Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

Drug Discovery using Deep Learning

[1st] Jayesh Sharma
Department of Computer Engineering
New Horizon Institute of Technology and Management Thane, India

[2nd] Shyam Mourya
Department of Computer Engineering
New Horizon Institute of Technology and Management Thane, India

[3rd] Rounak Rai
Department of Computer Engineering
New Horizon Institute of Technology and Management Thane, India

[4th] Rushikesh Nikam
Department of Computer Engineering
New Horizon Institute of Technology and Management Thane, India

***Abstract*— The SARS-CoV-2 infection has killed over 3.9 million people, indicating there is an urgent need for effective treatment. This, however, cannot be accomplished with present drug development or application systems, as it takes several years for newly discovered drugs to reach the market. In this project we have tried to identify commercially available anti-viral drugs, synthetic molecules that could potentially disrupt SARS-CoV-2's viral components. Our aim was to bind our molecule with the main enzyme of SARS-CoV-2 slowing the virus's replication process enabling our body to fight against the virus. We first took a large number of molecules and fed them to an RNN-LSTM. The molecules which would be fed would be in a format similar to a string. The RNN would then identify the patterns and rules from these molecules using them to generate molecules which are currently not in existence but could be synthesized later in the future. Later we combined the new molecules and the pre-existing molecules, forming a new set. A diverse subset is then selected on which molecular docking was then performed with the SARS-CoV-2 virus's main protease and potential inhibitors were identified.**

Keywords—Deep Learning, Natural Language Processing, Long Short-Term Memory, Drug Discovery, COVID-19

I. INTRODUCTION

The infection of SARS-CoV-2 which originated in Wuhan, China has infected more than 180 million people and more than 3.9 million have already lost their lives[1] due to the infection caused by it referred to as 'COVID-19' and the resulting complications arising from it. This virus does not discriminate between men and women, rich and poor, religion, ethnicity; anybody can get infected. The case of its infection can prove fatal if the individual is already suffering from illnesses such as high-blood pressure, diabetes, lung diseases etc. It is said that SARS-CoV-2 originated from bats and through an intermediatory animal found its way to humans. As of this moment there are no reliable treatment options available against this disease.

There are two approaches we can employ against this virus:

1. Vaccines

A vaccination is a biological preparation that induces successful acquired immunity to a specific infectious disease. A vaccine usually involves an agent that looks like a disease-causing microorganism and is mostly produced from damaged or destroyed versions of the microbe, its toxins, or one of its surface proteins. The agent activates the body's immune system to recognize and destroy the agent as a danger, as well as to recognize and destroy any microorganisms associated with that agent that it may meet in the future. If fully vaccinated a person is less likely to catch an infection or even if one catches it won't be that severe compared to the original infection. The problem with vaccines is that it won't help people that are currently getting infected and it takes a long time to fully-vaccinate the entire population. Also it is difficult for poorer nations to procure them.

2. Small-molecule drug development

A small molecule is an organic compound with a molecular weight less than 900 Daltons. Large structures such as nucleic acids, proteins and many polysaccharides are not considered small molecules despite the fact that their constituent monomers are often regarded as small molecules.

Our project takes the second approach to find such small molecules which potentially act against SARS-CoV-2. Our aim was to find such drugs which can be repurposed for the fight against COVID-19 and predict the structure of drugs which are not currently in existence but which could be synthesized later which can act on this virus. We wanted to find molecules which slow down the replication process of the virus enabling our immune system to catch up with the virus.

Let us understand how this virus enters the host cell. The virus's spike protein attaches to a protein on the surface of cells, called ACE2. It is normally involved in blood pressure regulation. As the coronavirus binds to it, chemical modifications occur that essentially fuse the membranes surrounding the cell and the virus, allowing the virus's RNA to penetrate the cell. The virus then hijacks the host cell's protein-making machinery in order to translate its RNA into new virus copies. A single cell may be forced to develop tens of thousands of new virions in just hours, which then infect other healthy cells. Parts of the virus's RNA often encode proteins that remain in the host cell. At least three of them are identified. One stops the host cell from alerting the immune system that it is under threat. Another induces the host cell to release newly formed virions. Another aids the virus's resistance to the host cell's innate immunity. The virus hijacks the host cell's protein-making machinery with the help of enzymes called proteases of the virus. By successfully binding our drug with the main-protease of the SARS-CoV-2 we can slow down the replication process of the virus enabling the immune system to catch-up and defeat the virus.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

