World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644

DESIGNING POTENTIAL DRUGS THAT CAN TARGET SARS-COV-2'S MAIN PROTEASE: A PROACTIVE DEEP TRANSFER LEARNING APPROACH USING LSTM ARCHITECTURE

1. Omar Dasser, PhD student *
Research Laboratory Mathematics, Computer Science and Engineering Sciences, Hassan I University- Settat 26002 – Morocco.
2. Moad Tahri, PhD student
Laboratory of informatics research and innovation, Hassan II University, Casablanca, Morocco.
3. Louay kila, MD
M.B.B.S, ECFMG Certified, Al Ain Hospital- Health Authority Abu Dhabi.
4. Abderrahim Sekkaki, PhD
Laboratory of informatics research and innovation, Hassan II University, Casablanca, Morocco.
Correspondence
Omar Dasser PhD student,
Research Laboratory Mathematics, Computer Science and Engineering Sciences, Hassan I University- Settat 26002 – Morocco.
Email: Dasseromar@gmail.com
Present address
Research Laboratory Mathematics, Computer Science and Engineering Sciences, Hassan I University- Settat 26002 – Morocco.
Funding information

Article history:
Received: May 1st 2022
Accepted: June 1st 2022
Published: July 6th 2022

Abstract:

On December 2019, the world entered a state of alarm and dismay with the outbreak of a severe acute respiratory syndrome coronavirus 2 (SARS-CoV 2) from Hubei-China and has infected as of the 1st October 2021 more than 233,770,079 people worldwide. This caused up to 4,782,608 deaths, and the World Health Organization (WHO) declared on January 2020 a global health emergency due to the rate at how much the infection is spreading and the mortality rate that approaches 4.5 percent [1]. It is considered to be extremely costly to bring a new drug to the market in terms of time and financial investment, which is, respectively, on average, around ten years and 1 billion dollars. Drug discovery alone can take up to 3 years which is a time we cannot accept in the context of a global pandemic.

Keywords:

INTRODUCTION

On December 2019, the world entered a state of alarm and dismay with the outbreak of a severe acute respiratory syndrome coronavirus 2 (SARS-CoV 2) from Hubei-China and has infected as of the 1st October 2021 more than 233,770,079 people worldwide. This caused up to 4,782,608 deaths, and the World Health Organization (WHO) declared on January 2020 a global health emergency due to the rate at how much the infection is spreading and the mortality rate that approaches 4.5 percent [1]. It is considered to be extremely costly to bring a new drug to the market in terms of time and financial investment, which is, respectively, on average, around ten years and 1 billion dollars. Drug discovery alone can take up to 3 years which is a time we cannot accept in the context of a global pandemic. Artificial intelligence methodologies proved to be very resourceful for solving many tasks, especially when it comes to computer vision, natural language processing, and solving core problems in biology, such as a gigantic leap in the prediction of the 3-D shapes of protein structures based on its amino-acids sequences [2] [3] and also using GAN architectures to search for new molecules [4]. Our main goal is to harvest the power of these methodologies in order to generate new molecules that can potentially treat the disease and thus contributing in reducing the time for the drug discovery process. The genetic code is oftentimes called the genetic blueprint as it contains all instructions that a cell would need to survive, proliferate, and perform its role in the organism. These instructions are found in the form of DNA; for them to become realized, they pass through two steps which are transcription and translation. In the flow of information, the first step is to transcribe the double-

World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644

stranded DNA (dsDNA) template to yield a single-stranded RNA (ssRNA) molecule, called Messenger RNA (mRNA). This mRNA will then carry the transcribed instruction from within the Nucleus into the Cytosol, where it will be Translated into Protein Product.

Transcription Process: The Enzyme RNA Polymerase-II (RNA pol-II) is required for transcription to occur, as it binds to the template DNA strand and catalyzes the formation of a complementary mRNA. In Eukaryotic Cells, there are three main different types of RNA Polymerase that exist. RNA pol I transcribe the genes that encode Ribosomal RNAs (rRNAs). RNA pol II transcribes mRNA, which will be translated, yielding protein products. RNA pol III transcribes the genes for Transfer RNAs which are essential in the translation process.

