Learning Planet Institute

Université Paris Cité

Master AIRE Life Sciences

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

Tony MAKDISSY

Team Members:

Milena MILOVANOVIC, Avishkar JADHAV, Louis ELVERSTON, Mostafa ELRAIES, Joann LEVAY, Laure Mourgue D'ALGUE, Dakshayani PINNINTI, Stasa RAKOCEVIC, and Hritika KATHURIA

Supervisors:

Ariel LINDNER, Ernest MORDRET, Helena SHOMAR, and Amir PANDI

July 2023 - November 2023

Contents

Abstract

The field of *de novo* protein design has witnessed significant advancements, evolving from manual crafting to sophisticated computational methodologies. This report details an exploration into de novo protein design, specifically focusing on the computational pipeline developed for generating potential protein binders. Leveraging tools like RFdiffusion, HDOCK, and ChimeraX. A theoretically automated workflow was employed to inspect target proteins, identify hotspots, and validate sequences through pull-down assays. Despite challenges in automating hotspot identification and GPU-dependent execution, the pipeline demonstrated promising results. Three sequences exhibited positive binding to the T4 bacteriophage, showcasing the feasibility of this approach. The success of this proof-of-concept provides a foundation for scaling up the design and exploring more complex protein structures.

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. Lubritect targeted 5 sustainable development goals:

- SDG3 (Good health and wellbeing)
- SDG4 (Inclusive and equitable quality education and promote lifelong learning opportunities for all)
- SDG5 (Achieve gender equality and empower all women and girls)
- SDG8 (Decent Work and Economic Growth)
- SDG12 (Responsible Consumption and Production)

All while trying to follow the Free and Open-Source Software (FOSS) philosophy [4] by using free and open-source tools and making sure that all our codes and results are available via the iGEM website.

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. With 4 supervisors who oversaw the project and offered scientific and emotional help [5]. My main role in this project was generating and *in-silico* testing of new protein structures to bind to the targets of interest.

2 Introduction

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?" [6] or, in broader terms, "Are natural proteins special? Can we do that?" [7].

2.1 History of *de novo* protein design

According to Korendovych and DeGrado [6], *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 [8]. 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 [9]. 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 [10].

2.2 Advancements in Molecular and Computational Biology

The advent of accurate protein structure prediction tools exemplified by AlphaFold2 [11], RoseTTAFold [12], and ESMFold [13], 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 screening 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 [14, 15], 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 [16] 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 [17]. 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 welcomed other ML approaches, such as RosettaSurf [18], 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) [19].

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) [15]. Also, we found a lot of resources and tutorials on how to use RFdiffusion (online seminars about RFdiffusion and ProteinMPNN). along with clean and well-documented code on GitHub.

2.5 Free and Open-Source Software (FOSS)

It is noteworthy that every tool utilized in this project adheres to the principles of open-source and/or free usage. Our team advocates for the openness and accessibility of scientific endeavors to a broader audience. We exclusively employed tools that align with the tenets outlined in "What is Free Software?" [4]. A comprehensive list of all the tools considered throughout the project is provided in Table 1. The entirety of our code is accessible on GitHub (`https://github.com/Tony-Makdissy/iGEM_2023`). Additionally, all the papers referenced in this report are free to access without a charge.

3 Methods

3.1 Design Considerations

The utilization of RFdiffusion was made possible through the Google Colab adaptation developed by Sergey Ovchinnikov [20], facilitating seamless access to RFdiffusion and Google's GPUs. However, owing to its inherent lack of accuracy, multiple sequences required testing to identify those binding to the targets of interest. This computationally intensive process, with a time complexity proportional to the square of the number of residues $O(N^2)$ (where N is the number of residues), prompted the need to reduce the target protein size, as recommended by RFdiffusion developers [21]. We also had to limit the length of the produced sequences. To navigate this, a tradeoff was defined between the number of sequences and their length.

Given the absence of clear guidance on the optimal binder size for this relatively new tool, we experimented with multiple lengths and selected what we considered the most suitable.

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 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.