The generation of new molecules with could act as potential inhibitors of SARS-CoV-2 was done with the help of an RNN-LSTM model. RNNs have been successful in field of NLP, translation, composing music. Much of this success of RNNs has been due to the use of LSTM (long short-term memory) cells. RNNs based on LSTMs have been successfully used to predict protein function from sequence, aqueous solubility of molecules with drug like properties. Here we used LSTMs to capture the chemical syntax of original molecules to generate new molecules.

Later, a diverse set of original and generated molecules was used in binding with the main protease of SARS-CoV-2 and their effectiveness was measured using a metric called binding affinity. Binding affinity is the strength of the binding interaction between a single biomolecule (e.g. protein or DNA) to its ligand/binding partner (e.g. drug or inhibitor). PyRx was used to calculate the binding affinity of molecules by which they were later ranked to find which were the best among them.

II. RELATED WORK

The existing methods and systems help us in providing us with the basic knowledge of how we can implement our project. We learn from various elaborate explanations and intend to improve the existing methodologies and hence come up with our system. Following are the various insights gathered from different sources which have proved helpful in our research.

Before starting with our we had to know about the SARS-CoV-2 in detail. The first paper we referred to was about the origin, transmission, and characteristics of Coronaviruses by M. A Sheeren from State Key Laboratory of Virology, College of Life Sciences, Wuhan University and his team. It summarizes and comparatively analyzes the emergence and pathogenicity of COVID-19 infection and previous human coronaviruses severe acute respiratory syndrome coronavirus (SARS-CoV) and middle east respiratory syndrome coronavirus (MERS-CoV). It also discusses the approaches for developing effective vaccines and therapeutic combinations to cope with this viral outbreak. It is also important to know how currently drug research and development takes place[2].

We have referred to a book, “Introduction to Biological Molecule Drug Research and Development”. This book offers an overview of the science that underpins effective pharmaceutical research and development projects. The book first outlines fundamental ideas before comparing and contrasting approaches to biopharmaceuticals (proteins) and small molecule drugs, providing an explanation of the market and management challenges involved with these approaches. The second half of the book contains deliberately chosen real-life case studies that demonstrate how the hypothesis outlined in the first half of the book is ultimately put into effect. Herceptin/T-DM1,erythropoietin, anti - HIV protease inhibitor Darunavir, and other drugs have been studied. By this we came to realize how time consuming, cumbersome and expensive the current Drug Discovery process is[3].

Thirdly we discuss a paper by Bo Ram Beck and his team that talks about drug repurposing. In this study, a pre-trained deep learning-based drug-target interaction model called Molecule Transformer-Drug Target Interaction (MT-DTI) to identify commercially available drugs that could act on viral proteins of SARS-CoV-2. Sadly this study does not focus new drugs that could act as potential cure for COVID-19 that are not currently in existence but which could be synthesized later[4].

III. METHODOLOGY

This section provides a detailed explanation of the flow of the project involving creation of the molecule generator and the which molecules bind best with the main protease of SARS-CoV-2.

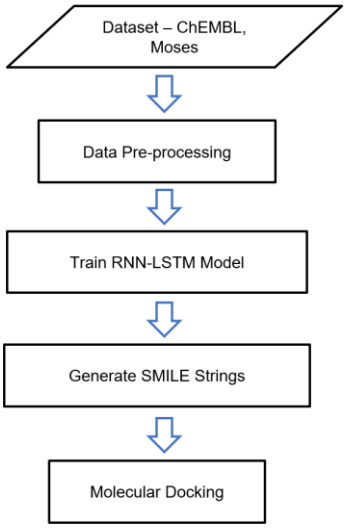

Fig. 1 Project Workflow

A. *Datasets*

In our project we worked with two datasets ChEMBL[5] and MOSES[6]. ChEMBL is a manually curated dataset containing biologically active compounds. It combines chemical, bioactivity, and genomic data to help in the processing of genomic data into successful new drugs. The MOSES dataset is based on the ZINC Clean Leads collection. It contains 4,591,276 molecules in total, filtered by molecular weight in the range from 250 to 350 Daltons, a number of rotatable bonds not greater than 7, and XlogP less than or equal to 3.5. Molecules containing charged atoms or atoms besides C, N, S, O, F, Cl, Br, H or cycles longer than 8 atoms have been removed. The molecules were filtered via medicinal chemistry filters (MCFs) and PAINS filters. filters.

