A Composite Deep Neural Network Solution to Predict and Generate Potential COVID-19 Antidotes

---

Our GitHub Implementation:
Github: https://github.com/rushirajsherlocked/SAMHAR-COVID-19
Colab Notebook: https://colab.research.google.com/drive/1NnsEVRVaD2dKnTWtP4p5iPtiQ8fxpkEG

Abstract:

A new coronavirus (CoV) identified as COVID-19 virus is the etiological agent responsible for the 2019-2020 viral pneumonia outbreak that commenced in Wuhan . Currently there are no targeted therapeutics and effective treatment options remain very limited. Rigorous efforts are continuously being made by scientists and researchers all over the world to find molecules which can serve as a potential antidote. But practically, there are millions of compounds and such an exhaustive approach manually is extremely difficult and infeasible. Moreover, it might be very well possible that all the current drugs may not be an effective cure for this pandemic disease. This is where the role of AI comes in to help in the pursuit of discovering/generating an antidote. We propose a novel Composite Deep Learning solution consisting of a predictive network architecture and a novel interleaved GAN architecture to ***Predict and Generate*** potential antidotes. This model learns from all the available compounds by combining various datasets like Moses, ChEMBL, Harmonizome, various government databases etc to predict an antidote from the existing drugs. At the same time, the model also generates new potential drugs using the Generative model which maps from the latent space of proven molecules like N3, Ebselen which act as a potential inhibitor. Finally, the predicted and generated molecules can be evaluated by docking them with various enzymes and proteins essential for survival of CoV(for instance M proteinase) of CoV and can be further sent for synthesis or clinical trials and FDA tests can be conducted.

1. INTRODUCTION

1.1 Importance of accurate Data

Data plays a pivotal role in designing any authentic AI solutions, especially when it comes to finding a potential drug which can help as a vaccine or drug for any diseases. In order to create a Deep Learning architecture, we reviewed various authentic research articles and journals which were recently published considering the protease structure of COVID-19 and its potential inhibitors. One of the most significant literature is "***Structure of Mpro from COVID-19 virus and discovery of its inhibitors***"
(Reference: https://www.nature.com/articles/s41586-020-2223-y).

Researchers have identified a ***mechanism-based inhibitor, N3,*** by computer-aided drug design and subsequently determined the crystal structure of COVID-19 virus Mpro in complex with this compound. ***"N3 is a potent irreversible inhibitor of COVID-19 virus Mpro"***

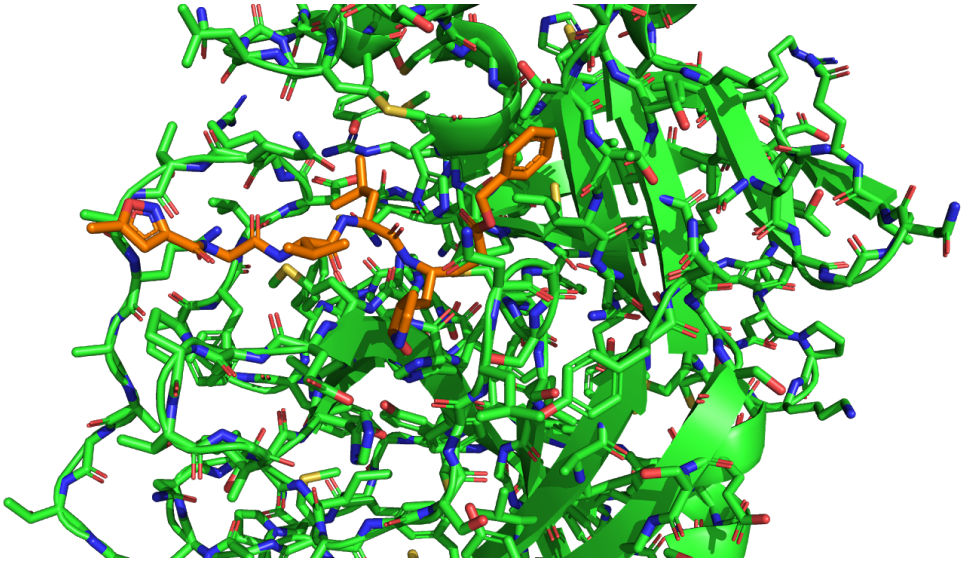
Next, through a combination of structure-based virtual and high-throughput screening, over 10,000 compounds have been assayed including approved drugs, drug candidates in clinical trials, and other pharmacologically active compounds as inhibitors of Mpro . Six of these compounds inhibited Mpro with IC50 values ranging from 0.67 to 21.4 μM. ***Ebselen*** also exhibited promising antiviral activity in cell-based assays.

