In silico prediction for potential non-phospholipidosis-inducing inhibitors against in vitro replication of SARS-CoV-2

Junqi Lu, Anish Karpurapu

Problem description

Since the outbreak and the spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the ongoing coronavirus disease 2019 (COVID-19) has been an extreme threat to public health worldwide—over 7.1 million people have been infected and more than 0.4 million deaths have been reported. Although vaccines and monoclonal antibody treatments against COVID-19 have been developed, small-molecule therapeutics are still highly limited despite their advantages. An antiviral drug can provide additional options for people who are not suitable for vaccination due to medical or religious reasons, have been infected by COVID-19 before they can get vaccinated timely, or already vaccinated but with low or no responses. Additionally, the manufacturing rate of small-molecule drugs is faster than those of vaccines and antibodies, let alone that small-molecule drugs have lower requirements for storage, transportation, and administration, which would boost their applications in remote and underdeveloped regions.

Under this urgent need for small-molecule drugs, Gao et al built a structural-based drug repositioning machine learning model to predict effective SARS-CoV-2 drugs by in silico screening potent inhibitors against SARS-CoV, the pathogen for the SARS pandemic in 2003.[1] They hypothesized that potent SARS-CoV inhibitors are also effective for inhibiting SARS-CoV-2, considering the sequence similarity between the two viruses. However, this logic is not necessarily always valid. Pathogenetic distinctions do exist between the two viruses and can impair the practical application of their findings. For example, asymptomatic transmission exists for SARS-CoV-2 but not SARS-CoV.[2] To obtain more SARS-CoV-2 relevant results, near the onset of the recent COVID-19 pandemic, Touret et al screened 1520 FDA-approved and EMA-approved drugs in vitro and identified 90 compounds as positive hits to specifically and potently inhibit SARS-CoV-2 replication.[3] Yet, in vivo screening for such a large scale is still time-consuming and labor-intensive. Consequently, by exploiting the power of machine learning, our proposed study aims to develop a model that can learn from the inherited knowledge buried within the in vitro data and make reliable predictions for SARS-CoV-2 inhibitors. The knowledge gained from this project can provide new opportunities to screen larger libraries in a context closer to in vitro and thus judge additional already-approved drugs for repurposing against COVID-19 and its variants.

Modeling approach

Our overall approach would consist of data preprocessing, training several machine learning models, and finally running our model on the ChEMBL dataset to try to identify novel antivirals that inhibit the SARS-CoV-2 virus.

Data preprocessing would involve cleaning the dataset, obtaining the SMILES, creating molecular descriptors, and train-valid-test splitting. We will use the Prestwick Chemical Library that Touret et al used in their SARS-CoV-2 inhibitor screening. This dataset contains the chemical names of the 1520 market-approved drugs with their inhibition indexes tested by Touret et al using a broad-spectrum antiviral drug called Arbidol (umifenovir) as a control. Moreover, the chemicals in this library have high chemical diversity and high pharmacological diversity and that adds to our model's robustness for processing new data. Using this dataset, we plan to first find SMILES information through ChEMBL lookups before creating molecular descriptors using Morgan Fingerprints. We will perform a dimensionality reduction to get the top 50 principal components. This will decrease the chance that we overfit the data as a result of having more descriptors than data points. We also plan to test other approaches for getting the descriptors, for example, by getting attributes by converting the SMILES to ConvMol objects (as we have done in Lab6). Lastly, to prepare the datasets, we will perform train-validation-test splitting stratified on the class label. 20% of the data will be held out for our validation set, and we will do five-fold cross-validation on the rest of our dataset to choose the optimal model architecture and hyperparameters. The inhibition indexes in the dataset are numerical and Touret et al defined a hit as a drug with an inhibition index higher than 1. We will use a step function to classify each compound as an inhibitor or a non-inhibitor and then use those classes for model stratification.

We will try a variety of machine learning regressors on this dataset including random forests, graph convolutional models, recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and variational Bayesian neural networks. Hyperparameter tunning will be applied for the optimal performance. We will use these machine learning models to regress the numeric inhibition index. Random forests will serve as a simple baseline that we hope to improve on by using more complex models mentioned above. We will train all these machine learning models using the five-fold cross-validation with splitters based on clustering algorithms including Butina and fingerprint differences to ensure the model robustness for new data, and test on the hold-out dataset to prevent any risks of data leakage. Depending on the amount of time we have left in the project, we will also seek to use explainable AI approaches to determine the most important features in predicting whether a compound can inhibit SARS-CoV-2.

