficient and drastically lower log-loss, meaning that the model is performing better than baseline, and is more confident in the results. We can also observe from the scores that the hyperparameter optimization for the Chemprop model helped in some aspects such as accuracy and f1 score, but hurt in others, such as log-loss or roc-auc. From the results it is not clear how much value the hyperparameter optimization added.

	Log Loss	ROC-AUC	Accuracy	F1	Matthew's Corr
Random Forest	0.57	0.63	0.62	0.72	0.12
SVC	0.71	0.53	0.64	0.76	0.09
Dummy Most Frequent	9.02	0.50	0.74	0.85	0.00
Logistic Regression	0.96	0.53	0.60	0.72	0.03
Gradient Boosted Tree	0.53	0.68	0.70	0.80	0.21
KNN	4.16	0.49	0.54	0.66	-0.04
Chemprop Optimized	0.78	0.70	0.74	0.83	0.28
Chemprop Default	0.48	0.76	0.69	0.78	0.23

Figure 12. Table showing the different BACE models and their respective scores on the validation data.

Figure 12 also allows us to compare some performance metrics between the simple models. The simple models did not appear to perform drastically better than the baseline Dummy model for accuracy or F1. However, in agreement with the roc-curves above, the Gradient Boosted Decision Tree and Chemprop models showed some of the best overall scores. Since many of the models were not particularly high-performing we wanted to explore the interpretation of these models to see if we could understand what could be happening as well as try to understand if there are any fundamental differences between the Gradient Boosted Tree and the optimized Chemprop model.

In examining some of the molecules that were predicted correctly or incorrectly, the example shown in Figure 13 is informative. The top molecule is inactive in the data, while the bottom two both belong to the positive class. Interestingly, the Gradient Boosted Tree model predicts all three to be active, while the Chemprop-optimized model predicts them all to be inactive. Given the only difference in the three structures is the location of the oxygen atom and the size of the ring on the left side, it is perhaps understandable that these would have very similar fingerprint based representations and therefore would be predicted similarly. This example highlights one of the key challenges in using 2D representations, where in 3D, the size of the ring might completely prohibit the molecule from binding the protein correctly. These subtle differences that have large impacts on the results is part of what makes chemical property prediction so challenging.

