Antibiotic Discovery Using Message Passing Graph Neural Networks

Karolina Gustavsson
Mahan Tourkaman
Viveka Brockman

ID2223
February–March 2022

1 Introduction

The purpose of this project is to understand how deep learning can be used in drug discovery, as applied to the problem of finding novel candidate small molecules with antibiotic properties.

All the datasets and scripts used in this project are available on GitHub.

In our proposal we stated a few points that we decided to change, with the main one being our investigation of RNA molecules. We decided not to pursue RNA molecules, despite the fact that it would be interesting to know whether ribozymes, for instance, could act as antibiotics. The main reason for this was the size of RNA molecules, as they tend to be rather large and would therefore require a lot of computing power. Additionally, it was difficult to find a suitable RNA dataset with SMILES strings. Instead, we decided to train our model using smaller molecules from the training dataset used in [1], in order to then predict the antibiotic properties of a subset of COCONUT [2] — the COlleCtion of Open Natural ProdUcTs. In order to improve our model and better understand the effect of its various hyperparameters we also preformed hyperparameter optimization.

The main tool used in this project is Chemprop [3, 4], a library for predicting molecular properties using message passing graph neural networks. The training process starts by first presenting the molecules of interest to Chemprop in the form of SMILES strings, along with a desired set of target numerical features. In contrast to older methods, Chemprop learns which functional groups (chemical features) are important (exhibit or lack bioactivity), and therefore there is no need for any feature engineering beforehand. The architecture behind Chemprop is implemented in PyTorch. SMILES files are representations of molecules in terms of ASCII strings, a standard notation in chemistry. In the graph analogy, the nodes represent the atoms, and the edges stand for the bonds. This graph representation is then converted into a vector which the learning algorithm can then learn the molecular structure from. Using this learned model of the molecules and the molecular properties used during training, various properties can be predicted. Such properties might include toxicity, antimicrobial characteristics, and solubility.

2 Datasets

We trained our model using the same dataset as in [1], the deduplicated version of the primary screening data with 2336 molecules for growth inhibition of E. coli, from an FDA-approved drug library, which also includes more than 300 compounds from a natural product collection. The dataset can be found in Table S1 under Supplemental Information, sheet S1B, in the aforementioned article. In our pipeline, we changed the labels from {Inactive, Active} to {0, 1}, and extracted the SMILES and Activity columns as CSV file, for use with Chemprop.

For our predictions, we created a dataset based on COCONUT. We used the advanced search functionality, in order to find compounds containing between 0 and 5 carbon atoms, downloaded the resulting SDF file and, using a custom Python script, extracted the SMILES strings and compound IDs and saved it as a CSV file.

Feed-Forward Layers	Hidden Size	Message-Passing Steps	Dropout	Validation AUC ▼
2	**900**	**4**	**0.15**	**0.992459**
1	500	4	0.05	0.992081
2	800	4	0.4	0.99095
1	1700	5	0.1	0.989442
3	700	2	0.35	0.989065
2	1100	3	0.3	0.987557
1	2300	6	0	0.986802
1	2400	6	0	0.986425
1	1400	3	0.25	0.986048
2	400	3	0.15	0.985671
2	1300	3	0.4	0.984917
1	2300	5	0	0.984917
1	2200	6	0.05	0.984163
3	1600	6	0.2	0.983409
2	1000	3	0.35	0.983032
3	1100	3	0.25	0.981523
2	1900	5	0.3	0.981146
2	2000	4	0.05	0.980392
2	1400	2	0.25	0.980015
3	2100	2	0.3	0.973605

Table 1: Validation AUC for various hyperparameter configurations, using Chemprop's default settings: 1-fold cross-validation, and 1-ensemble models (i.e. single models).

3 Method

Our training and prediction datasets were prepared as described in the previous section.

Using Chemprop's command line interface we trained several models in order to then perform hyperparameter optimization and ensembling using Chemprop's `chemprop_train` and `chemprop_hyperopt`.

During training, we kept the default recommendations of Chemprop: The training data was randomly split into 80% training, 10% validation, and 10% test. The model settings were as follows. 30 epochs, 1-fold cross validation, ReLU activation, hidden size 300, 2 feed forward layers, 3 message passing steps and 0 dropout. The ensemble parameter was set to 1, meaning only one model was trained.

Using the trained model, we generated predictions for our COCONUT dataset, and selected the ones with an activity over 0.5 as our candidates for novel antibiotics.

4 Results

4.1 Model Training

The model upon which the predictions were made was trained using no extra configurations than the defaults in Chemprop's `chemprop_train` command (1-fold cross-validation, single model (1-ensemble)), and for that training we obtained an AUC of 0.984917 on the validation set, and 0.842208 on test set. It took about 3 minutes to train this model.

