 doi:10.20944/preprints202010.0196.v1

Article

De Novo Drug Design using Artificial Intelligence ASYNT-GAN

Ivan Jacobs [1], **Manolis Maragoudakis** [2] *

[1] Ivan Jacobs; ivan.jacobs@ai4u.ai
[2] Manolis Maragoudakis; mmarag@ionio.gr
* Correspondence ivan.jacobs@ai4u.ai

Abstract: Computer-assisted de novo design of natural product mimetics offers a viable strategy to reduce synthetic efforts and obtain natural-product-inspired bioactive small molecules but suffers from several limitations. Deep Learning techniques can help address these shortcomings. We propose the generation of synthetic molecule structures that optimizes the binding affinity to a target. To achieve this, we leverage on important advancements in Deep Learning. Our approach generalizes to systems beyond the source system and achieves generation of complete structures that optimize the binding to a target unseen during training. Translating the input sub-systems into the latent space permits the ability to search for similar structures and the sampling from the latent space for generation.

Keywords: de novo drug design; synthetic molecule structure generation; deep learning;

1. Introduction

An outbreak of the novel coronavirus SARS-CoV-2 has caused worldwide social and economic disruption. The scientific community has limited knowledge of the molecular details of the infection. Currently there is a lack of common adaption of antiviral drugs nor are there vaccines for prevention.

The time and effort to create and market a drug or vaccine that can treat a certain infection can span over decades and millions of investments. Throughout the process of drug discovery, one of the main challenges is to identify a molecular structure that can attach itself to a target. The quality of the binding has a direct influence on side effects and effectiveness of the treatment. Multiple approaches exist to find or create these structures. On successful structure identification, the compounds structure is further built around it.

Computer-assisted de novo design permits to reduce efforts and obtain natural-product-inspired bioactive molecules. However, the currently applied methods have shown multiple limitations, in particular unsatisfactory scoring of biological activities. Machine Learning (ML) combined with computational power is a valid candidate to optimize these methods.

Three types of in silico drug-target interaction DTI prediction methods have been proposed in the literature: molecular docking, similarity-based, and deep learning-based models. Molecular docking is a simulation based on human predefined rules aiming to optimize the conformation of ligand and target protein. Although it can provide an intuitive visual interpretation, it is difficult to obtain a 3D structure of a feature and is challenging to scale to large datasets. To mitigate these problems, two similarity-based methods, KronRLS [1] and SimBoost [2] have been proposed using efficient machine learning methods.

Most machine learning approaches concentrate on binding affinity prediction. Where special attention is given to the compound sequence and the probability of high affinity score based on inhibition constant Ki, dissociation constant K_d, potency IC_{50} and selectivity with a target protein.

Transformer architectures [3] are used to leverage the advances of Natural Language Processing (NLP) and work with the SMILES representations of compounds. Following this approach, the compounds must be created manually or searched in the chemical space. The chemical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

space has been estimated to be in the order of 10^{63} organic compounds of size up to 30 atoms [4], which infers that iterating over the full space in search for a potential hit is unacceptable.

Reinforcement Learning techniques [5] have successfully been applied by splitting compounds into meaningful molecule fragments and adding a molecule fragment at the time to produce a score generated by human defined rules. The model however is not gaining insights from the training data but rather by humanly predefined rules for scoring.

For these reasons we propose the generation of synthetic small and sophisticated molecule structures that optimize the binding affinity to a target (ASYNT-GAN). To achieve this, we leverage on three important achievements in Machine Learning: Attention, Deep Learning on Graphs and Generative Adversarial Networks. Similar to question answer in NLP we generate a molecular architecture based on an existing target that functions as context. By exploring the latent space, created by the model, we propose a novel way of searching for candidate compounds suitable for binding.

2. Experimental Section

2.1. Methods

2.1.1. Method overview

We first learn the transformation of an input molecule into the latent space using an encoder decoder architecture with attentions. We show that by doing so we can sample from the latent space for generation purposes and find similar structures by their proximity in the latent space. We identify regions of interest and resample points from those regions. The resampled points are then used together with the initial activations to generate the final output by run through a second encoder decoder architecture that may share weights with the first one.

