Proceedings

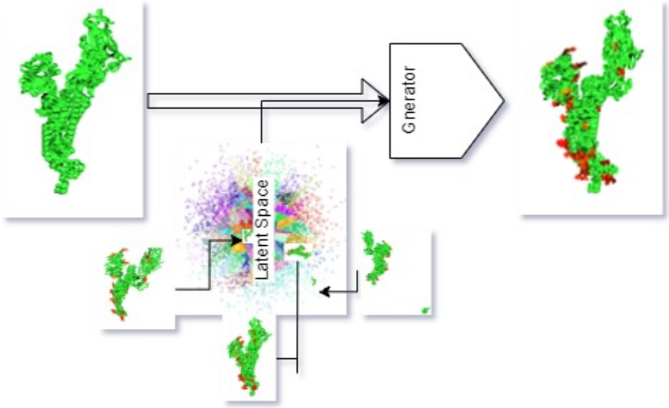
De Novo Drug Design using Artificial Intelligence ASYNT-GAN

Ivan Jacobs [1] and Manolis Maragoudakis [2]

[1] AI4U; ivan.jacobs@ai4u.ai

[2] Ionian University; mmarag@ionio.gr

[†] Presented at the 1st International Electronic Conference on Biomolecules: Natural and Bio-Inspired Therapeutics for Human Diseases, 1–13 December 2020; Available online: https://iecbm2020.sciforum.net/.



We learn and embedding of proteins and ligands in the latent space. By sampling from the latent space additional to a target system as an input we generate a system consisting of protein, ligands and bonds in the 3D space optimized for binding. We show that we are able to generalize to a variety of systems and their respective chains.

1. Introduction

An outbreak of the novel coronavirus SARS-CoV-2, the causative agent of COVID-19 respiratory disease, has infected over 10 million people since the end of 2019, killed over 500,000, and caused worldwide social and economic disruption. There are currently no antiviral drugs with proven efficacy nor are there vaccines for its prevention. Unfortunately, the scientific community has little knowledge of the molecular details of SARS-CoV-2 infection.

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. The synthesis of structurally intricate products may be challenging, rendering structure-activity relationship studies elaborate tasks. The 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.

Nevertheless, the current computational de novo design methods for generating natural-product-inspired small molecules suffer from several limitations, in particular unsatisfactory scoring of biological activities. Artificial Intelligence combined with computational power can help to optimize and shorten the process.

Throughout the process of building a drug, the main challenge is to identify a molecular structure that is able to attach itself to a target protein. The quality of the binding has a direct influence on side effects and effectiveness of the treatment.

Multiple approaches are taken to find or create these structures. Depending on the use case, a search for approved drugs with a similar structure can be executed or a new drug can be created.

All drugs available for therapy may be classified into two major classes, i.e., protein therapeutics and small molecule drugs. Despite the fact that the latter is currently the more predominant therapeutic agent in use, the impact of protein therapeutics is increasing, mainly due to the advances in recombinant DNA technology. Moreover, proteins exhibit high specificity, less immunogenicity and are widely used as therapeutic agents in the treatment of various disorders and diseases. Small drugs have limited surface area available to contact a target protein and, in most cases, need the presence of a deep hydrophobic pocket in the target protein for having a favourable interaction, thus limiting the number of potential druggable targets.

In contrast, protein drugs are large and do not have this limitation, thus making them important tools in human disease therapy, and treatment or management of some crippling diseases. Moreover, the higher binding selectivity and specificity of protein therapeutics aids in targeting specific steps in disease pathology, thus revising the treatment paradigm of certain diseases and subsequent supplementation or replacement of small molecule drug therapies.

A vaccine trains the body's immune system to recognize some signature viral protein called an antigen. Viruses do not make their own components. Instead, a virus enters into the cells by attaching through to them. Once inside, the viral RNA becomes part of the host cell's protein production machinery, and produces new copies of viral proteins and RNA which then assemble into thousands of new viruses to spread the disease. One way to stop a disease is to block the virus from entering the cells. Vaccines do that by training the body to identify and attack the virus before it can infect healthy human cells.

A vaccine is essentially a pure preparation of one or more key components of the virus – such as the envelope, spike or a membrane protein – that is injected in the body to give the immune system a preview of the virus without causing disease. This preview tells the immune system to seek out and attack the virus containing those specific proteins if the real virus ever shows up.