Though not used for consequent model training, the results of the hyperparameter optimization run are included in Table 1.

4.2 Predictions

The compounds with predicted activity levels above 0.5 are shown in Table 2.

SMILES	Activity ▼	Name
OC1(NN1)C2(O)NN2	0.9787	3-(3-hydroxydiaziridin-3-yl)diaziridin-3-ol
O=C(O)C1(O)NN1	0.8825	3-hydroxydiaziridine-3-carboxylic acid
S1SSSSSSS1	0.7282	octathiocane
O=C1N(C)C2SSC1N=C2O	0.6948	8-hydroxy-5-methyl-2,3-dithia-5,7-diazabicyclo-oct-7-en-6-one
ON(O)N1C=C(N)N(N)N1	0.5961	No name available
S1SSSCSS1	0.5432	1,2,3,4,5,6-hexathiepane
S(SSSC)SSC	0.5350	dimethylhexasulfane

Table 2: Molecules with predicted antibiotic activity levels above 0.5. The closer they are to 1, the more effective they are predicted to be. Sorted by descending activity.

5 Discussion

The obtained validation and test AUC results are fairly good: Validation AUCs are in the range 0.98-0.99 and test AUCs in the range 0.83-0.88. No single hyperparameter seems to stand out as the one contributing the most to these results, and we therefore conclude that it is perhaps more about the combination of them, as opposed to any individual one.

For the predicted activity, one interesting feature is that four out of seven of the candidate compounds contain sulfur. This is of interest since sulfonamides, which contain sulfur, historically have been used in antimicrobial drugs. This provides more credibility to our findings. However, it is worth considering that wet lab experiments are needed for complete validation.

While in silico methods for drug discovery (such as ours) show great promise, they have also been criticized for the disconnect that exists between performance of the models and ensuring that they are suitable for real life applications [5]. One has to consider removing problematic compounds, solubility, toxicology, manufacturing, patents and mechanism among other things. Furthermore, the labeling of the data can vary depending on assay setup and experimental conditions. In addition to this, there is a small amount of data that is relevant in vivo. Often the focus is less on endpoint application for decision making and more on model performance. Another concern is validation; it might be impractical to validate with assays in a wet lab.

6 Conclusion

Chemprop's message passing neural networks were used to find new antibiotic candidates. Seven compounds were chosen, four of which contained sulfur.

Further research is however needed in order to confirm the efficacy of these compounds in wet lab. It would also be interesting to investigate Chemprop's scaffold-based dataset split mode and the effect it would have on the type and quality of the predictions.

Other future directions might include trying other datasets, creating new datasets from a chemical database based on certain molecular properties, using RDkit for investigating larger molecules, validation using in silico structural biology methods — looking at mechanisms and binding abilities.

Datasets of interest include GDB [6], Zinc [7] and ChEMBL [8]. Another future direction could be the interpretation of the obtained model in order to find out what the compounds with antibiotic properties have in common.

There are lots of further investigations that can be done. For instance: trying other datasets, creating other datasets from a chemical database based on certain molecular properties, using RDkit for investigating larger molecules, validation using structural biology in silico methods — looking at mechanisms and binding abilities. If we had access to a GPU, we could have tried a larger dataset such as one of the GBD datasets. Datasets of interest for future studies include GDB, Zinc and ChEMBL. Another question that could be investigated further is interpreting the model i.e. what do the compounds with antibiotic properties have in common.

References

[1] Jonathan M. Stokes et al. “A Deep Learning Approach to Antibiotic Discovery”. In: *Cell* 180 (2020), 688–702.e13.

[2] M. Sorokina, P. Merseburger, and K. Rajan et al. *COCONUT online: Collection of Open Natural Products database*. https://doi.org/10.1186/s13321-020-00478-9. 2021.

[3] Kevin Yang, Kyle Swanson, and Wengong Jin et al. “Analyzing Learned Molecular Representations for Property Prediction”. In: *Journal of Chemical Information and Modeling* 59 (8 2019), pp. 3370–3388. URL: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.9b00237.

[4] Wengong Jin et al. *Chemprop*. https://github.com/chemprop/chemprop. 2018-2021.

[5] Andreas Bender. *How to Lie With Computational Predictive Models in Drug Discovery*. 2020. URL: http://www.drugdiscovery.net/2020/10/13/how-to-lie-with-computational-predictive-models-in-drug-discovery (visited on 03/09/2022).

[6] Reymond Group. *GDB Databases*. https://gdb.unibe.ch/downloads. 2005-2012.

[7] Sterling and J Irwin. *Zinc*. https://zinc15.docking.org. 2015.

[8] Davies M and Nowotka M et al. *ChEMBL*. https://www.ebi.ac.uk/chembl/. 2017.