Translation Process: As discussed above, the product of Transcription is the production of a single-stranded mRNA copy of the gene, which next must be translated into a protein molecule. Translation is the process which by the genetic code is translated into a sequence of Amino Acids, which consequently form proteins. [5]

METHOD

The .pdb (protein data bank) file, retrieved from the RCSB website[6], offers a digital representation of the protease that can be uploaded to a docking simulation tool representing the structure of the main protease (Mpr o) was loaded as a macromolecule to PyRx, with which we will simulate the docking of the newly generated molecule. Each simulation produces a metric called binding affinity score, also called the binding energy. Some already known drugs passing clinical trials, such as Hydroxychloroquine, gave a score of -5.3 Kcal/mol.

Pre-processing A first step into preprocessing the SMILES data was to create a tokenizing function that would convert a SMILES string into one hot-encoded vector from a set of all possible tokens that will be fed to the neural network. And another function that would decode the one hot-encoded vector to its corresponding SMILES for further processing. Most of the utility functions were implemented using the RDKit library on Python [7], and their main purpose was to:

- Converting SMILES to mols (and eventually determining whether the SMILES are valid or not).
- Converting mols to SMILES (mainly to ensure that we'd use only canonical SMILES).
- Calculating molecular weight.
- Calculate the number of atoms, number of Spiros, number of Chiral centers, number of bridgeheads number of macro- cycles in order to deduce the synthetic feasibility score.
- Writing Results to a chemical table file (.sdf) that'll be passed to PyRx.

Our goal is to generate SMILES that fits our needs; this can be done through different techniques. RNN with LSTMs has shown great success for text generation. Even though LSTM was first invented in 1997, training LSTMs with MLE still outperforms recent methods in text generation like Scheduling Sampling (SS), and it is also as good as some recent and complex architectures such as SeqGan [10]. LSTMs and its variants are known to alleviate the vanishing and exploding gradient problems due to a memory cell they contain[cite]. In the context of SMILES generation, these

models typically fail due to the errors that accumulates with each recursion [11] and can eventually lead to poor quality of the generated sequences. This phenomenon is known as the bias exposure problem [28][29]. To solve this issue, we will train our model following the maximum likelihood estimation. Doing so, our model opts to choose the token with the highest probability. However, in the sampling phase, we update our model using the temperature-decoding method, which shrinks or enlarges probabilities to ensure more flexibility in the search area of the best token and to produce distinct and diverse generations while sampling our SMILES.

We generate with this model a batch of molecules that will be filtered according to the swiss cheese principle (Fig .1); We remove duplicate molecules (SMILES sequence that can be generated twice or might be the same after the canonical form conversion). We also remove non-valid and erroneous molecules (molecules that cannot exist and don't obey to laws of physics). After that, we eliminate molecules that have a great molecular weight (MW > 850 Da) and molecules that are hard to synthesize (Synthetic accessibility score >3.5). We pass the final results into the PyRx tool and retrieve the top 100 molecules by binding affinity score, which will be used to finetune the model; the binding simulation is done using the following vina search space parameters:

- x: 51.3737 Å
- y: 66.9738 Å
- z: 59.6069 Å

And center values of:

- x : -25.9865 Å
- y : 12.5886 Å
- z : 59.1565 Å

We repeated the above tasks until we got our final results.

World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644

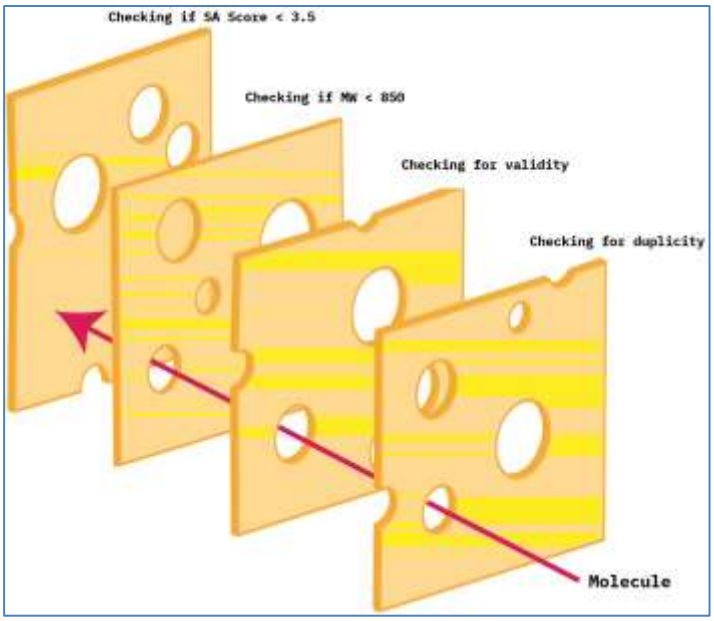


Figure 1- Swiss cheese principle for filtering and selecting molecules

The first step in our architecture is to train a generative model with the SMILES data representation of some existing pharmaceutical compounds, which serves as our base model. In order to generate a sequence, the model will be alimented in a first step with the BoS token (Beginning of Sequence) and will then produce a probability distribution over all the set of possible tokens at each time until the model predicts the EoS (End of Sequence). In order to alleviate the problem of bias exposure, we train our model through maximum likelihood estimation (eq.).

