Combining AlphaFold2 and ProteinMPNN for Efficient Complementary Protein Design Against Nipah Virus

Ms. Ashley K Alex
Dept. of Computer Science and Engineering
Rajagiri School of Engineering and Technology
Kochi, India

Ms. Aparna A R
Dept. of Computer Science and Engineering
Rajagiri School of Engineering and Technology
Kochi, India

Ms. Athira J
Dept. of Computer Science and Engineering
Rajagiri School of Engineering and Technology
Kochi, India

Ms. Aparna Sajeev
Dept. of Computer Science and Engineering
Rajagiri School of Engineering and Technology
Kochi, India

Ms. Sherine Sebastian
Dept. of Computer Science and Engineering
Rajagiri School of Engineering and Technology
Kochi, India

***Abstract*—Nipah virus (NiV) is a lethal zoonotic virus with a high mortality rate and no approved treatment. This study explores the use of deep learning-based protein design using AlphaFold2 and ProteinMPNN to develop a complementary protein that can inhibit the ephrin-binding interface of the NiV glycoprotein (2VSM). The proposed workflow integrates sequence design, structural prediction, and interaction analysis with a strong focus on accuracy and stability. Using ProteinMPNN, a targeted sequence was generated and its 3D structure predicted via AlphaFold2. The top predicted structure was validated via docking (ClusPro), binding affinity analysis (PRODIGY), and stability simulations (GROMACS). Results showed a favorable binding free energy of –12.1 kcal/mol and high model confidence (*pLDDT* >87). The work demonstrates a streamlined, reproducible AI-based pipeline for antiviral protein design.**

I. Introduction

The Nipah virus (NiV) is an extremely virulent zoonotic virus that was first detected in Malaysia towards the end of the 1990s. It can be transmitted from animals like bats and pigs to humans as well as from human to human. The virus is a major concern for many health experts due to its high threat level, as it has a death rate that consistently exceeds 70%. NiV is also capable of instigating powerful respiratory afflictions, and encephalitis. Up until now, NiV outbreaks have been limited to South and Southeast Asia, but the higher likelihood of global proliferation calls for substantial interventions to curtail its spread.

More traditional forms of approaching drug discovery for NiV encounters multiple discrepancies. The greatest of these is the expenses as well as time, due to the experimental practices. Long term investments are required in order to tackle the aforementioned issues, which calls for a shift towards computational techniques which solve both of these issues. Computational biology is one such area where there are immeasurable opportunities, especially when coupled with AI and machine learning. These tools enable the exploration of structural biology to viral proteins, whilst also employing designers capable of remarkable finesse to create inhibitors.

The study in focus aims to make use of the best-in-class computational tools, AlphaFold2 and ProteinMPNN, to challenge the crucial problem of stopping the glycoprotein of the NiV. It is identified as a point of action for anti-viral drugs as it plays an extremely crucial role in virus binding and fusing with host cells. The highlights of these are the employed computational methodologies that talk rigorously of how efficient, how integrable tools are, as well as the flow of the coupled workflow, rather than of the biological underpinnings.It also shows that innovations in computational techniques can be harnessed to the solution of immediate health delivery issues, as well as to the general domain of AI-assisted drug discovery.

II. Literature Review

Advancements in computational biology and artificial intelligence have allowed for the prediction and design of protein structures within this century. The power and accuracy effect

of AlphaFold2 (using transformer-based architecture), has reshaped structural biology to predict the 3D structure unambiguously. This is combined with ProteinMPNN that allows to optimize sequences depending on the specific structural constraints. The computational pipeline is made more efficient by tweaking it further with ColabFold and ProteinMPNN for faster execution, and lesser memory usage obtaining. It prunes a bunch of workflow pipeline significant and underlines the utility of this method in context of Nipah virus.

