ORIGINAL ARTICLE

Journal Section

Designing potential drugs that can Target SARS-COV-2's main Protease : A Proactive Deep transfer Learning Approach using LSTM Architecture

Omar Dasser, PhD student[1*†] | Moad Tahri, PhD student[2*] | Louay Kila, MD[3*] | Abderrahim Sekkaki, PhD[2*]

[1]Research Laboratory Mathematics, Computer Science and Engineering Sciences, Hassan I University- Settat 26002 - Morocco

[2]Laboratory of informatics research and innovation, Hassan II University, Casablanca, Morocco

[3]Faculty of Medicine, The National Ribat University, Khartoum, Sudan

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

Drug discovery is a crucial step in the process of delivering a new drug to the market that can take up to 2-3 years which can be more penalizing given the current global pandemic caused by the outbreak of the novel coronavirus SARS-CoV 2. Artificial Intelligence methodologies have shown great potential in resolving tasks in various domains such as image classification, sound recognition, also in the range of the previous years, Artificial Intelligence proved to be the go-to for generative tasks for use cases such as music sequences, text generation and solving also problems in biology. The goal of this work is to harvest the power of these architectures using generative recurrent neural network with long short-term memory (LSTM) gating techniques in order to generate new and non-existing molecules that can bind to the main COVID-19 protease, which is a key agent in the transcription and replication of the virus, and thus can act as a potential drug that can neutralize the virus inside of an infected host. As of today, there are no specific targeted therapeutic agents to treat the disease and all existing treatments are all very limited. Known drugs

[*]Equally contributing authors.

that are passing clinical trials such as Hydroxychloroquine and Remdesivir showed respectively a binding energy with SARS-CoV-2's main protease of -5.3 and -6.5, the results of the newly generated molecules exhibited scores ranging till -13.2.

KEYWORDS
Covid-19, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), SMILES, Generative recurrent neural networks, Protein-ligand docking, molecule generation, In-silico drug discovery

1 | 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 550,348,645 people worldwide. This caused up to 6,352,882 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 10 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 specially when it comes to computer vision, natural language processing, 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-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 transcribes 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]

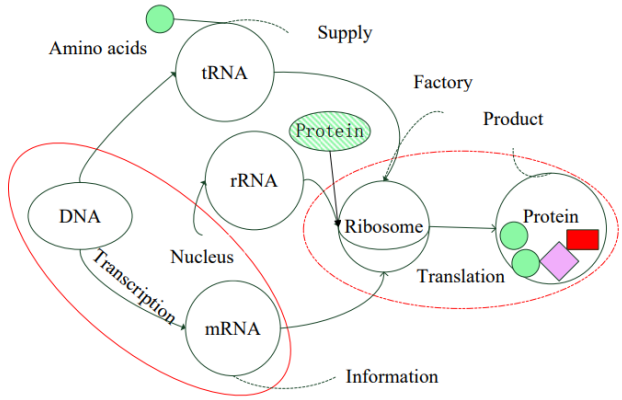

FIGURE 1 Gene transcription and translation.[6]

The lifecycle of the virus inside of an infected host is no different than the processes described above and thus one could stop the activity of the virus by stopping the main processes for its reproduction which are translation and replication. The novel SARS-CoV-2 possesses an enzyme called Main Protease (Mpro) which cuts the Polyproteins translated from viral RNA to yield function viral proteins necessary for viral replication. Currently, there's increasing evidence that some of the observed mutations may be capable of changing the Antigenic Phenotype of SARS-CoV-2, this consequently would affect Immune Recognition.[7] This implies that Vaccination alone is not a valid solution. It has been proposed in a previous study that the substrate-recognition pocket of Mpro is highly conserved among all Coronarviruses.[8] Making M_{pro} one of the most suitable options for pharmacological intervention. It is, therefore, of the utmost importance to specifically target and inhibit this enzyme, which in turn, will block viral replication. Furthermore to justify our position, currently, there are no known human proteases with similar cleavage specificity as that of M_{pro}, making toxicity due to M_{pro} Inhibitors to be unlikely.[9] To this end, we decided to study the Ligand Docking on the Main Protease of the virus as it is a sensible approach to halt the activities of the virus[10]. We decided to approach this study via an In-silico screening technique, using the PyRx tool[11] that uses Autodock vina as a screening method[12]. In our work, we opted for using LSTM architectures to create our generative model, in order to produce those ligands. The first step is to create a base model trained on the SMILES of existing pharmaceutical compounds and can generate valid and new molecules. The model will be then fine-tuned after each generation with molecules based on their binding affinity score with the main protease of the covid-19 (6LU7) and based on their synthetic accessibility and molecular weight. We choose the SMILES (Simplified Molecular Input Line Entry System) digital encoding as the main type of data that would represent molecules since that they represent and describe atoms and their bonds quite well and in a text line format that can be easily loaded to a generative recurrent network which has proved to be adequate and efficient in those kinds of tasks. The present document is divided into three main parts. The first section describes the problem including all the elements that we used, it also states the choice of the data type, data sources alongside the data modeling and generation. The second part focuses on the related works done within the same context in the field of the in-silico Protein-Ligand docking study, and also presents the data that we used within our work for the ligands and the macromolecules and also presents a dive-in into different methodologies followed, while in the third part, we present our results regarding the docking of the generated molecules with the Mpro (6LU7), their synthetic feasibility score, and their ADME properties. We conclude this work with a final part

discussing the work as a whole and some eventual future works to take into consideration

2 | METHODS

2.1 | Related work

A review of the state of the art on molecule generation using artificial intelligence has been conducted by Daniel C.Elton in 2019 [13] presents several approaches to generate molecules using RNNs (Recurrent Neural Networks), AutoEncoders and GANs (Generative Adversarial Networks). It also presents the different metrics (Diversity, Novelty, Stability, Synthesizability, Non-triviality, Good properties) to take into consideration in order to get a broader sense on the produced molecules. Moreover, these metrics can give us an overview on the quality of the generated molecules. They can also work as reward functions. Another study conducted by Esben Jannik Bjerrum in 2017[14] (Bjerrum 2017) focused on molecule generation for data augmentation using Variational AutoEncoders. Even though the studies covered many interesting parts, the scope that the study covered differs from our scope, as our goal doesn't only stop at generating valid molecules, but also to generate specific molecules that can inhibit a protein. A study was done by Anvita Gupta, Alex T. Müller et al. in 2017[15] in which they used Recurrent Neural Networks (RNNs) to generate molecules and fine-tuning in order to generate specific target molecules that would achieve a similarity score close to some known drugs. Many aspects of this study were beneficial to the success of ours nevertheless some aspects differed from our goal and the approach we used. We used the same techniques for the protein-ligand docking study and introduced other generation metrics such as synthetic feasibility that can be very beneficial since the current use case is to produce new molecules. Another interesting study was led by Bowen Tang et al. (AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2) [16] in which they leveraged the properties of advanced Deep Q learning to generate potential lead compounds that can target the SARS-COV-2 protease (mainly 3CLpro and M_{pro}) using the fragment drug-design (ADQN-FBDD).