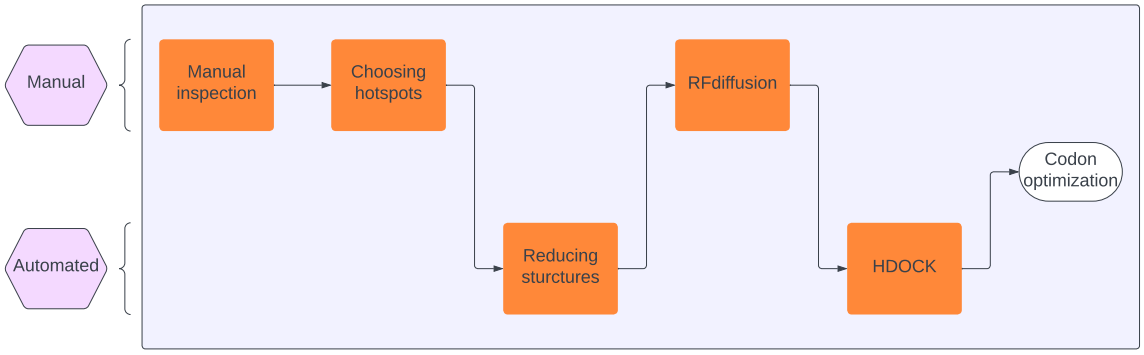


Figure 1: General organization graph of the iGEM project.

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 manual inspection to identify potential 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 [21] and previous studies in this field [22].

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.

Following the established criteria, we opted for Human Papillomavirus (HPV) as one of our targets due to its nature as a naked virus (non-enveloped) [23]. At the initial stages of our exploration of 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.

We also decided to use Bacteriophage T4 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 identify hotspots, we employed ChimeraX functionalities such as "hydrophobic" and "electrostatic" to produce an informative depiction of protein surfaces. Figure 2 illustrates the "hydrophobic" surface of the T4 bacteriophage capsid (Protein Data Bank id: 7vs5). The selected residues (hm19, hm22, hk48, hk49, and fh281) are located within a hydrophobic groove exposed to the solvent, aligning with our design rules.