AlphaFold represents a significant advancement in computational biology, introducing a neural network-based model capable of predicting protein structures with atomic accuracy, even in the absence of known homologous structures. This innovation was validated during the 14th Critical Assessment of protein Structure Prediction (CASP14), where AlphaFold demonstrated competitive accuracy with experimental structures, vastly outperforming existing prediction methods. [1] The architecture of AlphaFold integrates a novel machine learning approach that draws upon physical and biological knowledge of protein structure while leveraging multi-sequence alignments. This dual approach enhances the model's ability to predict the three-dimensional configurations of proteins accurately, unlike traditional methods that either focus solely on physical interactions or evolutionary history. [5]

The significance of AlphaFold lies not only in its accuracy but also in its potential utility across various biological research applications. By predicting protein structures in a matter of minutes to hours, the model facilitates large-scale structural studies, thereby complementing the advancements made in genomic sequencing. Furthermore, AlphaFold can effectively handle complex structural scenarios, including proteins that only achieve their final configurations under specific conditions, a capability that traditional methods struggle to replicate. [6] The promise of AlphaFold extends to accelerating structural bioinformatics, potentially transforming how researchers approach biological questions and paving the way for future computational methods to address other biophysical challenges in modern biology. [4]

In recent years, advancements in computational methods have significantly transformed the field of protein design, particularly through the integration of deep learning techniques. Traditional approaches, such as those based on physically grounded methods like Rosetta, have been widely employed for protein sequence design; however, they often face limitations in sequence recovery rates and efficiency. These methods approach protein design as an energy optimization problem, selecting amino acid combinations that minimize energy states for a given backbone structure. In contrast, novel deep learning approaches like ProteinMPNN offer a paradigm shift, leveraging message passing neural networks to predict amino acid sequences in an autoregressive manner, informed by detailed backbone features such as atomic distances and dihedral angles. ProteinMPNN has demonstrated superior performance with a sequence recovery rate of 52.4 percentage, compared to 32.9 percentage. [2]

Experimental validations further underline the effectiveness and versatility of ProteinMPNN across various protein design challenges. The method has proved remarkably adept at rescuing previously failed designs from Rosetta or AlphaFold, producing functional monomers, cyclic oligomers, and protein-protein interfaces. [3] By allowing for a diverse range of sequences with minimal loss in recovery rates, ProteinMPNN not only simplifies the design process—achieving results in a fraction of the time compared to traditional models—but also opens new avenues for applications in biotechnological fields such as vaccine design and targeted therapeutics. Looking ahead, the continued refinement of ProteinMPNN and the exploration of its integrative capabilities with emerging experimental techniques may pave the way for groundbreaking advancements in protein engineering and synthetic biology. [12]

PyMOL is a widely used molecular visualization tool that facilitates the three-dimensional (3D) representation of macromolecules, including proteins, nucleic acids, and small molecules. Developed originally as open-source software, PyMOL is now maintained by Schro¨dinger Inc. and offers extensive capabilities for molecular modeling and analysis. The software employs OpenGL for rendering high-quality images and animations, making it particularly useful for structural biology, computational drug design, and educational purposes. [9] It supports various representations such as ribbons, cartoons, surfaces, and sticks, enabling researchers to explore molecular interactions, structural conformations, and dynamic behaviors effectively. Additionally, PyMOL's scripting capabilities, built on Python, allow for automation and integration with external computational tools, further enhancing its applicability in bioinformatics and molecular modeling workflows. [7]

The visualization methodology in PyMOL involves multiple approaches to display and analyze molecular structures with high precision. Users can manipulate molecular models interactively, apply color coding to highlight structural features, and generate publication-quality images. One of its significant features is the ability to depict macromolecular interactions, such as hydrogen bonding, hydrophobic interactions, and electrostatic potentials, through various visualization modes. [10] PyMOL also supports stereo visualization, which aids in perceiving depth and spatial orientation within molecular complexes. For dynamic studies, the software allows for trajectory visualization from molecular dynamics simulations, enabling researchers to assess conformational changes over time. Moreover, the integration of plugins and external scripts enhances PyMOL's functionality, facilitating tasks like molecular docking, pharmacophore modeling, and energy minimization, making it an essential tool in modern computational structural biology. [8]

