A summer internship project report on

Ai based Drug Discovery for SARS-CoV-2

Submitted to

School of AIDE

Indian Institute of Technology Jodhpur

by

Harsh Khandelwal

106119048, 4th Semester

Department of Computer Science and Engineering

National Institute of Technology, Tiruchirappalli

Supervisor

Dr. Manish Agarwal, Faculty of Business Analytics

July 2021

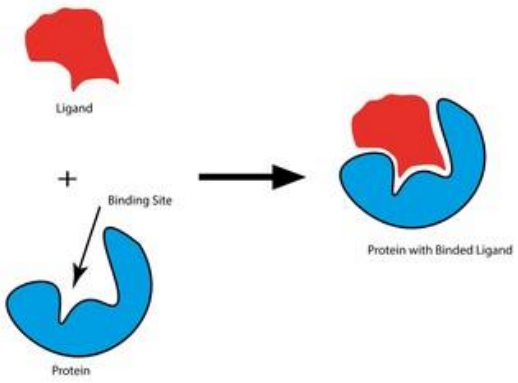
Introduction:

At the time of writing, we have faced many members of SARS family. SARS-CoV-2 can be stopped by blocking their active sites, this process is also known as protease inhibition. We are using multi-objective genetic algorithm NSGA-II with Adversarial Auto encoders to obtain a ligand with maximum affinity score and also following the constraints such as toxicity, practicality of synthesizing the ligand, etc. We used the PyRx open-source tool kit for calculating and optimizing the docking of the ligand with the protease using AutoDock vina.

Methodology:

Scoring Methods

Binding Affinity Score: This scoring function gives the estimate of how strong bond is formed between ligand and target protease; it tells the energy released for that bond also known as binding energy. In order to get a good binding affinity, the ligand has to work like a key while protease behaving as a lock, ligand has to properly oriented and at the correct site in order to make a strong bond. We used the virtual docking tool AutoDock vina which is included in software named PyRx.



Synthetic Accessibility: This scoring function allows us to get an estimate of the efforts required to synthesize the ligand. Since we want to mass produce the drug we will be able to decide which drug would be economically best fit for the nation.

Toxicity Estimation: We used the PAINS filter which is used to estimate the toxicity of the drug to human body.

Natural Product Likeness (NP): This scoring function tells how our drug is similar to the natural compounds which are produced naturally. In this way we can have a hold on the natural balance of our body. Since we are used to consume the products which are naturally available, therefore if the drug is having a good NP score implies that our body wouldn't have a problem to consume it.

Quantitative Estimate of Drug-likeness: The QED is based on a method for multi-parameter optimisation known as 'desirability functions'. This scoring function tells us whether a molecule can be used as drug. It uses Lipinski's rule of five, which includes physiochemical properties of molecules such as size, how often its used, etc.

Adversarial Autoencoders (AAEs)

This is a probabilistic autoencoder that uses Generative adversarial networks (GANs) and Variational autoencoders (VAEs). GANs are made of two parts generator and discriminator and the variational autoencoder has 3 major components encoder, decoder and latent vector. The VAEs are able to generate meaningful outputs only when we give an input which is from the same distribution as our training input, since we as humans don't know where that distributional space is so we use AAEs which learns the parameters of distribution instead of latent vector, and side by side for better generative outputs the discriminator is also be trained.