The output are the coordinates of the molecules in 3D space. This approach permits the generation of molecule sequences with specific attributes.

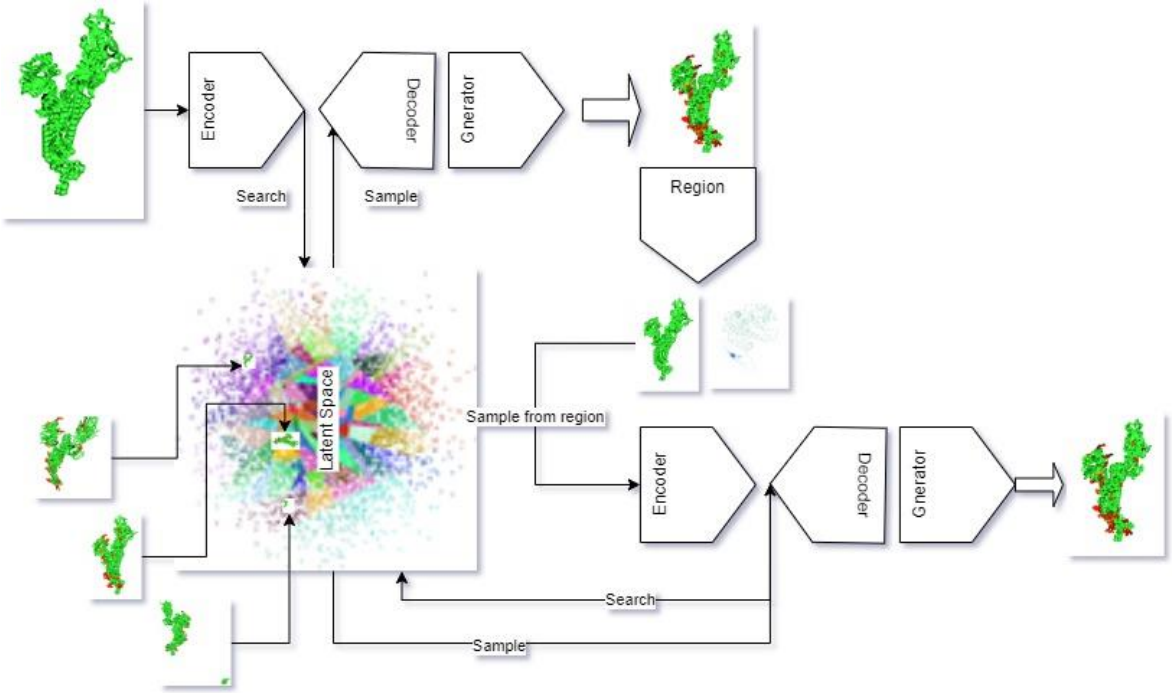


Figure 1: Architecture of the proposed solution. The model consists of an Encoder-Decoder architecture that translates the inputs into the latent space and a Generator that produces the 3D structure of the system. We propose a stacked Generator architecture that takes the first

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

output and calculates regions of interest. We use the regions of interest to re-sample and generate a second output that is concatenated to the first to produce the prediction of the 3D structure of the molecular system.

2.2. Learning a latent embedding

2.2.1. **Data**

Our embedding method is learned from a collection of systems comprised of proteins and ligands, small molecules, used in drug compounds. The systems are split in chains. Per chain we extract the proteins, ligands and their respective binding bonds as point clouds. During training we use as input the proteins and a sample Gaussian distribution or a limited number of points sampled from the point cloud of the ligand.

The systems are protein structures from the protein data bank RCSB [6]. The Protein Data Bank archives information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data. The ligands that we consider as valid are the ligands that are referenced as Chemical Component with a Drug Bank [7] identifier.

2.2.2. Attention Based Generator

We use U-Net architecture, as shown in **Figure 2**, U-Net [8] decorated with residual blocks and attention gates to encode the point cloud coordinates of the target protein and a PointNet [9] decoder for constructing the ligand structure fitting the binding in 3D space.