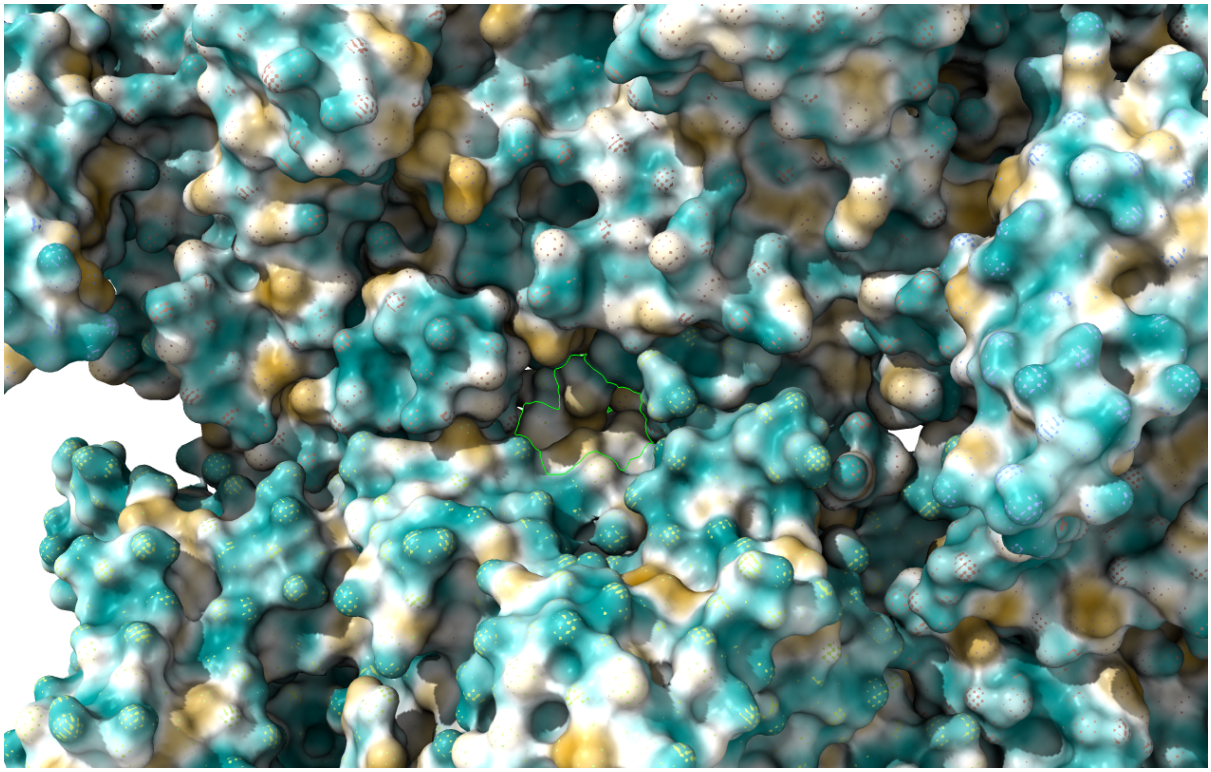


Figure 2: Hydrophobic surface of the T4 bacteriophage capsid, generated using ChimeraX. The hydrophobic surface is colored in yellow, while the hydrophilic surface is colored in blue. The selected residues are where we successfully designed two binders.

3.3 Running RFdiffusion

To reduce the target protein size, I developed a simple Python script, using the Biopython library [24]. The script takes as input a PDB file (or id), a list of hotspots and a radius (in

Angstroms). It then outputs a new PDB file containing only the residues within the specified radius of the hotspots. The script is available on the project's GitHub repository (https://github.com/Tony-Makdissy/iGEM_2023). We wanted to automate this step, in order to minimize the manual steps and the potential errors and biases introduced.

These reduced structures 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. Thankfully this step would not produce human biases, but it caused a significant delay in the process.

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 [25] known for its consistent performance in the CASP-CAPRI [26] competition over the years. HDOCK utilizes a Fast-Fourier-Transformation-based docking algorithm, making it a fast rigid body docking tool. Additionally, HDOCK can be used locally, so I developed a Python script to parallelize the process and get all the needed statistics. The script utilizes the "multiprocessing" Python built-in module [27].

By the end of this step, only the sequences deemed most likely to exhibit binding were retained for further experimental validation.

4 Results

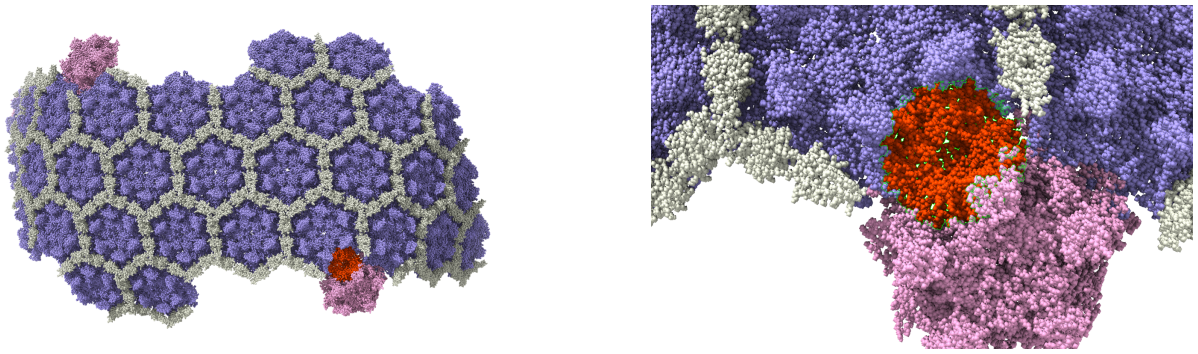
4.1 RFdiffusion Results

After manually inspecting four proteins:

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

I selected approximately 50 different sets of hotspots, each containing 2 to 6 hotspots, termed as a "run." Refer to the Supplementary Data for the complete table of manual hotspot choices can be found in the folder "Supplementary Data" under the name "Manual choices for hotsopts.csv"

In addition to specifying parameters such as hotspots and the number of sequences, I also uploaded the reduced structures to the Google Colab Virtual Machine (Figure 3).



(a) Original structure image.

(b) Structure image zoomed four times.

Figure 3: Comparison of the original image and a zoomed image of the structure. The full structure (PDBID 7vs5) has 102,943 residues, while the reduced structure (in orange) has only 514.

