Machine Learning and AI for SARS-CoV-2 Main Protease Inhibitor Discovery

Ammar Ali
Supervisor: Prof Garrett M. Morris

Project Aims

At the start of the project, the main goal was to determine if machine learning could help find which inhibitors might be most effective against the SARS-CoV-2 Main protease. The title of the project can be broken down into tool and a problem: a “Machine Learning and AI” tool and a “Main Protease Inhibitor Discovery” problem.

Main Protease (Mpro, Nsp5, Main Proteinase, 3CLpro) is one of the proteins that is produced by the translation of the viral RNA of SARS-CoV-2. It is responsible for hydrolysing peptide bonds at 11 different sites from pp1a and pp1ab. This allows the proteins to become separate entities and allows for the virus to carry out its function and replicate. Therefore if a suitable inhibitor for Mpro is found it will prevent the virus from replicating. Mpro is also an attractive enzyme to target for inhibition because it is not used within human cells, hence an inhibitor with fewer side effects will be easier to find.

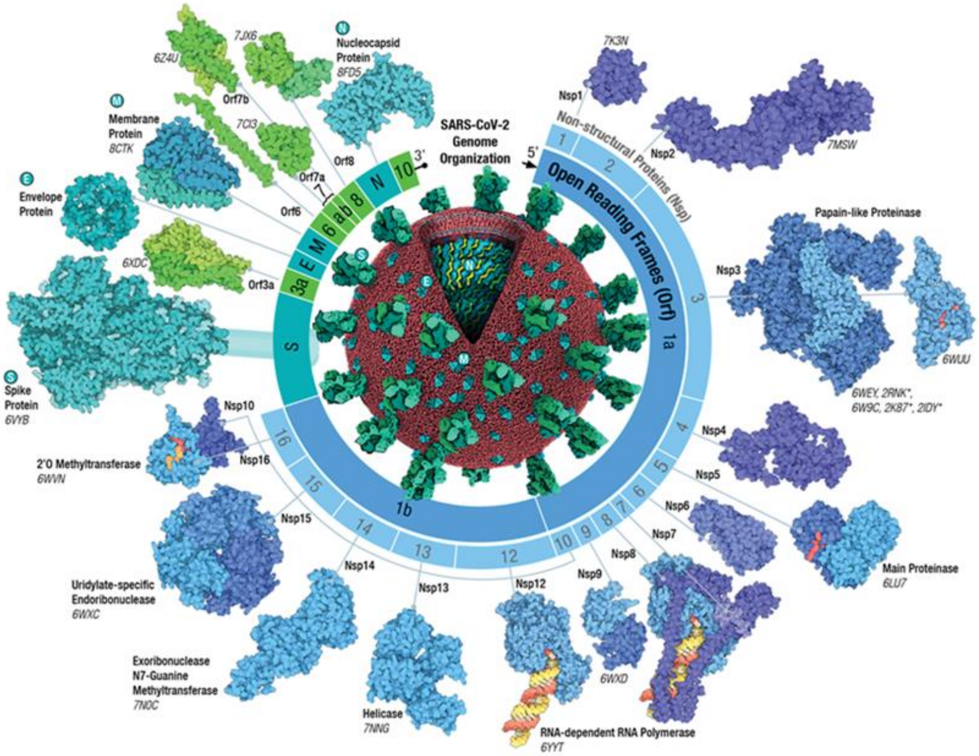

Figure 1 - Graphical representation of SARS-CoV-2 virus with the proteins that will be produced by the translation of the viral RNA

Various machine learning models were trialled using an automated machine learning package named FLAML. It was soon realised that gradient boosting machines were particularly suited for this kind of task. Although the initial plan was to delve deeper, using the details of how the inhibitor interacts with the Main Protease, and to craft a model with Graph neural networks and Protein Ligand Interaction Graphs, time constraints prevented this from occurring.

Instead, the focus remained on the gradient boosting machine, seeking a deeper understanding of its predictions. A SHAP analysis was then used to convert the features back to the features of the chemical that acted as positive predictors in the model. From this analysis, it was possible to combine these substructures and identify a general structure prevalent in many potent inhibitors.