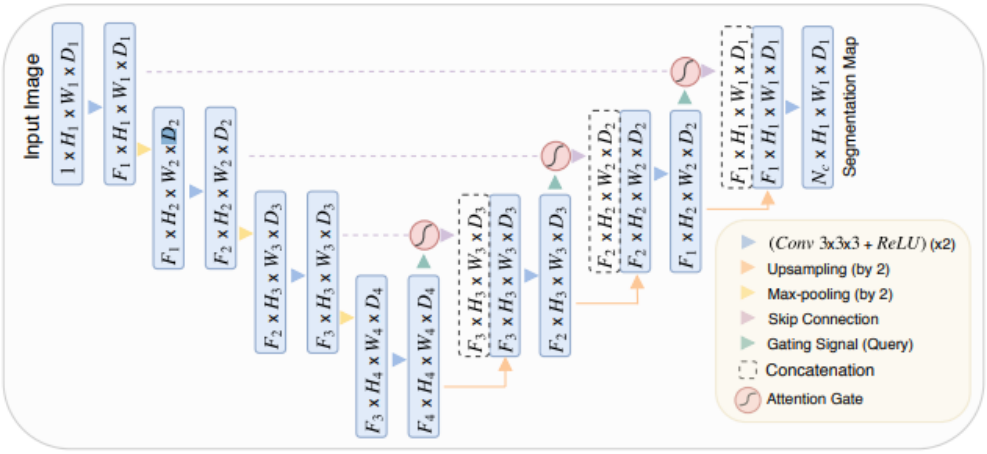


Figure 2: Attention U-Net: The inputs are down sampled to a common shape using Convolutional Layer followed by a Max-Pooling. The attention with the previous layer is calculated using the attention gate and the result is concatenated to the result of the up sampling.

Attention gates are used during the up sampling where the down sampled inputs from the ligand and the protein are concatenated and run through an attention gate. The produced activations are concatenated with previous up sampled activations. The concatenated result is up sampled by the following layer **Figure 3**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

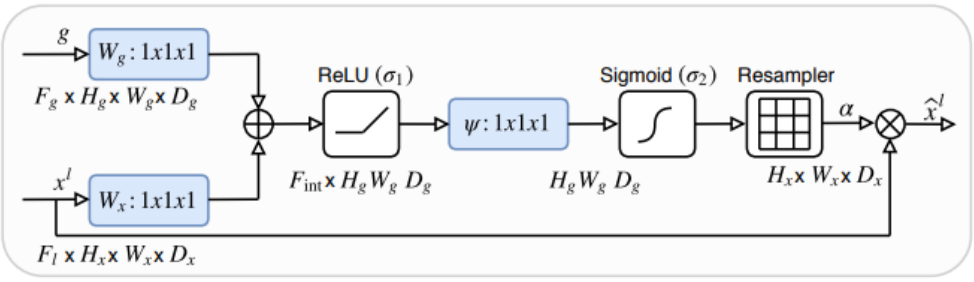


Figure 3: Attention Gate (AG) [8]: The input features (x^l) are scaled by the coefficients (α) computed in AG. Regions are selected by analysing the activations generated by the gating signal (g) which is collected from a coarser scale.

The decoder takes in the latent representation produced from the up sapling in the U-Net and produces the point cloud coordinates for the ligand. The decoder is a Point Net residual network with up sampling capacity in residual blocks **Figure 4**.