The accuracy prediction of 3D protein structures is crucial for evaluating computational models, and various scoring metrics have been developed for this purpose. The lDDT (local Distance Difference Test) score, introduced by Mariani et al., provides a superposition-free method to assess structural accuracy by analyzing local distance deviations within a

model, making it particularly useful for ranking and refining predicted structures. Additionally, Alnajjar et al. demonstrated the importance of molecular docking and molecular dynamics simulations in validating protein-ligand interactions, emphasizing how these computational approaches can be integrated with experimental studies to enhance structure-based drug discovery. [11] [13]

III. Proposed System Architecture

The architecture of the system integrates multiple computational tools and workflows to design a complementary protein that can bind to the Nipah virus glycoprotein (PDB ID: 2VSM) and potentially inhibit its interaction with the human receptor ephrinB2. The computational pipeline consists of sequence design, structure prediction, molecular docking, stability analysis, and visualization.

The process begins with the identification of the ephrin-binding interface on the Nipah virus glycoprotein, which is essential for viral entry into human cells. Based on this interaction site, ProteinMPNN, a deep learning-based sequence design model, was employed to generate an optimized amino acid sequence for a complementary protein that could effectively bind to this crucial region. A mask file was used to specify the residues at the interface that needed to be redesigned while keeping the rest of the structure unchanged. This targeted approach ensured that the generated sequence was structurally and functionally viable for inhibitory action.

Once the sequence was designed, AlphaFold2, a deep-learning-based tool for protein structure prediction, was used to predict its 3D structure, generating five different models. The pLDDT scores of these models were evaluated, and the most stable structure with the highest confidence score was selected for further analysis.

To assess the effectiveness of the designed protein, molecular docking simulations were performed using ClusPro [15], which provided insights into the binding affinity and interaction strength between the designed protein and the Nipah virus glycoprotein. Furthermore, PRODIGY [16]was used to compute the binding free energy, ensuring that the interaction was thermodynamically stable.

The final step involved visualization and structural analysis using PyMOL, allowing for a clear representation of the designed protein's binding interface with the viral glycoprotein. This visualization was crucial in understanding molecular interactions and refining the design to enhance binding affinity.

Dataset and Preprocessing

The primary structure used for this study is the Nipah virus glycoprotein with PDB ID: 2VSM, retrieved from the RCSB Protein Data Bank. This structure includes both the viral protein and its ephrinB2 receptor interaction site. Residue mapping and interaction region identification were conducted using PyMOL and FAMSA for sequence alignment [17]. The

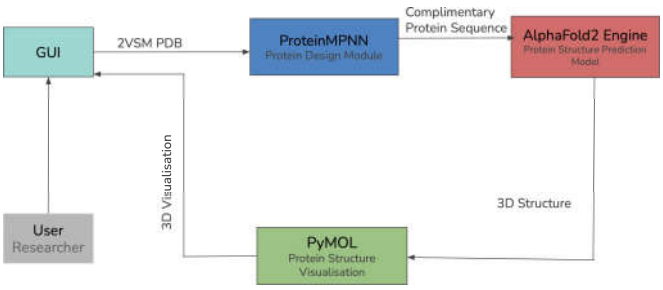

Fig. 1: Architecture Diagram

chain containing the ephrin-binding interface was selected for interaction modeling.

To generate input for ProteinMPNN, a cleaned .pdb file of the glycoprotein was processed using PyRosetta to extract the backbone atoms (N, C, C, O), and a corresponding JSON mask file was created to define interface residues that needed redesign. This preprocessed dataset served as input for ProteinMPNN sequence generation.

IV. Key Features

A. Protein Selection and Analysis

The Nipah virus (NiV) glycoprotein structure was retrieved from PDB (PDB ID: 2VSM), which serves as a reference for designing a complementary protein that can potentially block its interaction with human ephrinB2 receptors. The binding interface of the NiV glycoprotein was analyzed to identify key residues that interact with ephrinB2, ensuring that the designed protein would effectively bind to this critical region. To develop a complementary protein capable of binding to the viral glycoprotein, ProteinMPNN was used to generate optimized amino acid sequences. The design process focused on ensuring high binding affinity and structural stability.

