Generating Novel Compounds Targeting SARS-CoV-2 Main Protease Based On Imbalanced Dataset

Fan Hu[1†], Dongqi Wang[1,2†], Yishen Hu[1,2], Jiaxin Jiang[1], Peng Yin[1*]

[1]*Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China*

[2]*University of Chinese Academy of Science, Beijing, 100049 ,China*

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313317

***Abstract*—The *de novo* drug design plays an important role in the drug discovery. Recently deep learning based method has been popular as a promising approach for the design of novel drugs with desirable properties. However, conventional target-specific generative models mainly concentrate on the known inhibitors and thus produce similar molecules. And these derivatives of known inhibitors are probably negative against the same target. Considering the cost of chemical synthesis and experimental validation, the low false positive rate of generative molecules is very important. In this paper, we propose an efficient pipeline to generate novel SARS-CoV-2 3C-like protease inhibitors. Based on the GPT2 generator and the well performing multi-task predictor which achieves high precision on the highly imbalanced 3CL in vitro screening dataset (650 positive of 297,467 molecules), we acquired a number of novel 3CL-target compounds and analyzed their molecular properties. Moreover, we applied randomized SMILES for data augmentation of positive molecules to create larger chemical space for the generator. Finally, the selected positive compounds with desirable properties are exhibited, as well as their nearest neighbors of 3CL inhibitors which have already been verified in vitro.**

Keywords—de novo drug design, SARS-CoV-2, 3C-like protease, imbalanced dataset, transformer

I. INTRODUCTION

The ongoing COVID-19 pandemic has so far sickened more than twenty millions and killed hundreds of thousands people across globe, which is caused by SARS-CoV-2, a member of the family Coronaviridae [1]. Similar to MERS-CoV and SARS-CoV, SARS-CoV-2 can cause severe respiratory diseases in human. It is urgent to find some ways to inhibit the virus and get everything back on track. Given the fact that the development of effective and safety vaccine can take years, discovery of drugs that inhibit SARS-CoV-2 would be very important, especially for those people who had already been infected. However, the effect of commercial drug repurposing seems to be lower than expected [2, 3]. Perhaps the *de novo* design of novel compounds that specifically targeting SARS-CoV-2 may be a good choice. One main challenge for traditional drug design is multi-objective optimization problem, that is, chemists have to consider many parameters that are hard to explore systematically, as well as increasingly large and complex chemical space which has been now estimated more than 10^{60} molecules.

More recently, as a powerful tool for big data, deep learning has been introduced into the area of drug discovery and drug design. Researchers took large sets of existing compounds to train their model and then generated novel compounds according to the distribution. There are several ways for representing compounds for machine learning models. A broadly used input format is Simplified Molecular Input Line Entry Specification (SMILES), which encodes compound into a sequence of ASCII strings using a depth-first graph traversal. With SMILES as put, researchers has proposed many deep models for drug molecules generation. Segler et al. [4] introduced a recurrent neural network (RNN) based model to generate novel molecules, which is trained to predict the next character given previous characters in the input SMILES. They further fine-tuned the model on sets of known inhibitors and thus produced potential novel compounds against targets. Interestingly, they simulated the Design-Synthesis-Test Cycles to design positive molecules, but the generalization of target prediction part was critical. In another study, Gómez-Bombarelli et al. [5] reported a Variational Autoencoder (VAE) method consisting of three coupled parts: encoder, decoder, and predictor. The encoder converted the one-hot vector of SMILES into a real-valued dense vector, the decoder converted this vector back to SMILES representation vector, and the predictor estimated chemical properties based on the latent dense vector. Another promising method is to design drug using Generative adversarial network (GAN). Guimaraes et al. [6] proposed a method combined GAN and reinforcement learning (RL), which bias the model to generate molecules with desirable properties. RL has increasingly become a popular reward function, which can be combined or separately used, to fine tune the generator to produce molecules with high reward. For example, Zhavoronkov et al. [7] discovered discoidin domain receptor 1 (DDR1) portent novel inhibitor in 21 days by the combination of VAE and RL.