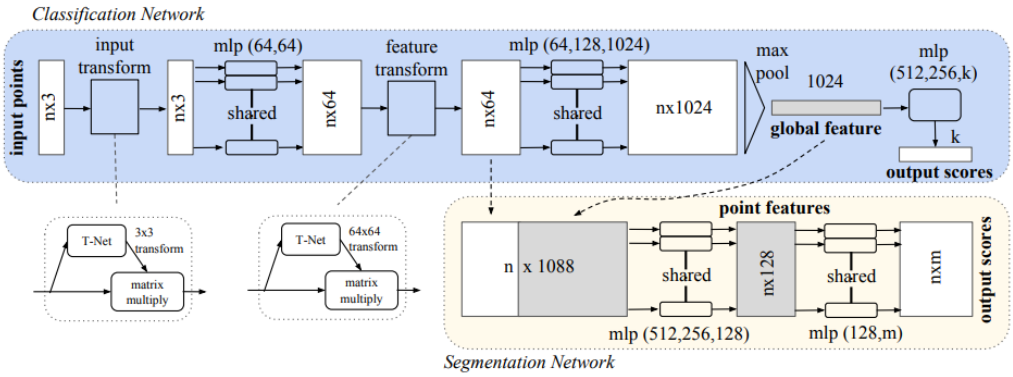


Figure 4: PointNet [9] Architecture takes points in 3d space as input applies feature transformation and applies aggregation with an Max-Pool layer.

During training, we sample coordinates from the target protein and train the network to produce point samples of the ligand that will produce the best binding affinity in 3D space.

The encoder is trained with a 2048 randomly sampled points from the protein and 64 randomly sampled points from a ligand or a 64 point sampled from the boundary with Gaussian-decaying probabilities. This is done with the purpose of simulating use cases where the ligand or part of it is known and provided as input.

The decoder is conditioned to generate the ligands structure based on the target structure as an input. We train the encoder and decoder with the *L2* loss from the Chamfer distance that produces the sum of closest point distances, with an additional latent regularization loss to constrain the latent space of the learned embeddings.

We use a symmetric version of the Chamfer distance [10], calculated as the sum of the average minimum distance from point set A to point set B and vice versa. The average minimum distance from one point set to another is calculated as the average of the distances between the points in the first set and their closest point in the second set, and is thus not symmetrical.

The loss is given as:

$$Ch(X,Y) = \sum_{x \in X} min_{y \in Y} \; ||x - y||_2^2$$

$$d_{CD}(X,Y) = Ch(X,Y) + \; Ch(X,Y)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

$$\mathcal{L}_{end}(\theta_e, \theta_d) = \frac{1}{|P||B|} \sum_{i \epsilon P} \sum_{j \epsilon B} L_c(d_{CD}(D_{\theta_d}\left(x_{i,j}, E_{\theta_e}(g_i)\right), t_{i,j})) + \lambda ||E_{\theta_e}(g_i)||_2$$

where P is the set of all training molecules in each mini-batch, B is the set of point samples sampled per target, $\mathcal{L}_c(.,.)$ is the $\mathcal{L}2$ loss, E_{θ_e} is the encoder parameterized by trainable parameters θ_e, D_{θ_d} is the decoder parameterized by trainable parameters θ_d, and g_i is the sampled point cloud for the *i-th* binding ligand structure.

2.3. Similarity Search

We translate the inputs, protein and ligands, into the latent space. We can use the properties of the encoder to index systems or part of systems and perform a search for similar systems.

The **embeddings** of all the systems are inserted into an index and searched for similarities using Approximate nearest neighbor [11].

An approximate nearest neighbor search algorithm can return points, whose distance from the query is at most *c* times the distance from the query to its nearest points.

The appeal of this approach is that, in many cases, an approximate nearest neighbor is almost as good as the exact one. In particular if the distance measure accurately captures the notion of user quality, then small differences in the distance should not matter.

Figure 2: Approximate nearest neighbor [10] search returns points, whose distance from the query is at most c times the distance from the query to its nearest points.

The search in latent space can be done during training or during inference. During training, if the ligand is partially known, its latent representation can be used to look for candidates instead of sampling from the Gaussian-decaying probabilities.

2.4. Experiments Data Generation and Model Training

2.5. Progressive Training

We introduce the notion of progressive training. During training, we progressively reduce the number of sampled points from the ligands. We start by sampling *1042* points from the point cloud of the point cloud of the ligand and gradually reduce to zero. When reduced to zero we sample from Gaussian-decaying probabilities as an input to the up-sampling part of the attention-based U-Net. We observe an overall stabilization and faster convergence of the generator.

