NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

SCSE21053 – Computational analysis to discover the potential neutralizing antibodies for novel coronavirus II

Presented by Chan Joshua Juan Yin

Supervised by Assoc Prof Kwoh Chee Keong

Abstract

The development of vaccination for novel coronavirus often requires screening of thousands of available strains of antibodies, which is prohibitively expensive and often not feasible due to a lack of available structures. Accordingly, Machine learning (ML) models can enable the rapid and inexpensive exploration of vast sequences with affordable resources. In this project, we utilized ML approaches to discover neutralizing antibody for potential novel coronavirus based on viral antigen-antibody binding properties. Building on from the paper by Magar et al. 2021, we introduced Mean and Max Pooling for data augmentation, applied Repeated Stratified K-fold strategy for splitting the training and test dataset, and employed 8 classifier models for ML training. Our model discovered that Logistic Regression (LR) with mean pooling could achieve the highest accuracy score(74.17%), F1 score (85.01%). We then conducted hyperparameter random searching to fine tune the selected model, achieving final Accuracy, F1 and ROC AUC score at 74.92%, 85.39% and 53.05%. Our research could provide insights to future ML prediction approaches to coronavirus antigen-antibody binding properties.

Data preparation

- We leveraged the database from Raybould et al 2021 (Oxford database). The database provided a comprehensive list of antibodies and nanobodies able to bind to coronaviruses.
- Collaborating with another team member of this research project, Ng Yue Hao Shaun, we extracted FASTA sequences of paratope and epitope (total 310 pairs), and data on whether each pair is neutralizing.
- We used RDKit library on Python to generated a feature matrix and an adjacency matrix for each paratope and epitope molecule. This approach is similar to the approach recorded in Magar et al. 2021 paper.

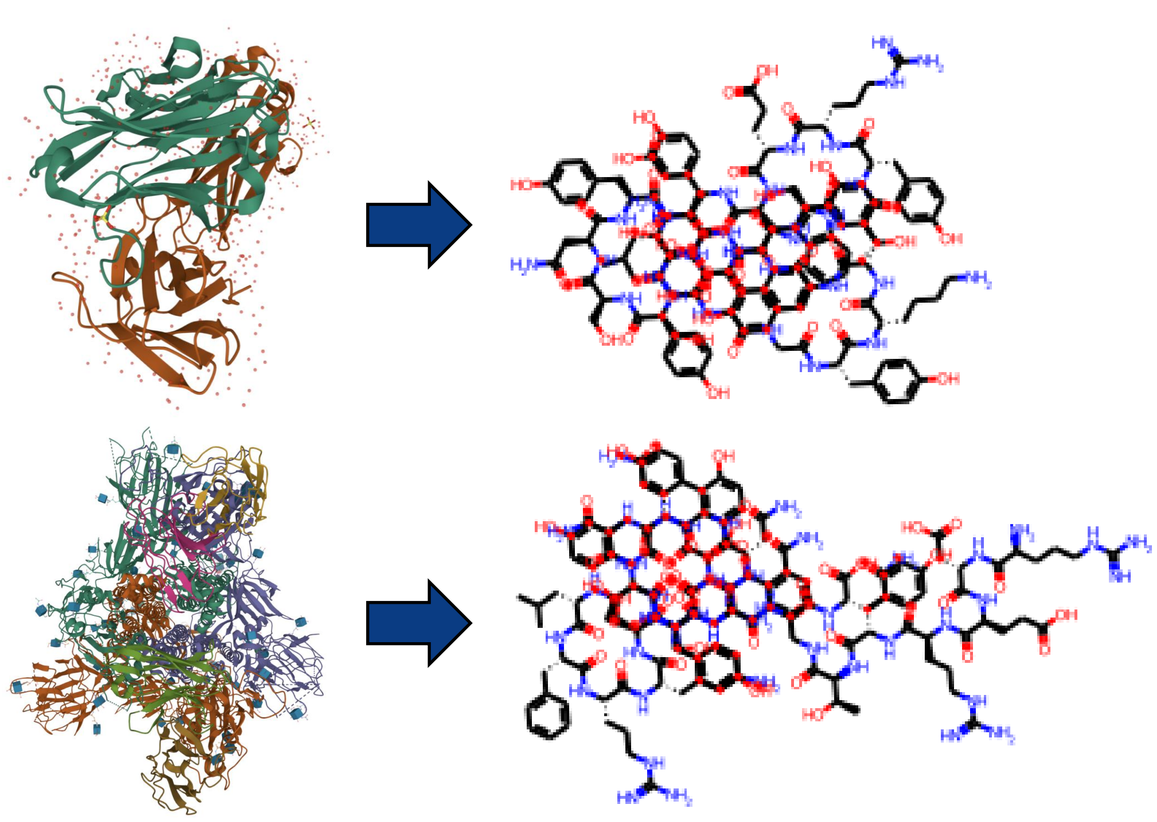

Figure 1: Extracting FASTA sequences of paratope and epitope from antibody and coronavirus molecular structures. **(Up)** 6XCA molecule, also named as C105 neutralizing antibody Fab fragment. **(Down)** 6XCM molecule, which is a SARS-CoV-2 spike glycoprotein. This pair of paratope-epitope is neutralising.

Result

- LR exhibited the highest average Accuracy score (74.17%) and F1 score (85.01%) when training on mean pooling matrix.
- SVN exhibited the highest average ROC AUC score (51.92%) when training on mean pooling matrix.
- LR was selected as the best model and mean pooling matrix was selected as the input data for training due to their performance.
- Hyperparameter Random Search was conducted on the selected model – LR. Result showed Accuracy score at 74.92%, F1 score at 85.39%, ROC AUC score at 53.05%.