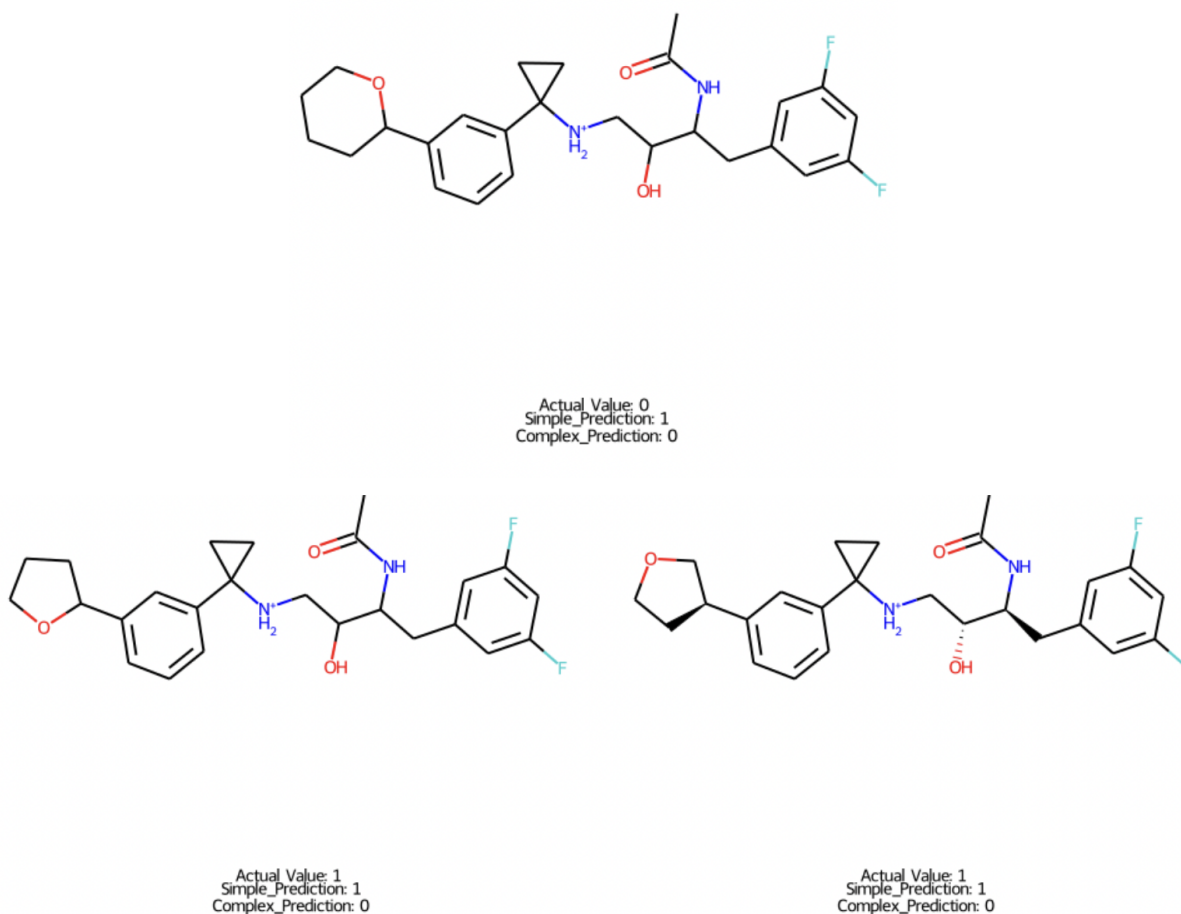


Figure 13. Example molecules from BACE dataset showing some of the subtlety that is often present in chemical structure to biological activity relationships.

Lastly, we wanted to try and understand to what extent the model interpretability might compare between the simple and complex models. In the chemical fingerprint based modeling space, interpretation is a very open and active question, with many papers only coming out in the last few years. [18, 19] However, many of these approaches involve more elaborate featurization and descriptors and typically utilize genetic algorithms or Monte Carlo simulations. For our example, we wanted to see how we could relate this to the circular fingerprints.

In order to explore the simple model interpretation. We extracted the feature importances and then ranked the top 20 values. These were then associated with certain positions in the fingerprint array using RDkit. [20] It should be noted that there is potential for some bit collisions, so the substructures might not be exactly the same for every molecule, but when we checked a handful of different molecules these structures consistently appeared. From Figure 14, we can see the top bits and then the associated substructures of these. Interestingly, some of these are quite simple, such as bit number 650, 623 and 904 which are simply a single oxygen, nitrogen or fluorine atom respectively. The highest importance bit was 444, which is simply a basic nitrogen atom. Other highly scored bits are

various parts of common substructures or ring systems. This result likely helps explain some of the above observations that the model is more learning general scaffolds, and without additional data might have a hard time distinguishing between subtle atomic features from a fingerprint alone. Bit 511 was present in this dataset, and was perhaps one of the more specific bits in that it requires a specific 5-membered ring to contain a nitrogen atom. The results from this exploration seem to suggest a higher importance of the features holistically over an individual few bits. The small nature of these might also suggest that exploring larger radius fingerprints could also contribute to a less localized interpretation of the chemical structure.

