An Active Learning Approach to Predicting Potency of Small Molecules for Antibacterial Screening

Sriram Kidambi, Daniel Lee, Mirudhula Mukundan

Abstract

Drug discovery and antibiotic screening are time-consuming and cumbersome wet-lab procedures. Screening for potency in general involves several small molecules and working with each of them at the lab-scale is an expensive ordeal. Conducting computational analysis before the lab screening can not only decrease the number of potential target molecules but also reduce expenditure by a huge margin. In this project, we explore the usage of Passive and Active Machine Learning approaches to identify potent compounds against the bacterium, Burkholderia *cenocepacia*, that is commonly known to be an opportunistic pathogen, mainly affecting immunocompromised patients suffering from cystic fibrosis.

1 Introduction

1.1 Active Learning

Active learning is a technique used in machine learning that iteratively selects a subset of data from a large sample of unlabeled data to be labeled via a domain expert or oracle, and then uses the newly labeled points to train machine learning models. In this way, models can achieve a high performance while utilizing fewer labeled data. Within active learning, there are multiple methods and algorithms to select the data subset to be labeled, with each giving different performance results.

1.2 Significance

The significance of active learning lies in the efficient computational cost as the method does not require all of the data to be labeled. Labeling a data sample may cost hundreds of thousands of dollars and therefore active learning utilizes only the most important samples to label, thereby reducing financial/computational expenses while achieving high performance results.

1.3 Objective

The objective of this project is to analyze the power of active learning on a large dataset by exploring different known active learning methods on a binary classification task. We will implement at least three different active learning methods and compare their classification accuracies against one another, as well as the baseline passive learning method (utilizing the entire dataset at once to train a model).

2 Background

Computational approaches to drug discovery has been around for a while, but it is only in recent years that machine and deep learning approaches have been implemented to narrow down the search space and make appropriate decisions for potential drug molecules. Several research have been conducted using the molecular structures of the small molecules as input to machine learning models, either as graphs or as vectors. The different techniques used and the types of inputs to each model has been well reviewed by Carracedo-Reboredo *et al.*[2] and Jukic *et al* [5]. A popularly used method for drug discovery method named ChemProp uses a novel technique called Message-Passing Neural Networks [4], that utilizes Convolutional Neural Networks. Another research group has also performed Deep Learning

techniques to discover molecules with bactericidal properties against Mycobacterium *tuberculosis* [9]. From recent research, a machine learning model was developed from the data obtained from high-throughput antibacterial screening to identify target molecules against Burkholderia *cenocepacia*, where they utilized ChemProp to model the data. As an off-shoot idea from their work, we used their dataset to run a simple machine learning model, but with active learning in order to train the model in such a way that minimum number of training samples is enough to achieve a good/desirable classification accuracy (or till the budget for conducting the experiments is finished). Utilization of active learning for drug discovery is still a new methodology to train models and a notable research by Naik *et al* [6], was one of the first practical demonstration of applying active learning for high throughput search of small molecules and their effect on a target (here, protein).

3 Proposed Methods and Partial Results

3.1 Data

3.1.1 Source

The data was retrieved from the high-throughput screening conducted by Selin *et al.* where they tried to identify growth inhibitory compounds against the highly infectious Burkholderia *cenocepacia* [8, 7]. It contains the SMILES (Simplified molecular-input line-entry system) formatted data for 29,537 small molecules along with their average B-score value for each molecule. The authors describe the B-score value as a measure of antibiotic property that is inversely proportional to the potency of the compound against the bacterium. From their research, the authors specify the threshold for B-score to be at -17.5, i.e. a B-score lower than this threshold is said to be active against the Burkholderia strain. Staying true to the essence of their research, we will use the same threshold and binarize the data as 0 for inactive and 1 for active compound. We visualize the data with their respective labels, according to the set threshold, in Fig.1.

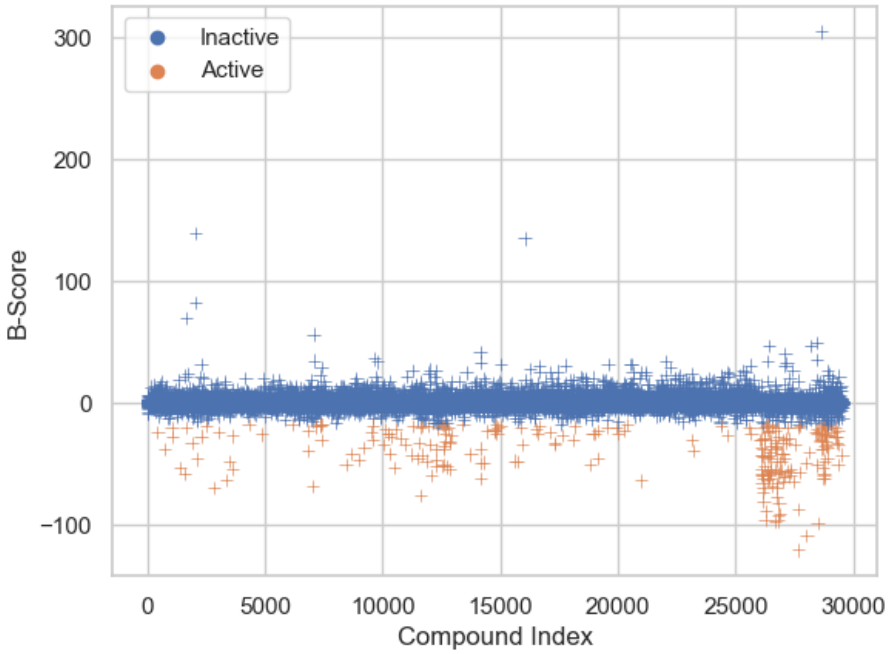

Figure 1: Visualization of all small molecules with respect to their Average B-score value. Blue crosses indicate B-score more than threshold value, and Red crosses indicate B-score lesser that the threshold value.