Each run can generate $S*Q$ sequences, where S represents the number of structure backbones predicted by the RFdiffusion algorithm in the initial step, and Q is the number of sequences generated per structure backbone using the NPMM step.

By the project's conclusion, around 1500 sequences were generated, posing a challenge for experimental testing. To address this, a filtration step was employed using HDOCK.

4.2 HDOCK Results

Following the sequence generation, we faced a substantial pool of thousands of sequences, necessitating a method to streamline experimental validation. An in-silico binding assay was chosen, drawing insights from the CASP-CAPRI competitions, renowned for assessing protein structure predictions.

After scrutinizing top-ranking servers in the CASP-CAPRI competition, HDOCK emerged as our choice, driven by its notable attributes:

- **Efficiency:** HDOCK utilizes a high-speed rigid-body docking algorithm grounded in Fast Fourier Transform (FFT).
- **Local Executability and Parallelizability:** In contrast to many web-hosted tools, HDOCK can be locally downloaded and executed, with docking processes parallelizable using the multiprocessing Python library.
- **Proven Performance:** HDOCK secured the 1st rank in CASP-CAPRI11 and continued to excel in subsequent competitions.

The selection criteria aimed at choosing best-scoring sequences while ensuring representation from various runs (Figure 4), enhancing the comprehensiveness of subsequent experimental validation.

One limitation of HDOCK is its lack of experimentally calibrated scores. Scores are not comparable between different target proteins, leading us to compare scores only within the same target. The scores themselves lack a universally meaningful interpretation.

4.3 Experimental Validation through Pull-Down Assays

To confirm binding capabilities, pull-down assays were performed on 60 generated sequences. Three showed positive binding to T4 bacteriophage (Figure 4), validating the computational design (see Table 2).

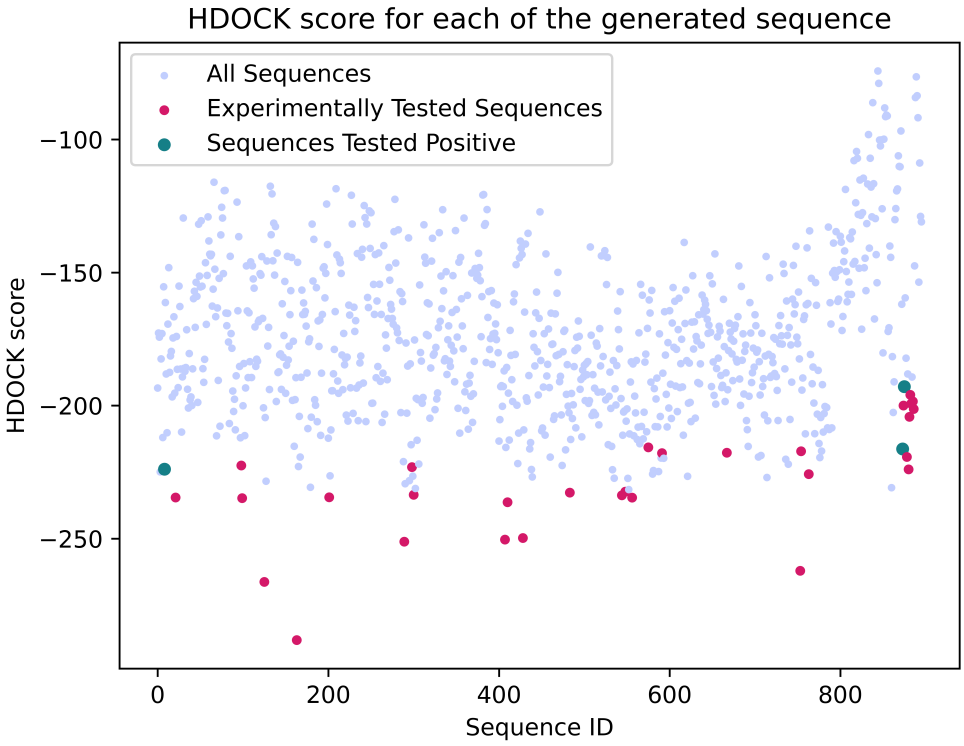


Figure 4: Universal scores against Sequence ID (as used in our tables). Red dots represent all tested sequences, while green dots highlight the three sequences exhibiting positive binding to T4 bacteriophage.