B. *Data Pre-processing*

Both of these datasets contain molecules in the form of SMILES which stands for simplified molecular line entry system, it is a format for representing molecules using short ASCII strings. We combined 1936962 molecules from MOSES and 556134 molecules from ChemBL and removed the molecules which were common and were left with 2493096 molecules, then we sampled 250000 molecules.

After this we performed pre-processing, we ran a script to filter out salts, nucleic acids, long peptides. As we were looking to generate new molecules which are not too large and not too small, we only retained molecules whose length was between 34-128 characters. In the end we were left with nearly 180000 SMILES which would be used for training our LSTM model.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

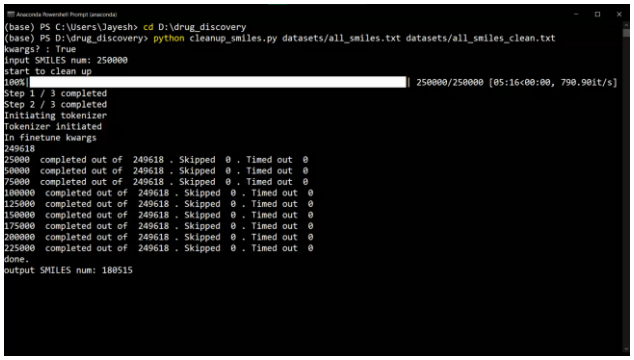

Fig. 2. Performing pre-processing

C. *Training our model to generate molecules*

Coming to our RNN-LSTM model. RNNs process a data sequence X = x1x2xn by taking each item xi in the sequence as input. The RNN processes the input through a series of gates to produce some hidden state hi and (optionally) an output vector y i. The hidden state hi is transferred from cell to cell and indicates which information the RNN has previously seen. Furthermore, recurrent connections enable RNNs to learn complex temporal relationships.

LSTMs have three gates input, output and forget. LSTMs are able to control which information passes to the next cell through hidden state hi. Important information can pass through successive cells unchanged. In this way LSTMs solve the vanishing gradient problem that RNNs face due to back propagation.

RNN models, which can produce a probability distribution across all potential tokens at each time step, may be used to construct sequences one token at a time. The RNN's goal is usually to predict the next token in a given input. It's worth noting that the input might be one or more tokens long; if there are m tokens in the input, the model will predict the (m+1)th token. Maximum likelihood estimation was used to train RNNs. The output vector yi is a probability distribution over the possible tokens, and the target vector yi is an array of one-hot encoded vectors, where each vector represents one token. Only one bit of a zero vector of the length of the number of tokens in the dataset is set (“hot”) in one-hot encoding. For each vector in the array, the model attempts to maximize the probability given to the right token.

Our LSTM model is made up of two layers, each with a 256-element hidden state vector that is regularized by dropout. A dense output layer and a neuron with a softmax activation function follow these two layers. The LSTM model receives a one-hot encoded sequence of the compound's SMILES string, with each string divided into tokens. A ‘G' token (for “go”) is added to the beginning of each SMILES string, and a ‘E' is appended to the end of the SMILES string. When padding was required, the token ‘A' was used.