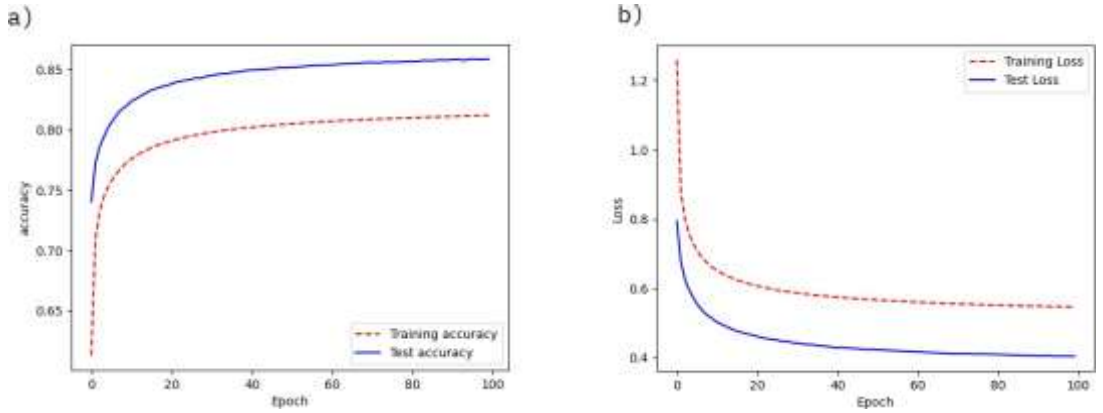
$$M\ LE =. P\theta\ (xt\ |X1{:}t\ -1) \qquad (25)$$

The loss function is calculated as the categorical cross-entropy between the actual value of the next token and the predicted one and then is averaged through all the predictions (eq. 26) [8] [9].

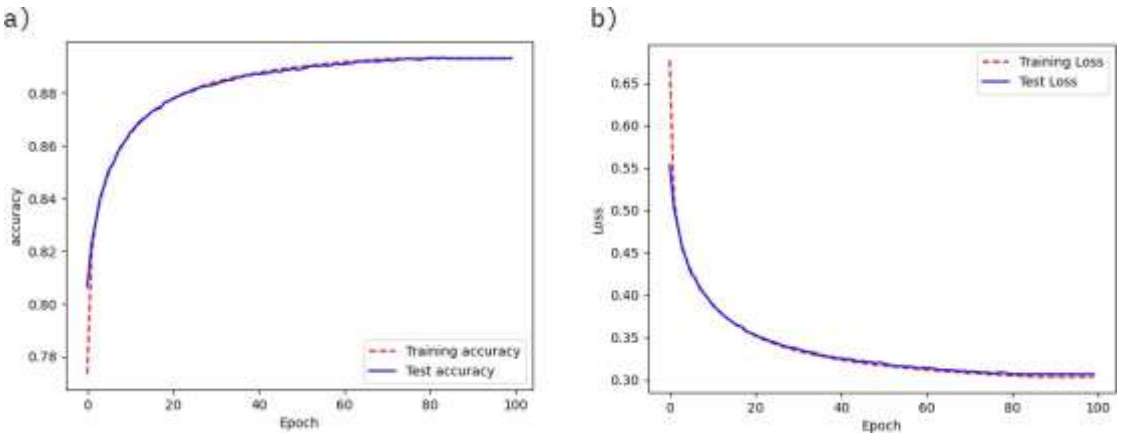
RESULTS

We evaluated both of our models (vanilla-LSTM and BN-LSTM) in order to choose the best model for our SMILES generations. Each model was constituted of 2 LSTM cells and one fully connected layer; we proceeded to remove the dropout from the BN-LSTM model; we conclude with the following results shown in figures