During the COVID-19 pandemic, these AI aided methods have naturally attracted attention for the fast design of novel antivirals and several teams have reported some target specific generative molecules for SARS-CoV-2. Tang et al. presented a deep Q-learning network combined with the fragment-based drug design for generating drugs and

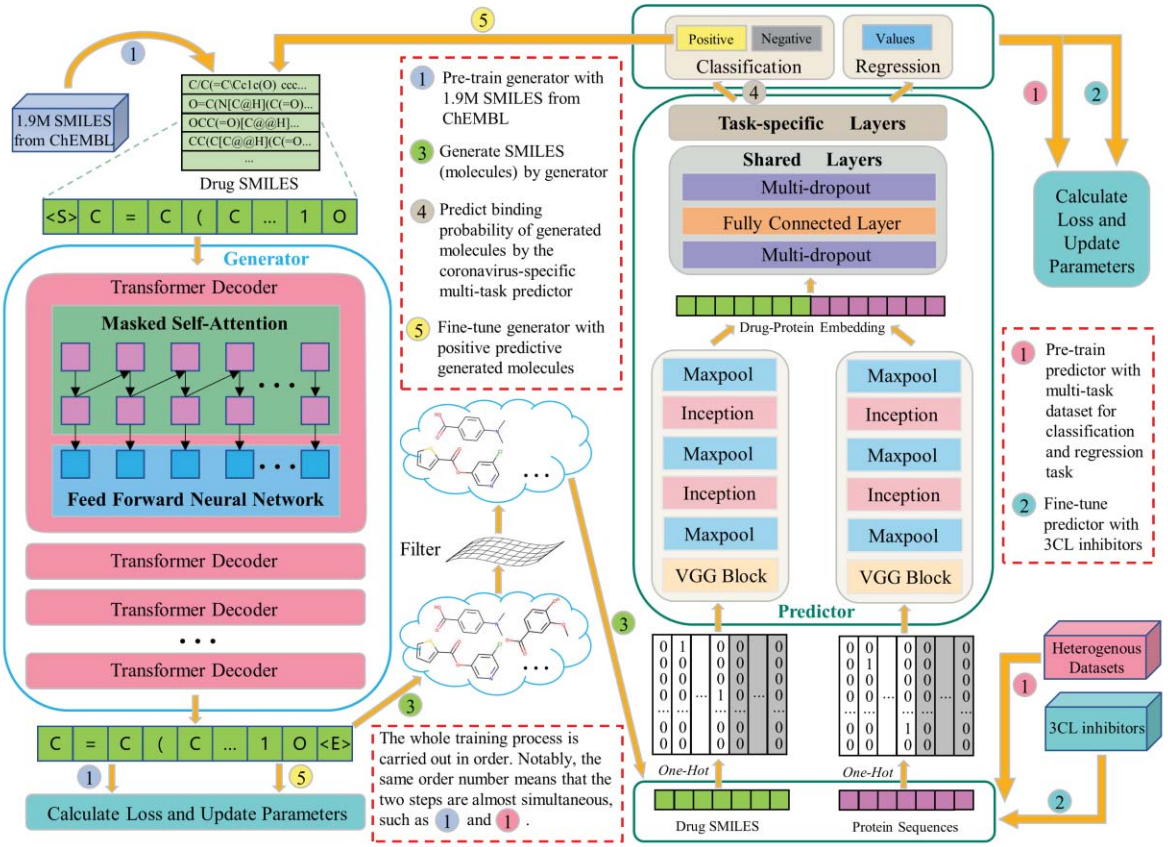

Fig. 1. The schematic of proposed method

exhibited 47 potential inhibitors targeting SARS-CoV-2 3CL protease [8]. Integrating protein pocket and ligand features, Zhavoronkov et al. showed 10 representative generated SARS-CoV-2 inhibitors that target on 3C-like protease [9]. Another study proposed a generative framework CogMol to design drug candidates targeting on three relevant proteins of SARS-CoV-2 with high off-target selectivity [10].

However, to acquire desirable inhibition activity, most target specific generative models were fine-tune on the known inhibitors against given targets and thus produced similar molecules. In fact, many derivatives of known inhibitors are probably to be negative, even against the same targets. Keeping the balance between novelty and binding affinity of generative molecules remains a challenge. Considering the cost of chemical synthesis and experimental validation, the low false positive rate of generative molecules is very important, that is, excluding molecules with low binding possibility as much as possible.

