CS230

Deep Learning Approaches for Predicting Drug Mechanisms of Action

(Healthcare)

Toren Fronsdal
Department of Statistics
Stanford University
toren@stanford.edu

Kai Fronsdal
Department of Computer Science
Stanford University
kaifronsdal@stanford.edu

Abstract

Discovering a drug's mechanisms of action (MOAs) is a critical first stage in the drug discovery process; however, this process is typically costly and time-intensive. Using a novel dataset of gene expressions and cell viability data, we evaluate algorithms to predict a drug's mechanism of action. This problem is a multi-label classification problem with a sparse, high-dimensional outcome space. We examine models that take advantage of the correlation structure of the 206 labels (MoAs). In addition to benchmarking gradient boosting decision trees (GBDTs) and neural networks for this problem, we propose a "sequential" neural network model, in which we first predict the type of interaction (e.g. agonist, inhibitor, antagonist, etc.) and use this prediction in the second-stage model to predict the MoA, explicitly leveraging the correlation across labels with the same type of interaction. We find that neural network approaches dominate GBDTs, and that the sequential neural network performs best of all the models we benchmarked.

1 Introduction

In an effort to accelerate drug discovery efforts, researchers have begun to exploit new computational methods rather than rely on more serendipitous means. Simulating and predicting the biological effects of potential drugs rather than testing each one directly may save researchers enormous amounts of time and resources. These computational methods show most promise in screening potential drugs to reduce the number of chemical combinations that are developed and tested. One important biochemical interaction of a drug is the mechanism of action, which identifies the molecular target to which the drug compound binds as well as the action that occurs there. For example, one drug may bind to a specific enzyme or receptor and inhibit this target's function, whereas another drug may activate the target. In recent years, efforts have been made to use machine learning techniques to predict the mechanism of actions that result from a drug. Since the process of discovering the mechanism of action directly is both costly and time-intensive, accurately predicting the mechanism of action can help accelerate scientific advancements to help treat diseases.

Formally, mechanism of action prediction is a multi-label classification problem, as each drug may have multiple mechanisms of action associated with it. In this paper, we will explore the application of deep learning techniques to this prediction problem. Using data on gene expressions and cell viability measures, which are easy and cheap for researchers to collect, we will build predictive models to tackle this problem and will benchmark the accuracy of different models in predicting mechanisms of action. Given that traditional machine learning methods like random forests and gradient boosted trees have tended to outperform deep learning techniques with tabular data, most research into predicting mechanisms of actions has involved these traditional machine learning models. However, we hypothesize that deep learning models may provide more accurate predictions for this specific problem. This is a multi-label classification problem with a sparse, high-dimensional outcome

CS230: Deep Learning, Winter 2021, Stanford University, CA.

space. The key challenge of this task will be leveraging the fact that a drug's mechanisms of action are highly correlated with one another. The best models will thus take into account this correlation when making predictions, and deep learning models that simultaneously predict all mechanisms of action can do just that. We propose a "sequential" neural network that first predicts a drug's type of interaction (e.g., inhibitor, agonist, antagonist, etc.) and then uses the predicted interactions to predict the mechanism of action. We find that our neural network approaches outperform gradient boosting methods at this prediction problem.

2 Related Work

Methods for predicting chemical effects are generally either similarity-based or feature-based approaches. Similarity-based methods are based off the assumption that similar compounds will have similar effects. This includes nearest neighbor algorithms like Ajmani et al. [2006] and Cao et al. [2012] as well as support vector machines like Darnag et al. [2010]. Feature-based methods include linear models like Sagardia et al. [2013] and random forests like Bauer et al. [2018]. However, the problem with feature-based methods is that they require deep insights into chemical and biological properties such as molecular interactions, reactions, and metabolic changes. Deep learning on the other hand allows us to automatically find important features and interactions between them.

Deep learning has also been applied to other chemical prediction problems. For instance Mayr et al. [2016] found that "deep learning excelled in toxicity prediction and outperformed many other computational approaches like naive Bayes, support vector machines, and random forests."