Taking a proper look at the research findings, it can be taken into consideration that most compound exhibit similar structure and properties to N3 and also certain FDA approved drugs which are currently being used for the cure of ***HIV*** *can act* as a potential inhibitor

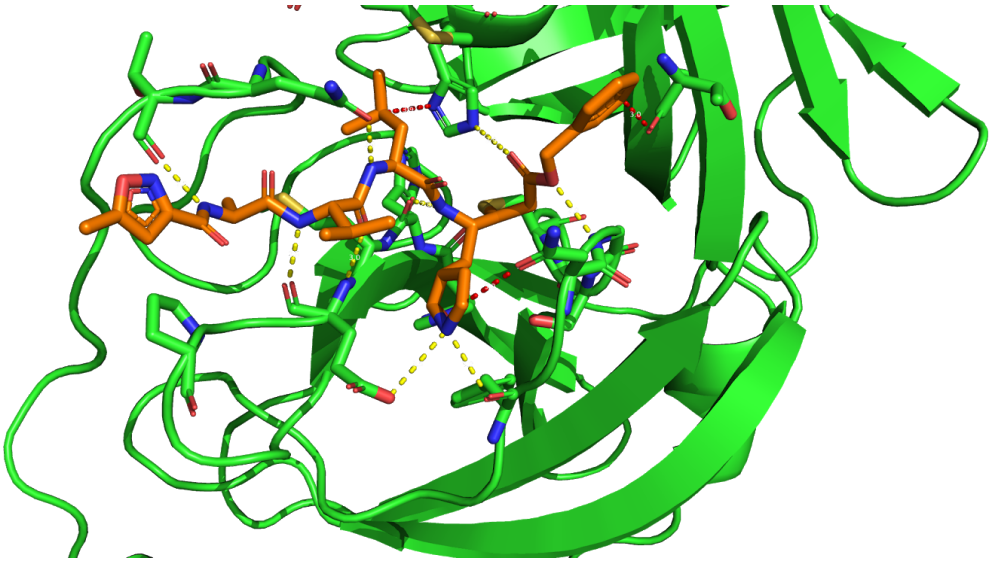
1.2 Visualizing The Target Virus

COVID-19 is the respiratory illness caused by the SARS 2 Coronavirus, it was unseen before in humans and rapid genetic data collection has shed some light into the origins. The genomes of several isolates of the virus are available, it is a ~30kB genome and can be found here: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512. Basic Local Alignment Search Tool (BLAST) results show close homology to the bat Coronavirus.

A crystal structure of the main protease of the virus was obtained by Liu et al., found at https://www.rcsb.org/structure/6LU7, corresponding to not-yet-published work. The structure is complexed with a ligand called N3, which serves as an excellent starting point for new drug candidate investigations. The following structure is visualized below using PyMol, which is excellent molecular visualization software available here: (Reference: https://pymol.org/2/ or https://github.com/schrodinger/pymol-open-source.)



The following photo shows the interactions happening in the binding site between the N3 ligand and the protein, with yellow interactions being Hydrogen Bonds, and Red interactions being sites where hydrogen bonds could be possible but are not occurring in the structure. It is also worth mentioning that the ligand shown here is what is called a "covalent inhibitor" meaning it is chemically bound to the protein. This can be seen on the right side of the orange ligand molecule, where it is connected to the yellow sulfur atom of the protein.