2.6. Stacked Generators

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

We introduce the notion of stacked generators. The points generated from the first generator layer are used as attention regions for the second generator layer. Points from the target protein are sampled from these regions, as shown in **Figure 5**, and used as input for the next Generator Layer.

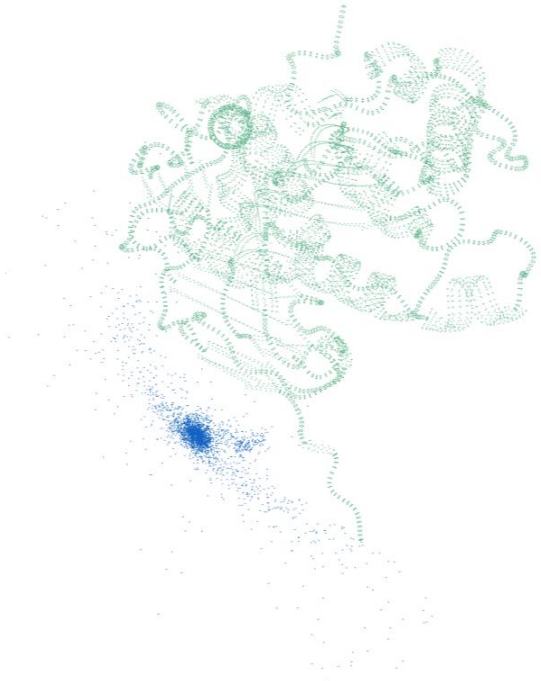


Figure 5: Show the regions of interest identified to resample data from those regions. The data is concatenated to the activations from the first generator and used as input by the stacked generator

We trained the generator layers with shared weights and separately. In both cases, we noticed a significant increase in accuracy in the second generative layer as well as a stabilization of the overall loss during the training as shown in **Figure 6** and **Figure 7**.

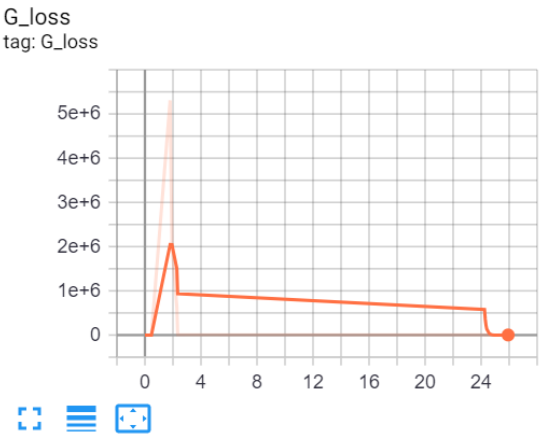


Figure 6: Depicts the Loss evolution and stabilization using the stacked progressive generator approach.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

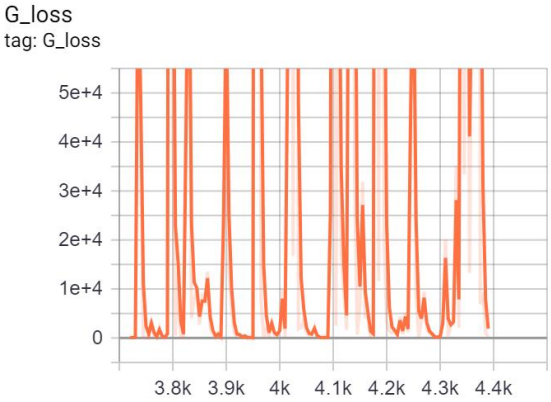


Figure 7: Depicts the loss evolution without using the stacked generator network approach.

2.7. Interpolation

We experimented with an interpolation approach that takes in the attention grids from the Attention U-Net and interpolates the input points. The interpolated points are fed into Residual Network [12] consistent of PointNet [9] Dense Layers. This approach has shown promising results in converging fast in coordinates of ligands.

