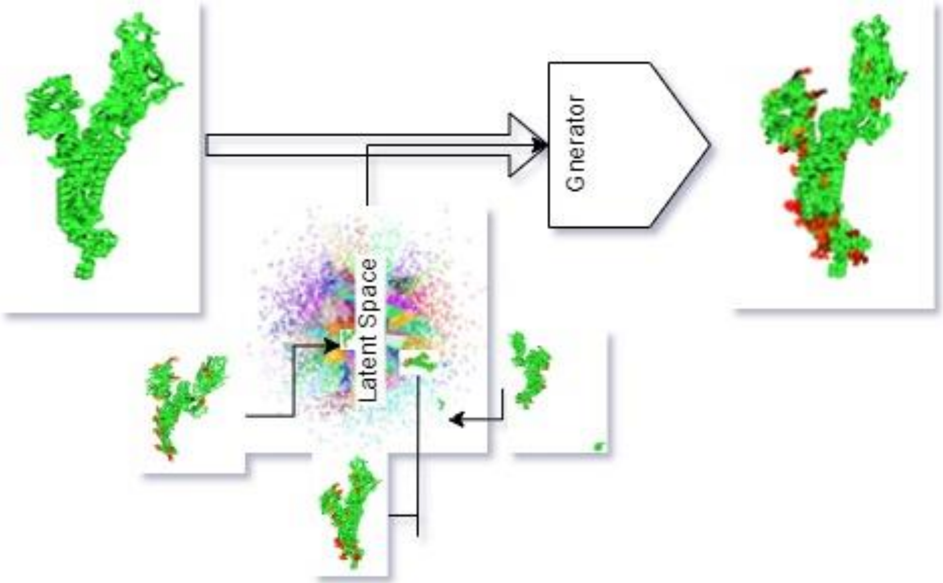
De Novo Drug Design using A.I.

Ivan Jacobs[1], Manolis Maragoudakis[2]
[1]AI4U, [2]Ionian University
ivan.jacobs@ai4u.ai, mmarag@ionio.gr



ASYNT-GAN

SARS-CoV-2

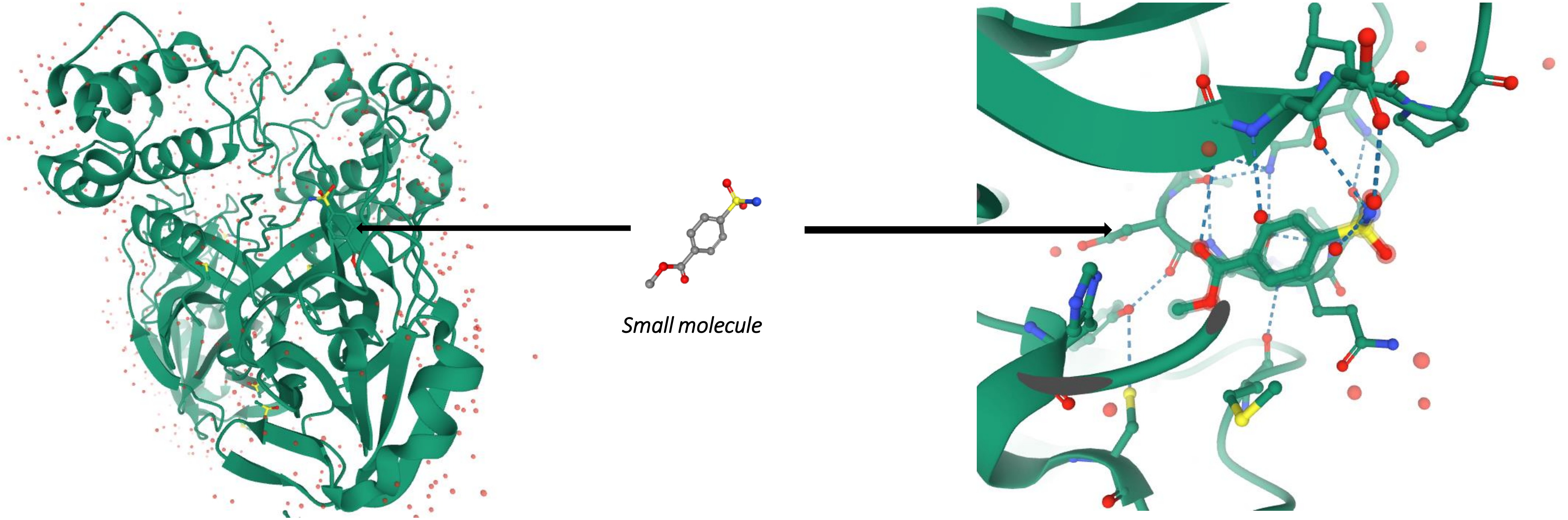
An outbreak of the novel coronavirus SARS-CoV-2 has infected millions of people, killed over half million and caused worldwide social and economic disruption.

There are currently no antiviral drugs with clinically proven efficiency nor are there vaccines for its prevention.

The time and effort to create and market a drug or vaccine can span over decades and millions of investment. It is estimated that the drug discovery and development process takes around 10-14 years and more than 1 billion dollars capital in total [1]

[1] Daina A., Blatter M.C., Baillie Gerritsen V., Palagi P.M., Marek D., Xenarios I. Drug Design Workshop: A Web-Based Educational Tool To Introduce Computer-Aided Drug Design to the General Public. J. Chem. Educ. 2017;94(3):335–344. doi: 10.1021/acs.jchemed.6b00596

Why is it hard



Small molecule

Structure: Protein + small molecules ligands

Ligand Interactions

Find or build a molecular structure that is able to attach itself to a target protein
It needs to fit within the 3D structure of the target
It needs to produce the correct chemical reactions with the target
The quality of the binding impacts side effects and effectiveness of the treatment

Images RCSB PDB. MOLID:5R80, Mol* (D. Sehnal, A.S. Rose, J. Kovca, S.K. Burley, S. Velankar (2018) Mol*: Towards a common library and tools for web molecular graphics MolVA/EuroVis Proceedings. doi:10.2312/molva.20181103)

Why is it hard

The computer-assisted de novo design of chemical structures offers a viable strategy to reduce efforts and obtain bioactive small molecules.

The current computational de novo design methods for generating small molecules suffer from several limitations.

Artificial Intelligence combined with computational power can enable the production of chemically correct structures with a planned biological activity.

Proposition ASYNT-GAN

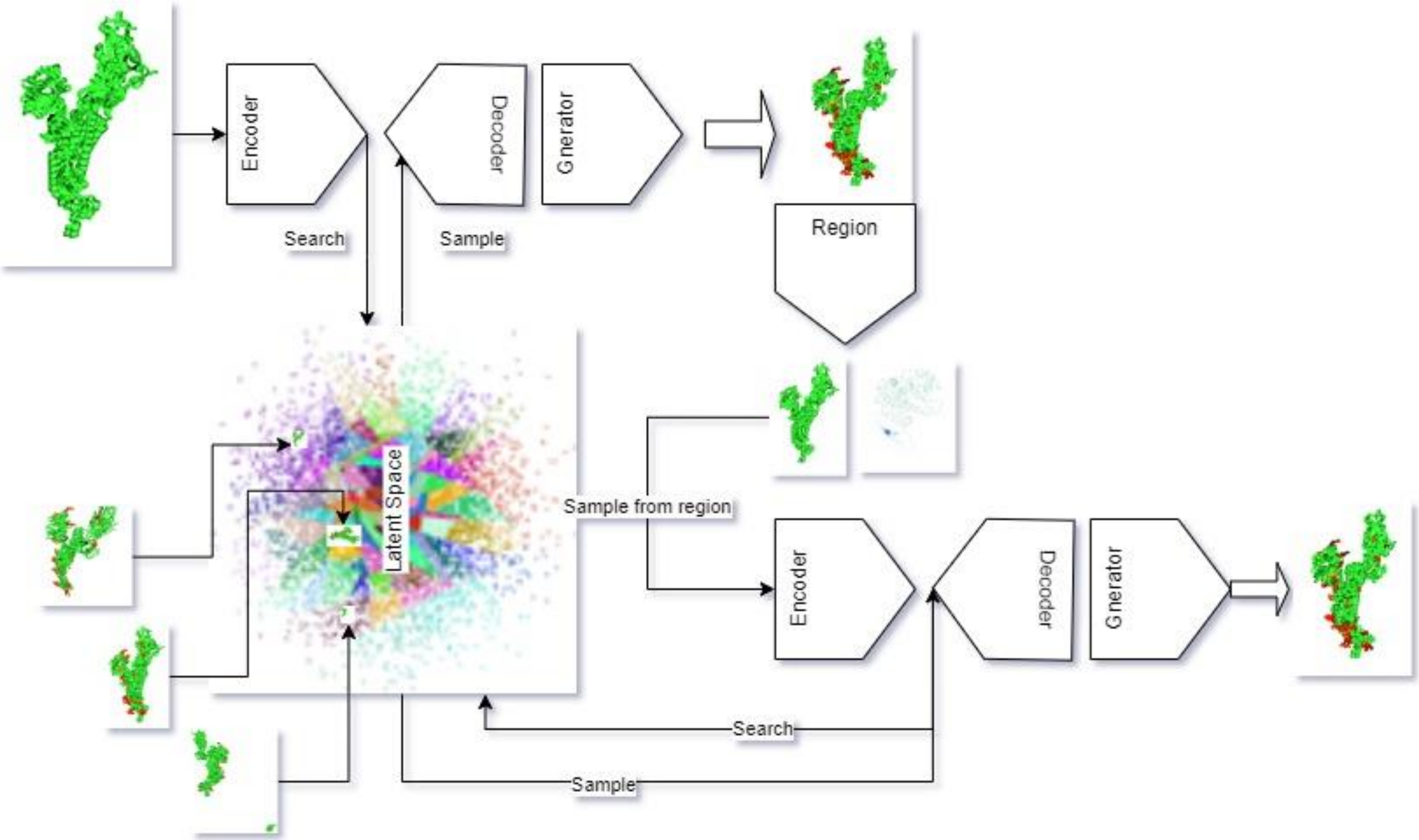
In this paper we propose the generation of synthetic molecule structures that optimize the binding affinity to a target (ASYNT-GAN).

We achieve this by leveraging on important milestones in Deep Learning:

- Attention
- Deep Learning on Graphs
- Generative Adversarial Neural Networks

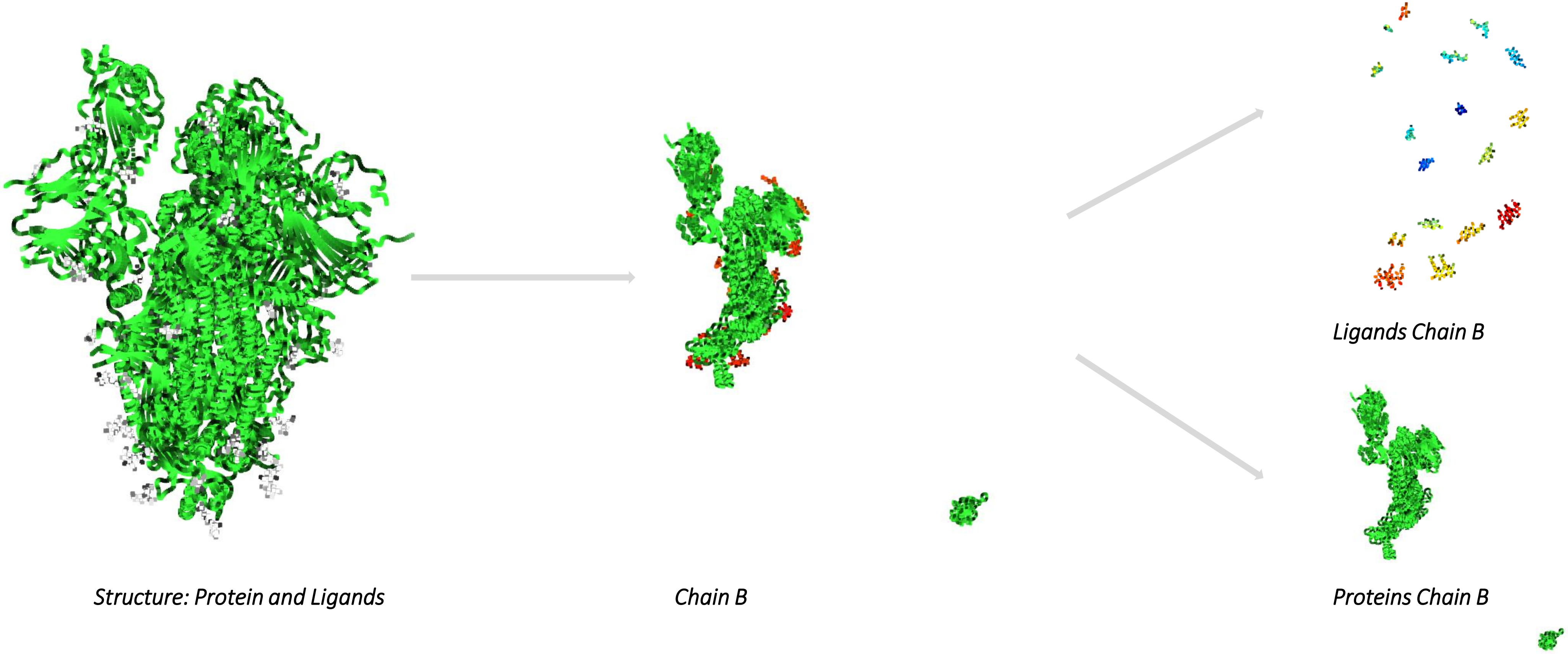
By adopting this approach, we propose a novel way of searching for existing compounds that are suitable candidates. Similar to question and answer in Natural Language Processing (NLP) we are able to find drugs with highest relevance to a target. We are able to identify substructures that are the most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

Method



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 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.

Data



Structure: Protein and Ligands

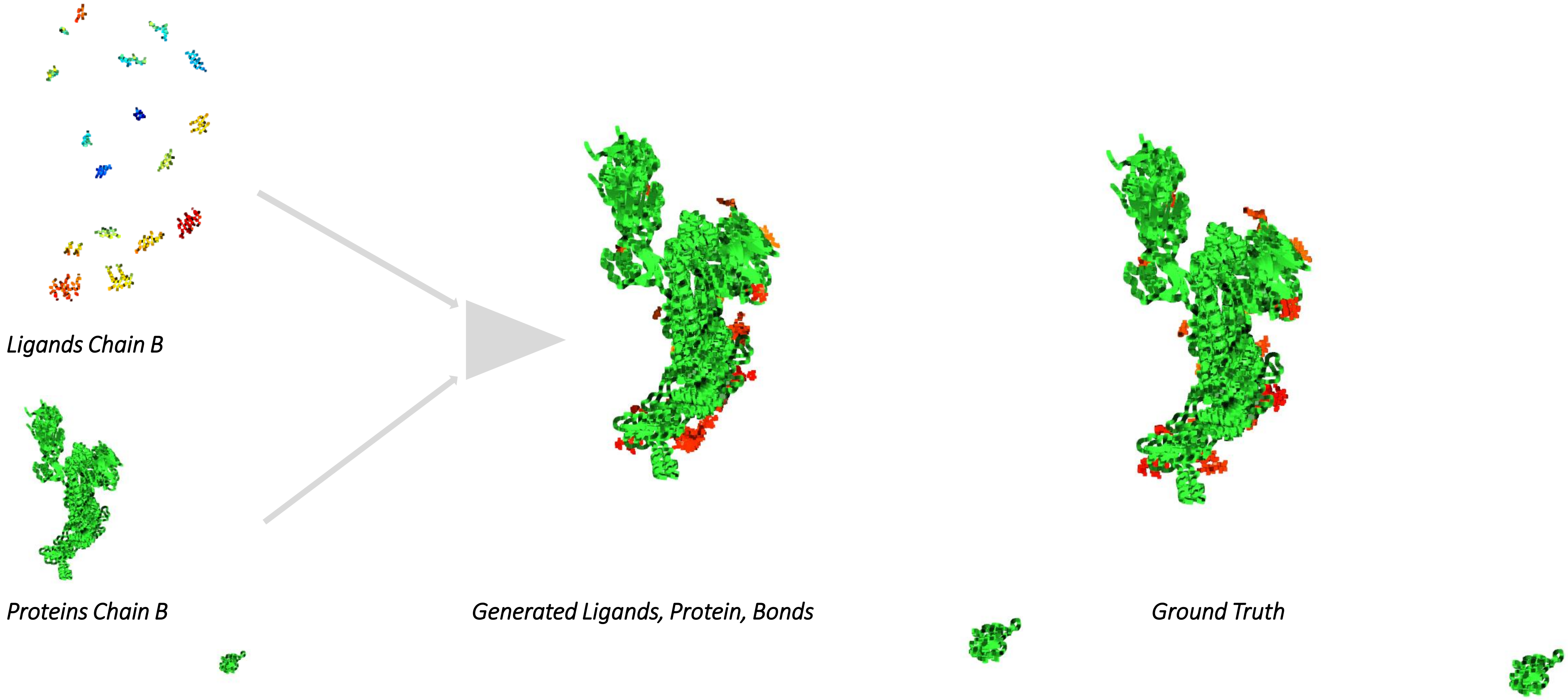
Chain B

Ligands Chain B

Proteins Chain B

We used a series of viral Proteins of the SARS-CoV-2 from data bank RCSB. The systems that we consider as valid contain ligands that are referenced as Chemical Component with a DrugBank ID. Each system is split into chains. Each chain is split into proteins and ligands.

Data



Ligands Chain B

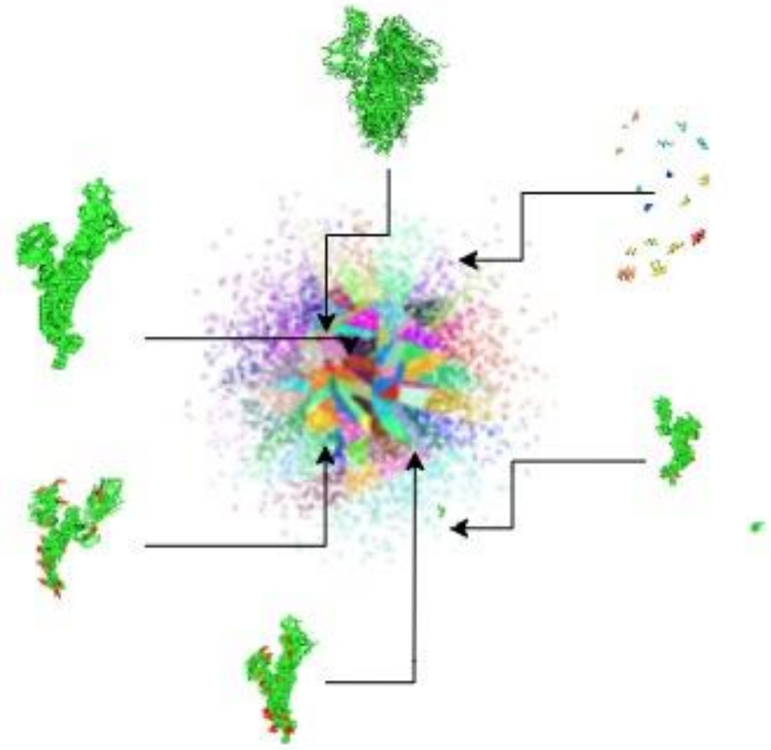
Proteins Chain B

Generated Ligands, Protein, Bonds

Ground Truth

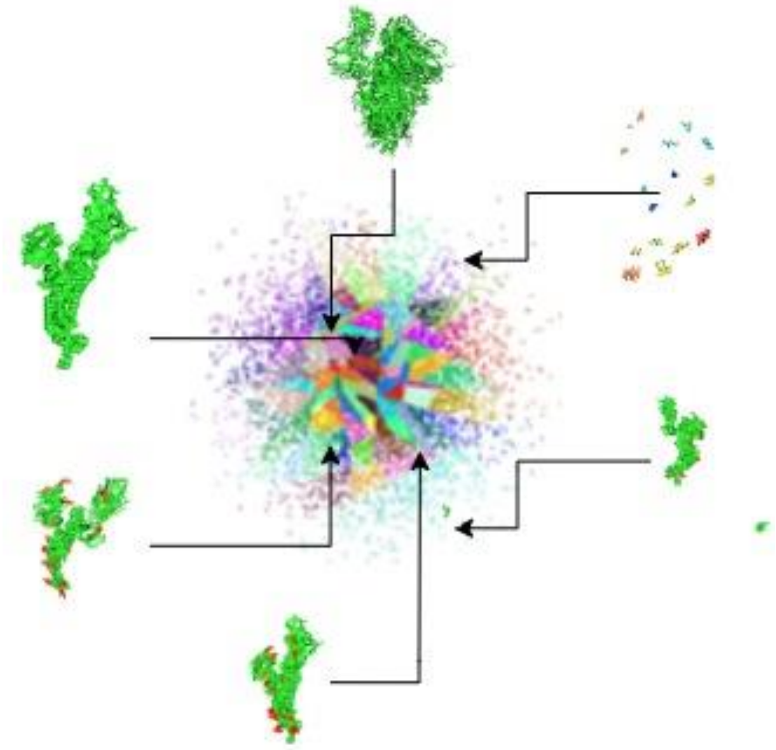
During training we use proteins as input 1 and a sample of the Gaussian distribution as input 2. We have experimented with approaches where Input 2 is a limited number of points sampled from the ligand.
The method generates a full system in 3D space comprising of Ligands, Protein and Ligand Bonds (Ligand Interactions)

Similarity Search



With our method we effectively translate the inputs into the latent space. We can use these properties to index full 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.

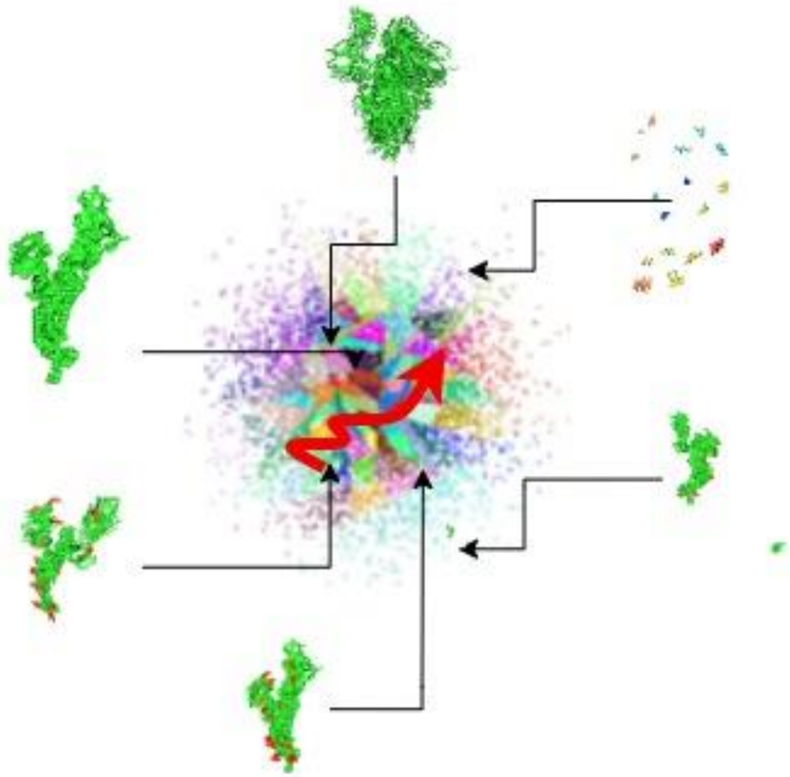
Similarity Search



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.

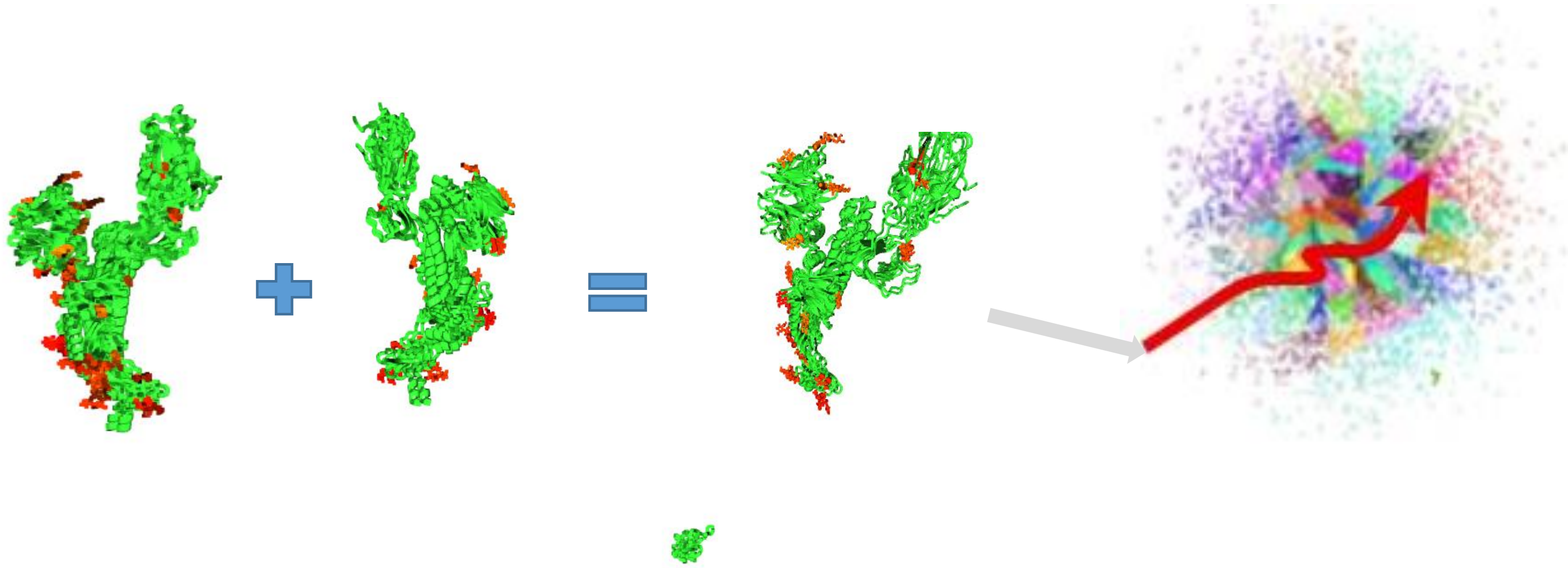
The appeal of this approach is that, in many cases, an approximate nearest neighbor is almost as good as the correct one. In particular, if the distance measure accurately captures the notion of user quality, then small differences in the distance should not matter.

Similarity Search



The latent space has structure that can be explored, such as by interpolating between points and performing vector arithmetic between points. For instance we can use the best match from the approximate nearest neighbor search as starting point for a walk through the latent space.

Similarity Search



We can perform vector arithmetic between points in latent space which have meaningful and targeted effects. The results can similarly be used to initiate a search and walk through the latent space. These characteristics can be used during training, inference and drug repurposing.

Data Generation Metrics

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

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

For our experiments, we evaluate the generative quality with Chamfer Distance (CD).

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.

We estimate CD using 1024 randomly sampled points from the ground truth and the generated systems.

Results

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

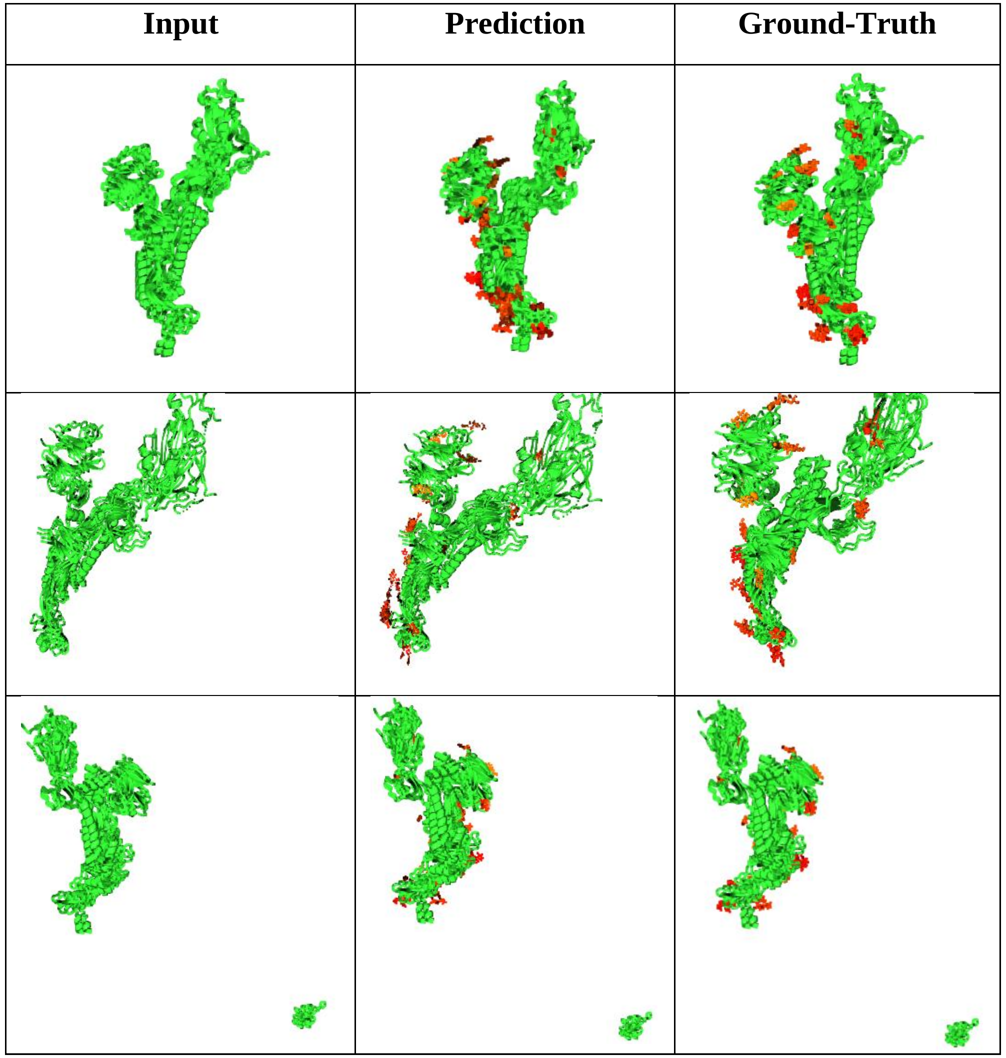


Table 2: Visual Representation

We evaluated our method on a series of viral Proteins of the SARS-CoV-2. We compare quantitative and qualitative performance in Table 1 and Table 2. Quantitative the difference between the generated systems and the ground truth is small. Qualitative our solution achieves good generation of complete structures. The learned representation does generalize to systems beyond the ones used during training.

Conclusions and next steps

Our experiments show that we are able generate complete systems and to generalize to structures of unseen systems. Translating the input systems into the latent space permits searchability for similar structures and sampling from the latent space for generation.

Topics for future work include integrating the search capabilities in the training process, exploring alternatives for sampling and generating from regions of interest.

Thank you

De Novo Drug Design using A.I.

ASYNT-GAN

Ivan Jacobs[1], Manolis Maragoudakis[2]
[1]AI4U, [2]Ionian University
ivan.jacobs@ai4u.ai, mmarag@ionio.gr