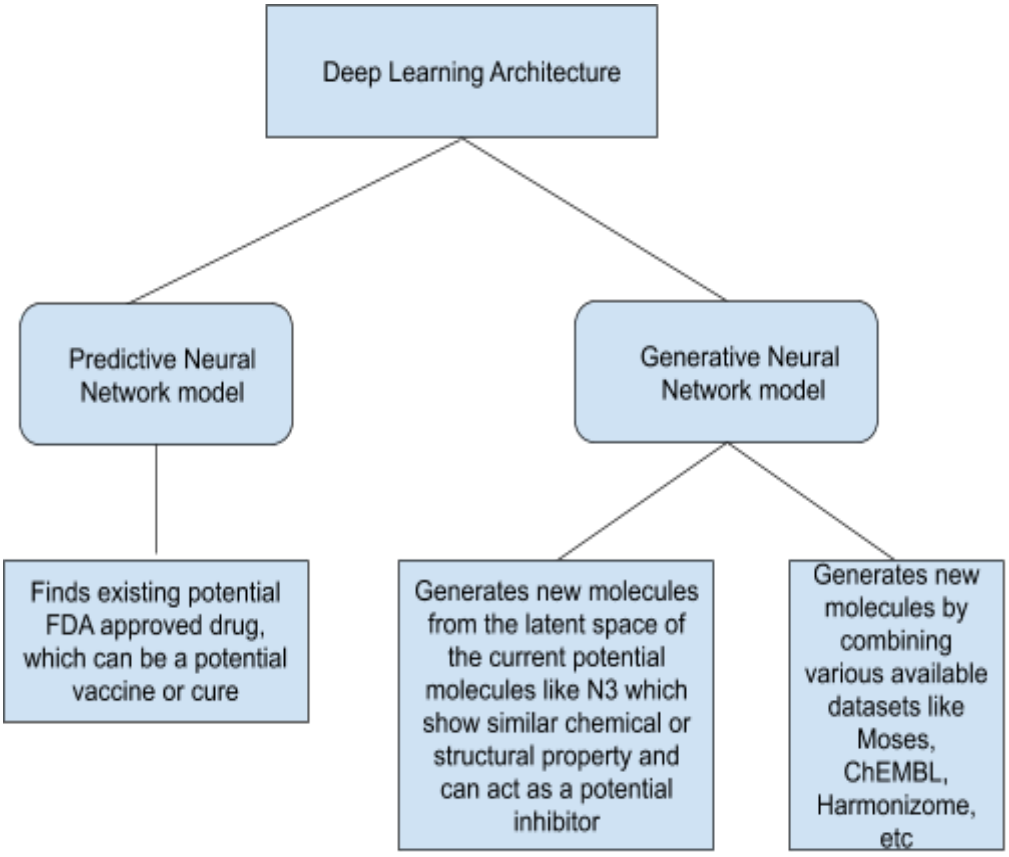


1.3 Proposed Solution

Keeping that in mind, here we propose a novel approach to combine various available datasets like Harmonizome, Moses, ChEMBL etc. Combining just two datasets - Moses and ChEMBL datasets give us 4 million compounds. Harmonizome contains 72 million functional associations as mentioned above. This provides us with a varied and a good amount of dataset which can be used to train neural networks especially the generative models which can further produce potential compounds from their latent space.

Further, the generated compounds can be docked with the enzyme using docking software like PyRx, Autodock Vina etc and their binding scores can be calculated. Out of all these generated compounds, some of them might show some similarity to the existing FDA approved drugs and some might be totally newly generated compounds. Steps can be taken further to synthesize such compounds and clinical trials can be carried out.

Here, we design a composite Deep Learning based Neural Architecture which uses two approaches - ***Predictive*** *and* ***Generative***. These can help us in predicting potential drugs from FDA approved drug databases using Predictive Network as well as in generating new molecules based on similar structure and chemical properties using a Generative network.



2.DATASETS

2.1 Dataset: Finding relevant compounds with activities

SARS and MERS are both coronavirus variants that are very similar and since their respective outbreaks, many biological assays have been done to test compounds on their main proteases. Bioactivities measured in papers by medicinal chemists and biochemists are tracked by The National Center for Biotechnology Information (NCBI) and are freely available. A database of protease inhibitors will be built using this data.

2.2 General Dataset Preparation

The dataset generation begins by simply searching the NCBI website, bioassay search found at https://www.ncbi.nlm.nih.gov/pcassay/advanced, to try and find relevant assays. An assay is needed to assess the *inhibition potential* of a molecule

We use the ***NCBI*** website to download different assays based on different targets, we searched various different kind of targets to get different assays and at last, we have a large list of assays and then get the molecular structure of the assays because the assays don't have SMILE strings in them, they just have compound IDs
(link: https://www.ncbi.nlm.nih.gov/pcassay/advanced)

As these targets are RNA proteinase, and RNA proteinase is also present in COVID1, the ligands which bind to this proteinase can bind to COVID19 proteinase up to some significant extent