Inspired by previous studies, we propose a pipeline to generate novel SARS-CoV-2 3CL protease inhibitors (as displayed in Fig. 1), which consists of generator (generate molecules), filter (control property) and predictor (get positive molecules). At the beginning, the GPT-2 generator is trained on 1.9 million bioactive molecules and then produces diverse new compounds. After passing filter, the remaining generative molecules would be predicted by the multi-task predictor and those positive molecules would be used to fine-tune the generator. For each epoch, generator produces 100,000 new molecules. These molecules would be assessed by several scoring metrics and the ratio of predicted positive would be calculated. Rather than fine-tuning generator directly on the known 3CL inhibitors, we prefer to fine-tune our generator based on the predicted positives. Most importantly, the well performing multi-task predictor is critical for this task, which is pre-trained on large heterogeneous datasets and fine-tuned by experimental 3CL protease inhibitors. The predictor achieves Precision=0.67 on the highly imbalanced 3CL inhibitors dataset (650 positive of 297,467 total), which means the false positive is relatively low. Notably, the ratio of predicted positives to total generations is extremely low at early epochs (e.g., 16 of 100,000 at 1st epoch), which is similar to real-world situation that the hit rates are very low. Thus we explore data augmentation for positive molecules by randomized SMILES technique to create larger chemical space for generator. Finally, the selected generative molecules are exhibited, as well as their corresponding nearest neighbors of the experimental 3CL inhibitors.

II. Methods

A. Data

ChEMBL-27, which contains 1.9 million bioactive molecules with drug-like properties, is used to pre-train the generator [11]. A SARS/MERS/SARS-CoV-2 3CL protease inhibitor specific dataset is collected and used to fine-tune the predictor, which contains 650 positive and 296,817 negative molecules. The inhibitors with binding affinity (IC50/Kd/Ki) lower than 10 μM are regarded as positive.

B. Generator

Here we use GPT-2 to generate molecules. GPT-2, the direct successor to GPT, consists of transformer decoder modules. Briefly, four basic components in transformer are

multi-head attention, residual connection, normalization and linear layer. Instead of performing a single attention function with d_1 -dimensional keys, values and queries, multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions [12]:

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^O \quad (1)$$

$$\text{where } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

Where the projections are parameter matrices $W_i^Q \in R^{d_1\times d_2}, W_i^K \in R^{d_1\times d_2}, W_i^V \in R^{d_1\times d_1}$, and $W^O \in R^{d_1\times d_1}$, in which $d_1 = hd_2$, and h is the number of heads. In practice, we compute each matrix of outputs as 'scaled dot-product attention':

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_i}})V \quad (3)$$

Equation (3) computes the compatibility function using a feed-forward network with a single hidden layer, and apply a softmax function to obtain the weights on the value.

GPT-2 is trained using causal language modeling, that is, GPT-2 learns to predict a word given only its left context. And this property makes GPT-2 very good at text generation.

C. Filter

The chemical validity and uniqueness of generated molecules are checked by RDKit package and the invalid and duplicated molecules are excluded. Then several metrics are used to filter these molecules to ensure the desired drug properties. Specifically, the molecular weight is in the range of 200 to 550 Dalton, the number of hydrogen acceptors is not more than 10, the number of rotatable bonds is not more than 10 and the octanol-water partition coefficient (logP) is not more than 5.

D. Predictor

Here we use a multi-task CNN model as predictor. Briefly, the model consists of two parts: shared layers and task-specific layers. The shared layers are designed to learn a joint representation for all tasks and the task-specific layers use the joint representation to learn the weights of specific blocks based on specific tasks. Here two related tasks are defined: binary classification and regression. The multi-task model was pre-trained by large amounts of data from various heterogeneous protein-ligand datasets (unpublished). Then the model is fine-tuned on the imbalanced 3CL inhibitors dataset.

E. Evaluation

The metrics Accuracy, Precision, Recall, F1score, Area under the curve (AUC) and Matthews correlation coefficient (MCC) are used to evaluate the performance of predictor. Formulas are listed below:

$$Accuracy = \frac{TP+TN}{P+N} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ score = 2\times\frac{precision\times recall}{precision+recall} \quad (7)$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8)$$

where TP is true positive, TN is true negative, FP is false positive, FN is false negative, P is positive, N is negative.

Besides, four objectives including solubility, drug-likeness, synthesizability and chemical similarity are used to evaluate generated molecules [13]. Briefly, solubility refers to octanol-water partition coefficient (logP), which is the concentration ratio of the substance between a water-saturated octanolic phase and an octanol-saturated aqueous phase, and indicates the lipophilicity of molecules [14]. The logP of a potential orally active drug should not more than five. Drug-likeness, which indicates how likely a molecule can be a drug, is scored by Quantitative Estimate of Druglikeness (QED) [15]. Usually the QED score is in the range of 0 to 1, the higher score the better. Synthesizability refers to how easy a molecule can be synthesis and is quantified by Synthetic Accessibility (SA) score [16]. The SA score is between 1 (easy to make) and 10 (very difficult to make). Chemical similarity, which indicates how similar is a molecule to another, plays an important role in drug discovery such as identifying compounds with similar bioactivities based on structural similarity. The Tanimoto similarity metric is used to evaluate the similarity between molecules.