2.2 | Ligands

The SMILES digital encoding offers a good representation of molecules since it can accurately describe the atoms and their bonds with a line text string which can serve afterward as input for Recurrent Neural Network. When choosing a molecule digital encoding three important properties should be taken into consideration[13] :

- **Uniqueness**: each Molecule structure has a unique represen-tation
- **Invertibility**: each representation is associated with only one Molecule
- **Representation type**: the representation can be either a sequence or a tensor

Representation technique	Is unique	Is invertible	type
SMILES	No	Yes	Sequence
Canonical SMILES	Yes	Yes	Sequence
InChI Keys	Yes	Yes	Sequence
Tensor Field Network	No	No	Tensor
MACC Keys	No	Yes	Sequence
Chemception images	Yes	Yes	Tensor
Tensors	No	Yes	Tensor

Table 1 : depicting molecules and some of their digital representation.

The choice of the canonical SMILES form over the regular SMILES representation has been done to ensure the bijectivity between the two sets of molecules and their sequence representation by ensuring uniqueness, each molecule is associated with one representation and invertibility, each representation is associated with a single molecule.

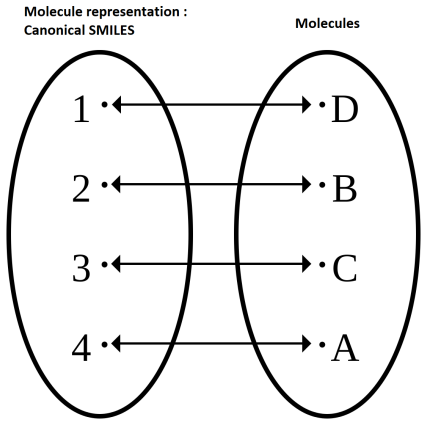

FIGURE 2 Relation between molecules and their SMILES representation.

Even though the InChI technique has proved to be very efficient with database indexing, Canonical SMILES offers a better representation when it comes to the different bond types and can be better suited for the model. To this end, we used SMILES data retrieved from the ChemBL databases[17](Mendez et al.,2019). The aggregation of these SMILES data sources gave us more than 600,000 SMILES. We also removed all the duplicates to avoid generating more redundant SMILES by our model, resulting in a lower uniqueness value. Finally, we removed all the long SMILES which would normally represent long molecules. We ended up with a training dataset that contains around 400,000 entry.

2.3 | Macromolecule

The Viral Protease is a common target for the development of Antiviral Drugs, as indicated by several studies on HIV, Hepatitis C Virus, and Ebola Virus.[18][19][20](Lv et al.,2015; De leuw et al.,2017 ; Nishimura et al.,2015). Regarding the Covid-19, M_{pro} is the Main Viral Protease, and it has been found to be conserved among the Coronaviruses (Coronaviridæ). Vaccination alone is not a valid solution since the viruses tends to reemerge and spike mutations do occur[21](Li et al.,2020). In the light of recent event, we can witness this happening in the United Kingdom, where a new variant of the SARS-CoV-2 has emerged, the variant is referred to as SARS-CoV-2 VUI 202012/01 (Variant Under Investigation, year 2020, month 12, variant 01) and is defined by multiple spike proteins mutations (deletion 69-70, deletion 144, N501Y, A570D, D614G, P681H,T716I, S982A, D1118H)[22](European Centre for Disease Prevention and Control, 2020) which can make a lot of vaccines passing clinical trials obsolete, this begs for the development of Antiviral medications that can inhibit a conserved target, we argue that Mpro is a suitable candidate and is a good choice to choose as our Macromolecule with which we will be performing our ligands docking simulation, since that it's considered to be an attractive drug target as it plays a key role for the replication of the Coronaviridæ[23][24], following the same rationale, Mpro would be a suitable target for inhibition.

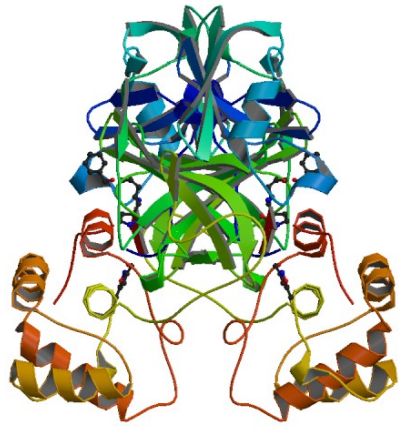

FIGURE 3 The crystal structure of COVID-19 main protease in complex with an inhibitor N3.

The .pdb (protein data bank) file, retrieved from the RCSB website[25], offers a digital representation of the protease that can be uploaded to a docking simulation tool representing the structure of the main protease (M_{pro}) 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.

2.4 | Pre-processing

A first step into preprocessing the SMILES data was to create a tokenizing function that would convert a SMILES string into a one hot-encoded vector, from a set of all possible tokens, that will be fed to 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[26] and their main purpose was to :

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

2.5 | Prerequisites

2.5.1 | Batch Normalisation

Let's consider the input of a given layer in a deep neural network with dimention n :

$$u = (u^{(1)}, ..., u^{(i)}, ..., u^{(n)}) \tag{1}$$

Let's denote $\mathscr{B}$ a mini-batch of size m.

$$\mathscr{B} = \{u_1, ..., u_m\} \tag{2}$$

The empirical values for the mean and variance of $\mathscr{B}$ are calculated as follows :

$$\mu_{\mathscr{B}} = \frac{1}{m}\Sigma_{i-1}^{m} x_i \tag{3}$$

$$\sigma_{\mathscr{B}} = \sqrt{\frac{1}{m}\Sigma_{i-1}^{m}(x_i - \mu_{\mathscr{B}})^2} \tag{4}$$

We can normalize each dimension $u^{(k)}$ as follows :

$$\overline{u}_i^{(k)} = \frac{u_i^{(k)} - \mu_{\mathscr{B}}}{\sqrt{\sigma_{\mathscr{B}}^{(k)^2} + \epsilon}} \tag{5}$$

Where $k \in \{1, n\}$ and $i \in \{1, m\}$ and ϵ is a small arbitrary positive constant added in the denominator to ensure numerical stability.
The previous normalization of $\overline{u}_i^{(k)}$ has the effect to reduce its representation ability. Batch Normalization restores the network representation power by introducing additional parameters $\gamma^{(k)}$ and $\beta^{(k)}$ that are subsequently learned during the training phase.
The Batch Normalization transformation is defined by :

$$BN_{\gamma^{(k)},\beta^{(k)}}(u^{(k)}) = \gamma^{(k)}\overline{u}_i^{(k)} + \beta^{(k)} \tag{6}$$

Where $\overline{u}^{(k)}$ remains internal to the current layer and $BN_{\gamma^{(k)},\beta^{(k)}}(u^{(k)})$ is passed to the next layer.