Methods, Sources and Approaches

Machine learning is a subset of artificial intelligence that focuses on the development of algorithms that allow a computer to perform tasks without being specifically programmed to do so. This is done by feeding data into a machine learning model and allowing the model to detect and find the underlying patterns within the data so that it can then generalise the information and make predictions using the generalised information.

Generally, there are three subcategories within machine learning. These are:

- Supervised Learning - The data that the model is trained on contains labels. Essentially, this means that the data included the outputs and the inputs and the model will try to find the relation between the two in order to make predictions on further new or unseen data.

- Unsupervised Learning - The data that the model is trained on does not contain labels. The purpose of unsupervised learning is therefore to find the relationships between the input features and structure the data by similarity or to reduce the dimensionality of the input feature space.

- Reinforcement Learning - There is no dataset in reinforcement learning usually. The model is trained by interacting with an environment. The model then receives either positive or negative feedback from the interaction. Using that information the model will then try to maximise the positive feedback by adjusting its interaction with the environment. An example of this is using a machine learning model to beat a game.

In the project both supervised and unsupervised learning was used. Supervised learning problems can be categorised into two further topics. These are:

- Classification - This is when the prediction from the model is a discrete value. For example in the context of the project, this could be assigning either a “strong inhibitor” or “weak inhibitor” value to each inhibitor that the model makes a prediction for.

- Regression - This is when the prediction of the model is a continuous value. Again in the context of the project, this would be assigning a predicted pIC50 value to each chemical (more details about what a pIC50 value is will be discussed later)

When using machine learning the process can be broken down in these general steps:

1. Data Collection
2. Data Preprocessing
3. Featurisation
4. Model Selection
5. Training
6. Evaluation
7. Tuning / Optimisation
8. Prediction and Interpretations

These are just general steps, each problem may require further steps that are not mentioned. Further details about each step and how they were used in the context of this project are discussed.

Data Collection and Data Preprocessing

The data that was used was from the PostEra COVID Moonshot Dataset, which was a data set that contained approximately 2000 small molecule inhibitors along with an IC50 value to represent the affinity that each molecule would have with the active site of Mpro. The structure of the inhibitor was given in their SMILES representation. SMILES is a text based format which can be understood by humans as well as computers. An example of a SMILES representation can be seen in Figure 2.

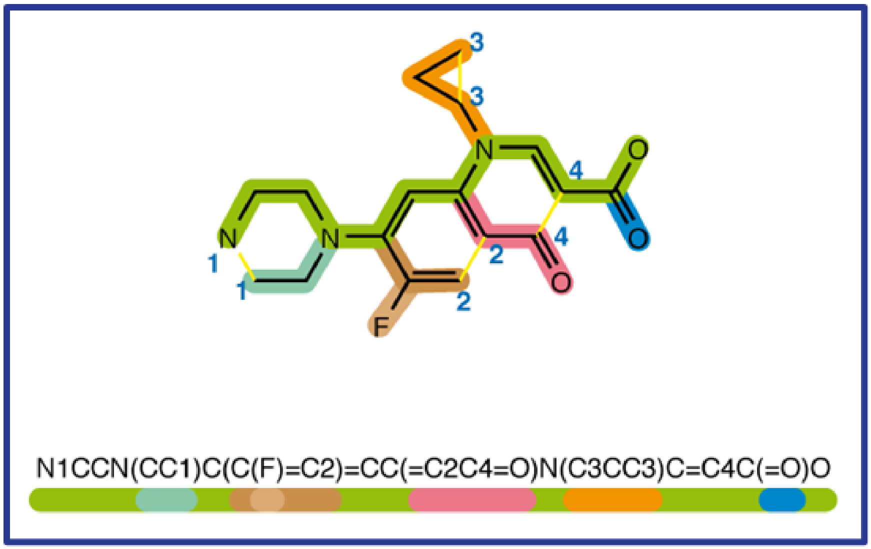

Figure 2 - An example of a molecule and it's representation in SMILES format

The IC50 of an inhibitor is the concentration at which it will inhibit 50 percent of the function of the enzyme that is being inhibited, which in this case is Mpro. In the dataset the IC50 value was found using two methods, which were rapid fire IC50 and fluorescence IC50. For the machine learning model only fluorescence IC50 values were used because most of the chemical had a fluorescence IC50 value.