The searches we used to generate a good AID (assay ID's) list are:

1. Protein target GI73745819 - SARS Protease
2. Protein target GI75593047 - HIV pol polyprotein
3. NS3 - Hep3 protease
4. 3CL-Pro - Mers Protease

3. METHODOLOGY

3.1 Solution Overview

We propose as a solution, a composite Deep Learning architecture which encompasses two approaches in a single model - *predictive approach* and a *generative approach.* As discussed earlier in the introduction, the predictive part of the solution is a Deep Neural Network trained on existing FDA approved drugs which predicts a drug that best reverses the collaterals of the COVID-19 virus. The generative part of the solution is a Generative Adversarial Network model trained on extensively large datasets with relevant filters, to ultimately generate a new drug which ideally reverses the effect of COVID-19 to a 100%. Eventually, both the generated and predicted samples are evaluated for their respective docking to the COVID-19 genome sequence using the PyRx tool and Autodock Vina. This associates a docking score with all generated and predicted samples, which can then be ranked for practical usage.

3.2 Predictive Approach

A Predictive Deep Neural Network :

A list of pubchem compounds with unknown activities is then used to predict activities.

(Reference: Predictive deep learning model using an "Edge Memory Neural Network", described herein https://chemrxiv.org/articles/Building_Attention_and_Edge_Convolution_Neural_Networks_for_Bioactivity_and_Physical-Chemical_Property_Prediction/9873599 and implemented at https://github.com/edvardlindelof/graph-neural-networks-for-drug-discovery.)

3.3 Generative Approach

We propose a novel ***"Inter-leaved GAN Architecture"*** (A new inner GAN Architecture inside the *Generator* of the outer GAN) .The aim of this model is to subsequently sample the latent space to generate new graphs based on the attributes of the compounds which show some inhibition property to the COVID-19 Mpro protease and also by combining two or more publicly available and approved datasets like Moses, ChEMBL, Harmonizome etc. The generated compounds may not be FDA approved as of now but can be synthesized further and lab trials can be conducted.

(Reference: Constrained Graph Variational Autoencoder", described herein https://arxiv.org/pdf/1805.09076.pdf)

3.4 Validating the predicted and sampled compounds

Validation - Molecular Docking Studies Using ***Autodock Vina***

Re-docking the N3 Ligand as a baseline

The first step in the validation of proposed structures is to re-dock the N3 ligand into the protease structure, to get a baseline for the energy score associated with their binding. It is important to note that the N3 ligand shown in the X-Ray structure is a covalent inhibitor, which means it actually reacts with the active site of the protein. This results in a much stronger bond between ligand and target than non-covalent inhibition.

This means that the re-docked structure may not be the same as the x-ray since it lacks the covalent bond to the protein. The following is a procedure for the docking of the N3 ligand into the protein receptor. The same procedure is used for all of the altar dockings of the candidate molecules. The procedure requires 3 programs:

