@cardonalab

@cardona_lab

sciforum-045166
May 8-May17, 2021

Prediction of Antibiotic Activity against *Burkholderia cenocepacia* Using a Machine Learning Model

A S M Zisanur Rahman[1], Chengyou Liu[2], Lukas Timmerman[1,3], Andrew Hogan[1], Rebecca Davis[4], Pingzhao Hu[2,3,5] , Silvia Cardona[1,6*]

[1]Department of Microbiology; [2]Department of Electrical and Computer Engineering; [3]Department of Computer Science; [4]Department of Chemistry; [5]Department of Biochemistry and Medical Genetics; [6]Department of Medical Microbiology & Infectious Diseases
University of Manitoba, Winnipeg, MB, Canada

*Silvia.Cardona@umanitoba.ca

University of Manitoba

Abstract

A fundamental challenge in antibiotic discovery is finding new bioactive compound classes. Due to the longer timeframe and higher cost associated with conventional approaches, it has become imperative to adopt alternative antibiotic discovery paradigms. In this study, we exploited the machine learning (ML) model's ability to make predictive models and applied it to predict growth inhibitory activity in chemical scaffolds outside the training dataset. We employed a Directed-Message Passing Neural Network (D-MPNN) approach to train binary classification and regression ML models on a high-throughput screening dataset performed against *Burkholderia cenocepacia* previously in our laboratory. The D-MPNN belongs to Spatial-based Convolutional Graph Neural Networks (ConvGNNs), an end-to-end neural network that generates the graph representation of a molecule after iterative message passing process through molecular bonds. To avoid over-fitting and enhance the accuracy of the prediction, we additionally fed the model with 200 global molecular descriptors. The model was further optimized using Bayesian hyperparameter optimization and ensembling. The trained model attained a receiver operating characteristic curve-area under the curve (ROC-AUC) of 0.823. As a proof of principle, we employed the trained ML model to predict the bioactivity of 1,615 FDA-approved compounds and tested the bioactivity of the top 100 ranked compounds in vitro. We found 17 growth-inhibitory compounds with a linear correlation between the predicted rank and the activity. This work highlights the application of ML approaches to rapidly explore chemically diverse, ultra-large compound libraries and discern potential compounds in an inexpensive fashion, thus increasing the chance to discover early lead compounds.

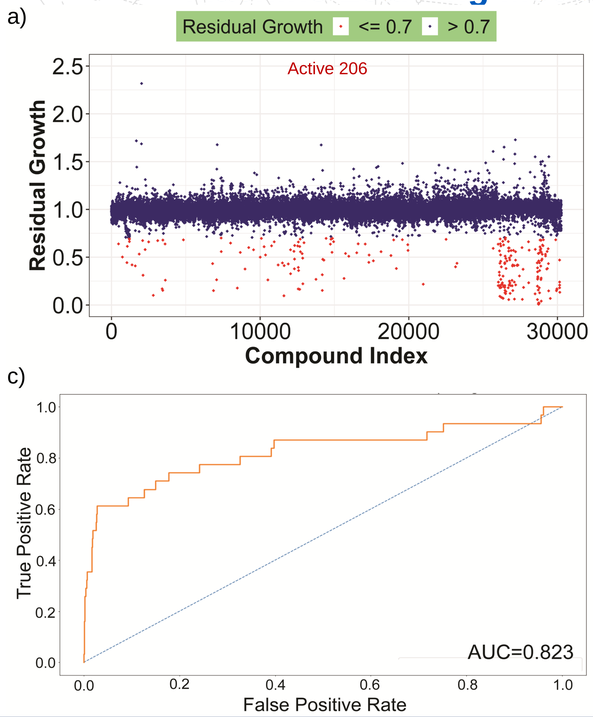
Introduction

The emergence of multidrug-resistant bacterial infections is one of the major health threats of modern times. The World Health Organization and the Public Health Agencies across the globe have highlighted the severity of the problem and the urgent need to develop new antibiotics[1–3]. To address this antibiotic resistance crisis, it is important to adopt novel antibiotic development scheme.

