RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Project Report on

Protein Structure Prediction

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering

By

Aparna A R (U2103044)
Aparna Sajeev (U2103045)

Ashley K Alex (U2103052)
Athira J (U2103054)

Under the guidance of

Ms. Sherine Sebastian

Department of Computer Science and Engineering

Rajagiri School of Engineering & Technology (Autonomous)

(Parent University: APJ Abdul Kalam Technological University)

Rajagiri Valley, Kakkanad, Kochi, 682039

April 2025

CERTIFICATE

This is to certify that the project report entitled ***"Protein Structure Prediction"*** *is a bonafide record of the work done by* ***Aparna A R (U2103044),Aparna Sajeev (U2103045)*** *,* ***Ashley K Alex (U2103052)*** *,* ***Athira J (U2103054)*** *submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in "Computer Science and Engineering" during the academic year 2024-2025.*

Project Guide
Ms. Sherine Sebastian
Assistant Professor
Dept. of CSE
RSET

Project Coordinator
Mr. Harikrishnan M
Assistant Professor
Dept. of CSE
RSET

Head of the Department
Dr. Preetha K G
Professor
Dept. of CSE
RSET

ACKNOWLEDGEMENT

We wish to express our sincere gratitude towards **Rev. Dr. Jaison Paul Mulerikkal CMI**, Principal of RSET, and **Dr. Preetha K G**, Head of the Department of Computer Science and Engineering for providing us with the opportunity to undertake our project, "Protein Structure Prediction".

We are highly indebted to our project coordinators, **Mr. Harikrishnan M**, Assistant Professor, Department of Computer Science and Engineering and **Ms. Sangeetha Jamal**, Assistant Professor, Department of Computer Science and Engineering for their valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our project guide **Ms. Sherine Sebastian**, Assistant Professor, Department of Computer Science and Engineering for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, we would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Aparna A R

Aparna Sajeev

Ashley K Alex

Athira J

Abstract

The project aims to design and evaluate complementary proteins capable of inhibiting the Nipah virus glycoprotein, a key viral component responsible for host cell attachment and entry. The approach begins by uploading the PDB structure of the Nipah virus glycoprotein–human receptor complex to the website. The website integrates ProteinMPNN, a deep learning-based protein design tool, which generates complementary protein sequences specifically optimized to bind the glycoprotein, using the human receptor as a structural reference.

The generated sequences are then processed through AlphaFold2, a protein structure prediction model, to determine the 3D conformation of the designed complementary proteins. These predicted structures are subsequently visualized using PyMOL, for interactive molecular visualization. To assess the structural stability of the designed proteins, multiple parameters were evaluated, including the number of hydrogen bonds, residue-residue interactions, and ionic bonds. These metrics were used to compute a stability score. For functional validation, molecular docking was performed using ClusPro and PRODIGY, which assessed the binding affinity and interaction interfaces between the designed complementary proteins and the viral glycoprotein.

Compared to traditional experimental approaches, this project significantly accelerates the design and optimization process. By combining machine learning-driven protein prediction with high-resolution visualization, it offers a faster, cost-effective framework for antiviral discovery. This approach has the potential to improve the therapeutic development for Nipah and similar viral threats, contributing to more efficient responses in pandemic preparedness.

Contents

List of Abbreviations

3D	-	Three-Dimensional
AI	-	Artificial Intelligence
CASP	-	Critical Assessment of Protein Structure Prediction
GNN	-	Graph Neural Network
GPU	-	Graphics Processing Unit
MSA	-	Multiple Sequence Alignment
PDB	-	Protein Data Bank
PyMOL	-	A molecular visualization system
RMSD	-	Root Mean Square Deviation
RMSF	-	Root Mean Square Fluctuation
R&D	-	Research and Development
SASA	-	Solvent Accessible Surface Area
UniProt	-	Universal Protein Resource
UCSF	-	University of California, San Francisco

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Infectious diseases, especially emerging viruses, have highlighted the urgent need for novel ways to treat and prevent them. The Nipah virus, one of the serious health threats, is capable of causing high death rates and potentially widespread outbreaks. While Southeast Asia has been most affected by the Nipah virus, recent cases have been reported internationally. This situation raises global concerns, as the virus severely impacts both the lungs and the brain. Traditional methods of creating vaccines and medicines are time-consuming and require a lot of effort and resources. This is where computational methods comes into play. Computational methods offer a faster and more efficient way which speeds up drug discovery using computer simulations. Adavancements in machine learning and artificial intelligence have revolutionized the study and design of drugs. For example tools like AlphaFold2,can predict the 3D structure of proteins from their amino acid sequence.This has been a breakthrough in helping scientist better understand viral proteins,which is crucial for designing new therapies.

However we need molecules that will attach to the proteins and inhibit the functioning of the virus as well as the shape of the protein. Towards this end, tools such as ProteinMPNN and other protein design programs can design novel polypeptides that may attach to and inhibit the activity of viral proteins. The aim of this project is to design a protein that could block the Nipah virus from infecting cells by binding its glycoprotein.

3D visualization tools such as PyMOL are employed to design the protein and analyze the binding of the protein to the virus, key sites of contact, and the stability of the protein. The development of these computational and modeling methods helps to shorten the time frames and reduce the costs of developing drugs, which is highly desirable in conditions where we are already coping with new viruses.