3.1.2 Representing SMILES Data as Molecular Fingerprints

Selecting an appropriate representation for the SMILES data is important in order to get informational features for downstream machine learning classification. One such method is to convert the molecular structure of the compound to a 2048-bit array. By this method, we can obtain a feature set size that is unchanging for all samples. This was achieved by using the Python package RDKit.

3.1.3 Resolving Class Balance Issues

The two classes - active and inactive, are heavily unbalanced. We observed that the data had only 256 active compounds. For this reason, we first did uniform random undersampling of the major class, i.e. the inactive compounds. This way we had equal number of compounds in each class. In order to eliminate any bias in the sampling of data, we simulate the model learning n number of times each with a different set of sampled data. Initially, we have set the value of n to 100. This may change for when we do a final cohort study.

3.2 Passive Learning

As a baseline model, we implemented a passive learning version with Logistic Regression Classifier as the base learner. Our model tries to learn which compounds are active and inactive against the bacterium. We sampled and modeled the data, with a 80:20 train-test split, for 100 simulations and observed an **average accuracy of 71.48**% with a **standard deviation of 0.04**, across all simulations.

3.3 Query by Committee

Query by Committee(QBC) query selection method selects the most informative instances for labeling by training numerous models through bootstrapping train data, then using the disagreement of those models of the query as uncertainty measure. Instance with the highest disagreement will be added to the train data. Logistic regression will be the base learner for our QBC method. QBC would be a good method to try first as our data may not be linearly separable. Sampling instances in the regions with the most uncertainty would define the decision boundary well and so QBC is expected to have a better performance than our passive learning.

3.4 Expected Model Change

Expected Model Change(EMC) query selection method chooses the most informative instances for labeling by computing the expected change in the model's performance by the instance. Instances that will bring the largest expected change will be selected sequentially from the pool of data. Our base learner is logistic regression and the norm of the derivative of the LR mean-squared loss will be used to compute the expected change. EMC will be helpful in capturing the change brought to the model by each instance, so this feature could be useful for a small train-test set like in our procedure.

3.5 A^2 algorithm

This algorithm would work well for binary classifiers and non-separable data, making it a good candidate algorithm for modeling our data [1]. This algorithm is known to be quite powerful, from literature, so we expect it to show the best performance.

3.6 Additional: Learning Active Learning through Reinforcement learning

This is an **ambitious** method that we may or may not have the time to implement but would like to explore. There are a few pre-prints [3, 10] that have explored this topic with respect to Natural Language Processing (NLP) and Image Classification, but none with a biological context like our data. In our case, a good policy or learning rule for the RL algorithm would be to make good queries or selections from the unlabeled data pool. We expect this method could make querying from unlabeled samples much easier and produce a better outcome than random sampling.

3.7 Code Availability

The code is public and available at the following github repository.

3.8 Potential problems with our approach

Even though our result from the passive learning problem is well above chance value, the accuracy is still on the lower end of the spectrum of a "good classification accuracy". Additionally, there are other factors apart from compound structure that could potentially be involved in potency against a particular bacterium. Solely depending on the compound structure could skew our model, producing results that may not reflect real-world scenarios. This is an unwanted outcome, potentially leading to squandering of resources when doing antibiotic discovery.

4 Mid-term and Final check

Currently, we have already implemented a passive learning version. For the mid-term check, we expect to have completed at least one of the active learning methods (potentially QBC). By the final check, we expect to have completed three different active learning methods along with the passive learning model, and performed a comparative analysis between the four techniques.

References

[1] Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning* (2006), pp. 65–72.

[2] Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F. J., Carballal, A., Maojo, V., Pazos, A., and Fernandez-Lozano, C. A review on machine learning approaches and trends in drug discovery. *Computational and structural biotechnology journal 19* (2021), 4538–4558.

[3] Fang, M., Li, Y., and Cohn, T. Learning how to active learn: A deep reinforcement learning approach. *arXiv preprint arXiv:1708.02383* (2017).

[4] Heid, E., and Green, W. H. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling 62*, 9 (2021), 2101–2110.

[5] Jukic, M., and Bren, U. Machine learning in antibacterial drug design. *Frontiers in Pharmacology* (2022), 1284.

[6] Naik, A. W., Kangas, J. D., Sullivan, D. P., and Murphy, R. F. Active machine learning-driven experimentation to determine compound effects on protein patterns. *Elife 5* (2016), e10047.

[7] Rahman, A. Z., Liu, C., Sturm, H., Hogan, A. M., Davis, R., Hu, P., and Cardona, S. T. A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery. *PLOS Computational Biology 18*, 10 (2022), e1010613.

[8] Selin, C., Stietz, M. S., Blanchard, J. E., Gehrke, S. S., Bernard, S., Hall, D. G., Brown, E. D., and Cardona, S. T. A pipeline for screening small molecules with growth inhibitory activity against burkholderia cenocepacia. *PLoS One 10*, 6 (2015), e0128587.

[9] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. A deep learning approach to antibiotic discovery. *Cell 180*, 4 (2020), 688–702.

[10] Werner, T. Reinforcement learning approach to active learning for image classification. *arXiv preprint arXiv:2108.05595* (2021).