Related work in MoA prediction includes Bauer et al. [2018], a decision tree using structural alerts (high chemical reactivity molecular fragments or fragments that can be transformed by enzymes into fragments with high chemical reactivity) achieving an accuracy of 92.2%, and Warchal et al. [2019], a gradient boosted tree classifier achieving an accuracy of 80%. Bauer et al. [2018] only uses 301 chemicals in their testing set. We believe that our larger dataset will enable a deep learning model to both take into account more complicated structures in the data such as correlations in MoAs as well as generalize better.

3 Data

The dataset we will be using for this project is provided by the Connectivity Map, a joint project created by the Broad Institute, the Laboratory for Innovation Science at Harvard, and the NHS LINCS program.[1] This paper is the first publication analyzing this dataset to predict mechanisms of action. The dataset is in a tabular form and consists of a total of 39,650 observations, with an average of six observations per drug, and 206 mechanisms of action annotations per observation. A mechanism of action is enzyme or protein that is affected by a drug in conjunction with the action that occurs there (e.g., acetylcholine receptor agonist, adenylyl cyclase activator, and tubulin inhibitor). Each unit of observation is a drug-dose-time combination, where dose and time are specific to one of the six experiments performed on each drug. The inputs relevant for predicting the mechanisms of action are gene expressions and cell viability measures as well as experiment specific information (duration of treatment and dosage). There are 772 gene expression features and and 100 cell viability features per observation. The diagram below illustrates 6 observations for a given drug.

Drug	Time	Dose	Gene Expressions				Cell Viability			
			G1	G2	...	G772	C1	C2	...	C100
A	24	D1	-0.854	0.276	...	-1.074	0.898	-1.994	...	0.751
A	48	D1	0.336	0.462	...	-0.927	0.816	0.632	...	0.090
A	72	D1	0.198	0.533	...	-0.232	0.989	-0.651	...	-0.464
A	24	D2	-1.660	0.256	...	-1.681	0.137	0.142	...	0.172
A	48	D2	1.762	1.101	...	1.828	0.704	0.631	...	-1.076
A	72	D2	1.594	0.951	...	0.828	0.386	-0.036	...	1.031

The gene expression and cell viability features are normalized by comparing to appropriate controls for the same plate experiments and normalized using luminex invariant set (LISS) normalization and

[1] An overview of the data is provided by the Laboratory for Innovation Science here.

quantile normalization. See the Broad Institute's Connectopedia for more on the data processing steps. We perform additional data processing by performing one-hot encoding for the time feature.

4 Methodology

We benchmark multiple learning algorithms and see which provides the greatest predictive accuracy. Despite algorithms like gradient boosting and support vector machines typically dominating deep learning algorithms for problems with tabular data, we expect that the reverse will be true for this particular problem since neural networks will allow the model to capture and take advantage of the correlations across labels as we outlined above. Thus, we compare the performance of gradient boosting decision trees—using the gradient boosting framework LightGBM—to the performance of neural network models.

We use approximately 85 percent of the data for the training set and the remaining 15 percent of the data for the test set. Each drug is represented six times in the data, for each drug-dose-time combination, so we stratify the train-test split based on drugs to prevent data leakage in which the same drug appears in both the train set and the test set. For the training set, we perform model selection and tuning using five-fold cross validation. Similarly to the train-test split, the five folds are split with stratification by drugs to prevent data leakage across folds. We use the five-fold cross validation to search over model architectures and hyperparameters using grid search. For each model we build, we outline briefly the hyperparameters we tuned below. A full list of hyperparameters we tuned, with values we searched over, can be found in Appendix A.