Given the recent advancement of machine learning algorithms, it can be used for *in silico* exploration of vast, diverse chemical spaces that are otherwise unprocurable. This approach will increase the chance and rate of early structurally novel scaffold discovery with desired bioactivity while simultaneously decreasing the associated cost and time. Here, we applied the Directed-Message Passing Neural Network (D-MPNN)[4] to train binary classification and regression models using a high-throughput screening dataset performed against *B. cenocepacia* K56-2 wild type[5]. In D-MPNN, the molecular information is propagated from edges to edges and constructed into a continuous vector for each molecule at the end of the message passing phase (**Fig. 1b**). Then the representation vector for each molecule is fed into the readout phase, which is a Feed-Forward Neural Network (FFN) that generates the final prediction (**Fig. 1b**). To enhance the model's accuracy and avoid over-fitting, additional molecule-level features such as molecular descriptors were provided into the model. Models were further optimized by Bayesian hyperparameter optimization and ensembling.

As a proof of principle, we applied the trained model on an FDA-approved compound library to predict their growth inhibitory activity. We empirically tested the top 100 ranked compounds and identified 17 active compounds. We observed a linear correlation between the predicted rank and bioactivity. Future work will use the model to predict bioactivity of large and diverse compound libraries in order to discover novel bioactive compounds.

Machine Learning Model Training and Evaluation



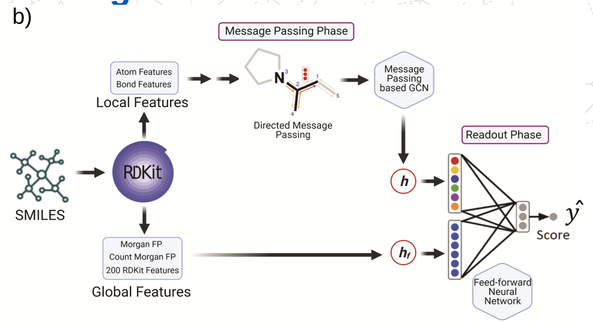
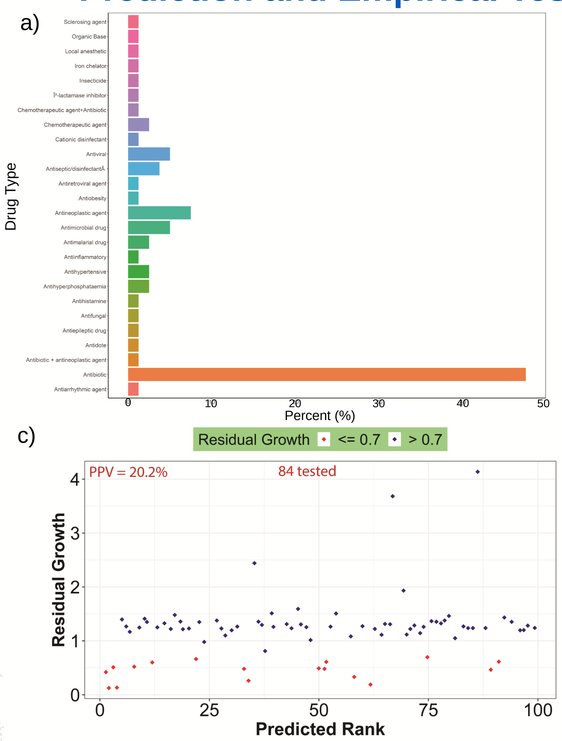


Fig. 1: Initial training and performance evaluation of the machine learning model. a) High-throughput screening data generated by screening a compound library of 30,259 compounds against *Burkholderia cenocepacia* K56-2 wild-type. Using 0.7 as threshold, the screening yielded 206 active compounds. Darkblue are inactive and red are active compounds. **b)** Machine learning model was trained using a Directed-Message Passing Neural Network (D-MPNN) approach which extracts compounds local features such as atom and bond features. The model was fed more than 200 additional global molecular descriptors to further increase the accuracy. Dataset was split into 80:10:10 to train, validate and test the model. **c)** ROC-AUC plot evaluating model performance after training. The model attained a ROC-AUC of 0.823.