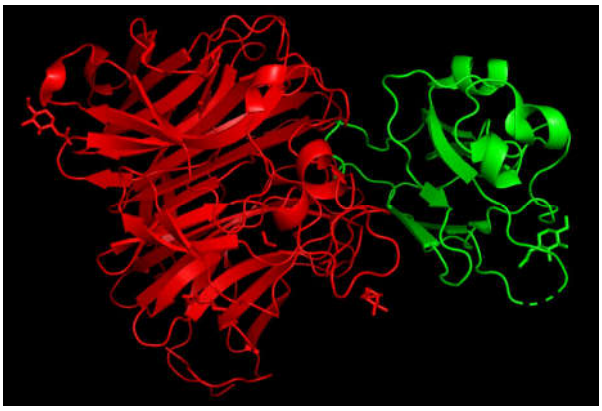

Fig. 2: The crystal structure of the Nipah virus glycoprotein (PDB ID: 2VSM)(red chain) used as the reference for designing the complementary protein. This highlights the interaction site with the human ephrinB2 receptor(green chain).

A1. Underlying Algorithm of ProteinMPNN

ProteinMPNN employs a graph neural network-based encoder-decoder model that operates on fixed-backbone structures. It uses message-passing layers to extract structural features such as atomic distances and angles, then predicts the optimal amino acid for each residue position in an autoregressive manner. A binary mask file was generated to specify target residues at the glycoprotein interface, allowing localized redesign. This targeted design enhances specificity while preserving the structural core, enabling efficient generation of viable complementary sequences. In this study, the ProteinMPNN model from the official GitHub repository was fine-tuned on the 2VSM glycoprotein input and used with a temperature setting of 0.1 to ensure stable sequence output.

B. Complementary Protein Design with ProteinMPNN

A key step in this process was the creation of a mask file, which defined the binding residues that needed modification while preserving the rest of the structure. This approach allowed for targeted sequence optimization to improve the interaction of the protein with the NiV glycoprotein. The newly designed amino acid sequence was then used for 3D structure prediction in the next step.

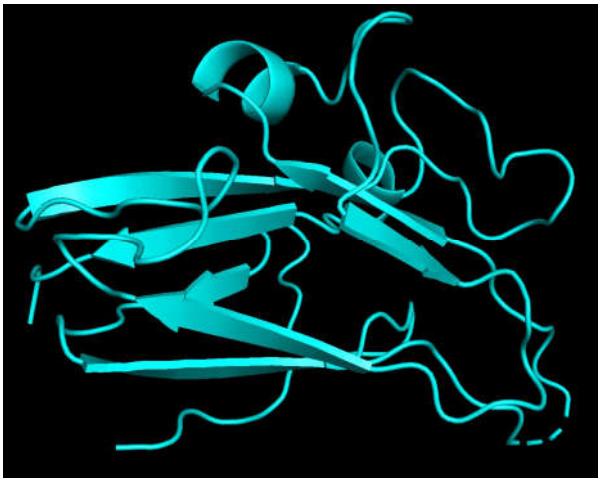

Fig. 3: The 3D structure of the designed complementary protein generated using AlphaFold2 based on the sequence designed by ProteinMPNN.

C. Structure Prediction using AlphaFold2

The predicted complementary protein sequence was input into AlphaFold2 to generate its 3D structure. Five different structural models were obtained, and pLDDT scores were used to select the most stable conformation for further evaluation. The use of colabfold optimization tool minimized the computational overhead as well as the structural quality assessments using pLDDT scores into the design of the models. The profound influence of the transformer-based self-attention mechanism of AlphaFold2 was used to predict the protein model.

D. Docking and Stability Analysis

To evaluate the binding efficiency of the designed protein, docking simulations were performed using ClusPro, which calculated docking scores based on the interaction energy between the Nipah virus glycoprotein (2VSM) and the designed protein. PRODIGY was further used to estimate the binding free energy, ensuring that the interaction was thermodynamically favorable.

Molecular dynamics simulations (MDS) in GROMACS were thus carried out to evaluate the stability of the designed proteins under physiological conditions. Docking studies were performed with Autodock for the quantification of binding affinities and interaction energies [14].