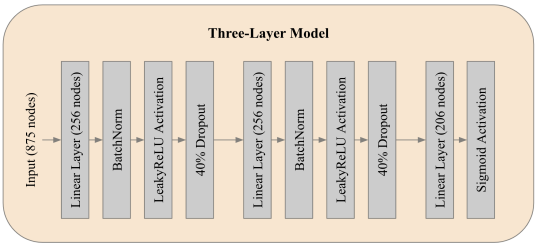
To evaluate the accuracy of our algorithms, we calculate the mean negative log loss across all 206 labels on the test set. Let N be the number of labels (mechanisms of action) and let M be the number of observations. Then we minimize

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\left[y_{i,j}\log(\hat{y}_{i,j}) + (1-y_{i,j})\log(1-\hat{y}_{i,j})\right].$$

4.1 Models

Gradient Boosting Decision Trees — We use the gradient boosting framework LightGBM (Ke et al. [2017]) to estimate 206 separate gradient boosting decision trees, one for each label in this multi-label problem. For each of the 206 models, we ensemble 100 trees. By grid search cross validation, we selected a learning rate of 0.01, a maximum number of leaves of 50, and a maximum tree depth of 4. The key disadvantage of this approach is that these models each predict a separate label and thus cannot share information across labels. Further, the hyperparameters were turned to minimize the average negative log loss across all models; tuning each model separately may yield superior results.

Three-layer Neural Network — Overall, we found that shallower nets performed better than deep ones. Ultimately, the best neural network model selected based on grid search cross validation was one with two hidden layers and 256 nodes in each hidden layer. This model was trained for 20 epochs with a batch size of 128, a learning rate of 0.01, weight decay of 1e-5, and a leaky ReLU activation function for all but the last layer, which has a sigmoid activation function. Each hidden layer was followed by batch normalization and a dropout with a dropout rate of 0.4. We use an Adam optimizer. Further, we allowed for dynamic learning rate reductions by a factor of 0.1 if there are 10 epochs for which the loss improves by no more than 1e-4.



Three-layer Neural Network with Label Smoothing — Given that there is a significant class imbalance in the labels, one method for ensuring that predictions are not over-confident and will better

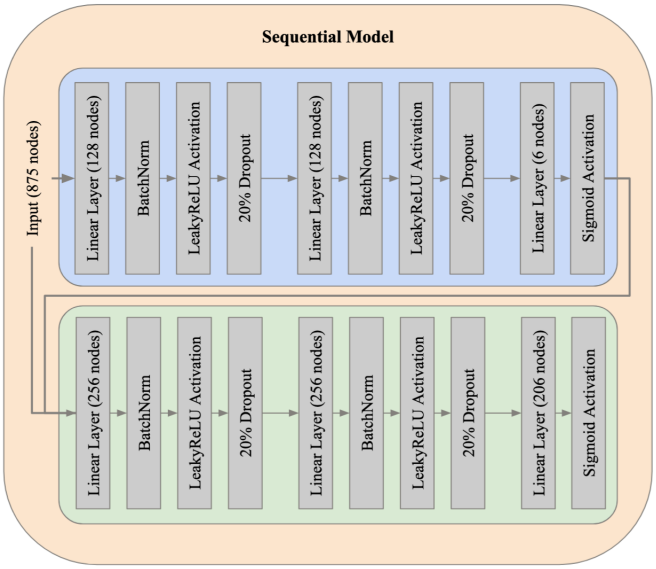
generalize to newly developed drugs is label smoothing. By artificially softening the targets, smoothing the labels can help prevent the network from becoming over-confident (Müller et al. [2019]). We benchmark a model that includes label smoothing as an additional form of regularization. For a label smoothing of parameter α and given that each label has two classes (0 or 1), the ith smoothed label is given by

$$y_i^{LS} = y_i(1 - \alpha) + \alpha/2.$$

Of the potential smoothing parameters we searched over, the optimal smoothing parameter was $\alpha = 0.001$. Grid search cross validation selected the identical parameters for this model as it did for the three-layer neural network model without label smoothing.

Sequential Neural Networks — Lastly, we propose a novel approach to predicting mechanisms of action in which we construct a two-stage sequential neural network. We developed this model structure to leverage the fact that there is particularly high correlation across labels with the same type of interaction (e.g., activator, inhibitor, etc.) regardless of the location of the interaction. Thus, in the first stage of this two-stage model we predict the type of interaction to occur at *any* enzyme or protein. The six labels predicted in the first-stage model are "activator," "agent," "agonist," "antagonist," "blocker," "inhibitor," and "other," where "other" encapsulates any form of interaction that appears only once across all 206 mechanisms of action. We then use this prediction as an additional input into the second-stage neural network model, which predicts the full mechanisms of action.