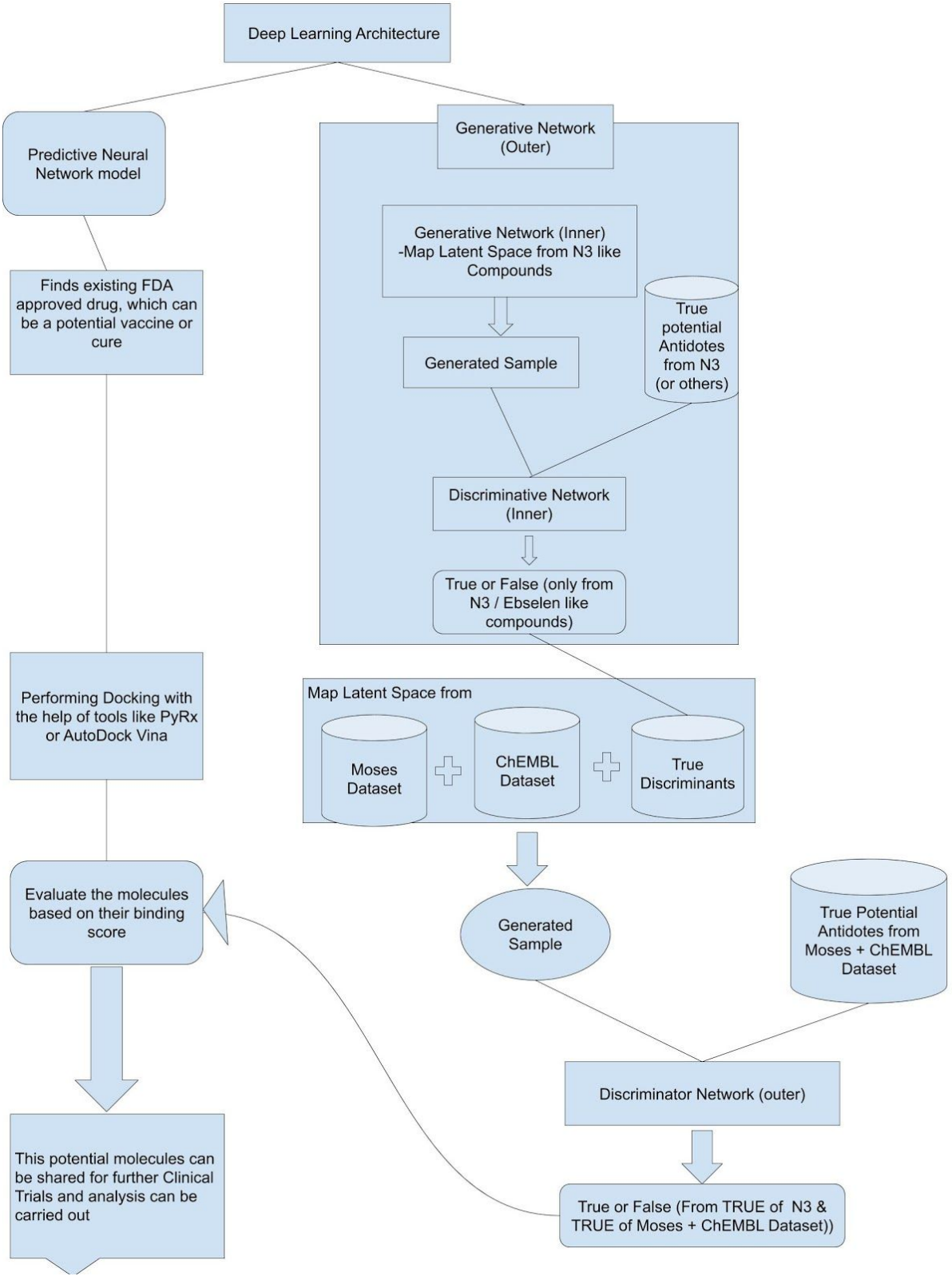
Pymol - https://pymol.org/2/ or https://github.com/schrodinger/pymol-open-source

Autodock Vina - http://vina.scripps.edu/download.html

Autodock Tools, found in MGL tools - http://mgltools.scripps.edu/downloads

3.5 Detailed Workflow

The complete Deep Learning Architecture along with the detailed workflow is as below:



3.5.1 Approach based on workflow:

(1) Data Preparation

The original network was only trained on ~450k unique smiles. Our first goal was to train a network from scratch that would be highly adept at generating robust, realistic molecules.
We combined data sets from two sources: i) Moses data set and ii) ChEMBL data set. Together these two data sets represented about 4 million smiles.

After cleaning the smiles using the cleanup_smiles.py script and only retaining smiles between 34 to 128 characters in length, './datasets/all_smiles_clean.smi' contains the final list of ~2.5 million smiles on which the initial network was trained.

(2) Neural Network Training

This is used to train an RNN on the known universe of SMILES to learn to very accurately generate novel small molecules. We then use this initial network to generate our candidate molecules.

Further training of the proposed workflow requires more computing power to be time effective. Given the proper constraints of GPU and high clock speed, the whole composite deep learning model (Predictive + Generative) can be trained on a very extensive combination of large datasets and robust conclusions can be drawn.

Below mentioned steps are the ideal sequence of our plan

(3) Generating Universe of SMILES

After completing the training, the new network would generate thousands of smiles. The generation process might take several hours though

(4) Finding Top Candidates from Initial Universe of SMILES

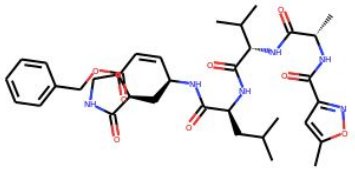
Out of all the generated compounds, we have two options either we can check all of the compounds by docking them with PyRx (suggested only if sufficient hardware resources are present) or we can use some evaluation technique which filters out the compounds which might be of less activity or not required by us

Finally, after selecting a particular set of smiles, save them to a CSV file and generate corresponding molecular structures of them. Save those structures in an SDF file. Further, those SDF files can be opened in PyRx and the docking score can be evaluated by docking them with COVID-19 protease.
Select, the molecule with the highest docking score(the more negative, the better)
Send those molecules for further analysis (like synthesis, clinical trials etc.)