World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644



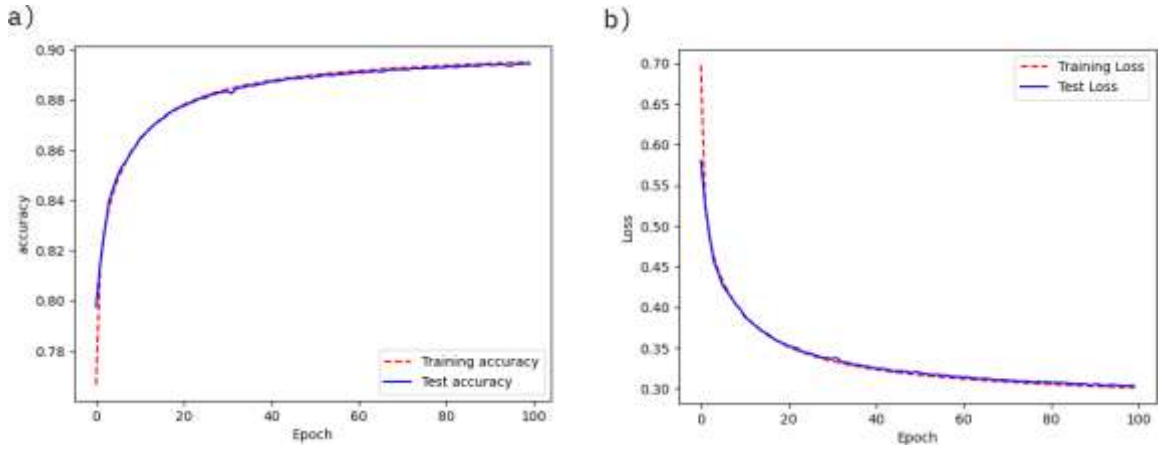
F I G U R E 2 Vanilla model accuracy (a) and training loss(b) evolution per epoch. The produced model resulted in a validity value of 43.10%, a uniqueness value of 99.88%, and an originality value of 99.42% within the first generated set. We observe that the validation loss is lower than the training loss. This behavior of the model is due to the fact that the dropout regularization is applied during training but not during testing. This implies that our model is underfitting and is not able to perform well on the training set. Therefore, such behavior explains why the model couldn't produce a higher validity value.



F I G U R E 3 Batch Normalization model accuracy (a) and training loss(b) evolution per epoch. The produced model resulted in a validity value of 90.98%, a uniqueness value of 98.36%, and an originality value of 90.37% within the first generated set.

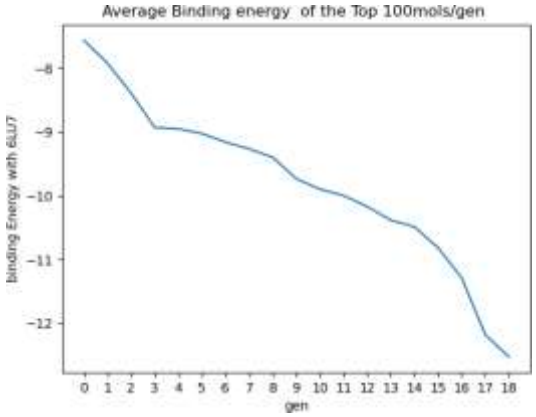
World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644

We retrain our model, using this time an orthogonal initialization for all the weights in our model instead of the normal weight initialization. We have noticed a slight improvement on both the loss and accuracy of the model, as well as an increase on the validity of the generated set. The figure below(fig.9) depicts the result acquired.



F I G U R E 4 Batch Normalization model with orthogonal weight initialization accuracy (a) and training loss(b) evolution per epoch. The produced model resulted in a validity value of 92.76%, a uniqueness value of 98.16%, and an originality value of 90.63% within the first generated set.