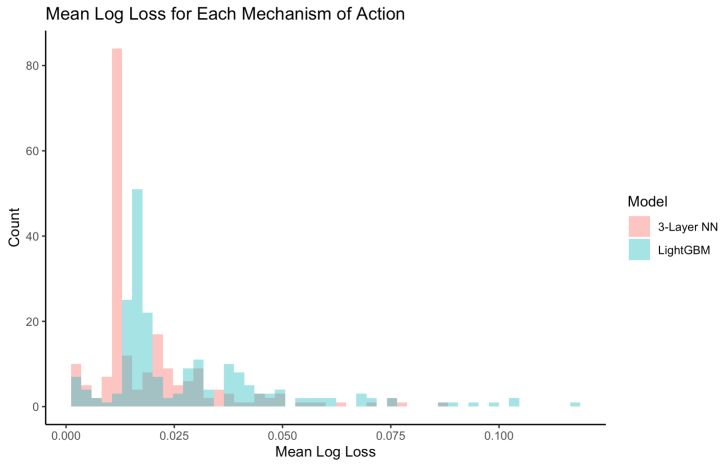
Like the previous model, both stages of this two-stage model also used label smoothing with parameter $\alpha = 0.001$. Both the first and the second stage models have three layers. The first-stage model was trained for 20 epochs with a batch size of 128, a learning rate of 0.01, weight decay of 1e-4, and a leaky ReLU activation function for all but the output layer, which has a sigmoid activation function. The two hidden layers have 128 nodes each while the output layer has six (for each of the six labels). Each hidden layer is followed by batch normalization and a dropout with dropout rate 0.2. The second-stage model has the same parameters as the regular three-layer neural network outlined above, with the exception of the dropout rate, which becomes 0.2.



5 Results

As we hypothesized, the deep learning approaches outperformed gradient boosting. The gradient boosting decision trees performed the worst of the models we benchmarked. As highlighted in the histogram below, gradient boosting decision trees resulted in large outliers in the mean negative log loss for some mechanisms of action. This longer right tail can be attributed to that fact that there was little predictive signal from the gene expression and cell viability data for some mechanisms of action. However, it appears that when sharing information across mechanisms of action, predictions for these outliers improve considerably. The gains to the incorporation of correlations across labels with the

neural network approaches are primarily driven by reduction in the loss for the mechanisms of action that are hardest to predict.



The mean negative log loss is presented for each of the models we benchmarked below.

Model	**CV Loss**	**Test Loss**
LightGBM	0.02098	0.02123
3-Layer NN	0.01781	0.01833
3-Layer NN w/ Label Smoothing	0.01769	0.01828
Sequential 3-Layer NN w/ Label Smoothing	**0.01765**	**0.01802**

All three neural network models performed significantly better than the gradient boosting decision tree model. The regular three-layer neural net had a 13.6 percent lower loss than the LightGBM model. The label smoothing lowered the test loss slightly relative to the neural network without label smoothing. The best performing model was the sequential model, with a test loss of 0.01765, slightly better than the three-layer model with label smoothing and 15.1 percent lower than the loss for the LightGBM model.

Leveraging the correlation structure across labels appears to be the primary driver of improvements for these models. While all of the neural networks implicitly encode any correlations across labels, our sequential model explicitly leverages the fact that there is a correlation in the type of interactions occurring across enzyme and protein targets. Adding this two-stage structure to the neural network model was the primary contribution of this paper.

6 Conclusion

Overall, the results in this paper shows the promise of deep learning techniques in aiding the complex and costly drug discovery process. These deep learning algorithms are particularly useful when researchers want to quickly predict the effect of a drug on a large number of enzymes and proteins simultaneously. Future research should harness more data from a larger sample of drugs and a larger number of mechanisms of actions. Further, future research should explore the use of deep learning for predicting mechanisms of action with other feature variables. While gene expressions and cell viability data are easy and cheap to collect, other information can certainly be used to predict the mechanisms of action of a drug. While the models presented in this paper are not perfectly accurate, these algorithms may provide enough signal to allow researchers to better target their future drug research.