This approach would be a good fit when systems are split into meaningful sub systems and generation is done in particular 3D sub spaces. As shown in **Figure 8** the interpolation produces less noise and in contrast to **Figure 9**. However not enough points are generated.

Figure 8: Generated using interpolation approach produces points with less noise.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

Figure 9: Generated using the stacked generator approach is noisier but produces more details

3. Results

3.1. Metrics

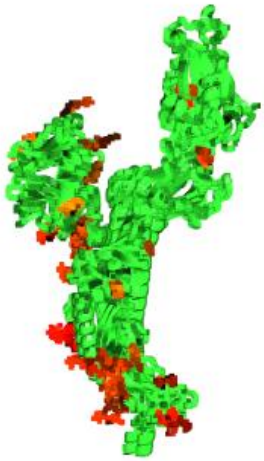
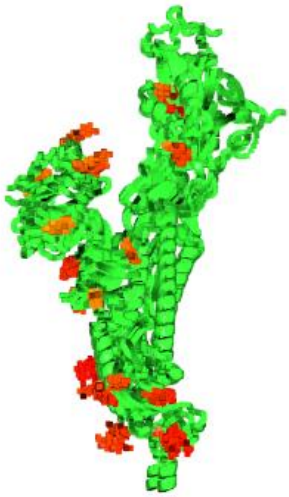
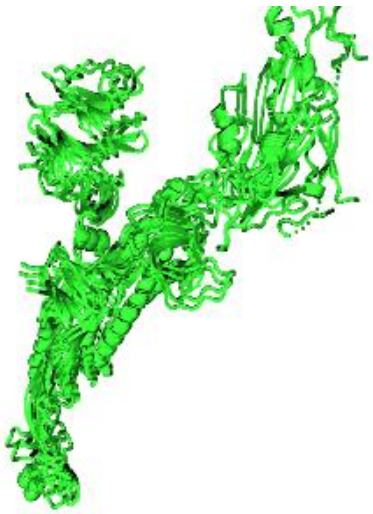
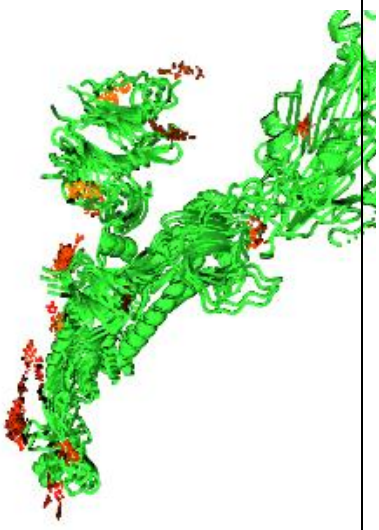
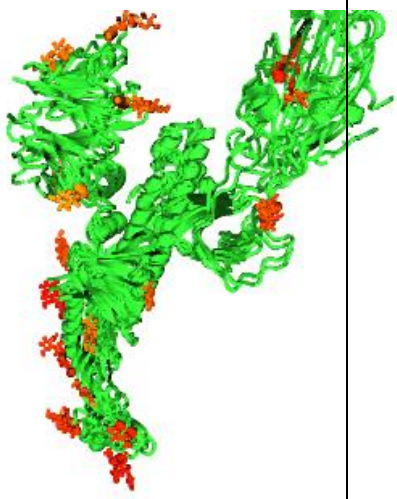
For our experiments, we evaluate the generative quality with Chamfer Distance [10] (CD). We estimate CD using *1024* randomly sampled points on the ground truth and generated systems. We have tested the Chamfer Distance [10] on a series of viral Proteins of the Severe acute respiratory syndrome coronavirus 2.

