BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Web application for protein structure prediction

Romana Ďuráčiová*

Abstract

The breakthrough in protein structure prediction came in 2021 with Google DeepMind's AlphaFold, which streamlined lab research using deep learning software. However, accessibility to this tool is a challenge for scientists who are not familiar with the command line. This issue can be addressed by developing a user-friendly interface connected to high-performance hardware for easy computation submission. The input of this work is a familiar solution that no longer suits the current technical, security, and design requirements, and the goal is to recreate a similar but improved solution. Making this technology accessible contributes to science in different ways, such as drug discovery in medicine and the design of novel enzymes in biotechnology.

*xdurac01@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

In biochemistry, knowing protein structures is crucial, but lab research can be costly and slow. Advancements in deep learning software offer a breakthrough, saving time and money while providing highly accurate results.

AlphaFold-based tools are command-line scripts that require complex installation and significant resources, making them less accessible for many researchers. A potential solution is to develop a user-friendly interface for easier computation.

Existing solutions offer such an interface, but most of their functionality is behind a paywall, or the computations are limited to a certain number per day or month (such as Neurosnap or AlphaFold Server). There is also an existing AlphaFold-based platform by the CERIT-SC research group that will serve as the basis of the new solution.

By leveraging the existing tools, the new platform will offer a high-quality experience in submitting protein prediction computations using a variety of models such as AlphaFold3 [1], AlphaFold2 [2], ColabFold [3], OmegaFold [4], and ESMFold[5] and an intuitive interface for scientists and researchers.

2. From primary to quaternary structure

The protein structures are organized into four hierarchical levels shown in Figure 1.

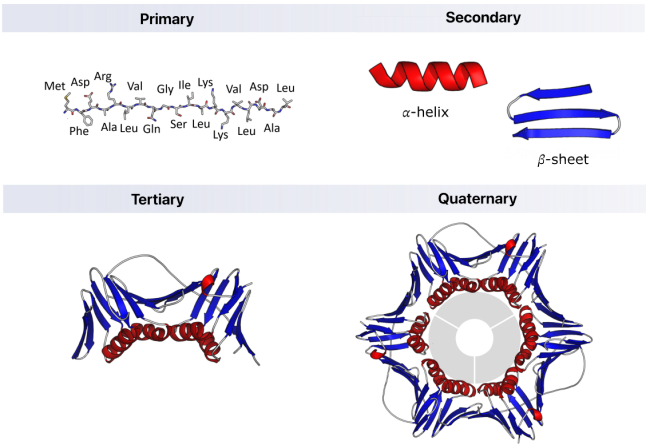

Figure 1. Protein Structure Levels. *Jmarchn, from Thomas Shafee, CC BY-SA 3.0, via Wikimedia Commons*

The **primary structure** of a protein refers to its linear sequence of amino acids and is the simplest form of the protein. **The secondary structure** consists of α-helices and β-sheets that give the protein structure and stability. The protein's unique three-dimensional shape is given by the **tertiary structure**, and a functional protein complex generally described as a **quaternary structure** occurs when two or more chains engage with each other and operate as a single functional unit [6].

2.1 Formal Representation of Protein Data

The computation begins by inputting a sequence of amino acids (the primary structure) in the FASTA for-

mat. The prediction tool uses this data to derive the protein's three-dimensional tertiary structure model.

Output formats vary by the tool: AlphaFold 2 generates a PDB file, while AlphaFold 3 provides a more detailed file in the MMCIF format. A visual overview is illustrated in Figure 2

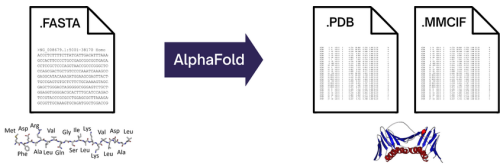

Figure 2. Input and Output Data Formats.

The computation provides basic and advanced settings that users can specify for their research. These parameters are passed during script execution and saved in **JSON format**.

3. Data Visualization

The 3D model of the protein is visualized using the **Mol* Viewer** [7] (see example in Figure 3), which offers various features and can be easily integrated into a web environment as a single component.

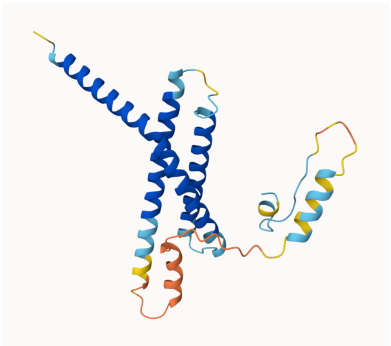

Figure 3. Protein Visualization.

Along with the 3D models, the prediction tools provide additional confidence measures that utilize the Recharts library for visualization.

4. Behind the Scenes

The architecture of the application illustrated in the Figure 4 consists of three parts:

- **Front-End Client:** Developed with Next.js and React, the front-end handles the user interface that allows users to **submit** computational requests and **view results**. Interaction with the back-end via HTTP requests and responses.
- **Back-End:** The server, built with Python and Flask, receives user requests, **coordinates** computational tasks via the AlphaFold engine, and **returns** the results to the front end for display.
- **Computational Engine (AlphaFold):** This module performs essential computations for predicting protein structures and delivering results for the back-end to manage.

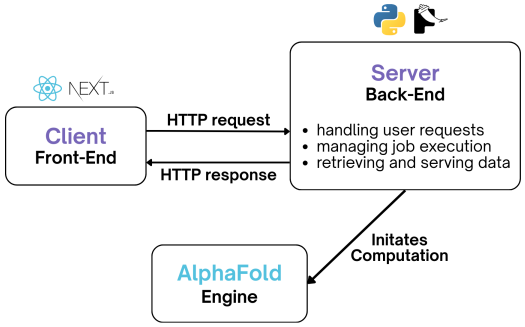

Figure 4. Application Architecture.

5. Conclusions

By integrating protein prediction processes with web development technologies, I developed an intuitive web application that effectively streamlines computational protein prediction for scientists. It provides **various prediction tools**, allows simple computation **submissions** with appropriate **feedback** and **error handling**, features **data visualization**, and offers an overview of computed proteins that can be **downloaded locally**.

The application is deployed in the Kubernetes cluster and is currently being used by registered users of Metacentrum. After five months of operation, it has successfully completed over **700 predictions**.

It has the potential to be extended with more prediction tools such as Chai 1 or Boltz 1. More visualizations of the provided data, e.g., PAE (Predicted Aligned Error), can be added. Further modularization and refactoring of the code would be beneficial, and the simplification of adding new tools has significant potential for development.

Acknowledgements

I would like to thank my supervisor, Ing. Ivana Burgetová, Ph.D., for her help. I also extend my thanks to my external consultant, RNDr. Lukáš Hejtmánek, Ph.D., for his mentorship and guidance. Additionally, I would like to thank CERIT-SC for providing the computational resources necessary to deploy this tool.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Evans, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493—-500, 2024.

[2] John Jumper, Richard Evans, Alexander Pritze, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[3] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature*, 19:679–682, 6 2022.

[4] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022.

[5] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019.

[6] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 4 edition, 2002.

[7] David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, and Alexander S Rose. Mol* viewer: modern web app for 3d visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1):W431–W437, 05 2021.