Evaluation approach

In the original dataset, Touret et al included the calculated inhibition indexes and defined drugs with an inhibition index higher than 1 are effective for inhibiting SARS-CoV-2 replication. Following the same logic, our regressor will output predicted inhibition indexes according to the input descriptors, and we will also use a step function to classify the outputs into two classes: effective and non-effective. By comparing the true inhibition indexes and the predicted inhibition indexes, we will assess the regressor performance by scatter plots visually for correlations between predictions and true labels, calculating the Pearson correlation coefficients to estimate potential linear relationship, and calculating the Spearman rank correlation to check whether the predictions are ranked in the same way as the training data points. We will also report the R^2 with MAE and RMSE.

Additionally, we will convert the predicted inhibition indexes into the predicted classes and compare that to the original classification in the training data. Our evaluation metrics relevant for classification will include precision and recall. Since our dataset has more non-effective molecules than the effective ones, we will also include the balanced accuracy and Mathews correlation coefficient. Moreover, ROC-AUC, enrichment factors, and BEDROC will be calculated and generated to assess the predictions with low uncertainty. All these tests will reveal how well our model does in predicting inhibitors versus non-inhibitors of SARS-CoV-2.

Clustering and dimensionality reduction plots, such as PCA, will be used as auxiliaries to examine the performance. Judgments above will be made based on results from sufficient times of performances to ensure the inter-performance consistency and the statistical significance for comparing different models' performances. T-test and Wilcoxon signed-rank test will be chosen based on the outputs' continuity and distribution accordingly. A p-value less than 0.05 will be considered statistically significant.

Expected challenges and alternative approaches

A primary challenge of our dataset is that it is not incredibly large with 1520 data points. If we do not achieve promising results in our initial data analysis, we will consider adding in extra in vitro data from other sources to create a bigger dataset. We will also consider switching datasets if we find a significantly larger dataset that we cannot combine with our current data.

Another limitation is that our dataset has only the gross measurement of the antiviral effects without specific mechanistic explanations. One concern we are having for our prediction to proceed downstream of the drug development is phospholipidosis, a drug-induced abnormal accumulation of intracellular phospholipids that can lead to lysosomal storage disorder in vitro. This can be a confounding variable for determining whether an antiviral molecule is virus specific and it might be an obstacle to translate the findings to in vivo. If our model predicted more drugs that tend to induce phospholipidosis, we would develop another machine learning model that predicts whether a drug is likely to cause phospholipidosis to exclude it from the output. Multiple datasets of drugs that can induce phospholipidosis are publicly available.[4] Furthermore, we can verify if this model is performing well by searching through our top results and determining if they correspond to experimental findings as reported in papers describing phospholipidosis.

Expected outcomes and definitions of success

Regarding the performance of the model, we define the success of our model more in a continuum sense. At least, we primarily expect our model to perform better than a random model does for correct predictions. That will be judged by whether our model has a ROC-AUC that is larger than 0.5 and an enrichment factor that is larger than 1. Further, we expect our model to learn from the training data and thus, avoid the overfitting issue. If our model's prediction has a smaller R^2 on a testing dataset than it does on a training dataset, then the model is overfitting on the training dataset. The closer the two R^2 from the training and the test datasets are, the better the model's performance will be. We expect to have a balanced accuracy closer to 1 and an enrichment factor as large as possible. Since we proposed a drug discovery project, we expect our model to have a high precision and a low recall to properly deal with the precision/recall trade-off as a conservative model. We will also perform adversarial controls by shuffling the molecular fingerprints corresponding to the inhibition index to see if our model performs better than scrambling. Finally, considering the small size of our dataset, we will check whether our model predicts SARS-CoV-2 inhibitors more accurately than a random forest model does. For all the judgments mentioned above, we expect statistical consistency and significance with p-values smaller than 0.05 to further endorse our conclusions.

References

1. Gao, K., et al., *Repositioning of 8565 Existing Drugs for COVID-19.* J Phys Chem Lett, 2020. **11**(13): p. 5373-5382.
2. Johansson, M.A., et al., *SARS-CoV-2 Transmission From People Without COVID-19 Symptoms.* JAMA Netw Open, 2021. **4**(1): p. e2035057.
3. Touret, F., et al., *In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication.* Sci Rep, 2020. **10**(1): p. 13093.
4. Lowe, R., R.C. Glen, and J.B. Mitchell, *Predicting phospholipidosis using machine learning.* Mol Pharm, 2010. **7**(5): p. 1708-14.