Protein-based vaccines require mass production of viral proteins in facilities, which can guarantee their purity. Growing the viruses and purifying the proteins at medically acceptable pharmaceutical scales can take years. In fact, for some of the recent epidemics, such as AIDS, Zika and Ebola, to date there are no effective vaccines.

Antibodies are the immune system's warriors. Their role is to pinpoint disease pathogens, attaching to them and neutralizing their effects. Though antibodies are of great value for biomedical research, the process of creating them has been time-consuming and tedious.

Since the last few decades, proteins have emerged as the major class of pharmaceuticals with more than 200 protein-based products currently available in the market, of which 90% are used as therapeutics. The protein engineering market is bolstered by the need for drugs with improved efficiency, specificity, technological capabilities, rise of antibody based drugs and steady growth in the therapeutic market. Monoclonal antibodies (mAbs) is the fastest growing segment in the therapeutic market, though the other segments comprising non-mAb recombinant proteins like Insulin, Erythropoetin (EPO), Interferons (INF), Interleukins (ILs) and Somatotropin (hGH) are also in great demand for therapy.

In this paper we propose the generation of synthetic small and more sophisticated molecule structures that optimize the binding affinity to a target (ASYNT-GAN). To achieve this we leverage on three important achievements in A.I.: Attention, Deep Learning on Graphs and Generative Adversarial Networks. Similar to text generation based on parts of text we are able to generate a molecule architecture based on an existing target.

By adopting this approach, we propose a novel way of searching for existing compounds that are suitable candidates. Similar to question and answer Natural Language solutions we are able to find drugs with highest relevance to a target. We are able to identify substructures of the molecular structure that are the most suitable for binding.

In addition, we are proposing a novel way of generating the molecule in 3D space in such a way that the binding is optimized. We show that we are able to generate compound structures and protein structures that are optimised for binding to a target.

2. Related Work

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 (B. Shin§†, 2019) models. Molecular docking (Olson, 2010) (al. L. e., 2016) is a simulation-based method using the 3D structured features of molecules and proteins. Although it can provide an intuitive visual interpretation, it is difficult to obtain a 3D structure of a feature and cannot scale to large datasets. To mitigate these problems, two similarity-based methods, KronRLS (al. P. e., 2014) and SimBoost (al. H. e., 2017) have been proposed using efficient machine learning methods.

Most machine learning methods concentrate on binding affinity prediction. Where special attention is given to the compound sequence and the probability of high affinity score based on Ki (inhibition constant), Kd (dissociation constant) and IC50 i.e., potency and selectivity with a target protein.

The authors of the (Bonggun Shin, 2019) successfully use transformer architectures to leverage the advances of NLP to work with the SMILES representations of compounds. They adopt the strategy pre-train the model on a large corpus as an effective way of learning a chemical structure. The authors unfortunately did not take the same approach for the protein sequences and trained a second model using Convolutional Neural Networks to analyse the protein sequence. The activations of both models are concatenated and used as an input for fully connected layers. The model is then fine-tuned for the task of producing an affinity score.

This approach still means that the compounds need to be manually created or a preselected subset of compounds should be evaluated against possible targets. The chemical space is estimated to be in the order of 10^{60} molecules, which means that pre-processing work needs to be done.

It also means that the "dynamic" between compound and target protein is not calculated using the self-attention mechanism of transformers. A number of tactics could have been adopted during the pre-training and fine-tuning to identify and highlight the regions of ligand binding in the purpose to optimize prediction accuracy.

Another interesting approach by the authors of (al. B. T., 2020) using Reinforcement Learning where the compounds are split into meaningful molecule fragments and adding a molecule fragment at the time.

Using Reinforcement Learning, they give a reward at every step based on the validity and completeness of the chemical compound at the end of each state. A reward at the end state is given based on particular criteria that can be edited per use case. The down side of this approach is that the model is not gaining knowledge from the data it is working on but more from the humanly predefined rules for scoring.

3. Methods

3.1. 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 are able to sample from the latent space for generation purposes and find similar structures by their proximity in the latent space.

Additionally we show that by providing an additional class label as input we are able to condition the generation. The first output are the coordinates of the molecules in 3D space. The second output is a probability of the output belonging to a class or in our case the SMILES string. The simplified molecular-input line-entry system (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings. SMILES strings can be imported by most molecule editors for conversion back into two-dimensional drawings or three-dimensional models of the molecules. This approach permits the generation of molecule sequences with specific attributes.

4. Learning a Latent Embedding

4.1. 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. 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 DrugBank ID. The following figure depicts the architecture of the process that was followed.