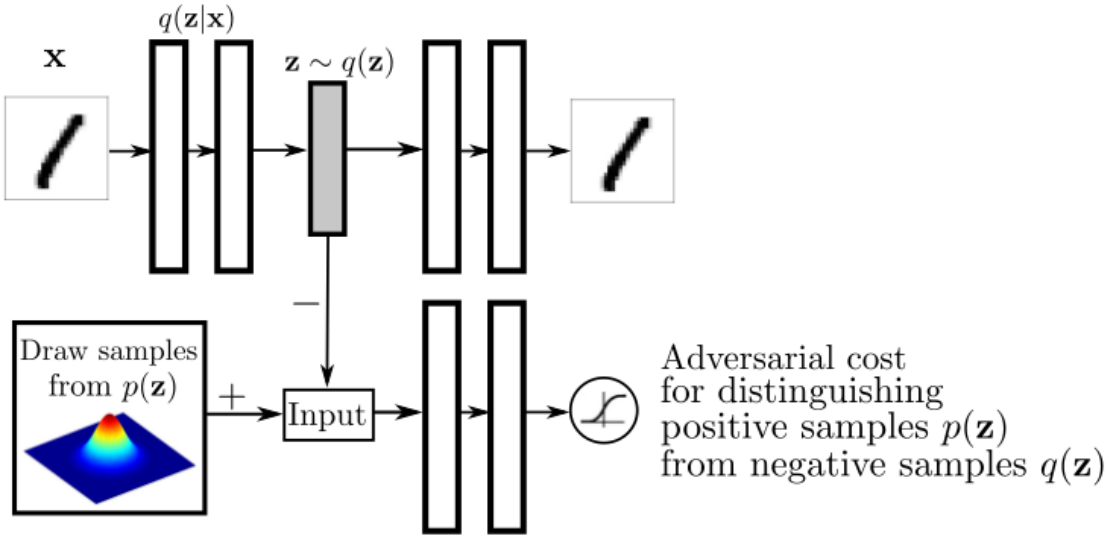


Figure 1: https://towardsdatascience.com/paper-summary-adversarial-autoencoders-f89bfa221e48

Evolution

We used the simplified molecular-input line-entry system (SMILIES) representation for the ligands. Previously the new generation was formed by modifying the parent molecule by means of replacing, inserting or deleting the symbol from the molecule but that could generate invalid molecules very often and also, we will not be able to know weather we are improving globally or not. Therefore, we used Adversarial Auto encoders which will be trained on the previous generation and their scores to create valid molecules which could surpass the old generation. Since the Neural networks are not good with Graph's data structure, we used the vocabulary to covert those graphs into vectors that neural network can understand.

NSGA-II

We start with the pre-existing drugs as our initial population then with the help of different scoring function we will perform non-domination sort and crowding distance sort, then we take the top drugs among them and we apply our new evolution technique in order to generate new samples. After generating and validating new samples, we than combine them with the best from old population and discard rest of them. After which the AAEs are trained on the new population. Combining the best old population ensures us that the model wouldn't lose the track of the improvement.

And we keep on doing this until our expectations are reached. Passing the best ones from old generations will help our neural network on track.