III. Results

A. Predictor Training

A well performing predictor is extremely important for this task. We are more depending on the performance of predictor to tell us whether the generated molecules are "true", compared to an image discriminator. A recent study showed that the imbalanced nature of the datasets has a negative impact on the classification performance of machine learning algorithms [17]. In their study, the precision scores are extremely low (most lower than 0.01) on several highly imbalanced drug datasets, although several tricks has been tried.

Here we use a multi-task CNN model as predictor. The multi-task predictor was previously pre-trained by large amounts of data from various heterogeneous protein-ligand datasets (unpublished). To verify the effectiveness of fine-tune (in which the model has learned latent "knowledge" from heterogeneous protein-ligand interactions), we compare the performance of models either fine-tuned or trained from scratch on 3CL-inhibitor dataset. Several metrics are used to evaluate the model performance. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications and is commonly used in the imbalanced data classification. It returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

Different from directly fine-tuned generator on the whole 3CL inhibitors, we explore the possibility of focusing our generator on the predicted positives by predictor. One advantage is that the false positive of the predicted positive molecules would be very low, because the predictor has

"seen" large amounts of experimental negatives and achieves high precision on the highly imbalanced 3CL-inhibitor dataset. But it should also be noted that the coverage probability of experimental positives might be low, as can be inferred from the low recall. At last, the fine-tuned model is used as our predictor.

B. Generation of Positive Compounds

For a more quantitative assessment, we compare the positive predictive generated molecules with 650 positive 3CL inhibitors using the Tanimoto similarity score. Fig. 2 shows the distributions of the nearest neighbor Tanimoto similarity score, which is generated by comparing generated molecules and their nearest neighbors in the 3CL inhibitors. As epochs goes on, the model produces more and more similar molecules to those inhibitors, and most generated molecules with Tanimoto similarity score more than 0.7.

The rediscovery (Tanimoto similarity score=1) means the generated molecules has exactly the same substructure fingerprint as one positive inhibitor. Interestingly, the model have rediscovered 32 different inhibitors within 10 epochs. It should be noted that, rather than fine-tuning directly on 3CL inhibitors, our generator only "see" positive molecules predicted by predictor. The generator is trying to replicate the probability distribution of the predictor. Moreover, 38 and 49 different inhibitors have been re-produced when Tanimoto similarity score thresholds are set to 0.9 and 0.8, respectively.

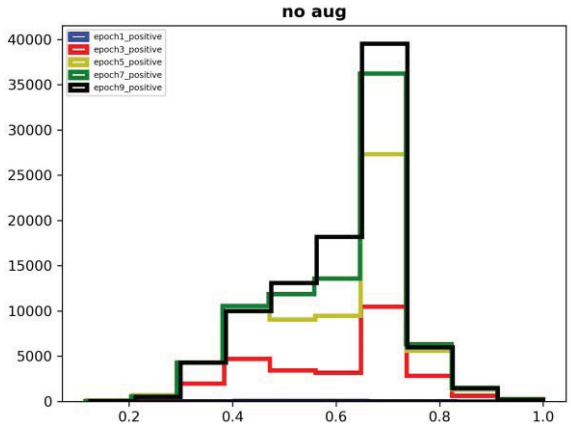

Fig. 2. Nearest neighbor Tanimotosimilarity of generated molecules for 3CL-inhibitors. Coordinates of x and y: nearest neighbor Tanimoto similarity score and the numbers of generated molecules.

C. Data Augmentation for Positives

As mentioned above, the ratio of predicted positives to total generations is extremely low at early epochs (e.g., 16 of 100,000 at epoch 0), which is similar to real-world situation that the hit rates are very low. We have explored a data augmentation method for these positive predictive generated molecules. The numbers of molecules used for fine-tune are amplified 10 times each epoch by using randomized SMILES technique [18]. Briefly, multiple SMILES strings of a molecule can be achieved by randomizing the atom ordering, which does not alter the way the molecule graph is traversed but changes only the starting point and the order of branching paths. Thus a maximum of n! different SMILES strings could be generated for a molecule with n heavy atoms.