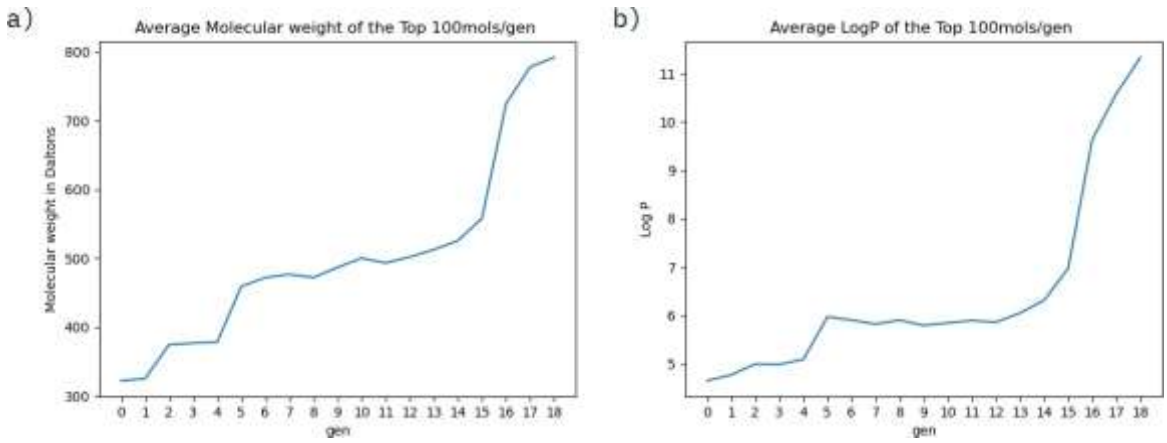
When calculating the average binding energy of the top 100 candidates of each generation, we can see that we're getting better results in terms of binding affinity score with M_{pro} (6LU7) within each generation (Fig.10)



F I G U R E 5 Binding energy evolution of the top 100 /gen. The figure shows that with each generation, we get new generated molecules that achieves a better(smaller) binding affinity score with 6LU7.

Even though, as shown in the figure above, we got lower binding affinity scores within each generation, we stopped all the iterations in generation 18 as all the newly generated molecules has undesirable pharmacokinetic properties such as high molecular weight and high lipophilicity, which can lead in general to a lower solubility, high turnover, low absorption and can also lead in some cases to toxicity and metabolic clearance [35]. Below we describe some of the generated molecules that had interesting assets; Binding energy with 6LU7, Synthetic Accessibility score, and ADME Properties (Molecular Weight, LogP, H-Bond donor, H-bond acceptor). Many ADME Properties were calculated using the SwissADME web tool [36]

World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644

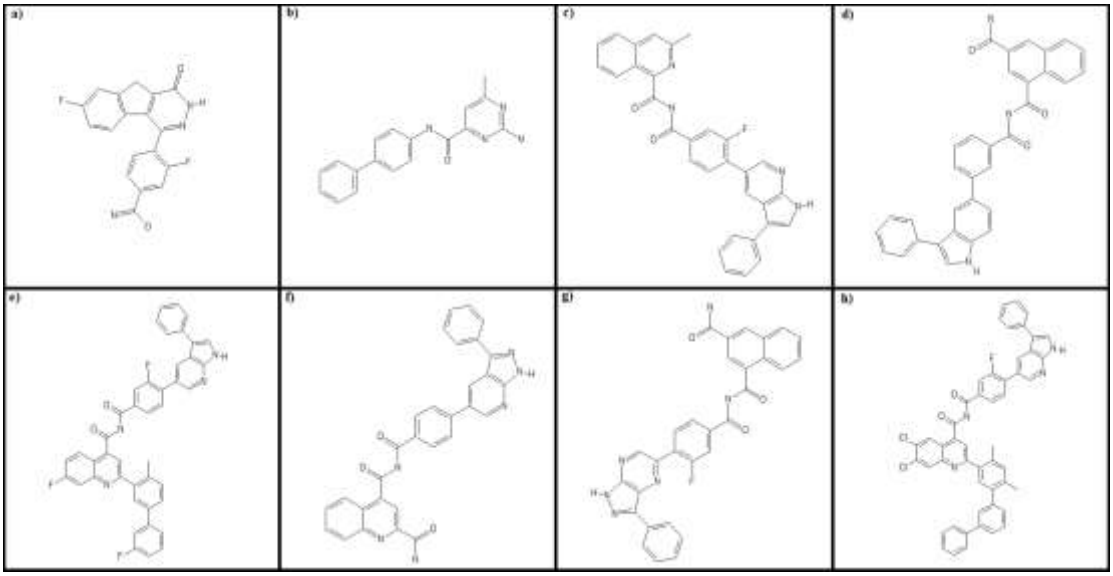


F I G U R E 6 Average Molecular Weight and logP of the top 100mols/gen. The figure shows that the top 100 of the generated molecules becomes more and more heavier in terms of Molecular weight(a) and increases also in the value of the calculated logp (b) within each generation.