3.2. Results Discussion

Protein	Chain	Chamfer Distance
6VYB	B	49.71
6VYB	A	90.33
6VW1	A	5.95
6VW1	B	16.87
6WPT	A	78.67
6WPT	B	29.55
6WPT	C	58.04
6WPT	E	48.25
6VXX	A	69.02
6VXX	B	107.20

Table 1: Quantitative Representation of the metric applied to the results produced.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

Input	Prediction	Ground-Truth
		
		

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2020 doi:10.20944/preprints202010.0196.v1

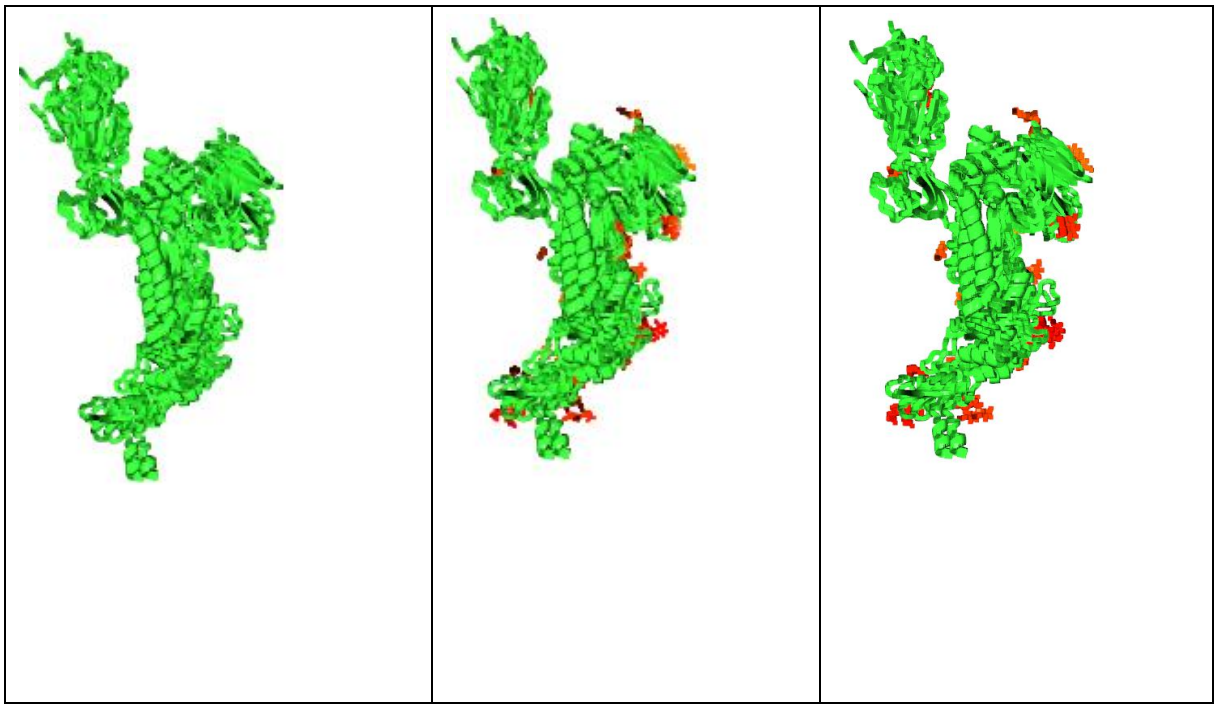


Table 2: Visual Representation of the produced results.

We quantitatively and qualitatively compare performances in **Table 1** and **Table 2**, respectively. Given our solution is trained to learn a latent representation of ligands; the learned representation does generalize to systems and chains beyond the source system. Visually, as shown in **Table 2** our solution achieves good generation of complete structure that optimizes the binding molecules in the system (e.g., ligand and protein), but performs poorly in terms of generating point clouds without noise as shown in **Figure 10**.

