1. RESEARCH IDEA

Research title: Deep learning applied to drug discovery and the life sciences

Keywords: Artificial Intelligence, Deep Learning, Representation Learning, Drug Discovery, Protein Structure Prediction, Protein Folding, Antibody Structure Prediction, Protein Design, Antibody Design, Protein-Protein Interaction

Desired date for the research project to be launched in 2021: 10-01-2022

Purpose

Artificial Intelligence (AI) has been applied successfully in different areas, and recently achieving breakthrough results to protein structure prediction with AlphaFold. In this letter of intent, we describe our research proposal which main goal is to shorten the time needed at the early drug discovery phase by developing AI methods and frameworks for tasks such as protein/antibody structure prediction, protein/antibody sequence design, protein-protein interaction, and evaluation of target candidates.

Research Idea:

Recently, our society has been facing a great challenge with the Covid-19 pandemic. For two years, we have been living with actions that range from mandatory quarantines, travel restrictions and limitations on social gatherings. At the beginning of the pandemic, one of the main concerns was the time that would take to develop a vaccine that could target the SARS-CoC-2 virus. Usually, a vaccine takes 5-10 years to be developed [1], but what we observed was the fast development (in around 12 months) of vaccine candidates by companies such as Pfizer, Moderna, and Janssen. This rapid development process was crucial for alleviating restrictions worldwide. The rapid development vaccines were possible because SARS-CoV-2 is a member of the coronavirus family, that have been studied by decades, so that data on structure, genome, and life cycle of this type of virus were already known and offered a strong literature for scientists [2]. To be prepared for another pandemic, in which there is the possibility of less knowledge regarding the virus is available, it is crucial that we shorten the time needed for the discovery and development of new drugs. An important factor on achieving this objective is the effective applications of computational methods that can accelerate the early drug discovery process.

Artificial Intelligence, especially a branch named as Deep Learning (DL) that stack layers of neural networks and non-linear activation functions, has been regarded as an important method achieving high performance in tasks which representations involves images, natural language, and speech. Lately, an AI-based method called AlphaFold [3] was able to model a problem that has been a great challenge for scientists in the life sciences, protein structure prediction. Experimentally determining the structures of proteins was a laborious and time-consuming task, and, before AlphaFold in silico computational methods were not accurate for doing accurate predictions in various scenarios. With the development of AlphaFold and its release for the community as an open source software with ColabFold [4] and OpenFold[5], researchers have been able to obtain big advances on drug development and understanding more about

the protein functions. In parallel to the development of AlphaFold and other AI-based structure prediction networks [6], deep learning has also been successfully applied recently to protein sequence design [7, 8]. For the protein sequence design problem, a method named ProteinMPNN has obtained not only high recovery rate, but also, the sequences designed by network led to high developability in wet lab experiments achieving proteins with good properties, such as solubility. After the experimental results obtained using the method, ProteinMPNN is currently the main design tool utilized at the Institute for Protein Design [8], replacing other traditional computational methods. The recent advances in protein structure prediction and protein sequence design using deep learning, with the open sourcing of these methods to the research and industry communities, have given powerful tools that can influence and accelerate new discoveries that can change the life of our society in the near future.

However, still many possibilities on the application of deep learning methods and the development of novel computational biology methods are to be explored in the life sciences, especially tackling early drug discovery processes. Usually, the application of a protein or antibody with a therapeutics purpose lead to its interaction with another protein with the goal of neutralizing the activities of that protein, e.g. a specific antigen. For effectively tackling this generative problem, the research in areas such as interactor-conditioned protein design and antigen-conditioned antibody design is crucial and still AI-based methods that achieve reasonable performance are needed. Other challenging problems in drug discovery in which deep learning can be applied involves docking, functionality prediction, affinity maturation, developability prediction, and the development of new in silico evaluation methods. The development of new AI-based methods that tackle these problems are the main objective of this research proposal and will be detailed next.

The main part of the proposal is organized into four components: 1) Data creation, data augmentation and representation learning for proteins, 2) Deep learning-based applications in bioinformatics, 3) Generative models for conditioned protein and antibody design using deep learning, 4) Evaluation methods: in silico and wet lab experiments. Each section includes a brief description and literature review and what are the main problems that we aim to tackle during the Young Scientist Fellowship program.

Finally, I will conclude this letter of intent on why I would like to continue my research in IBS and within the IBS Data Science Group.

1.Data creation, data augmentation and representation learning for proteins

Training DL algorithms that achieve state-of-the-art performance usually requires highly structured and curated datasets. The recent advances obtained by AlphaFold were made possible by scientists working in the life sciences by organizing the data obtained in wet lab experiments around different universities and research institutes for decades and sharing the data with the research community. For example, AlphaFold and ProteinMPNN are trained using protein structures experimentally collected during the last 50 years deposited in the Protein Data Bank (PDB) [9]. The PDB dataset contains hundreds of thousand of protein structures, still, for specific types of proteins, such as antibodies, the number of examples is highly reduced (to a few thousands of structures), what affects the application of methods trained with this data for several applications. Given this problem, data creation and data curation for novel protein datasets especially modeled for drug discovery applications is important to be considered by the scientific community.

Protein data is affected by evolution, i.e. evolutionary related proteins have similar sequences, structures, and functions. When we analyze the PDB data, for example, data is biased by protein families that contain more examples in the dataset. Therefore, usually, when pre-processing protein data, scientists cluster data from the same protein families to alleviate the bias problem. This leads to, for example, after the pre-processing of the PDB data, less than fifty thousand experimental examples are used to train protein sequence design methods such as ProteinMPNN. It is important, then, to explore novel methods to augment the datasets available. Recently, AlphaFold has been applied to generate an augmented dataset that provides open access to more than 990 thousand protein structure predictions [10]. An antibody structure prediction network named IgFold also creates an augmented set of more than 100 thousand antibody structures [11]. The augmented data generated by these works can be an important part of improving the understanding of nature and developing deep learning methods for novel applications in which data was a prohibitive factor.

Finally, another important factor when modeling the dataset and a deep learning-based model is the data representation. As proteins are represented by a sequence of amino acids from a limited discrete set, they can be simply encoded by one-hot encoding vectors. Given the vast availability of protein sequence data and the recent advances of natural language processing algorithms, recently the application of AI-based large language model architectures, such as BERT and GPT-2, have been also used as a protein representation achieving good results [12]. These sequence-based representation methods have been applied to different life sciences problems ranging from protein functionality prediction [13] to affinity maturation [14]. Different types of structure-based representations, such as graphs [7], distance matrices [15], and point clouds [16], have been also being used to represent proteins. To have an accurate structure-based representation is an important part of the development of sequence design methods, i.e. designing an amino acid sequence that represents a target sequence. With the high accuracy and computational efficiency of structure prediction networks, structure-based representations have also been currently applied to predict structure from sequence in inference time for protein functionality prediction [17].

2.Deep learning-based applications in bioinformatics

Deep learning has been applied for various challenges in bioinformatics. These deep learning-based applications are currently being used as powerful tools for different pipelines. Next, we present a brief literature review on some of the recent developments to solidify the potential of research in the field. Novel methods are expected to be developed and applied to areas that are still incipient.

Firstly, we analyze the application of DL methods to the structure prediction problem, i.e. predicting structure from sequence, for proteins and antibodies. The trRosetta network proposed in [16] considers structure prediction as a classification problem of four features that represents the distance and angles between amino acids. As another view on this problem, AlphaFold [3] views structure prediction as a graph inference problem. The architecture is divided in three modules, the first pre-process the input sequence creating multiple sequence alignments (MSAs) and templates, the second contain stacked Evoformer blocks to refine sequence and pair representations, and, finally, the third is a structure module that predicts the final structure. In RosettaFold [6] a ‘three-track’ neural network is used to prediction. Similar to AlphaFold, one module is used for sequence, one module for pair wise representation, and one module for structure. The architecture of RosettaFold include graph transformers and SE(3)-transformers.

Given the hypervariability of antibodies and that only few thousands of antibody structures are contained in the PDB, antibody structure prediction is still a challenge. In [18] a computationally efficient neural network named IgFold is proposed. The architecture of IgFold consists of a pre-trained antibody language model and a graph neural network (GNN) that directly tries to predict atom coordinates. The recent advances in structure prediction networks provided a powerful tool for researchers, and still more advances especially related to removing the need of MSAs, reducing inference time, and creating reversible architectures is expected in the near future.

Next, we describe the application of DL methods to the protein sequence design problem, i.e. predicting structure from sequence. An important aspect of being able to generate a protein sequence given a conditioned model is how the protein is represented. In [7] a graph representation is used and the protein design problem is casted as a language modeling conditioned on an input graph. In [19] a method named ProteinSolver also uses a deep GNN to model protein design as a constraint satisfaction problem, similarly to the puzzle Sudoku. McPartlon et al [20] utilizes a generative SE(3)-equivariant graph transformer architecture that uses a partial masking scheme to predict each amino acid identity and side-chain conformation. Yang et al [21] utilizes additional information parameterized by a masked language model to condition the structure of an architecture similar to [7] for protein sequence design. In [22] the training data is augmented by nearly three orders of magnitude using predicted from AlphaFold. Experimental and augmented data are used to train a geometric vector perceptron that is combined with GNNs and transformers for training. ProteinMPNN [8] improves the performance of the architecture proposed in [7] by adding backbone atom coordinates to the protein representation. An order agnostic autoregressive model is also proposed to enable the application of the method to multi-chain design problems. ProteinMPNN achieved high developability and solubility rates in experimental tests. Protein sequence design is a crucial problem to the development of new drugs and have achieved results proven experimentally, but still additional advancements are needed in interactor-conditioned design and on the application of these methods to antibody sequence design.

Language models (LMs) have achieved outstanding performance recently into various natural language processing tasks. These models have the ability to act as generative methods, i.e. generate new text based on a condition or on what was generate before, and also to effectively learning a representation space for words and sentences, i.e. representing a word as a vector of floats that can encode semantic and syntactic patterns. A protein is a sequence of a discrete set of amino acids, in a way that a protein sequence can be interpreted as a large word. Given the similarity of both sequence-based protein representation and character-level language representation, language models were trained to encode protein representations. Unirep [23] learns a fixed-length vector representation of proteins by training a recurrent neural network (RNN) architecture on a set with 24 million protein sequences. ProteinBERT [12] improves upon a classic Transformer/BERT architecture by separating local (amino acid level) and global (entire protein sequence) representations, supporting different types of representation for local and global tasks. ProteinBERT is trained on 106 million protein sequences. ESM-1b [24] train a transformer-based language model across 250 million protein sequences. Differently from the previous methods that rely on RNN archiectures and attention mechanisms, CARP [25] proposes the use of convolutional neural networks (CNNs) to learn protein representations to improve computational efficiency while achieving similar results to other pretrained protein language models. Language models have also been applied to learn antibody representation. IgLM [26] trains a deep generative language modeling on 558M antibody heavy-chain and light-chain variable sequences. Each sequence generation is conditioned on the chain type and

species-of-origin. ProGen2 [27] develops a suite of language models with different number of parameters to proteins and antibodies. ProGen2 uses an autoregressive transformers architecture and, for antibodies, is trained on data similar to [26]. Recent work has been also investigated to understand the capacity of these protein and antibody language models to learn structural properties only from sequence data. It was also shown in [27] that training with more data can hurt the language model performance, and more emphasis should be put into the data distribution. The understanding of the representations generated by these models and the combination with structure-based representation will be crucial to solve various tasks, such as functionality prediction.

In addition to the aforementioned tasks, DL has also been successfully applied to many other important applications in the life sciences. Next, we will briefly present additional methods relevant to drug discovery. MaSIF [28] proposes a geometric deep learning method that use a surface representation including chemistry features and geometry features to three tasks: protein pocket-ligand prediction, protein-protein interaction site prediction, and prediction of protein-protein complexes. dMaSIF [29] extends the method presented in [28] by eliminating the need for pre-computed surface features in a way that the model can efficiently be trained in an end-to-end manner. In [30], Jha et al utilizes both sequence and structure features as input for a graph convolutional network and a graph attention network for protein-protein interaction prediction. LM-GVP [31] also focus on combining sequence and structure features for protein functionality prediction using language models and a geometric vector perceptron architecture. SMCDiff [32] generates a scaffold conditioned on a given motif using diffusion models. SMCDiff is able to generate scaffold with up to 80 amino acids sampling from a neural network trained to learn a distribution over protein backbones. Affinity maturation is performed using an antibody language model in [33]. It is shown in [33] that, using an antibody LM, affinity can be improved even without using the antigen information. Still, many more DL-based methods are being developed and applied in the life sciences, with a huge potential for applications in which experimental data is available.

3.Generative models for conditioned protein and antibody design using deep learning

Pivotal for the development of a new therapeutics is the design of a protein that interacts with another specific target protein. For example, when developing an antibody therapeutic, there exists an antigen that we want to target and neutralize. It is also a possibility that the region that we want to interact with to neutralize the antigen is known during the design process. In this research proposal, the main focus is on the investigation of generative models that are conditioned to a target protein. We define this problem as follows. We first assume that a target protein/antigen is given. The region of the target protein/antigen in which we want to interact might be given depending on the properties of the disease being investigated. The main objective in this part is to develop an interactor-conditioned generative model to design protein/antibodies using deep learning.

Here, we review different deep learning-based conditioned generative methods developed recently by the research community as possible options to tackle this problem. As proteins have different types of possible representations, such as sequence, structure, and surface, methods for applications with different conditioning and generation targets are described. First, we think about an application in which the condition is a sequence and the output of the generation is also a sequence, e.g. neural machine translation (NMT) or question and answering (QA). Usually, NMT solutions are based on an encoder-decoder architecture in which each block (encoder and decoder) are implemented using Transformer

models [34]. In this way, the encoder learns a representation for the sentence in the source language that conditions the generation. The decoder, on the other hand, is responsible to generate the sentence in the target language using this representation and the information regarding the source language and the words already generated in the target language. For the QA task, in [35] a method utilizing two types of contextual word embeddings, ELMo and BERT, integrated with a transformer encoder is proposed. This method when fined-tuned with RoBERTa obtain state-of-the-art results for QA. Lately, text-to-image generation models have achieved breakthrough results that can change the art industry. The input of these systems is an input sentence that conditions the generation of an image. The image should be generated in a way that is well suited for that input sequence (prompt). These methods, e.g. DALLE [36] and Imagen [37], relies on large language models to encode the input sentence and on diffusion models for high-fidelity image generation. Diffusion models for the generation of 3D point clouds have also been investigated and are a promising direction for modeling 3D shapes [38]. A two-way flow-based generative model named C-Flow was proposed in [39] to generate 3D point clouds from images and images from 3D point clouds. The successful application of these methods of interactor-conditioned protein/antibody design has the potential of revolutionize the drug discovery process and is the main goal of this research proposal.

To apply these methods to protein design, various challenges should be addressed. Ongoing investigations are being performed at our research group testing various approaches on how to tackle these challenges. Next, some of these challenges are described. The first challenge is choosing the type of representation to use for the encoder and decoder parts. Recent results obtained by protein sequence design methods like [8] suggests that to generate a structure-based generation might be desired in these applications. The second challenge is what to condition on and what to generate. While conditioning and generating the entire protein is desirable, it is computationally demanding. So, only conditioning/generating based on epitope and paratope combined with a general representation of the proteins is a possibility. If a structure-based representation is used, the protein generated should consider backbone structure constraints. Finally, another challenge for this application is the data available of protein complexes and antigen-antibody pairs. Datasets, such as the one provided in [40], still have a low number of examples compared to the data needed to cover the protein space and for the development of deep learning models that can effectively generalized to different interactors. Additional datasets specifically designed for interactor-conditioned protein design should be created. A possibility to alleviate this problem is to use protein complexes predicted by AlphaFold to augmented this dataset.

4.Evaluation methods: in silico and wet lab experiments

In the end, after designing a protein or an antibody that is a potential therapeutic drug, the evaluation of the final design should be made in wet lab experiments. As AI-based generative methods have the property of generating a large number of candidates, it is not feasible to test all the possible candidates experimentally. Therefore, the development of in silico methods for evaluating potential candidates is important to reduce the cost and time spent for the design of new drugs. The development of novel in silico methods is an interdisciplinary research that involve the collaboration of specialists in the training of AI models and with experience on drug discovery using traditional pipelines.

The improvement of the evaluation methods involves different possibilities. These involve, for example, the development of more accurate energy functions that can predict more accurately the stability of a

protein or the interaction between multiple proteins in protein complexes. Preliminary research is also being developed applying AI-based models to the modeling of these functions. Another possibility is the development of in silico software that can model the processes done experimentally, such as the in silico evaluation improves the resemblance to what is done in wet lab experiments, this include the influence of solvents in protein expression. Finally, it is necessary the development of better evaluation criteria given a set of parameters to select the candidates that will be selected for further experimental evaluations.

5. Concluding remarks

In the next few years, I would like to study the development of new AI-based models that can be applied to drug discovery and the life sciences. My main research interests and research topics include (but are not limited to):

-Data creation, data augmentation and representation learning for proteins

-Deep learning-based applications in bioinformatics

-Generative models for conditioned protein and antibody design using deep learning

-Evaluation methods: in silico and wet lab experiments

The main outcome of my research is to shorten the time needed during the early drug discovery process by the development of new AI-based models. These models can also be important and open the possibility to tackle rare diseases that are not deeply investigated and that currently do not have medical treatment. We hope that our research contributes to the bioinformatics and the computational biology communities. We expect that the algorithms and the software developed during the research will have general properties that will allow its use to application in different research areas. Given the novelty of the research proposal, it is expected that our contributions also extend to the Artificial Intelligence and Machine Learning communities.

The IBS Data Science Group is the best place to conduct this research topic since our group has close interdisciplinary collaboration with other IBS centers, including an ongoing collaboration with the IBS Protein Communication Group working on the development of new therapeutics. The development of AI-based models involves the need of high computational power. The infrastructure provides by IBS and the IBS Data Science Group is crucial for the development of this project. In addition, part of this project involves the need of wet lab experiments in an interdisciplinary collaboration with biologists and professionals from the life sciences. The current collaboration with the IBS Protein Communication Group makes IBS the right environment for the development of this research project.

In IBS there is a unique opportunity to collaborate with top students, from universities like KAIST and UST, and with top researchers in their field. Currently, one graduate student and six undergraduate students are under my direct supervision. As a researcher, I am passionate to mentor students to achieve their personal and professional goals. Having the possibility to mentor and collaborate with students and researchers is a unique opportunity to my personal development and to my professional career.

With the generous support of the Young Scientist Fellowship program, I hope to achieve accelerate the process of drug discovery and influence the future of our society.

References:

[1] https://coronavirus.jhu.edu/vaccines

[2] https://www.medicalnewstoday.com/articles/how-did-we-develop-a-covid-19-vaccine-so-quickly#Other-coronaviruses

[3] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.

[4] Mirdita, Milot, et al. "ColabFold: making protein folding accessible to all." *Nature Methods* (2022): 1-4.

[5] https://openfold.io/

[6] Baek, Minkyung, et al. "Accurate prediction of protein structures and interactions using a three-track neural network." *Science* 373.6557 (2021): 871-876.

[7] Ingraham, John, et al. "Generative models for graph-based protein design." *Advances in neural information processing systems* 32 (2019).

[8] Dauparas, Justas, et al. "Robust deep learning based protein sequence design using ProteinMPNN." *bioRxiv* (2022).

[9] Bank, Protein Data. "Protein data bank." *Nature New Biol* 233 (1971): 223.

[10] Varadi, M *et al*. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* (2021).

[11] Ruffolo, Jeffrey A., and Jeffrey J. Gray. "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies." *Biophysical Journal* 121.3 (2022): 155a-156a.

[12] Brandes, Nadav, et al. "ProteinBERT: A universal deep-learning model of protein sequence and function." *Bioinformatics* 38.8 (2022): 2102-2110.

[13] Wang, Yiquan, et al. "A large-scale systematic survey reveals recurring molecular features of public antibody responses to SARS-CoV-2." *Immunity* (2022).

[14] Ruffolo, Jeffrey A., Jeffrey J. Gray, and Jeremias Sulam. "Deciphering antibody affinity maturation with language models and weakly supervised learning." *arXiv preprint arXiv:2112.07782* (2021).

[15] Yang, Jianyi, et al. "Improved protein structure prediction using predicted interresidue orientations." *Proceedings of the National Academy of Sciences* 117.3 (2020): 1496-1503.

[16] Wang, Yeji, et al. "A point cloud-based deep learning strategy for protein–ligand binding affinity prediction." *Briefings in Bioinformatics* 23.1 (2022): bbab474.

[17] Lee, Minji & Rzayev, Anar & Jung, Hyunkyu & Vecchietti, Luiz Felipe et al. “Structure-based representation for protein functionality prediction using machine learning.” in Proceedings of the Korea Computer Congress (KCC), (2022).

[18] Ruffolo, Jeffrey A., and Jeffrey J. Gray. "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies." *Biophysical Journal* 121.3 (2022): 155a-156a.

[19] Strokach, Alexey, et al. "Fast and flexible protein design using deep graph neural networks." *Cell systems* 11.4 (2020): 402-411.

[20] McPartlon, Matt, Ben Lai, and Jinbo Xu. "A Deep SE (3)-Equivariant Model for Learning Inverse Protein Folding." *bioRxiv* (2022).

[21] Yang, Kevin K., Niccolò Zanichelli, and Hugh Yeh. "Masked inverse folding with sequence transfer for protein representation learning." *bioRxiv* (2022).

[22] Hsu, Chloe, et al. "Learning inverse folding from millions of predicted structures." *bioRxiv* (2022).

[23] Alley, Ethan C., et al. "Unified rational protein engineering with sequence-based deep representation learning." *Nature methods* 16.12 (2019): 1315-1322.

[24] Rives, Alexander, et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." *Proceedings of the National Academy of Sciences* 118.15 (2021): e2016239118.

[25] Yang, Kevin K., Alex X. Lu, and Nicolo K. Fusi. "Convolutions are competitive with transformers for protein sequence pretraining." *bioRxiv* (2022).

[26] Shuai, Richard W., Jeffrey A. Ruffolo, and Jeffrey J. Gray. "Generative language modeling for antibody design." *bioRxiv* (2021).

[27] Nijkamp, Erik, et al. "ProGen2: Exploring the Boundaries of Protein Language Models." *arXiv preprint arXiv:2206.13517* (2022).

[28] Gainza, Pablo, et al. "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning." *Nature Methods* 17.2 (2020): 184-192.

[29] Sverrisson, Freyr, et al. "Fast end-to-end learning on protein surfaces." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[30] Jha, Kanchan, Sriparna Saha, and Hiteshi Singh. "Prediction of protein–protein interaction using graph neural networks." *Scientific Reports* 12.1 (2022): 1-12.

[31] Wang, Zichen, et al. "Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction." *Scientific reports* 12.1 (2022): 1-12.

[32] Trippe, Brian L., et al. "Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem." *arXiv preprint arXiv:2206.04119* (2022).

[33] Ruffolo, Jeffrey A., Jeffrey J. Gray, and Jeremias Sulam. "Deciphering antibody affinity maturation with language models and weakly supervised learning." *arXiv preprint arXiv:2112.07782* (2021).

[34] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[35] Laskar, Md Tahmid Rahman, Xiangji Huang, and Enamul Hoque. "Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task." *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020.

[36] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *International Conference on Machine Learning*. PMLR, 2021.

[37] Saharia, Chitwan, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." *arXiv preprint arXiv:2205.11487* (2022).

[38] Luo, Shitong, and Wei Hu. "Diffusion probabilistic models for 3d point cloud generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[39] Pumarola, Albert, et al. "C-flow: Conditional generative flow models for images and 3d point clouds." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[40] Akbar, Rahmad, et al. "A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding." *Cell Reports* 34.11 (2021): 108856.

Additional Remarks

For the development of novel AI-based models for drug discovery, as described in this letter of intent, the computational resources offered by the IBS Data Science Group will be a great asset. Also, for experimentally evaluating the models designed by this project, the interdisciplinary collaboration between the IBS Data Science Group and the IBS Protein Communication Group is an important factor on why IBS is the right place to conduct this research.