The IC50 was then converted to a pIC50 value and then the inhibitors with a pIC50 value greater than 6 were classified as strong inhibitors, whilst the inhibitors with a value less than 6 were classified as weak inhibitors. These classifications are the labels that will be used in the machine learning model.

Featurization

The machine learning model needs a way to measure the similarity and differences of each molecule. There are multiple ways of doing this, for this project Morgan Fingerprints were used. Morgan Fingerprints work by indexing through each atom in a molecule and then looking at the atoms that are in a set radius next to the indexed atom, this will be one substructure. The set radius used in the project was 2. An algorithm is then used to find the relations between the atoms and allow each to be assigned a unique value. The values are then hashed into a 2048 long bit vector and each bit will correspond to one or multiple substructures. If a value of one is recorded for a bit in the fingerprint of a chemical, this means that it contains a substructure that corresponds to the active bit.

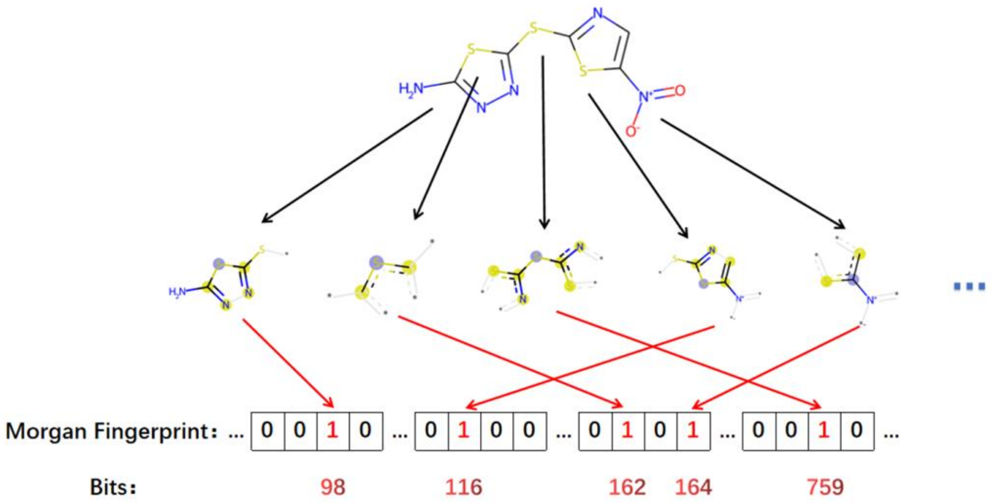
H₂N
S
N
N
S
N
S
N⁺=O
O⁻
...
Morgan Fingerprint: ... 0 0 1 0 ... 0 1 0 0 ... 0 1 0 1 ... 0 0 1 0 ...
Bits: 98 116 162 164 759

Figure 3 - Visual example of how Morgan Fingerprints are calculated

To determine if the Morgan fingerprints contain enough information to separate strong and weak inhibitors a UMAP projection of the data can be performed. A UMAP projection essentially groups the data points based on their similarities using the input features only. More similar points will be closer together. The UMAP projection from using Morgan fingerprints can be seen in Figure 4,5 and 6.