2.5.2 | LSTM

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 [27]. 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 fails due to the errors that accumulates with each recursion[13] and can eventually lead to a 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 opt to choose the token with the highest probability. However, in the sampling phase we update our model using temperature-decoding method, which shrink or enlarge probabilities to ensure more flexibility in the search area of the best token and to produce distinct and diverse generations while sampling our SMILES.

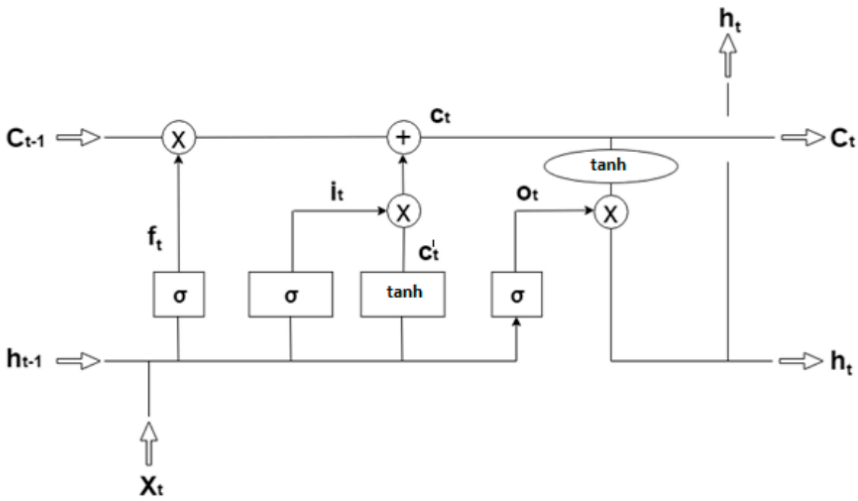

FIGURE 4 LSTM cell representation[30]

The LSTM cell an be represented by :

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i) \tag{7}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f) \tag{8}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{9}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{10}$$

$$h_t = o_t \tanh(c_t) \tag{11}$$

2.5.3 | Batch Normalization applied to LSTM

A common phenomenon that occurs within machine learning is the covariate shift[31][32], which changes the distribution of the inputs presented. In order to remediate to this issue we decide to include batch normalization layers, which is an approach for network reparametrization that standardizes the outputs using estimations of their means and standard deviations. We apply a batch normalization variant as in from input-to-hidden and hidden-to-hidden layers. So our LSTM model will be defined as follows :

$$\hat{i}_t = \sigma(BN(W_{xi}x_t) + BN(W_{hi}h_{t-1}) + W_{ci}C_{t-1} + b_i) \tag{12}$$

$$\hat{f}_t = \sigma(BN(W_{xf}x_t) + BN(W_{hf}h_{t-1}) + W_{cf}C_{t-1} + b_f) \tag{13}$$

$$\hat{c}_t = \hat{f}_t c_{t-1} + i_t \tanh(BN(W_{xc}x_t) + BN(W_{hc}h_{t-1}) + b_c) \tag{14}$$

$$\hat{o}_t = \sigma(BN(W_{xo}x_t) + BN(W_{ho}h_{t-1}) + W_{co}c_t + b_o) \tag{15}$$

$$h_t = \hat{o}_t \tanh(\hat{c}_t) \tag{16}$$

2.6 | Proposed approach

2.6.1 | Generations and transfer learning

The figure below(Fig.5) depicts the main architecture that we followed all along our work.