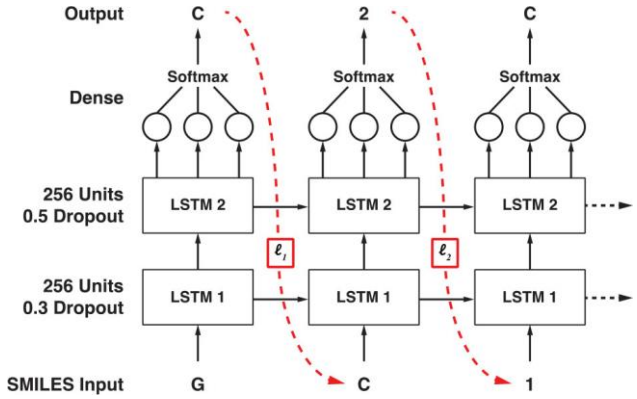

Fig. 3. LSTM Architecture

Each molecule was padded to the length n of the longest SMILES string for training purposes (padding signified by the token ‘A'). The input was the first n-1 characters, and the destination was the last n-1 characters. To begin Sampling, the sentinel token "G" was given. The last sampled character is used as the next character in the created sequence at each step of the sampling process. The sampling process continues until the token ‘E' is generated, which denotes the end of the sequence. Below are the equations for calculating the loss error L and the softmax function P(yi) with temperature factor T.

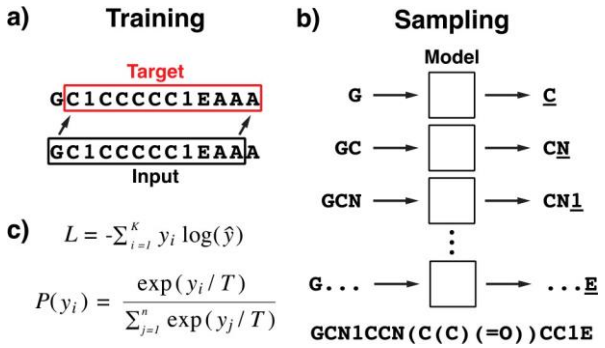

Fig. 4. A) Training procedure B) Sampling Procedure C) Equations for the loss error L

The model was used to generate 10000 molecules and their validity, uniqueness and originality was checked, those which failed to meet the criteria were removed. About 9700 molecules were generated which met the criteria. In order to compare the generated molecules to the original molecules used for RNN training, 24 common physiochemical features for the data were calculated. PCA was performed on these 24 generated features from the training molecules, and the first two principal components (PC1, PC2) were selected. The coordinates of the generated molecules were transformed accordingly. Figure 6 below shows that there is an overlap in the chemical subspace between these two sets of molecules indicating that they follow the same rules of chemistry as the original molecules.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

Fig. 5. Generating Molecules

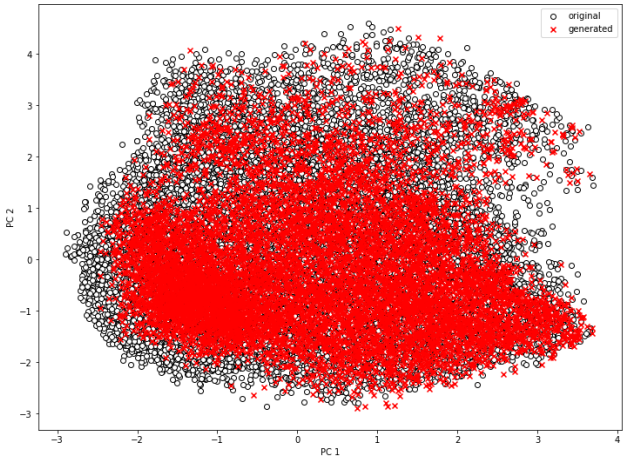

Fig. 6. Comparing generated and original molecules using PCA.

D. Molecular Docking

After generating molecules our next task was to find out molecules with high binding affinity to the viral protease. We could not bind all molecules as it would take very long time so we decided to take a diverse subset of molecules for binding, the similarity metric used for this purpose was Tanimoto similarity[7].

$$T(a,b) = \frac{N_c}{N_a + N_b - N_c}$$

We wrote a function that intakes a list of smiles, randomly shuffles them, then adds first thirty, then sets a max-similarity threshold between any new molecule and existing list and iteratively increases the threshold until X components are picked to ensure diversity. For better representation of each molecule in PyRx we created a four-letter code to order each molecule. We later converted all molecules to a sdf file which is understood by PyRx.

