# **Repurposing of Drugs for SARS CoV 2 Virus using Ensemble Learning and Generative Adversarial Networks to create similar and high performing Molecules**

Viren Varma | Mohd. Adain W

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2 virus), responsible for the Coronavirus disease 2019 (COVID-19) disease has caused massive loss to life and property. With still an unclear way of treatment, the virus is continuously creating newer more powerful variations in mere months. Right now there is no specific medicine to treat the covid-19 virus in infected patients with modern drug development procedures taking around 5-10 years to get a drug to the market. Drug Repurposing is a method to test drugs currently in circulation in the market against the new virus to create a quick and safe line of treatment. In this paper, we propose a method for the repurposing of drugs using an Artificial Neural Network (ANN) using the Ensemble Learning Algorithm. And Then using such a drug's molecular fingerprint as a dataset to train a Generative Adversarial Network as the differentiator to create more similar drugs. Which can then be created in a lab and tested for further use.

Key Terms

Drug Repurposing, Ensemble Learning, Generative Adversarial Network, Random Forest Regressor

1. INTRODUCTION

Even though the close relation of SARS CoV -2 having close relation to SARS CoV - 1 there have been no specific drugs available for it’s treatment. As mentioned earlier with newer drugs taking 5-10 year to be developed needing millions of dollars in funding. Creating novel inhibitors might be a very good option but with the rapid variants of the virus we need an immediate line of treatment.

The inhibition of viral proteases necessary for proteolytic processing of polyproteins has been a successful strategy in the pharmacological treatment of human immunodeficiency virus (HIV) and hepatitis C respectively, proving the capability of protease inhibitors for the treatment of viral infections. Similarly, the main protease of SARS-CoV-2 is thought to be essential for viral replication and, therefore, is regarded as a promising target for antiviral pharmacotherapy. The crystal structure of the SARS-CoV-2 main protease was recently solved16 enabling the rational design of specific inhibitory compounds.

The SARS CoV -2 virus and its variants need a main Protease enzyme in the body to multiply. Computational Drug Repurposing uptil now has screened the compounds for binding at the recently solved crystal structure of the main protease (Mpro) of SARS-CoV-2. Post the shape screening process, multiple docking protocols are applied followed by locating the molecular descriptors relevant for pharmacokinetics to narrow down the number of initial hits.

Molecular dynamics simulations are conducted next to confirm the stability of docked binding modes and comprehensively quantify ligand binding energies. Post the evaluation of off-target binding, a list of commercially viable compounds, with binding affinity to the target protease that is predicted to be more favorable than that of the co crystallized peptidomimetic compound, is released.

In this study, Molecular Fingerprinting is done using the SMILES compound formula of the molecule and passed in an Artificial Neural Network to calculate its binding affinity to Mpro enzyme and graphed on the basis of its affinity to bind with the active points on the enzyme.

Generative modelling comes under unsupervised learning in machine learning that involves generating and learning the regularities from input data so that the model outputs new examples that could have been reasonably drawn from the original dataset. The generator model that we train to generate new Molecular Fingerprints of molecules with binding affinities comparable to the repurposed drugs, and the discriminator model (ANN previously trained) that tries to reinforce higher value binding affinity molecules.

To the best of our knowledge, no specific drugs have been created or are available to treat SARS-CoV-2 despite its close relation to the SARS-CoV virus. Thus, there remains an urgent need

for the development of specific anti-viral therapeutic drugs to overcome SARS-CoV-2. The proposed project aims to transform drug discovery in the context of treating COVID-19 from a slow and expensive process to a rapid and inexpensive one. Machine Learning tools coupled with molecular dynamics simulations on High-Performance Computing systems will significantly accelerate the pre-clinical drug-design procedure by an efficient drug-screening procedure and generation of novel drug compounds that can act as potent medicines. The objectives of the proposal are in complete alignment with the scope of the present call.