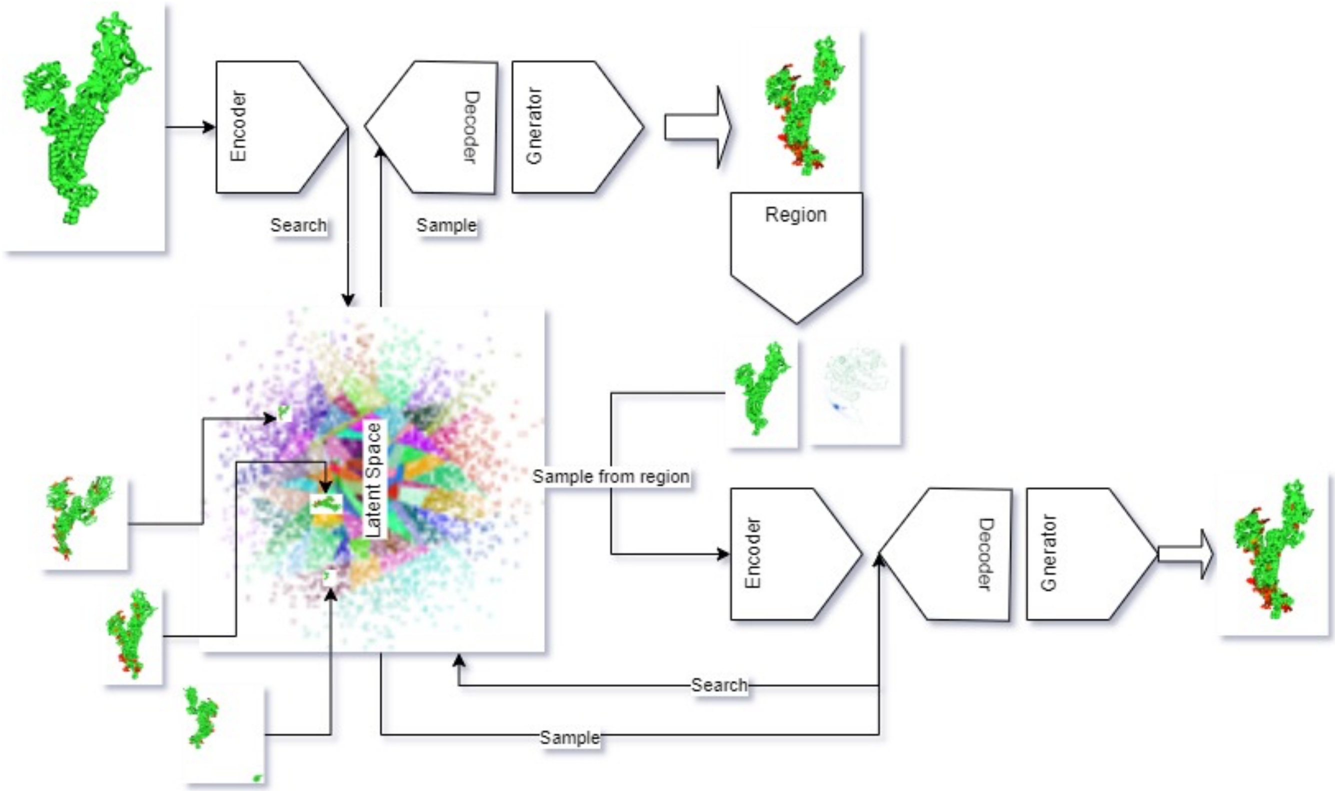


Figure 1. Architecture of the proposed solution.

4.2. Attention Based Generator

We use U-Net architecture, as shown in Figure 2, (Olaf Ronneberger, 2015) decorated with residual blocks and attention gates to encode the point cloud coordinates of the target protein and a PointNet (Charles R. Qi, 2016) decoder for constructing the ligand structure fitting the binding in 3D space.