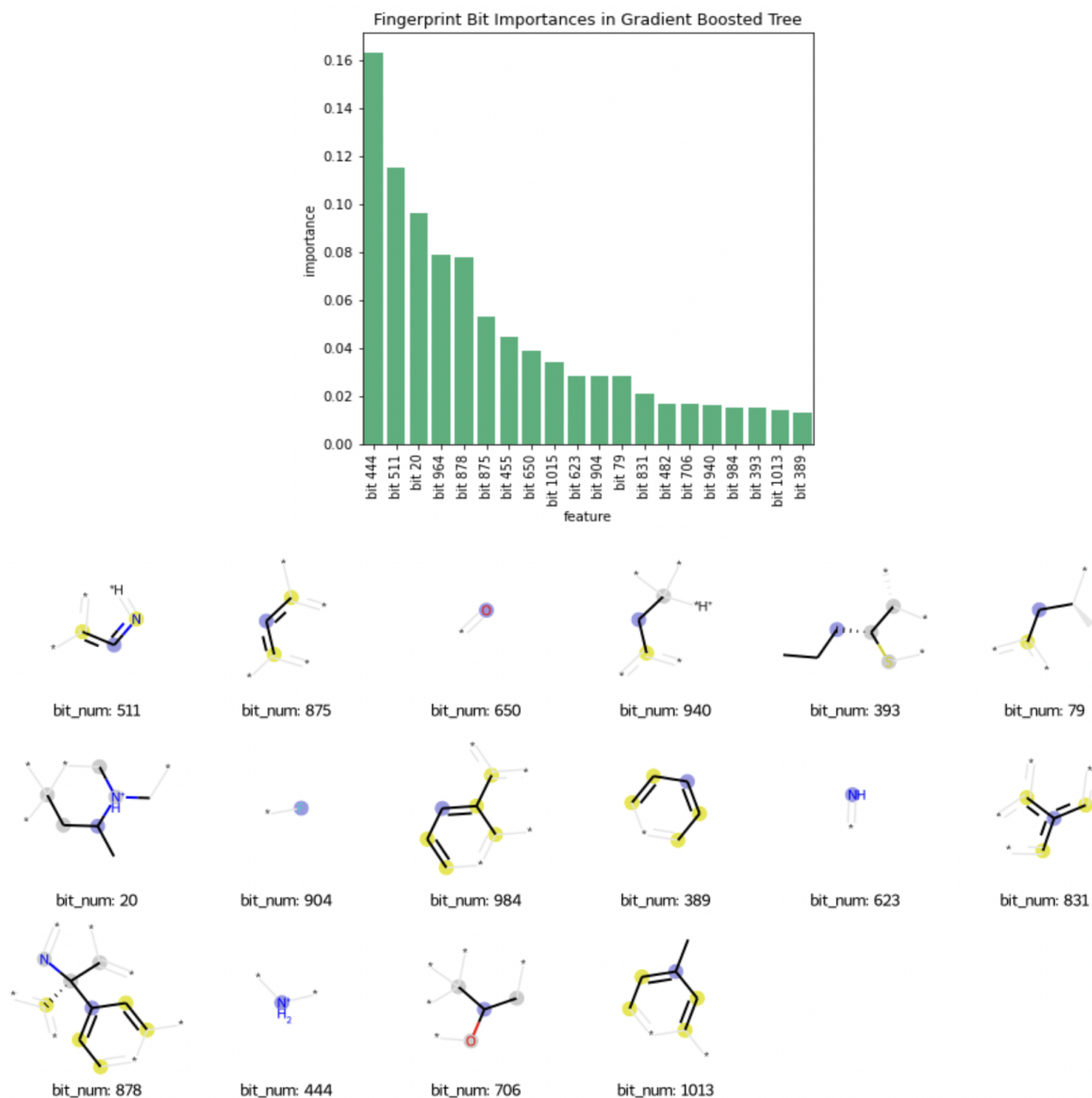


Figure 14. The feature importances and the associated fingerprint bits for the top gradient boosted tree predictions.

To explore how interpretability would work in a complex model space, we utilized the Chemprop interpret function. This function utilizes a Monte Carlo simulation over the different models and substructure types to generate a minimum substructure leading to a positive result as well as a measure of how much that substructure contributes to the prediction. In practice, this function was of limited usefulness in our hands as the simulation took almost 30 minutes to run on each compound. From the interpretation results, shown in Figure 15, the substructure that was contributing significant activity to the prediction, was also present in several of the other molecules that were predicted to be inactive. It is possible that this is a result of the model not being particularly performant, or that there were other more negatively correlated features on these molecules as well. From our work here, we did not find the fingerprint based interpretability of either the simple or complex models to add significant insight to the problem of BACE activity prediction in such a way that it would be easily actionable.

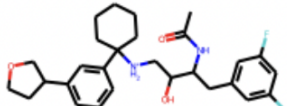
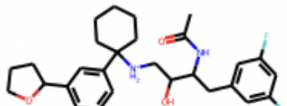
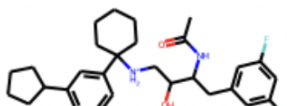
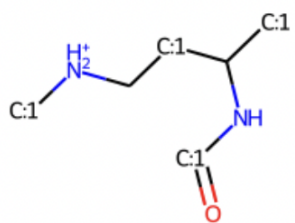
	Structure	prediction	substructure	substructure_activity
0		0.104	NaN	NaN
1		0.193	NaN	NaN
6		0.633		0.961

Figure 15. The results of a Chemprop interpretation of some of the molecules. The row highlighted in blue shows the only positive example here, which is why it is the only one with a shown substructure.