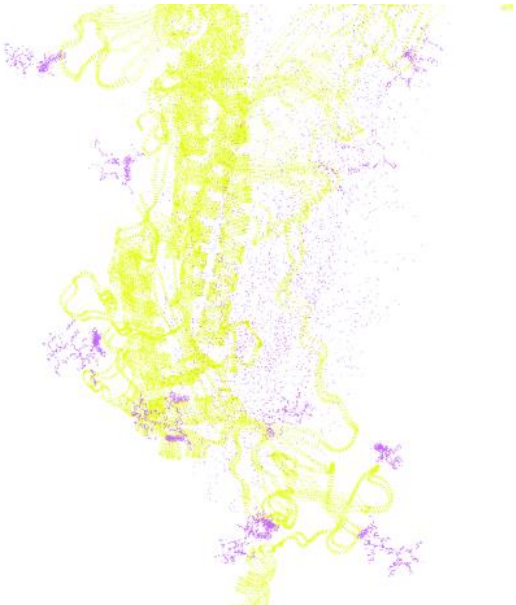


Figure 10: Depicts the amount of noise in Generated structure

4. Discussion

The generation of synthetic small and more sophisticated molecule structures that optimize the binding affinity to a target (ASYNT-GAN) through encoding a protein and generating a system

comprised of a ligand and a protein. Experiments show that ASYNT-GAN is able to generate ligand structures for proteins unseen during training. Translating the input sub-systems into the latent space permits the reachability for similar structures and the sampling from the latent space for generation. Topics for future work include ways of integrating the search capabilities in the training process, explore alternatives for sampling and generating points ASYNT-GAN from regions of interest, provide for ability to generate alternative variants of proteins to predict mutations.

Author Contributions: Conceptualization, I.J.; methodology, I.J.; software, I.J.; validation, M.M.; formal analysis, I.J. and M.M.; investigation, I.J. ; resources, I.J.; data curation, I.J.; writing—original draft preparation, I.J.; writing—review and editing, I.J. and M.M; visualization, I.J.; supervision, M.M.; project administration, I.J.; . All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, Sep. 2018, doi: 10.1093/bioinformatics/bty593.

[2] "(PDF) SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines," *ResearchGate*. https://www.researchgate.net/publication/316235177_SimBoost_a_read-across_approach_for_predicting_drug-target_binding_affinities_using_gradient_boosting_machines (accessed Sep. 27, 2020).

[3] B. Shin, S. Park, K. Kang, and J. C. Ho, "Self-Attention Based Molecule Representation for Predicting Drug-Target Interaction," *ArXiv190806760 Cs Stat*, Aug. 2019, Accessed: Sep. 27, 2020. [Online]. Available: http://arxiv.org/abs/1908.06760.

[4] "Organic Chemistry - Paperback - Jonathan Clayden, Nick Greeves, Stuart Warren - Oxford University Press." https://global.oup.com/ukhe/product/organic-chemistry-9780199270293?cc=lu&lang=en& (accessed Sep. 27, 2020).

[5] 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, Mar. 2020, doi: 10.1101/2020.03.03.972133.

[6] R. P. D. Bank, "RCSB PDB: Homepage." https://www.rcsb.org/ (accessed Sep. 27, 2020).

[7] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. Database issue, pp. D1074–D1082, Jan. 2018, doi: 10.1093/nar/gkx1037.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *ArXiv150504597 Cs*, May 2015, Accessed: Sep. 27, 2020. [Online]. Available: http://arxiv.org/abs/1505.04597.

[9] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," *ArXiv161200593 Cs*, Apr. 2017, Accessed: Sep. 27, 2020. [Online]. Available: http://arxiv.org/abs/1612.00593.

[10] A. Hajdu, L. Hajdu, and R. Tijdeman, "Approximations of the Euclidean distance by chamfer distances," *ArXiv12010876 Cs Math*, Jan. 2012, Accessed: Sep. 27, 2020. [Online]. Available: http://arxiv.org/abs/1201.0876.

 doi:10.20944/preprints202010.0196.v1

[11] A. Andoni, P. Indyk, and I. Razenshteyn, "Approximate Nearest Neighbor Search in High Dimensions," *ArXiv180609823 Cs Stat*, Jun. 2018, Accessed: Sep. 27, 2020. [Online]. Available: http://arxiv.org/abs/1806.09823.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *ArXiv151203385 Cs*, Dec. 2015, Accessed: Sep. 27, 2020. [Online]. Available: http://arxiv.org/abs/1512.03385.