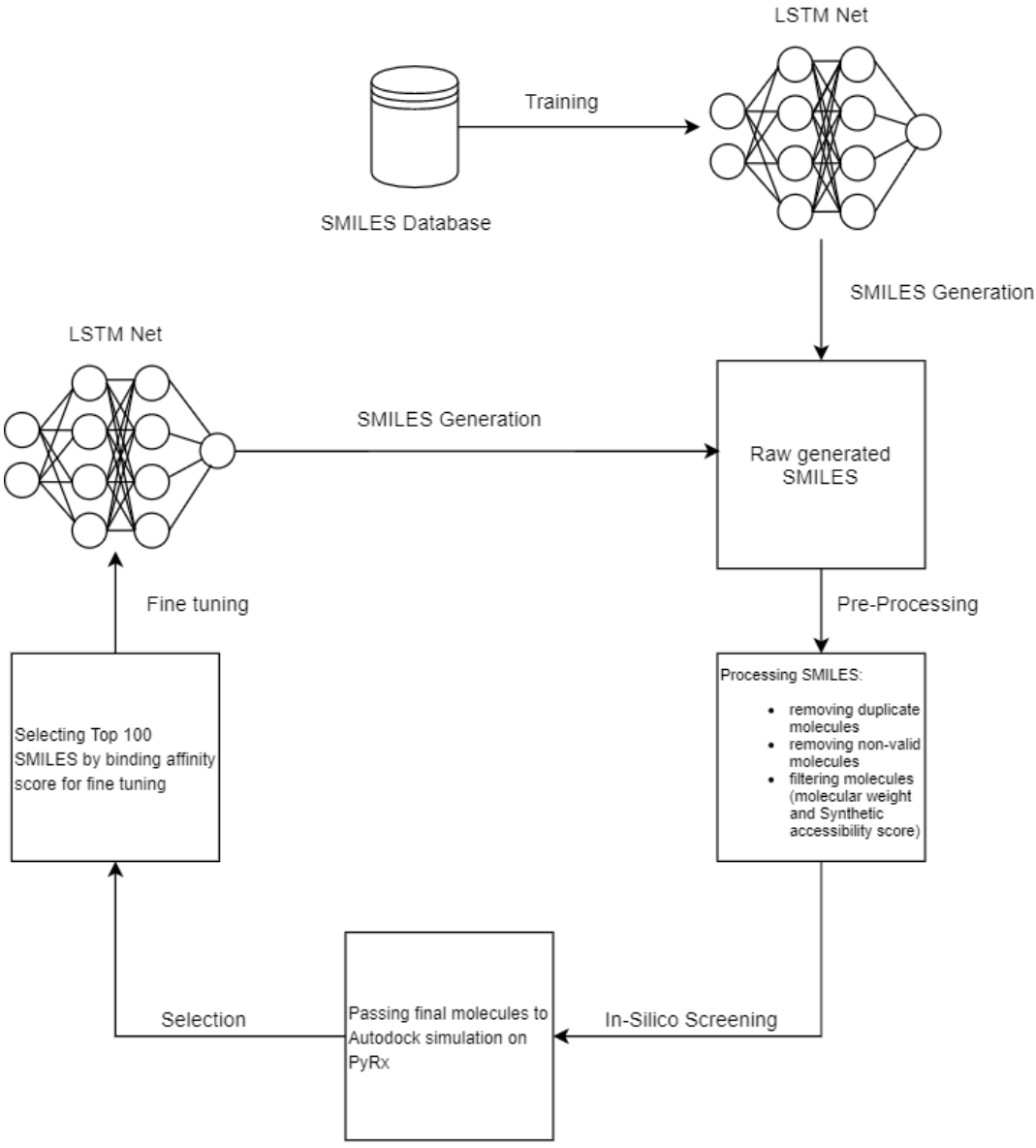

FIGURE 5 Main architecture explaining the adopted approach.

We generate with this model a batch of molecules that will be filtered according to the swiss cheese principle(Fig.6); 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.

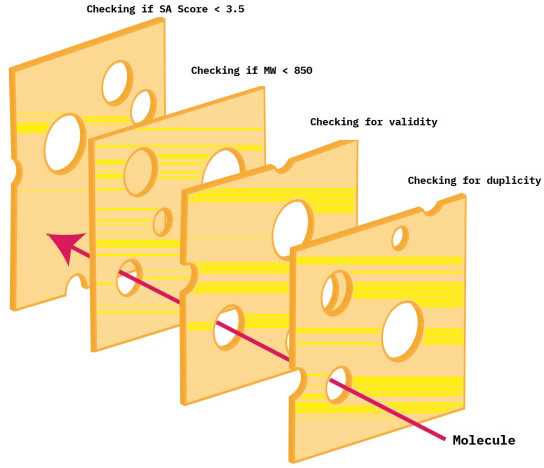

FIGURE 6 Swiss cheese principle for filtering and selecting molecules.

To give some semantics into the generated molecules we implement many metrics[13] such as :

- **Uniqueness** :
  Uniqueness is a metric that describes the number of unique SMILES within one generation batch. And is described as follows:

$$R_{uniqueness} = \frac{|set(A)|}{|A|}, R_{uniqueness} \in (0, 1] \tag{17}$$

  where : set(A) is the set of unique SMILES within A.
- **Validity** :
  Validity is a metric that describes the number of valid molecules on the generated set.

$$R_{validity} = \frac{|Valid(A)|}{|A|}, R_{Validity} \in [0, 1] \tag{18}$$

  where : A is the set of generated molecules and valid(A) is the set of valid SMILES within A.
- **Originality or novelty** :
  Originality is a metric that can describe if the model is generating new molecules and not only reproducing the molecules he has seen on the training set.

$$R_{Originality} = 1 - \frac{|G(A) \cap T(A)|}{|T(A)|}, R_{Originality} \in [0, 1] \tag{19}$$

  where : G(A) is the set of the generated molecules. T(A) is the training set of the model.
- **Synthetic feasibility** :
  Our main aim is to generate new and non-existing molecules. Thus, we need not only to generate molecules that can bind easily to M_{pro} but also molecules that would be synthetically accessible in order to deliver the molecule quickly to the market. To this end, we introduced synthetic feasibility which is a metric based on the Synthetic accessibility score[33], that represents and estimates the synthetic feasibility of a given molecule and produces an output between 1 and 10 (1 being the easiest to make and 10 the hardest) this score takes into consideration various penalties and metrics such as :

$$SizePenalty = N_a^{1.005} - N_a \tag{20}$$

$$StereoPenalty = \log_{10}(N_c + 1) \tag{21}$$

$$SpiroPenalty = \log_{10}(N_s + 1) \tag{22}$$

$$BridgePenalty = \log_{10}(N_B + 1) \tag{23}$$

$$MacrocyclePenalty = \begin{cases} \log_{10}(2) \text{ , if } N_M > 0, \\ 0 \text{ , if } N_M = 0. \end{cases} \tag{24}$$

Such as :

N_a : Number of atoms

N_C : Number of Chiral Centers

N_S : Number of Spiros

N_B: Number of Bridge heads

N_M: Number of Macrocycles

We opted towards using fine-tuning as a method for transfer learning to produce with each new generation molecules similar to the ones that achieved good binding affinity scores in the previous generations. To this end, we pick the top 100 molecules with each generation and add a smaller sample generated from the base model (gen 0) so that it promotes diversity. As the model is retrained with similar data, it starts to produce a molecule sample with a higher similarity score. Thus, the validity of the model starts to increase, and the uniqueness starts to decrease. We also increase the sample size within each generation in order to maintain a representative sample size of the pre-processed molecules.

2.6.2 | Training phase

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.25).

$$MLE = \prod_t P_\theta(x_t | X_{1:t-1}) \tag{25}$$

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)[15][34].

$$L = -\Sigma_{X \in \gamma} \Sigma_{t=1}^{T} log P_\theta(x_t | X_{1:t-1}) \tag{26}$$

Since the MLE optimization tends to capture only the peaks of the distribution and neglects the tails, we employed a multinomial sampler with a sampling temperature in order to rescale the distributions away from the peaks. It is done as follows (eq.27):

$$P_i^{new} = \frac{\exp(\frac{P_i}{T})}{\Sigma_j \exp(\frac{P_i}{T})} \tag{27}$$

where T is the sampling temperature. When generating with low values of T, the generated molecules are usually not diverse and close to molecules seen on the training set. Whereas higher values of T can lead to a new diverse set but can also cause a low validity rate on the generated set. Nevertheless, it is more probable to produce erroneous and nonsensical results when selecting from a wider space.

3 | EXPERIMENTS

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 1 fully connected layer, we proceed to remove the dropout from the BN-LSTM model, we conclude to the following results shown in figure 7 and 8.

