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

Tony MAKDISSY*

August 2023 – November 2023

Abstract

De novo protein design, the art of creating functional proteins from scratch, has evolved over the years, transitioning from manual structure crafting to sophisticated computational approaches. This report details our exploration in *de novo* protein design, specifically focusing on the computational design pipeline developed to generate potential protein binders for immobilizing sexually transmitted infection (STI) proteins within a gel mesh network. Leveraging tools like RFdiffusion, HDOCK, and ChimeraX, we employed a theoretically automated workflow to inspect target proteins, select hotspots, and validate sequences through pull-down assays. Despite facing challenges in fully automating hotspot identification and GPU-dependent execution, our pipeline demonstrated promising results, with three sequences exhibiting positive binding to the T4 bacteriophage. The success of this proof-of-concept provides a foundation for scaling up the approach and exploring more complex protein designs.

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].

*Learning Planet Institute

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?" [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 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 [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 welcomed 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.

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?" [18]. 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, available here. However, it requires further refinement and documentation. Additionally, all the papers and documents referenced in this report are freely accessible without the need for platforms like Sci-Hub.

3 Methods

3.1 Design Considerations

The utilization of RFdiffusion was made possible through the Google Colab adaptation developed by Sergey Ovchinnikov [19], 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 [20]. We also had to limit the length of the produced sequences.

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	Website: URL needed
BioPython	Website: URL needed

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 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.

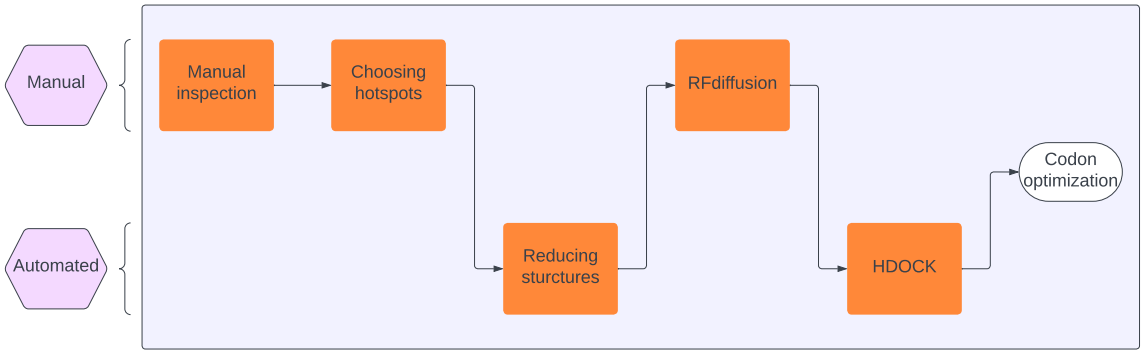

Figure 1: General organization graph of the iGEM project.

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 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 [20] and previous studies in this field [21].

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) [22]. 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.