Conclusion

Next Steps

While these results are interesting, we recognize that there are many future directions that would be relevant to explore. From this work, we were only able to perform minimal hyperparameter tuning, and only on a single dataset. It would be useful to explore whether the results from the BACE dataset, where Chemprop performance did not improve significantly with the tuning, holds true for other datasets. While some parameters were explored for the simple models, these could also benefit from additional tuning of the parameters. In particular, the depth and minimum sample parameters in the tree-based methods can often modulate overfitting and might help improve performance further. The support vector classifier could also benefit from tuning the kernel function and the regularization parameter since its performance was lower than expected overall based on the previous literature.

Another opportunity for future work would be to expand the number of complex models. While the setup and configuration of these can take time, it would be interesting to understand how these more complex models relate to one another. For instance, other neural network (NN) modeling approaches such as Deepchem [\[13\]](#) or Few-Shot Mol [\[12, 21\]](#) would be relevant to compare. While the software is often proprietary and costly, the results could also be compared to 3-dimensional docking [\[22\]](#) or quantum mechanical based Free Energy Perturbation (FEP). [\[23\]](#) The idea of multi-task learning, where representations learned on one task can then help improve predictions for another task is also implemented in chemprop, but was not tested for this project, but could have an impact on the performance.

Lastly, the featurization of the molecules is of the utmost importance to model performance. There are some opportunities to leverage calculated molecular properties, such as through the descriptastorus package [\[24\]](#), which might help improve generalizability. Moreover, other fingerprinting techniques or count based fingerprinting [\[25\]](#) could also be informative to compare. While Chemprop provides a deep-learning approach to the chemical representation with the graph-based approach, it could be informative to consider undirected graph based or other deep-learning approaches as well. While difficult computationally, quantum mechanical descriptions based on molecular dynamics are also becoming increasingly popular and less costly to obtain.

While there are many directions to explore further, it could be of interest to consider various ensemble methods where one does not need to look for a “best” model but instead can capture the benefits of several modeling approaches at once. This is further supported by the iterative and uncertain nature of drug discovery where many experiments are still likely to be run, so even if models can enrich the target molecules in ones that are higher value it will help discover new drugs faster and with reduced cost.

Conclusion

From the above results, it is clear that data science in the drug discovery context is not a trivial question. While it is hard to say with confidence whether simple or complex models are “better” we have found results that in certain cases, a complex model such as Chemprop can be higher performing than the simple model types. However, there is also clear evidence that Chemprop can also overfit and for other datasets it is not as high performing as some of the simpler models. Combined with some of the complexity to setup, train and run a Chemprop model there is also strong justification to utilize simpler modeling approaches at first as well. Furthermore, within difficult problems faced in drug discovery, models have not yet consistently demonstrated very high performance as evidenced by our predictions on the BACE target. As a result, careful review and curation of the results is important in addition to looking at aggregated scores. Our results between the different data splitting techniques also highlight the need for caution when moving to newer chemical structures that the models might not be as performant on. Overall, we hope that this framework has proven useful for future datasets and predictions as they can be easily integrated into the pipeline. We also believe that this approach can provide additional context and is in-line with the views of many in the chemistry community of utilizing a range of common and standardized datasets to provide context when developing new modeling techniques to better understand where and when to apply them. [1]