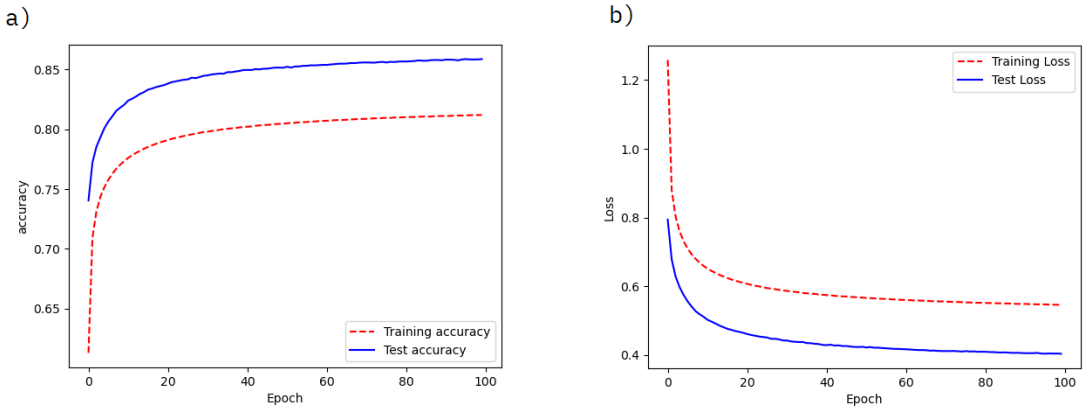

FIGURE 7 Vanilla model accuracy (a) and training loss(b) evolution per epoch. The produced model resulted in a validity value of 43.10%, a uniqueness value of 99.88% and an originality value of 99.42% within the first generated set. We observe that the validation loss is lower than the training loss. 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.

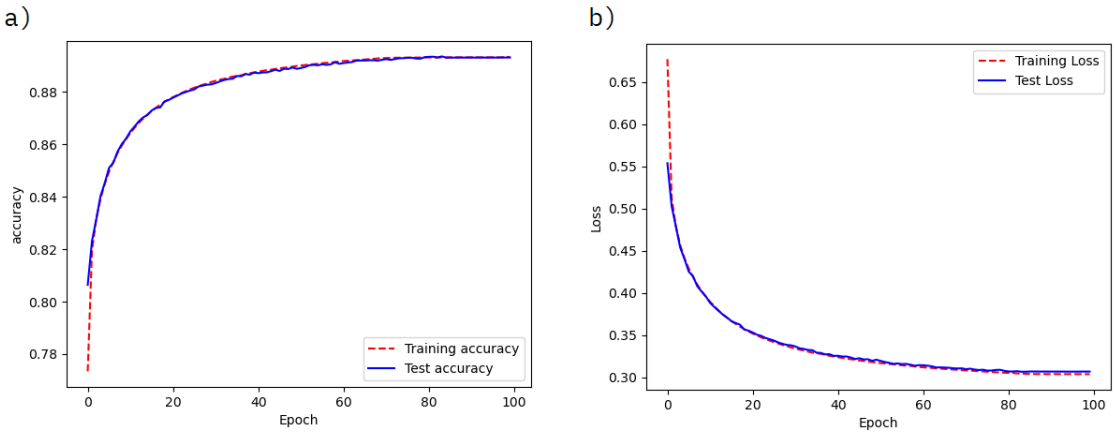

FIGURE 8 Batch Normalization model accuracy (a) and training loss(b) evolution per epoch. The produced model resulted in a validity value of 90.98%, a uniqueness value of 98.36% and an originality value of 90.37% within the first generated set.

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.

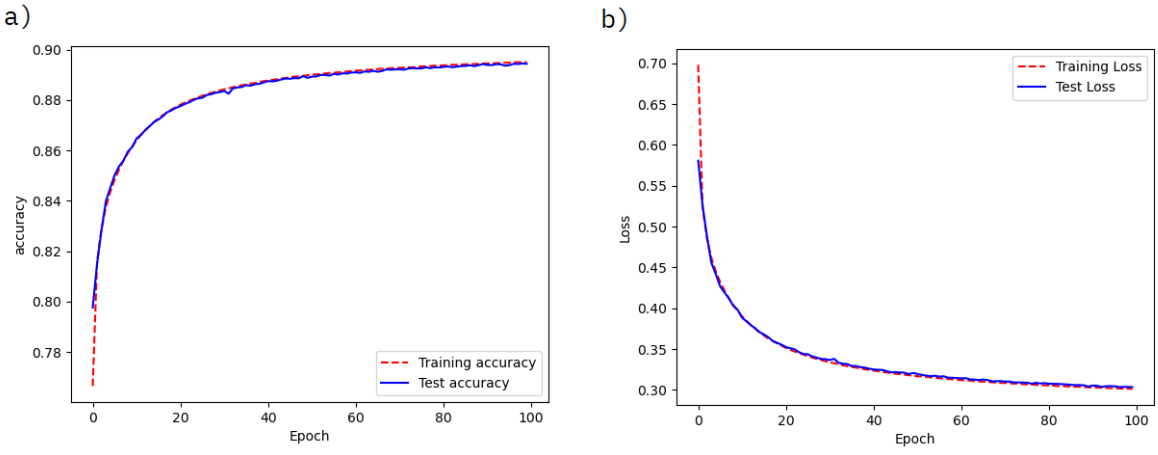

FIGURE 9 Batch Normalization model with orthogonal weight initialization accuracy (a) and training loss(b) evolution per epoch. The produced model resulted in a validity value of 92.76%, a uniqueness value of 98.16% and an originality value of 90.63% within the first generated set.

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).

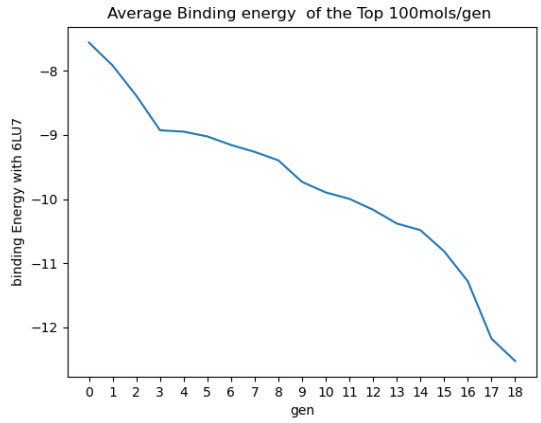

FIGURE 10 Binding energy evolution of the top 100 /gen. The figure shows that with each generation we get new generated molecules that achieves a better(smaller) binding affinity score with 6LU7.

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, high lipophilicity (Fig.11) 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].