With the help of PyRx binding affinity of these compounds to the protease was calculated and results were tabulated and saved in a .csv file. This can be said to be our gen0 results. After this the top 35 ranked molecules were selected along with the molecules similar to them to generate new molecules. A diverse subset of training molecules and generated molecules was again selected and binding was performed. This iterative process was repeated 10 times in total to get the final results.

IV. RESULTS

In this section we have presented small molecules in SMILES format which have the highest binding scores with the main protease of SARS-CoV-2, also we have discussed what these results convey.

The following table lists the molecules with the highest binding affinities in descending order of their negativity. More negative means it binds more strongly.

TABLE I. Generated molecules with highest binding scores

Sr. No	SMILE	Binding Affinity (kcal/mol)
1	O=C(NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O)C1CC2CCC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.8
2	O=C(NC1CC2CCC(C1)N2CCc1ccccc1)C1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.7
3	O=C(NC1CC2CCC(C1)N2CCc1ccccn1)C1CC2CCC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.6
4	O=C(NC1CC2CCCC(C1)N2CCc1ccccn1)C1CC2CCC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.6
5	O=C(NC1CC2CCCC(C1)N2CCc1ccccc1)C1CC2CCC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.5
6	O=C(NC1CC2CCC(C1)N2CCc1ccccn1)C1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.5
7	O=C(NC1CC2CCCC(C1)N2CCc1ccccc1)C1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.4
8	O=C(NC1CC2CCCC(C1)N2CCc1ccccc1)C1COC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.3
9	O=C(NC1CC2CCC(C1)N2CCc1ccccc1)C1CC2CCC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.2
10	O=C(NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O)C1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.2
11	O=C(NC1CC2CCC(C1)N2CCc1ccccc1)C1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.2
12	O=C(NC1CC2CC(C1)N2CCc1ccccn1)C1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-17.1
13	NC(=O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)C1C2CC(NC(=O)C(Cc3ccccc3)C(=O)NC(Cc3ccccc3)C(=O)O)CC1C2	-17
14	O=C(O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)C1C2CC(NC(=O)C(Cc3ccccc3)C(=O)NC(Cc3ccccc3)C(=O)O)CC1C2	-17
15	CC(O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)C1C2CC(NC(=O)C(Cc3ccccc3)C(=O)NC(Cc3ccccc3)C(=O)O)CC1C2	-16.9
16	O=C(O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)C1C2CC(C(O)CC3CC4CCC(C3)N4CCc3ccccc3)CC1C2	-16.9
17	CC(C)(C)NC(=O)C1CC2CCC(CC1C(=O)NC(Cc1ccccc1)C(=O)O)C2CC(O)C(Cc1ccccc1)NC(=O)C1C2CC(NC(=O)C(Cc3ccccc3)C(=O)O)CC1C2	-16.7
18	O=C(NC(Cc1ccccc1)C(=O)O)C1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-16.5
19	O=C(NCC1(c2ccccn2)CC1)NC1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-16.3
20	O=C(NCC1(c2ccccn2)CC1)NC1CC2CCC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-16.3
21	O=C(NCC1(c2ccccc2)CC1)NC1CC2CCC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-16.3
22	O=C(NCC1(c2ccccc2)CC1)NC1CC2CC(C1)C2C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O	-16.2
23	COC(=O)N1CC2CCC(CC1C(=O)NC(Cc1ccccc1)C(=O)O)C2CC(O)C(Cc1ccccc1)NC(=O)C1C2CC(NC(=O)C(Cc3ccccc3)C(=O)O)CC1C2	-15.6