Prediction and Empirical Testing of FDA Approved Compounds



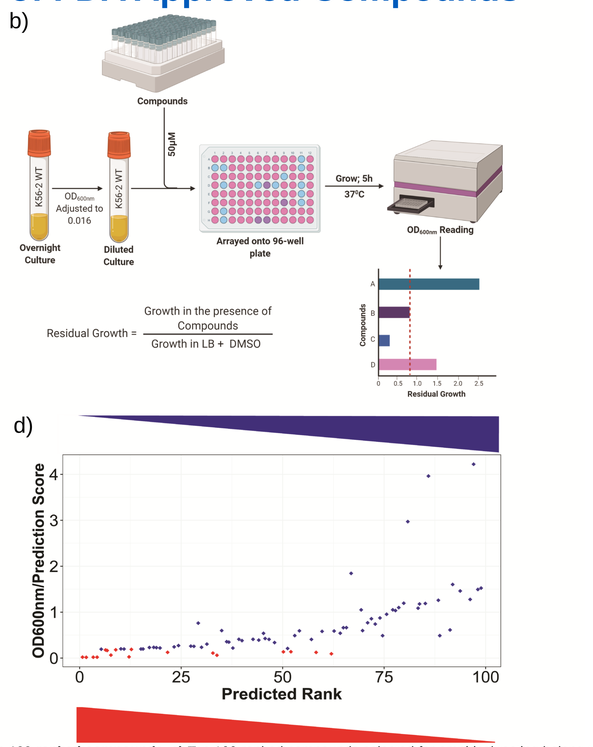


Fig. 2: Bioactivity prediction of an FDA approved compound library and *in vitro* testing of the top 100 ranked compounds. a) Top 100 ranked compounds selected for empirical testing belong to different drug families. **b)** A schematic of the screening protocol. 84 commercially available compounds (from the top 100) were screened. **c)** The screening identified 17 bioactive compounds with positive predictive value (PPV) of 20.2%. Darkblue are inactive and red are active compounds. **d)** The ratio of OD_{600nm} and prediction scores were plotted against the predicted rank of the corresponding compounds. The results show a linear correlation between the prediction score and bioactivity. Darkblue and red indicate compounds' probability of being inactive and active, respectively.

Bioactive Compounds Belong to Diverse Drug Families

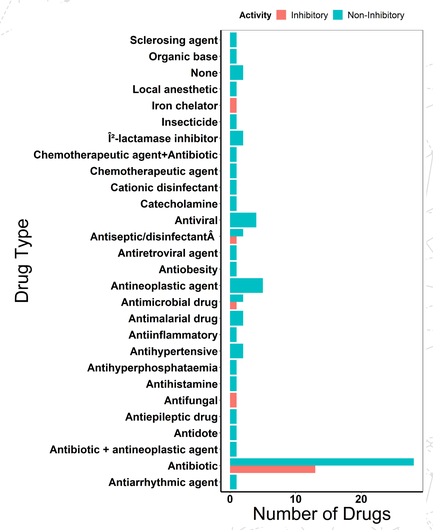


Fig. 3: *In vitro* growth inhibitory activity of compounds belonging to different drug families. As expected, most of the compounds exhibiting bioactivity were antibiotics or antimicrobial compounds.

Summary and Significance

- A hybrid molecular representation approach was utilized to develop a machine learning model which achieved a ROC-AUC of 0.823 on the test dataset.
- The trained model was used to predict growth inhibitory activity of an FDA approved compound library.
- *In vitro* screening of the 84 top ranked compounds yielded 17 growth inhibitory compounds, increasing the screening hit rate to 20.2%.
- The model can predict bioactivity of compounds outside of the training dataset, highlighting the ability of the model to generalize.
- Our work highlights the application of machine learning approach to rapidly explore chemically diverse compound libraries and may empower novel antibiotic discovery.

Future Work

- Future work will utilize the trained model for *in silico* screening of unprecedented chemical libraries to identify new antibiotic candidates.
- We anticipate that our model will be able to provide tractable bioactivity predictions of compounds with low structural similarity.

References and Acknowledgements

1. Brown, E. D. & Wright, G. D. *Nature* 529, 336–343 (2016).
2. Billington, J. K. *Am J Law Med* 42, 487–523 (2016).
3. Talebi Bezmin Abadi, A. et al. *BioNanoSci*. 9, 778–788 (2019).
4. Yang, K. et al. *J Chem Inf Model*. 59, 3370–3388 (2019).
5. Selin, C. et al. *PLoS ONE* 10, e0128587 (2015).

Cystic Fibrosis Canada
CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada
UMGF
CYSTIC FIBROSIS FOUNDATION
Bourses d'études supérieures du Canada Vanier Canada Graduate Scholarships