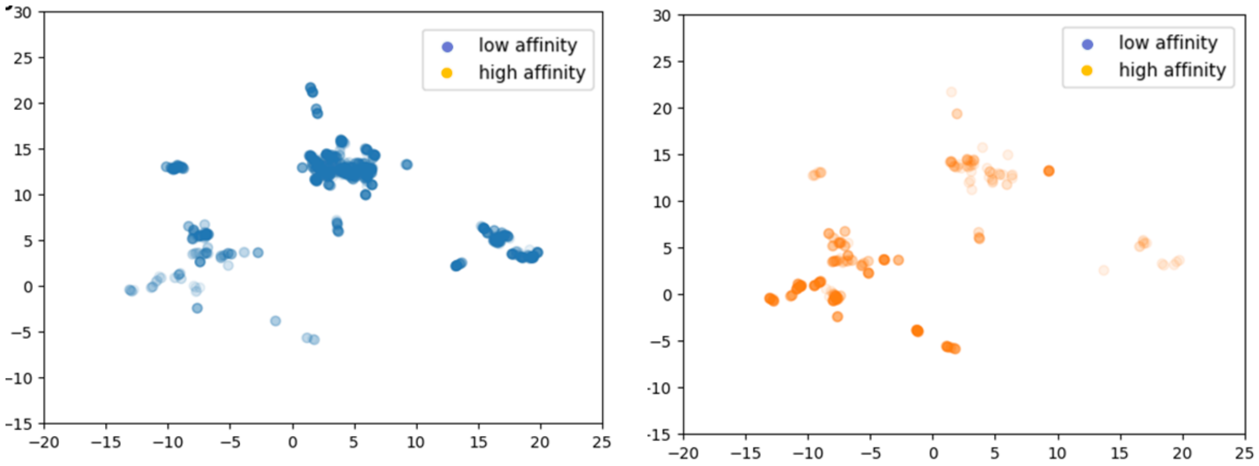

Figure 4 and 5 - UMAP projection/ Heatmap of low and high affinity inhibitors respectively

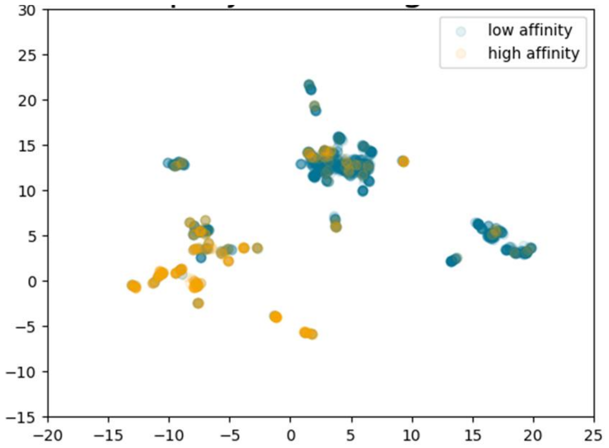

Figure 6- High affinity projections on top of low affinity projections to highlight the distinction between them.

Model Selection and Training

The machine learning model that was selected was an XGBoost machine this was decided after trying out different machine learning model with the automated machine learning package FLAML. The two best models were LGBM and XGBoost.

Evaluations / Tuning

After selecting XGBoost as the model then next step was to optimise the hyperparameters of the model. To do this, multiple techniques can be used. For this project Optuna, an automated hyperparameter optimising python module was used. To determine which hyperparameter was best, a cross validation metric using the F1 score of the model was used.

Results

Multiple metrics were used to determine the performance of the final model. These include:

Precision

Definition: Precision measures the number of true positives (items correctly labelled as positive) divided by the number of items labelled as positive (true positives + false positives).
Formula:

$$Precision = \frac{TP}{TP + FP}$$

Where:
TP = True Positives,
FP = False Positives

Interpretation: Higher precision means that more of the predicted positives are actually positive.

Recall

Definition: Recall measures the number of true positives over the number of false negatives and true positives

$$Recall = \frac{TP}{TP + FN}$$

Where:
TP = True Positives,
FP = False Positives

Interpretation: Higher precision means that more of the predicted positives were correctly predicted

F1 Score

Definition: F1 Score is the harmonic mean of precision and recall. It provides a single metric that balances the trade-off between precision and recall.
Formula:

$$F1 = 2 \times \frac{Precision + Recall}{Precision \times Recall}$$

Interpretation: F1 Score is particularly useful when classes are imbalanced. A higher F1 Score indicates better balance between precision and recall.

Confusion Matrix

Definition: A table used to describe the performance of a classification model on a set of data for which the true values are known. The matrix generally consists of four values: TP, TN, FP, FN.

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

Interpretation: Provides a detailed breakdown of the model's predictions vs. actual outcomes. It helps in understanding the types of errors a model makes.

AUCROC (Area Under the Receiver Operating Characteristic Curve)

Definition: AUCROC is a performance measurement for classification problem at various thresholds settings. The ROC is a probability curve and AUC represents the degree or measure of separability.

Interpretation: The higher the AUC, the better the model is at distinguishing between the positive and negative classes. An AUC of 0.5 suggests no discrimination (equivalent to random guessing), while an AUC of 1.0 indicates perfect discrimination.