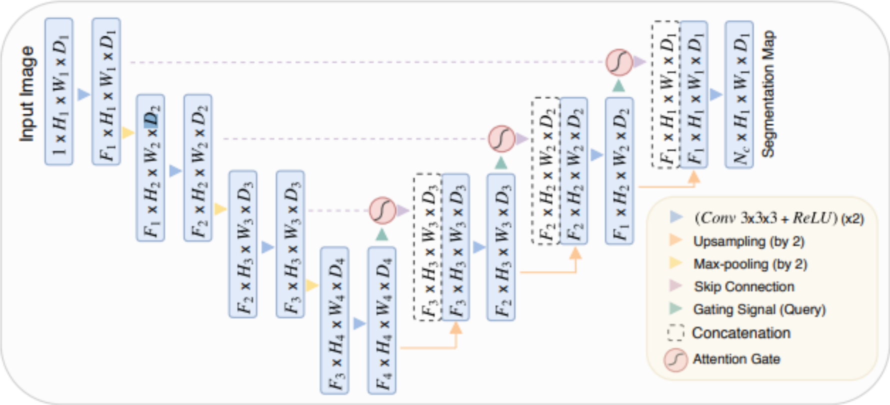

Input Image
$1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_4 \times W_4 \times D_4$
$F_4 \times H_4 \times W_4 \times D_4$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_2 \times H_3 \times W_3 \times D_3$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_1 \times H_2 \times W_2 \times D_2$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$N_c \times H_1 \times W_1 \times D_1$
Segmentation Map
(Conv 3x3x3 + ReLU) (x2)
Upsampling (by 2)
Max-pooling (by 2)
Skip Connection
Gating Signal (Query)
Concatenation
Attention Gate

Figure 2. Attention U-Net.

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)

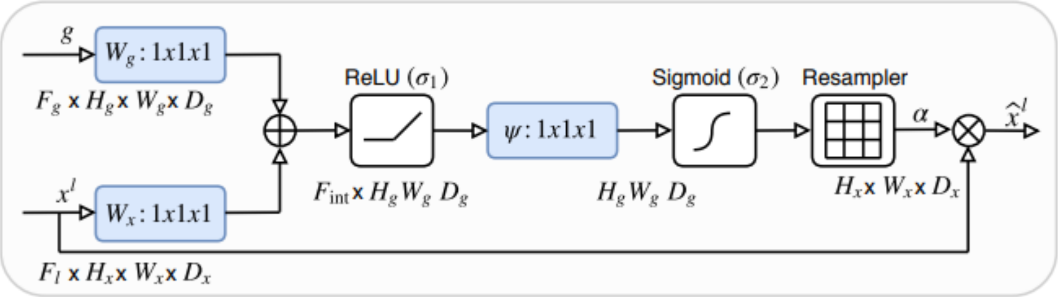

g
$W_g: 1x1x1$
$F_g \times H_g \times W_g \times D_g$
x^l
$W_x: 1x1x1$
$F_l \times H_x \times W_x \times D_x$
ReLU (σ_1)
$F_{\text{int}} \times H_g W_g D_g$
$\psi: 1x1x1$
Sigmoid (σ_2)
$H_g W_g D_g$
Resampler
α
$H_x \times W_x \times D_x$
$\hat{x}^l$

Figure 3. Attention Gate.

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).

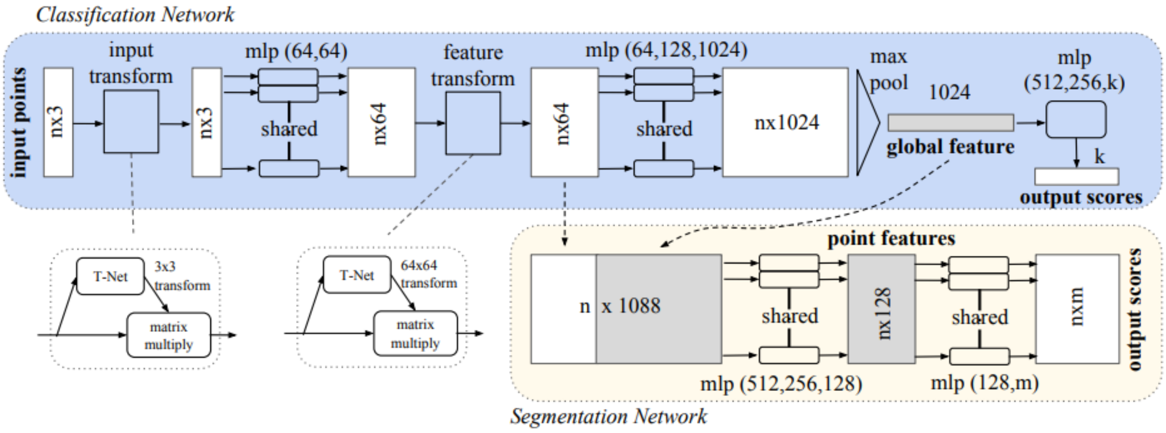


Figure 4. PointNet Architecture.