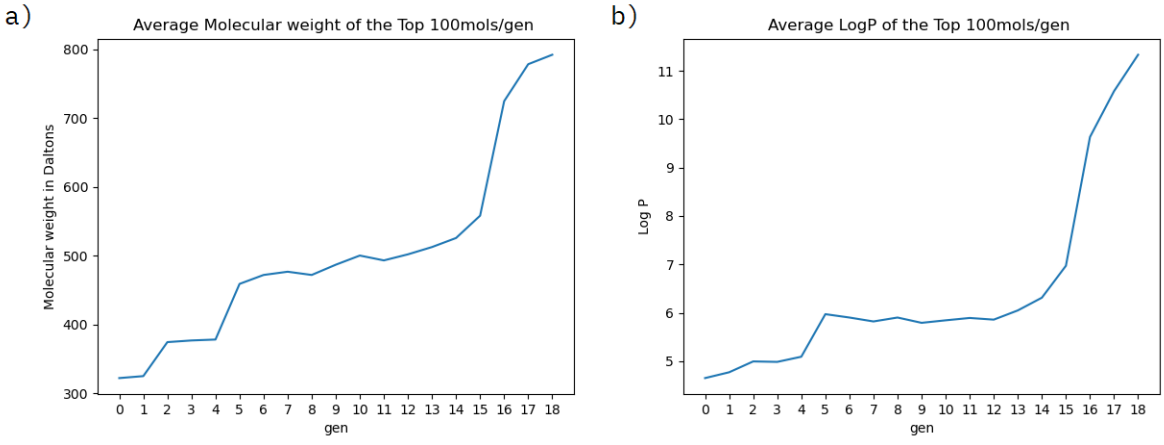

FIGURE 11 Average Molecular Weight and logP of the top 100mols/gen. The figure shows that the top 100 of the generated molecules becomes more and more heavier in terms of Molecular weight(a) and increases also in the value of the calculated logp (b) within each generation.

The table below (table.2) presents the different molecular properties of the generated molecules and some existing pharmaceutical compounds that are being evaluated to treat covid-19 such as Azythromicine [37], Remdisivir [38], Nitazoxanide [39], Lopinavir [40], Hydroxycholoroquine [41] and Chloroquine [42]. We present some of the relevant metrics for our study like the synthetic accessibility score and binding energy with M_{pro} and also some metrics relative to drug likeness and desirability like MW, cLogP, HBD, HBA mentioned on the Lipinski rule of 5, also called ro5 [43][44], and also QED or quantitative estimate of drug-likeness [45] which is a metric that reflects the underlying distribution of molecular properties relevant to drug likeness.

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

Table 2: ADME properties & metrics of the generated molecules

The figure below(Fig.12) depicts the 2-D structure of the generated molecules described in Table 2.

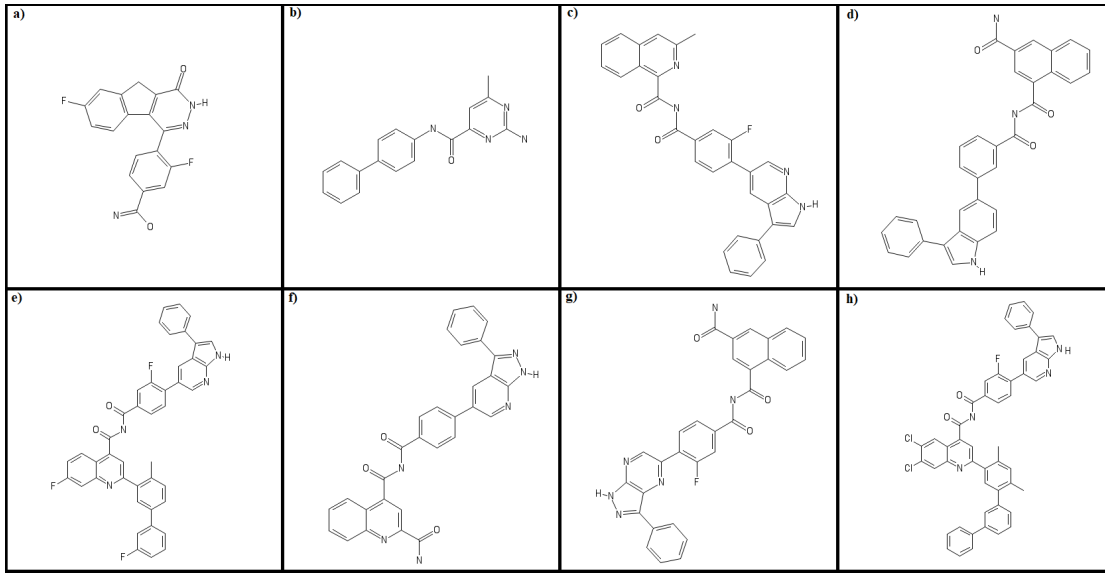

FIGURE 12 2-D structure of the generated molecules. a) depicts the molecules described in table.2 as mol1, b) depicts the molecules described in table.2 as mol2, c) depicts the molecules described in table.2 as mol3 d) depicts the molecules described in table.2 as mol4 e) depicts the molecules described in table.2 as mol5, f) depicts the molecules described in table.2 as mol6, g) depicts the molecules described in table.2 as mol7 , h) depicts the molecules described in table.2 as mol8. All the 2-D structures rendering were carried out using the pubchem sketcher web tool [46].

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 3 : SMILES representation of the generated molecules

The figures below depicts the 2-D structures of the protein-ligand interaction between the generated ligands and the M_{pro}. These figures were generated using PyMOL[47] and LigPlot+[48].

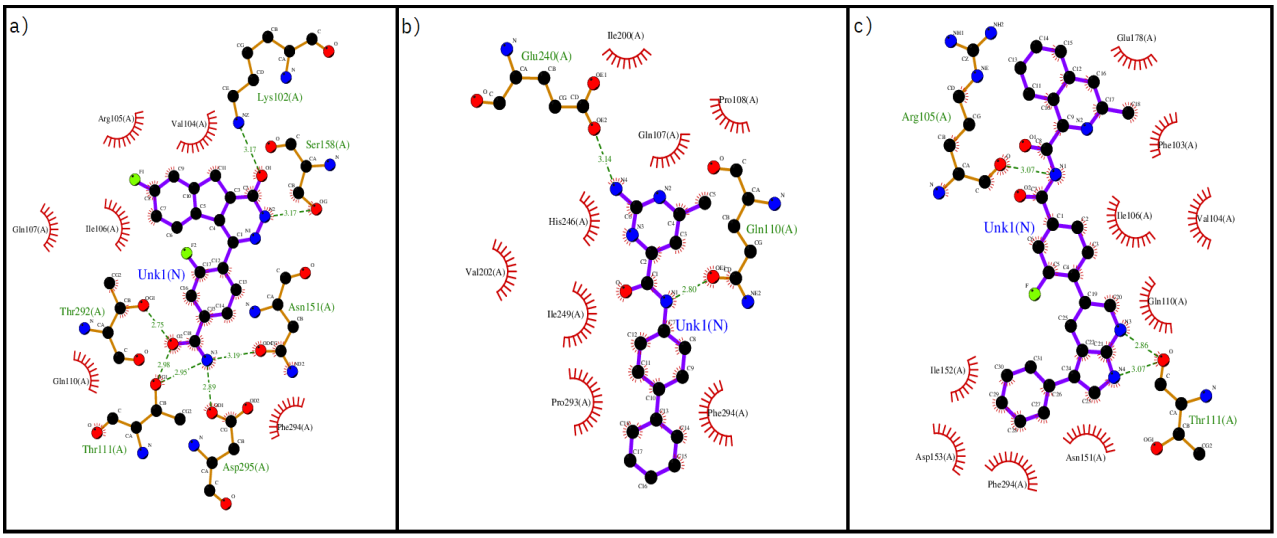

FIGURE 13 The figure shows the 2-D structure of the protein-ligand interaction between the mol1(a), mol2(b) and mol3(c) described in Table 2 and M_{pro}(6LU7).