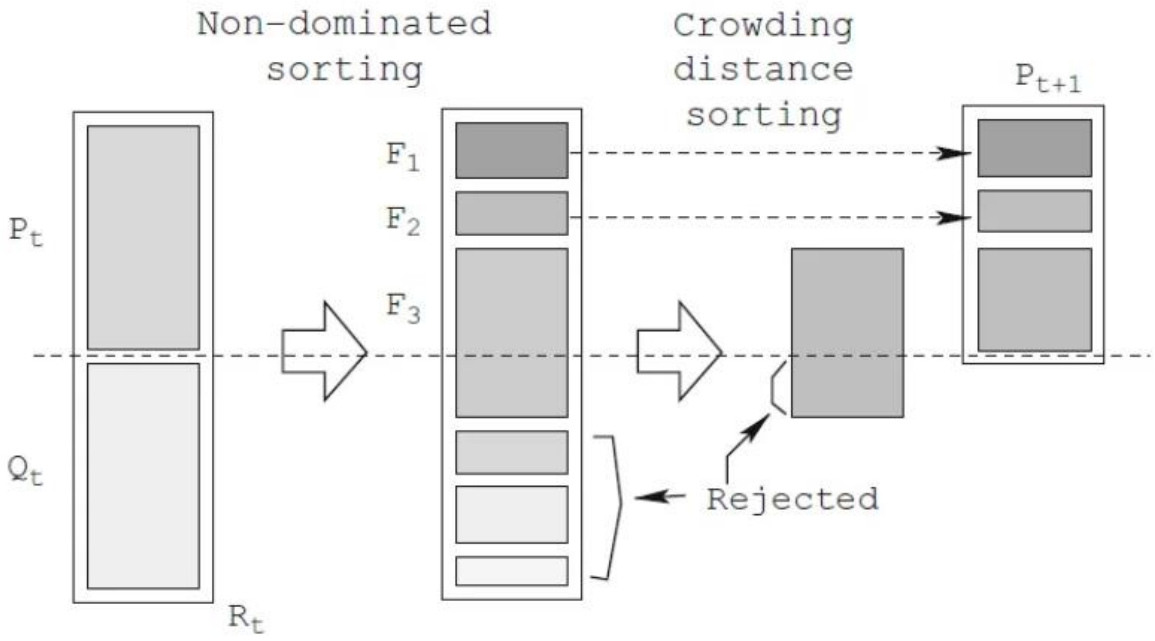


Figure 2: http://oklahomaanalytics.com/data-science-techniques/nsga-ii-explained/

Results and discussion:

I was able to implement the above methodology on small scale due to hardware and time constraints. The model was trained on a subset of MOSES dataset after few generations we were able to get a medium binding affinity score (around -7.1 kcal/mol) and other scores like NP, SA scores.

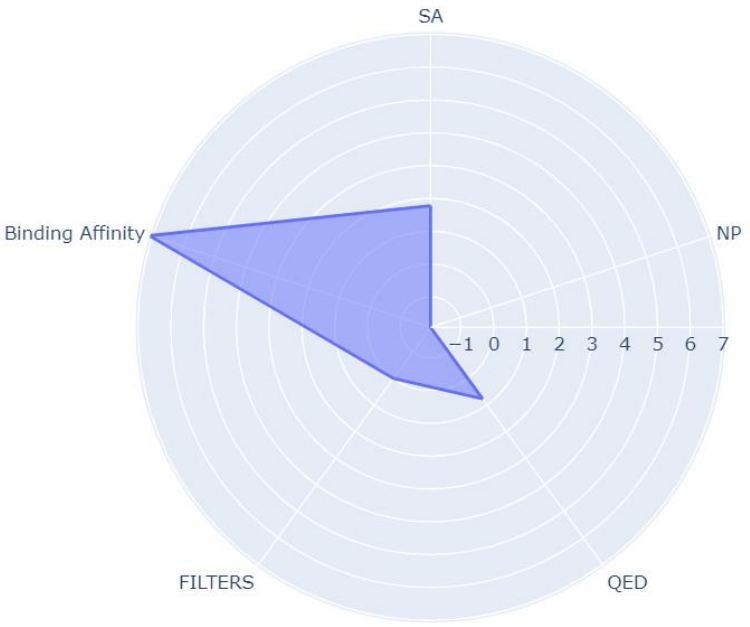


Figure 3 SMILE: 'O=C(Nc1ccccc1)c1ccc(COc2cccc(F)c2)o1'

There is a scope to train the model on complete dataset and for more number of generations for better results. This method can be applied to do drug discovery for such viruses just by changing the constraints of scoring functions and affinity scores which can be generated for every population at once using PyRx. I used ANNs but we can use a bunch of different generative networks such as Variational Autoencoder, etc.

Conclusions:

We tried to improve the generation of new drugs in the paper NSGA-II for SARS-CoV-2 [1]. Instead of replacing, deleting, inserting new symbols we used Adversarial Autoencoders which will create valid compounds more often and we could have a track on weather our generated drugs are getting better or not. This approach is not just restricted to Covid cure, we can extend this approach further as discussed above.

References:

1. Project Github: https://github.com/HarshKhandelwal1552/Drug-discovery-for-SARS-CoV-2-using-NSGA-II-and-Autoencoders
2. NSGA-II: https://www.iitk.ac.in/kangal/Deb_NSGA-II.pdf
3. NSGA-II for SARS-CoV-2: https://arxiv.org/pdf/2005.02666.pdf
4. SMILES format: https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system
5. Moses: https://arxiv.org/pdf/1811.12823.pdf
6. Adversarial Autoencoder: https://arxiv.org/abs/1511.05644
7. PyRx: https://pyrx.sourceforge.io
8. All SMILES Variational Autoencoder: https://arxiv.org/pdf/1905.13343v2.pdf