The results displayed in Fig. 3 indicate that, this data augmentation experiment performs worse than experiment without data augmentation. Most generated molecules have nearest neighbor Tanomi similarity lower than 0.5, which indicates that the distribution of generator is not getting close to that of the predictor. The data augmentation group has rediscovered 16 different inhibitors, and 23 and 29 different inhibitors when the Tanimoto similarity thresholds are set to 0.9 and 0.8, respectively.

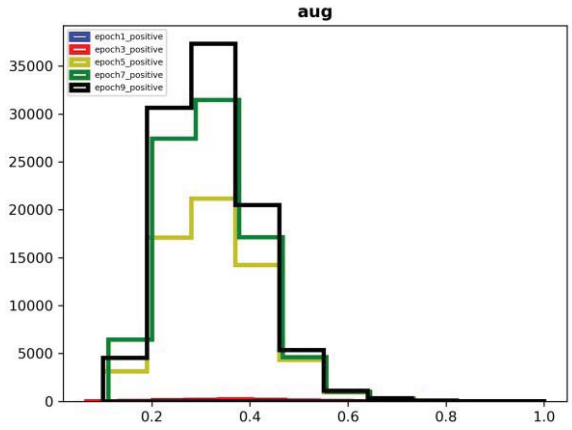

Fig. 3. Nearest neighbor Tanimoto similarity of generated molecules for 3CL-inhibitors (with data augmentation for positive predictive molecules). Coordinates of x and y: nearest neighbor Tanimoto similarity score and the numbers of generated molecules.

D. Property of Generated Molecules

According to the Lipinski's rule of five, a likely orally active drug for human needs to have some limited chemical properties and physical properties including: no more than 5 hydrogen bond donors, no more than 10 hydrogen bond acceptors, a molecular mass less than 500 Daltons, an octanol-water partition coefficient (logP) that does not exceed 5. After passing the filter, all the remaining generated molecules are eligible.

Validity (%) is the ratio between the number of valid molecules and all generated molecules and uniqueness (%) is the ratio between the number of unique molecules and all valid molecules. As epoch goes on, the validity of no augmented group continue increasing, whereas the validity of augmented group decreases before it begins to increase. This difference suggests that maybe the data augmentation introduced some noise to generator in the fine-tuning process. As for uniqueness, the data augmentation group performs better, which indicates that the degree of variety in this group is higher. But the generator may be much easier to get stuck in the characteristics of a small group of augmented inhibitors, as can be inferred from the number of rediscovered inhibitors. The no augmented group has totally rediscovered 32 unique inhibitors whereas the rediscovered unique inhibitors by the augmented group is 16.

SA score, which refers to Synthetic Accessibility (SA) score, is between 1 (easy to make) and 10 (very difficult to make). QED score is in the range of 0 to 1, is used to evaluate the drug-likeness of a molecule, the higher score the better. The SA scores of augmented group are generally higher than those of the no augmented group, which indicates the generated molecules in augmented group are a bit more difficult to synthesize. The QED scores of both

groups increase along epoch, suggesting a trend toward more drug-likeness.

E. Novel Molecules Display

Some of the selected molecules and their nearest neighbors in the known 3CL-inhibitors are also shown in Fig. 4. Some generated novel molecules have higher QED scores than their nearest neighbors, suggesting a higher drug-likeness. As mentioned above, the filter allows only molecules which follow the Lipinski's rule of five to pass, it is not surprising that these novel molecules have better properties than the known inhibitors.

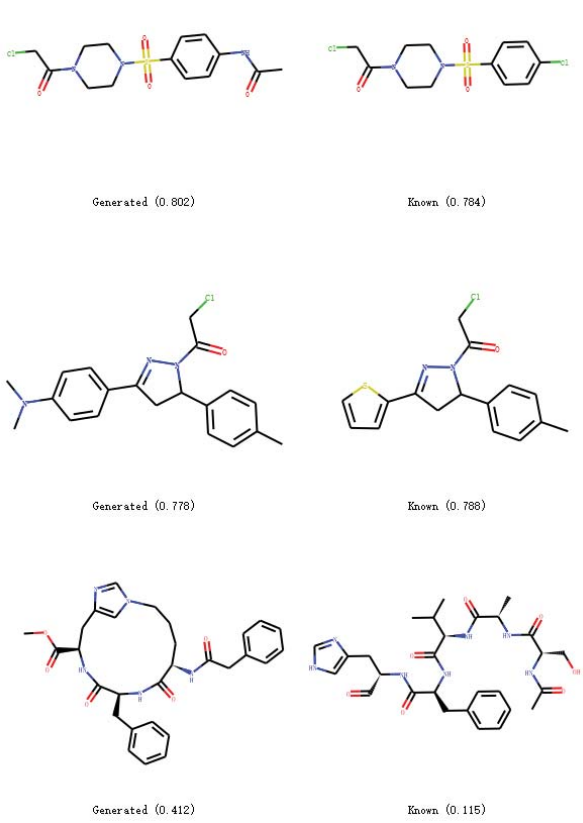

Fig. 4. Selected generated molecules with their nearest neighbors in the known 3CL-inhibitors. The number in bracket is the corresponding QED.