1. BACKGROUND
2. Drug Repurposing: Drug repurposing has the capability to make available medications with government approved safety protocols to new patient populations. Countless examples exist for the identification of signs for existing molecules, a lot many coming from accidental findings or from focused recent efforts specifically towards the mode of action of a specific drug. In recent times, there is a need for new approaches to research and development of drugs. With the emergence of big data repositories and connected analytical methods the integration of Artificial Intelligence and Machine Learning are revolutionizing the world of pharmaceuticals and drug discovery. Repurposed drugs can not only save valuable time and money by developing drugs faster but them being approved and well tested can also help prevent any side effects or complications.
3. Molecular Fingerprinting: involves unsupervised machine learning algorithms to learn vector representations of molecular substructures that point in similar directions for chemically related substructures. Compounds are encoded as vectors by summing the vectors of the individual substructures to predict compound properties. The basic substructure vector embeddings are then generated by training unsupervised machine learning algorithms on a database of compounds that consists of all available chemical compounds. The once pretrained resulting model yields excellent vector representations, and overcomes drawbacks of common compound feature representations like sparseness and bit collisions.
4. Ensemble Learning: is a Machine Learning Algorithm that takes predictions from multiple other support vector machines or Machine Learning Models or multiple readings from the same machine learning model to average out and give its output to increase the overall accuracy. In this paper we will be using the bagging and stacking methods in ensemble learning to create the discriminator and generator in the generative adversarial network.
5. Recurrent Neural Networks: A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. As our problem statement describes sentence type input data (Natural Language Processing) where sequence matters we are using an RNN. Unlike feedforward and convolutional neural networks (CNNs), recurrent neural networks are distinguished by their “memory” as they take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depend on the prior elements within the sequence. While future events would also be helpful in determining the output of a given sequence, unidirectional recurrent neural networks cannot account for these events in their predictions.
6. Generative Adversarial Networks: Generative Adversarial Networks (GAN), are an approach to generative modelling using deep learning methods to train a generative model by framing the problem statement similar to a supervised learning problem with two models: the generator model that we train to generate new data, and the discriminator model that attempts to classifies this new data as either from the domain or synthesized. The two models are trained simultaneously in an adversarial, until the discriminator model approves the created data about half the time, meaning the generator model is generating plausible data confirmed by the generator model.

III. SETUP AND COLLECTION OF DATA AND MOLECULAR FINGERPRINTING

ChEMBL is a systematised database of bio-active compounds with drug-like properties. It collects chemical, bioactivity and genomic data to help the decoding of genomic information into effective new drugs. It has over 2 million molecules of data saved. For our current project we extract the target drug molecules data for SARS CoV 2 molecule, which is our main protein.

We have consolidated a dataset comprising of the docking scores and binding energies of PIs of the SARS-CoV-2 virus obtained from various docking simulations (Pant et al.,2020; Fischer et al., 2020; Kouznetsova et al., 2020; Agrawal et al., 2020).

We then iterate and remove data with missing items and reiterate them with their canonical SMILES values for 9001 molecules. SMILES (Simplified Molecular Input Line Entry System) is a chemical annotation that allows a user to represent chemical compound structures in vector form that can be used by the computer. SMILES is an easily learned and workable annotation. We save the database as CSV and use the smiles value for Molecular Fingerprinting.

Inspired by natural language processing techniques we use the mol2vec introduced by Samo Turk to convert SMILES data for each drug into molecular substructures that point in similar directions for chemically related substructures. Compounds are encoded by adding together the vectors of each substructure and, for instance, being fed into a supervised machine learning algorithm to predict compound properties. The Mol2vec model which is once pretrained, yields dense vector representations, and overcomes the issues of common compound feature representations such as sparseness and bit collisions. The prediction potential is demonstrated on several compound property and bioactivity data sets and then cross referenced with results obtained for Morgan fingerprints as a reference compound representation.

IV. RANDOM FOREST REGRESSOR- DISCRIMINATOR

Autodock Vina is a very popular, and highly cited, open source docking program of scoring, minimization, and re-docking on carefully curated training datasets allowed to develop a simplified scoring function with optimum docking performance to empirically score the binding affinity. Our data uses this score to represent the binding affinity of each molecule with the Main protease molecule.

