AI-assisted Design of virus-binding proteins for the International Genetically Engineered Machine comptetion

Tony MAKDISSY* Amir PANDI† Ariel LINDNER‡ Ernest MORDRET§
Helena SHOMAR¶

August 2023 – November 2023

Abstract

Contains in clear language the problem statement, the indication of methodology, the main findings and the principal conclusion.

1 General Context

This report details my involvement in the Paris-Bettencourt team's collaborative participation in the International Genetically Engineered Machine (iGEM) contest [1]. The iGEM competition, held annually, is a globally recognized Synthetic Biology competition, uniting participants across three distinct age groups: high school, undergraduate, and graduate students from all over the world [2]. The topics covered by the competition are divided into 15 themes called villages [3]. Paris-Bettencourt project, named "Lubritect", was part of the Therapeutics village.

Lubritect is designed to be an innovative solution that combines mucin-based hydrogel with AI-generated protein structures, aiming to reduce the transmission of sexually transmitted infections (STIs). This approach leverages *de novo* protein design for versatility against various pathogens.

Paris-Bettencourt team is hosted by Learning Planet Institute (LPI), consisting of 7 LPI students, and 3 Non-LPI students. Each has a specific role like wet-lab, dry-lab, or human practices ...etc. My main role in this project was generating and *in-silico* testing of new protein structures to bind to the targets of interest. The team is supervised by 4 supervisors: Ariel Lindner, Ernest Mordret, Helena Shomar, and Amir Pandi [4]. Should I talk more about the supervisors ?

I should state clearly that in this report I will only talk about my work for *de novo* protein design and not about the other parts of the project. with mentioning a bit about the overall process.

*Learning Planet Institute
†Affiliation and role
‡Affiliation and role
§Affiliation and role
¶Affiliation and role

1.1 To delete

1.1.1 Task description

Describe what your laboratory, company, or institution is doing in general. Describe precisely what your team does and what their expertise is in a few lines. (don't copy-paste a boiler). Who are the key people you worked with and how did they acquire their expertise (mention their degrees, and/or professional experience...). Mention any other people you or your team collaborated with during your internship (limited to the particular topic you worked on). Add 3-4 lines on your integration to the institution and the work environment. Connection to the sustainable development goals (list the SDGs), if any.

2 Introduction

De novo Should I capitalize the 'D" in de novo? protein design is the field of science that addresses the fundamental question: "Is our knowledge of the principles of folding and function sufficient to design proteins from scratch?" [5] or, in broader terms, "Are natural proteins special? Can we do that?" [6].

2.1 History of *de novo* protein design

According to Korendovych and DeGrado [5], de novo protein design has evolved through several stages. In the nascent days of *de novo* protein design, researchers manually crafted protein structures. A landmark achievement occurred in 1983 when Moser et al. successfully designed a DDT binder through manual intervention [7]. The field transitioned towards computational approaches, where proteins are designed using computers and guided by physicochemical principles of protein folding. One notable example is the protein designed by DeGrado, Regan, and Ho in 1987 [8]. In this groundbreaking work, the team successfully crafted a 4-helix homomultimer, with each helix comprising 16 residues. The helices were strategically composed, featuring Leucine for hydrophobicity in the inner lumen of the helix, and Glutamic acid and Lysine for the external region of the helix. Additionally, Glycine was employed to disrupt the helix. This meticulous design strategy significantly reduced the potential residues' composition from 20^{16} (approximately $6.5 * 10^{20}$) possibilities to fewer than a thousand, thereby narrowing the design space significantly. Later on, in the early 2000s, with the accumulation of a vast repository of crystallized protein structures, marked the advent of fragment-based and bioinformatics-informed methods. Notably, this era saw the design of a TOP7 protein, incorporating fragments from the Protein Data Bank to construct a structure not observed in nature [9].

2.2 advancements in Molecular and Computational Biology

The advent of accurate protein structure prediction tools exemplified by AlphaFold2 [10], RoseTTAFold [11], and ESMFold [12], significantly impacted *de novo* protein design. These tools facilitated in-silico testing of designed protein structures, reducing the need for extensive experimental validation and enabling exploration beyond natural sequences. Along with the advancements in computational capabilities, DNA synthesis and high-throughput screen-

ing methods have accelerated *de novo* protein design. Recent achievements include the creation of mechanically coupled axle-rotor proteins by a team from the Baker lab, University of Washington [13], where the team discovered a wide variety of designs aided by computational simulations.