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, 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:

$$Ch(X,Y) = \sum_{x \in X} min_{y \in Y}\ ||x - y||_2^2 \tag{1}$$

$$d_{CD}(X,Y) = Ch(X,Y) +\ Ch(X,Y) \tag{2}$$

$$\mathcal{L}_{end}(\theta_e, \theta_d) = \frac{1}{|P||B|}\sum_{i\epsilon P}\sum_{j\epsilon B} L_c(d_{CD}(D_{\theta_d}\left(x_{i,j}, E_{\theta_e}(g_i)\right), t_{i,j})) + \lambda\, ||E_{\theta_e}(g_i)||2 \tag{3}$$

where *P* is the set of all training molecules in a given 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, $D_{\theta_d}$is the decoder parameterized by trainable parameters θ_d, and g_i is the sampled point cloud for the *i-th* binding ligand structure.

4.3. Merging and Refining

With the encoder-decoder, we can generate a smooth point cloud predicting the overall shape. However, the encoder-decoder may neglect some structures. We merge the output from the decoder with the second input and then learn a point-wise residual for the combination. Since the density of the two point clouds may be different and there may be overlapping between them, the merged point cloud is probably unevenly distributed. Existing sampling algorithms for point clouds, such as the farthest point sampling (FPS) and Poisson disk sampling (PDS) (Wei 2008), cannot guarantee the global density distribution of the results.

We use minimum density sampling (MDS) introduced by [ref. We denote the i-th sampled point as pi and the set of first i- sampled points as

$$\mathrm{Pi}\ =\ \{p_j\ |1\ \leq\ \mathrm{j}\ \leq\ \mathrm{i}\} \tag{4}$$

Unlike FPS returning the farthest point from P_{i-1} as as p_i, in each iteration, MDS returns a point that has the minimum "density":

$$p_i = \mathrm{argmin}_{\ \mathrm{x} \notin \mathrm{P_{i-1}}} \sum_{\mathrm{p_j \epsilon P_{i-1}}} \exp(-||\mathrm{x} - \mathrm{p_j}\,||^2/(2\sigma^{\,2})) \tag{5}$$

Taking the evenly distributed subset point cloud as input, we then learn a point-wise residual for refinement, which enables the generation of fine-grained structures. The architecture of the residual network resembles PointNet (Qi et al. 2017a), which consumes a point cloud and outputs a three-channel residual. We output the final point cloud after adding the residual point by point.

Our joint loss function $\mathcal{L}$ thus can be calculated as:

$$\mathcal{L} = \mathcal{L}_{end}(S_{endout}, S_{gt}) + \alpha\mathcal{L}_{end}(S_{final}, S_{gt}) \tag{6}$$

Where S_{endout} denotes the output from the decoder, S_{final}, denotes the final output, and S_{gt} denotes the ground truth. α is a weighting factor.

5. 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.

An approximate nearest neighbor search algorithm is allowed to 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.

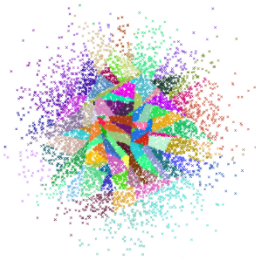


Figure 4. Approximate nearest neighbour.

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.

6. Experiments Data Generation and Model Training

6.1. 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.

6.2. Stacked Generators

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.

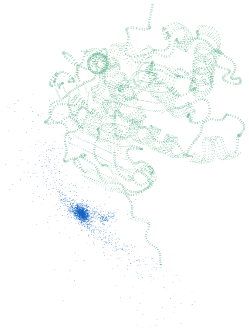


Figure 5. Regions of interest.

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 stabilisation of the overall loss during the training as shown in Figure 6 and Figure 7.

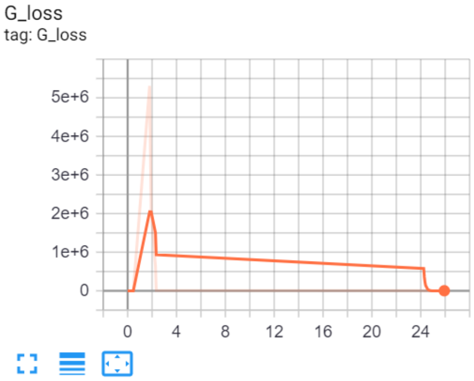


Figure 6. Loss with stacked progressive generator.