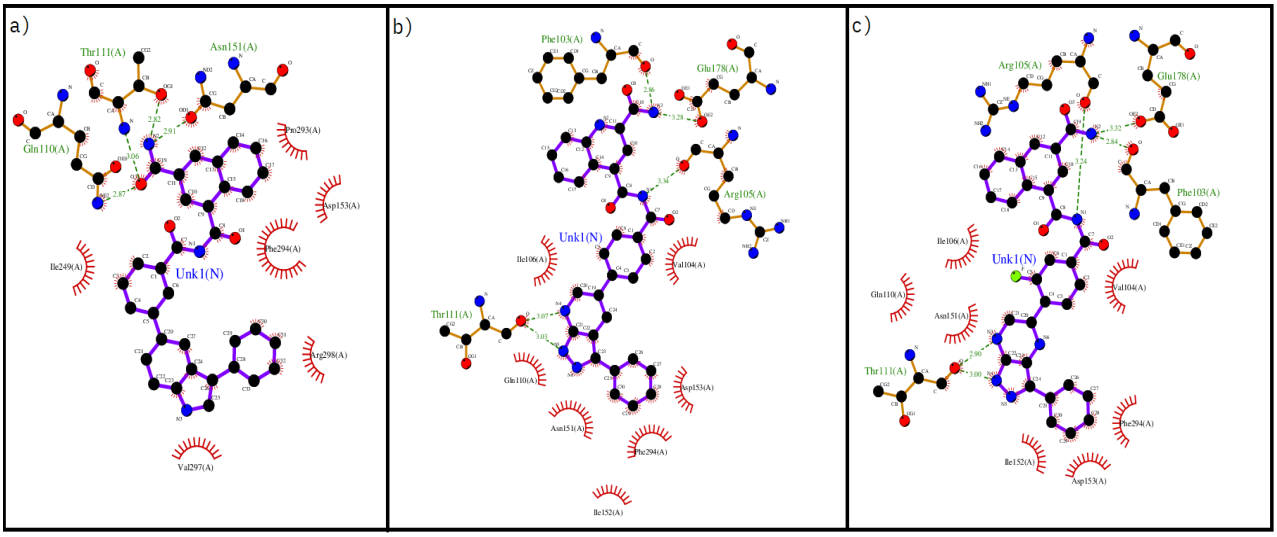

FIGURE 14 The figure shows the 2-D structure of the protein-ligand interaction between the mol4(a) and mol5(b), mol6(c) described in Table 2 and M_{pro}(6LU7).

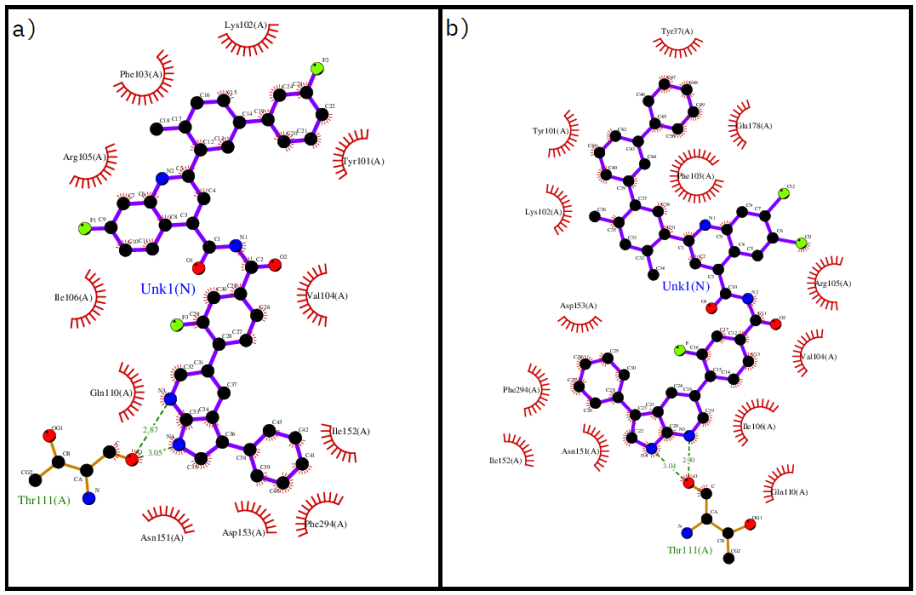

FIGURE 15 The figure shows the 2-D structure of the protein-ligand interaction between the mol7(a) and mol8(b) described in Table 2 and M_{pro}(6LU7).

4 | 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 a deep proactive transfer learning. We trained an LSTM architecture with SMILES representation of existing pharmaceutical compounds to produce our base model which has as 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 afterwards.

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] Hou, Y., Linhong, J.: Gene transcription and translation in design. In: Volume 7: 27th International Conference on Design Theory and Methodology. American Society of Mechanical Engineers (Aug 2015)

[7] Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J., and, D.L.R.: SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology 19(7), 409–424 (Jun 2021)

[8] Yang, H., Xie, W., Xue, X., Yang, K., Ma, J., Liang, W., Zhao, Q., Zhou, Z., Pei, D., Ziebuhr, J., Hilgenfeld, R., Yuen, K.Y., Wong, L., Gao, G., Chen, S., Chen, Z., Ma, D., Bartlam, M., Rao, Z.: Design of wide-spectrum inhibitors targeting coronavirus main proteases. PLoS Biology 3(10), e324 (Sep 2005)

[9] Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., Hilgenfeld, R.: Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science 368(6489), 409–412 (Mar 2020)

[10] Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., Duan, Y., Yu, J., Wang, L., Yang, K., Liu, F., Jiang, R., Yang, X., You, T., Liu, X., Yang, X., Bai, F., Liu, H., Liu, X., Guddat, L.W., Xu, W., Xiao, G., Qin, C., Shi, Z., Jiang, H., Rao, Z., Yang, H.: Structure of mpro from SARS-CoV-2 and discovery of its inhibitors. Nature 582(7811), 289–293 (Apr 2020)

[11] Dallakyan, S., Olson, A.J.: Small-molecule library screening by docking with PyRx. In: Methods in Molecular Biology, pp. 243–250. Springer New York (Dec 2014)