Table 1: ADME properties & metrics of the generated molecules

Molecule	Mol Weight	Log P	H-Bond Donor	H-Bond Acceptor	Binding energy (Kcal/mol)	Synthetic Accessibility Score	QED
Mol 1	339.082	3.17	3	3	-9.6	2.828	0.858
Mol 2	304.132	3.287	2	4	-9.3	1.831	0.959
Mol 3	500.165	6.463	2	4	-10.5	2.613	0.214
Mol 4	509.174	6.324	3	3	-11.0	2.402	0.195
Mol 5	512.16	4.509	3	6	-10.4	2.581	0.173
Mol 6	530.15	4.648	3	6	-10.5	2.758	0.147
Mol 7	688.209	10.075	2	4	-12.2	2.962	0.051
Mol 8	810.196	13.079	2	4	-13.2	3.213	0.04
Azithromycin	748.509	1.901	5	14	-7.6	nan	0.039
Remdisivir	602.225	2.312	4	13	-5.1	nan	0.059
Ritonavir	720.313	5.905	4	9	-5.1	nan	0.046
Hidroxy-Chloroquine	335.176	3.783	2	4	-6.2	nan	0.918
Chloroquine	319.182	4.811	1	3	-6.7	nan	0.942
Nitazoxanide	307.026	2.229	1	7	-7.9	nan	0.83

World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644



F I G U R E 7 structure of the generated molecules. a) depicts the molecules described in the table. Two as mol1, b) depicts the molecules described in the table. Two as mol2, c) depicts the molecules described in the table. Two as mol3 d) depicts the molecules described in the table. Two as mol4 e) depicts the molecules described in the table. Two as mol5, f) depicts the molecules described in the table. Two as mol6, g) depicts the molecules described in the table. 2 as mol7, h) depicts the molecules described in the table. Two as mol8. All the 2-D structure rendering was carried out using the PubChem sketcher web tool [12].

The table below shows the SMILES representations of the generated molecules described in Table.2

Molecule	SMILES
Mol 1	N=C(O)c1ccc(-c2n[nH] c(=O) c3c2-c2ccc(F)cc2C3) c(F)c1
Mol 2	Cc1cc (C (=O) Nc2ccc (-c3ccccc3) cc2) nc (N) n1
Mol 3	Cc1cc2ccccc2c(C(=O)NC(=O)c2ccc(-c3cnc4[nH]cc(-c5ccccc5)c4c3)c(F)c2)n1
Mol 4	NC(=O)c1cc(C(=O)NC(=O)c2cccc(-c3ccc4[nH]cc(-c5ccccc5)c4c3)c2)c2ccccc2c1
Mol 5	NC(=O)c1cc(C(=O)NC(=O)c2ccc(-c3cnc4[nH]nc(-c5ccccc5)c4c3)cc2)c2ccccc2n1
Mol 6	NC(=O)c1cc(C(=O)NC(=O)c2ccc(-c3cnc4[nH]nc(-c5ccccc5)c4n3)c(F)c2)c2ccccc2c1
Mol 7	Cc1ccc(-c2cccc(F)c2)cc1-c1cc(C(=O)NC(=O)c2ccc(-c3cnc4[nH]cc(-c5ccccc5)c4c3)c(F)c2)c2ccc(F)cc2n1
Mol 8	Cc1cc(C)c(-c2cc(C(=O)NC(=O)c3ccc(-c4cnc5[nH]cc(-c6ccccc6)c5c4)c(F)c3)c3cc(Cl)c(Cl)cc3n2)cc1-c1cccc(-c2ccccc2)c1

Table 2: SMILES representation of the generated molecules

The figures below depict the 2-D structures of the protein-ligand interaction between the generated ligands and the M_{pro}. These figures were generated using PyMOL [13] and LigPlot+ [14].