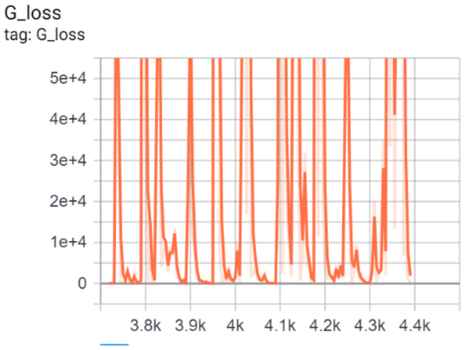


Figure 7. Loss with normal training.

6.3. 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 (Kaiming He, 2015) consistent of PointNet 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.

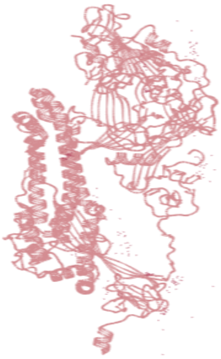


Figure 8. Interpolation.

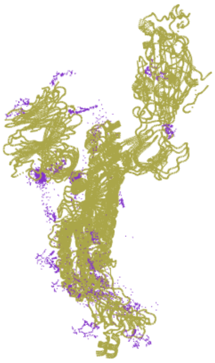


Figure 9. Stacked Generator.

6.4. Metrics

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

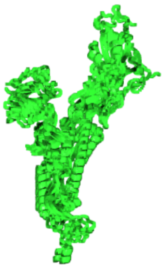
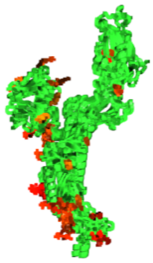
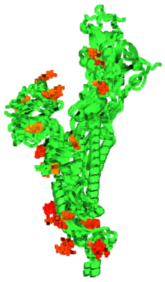
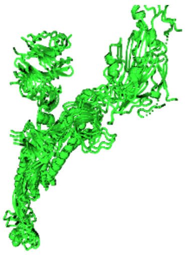
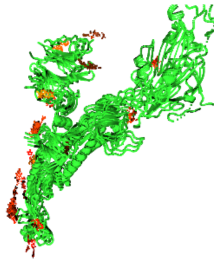
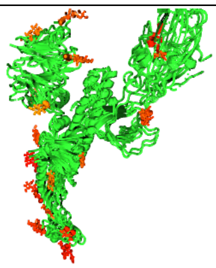
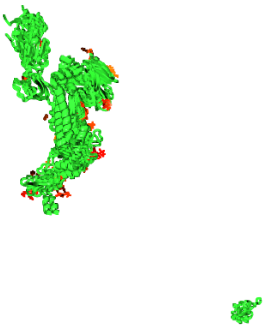
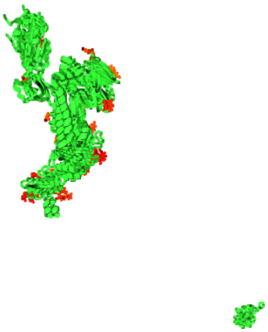
6.5. Results Discussion

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 generalizes 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.

Table 1. Quantitative Representation.

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 2. Visual Representation.

Input	Prediction	Ground-Truth
		
		
		

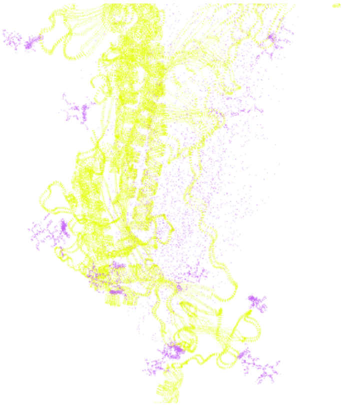


Figure 10. Noise in Generated structure

7. Discussion and Future Work

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.

References

1. al., B. T. (2020). AI-aided design of novel targeted covalent inhibitors against. bioRxiv.
2. al., H. e. (2017).
3. al., L. e. (2016).
4. al., P. e. (2014).
5. B. Shin§†, S. P. (2019). arXiv.
6. Bonggun Shin, S. P. (2019). Self-Attention Based Molecule Representation for Predicting Drug-Target Interaction.
7. Charles R. Qi, H. S. (2016). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. arXiv.
8. Kaiming He, X. Z. (2015). Deep Residual Learning for Image Recognition. arXiv.
9. Olaf Ronneberger, P. F. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation.
10. Olson, T. a. (2010).