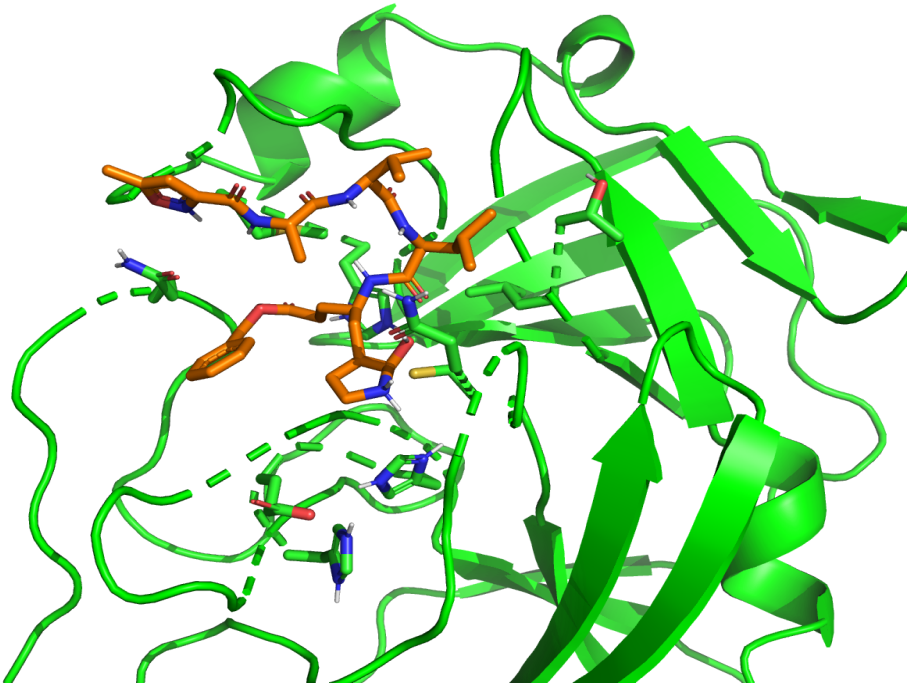
4. SOME PRIMARY RESULTS

4.1 Visualising the N3 ligand

```
#Visualizing the N3 ligand
Chem.MolFromSmiles("CC(C)C[C@H](NC(=O)[C@@H](NC(=O)[C@H](C)NC(=O)c1cc(C)on1)C(C)C)C(=O)N[C@@H](C[C@@H]2CCNC2=O)\C=C/
```



Following this procedure for the N3 ligand, we end up with a final lowest energy minimum of around -7.9kcal/mol. The exact value doesn't tell us much, because the specific parameters of the docking scoring function can vary, but this serves as a baseline for comparison of later candidates. The following is the lowest energy structure. We can see that it is in fact very different from the X-ray structure due to the lack of the covalent bond to the protein, with the N3 ligand sort of "bending back" in this conformation



4.2 Docking Potential Antidotes

Now for docking the candidates. The same procedure needs to be followed for each of the candidates, with the additional step below of loading the structures and saving them as PDB files, to be opened in AutoDockTools