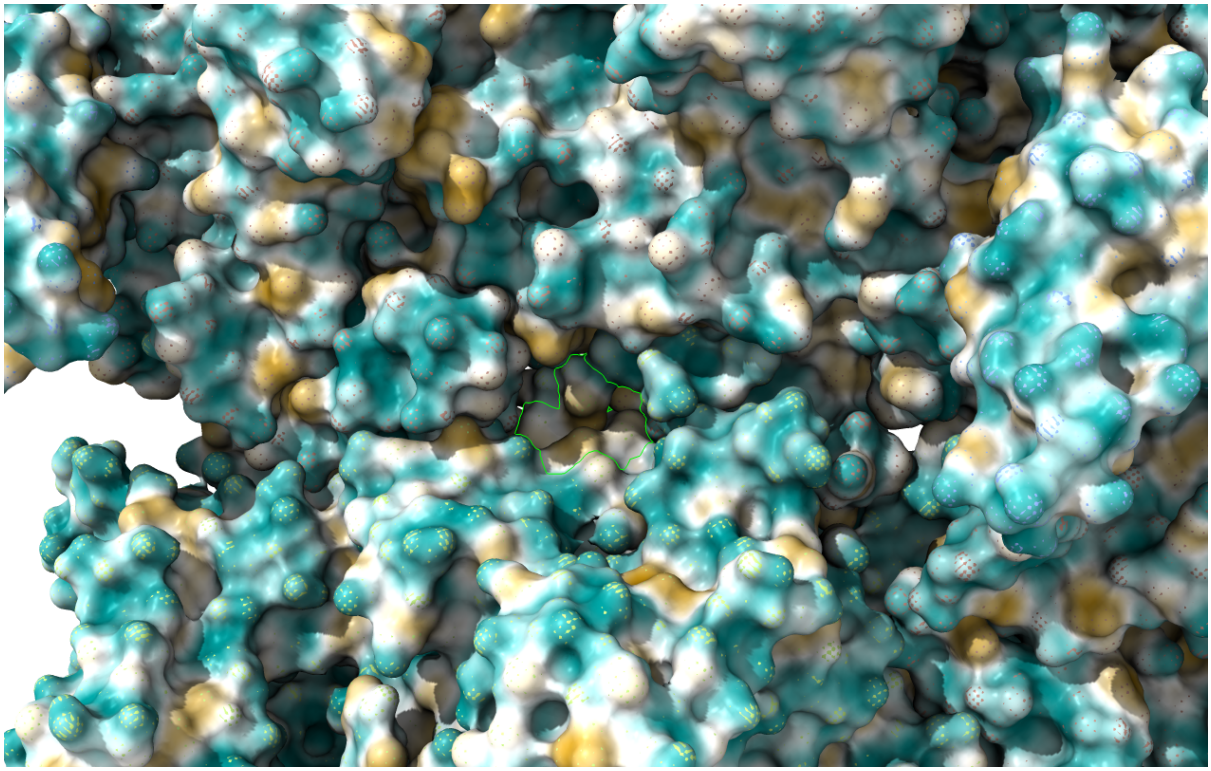

Figure 2: Hydrophobic surface of the T4 bacteriophage capsid, generated using ChimeraX. The hydrophobic surface is colored in yellow, while the hydrophilic surface is colored in blue. The selected residues are where we successfully designed two binders.

3.3 Running RFdiffusion

To reduce the target protein size, I developed a simple Python script, using Biopython library [23]. 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 GitHub here. We wanted 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 [24] known for its consistent performance in the CASP-CAPRI [25] 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 [26]. The code is available on GitHub here.

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

4 Results

After manually inspecting the following proteins:

- HPV major capsid protein (Protein Data Bank id: 7kzf)
- Bacteriophage T4 capsid (Protein Data Bank id: 7vs5)
- Bacteriophage T4 Long-Tail (Protein Data Bank id: 2xgf)
- GFP protein (Protein Data Bank id: 5b61)

I picked around 50 different sets of hotspots. Each set contains 2 to 4 hotspots. Each set of hotspots will be called a "run" from now on. A full table can be found in the attached table "Supplementary Data/Manual choices for hotsopts.csv"

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

By the end of the project, I have generated around 1500 sequences. Which are not feasible to test experimentally. So I had to filter them out using HDOCK.

4.1 HDOCK results

4.1.1 Introduction

Upon completing the preceding steps, we encountered a substantial pool of sequences, numbering in the thousands. The prospect of experimental testing for all these sequences posed a significant resource and time challenge.

To streamline and curtail the number of sequences for experimental validation, we sought an in-silico binding assay. A comprehensive analysis of results from the Critical Assessment of Protein Structure Prediction (CASP) and Critical Assessment of Prediction of Interactions (CAPRI) collaborations, particularly the CASP-CAPRI competitions, proved pivotal. CASP and CAPRI orchestrate these competitions, engaging participants to predict structures

of protein-protein complexes. Participants submit their predictions for evaluation, and rankings are made publicly accessible.

Following an exhaustive examination of the top-ranking servers in the CASP-CAPRI competition, our selection converged on HDOCK, driven by its commendable attributes:

- **Efficiency:** HDOCK employs 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. Furthermore, the docking process can be parallelized using the multiprocessing Python library.
- **Proven Performance:** HDOCK achieved preeminence by securing the 1st rank in CASP-CAPRI11 and has continued to play a pivotal role in subsequent competitions.

The selection process involved choosing the best-scoring sequences while endeavoring to include sequences from almost all the runs. This strategic approach ensured a comprehensive representation for subsequent experimental validation.

4.2 Experimental Validation through Pull-Down Assays

To experimentally confirm the binding capabilities of the generated sequences, pull-down assays were employed. Among the 30 tested sequences, three demonstrated positive binding to the T4 bacteriophage, affirming the effectiveness of the computational design approach in identifying potential binding sequences. In Figure 3, the universal scores of the generated sequences are depicted, with those exhibiting positive results highlighted in green.

5 Discussion

The success of our computational design pipeline, utilizing open-source tools and following a theoretically fully automated workflow, marks a significant achievement. The integration of RFdiffusion, HDOCK, and ChimeraX enabled the generation and prioritization of protein sequences with potential binding affinity.