For infectivity assessment, a pull-down phage retention assay was adapted. His-tagged proteins were incubated with T4 phage, and fractions were collected through an affinity purification column. Phage concentrations were quantified using a plaque assay. This setup, inspired by existing literature, provides insights into the designed protein's impact on viral infectivity.

I did not contribute to this step; credit goes to my friends Avi, Mostafa, and Louis for handling wet lab activities.

4.4 Further Analysis

After validating three sequences, a BLAST search yielded no matches with known proteins under default parameters. This supports RFdiffusion's potential for generating novel sequences.

A video demonstrating a hypothetical binding scenario is available in the "Supplementary Data" folder as "Binding_video.mp4." The binder, designed based on an AlphaFold2 prediction, is expected to bind to the area shown in Figure 2 (T4 bacteriophage capsid structure from PDB ID 7vs5). The code used to generate the video is available on the following GitHub repository: https://github.com/Tony-Makdissy/iGEM_final_presentation

4.5 iGEM Results

Despite achieving promising results in our project, we faced a setback as we failed to complete the required paperwork within the specified time, ultimately leading to our disqualification ☹.

5 Discussion

Our study successfully demonstrates the design and experimental testing of several proteins that do not exist in nature within an unusually short timeframe. This achievement is a testament to the potential of our approach in expanding the repertoire of available proteins.

Despite the success, challenges remain in fully automating the process. Computational resources, in theory, could support automation; however, manual inspection remains a necessity. A profound understanding of differential geometry is crucial for defining hydrophobic grooves and other geometric features. This manual intervention represents a current bottleneck in the workflow.

Our work lays a robust foundation for further research in this field. While the number of positively binding proteins is promising, achieving statistical significance requires more comprehensive studies. The sequences demand further investigation into their true structure and binding kinetics, as these structures were only assayed for binding. Additionally, optimizing the code remains an avenue for future improvements.

Moreover, it is crucial to acknowledge that the scope of our work extends beyond Lubretict. The protein generation and *in-silico* binding assays are blind to the overall scope. This pipeline holds potential beyond our current application and can be adapted for the generation of binders in diverse contexts.

De novo proteins require extra caution in their potential applications, especially regarding their interactions with living organisms, such as humans. The implications of introducing these novel sequences into biological systems necessitate comprehensive studies to understand their safety and efficacy. Additionally, even though a cell can produce a given sequence, the efficiency of production and the ability to fine-tune these sequences remain unknown. Ongoing studies in this field aim to shed light on these aspects, contributing to a more comprehensive understanding of the practical implications of de novo protein design.

6 Conclusion

In conclusion, our research not only marks a significant step forward in protein design and testing but also outlines a versatile pipeline with applications that stretch beyond the immediate scope.

Collaborative opportunities, practical implications, and the potential for future advancements further underscore the significance of our work in the broader scientific landscape.

Acknowledgments

We extend our gratitude to Minc's Lab, Institute Jacques Monod, and IFB - Institut Français de Bioinformatique for server access crucial to our computational tasks.

Special thanks to Alex and Sergey for invaluable assistance in protocol design, and to the Life Sciences Institute (LPI) team for continuous support, and fostering innovation.

Data Availability

As the team is planning to publish the results of this project in a peer-reviewed journal, the data is not available yet. However, the code is available on GitHub (https://github.com/Tony-Makdissy/iGEM_2023). Along with the successfully validated sequences (Table 2) and the list of run's parameters (see Supplementary Data).

A full list of all the runs' sequences with their corresponding statistics and experimental data will be available after the publication of the paper.

For this report, you can access the LaTeX code used to create it following the link https://github.com/Tony-Makdissy/internship_report_2024_autumn which contains all the code and data used to create this report.

Supplementary Data will be provided as a separate file along with the report and can be accessed on the GitHub Repository.

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. GNU. *GNU, "What is Free Software?"* https://www.gnu.org/philosophy/free-sw.html (2024).
5. IGEM Team, P. B. *Paris Bettencourt Team page* https://2023.igem.wiki/paris-bettencourt/team (2024).
6. Korendovych, I. V. & DeGrado, W. F. De novo protein design, a retrospective. *Quarterly reviews of biophysics* **53,** e3 (2020).
7. 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).
8. Moser, R., Thomas, R. M. & Gutte, B. An artificial crystalline DDT-binding polypeptide. *FEBS Letters* **157,** 247–251 (1983).