The XGBoost machine had the final metric values:

Precision: 84%
Recall: 82%
F1 Score: 83%
AUCROC: 94%

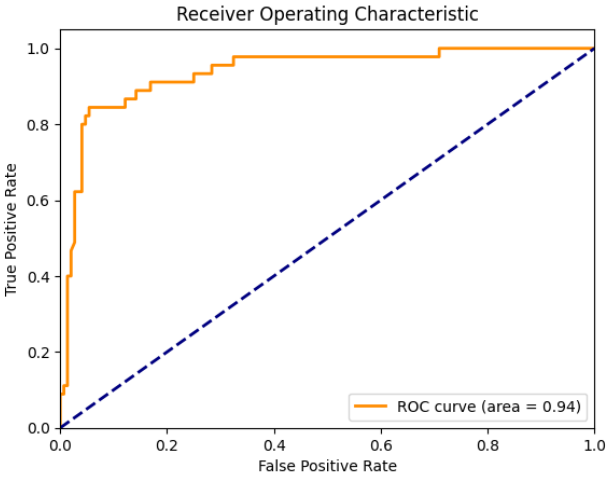

Figure 7 - ROC curve

Confusion Matrix:

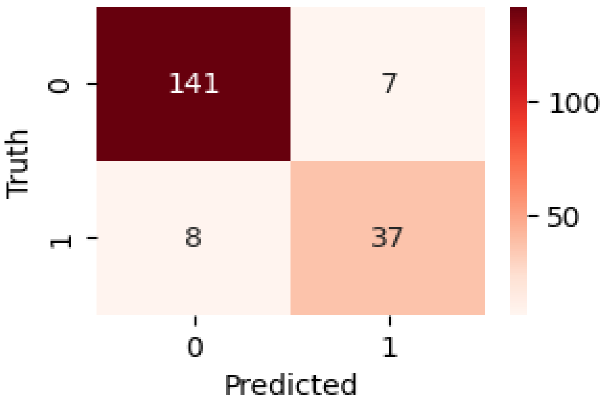

Figure 8 - Confusion matrix

Interpretation

In order to try to understand the decision process behind model, a SHAP analysis can be performed on the model, this entails finding the features which had the greatest impact on the model's prediction. By looking at the features that had the greatest positive impact on the model a molecule structure can be found with the arrangement of atoms and functional groups seen in Figure 9

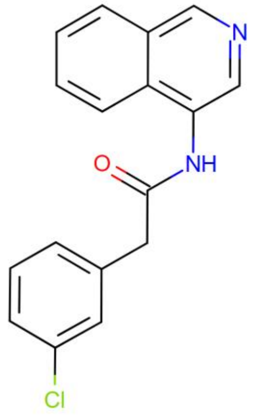

Figure 9 – Overlapping of features to give the molecule 2-(3-chlorophenyl)-N-(isoquinolin-4-yl)acetamide

Important lessons

Personally, I found every aspect of the project valuable including all the mistakes I made so I don't think I would do much differently. In terms of actual lessons that I learned, I've found that I would enjoy working within a research-based role and trying and testing new ideas is something that I really enjoyed. If I had more time for the project I would try to experiment with different ideas, an example of one would be using graph neural networks and using the information from the interaction of the inhibitors with Main protease. I did manage to complete the reading and understood the theory of using graph neural networks, but there wasn't enough time for me to implement the idea to make a machine learning model. This is something that I will probably try to do when I have some free time.

Throughout the internship I learned a lot about myself as well as researching and machine learning. I am much more confident in my abilities and applying for a PhD/ Masters. This experience has confirmed my interest in the field and made me confident in going through the academia route, and I believe it is a path that I would find fulfilling and enjoyable for me. This experience had an impact on my decision on which sector I would like to join; I feel like I would enjoy the bioengineering and modelling route now.

Resources used:

https://github.com/gmm/RDKit-on-Colab/blob/main/RDKit_Google_Colab.ipynb

https://pubs.acs.org/doi/pdf/10.1021/acs.jmedchem.2c00487

https://www.biorxiv.org/content/10.1101/2022.03.04.483012v1

https://github.com/MarcMoesser/Protein-Ligand-Interaction-Graphs