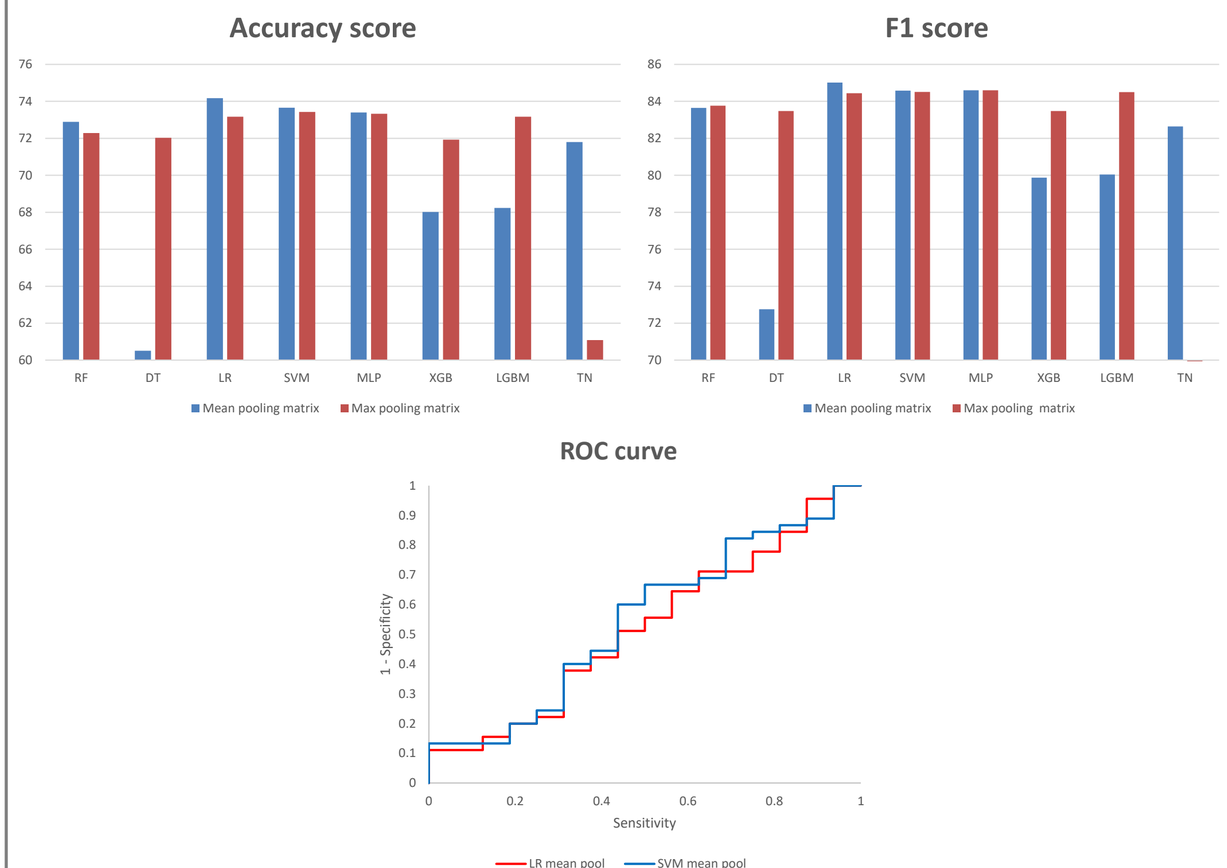

Figure 2: (Up left) Comparison of average Accuracy score of all models. **(Up right)** Comparison of average F1 score of all models. **(Bottom)** ROC curve of LR and SVN on mean pooling matrix.

Methodology

1. For each paratope and epitope, conducted matrix multiplication on the adjacency matrix with the feature matrix.
2. Conducted mean pooling and max pooling across all rows (atoms) in each product matrix. This generated 2 vectors: the mean pooling vector and the max pooling vector.
3. Concatenated the mean pooling vector of each paratope with its pair epitope; did the same for their max pooling vectors.
4. Concatenated mean pooling vectors of all molecules into a mean pooling matrix. Did the same to generate the max pooling matrix.
5. Applied 8 classifiers models: Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), XGBoost (XGB), LightGBM (LGBM), TabNet (TN).
6. For each classifier model, applied Random Stratified 5-Fold (repeat 10 times) to split the train and test data. Total 50 trainings per model.
7. Computed the average Accuracy, F1 and ROC AUC score of each classifier model. Compared their performance.

Conclusion

- Our ML model could predict neutralizing effects of a paratope-epitope pair at Accuracy closed to 75%, F1 score 85% and ROC AUC at 53%.
- The above research provides insights to future ML approaches to predict coronavirus antigen-antibody binding properties.

Future work

- Our model exhibit high bias given its Accuracy, F1 and ROC AUC score.
- Graphical Neural Network (GNN) or sequential models, e.g. GRU or LSTM, can reduce bias, as the have deeper neural network.
- Expanding our dataset to include new virus variants e.g. Delta/Omicron

References

- Magar, R., Yadav, P. & Barati Farimani, A. Potential neutralizing antibodies discovered for novel corona virus using machine learning. Sci Rep 11, 5261 (2021). https://doi.org/10.1038/s41598-021-84637-4.
- Matthew I. J. Raybould, Aleksandr Kovaltsuk, Claire Marks, Charlotte M. Deane (2021) CoV-AbDab: the Coronavirus Antibody Database. Bioinformatics. 37(5):734-735. doi = 10.1093/bioinformatics/btaa739.