Works Cited

1. Andreas Bender et al., “Evaluation Guidelines for Machine Learning Tools in the Chemical Sciences,” *Nature Reviews Chemistry* 6, no. 6 (June 2022): 428–42, <https://doi.org/10.1038/s41570-022-00391-9>.
2. “Datasets,” accessed August 21, 2022, <https://moleculenet.org/datasets-1>.
3. “Chemprop Documentation,” Swanson et al. accessed August 21, 2022, <https://chemprop.readthedocs.io/en/latest/>
4. Kevin Yang et al., “Analyzing Learned Molecular Representations for Property Prediction,” *Journal of Chemical Information and Modeling* 59, no. 8 (August 26, 2019): 3370–88, <https://doi.org/10.1021/acs.jcim.9b00237>.
5. “What Is PIC50? - Collaborative Drug Discovery Inc. (CDD),” July 31, 2018, <https://www.collaborativedrug.com/what-is-pic50-2/>.
6. “Tox21,” accessed August 21, 2022, <https://tripod.nih.gov/tox/assays>.
7. Ketan T. Savjani, Anuradha K. Gajjar, and Jignasa K. Savjani, “Drug Solubility: Importance and Enhancement Techniques,” *ISRN Pharmaceutics* 2012 (July 5, 2012): 1–10, <https://doi.org/10.5402/2012/195727>.
8. “Lipinski’s Rule of Five,” in *Wikipedia*, August 2, 2022, https://en.wikipedia.org/w/index.php?title=Lipinski%27s_rule_of_five&oldid=1101947937.
9. “Chemprop.Data.Scaffold — Chemprop 1.5.2 Documentation,” accessed August 21, 2022, <https://chemprop.readthedocs.io/en/latest/modules/chemprop/data/scaffold.html>.
10. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — Umap 0.5 Documentation,” accessed August 21, 2022, <https://umap-learn.readthedocs.io/en/latest/>.
11. Benjamin Obi Tayo Ph.D, “Simplicity vs Complexity in Machine Learning — Finding the Right Balance,” Medium, November 11, 2019,

- <https://towardsdatascience.com/simplicity-vs-complexity-in-machine-learning-finding-the-right-balance-c9000d1726fb>.
12. “FS-Mol: A Few-Shot Learning Dataset of Molecules,” Python (2021; repr., Microsoft, August 17, 2022), <https://github.com/microsoft/FS-Mol>.
 13. “DeepChem,” accessed August 21, 2022, <https://deepchem.io/>.
 14. “Classifier Comparison,” scikit-learn, accessed August 21, 2022, https://scikit-learn/stable/auto_examples/classification/plot_classifier_comparison.html.
 15. Roman M. Balabin and Ekaterina I. Lomakina, “Support Vector Machine Regression (LS-SVM)—an Alternative to Artificial Neural Networks (ANNs) for the Analysis of Quantum Chemistry Data?,” *Physical Chemistry Chemical Physics* 13, no. 24 (2011): 11710, <https://doi.org/10.1039/c1cp00051a>.
 16. Hongdong Li, Yizeng Liang, and Qingsong Xu, “Support Vector Machines and Its Applications in Chemistry,” *Chemometrics and Intelligent Laboratory Systems* 95, no. 2 (February 2009): 188–98, <https://doi.org/10.1016/j.chemolab.2008.10.007>.
 17. Jason Brownlee, “A Gentle Introduction to Threshold-Moving for Imbalanced Classification,” *Machine Learning Mastery* (blog), February 9, 2020, <https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/>.
 18. Matveieva, M., Polishchuk, P. Benchmarks for interpretation of QSAR models. *J Cheminform* 13, 41 (2021). <https://doi.org/10.1186/s13321-021-00519-x>
<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00519-x>
 19. Shoombuatong, Watshara, Philip Prathipati, Wiwat Owasirikul, Apilak Worachartcheewan, Saw Simeon, Nuttapat Anuwongcharoen, Jarl E. S. Wikberg, and Chanin Nantasenamat. “Towards the Revival of Interpretable QSAR Models,” 24:3–55. DORDRECHT: Springer Nature, 2017.
https://doi.org/10.1007/978-3-319-56850-8_1.
 20. Greg Landrum, “RDKit: Using the New Fingerprint Bit Rendering Code,” *RDKit* (blog), October 16, 2018, <http://rdkit.blogspot.com/2018/10/using-new-fingerprint-bit-rendering-code.html>.
 21. Megan Stanley et al., “FS-Mol: A Few-Shot Learning Dataset of Molecules,” n.d., 13. <https://openreview.net/forum?id=701FtuyLIAd> Accessed August 2022.
 22. “Drug Discovery Workflows | Schrödinger,” accessed August 21, 2022, <https://www.schrodinger.com/platform/smdr/drug-discovery-workflows#hit-identification>.
 23. “Free Energy Perturbation (FEP): Another Technique in the Drug Discovery Toolbox,” accessed August 21, 2022, <https://www.cresset-group.com/about/news/fep-drug-discovery-toolbox/>.
 24. Brian Kelley, “DescriptaStorus,” Python, accessed August 21, 2022, <https://github.com/bp-kelley/descriptastorus>.
 25. Gregory Landrum, “Fingerprints in the RDKit,” n.d., 23. , accessed August 21, 2022, https://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf