Research Plan - “Accelerating the discovery of high-quality drug candidates via deep learning”

Introduction

Recently, the Covid-19 pandemic highlighted the need for faster development of therapeutics and vaccines. At the beginning of the pandemic, one of the main concerns was the time needed to develop a vaccine that could target SARS-CoV-2. Usually, a vaccine takes 5-10 years to be developed [1], yet various companies were able to develop vaccine candidates in 12 months. The rapid development of vaccines was possible because SARS-CoV-2 is a member of the coronavirus family. This virus family has been studied for decades, so its structure, genome, and life cycle are well-documented to scientists [2]. To be prepared for another pandemic in which we may know less about the target virus, we must speed up drug discovery and development. For that, the effective application of computational methods is an important factor.

Artificial Intelligence (AI), especially a branch named Deep Learning (DL), has been regarded as an important research area achieving high performance in tasks whose representations involve images, natural language, and speech. For example, AlphaFold [3] made a leap in solving a problem that has been a great challenge for scientists in the life sciences, protein structure prediction. Experimentally determining the structures of proteins is laborious and time-consuming. Before AlphaFold, in silico computational methods could not accurately predict protein structures. With the development of AlphaFold and its release to the research community, researchers have made big advances on drug discovery. Deep learning has also been successfully applied to other protein related tasks, such as sequence design [4, 5]. For the protein sequence design problem, ProteinMPNN has high sequence recovery accuracy and high developability in wet lab experiments, achieving proteins with good properties, such as solubility. After the success achieved in wet lab experiments, ProteinMPNN has now become the main design tool utilized at the Institute for Protein Design [5], replacing other traditional methods.

The application of deep learning methods and the development of novel computational biology methods are only at the beginning stage and new exciting ventures need to be explored in the life sciences, especially for the early drug discovery process. Usually, the application of a protein or antibody for a therapeutics purpose leads to its interaction with another target protein. The goal of this interaction is to neutralize the target protein, such as an antigen. For effectively tackling this generative problem, research in interactor-conditioned protein design and antigen-conditioned antibody design is crucial. I am eager to investigate how AI-based methods may facilitate these tasks. Docking, affinity maturation, developability prediction, and the development of new in silico evaluation methods are other drug discovery challenges that deep learning may solve. The development of novel AI-based methods to tackle the aforementioned challenges is crucial to achieving my research goal.

In this research proposal, a main objective is to shorten the time needed at the early drug discovery process by improving existent and developing novel AI models for antibody structure prediction, antibody sequence design, and antigen-antibody interactions. These methods will be evaluated by standard in silico methods and the best candidates verified in wet lab experiments in an interdisciplinary collaboration between two IBS research groups: Data Science Group and Protein

Communication Group. Here are the individual objectives to be achieved with this Young Scientist Fellowship (YSF) program.

- Develop an AI-based target-conditioned antibody design method.
- Develop a validation software system that combines AI-based methods to accelerate the search for potential candidates in the early drug discovery phase.
- Improve AI-based methods for antibody related tasks, such as antigen-antibody docking, affinity maturation, and developability prediction. This improvement is to be achieved by algorithmic improvements, application of interpretable DL methods, and computational efficiency optimization of recent state-of-the-art methods being applied in the early drug discovery phase.

The main part of this proposal is organized as follows: 1) Data creation, data augmentation and robust representation learning for antibodies, 2) Target-conditioned antibody design using deep learning, 3) Evaluation methods: in silico and wet lab experiments. Each section includes a brief description and literature review regarding the main problems that we aim to tackle during the YSF program. In addition to introducing recent studies, we discuss potential weaknesses that can be tackled with the aim of proposing novel methods to accelerate drug discovery and to have candidates that translate successfully to wet lab experiments.

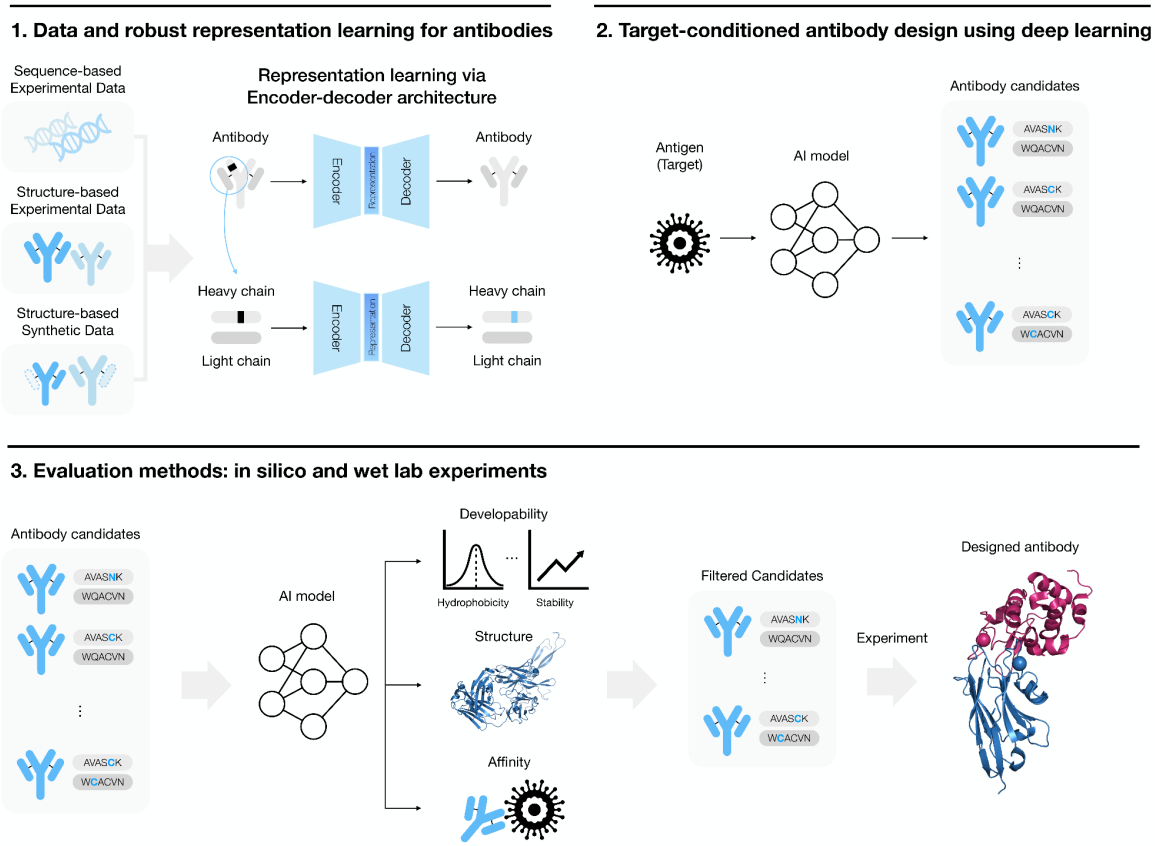

Fig. 1 - Research Overview

Data creation, data augmentation and robust representation learning for antibodies

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 the availability of extensive experimental data compiled over different universities and research institutes for decades. For example, AlphaFold and ProteinMPNN are trained using protein structures experimentally collected and deposited in the Protein Data Bank (PDB) [6]. The PDB dataset contains hundreds of thousands of protein structures. For specific types of proteins, such as antibodies, however, the number of examples is reduced to a few thousands, which becomes small for data-driven learning. Given this problem, creation and curation of antibody datasets will help facilitate many drug discovery applications and hence be an important contribution.

Data augmentation is an emerging alternative for applications where data is scarce. Protein data is affected by evolution, i.e. evolutionary related proteins have similar sequences, structures, and functions. As a result, protein families with more examples in the PDB dataset can bias the data. To alleviate this problem, usually, when preprocessing the data, scientists have clustered datasets by protein families. After this preprocessing step, less than fifty thousand experimental protein structures are available to train sequence design methods. This data shortage is what motivates my research of data creation and data augmentation. To augment available datasets, AlphaFold generated an augmented dataset with open access to over 200 million protein structure predictions [7], and IgFold created an augmented set of over 100 thousand antibody structures [8]. By studying these augmented datasets, together with my colleagues, I would like to contribute in a better understanding of biological structures and developing AI-based methods for expanding available datasets that could be widely used in research. We ought to investigate how these synthetic datasets can be used effectively to improve generalization of structure prediction, sequence design models, and how these can be applied to learn about protein-protein interactions.

When building AI-based models for drug discovery, it is crucial to have a robust protein representation. Given the vast availability of protein/antibody sequence data and the recent advances of natural language processing (NLP) algorithms, AI-based large language model architectures are being used to learn a compact sequence-based protein representation [9]. For protein structures, different types of representations, such as graphs [4], distance matrices [10], and point clouds [11], have been used. These representations need to fulfill importante properties such as being translation invariant and rotation invariant. Having a robust structure-based representation is essential for protein and antibody sequence design, i.e. designing an amino acid sequence that encodes a target antibody structure. I plan to investigate robust structure-based antibody representations and train models to encode antibody structures in a compact manner, and help the team develop a target-conditioned antibody design method. One immediate idea is to use a variational auto-encoder architecture to encode an antibody structure, and from this learned representation decode again the same antibody structure. This compact representation will be combined with generative models for antibody design later in this proposal.

Target-conditioned antibody design using deep learning

Pivotal for the development of new therapeutics is the design of a protein that interacts with another specific target protein. This is the main purpose of this research proposal. For example, when developing an antibody therapeutic, there exists an antigen that we want to target and neutralize. It is also assumed that the desired antigen region is known during design. My goal is to develop generative models that are conditioned to a target protein. More specifically, given a target protein/antigen and the desired interaction region, the main objective is to develop an AI model to generate an antibody that interacts with the target protein at the target binding location.

I will use deep learning to implement a target-conditioned generative model to design antibodies. I am currently reviewing different deep learning-based conditioned generative methods to get insights on how to effectively tackle this problem. As proteins have different types of possible representations, such as sequence, structure, and surface, methods with different representation properties are to be investigated. First, we introduce applications in which the condition is a sequence and the output of the generation is also a sequence, e.g. neural machine translation (NMT). Usually, NMT solutions are based on encoder-decoder architectures using Transformer models [12]. The encoder learns a representation for the sentence in the source language and the decoder is responsible for generating the sentence in the target language. More recent applications envision text-to-image generation models, which is a breakthrough idea. The input of these systems is an input sentence that conditions the generation of an image, where the image should be generated in a way that is well suited for that input sequence (prompt). These methods, e.g. DALL-E [13], rely 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 [14]. The successful application of these methods for target-conditioned antibody design has the potential to revolutionize drug discovery and is the main goal of this research proposal.

To apply generative methods to target-conditioned antibody design, various challenges should be addressed. 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 [5] suggest that a structure-based generation might be desired in these applications. I would also like to pursue this direction. The second challenge is what to condition on and what to generate, because conditioning and generating the entire protein/antibody will be computationally demanding. An alternative method will be to only condition on the epitope and on a general representation of proteins, which is the direction I would like to investigate further. If a structure-based representation is used, the antibody generated should consider backbone structure constraints. Finally, another challenge is the limited data availability of antigen-antibody pairs and protein complexes. Datasets, such as the one provided in [15], have a low number of examples compared to the data typically needed to cover the protein space and for the development of deep learning models that can effectively generalize to different targets. I hope to contribute to expanding the dataset for target-conditioned antibody designs. As a start, I would like to use proteins with loop-like conformations and protein complexes predicted by AlphaFold-Multimer [3] or UniFold-Multimer [16] for data augmentation.

Evaluation methods: in silico and wet lab experiments

One of the most exciting aspects of my research is the collaboration between computer science and medical science. Thanks to this collaboration, I can go a step further in designing AI models and have the opportunity to evaluate the generated outcomes as potential therapeutic drugs 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 in early drug discovery. The development of novel in silico methods is an interdisciplinary endeavor requiring the collaboration of specialists in AI models and drug discovery using traditional pipelines. Our goal is to develop an evaluation software system, involving multiple AI-based evaluation methods, to filter high-quality candidates for further experiments. An example of an AI-based evaluation methodology that can be applied to this system is the usage of confidence measures predicted from AlphaFold [3], such as the predicted local distance difference test (pLDDT) and the predicted aligned error (pAE), to rank potential candidates. For a target antigen, a total of 10-15 candidates generated by the proposed target-conditioned antibody design method and filtered by the evaluation software system implemented are to be expressed for experimental validation.

The development of new AI-based evaluation methods also involves additional possibilities, ranging from improving physics-based energy models to the development of new in silico software. The development of more accurate energy functions is an active area of investigation. The objective is to predict more accurately the stability of a protein or the interaction between multiple proteins in protein-protein and antigen-antibody complexes. Preliminary research is also being developed applying AI-based models to the modeling of these functions. The development of in silico software involves the ability to model the processes done experimentally, such as to improve the resemblance of the simulator to what is done in wet lab experiments. This includes, for example, the influence of solvents in protein expression.

Concluding Remarks

I am excited with the opportunity to develop novel 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 robust representation learning for antibodies
- Target-conditioned antibody design using deep learning
- Evaluation methods: in silico and wet lab experiments

The main outcome of this research proposal is a set of AI-based models that can speed up the early drug discovery process. These models can also help tackle rare diseases that are not deeply investigated and that currently do not have medical treatment. I truly believe these research goals will have a positive impact in society and contribute to advancing bioinformatics, computational biology, and artificial intelligence. The algorithms and software programs that come out of this research will be shared to benefit the wider research community across multiple domains.