World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644

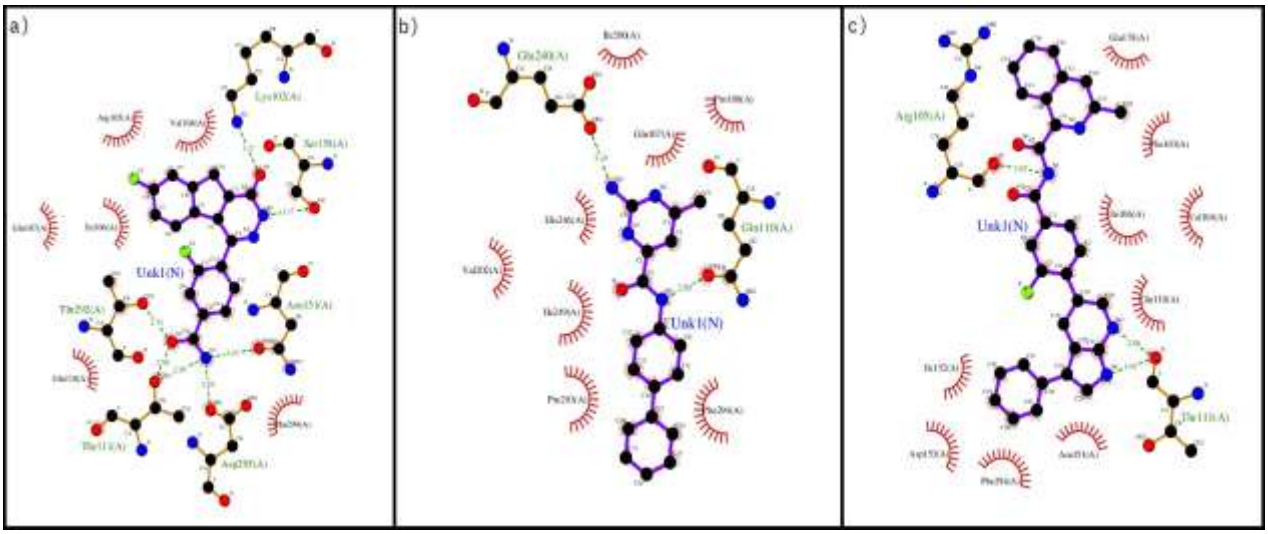


FIGURE 8 The figure shows the 2-D structure of the protein-ligand interaction between the mol1 (a), mol2 (b), and mol3 (c) described in Table 2 and M_{pro} (6LU7).

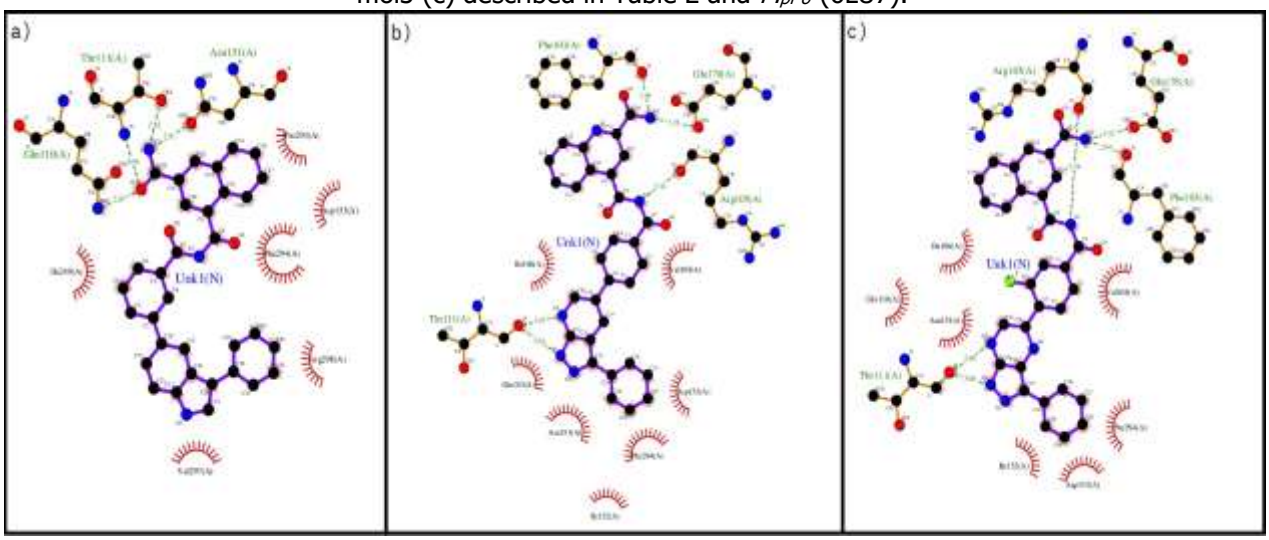


FIGURE 9 The figure shows the 2-D structure of the protein-ligand interaction between the mol4(a) and mol5(b), mol6(c) described in Table 2 and M_{pro} (6LU7).