Our first milestone in our project is to create a Discriminator Network that is able to calculate the Autodock Vina score (Binding affinity) of the molecule from its molecular fingerprint. Once the compound’s SMILES scores are converted to vector format they are then run on a custom built Random Forest Regressor Model of maximum depth 7 and random state 1. A Random Forest Generator here acting like our discriminator is an ensemble learning algorithm composed of homogeneous strong learning algorithms that are stacked by the Bagging Method. The algorithm bootstraps the current data data to create data trees. Trees of shallow depth cause low variance and a high bias and whereas trees with high depths cause high variance and low bias.

Here we required a low variance so the max depth of every tree has been restricted to only 7. The model is fitted over vectorized values of the compounds and their Autodock vina score which is the binding affinity of the compound to the main protease.The model is further tested on the test set and other molecules to detect other drugs that can bind to the main protease.

With the input as the fingerprinted vectors of the canonical SMILES values to the model, the model outputs the predicted Autodock vina score which is mapped to the actual values in our dataset and MAE (Mean Average Error) and MSE (Mean Standard Error) are calculated.

V. GENERATIVE ADVERSARIAL NETWORK

Using the Deep Reinforcement Learning for De novo drug design library created in 2018 by Olexandr Isadev we move forward to our second milestone in creating a Generative Adversarial Network to generate synthetic synthetic data that closely follow the underlying probability distribution of the data on which they were trained.Wherein the deep learning model developed in the first milestone in generating the Regressor Network acts as a Discriminator.

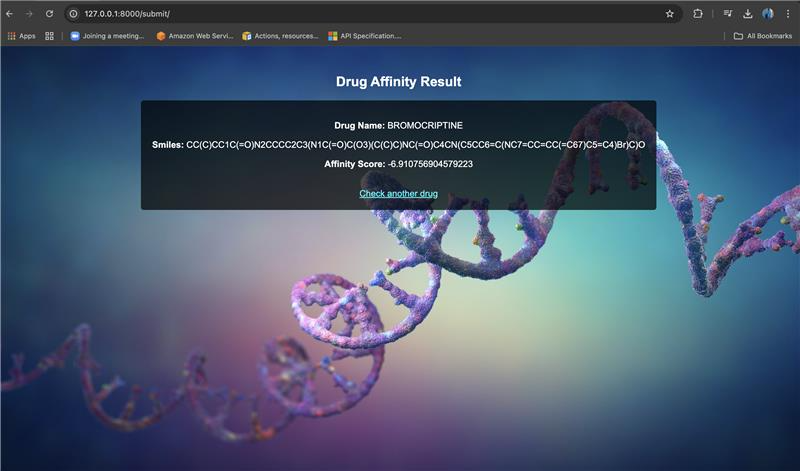
We use alphanumeric characters and special characters to create our token to form our generator data to use. As we are dealing with sequential data in a form of natural processing language we use the generator model in our GAN as a stack augmented Recurrent Neural Network.

The generator network is designed to have 1500 as the width, and 200 layers as depth of augmented stack memory. Here stack depth is determined by the length of the longest subsequence that needs to be generated. The network has 1500 hidden layers as we want the neural network to understand a very complex nonlinear function and wherein deeper neural networks generally tend to have a better performance. Specifying a learning rate of 0.01 we use the Adadelta optimiser for training. We specify our layer as GRU which uses a gating mechanism in recurrent neural networks which is like a long term - short term memory (LTSM) with a forget gate but without an output gate and lacks the number of parameters than LTSM.

Finally we run our training data of all compounds with their molecular fingerprints against their Autodock vina score using Random Forest regressor as a discriminator and create RNN model as generator to output the generated binding affinity which is the calculated Autodock vina score and reinforce this value to become maximum. We create a total 10000 molecules and consider only the valid Compound formulas generated.

VI. WEB APPLICATION

To enhance accessibility and facilitate rapid drug screening, we developed a web application powered by FastAPI for high-performance, asynchronous processing. The application enables researchers to input chemical formulas, SMILES representations, or drug names to predict their binding affinity to the SARS-CoV-2 main protease in real time. The backend integrates our Random Forest Regressor model for affinity prediction, providing an efficient and scalable solution for drug repurposing. Additionally, the system supports comparative analysis of multiple compounds and visualization of molecular properties. Future enhancements include real-time docking simulations, drug similarity searches, and an API for high-throughput screening, further streamlining AI-driven drug discovery.