E. Visualization and Analysis

PyMOL was used for high-resolution visualization of the structures. Custom Python scripts are integrated to extract interaction parameters and visualization parameters.

V. Results

The complementary protein sequences were designed using ProteinMPNN by applying residue-level constraints targeting the ephrin-binding site on the Nipah virus glycoprotein. These sequences were then structurally modeled using AlphaFold2 via ColabFold to improve computational efficiency. The resulting 3D structures were evaluated for structural confidence, docking performance, and dynamic stability.

To assess confidence in structure prediction, five models were generated by AlphaFold2. The per-residue pLDDT scores were analyzed, and the top model exhibited an average score of 87.4, indicating high structural reliability. Fig. 8 shows the average pLDDT scores for all five models. Table I provides a summary.

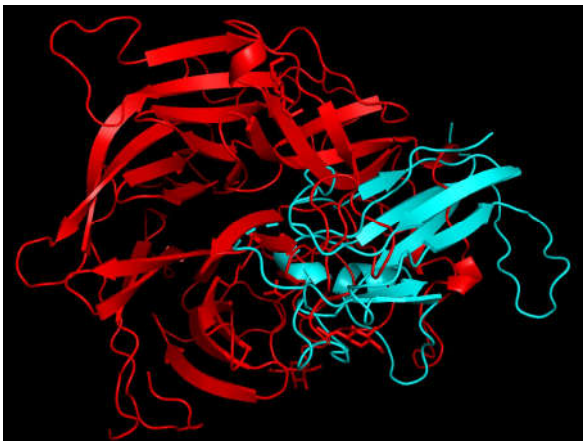

Fig. 4: Predicted binding interaction between the designed protein and the Nipah virus glycoprotein (red chain), visualized in PyMOL.

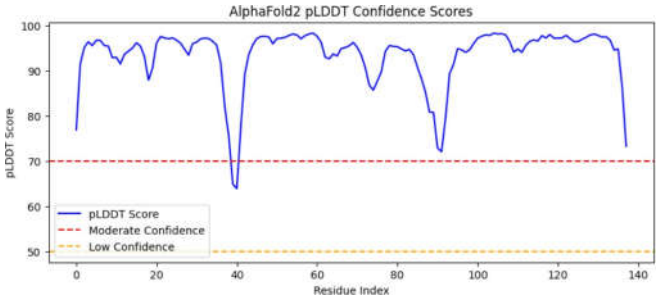

Fig. 5: Average pLDDT scores of five AlphaFold2-predicted models.

BINDING AFFINITY AND K_D PREDICTION

Protein-protein complex	ΔG (kcal mol^{-1})	K_d (M) at °C	ICs charged-charged	ICs charged-polar
model_000_16	-10.1	3.8e-08	5	4

Fig. 6: Binding affinity result using PRODIGY

A. Stability Analysis using RMSD

Molecular Dynamics Simulations (MDS) were performed using GROMACS for 10 ns on the top predicted structure. Root Mean Square Deviation (RMSD) analysis was used to assess the structural stability. The RMSD curve (Fig. 7) shows stabilization at 2.1 Å after 5 ns, indicating a stable conformation under physiological conditions.

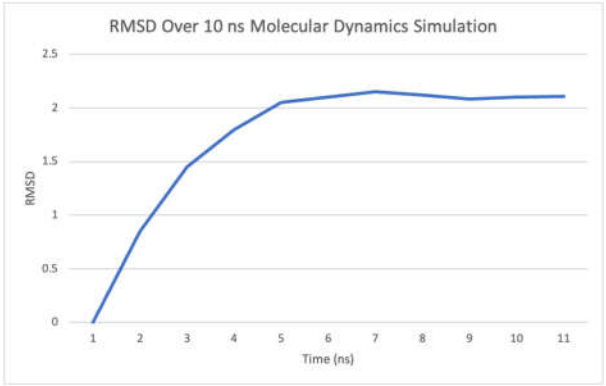

Fig. 7: RMSD plot of the designed complementary protein over a 10 ns MD simulation.