IV. Conclusion

In this study, we introduced a novel pipeline to generate novel SARS-CoV-2 3CL protease inhibitors. Based on the GPT2 generator and the well performing multi-task predictor on the highly imbalanced 3CL dataset, the false positive rate of our generative 3CL targeted molecules is presumably low. Additionally, we have leveraged randomized SMILES method to augment positive molecules, which may create larger chemical space for the generator. Finally, several selected generative molecules with potential inhibitory activity against 3CL protease are shown, along with their nearest neighbors of known inhibitors. We hope these results would be helpful in the fight against COVID-19. Future work will focus on improving the generalizability of predictor by using more virus specific data, and the optimization of property control process generation process by combining reinforcement learning.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NO. 11801542), the Strategic Priority Research Program of Chinese Academy of Sciences (NO. XDB 3804020), the Shenzhen Fundamental Research Projects(JCYJ20170818164014753, JCYJ20170818163445670 and JCYJ20180703145002040).

References

[1] A. E. Gorbalenya *et al.*, "Severe acute respiratory syndrome-related coronavirus: The species and its viruses-a statement of the Coronavirus Study Group," *bioRxiv*, 2020.

[2] Y. Wang *et al.*, "Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial," *Lancet*, vol. 395, no. 10236, pp. 1569–1578, May 2020.

[3] B. Cao *et al.*, "A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19," *N. Engl. J. Med.*, vol. 382, no. 19, pp. 1787–1799, May 2020.

[4] M. H. S. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS Cent. Sci.*, vol. 4, no. 1, pp. 120–131, 2018.

[5] R. Gómez-Bombarelli *et al.*, "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules," *ACS Cent. Sci.*, vol. 4, no. 2, pp. 268–276, 2018.

[6] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, and A. Aspuru-Guzik, "Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models," *Arxiv*, May 2017.

[7] A. Zhavoronkov *et al.*, "Deep learning enables rapid identification of potent DDR1 kinase inhibitors," *Nat. Biotechnol.*, vol. 37, no. 9, pp. 1038–1040, 2019.

[8] B. Tang, F. He, D. Liu, M. Fang, Z. Wu, and D. Xu, "AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2.," *bioRxiv*, p. 2020.03.03.972133, 2020.

[9] A. Zhavoronkov *et al.*, "Potential Non-Covalent SARS-CoV-2 3C-like Protease Inhibitors Designed Using Generative Deep Learning Approaches and Reviewed by Human Medicinal Chemist in Virtual Reality," *chemRxiv*, 2020.

[10] V. Chenthamarakshan *et al.*, "Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models," *Arxiv*, no. April, 2020.

[11] D. Mendez *et al.*, "ChEMBL: towards direct deposition of bioassay data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D930–D940, Jan. 2019.

[12] A. Vaswani *et al.*, "Attention Is All You Need," *Arxiv*, Jun. 2017.

[13] D. Polykovskiy *et al.*, "Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models," *Arxiv*, pp. 1–20, 2018.

[14] S. A. Wildman and G. M. Crippen, "Prediction of Physicochemical Parameters by Atomic Contributions," *J. Chem. Inf. Comput. Sci.*, vol. 39, no. 5, pp. 868–873, Sep. 1999.

[15] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, "Quantifying the chemical beauty of drugs," *Nat. Chem.*, vol. 4, no. 2, pp. 90–98, Feb. 2012.

[16] P. Ertl and A. Schuffenhauer, "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions," *J. Cheminform.*, vol. 1, no. 1, p. 8, Dec. 2009.

[17] S. Korkmaz, "Deep Learning-Based Imbalanced Data Classification for Drug Discovery," *J. Chem. Inf. Model.*, 2020.

[18] J. Arús-Pous *et al.*, "Randomized SMILES strings improve the quality of molecular generative models," *J. Cheminform.*, vol. 11, no. 1, pp. 1–13, 2019.