VII. RESULTS AND FUTURE SCOPE

The results of the project were determined in two phases. After the first milestone of creating a Discriminator model : Random Forest Regressor was mapped on selected test data which was 40 percent of the original data. Mean Average Error was found to be 0.3443, Mean Squared Error was 0.4737 and R2 value was found to be 0.8187 are calculated over the actual Autodock values versus the predicted values as shown in fig1.

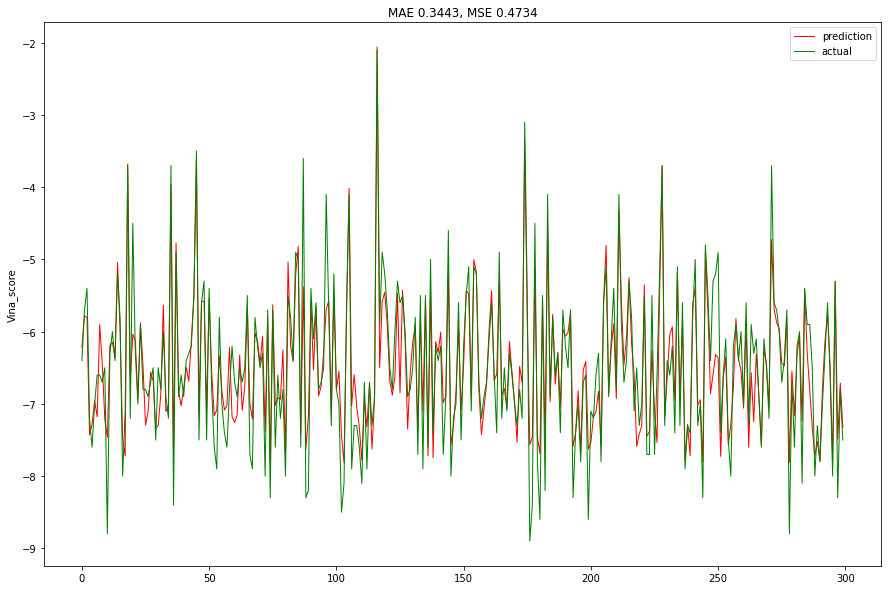


Fig 1: Autodock Vina score vs the index of the molecule. Red: Predicted, Green: Actual

The difference between the actual data collected from docking calculations over supercomputers versus the predicted data of our model are extremely close and the discriminator model is able to map all the variations in the data.

After the completion of the second phase in the project with the molecules being generated we need to check if the model had reinforced the creation of molecules with higher binding affinity to Main protease of the SARS COV 2. For which we first iterate to all the molecules and remove invalid Molecular compounds created and calculate the binding affinity of the valid molecules using our discriminator network.

Out of 10000 molecules generated by the generator network 6321 molecules were found to be valid and their respective Autodock vina score was calculated and mapped as shown in fig2.

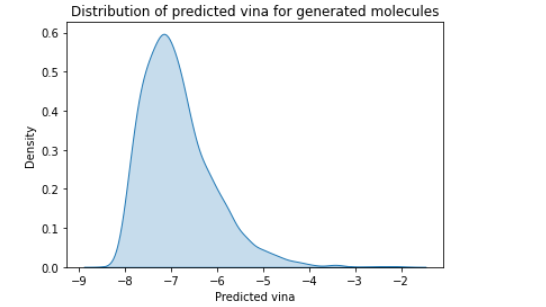


Fig 2: Density of molecules Vsr Auto Dock Vina

Further these molecules with the designer binding affinity can be created in a lab and tested over a period of time to come up with a more permanent solution to the crisis of the treatment of this deadly virus. And until then repurposed drugs already in circulation in the market can be a viable solution.

The drug databases consist of a large number of drugs and biomolecules, however, screening each of the drugs from these databases to identify the potential protease inhibitor experimentally, or computationally, is an expensive and time-consuming process. The machine learning algorithms can be executed on high-performance GPUs, and therefore will be able to significantly accelerate the drug-screening process.