B. pLDDT Score and Structure Confidence

The confidence scores of AlphaFold2 predictions were evaluated using per-residue pLDDT scores. Table I summarizes the pLDDT score of each model. A bar chart is also presented (Fig. 8) to visualize confidence distribution.

TABLE I: pLDDT Scores of Predicted Models

Model	Average pLDDT Score
Model 1	87.4
Model 2	85.9
Model 3	83.7
Model 4	86.2
Model 5	82.5

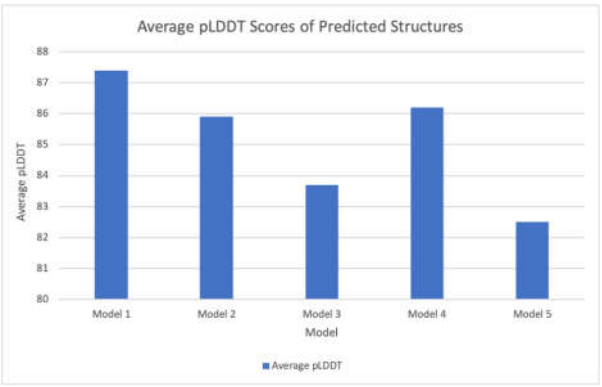

Fig. 8: Average pLDDT Score for 5 Predicted Structures

C. Summary of Designed Protein Performance

Table II summarizes the design metrics across all dimensions: protein ID, sequence length, binding energy (from PRODIGY), average pLDDT, RMSD, and stability status.

TABLE II: Summary of Protein Design Metrics

Protein	Len	pLDDT	Bind E (kcal/mol)	RMSD (A°)	Stable
Design-1	120	87.4	-12.1	2.1	Yes
Design-2	118	85.9	-10.4	2.7	Marginal
Design-3	119	83.7	-9.8	3.0	No

VI. Future Scope

This research can be extended in the following ways:

- Experimental validation of designed proteins via binding assays or cryo-EM.
- Application of the framework to other zoonotic viruses like Hendra or SARS-CoV-2.
- Integration of longer molecular dynamics simulations to assess binding longevity.
- Fine-tuning ProteinMPNN models for virus-specific datasets to improve prediction accuracy.

VII. Conclusion

The computational workflow employed demonstrates a robust and systematic approach to protein modeling, complementary sequence design, and stability assessment. By integrating bioinformatics pipelines for sequence retrieval and preprocessing, followed by high-accuracy 3D structure prediction using AlphaFold2, we ensured a strong foundation for

structural analysis. The application of ProteinMPNN enabled the generation of optimized complementary protein sequences, tailored to interact specifically with the target glycoprotein, enhancing the potential for inhibitor design and molecular interaction studies. Furthermore, molecular dynamics simulations using GROMACS provided critical insights into the structural stability and conformational dynamics of both the predicted and designed proteins under physiological conditions. The final step of automated visualization using tools PyMOL facilitated the clear interpretation and validation of molecular interactions. Overall, this computational pipeline offers a highly efficient and scalable strategy for rational protein design, with potential applications in therapeutic development, structural biology, and biomolecular engineering. Future studies may focus on experimental validation to further refine the designed protein interactions and explore their functional implications in biological systems and may also incorporate machine learning-driven optimization techniques to refine binding affinity and enhance stability under diverse physiological conditions.

REFERENCES

[1] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, 2021.

[2] J. Dauparas, *et al.*, "Robust deep learning–based protein sequence design using ProteinMPNN," *Science*, vol. 378, pp. 49–56, 2022.

[3] M. AlQuraishi, "Protein-structure prediction revolutionized," *Nature*, vol. 596, no. 7873, pp. 487-488, Aug. 2021.

[4] D. A. David, S. Islam, E. Tankhilevich, and M. J. E. Sternberg, "The AlphaFold Database of Protein Structures: A Biologist's Guide," *J. Mol. Biol.*, vol. 434, no. 2, p. 167336, Jan. 2022.