9. 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.
10. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *science* **302,** 1364–1368 (2003).
11. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).
12. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876 (2021).
13. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* **2022,** 500902 (2022).
14. Courbet, A. *et al.* Computational design of mechanically coupled axle-rotor protein assemblies. *Science* **376,** 383–390 (2022).
15. Lab, B. *Baker Lab Main page* https://www.bakerlab.org/ (2024).
16. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378,** 49–56 (2022).
17. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620,** 1089–1100 (2023).
18. Scheck, A. *et al.* RosettaSurf—A surface-centric computational design approach. *PLOS Computational Biology* **18,** e1009178 (2022).
19. IGEM Team, P. B. *Paris Bettencourt Project page* https://2023.igem.wiki/paris-bettencourt/description (2024).
20. Ovchinnikov, S. *RFdiffusion Google Colab notebook* https://colab.research.google.com/github/sokrypton/ColabDesign/blob/main/rf/examples/diffusion.ipynb#scrollTo=TuRUfQJZ4vkM (2024).
21. RosettaCommons. *RFdiffusion GitHub repository* https://github.com/RosettaCommons/RFdiffusion (2024).
22. Chen, J., Sawyer, N. & Regan, L. Protein–protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Science* **22,** 510–515 (2013).
23. Morshed, K., Polz-Gruszka, D., Szymański, M. & Polz-Dacewicz, M. Human papillomavirus (HPV)–structure, epidemiology and pathogenesis. *Otolaryngologia Polska* **68,** 213–219 (2014).
24. contributors, B. *Biopython Main page* https://biopython.org/ (2024).
25. Yan, Y., Zhang, D., Zhou, P., Li, B. & Huang, S.-Y. HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic acids research* **45,** W365–W373 (2017).
26. Institute, E. B. *CASP-CAPRI About page* https://www.ebi.ac.uk/pdbe/complex-pred/capri/casp-capri/ (2024).
27. Foundation, P. S. *Python multiprocessing Main page* https://docs.python.org/3/library/multiprocessing.html (2024).

Annexes

Table of free and open-source tools used in the project

Table 1: List of Open-Source Tools Used in the Project

Tool name	Links
RFDiffusion	GitHub Repository: https://github.com/RosettaCommons/RFdiffusion Google Colab Notebook: https://colab.research.google.com/github/sokrypton/ColabDesign/blob/main/rf/examples/diffusion.ipynb
HDOCK	Website: http://hdock.phys.hust.edu.cn/
AlphaFold2	GitHub Repository: https://github.com/google-deepmind/alphafold Google Colab Notebook: https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb
RosettaSurf	GitHub Repository: https://github.com/LPDI-EPFL/RosettaSurf
ESMFold	GitHub Repository: https://github.com/facebookresearch/esm Google Colab Notebook: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/ESMFold.ipynb
RoseTTAFold	GitHub Repository: https://github.com/RosettaCommons/RoseTTAFold Google Colab Notebook: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/RoseTTAFold.ipynb
ChimeraX	GitHub Repository: https://github.com/RBVI/ChimeraX Website: https://www.rbvi.ucsf.edu/chimerax/

Table of experimentally validated sequences

Table 2: Experimentally validated sequences and their corresponding target Protein Data Bank (PDB) IDs

Target PDBID	**Sequence**
2xgf	DLEALRAAIRAEADARAAAFVARPPLTPAERAALAA RLRARLAGRPDADARVAALRRLSPVAQLAERYRREA AERAAEVAALIPEGPEVAAYILQRANDAAATLRAAA
7vs5	MSNTLEQKIISARAVDADALLARVLARLDALAAEDPS HAAELAALRATIAEAQALAATAAPIPLRDLAIALRER ARALRAKDSAKNRRLVRLTDEAADLVRVLIAQALA
7vs5	MSSTLEQKIISSKAVDVEELLKRVLERLEELEKKDPKH KEKLEELRKKIEKALELAKTSTYVPLLDLAIELENEA QKLRGENAKENSELVRLTDEAADLVRAMISKALQ