Despite the advancements, achieving complete automation faces two challenges. Firstly, the reliance on a local GPU for RFdiffusion execution would enhance efficiency, reducing the manual steps involved in uploading results to Google Colab. Secondly, the identification of hotspots using manual rules of thumb remains a semi-manual step, requiring expertise in differential geometry.

However, the positive experimental results obtained from pull-down assays provide a robust proof of concept. The success in identifying binding sequences, especially with the constraints and limitations faced by the team, demonstrates the feasibility and potential scalability of the approach. Scaling up the pipeline, generating longer sequences, and testing them will be the logical next steps, guided by the promising outcomes observed in this initial phase.

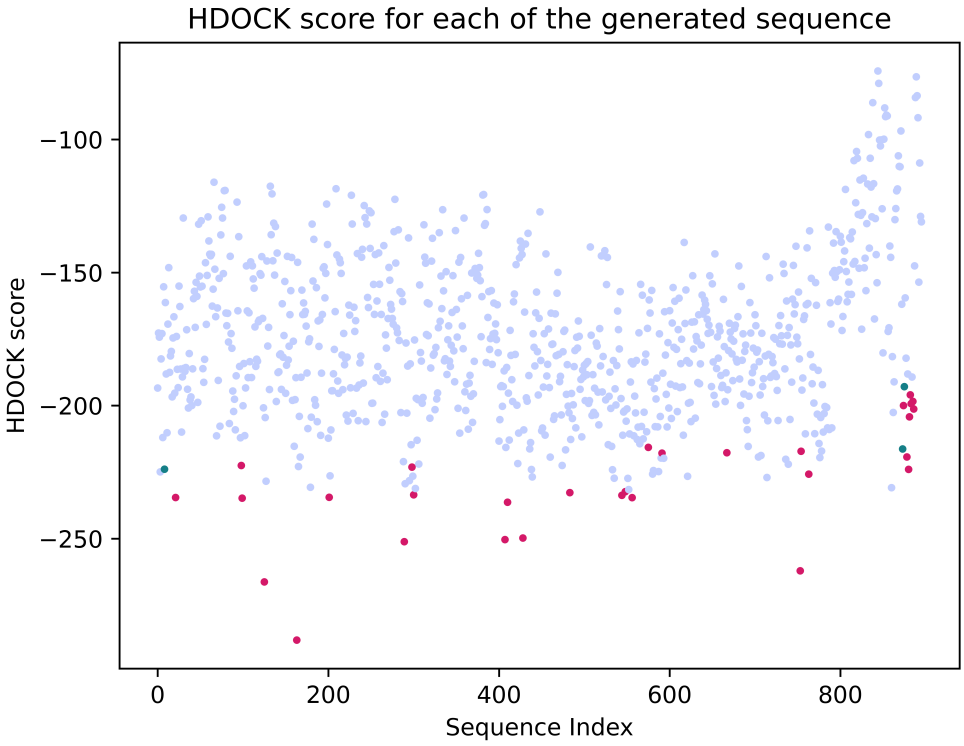

Figure 3: Graph showing the universal score against Sequence Global ID.

Acknowledgments

We extend our sincere gratitude to the Minc's Lab team and the Institute Jacques Monod for their generosity in providing access to their server, which played a crucial role in the execution of computational tasks for this project.

Special thanks to Alex and Sergey for their invaluable assistance during the protocol design phase. Their expertise and guidance significantly contributed to the development of an effective and robust computational pipeline.

We also express our appreciation to the entire Life Sciences Institute (LPI) team for their continuous scientific and emotional support throughout this endeavor. The collaborative spirit within the team fostered an environment conducive to exploration and innovation.

IFB

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. GNU. *GNU, "What is Free Software?"* https://www.gnu.org/philosophy/free-sw.html (2024).
19. Ovchinnikov, S. *RFdiffusion Google Colab notebook* https://colab.research.google.com/github/sokrypton/ColabDesign/blob/main/rf/examples/diffusion.ipynb#scrollTo=TuRUfQJZ4vkM (2024).
20. RosettaCommons. *RFdiffusion GitHub repository* https://github.com/RosettaCommons/RFdiffusion (2024).
21. 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).
22. Morshed, K., Polz-Gruszka, D., Szymański, M. & Polz-Dacewicz, M. Human papillomavirus (HPV)–structure, epidemiology and pathogenesis. *Otolaryngologia Polska* **68,** 213–219 (2014).
23. contributors, B. *Biopython Main page* https://biopython.org/ (2024).
24. 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).

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

26. Foundation, P. S. *Python multiprocessing Main page* https://docs.python.org/3/library/multiprocessing.html (2024).