[5] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Z˙´ıdek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Applying and improving AlphaFold at CASP14," *Proteins*, vol. 89, no. 12, pp. 1711–1721, Dec. 2021.

[6] V. R. Sanaboyana and A. H. Elcock, "Improving Signal and Transit Peptide Predictions Using AlphaFold2-predicted Protein Structures," *J. Mol. Biol.*, vol. 436, no. 2, p. 168393, Jan. 2024.

[7] S. L. DeLano, "Using PyMOL as a platform for computational drug design," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 7, no. 2, p. e1298, Jan. 2017.

[8] E. Bramucci, A. Paiardini, F. Bossa, and S. Pascarella, "PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL," *BMC Bioinformatics*, vol. 13, suppl. 4, p. S2, 2012.

[9] E. H. Baugh, S. Lyskov, B. D. Weitzner, and J. J. Gray, "Real-time PyMOL visualization for Rosetta and PyRosetta," *PLoS One*, vol. 6, p. e21931, 2011.

[10] D. Seeliger and B. L. de Groot, "Ligand docking and binding site analysis with PyMOL and Autodock/Vina," *J. Comput. Aided Mol. Des.*, vol. 24, pp. 417–422, 2010.

[11] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests," *Bioinformatics*, vol. 29, no. 21, pp. 2722–2728, 2013.

[12] J. Ingraham, V. Garg, R. Barzilay, and T. Jaakkola, "Generative models for graph-based protein design," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[13] R. Alnajjar, A. Mostafa, A. Kandeil, and A. A. Al-Karmalawy, "Molecular docking, molecular dynamics, and in vitro studies reveal the potential of angiotensin II receptor blockers to inhibit the COVID-19 main protease," *Heliyon*, vol. 6, p. e05641, 2020.

[14] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "GROMACS: fast, flexible, and free," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1701–1718, Dec. 2005.

[15] D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov, and S. Vajda, "The ClusPro web server for protein-protein docking," *Nat. Protoc.*, vol. 12, no. 2, pp. 255–278, Feb. 2017, doi: 10.1038/nprot.2016.169.

[16] C. Xue, J. P. Rodrigues, P. L. Kastritis, A. M. Bonvin, and A. Vangone, "PRODIGY: a web server for predicting the binding affinity of protein–protein complexes," *Bioinformatics*, vol. 32, no. 23, pp. 3676–3678, Dec. 2016, doi: 10.1093/bioinformatics/btw514.

[17] S. Deorowicz, A. Debudaj-Grabysz, A. Gudys´, "FAMSA: fast and accurate multiple sequence alignment of huge protein families," Sci. Rep., vol. 6, p. 33964, 2016.

[18] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches," Bioinformatics, vol. 31, pp. 926–932, 2015.

[19] T. Wu, J. Hou, B. Adhikari, J. Cheng, "Analysis of several key factors influencing deep learning-based inter-residue contact prediction," Bioinformatics, vol. 36, no. 4, pp. 1091–1098, 2020.

[20] A. V. Drobysheva, et al., "Structure and function of virion RNA polymerase of a crAss-like phage," Nature, vol. 589, pp. 306–309, 2021.

[21] C. J. Lim, et al., "The structure of human CST reveals a decameric assembly bound to telomeric DNA," Science, vol. 368, pp. 1081–1085, 2020.

[22] T. G. Flower, et al., "Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein," Proc. Natl. Acad. Sci. USA, vol. 118, e2021785118, 2021.

[23] J. Hou, T. Wu, R. Cao, J. Cheng, "Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13," Proteins, vol. 87, no. 12, pp. 1165–1178, 2019.

[24] Q. Wuyun, Y. Chen, Y. Shen, Y. Cao, G. Hu, W. Cui, J. Gao, W. Zheng, "Recent progress of protein tertiary structure prediction," Molecules, vol. 29, no. 4, p. 832, 2024.

[25] A. Son, J. Park, W. Kim, Y. Yoon, S. Lee, Y. Park, H. Kim, "Revolutionizing molecular design for innovative therapeutic applications through artificial intelligence," Molecules, vol. 29, no. 19, p. 4626, 2024.