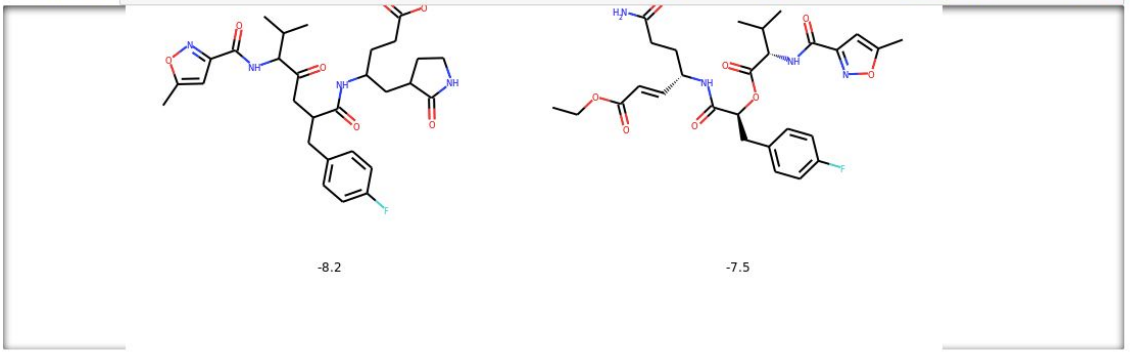
The generated compounds from this method are very strange for the most part, many with large strange rings. This is interesting because this paper is a relatively early example of generating molecular graphs and in the past little while have used a penalty on large rings such as the ones seen in these compounds. However, all is not lost, because several of these compounds still have interesting structure and are small and comparable to the N3 ligand.

4.2.1 Scores of Various compounds

Scores of various compounds

For the predicted compounds

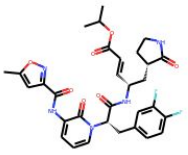
```
In [55]: rdkit.Chem.Draw.MolsToGridImage(best_predicted_mols, molsPerRow=2, maxMols=100, subImgSize=(400, 400), legends=["-8.
```



The Best predicted molecule from the Predictive Neural Network and the Generative Neural Network is shown below

```
In [59]: best_predicted_mols[7]
```

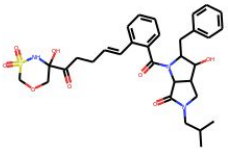
Out[59]:



The best one from the generative mthod is shown below, with a score of -9.8 kcal/mol

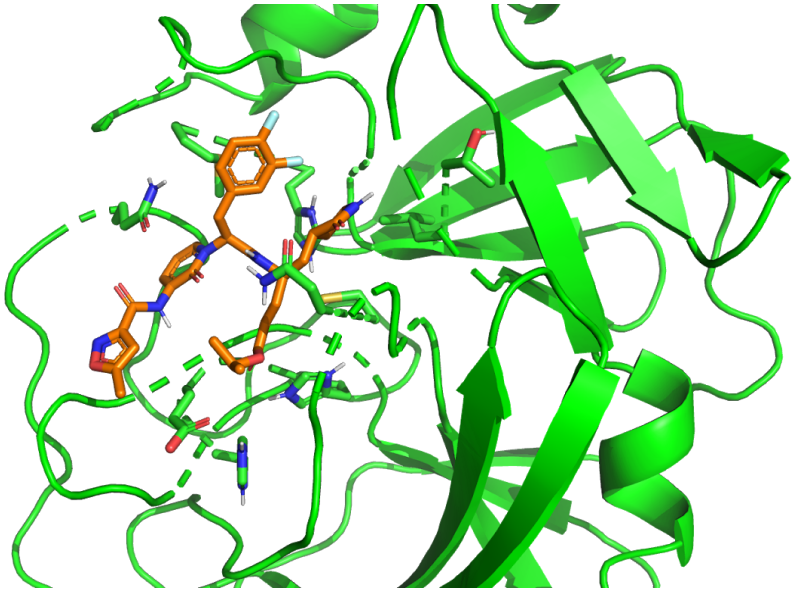
```
In [61]: generated[32]
```

Out[61]:

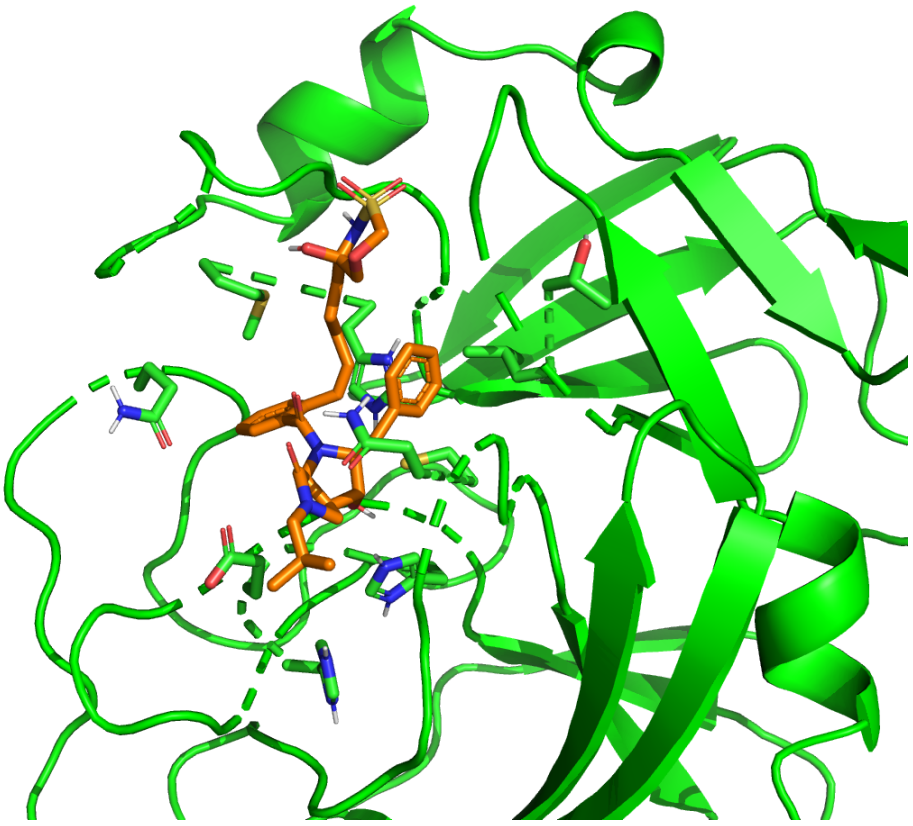


4.2.2 Visualizing the high scoring compounds in the active site

The following is the highest-scoring ***Predicted Compound*** mentioned above.



The following is the highest-scoring ***Generated Compound*** mentioned above.



5. CONCLUSION

A predictive deep learning model was trained on a self-generated set of protease inhibitors, and the PubChem literature was searched for 183 compounds that are similar to the n3 ligand. Predictions were made on these compounds and those with the 10 best predictive scores were docked to the ligand. The highest scoring compound is shown above and has a score of -8.7kcal/mol

A generative deep learning model was trained on a self-generated set of protease inhibitors, and 50 new compounds were sampled from the latent space of the model. The 10 most promising compounds were docked to the ligand. The highest scoring compound is shown above and has a score of -9.8kcal/mol

The best compounds from each method show significant gains over the baseline score of -7.9kcal/mol for the n3 ligand.

As Gujarat Biotechnology Research Centre (GBRC) have decoded the whole genome sequence of coronavirus, it might be more helpful to train a more accurate GAN model to generate some new compounds directly from the latent space of the coronavirus genome rather than using the SARS-2 family genome sequence when the dataset of the genome sequence is made public

The training of the above proposed neural networks is pretty intense and requires a lot of computational power which is currently unavailable to us, so we have trained and tested the above results for only structures similar to N3 but there are many other molecules that might be related and needed to be found out. Given proper computational resources, this task can be further enhanced and the speed can be increased drastically providing new insightful results.

If any chemical compound having sufficient affinity and resembles the active ingredient of any Ayurvedic medication/ Plant extract / Herbal medication can also be a promising Ayurvedic treatment for COVID-19 as this path has not been foreseen much in detail

REFERENCES

We gratefully thank all the researchers and developers who constantly work hard and open-source their findings to develop the solution to battle this pandemic to find better cure or vaccination. The above-proposed work is made possible only because of their initial and continuous research and findings.

https://www.nature.com/articles/s41586-020-2223-y

https://chemrxiv.org/articles/Building_Attention_and_Edge_Convolution_Neural_Networks_for_Bioactivity_and_Physical-Chemical_Property_Prediction/9873599

https://github.com/edvardlindelof/graph-neural-networks-for-drug-discovery

https://arxiv.org/pdf/1805.09076.pdf
https://github.com/microsoft/constrained-graph-variational-autoencoder

https://www.rcsb.org/structure/6LU7

https://github.com/schrodinger/pymol-open-source

https://pymol.org/2/

https://www.ncbi.nlm.nih.gov/pcassay/advanced

http://vina.scripps.edu/download.html

Generative Recurrent Networks for De Novo Drug Design

https://github.com/topazape/LSTM_Chem

https://github.com/molecularsets/moses

https://www.ebi.ac.uk/chembl/

http://amp.pharm.mssm.edu/Harmonizome/

https://pymol.org/2/

tinkavidovic/competition