TABLE 2. Pre-existing molecules with highest binding scores

Sr. No	SMILES	Binding Affinity (kcal/mol)
1	CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(O)C(Cc1ccccc1)NC(=O)C(CC(=O)O)NC(=O)c1ccc2ccccc2n1	-10.9
2	CC(C)CN(CC(O)C(Cc1ccccc1)NC(=O)OC1COC2OCCC12)S(=O)(=O)c1ccc2nc(N)sc2c1	-10
3	CC(C)CN(CC(O)C(Cc1ccccc1)NC(=O)OC1COC2OCCC12)S(=O)(=O)c1ccccc1	-10
4	Cc1ccc(S(=O)(=O)N(CC(C)C)CC(O)C(Cc2ccccc2)NC(=O)OC2COC3OCCC23)cc1	-9.9
5	CC(C)CN(CC(O)C(Cc1ccccc1)NC(=O)OC1COC2OCCC12)S(=O)(=O)c1ccc(O)cc1	-9.9
6	CC(C)CN(CC(O)C(Cc1ccccc1)NC(=O)OC1COC2OCCC12)S(=O)(=O)c1ccccc1	-9.3
7	CC(C)CN(CC(O)C(Cc1ccccc1)NC(=O)OC1COC2OCCC12)S(=O)(=O)c1cccc(N)cc1	-9.2

These results were obtained using PyRx, which is a Virtual Screening software for Computational Drug Discovery that can be used to screen libraries of compounds against potential drug targets.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

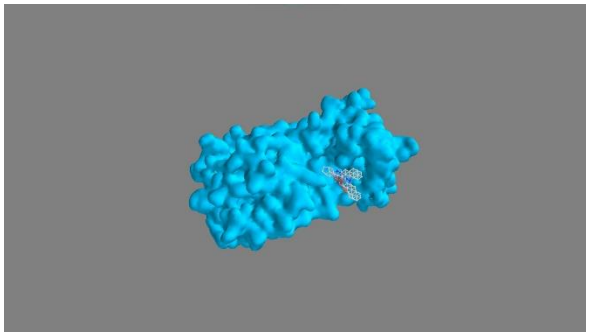

Fig 7. Binding our generated molecule with the main protease of SARS-CoV-2. This generated molecule received the highest binding score

Some of the molecules that were generated have significantly higher binding capacity than Remdesivir which has binding affinity of about -6.4 kcal/mol which is being given to COVID-19 patients. Also these molecules have better binding affinity than Pegylated Interferon alpha-2b which has binding affinity close to -5.7 kcal/mol. It was recently approved by Drug Controller General of India (DCGI) for emergency use to treat 'moderate COVID-19 cases' by Indian pharma company Zydus Cadela[8].

V. CONCLUSION

So with the help of our project we have realized how we can use deep learning techniques in the field of chemogenomics to identify candidate drugs which can act against SARS-CoV-2, slowing down its replication and giving our opportunity to catch up with the virus and finish it. We with the help of LSTMs we were able to generate unique and diverse molecules which followed the same rules of chemistry to a very high extent. In the end we believe these compounds that we found can developed by pharmaceutical companies and can be given to patients after proper clinical trials proving their effectiveness.

VI. FUTURE SCOPE

With the help of the approach that was used in our project in future we can similarly find anti-viral drugs for other viral infections. We can further in future use natural processing techniques to predict changes in viral RNA/DNA so that we are better prepared against the virus in the future.

REFERENCES

[1] COVID Live Update. (n.d.). Worldometer. Retrieved June 25, 2021, from https://www.worldometers.info/coronavirus/

[2] Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. J Adv Res. 2020 Mar 16;24:91-98. doi: 10.1016/j.jare.2020.03.005. PMID: 32257431; PMCID: PMC7113610.

[3] Ganellin, R. C., Jefferis, R., & Roberts, S. M. (2013). *Introduction to Biological and Small Molecule Drug Research and Development: Theory and Case Studies* (1st ed.). Elsevier.

[4] Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput Struct Biotechnol J. 2020 Mar 30;18:784-790. doi: 10.1016/j.csbj.2020.03.025. PMID: 32280433; PMCID: PMC7118541.

[5] Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M., Mosquera, J., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C., Segura-Cabrera, A., . . . Leach, A. (2018). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, *47*(D1), D930–D940. https://doi.org/10.1093/nar/gky1075

[6] Sterling, T., & Irwin, J. J. (2015). ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, *55*(11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559

[7] Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. *J Cheminform* **7,** 20 (2015). https://doi.org/10.1186/s13321-015-0069-3

[8] *Zydus Cadila receives Emergency Use Approval for Virafin*. (2021, April 23). Business Standard. https://www.business-standard.com/article/news-cm/zydus-cadila-receives-emergency-use-approval-for-virafin-121042300715_1.html