7 Contributions

The Data and Methodology sections were written by Toren Fronsdal; the Related Work and Results sections were written by Kai Fronsdal; and the Introduction and Conclusion were written jointly. The code and the presentation were written as a joint effort. The data was acquired by Toren Fronsdal.

8 Code

The code for this paper is available at this link: https://code.stanford.edu/toren/cs-230-project

References

Subhash Ajmani, Kamalakar Jadhav, and Sudhir A. Kulkarni. Three-dimensional qsar using the k-nearest neighbor method and its interpretation. *Journal of Chemical Information and Modeling*, 46(1):24–31, 2006. doi: 10.1021/ci0501286. URL https://doi.org/10.1021/ci0501286. PMID: 16426036.

Franklin J. Bauer, Paul C. Thomas, Samuel Y. Fouchard, and Serge J.M. Neunlist. High-accuracy prediction of mechanisms of action using structural alerts. *Computational Toxicology*, 7:36 – 45, 2018. ISSN 2468-1113. doi: https://doi.org/10.1016/j.comtox.2018.06.004. URL http://www.sciencedirect.com/science/article/pii/S2468111318300185.

Dong-Sheng Cao, Jian-Hua Huang, Jun Yan, Liang-Xiao Zhang, Qian-Nan Hu, Qing-Song Xu, and Yi-Zeng Liang. Kernel k-nearest neighbor algorithm as a flexible sar modeling tool. *Chemometrics and Intelligent Laboratory Systems*, 114:19–23, 2012. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2012.01.008. URL https://www.sciencedirect.com/science/article/pii/S0169743912000196.

Rachid Darnag, E.L. Mostapha Mazouz, Andreea Schmitzer, Didier Villemin, Abdellah Jarid, and Driss Cherqaoui. Support vector machines: Development of qsar models for predicting anti-hiv-1 activity of tibo derivatives. *European Journal of Medicinal Chemistry*, 45(4):1590–1597, 2010. ISSN 0223-5234. doi: https://doi.org/10.1016/j.ejmech.2010.01.002. URL https://www.sciencedirect.com/science/article/pii/S022352341000036X.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017. URL https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. Deeptox: Toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016. ISSN 2296-665X. doi: 10.3389/fenvs.2015.00080. URL https://www.frontiersin.org/article/10.3389/fenvs.2015.00080.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019. URL https://papers.nips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf.

Iban Sagardia, Rubén H. Roa-Ureta, and Carlos Bald. A new qsar model, for angiotensin i-converting enzyme inhibitory oligopeptides. *Food Chemistry*, 136(3):1370–1376, 2013. ISSN 0308-8146. doi: https://doi.org/10.1016/j.foodchem.2012.09.092. URL https://www.sciencedirect.com/science/article/pii/S0308814612014963. ASSET 2011.

Scott J. Warchal, John C. Dawson, and Neil O. Carragher. Evaluation of machine learning classifiers to predict compound mechanism of action when transferred across distinct cell lines. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 24(3):224 – 233, 2019. doi: 10.1177/2472555218820805. URL https://doi.org/10.1177/2472555218820805. PMID: 30694704.

A Grid Search and Hyperparameter Tuning

For the LightGBM model, the grid search searched over the following parameter values:

- Maximum number of leaves: 10, 50, 100
- Maximum tree depth: 3, 4, 10
- Learning rate: 0.01, 0.001

For the LightGBM model we were computationally constrained from considering more parameter values given that 206 different models must be trained, one for each label.

For all neural network models, the grid search searched over the following parameter values:

- Number of layers: 3, 4, 5
- Epochs: 15, 20, 30, 50
- Batch normalization: Yes, No
- Number of nodes per hidden layer: 128, 256, 512
- Dropout: 0.0, 0.2, 0.3, 0.4
- Learning rate: 0.01, 0.001
- Weight decay: 1e-4, 1e-5, 1e-6

The parameters that were not searched over included: parameters for the learning rate scheduler, the batch size, activation functions, and the optimizer.