World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644

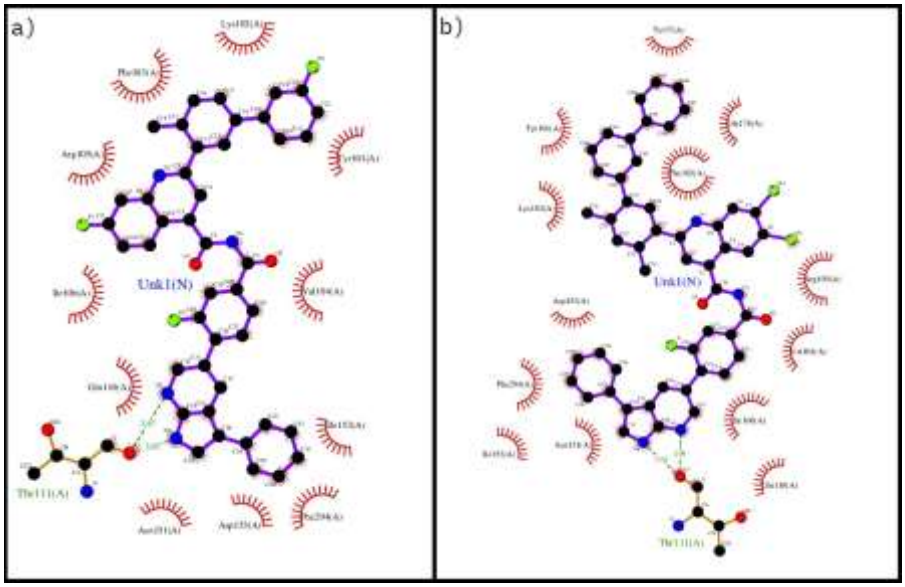


FIGURE 10 The figure shows the 2-D structure of the protein-ligand interaction between the mol7(a) and mol8(b) described in Table 2 and M_{pro} (6LU7).

CONCLUSION

In this work, we successfully produced a model capable of generating molecules that can inhibit SARS-CoV-2 main protease, as shown in our simulation-based on deep, proactive transfer learning. We trained an LSTM architecture with SMILES representation of existing pharmaceutical compounds to produce our base model, which has a goal only to generate valid molecules. We proceeded afterward to fine-tune the model with the SMILES representation of the best molecules that met filtering criteria such as molecular weight, feasibility to synthesize, and most importantly, the binding affinity score with the main protease. Further tests, such as in-vitro and in-vivo tests, should be made to retrieve more insights and findings of the above molecular results. Within the results, we found a common and shared fragment in many molecules; although with our approach, we can't grow molecules in more than one direction, this seems as a viable track to cover afterward.

REFERENCES

1. Pooja Singh, Sharma, A., Nandi, S.P.: Identification of potent inhibitors of COVID-19 main protease enzyme by molecular docking study (Apr 2020)
2. Hutson, M.: AI protein-folding algorithms solve structures faster than ever. Nature (Jul 2019)
3. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D.: Improved protein structure prediction using potentials from deep learning. Nature 577(7792), 706–710 (Jan 2020)
4. Blanchard, A.E., Stanley, C., Bhowmik, D.: Using GANs with adaptive training data to search for new molecules. Journal of Cheminformatics 13(1) (Feb 2021)
5. Ilona, M., Iorrie, L.: A brief history of genetics: Defining experiments in genetics. unit 5.6 (2010)
6. Berman, H.M.: The protein data bank. Nucleic Acids Research 28(1), 235–242 (Jan 2000)
7. Landrum, G.: Rdkit: Open-source cheminformatics, http://www.rdkit.org
8. Gupta, A., Müller, A.T., Huisman, B. J. H., Fuchs, J.A., Schneider, P., Schneider, G.: Generative recurrent networks for de novo drug design. Molecular Informatics 37(1-2), 1700111 (Nov 2017)
9. Kawthekar, P., Rewari, R., Bhooshan, S.: Evaluating generative models for text generation (2017)
10. Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: Sequence generative adversarial nets with policy gradient. CoRR abs/1609.05473 (2016)
11. Dallakyan, S., Olson, A.J. Small-molecule library screening by docking with PyRx. In:

World Bulletin of Public Health (WBPH)
Available Online at: https://www.scholarexpress.net
Volume-12, July 2022
ISSN: 2749-3644

Methods in Molecular Biology, pp. 243–250. Springer New York (Dec 2014)
12. Liu, J., Cao, R., Xu, M., Wang, X., Zhang, H., Hu, H., Li, Y., Hu, Z., Zhong, W., Wang, M.: Hydroxychloroquine, a less toxic derivative of chloroquine, is effective in inhibiting SARS-CoV-2 infection in vitro. Cell Discovery 6 (1) (Mar 2020)
13. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Research 49 (D1), D1388–D1395 (Nov 2020)
14. Schrödinger, LLC: The PyMOL molecular graphics system, version 1.8 (November 2015)