Relevance to the Objectives of the Research Center

The IBS Center for Mathematical and Computational Sciences (Data Science Group) is dedicated to using data science and artificial intelligence to solve global societal problems. We believe that this proposal is an excellent fit and work towards to achieve this bigger and common objective. The research topics proposed are state-of-the-art research algorithms which require cutting-edge technology in data science, applied AI and AI theory. In addition, to solve global societal problems, it is crucial to perform high-quality interdisciplinarity research that is one of the main objectives of IBS, the Center for Mathematical and Computational Sciences and the Data Science Group.

In the IBS Data Science Group we are tackling diverse problems involving data science and AI to a range of different problems, such as climate change and protein design. For the past year, I have been guiding a team that is involved in an interdisciplinary collaboration with the IBS Protein Communication Group to apply AI methods to protein and antibody design. By integrating both areas, we have been working to propose novel AI methods and facilitate the understanding and use of a gamma of AI methods proposed recently. This YSF will allow me to continue and strengthen the on-going collaboration and to develop computational methods that can make a broader impact to the overall research community.

Budget Plan

The summary of my budget plan is as follows.

Salary for Young Scientist Fellow	60 million (+ social insurances) = 73 million
Three Research Assistants (Intern, MSc student, PhD student)	9.6 million x 1 + 14.4 million x 1 + 24 million x 1 = 48 million
Attending conferences	5 million x 2 = 10 million
Seminar + visitors	4 million
IT equipments + maintenance	25 million
Total	160 million / year

My first plan is to hire additional junior researchers, including BSc, MSc, and PhD level student assistants. With the YSF budget, the plan is to support up to three student assistants with various backgrounds, including computer science, biology, and bio and chemical engineering. Moreover, by supervising them I will be preparing my skills for the next stage of my life, in which I want to be a PI or a faculty member. I plan to use the IBS and IBS Data Science Group workstations to run experiments, and, given the recent release of big synthetic datasets in the field, use additional resources to buy storage, graphic cards and laptops for my student assistants. In addition, I plan to attend top workshops and conferences yearly to exhibit my research outcomes, such as NeurIPS, ICML, ICRL, and AAAI, and in computational biology, such as RosettaCon and ICSB.

Organization Plan

The YSF program will give me the opportunity to conduct rigorous, creative and high-quality research. Research ideas described in this proposal have the potential to influence life sciences and computer science research communities. The research will also benefit by the interdisciplinary collaboration with multiple IBS research teams. I will target challenging conference venues for artificial intelligence and computational biology, such as NeurIPS, ICML, ICLR, AAAI, RosettaCon, and ICSB, and also prestigious venues for the life sciences, such as Nature, Nature Machine Intelligence, Bioinformatics, and PNAS. The intended plan described by year is as follows.

(Year 1) My main focus will be in the development of AI-based target-conditioned antibody design methods. Curation and augmentation of existing datasets with the target of generating antibodies using sequence-based and structure-based representations will both be investigated. Moreover, I plan to add recent state-of-the-art models to create an evaluation software system that aims to filter pre-candidates to wet lab experiments.

(Year 2) My focus will be to validate the methods and systems created experimentally. This will be a way to re-evaluate our research directions and prove the proposed methodology. For the research of novel AI methods, the focus will be on models that can predict direct metrics that correlates with experience in wet lab experiments, such as developability. The plan is to also compare sequence-based and structure-based representations to choose a target-conditioned antibody design methodology that will be prioritized.

(Year 3) The main objective is to focus on models that improve antigen-antibody interactions. In this case, methods that involve affinity maturation and improve protein-protein interactions are to be prioritized. It is expected that the models proposed during this YSF program are also integrated to the evaluation system created. I plan to also write a review of the above studies and document the created systems for the use of the research community.

Collaboration Plan

The IBS Center for Mathematical and Computational Sciences (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 Center for Biomolecular and Cellular Structure (Protein Communication Group) working on the development of new therapeutics. 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 Center for Biomolecular and Cellular Structure (Protein Communication Group) makes IBS the right environment for the development of this research proposal.

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

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

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

[7] 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).

[8] 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.

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

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

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

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

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

[14] 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.

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

[16] Li, Ziyao, et al. "Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold." *bioRxiv* (2022).