2.3 Machine Learning in *de novo* protein design

Despite these advancements, the field is limited by the vast number of possible protein structures. Machine learning (ML) approaches aim to overcome this challenge by training models to design proteins with specific structures or functions. Message Passing Neural Networks (MPNNs), exemplified by ProteinMPNN [14] developed by Baker lab, play a crucial role in predicting amino acid sequences starting from a given structure. In March 2023 Baker Lab published RFdiffusion, a successor of ProteinMPNN, which is a diffusion model trained on protein sequences and structures from Protein Data Bank (PDB) with structures generated by RoseTTAFold and AlphaFold2. RFdiffusion stands out as an innovative tool for *de novo* protein design. It allows for the generation of new protein sequences based on specified constraints such as sequence length, binding properties, and amino acid sequences [15]. It does so by first predicting a suitable structure for the given constraints, then generating a sequence that folds into the predicted structure using ProteinMPNNs. The field of *de novo* protein design also welcomes other ML approaches, such as RosettaSurf [16], which utilizes a surface-centric computational design approach, unlike ProteinMPNNs, which are position-centric (i.e. try first to predict the position of the amino acids and allosteric angles).

2.4 Lubritect project

Our team decided to go on a search to design protein binders aimed at immobilizing STIs' proteins. within a mesh network of a gel, ultimately creating an anti-STI lubricant to add an extra layer of protection alongside existing methods like condoms. Lubritect was designed as an answer to the alarming statistics regarding STIs, with high incidence (1 million new sexually transmitted infections every day), prevalence (80% of sexually active individuals will acquire human papillomavirus by 45) and disease burden (82,000 deaths in 2019 from hepatitis B) [17].

To achieve this goal, we used RFdiffusion to generate protein sequences. We made this decision because of the impressive history of Baker lab (creating TOP7. RoseTTAFold, ProteinMPNN, the design of self-assembly mechanically coupled axle-rotor proteins, and many more). Also we found a lot of resources and tutorials on how to use RFdiffusion (online seminars about RFdiffusion Link and ProteinMPNN Link). along with clean and well-documented code on GitHub Link.

NOTE: I should mention that all the discussed tools in this report are open-source and/or free to use. Also, all the references used are free to access.

2.5 To delete

2.5.1 Task description

Past research or work in the field. Define precisely what are the questions, objectives and tasks you were given. Connect them with a scientific or technical context. What are the approaches

generally used to solve the problem? (Reference them) What are the underlying assumptions or hypotheses, if any? Question raising. What are their limitations? Is there a gap? Which one? Purpose of the present research or work and experimental strategy chosen to address your scientific question Literature review. It is an echo of the points raised in the introduction. You can reference findings and describe the state of the art.

2.5.2 Key points

- Past research: how people used to find new protein structures. Needs a bit of research! HARD.
- Now you realize that this is super slow ...
- Mention the idea behind RFdiffusion. Maybe talk a bit about the lab's previous work (in one to two sentences).
- Talk about Diffusion models. ChatGPT, DALLE-E ... etc.
- Talk about MPNNs. What is it used for? How is it a better alternative to CNNs?
- Now say that we hoped to test this new tool on a real problem. And that's why we chose to work on the iGEM project. It's new so not a lot of research has been done on it.
- Mention other available tools like the one from EPFL from the lab you tried to apply to once. Also the new Atom Diffusion model
- One big limitation is that since there's not a lot of research done on this topic, there are not a lot of examples available. So we had to come up with our own protocols, NOTE: you can show search terms on academic search engines for words like "*de novo* protein design" and "protein structure prediction" and show that there are not a lot of results.
- Now talk that we had limited resources so we had to develop tools that can run in parallel, and run on specific environments like Google Colab and that can be run on a GPU and things like that.
- what about HDOCK, or the Codon Optimization tool?

3 Methods

3.1 Design Considerations

The application of RFdiffusion was facilitated by the Google Colab adaptation developed by Sergey Ovchinnikov [18], allowing seamless utilization of RFdiffusion and Google's GPUs. However, due to its inherent lack of accuracy, multiple sequences needed testing to identify those binding to the targets of interest. This computationally intensive process, with time complexity proportional to $O(N^2)$ (where N is the number of residues), prompted us to reduce the target protein size. This reduction, recommended by RFdiffusion developers [19], required defining a tradeoff between the number of sequences and their length, as there was no clear guidance on the optimal binder size for this relatively new tool.

In an effort to maximize efficiency with limited lab resources, an orthogonal approach was employed. In-silico docking experiments using the HDOCK binding algorithm were conducted to filter and prioritize sequences more likely to bind to the target of interest.

These considerations guided the development of the following pipeline (refer to Figure 3.1):

1. Inspecting the target protein using ChimeraX to identify specific regions of interest.
2. Selecting plausible regions of interest.
3. Truncating the target protein around the identified regions.
4. Generating sequences using RFdiffusion.
5. Filtering results through in-silico docking experiments using HDOCK.

The pipeline outputs a list of sequences with higher likelihoods of binding to the target protein, which are subsequently subjected to experimental testing after codon optimization.

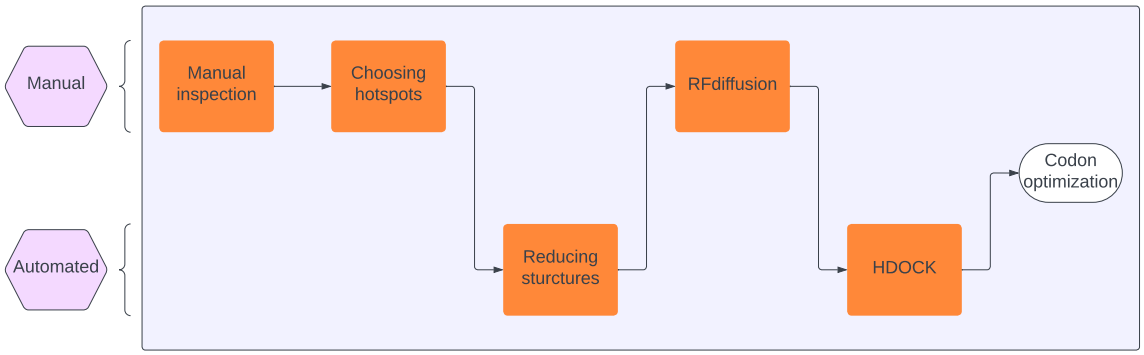

Figure 1: General organization graph of the iGEM project.

3.2 Target Protein Inspection and Hotspot Selection

Commencing with the delineation of a hotspot, we define it as a specific residue (amino acid) on the target protein that we aim for the designed binders to effectively bind to. Despite the potential ambiguity of this term, our adoption of RFdiffusion's terminology guides our terminology usage.

Our strategy for selecting potential binding sites involved both manual inspection and the identification of hotspots based on specific rules of thumb. We developed these rules of thumb through a combination of manual exploration and adjustments based on RFdiffusion guidance.

Alexis Courbet, a former member of the Baker lab, significantly contributed to shaping our guiding principles for binder design. Leveraging his expertise, we established a set of criteria to guide our manual inspection process, focusing on identifying residues that align with the following:

- Target regions exhibiting high solvent accessibility.

- Regions characterized by noticeable hydrophobic patches.
- Grooves within the target structure.

In accordance with the established criteria, we opted for Human Papillomavirus (HPV) as one of our targets due to its nature as a naked virus (non-enveloped) /citemorshed2014human. At the initial stages of our exploration with RFdiffusion, the potential interference of glycoproteins with our study was uncertain. To address safety concerns, our strategy involved expressing the proteins in an alternative vector rather than utilizing the actual virus.

The selection of Bacteriophage T4 was motivated by its availability in our laboratory and the associated safety considerations. Unlike HPV, using the virus in its native form was feasible and, with HPV proteins, presented a robust proof of concept for our study.

To select hotspots, ChimeraX functionalities such as "hydrophobic" and "electrostatic" were utilized, generating color-coded surfaces to visualize hydrophobic to hydrophilic and negative to positive surfaces on the protein structure. Ultimately, ChimeraX proved to be a valuable open-source tool for implementing our guidelines, enabling the effective identification of essential hotspots.

3.3 Running RFdiffusion

Due to the substantial size of capsids, which are intricate protein assemblies, computational limitations necessitated the pruning of structures around identified hotspots to prevent out-of-memory errors. To streamline this process and minimize potential errors and biases introduced by manual steps, a custom code was developed using the BioPython library. The code automates the reduction of structure based on predefined criteria, ensuring a more efficient and standardized approach.

These reduced sturcte are then manually used as input for RFdiffusion. To comply with legal restrictions, local code execution or connecting a local session to Google Colab is prohibited, as outlined in the Google Colab disallowed activities. Therefore, the results of the structure-reducing scripts were manually uploaded to the Google Colab Virtual Machine. If access to a robust local GPU were available, further automation of the entire process, after choosing the hotspots.

3.4 HDOCK: an orthogonal validation approach

After completing the previous steps, the number of generated sequences reached the order of thousands. However, the practical constraints of our team's budget made it unfeasible to test all these sequences experimentally. To prioritize potential designs and narrow down the candidates for experimental validation, we sought an orthogonal approach for assessment.

Through extensive research, we chose to employ HDOCK, an *in-silico* binding tool known for its consistent performance in the CASP-CAPRI [20] competition over the years. HDOCK utilizes a fast rigid-body docking algorithm, which can be parallelized on a server to enhance throughput, enabling the scanning of numerous potential designs within a reasonable timeframe.

We applied an *in-silico* filtration steps using a customized code that I develpoed to parallelize the process and get all the needed statistics. Only the sequences deemed most likely to exhibit

binding were retained for further experimental validation.

3.5 To delete

3.5.1 Task description

(1-2 pages with figures if relevant): (Methods is a section valid for any internship, not only research-oriented, experimental or theoretical. You did something according to a method, which is what you should understand and present here.) Present in detail the tools, techniques and methods you have used and why (giving a list of library's or equipment's name is not a presentation). Describe the scientific and/ or technical background with clear explanations, references, equations and/ or schematics or pictures.

3.5.2 Key points

- This section should be prttey consise in terms of word, but full of screenshots about the used tools.
- Here I can totally start pointing out to the graph where I have the entire pipline.
- Talk about RFdiffusion. How it works, talk about Sergey's Google Colab adaptation.
- Talk BRIEFLY about ChimeraX. and how you used it to find the targets of interest.
- Talk about different docking tools. Also the two contests for docking.
- Mention how HDOCK scored pretty good in these ones, and also how is it free to download.
- Talk about parrallel computing, and talk about parrallel computing using Python. and how you used this to further speed up the process of docking.

4 Results

After manually inspecting the following proteins:

- HPV major capsid protein (PDBID 7kzf)
- Bacteriophage T4 capsid (PDBID 7vs5)
- Bacteriophage T4 Long-Tail (PDBID 2xgf)
- GFP protein (PDBID 5b61)

I picked around 50 different sets of hotspots. Each set contains 2 to 4 hotspots. Point to the table. Each set of hotsopt will be called a "run" from now on.

Each run can produce $S * Q$ sequences, where S is the number of structure backbones that RFdiffusion algorithm should predict in the first step, and Q is the number of sequences generated per structure backbone using the NPMM step. generate

4.1 To delete

4.1.1 Task description

(2-3 pages with figures): Precisely describe the results of your work with clear explanations of the analysis, schematics and figures. Describe what worked, what didn't and why.

4.1.2 Key points

- Here there's two catches:
  - I have generated data for two targets but only one was tested.
  - I have really results that can be described totally in few sentecs.
- Talk about how many structures you generated.
- Talk about the tendecy of short ones to have repeated sequences. Come up with some metrics to measure this.
- just come up with some metrics to fill the page.
- Talk about the docking results. How good, bad they were.
- Talk about the i_PAE score, why you didn't use it from the begining?
- Show some nice graphs about the docking results. with color coding everhting, and the ones picked.
- Ask Louis, Momo, and/or Avi for help on the wet lab part (it should be a small subsection).
- There's a side hustles, a result should be visually appealing. So talk how I used ChimeraX to generate some cool animations. Talk about other new tools in development like the one I found during a lab meeting at Jussieu (I don't if the name is correct).

4.1.3 Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

5 Discussion

5.1 To delete

5.1.1 Task description

(1 page): Review findings. Discuss outcomes. Do your results make sense ? Evaluate them Do they provide elements towards solving your problem ? Which ones ? Do they open up new questions (scientific or technical).

5.1.2 Key points

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

5.1.3 Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

6 Conclusion

6.1 To delete

6.1.1 Task description

($\frac{1}{2}$ to 1 page maximum): Conclude regarding the missions and tasks you were given and the results you obtained . Mention if they will be used by your team. What are the limitations of your results ? What are the future directions (questions, implementation...) ?

6.1.2 Key points

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

6.1.3 Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

7 Acknowledgments

7.1 To delete

7.1.1 Task description

Also conclude regarding your work at the institution, your integration to the team, what you brought, what you have learned, what you need to improve.

7.1.2 Key points

- My colleages.
- Minc lab.
- two people whom I emailed for help.

- Others.

7.1.3 Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

References

1. IGEM Foundation. *iGEM foundation Main page* `https://igem.org/` (2024).
2. IGEM Foundation. *iGEM competition About page* `https://competition.igem.org/participation/introduction` (2024).
3. IGEM Foundation. *iGEM competition Villages page* `https://competition.igem.org/participation/villages` (2024).
4. IGEM Team, P. B. *Paris Bettencourt Team page* `https://2023.igem.wiki/paris-bettencourt/team` (2024).
5. Korendovych, I. V. & DeGrado, W. F. De novo protein design, a retrospective. *Quarterly reviews of biophysics* **53,** e3 (2020).
6. Hecht, M. H., Zarzhitsky, S., Karas, C. & Chari, S. Are natural proteins special? Can we do that? *Current Opinion in Structural Biology* **48,** 124–132 (2018).
7. Moser, R., Thomas, R. M. & Gutte, B. An artificial crystalline DDT-binding polypeptide. *FEBS Letters* **157,** 247–251 (1983).
8. DeGrado, W., Regan, L. & Ho, S. *The design of a four-helix bundle protein* in *Cold Spring Harbor symposia on quantitative biology* **52** (1987), 521–526.
9. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *science* **302,** 1364–1368 (2003).
10. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).
11. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876 (2021).
12. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* **2022,** 500902 (2022).
13. Courbet, A. *et al.* Computational design of mechanically coupled axle-rotor protein assemblies. *Science* **376,** 383–390 (2022).
14. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378,** 49–56 (2022).
15. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620,** 1089–1100 (2023).
16. Scheck, A. *et al.* RosettaSurf—A surface-centric computational design approach. *PLOS Computational Biology* **18,** e1009178 (2022).
17. IGEM Team, P. B. *Paris Bettencourt Project page* `https://2023.igem.wiki/paris-bettencourt/description` (2024).
18. Ovchinnikov, S. *RFdiffusion Google Colab notebook* `https://colab.research.google.com/github/sokrypton/ColabDesign/blob/main/rf/examples/diffusion.ipynb#scrollTo=TuRUfQJZ4vkM` (2024).

19. RosettaCommons. *RFdiffusion GitHub repository* `https://github.com/RosettaCommons/RFdiffusion` (2024).

20. Institute, E. B. *CASP-CAPRI About page* `https://www.ebi.ac.uk/pdbe/complex-pred/capri/casp-capri/` (2024).

8 Annexes

8.1 To delete

8.1.1 Task description

(3 pages maximum, if needed) Add experimental details, code, algorithms, figures, pictures and any additional information we need to understand your work. Links welcome. Organize them freely but clearly.

8.1.2 Key points

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

8.1.3 Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

9 What else?

Data. Why do you need them ? How and when were they collected, stored, accessed, selected and sorted ? Describe the data you used (information they contain, length, format, database,...) . Describe their limitations. Precise the legal and ethical context.

Implementation. Describe the operation pipelines with clear explanations and schematics. What is your detailed contribution here (experiment, code, algorithm, operation...) ? What have you implemented (and modified) ? How have you implemented it ?