[12] Trott, O., Olson, A.J.: AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry pp. NA–NA (2009)

[13] Elton, D.C., Boukouvalas, Z., Fuge, M.D., Chung, P.W.: Deep learning for molecular design—a review of the state of the art. Molecular Systems Design & Engineering 4(4), 828–849 (2019)

[14] Bjerrum, E.J.: Smiles enumeration as data augmentation for neural network modeling of molecules. arxiv (Nov 2017)

[15] 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)

[16] Tang, B., He, F., Liu, D., Fang, M., Wu, Z., Xu, D.: AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2 (Mar 2020)

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

[18] Wang, Y., Lv, Z., Chu, Y.: HIV protease inhibitors: a review of molecular selectivity and toxicity. HIV/AIDS - Research and Palliative Care p. 95 (Apr 2015)

[19] de Leuw, P., Stephan, C.: Protease inhibitors for the treatment of hepatitis c virus infection. GMS Infectious Diseases; 5:Doc08 (2017)

[20] Nishimura, H., Yamaya, M.: A synthetic serine protease inhibitor, nafamostat mesilate, is a drug potentially applicable to the treatment of ebola virus disease. The Tohoku Journal of Experimental Medicine 237(1), 45–50 (2015)

[21] Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., Qin, H., Wang, M., Lu, Q., Li, X., Sun, Q., Liu, J., Zhang, L., Li, X., Huang, W., Wang, Y.: The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 182(5), 1284–1294.e9 (Sep 2020)

[22] for Disease Prevention, E.C., Control.: Rapid increase of a sars-cov-2 variant with multiple spike protein mutations observed in the united kingdom – 20 december 2020 (2020), https://www.ecdc.europa.eu/en/publications-data/threat-assessment-brief-rapid-increase-sars-cov-2-variant-united-kingdom

[23] Shirato, K., Kawase, M., Matsuyama, S.: Middle east respiratory syndrome coronavirus infection mediated by the transmembrane serine protease TMPRSS2. Journal of Virology 87(23), 12552–12561 (Sep 2013)

[24] Zumla, A., Chan, J.F.W., Azhar, E.I., Hui, D.S.C., Yuen, K.Y.: Coronaviruses — drug discovery and therapeutic options. Nature Reviews Drug Discovery 15(5), 327–347 (Feb 2016)

[25] Berman, H.M.: The protein data bank. Nucleic Acids Research 28(1), 235–242 (Jan 2000)

[26] Landrum, G.: Rdkit: Open-source cheminformatics, http://www.rdkit.org

[27] Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: Sequence generative adversarial nets with policy gradient. CoRR abs/1609.05473 (2016)

[28] Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks (01 2016)

[29] Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. p. 1171–1179. NIPS'15, MIT Press, Cambridge, MA, USA (2015)

[30] Poornima, S., Pushpalatha, M.: Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units. Atmosphere 10(11), 668 (Oct 2019)

[31] Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference 90(2), 227–244 (Oct 2000)

[32] Cooijmans, T., Ballas, N., Laurent, C., Courville, A.C.: Recurrent batch normalization. CoRR abs/1603.09025 (2016)

[33] Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics 1(1) (Jun 2009)

[34] Kawthekar, P., Rewari, R., Bhooshan, S.: Evaluating generative models for text generation (2017)

[35] Tsaioun, K.: Evidence-based absorption, distribution, metabolism, excretion (ADME) and its interplay with alternative toxicity methods. ALTEX pp. 343–358 (2016)

[36] Daina, A., Michielin, O., Zoete, V.: SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Scientific Reports 7(1) (Mar 2017)

[37] Furtado, R.H.M., Berwanger, O., Fonseca, H.A., Corrêa, T.D., Ferraz, L.R., Lapa, M.G., Zampieri, F.G., Veiga, V.C., Azevedo, L.C.P., Rosa, R.G., Lopes, R.D., Avezum, A., Manoel, A.L.O., Piza, F.M.T., Martins, P.A., Lisboa, T.C., Pereira, A.J., Olivato, G.B., Dantas, V.C.S., Milan, E.P., Gebara, O.C.E., Amazonas, R.B., Oliveira, M.B., Soares, R.V.P., Moia, D.D.F., Piano, L.P.A., Castilho, K., Momesso, R.G.R.A.P., Schettino, G.P.P., Rizzo, L.V., Neto, A.S., Machado, F.R., Cavalcanti, A.B.: Azithromycin in addition to standard of care versus standard of care alone in the treatment of patients admitted to the hospital with severe COVID-19 in brazil (COALITION II): a randomised clinical trial. The Lancet 396(10256), 959–967 (Oct 2020)

[38] Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., Shi, Z., Hu, Z., Zhong, W., Xiao, G.: Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Research 30(3), 269–271 (Feb 2020)

[39] Cao, J., Forrest, J.C., Zhang, X.: A screen of the NIH clinical collection small molecule library identifies potential anti-coronavirus drugs. Antiviral Research 114, 1–10 (Feb 2015)

[40] Marzolini, C., Stader, F., Stoeckle, M., Franzeck, F., Egli, A., Bassetti, S., Hollinger, A., Osthoff, M., Weisser, M., Gebhard, C.E., Baettig, V., Geenen, J., Khanna, N., Tschudin-Sutter, S., Mueller, D., Hirsch, H.H., Battegay, M., Sendi, P.: Effect of systemic inflammatory response to SARS-CoV-2 on lopinavir and hydroxychloroquine plasma concentrations. Antimicrobial Agents and Chemotherapy 64(9) (Aug 2020)

[41] 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)

[42] Vincent, M.J., Bergeron, E., Benjannet, S., Erickson, B.R., Rollin, P.E., Ksiazek, T.G., Seidah, N.G., Nichol, S.T.: Virology Journal 2(1), 69 (2005)

[43] Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1pii of original article: S0169-409x(96)00423-1. the article was originally published in advanced drug delivery reviews 23 (1997) 3–25. 1. Advanced Drug Delivery Reviews 46(1-3), 3–26 (Mar 2001)

[44] Benet, L.Z., Hosey, C.M., Ursu, O., Oprea, T.I.: BDDCS, the rule of 5 and drugability. Advanced Drug Delivery Reviews 101, 89–98 (Jun 2016)

[45] Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L.: Quantifying the chemical beauty of drugs. Nature Chemistry 4(2), 90–98 (Jan 2012)

[46] 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)

[47] Schrödinger, LLC: The PyMOL molecular graphics system, version 1.8 (November 2015)

[48] Laskowski, R.A., Swindells, M.B.: Ligplot+: Multiple ligand-protein interaction diagrams for drug discovery. Journal of Chemical Information and Modeling 51(10), 2778–2786 (Oct 2011)