1.2 Problem Definition

The understanding of the biology of a cell lies in the correct prediction of the structure of a protein, enabling the design of targeted therapies for a large number of diseases. It will be important to state that considering the variety of posed emerging threats, Nipa virus induced ones will be able to emphasize the importance of effective computational protein structure prediction and analysis tools. This project seeks to address this shortcoming by employing cutting edge technologies such as AlphaFold2 which can model the 3D structures of viral proteins from their amino acid sequences. Indeed, a structure of a protein can be used to determine the most critical features that need to be examined, which include the active and the binding sites of the protein. This is very critical for the design of drugs. PyMOL is another visualization tool that details the structures based on regions of interaction, accessibility on the surface, and the general shape of the molecule which are all important parameters for the ideal development of drugs.

The project focuses to develop therapeutic drugs as solutions for Nipah virus by accurately predicting the protein struture by making use of AlphaFold2 and visualisation tool PyMOL. By identifying binding site it is possible to develop complementary protein which gets binded to binding site and thus prevents from attack of viral proteins.

1.3 Scope and Motivation

The project primarily deals with developing a software prediction along with an experimental determination of three-dimensional (3D) structural structure of proteins, particularly the glycoprotein of Nipah virus, using AlphaFold or other computational predictive methods. The structure will be deployed for visualization prediction and the analysis of its stability to identify likely areas of vulnerability. The target proteins will also be designed to create proteins that will complementally bind with the binding site of the chosen target-inhibitor protein. These efforts will serve as preliminary steps towards finding novel therapeutic interventions in structural biology in general. Probing these capabilities should be beneficial for drug discovery and vaccination applications as well.

The inspiration for this research study is the requirement for developing effective

treatments for the prevention and management of virus-caused diseases, such as Nipah. Comprehensive insight into the corresponding structure of viral glycoproteins can throw light on their function and indeed interaction with host cells. This research identifies possible inhibitor binding sites that may be required in the further development of targeted therapy. Current challenges with viral infections amplify the need for uncovering new therapeutic strategies. This project will, therefore, contribute to an international worldwide effort toward preventive measures against viral events using innovative structure-based drug designing.

1.4 Objectives

- Utilize the 2VSM structure as the receptor to design complementary proteins that can potentially inhibit the Nipah virus glycoprotein.
- Generate complementary protein sequences using ProteinMPNN to bind effectively to the 2VSM structure.
- Predict the 3D structure of the generated complementary protein sequences using AlphaFold2.
- Visualize the predicted protein structures using PyMOL to analyze key features and conformational stability.
- Evaluate the stability and binding affinity of the designed complementary proteins using molecular docking and energy calculations.
- Identify the most stable and effective complementary protein that can inhibit the glycoprotein function, potentially aiding in antiviral drug development.

1.5 Challenges

The project on protein structure prediction and complementary protein design involves addressing multiple technical and scientific challenges.

1. **Computational Complexity:**

   Protein structure prediction using AlphaFold2 and complementary protein design

with ProteinMPNN require substantial computational resources. Running these tools on standard hardware may lead to prolonged processing times or limited scalability. Visualization of large molecular structures using tools like PyMOL can be resource-intensive, particularly for high-resolution rendering or working with extensive datasets like cryo-EM maps.

2. **Accuracy and Stability:**

   While AlphaFold2 provides accurate predictions for many proteins, structural instability in certain regions or discrepancies in predicted binding sites can affect downstream design efforts. Ensuring that the complementary proteins designed by ProteinMPNN are not only structurally compatible but also biologically functional is another significant challenge.

3. **Data Quality and Availability:**

   Reliable 3D protein structures require high-quality input data. Errors in the amino acid sequence or incomplete structure predictions may compromise the design process. Publicly available datasets, such as the Protein Data Bank (PDB), may lack adequate structural information for certain target proteins like those of the Nipah virus.

4. **Integration of Tools:**

   Integrating multiple tools, including AlphaFold2, ProteinMPNN, and visualization software like PyMOL, involves developing workflows that ensure seamless data transfer and compatibility between formats (e.g., PDB, FASTA). Misalignments between tools or errors during data conversion can result in incorrect or incomplete outputs.

5. **Validation Challenges:**

   Computational results must be experimentally validated to ensure real-world applicability. However, wet-lab experiments to confirm binding affinity or functional efficacy may not always be feasible during the computational phase.

6. **Interdisciplinary Knowledge Requirements:**

   The project requires expertise in multiple domains, including computational biology, bioinformatics, and structural biology, as well as a solid understanding of machine

learning tools. Translating computational insights into biological relevance can be challenging without adequate domain knowledge.

7. **Software Limitations:**
   While tools like AlphaFold2 and PyMOL are powerful, they may not address all use cases. For example, AlphaFold2 is not designed to predict protein-protein docking or analyze long-term stability.

1.6 Assumptions

- **Accurate Protein Sequence Data** The 3D structure of the Nipah virus glycoprotein, 2VSM, is assumed to be accurate and complete. Any errors or ambiguities in the sequence could compromise the quality of the predicted structure and the subsequent design of complementary proteins.

- **AlphaFold2 Prediction Accuracy** It is assumed that AlphaFold2 will generate a reliable 3D structure for the glycoprotein, given its proven success with similar viral proteins. While AlphaFold2 provides high accuracy, there may still be regions of structural uncertainty, particularly in flexible loops or unstructured regions.

- **Complementary Protein Design Feasibility**The project assumes that the design of complementary proteins or peptides using ProteinMPNN will result in functional and stable candidates. It is assumed that these designed proteins will exhibit binding affinity and functionality when experimentally tested.

- **Computational Resources**The project assumes that sufficient computational resources will be available for running AlphaFold2 and ProteinMPNN simulations, despite the high computational demands.

1.7 Societal / Industrial Relevance

This project is relevant across various key societal-industrial domains, especially in public health and biotechnology. The Nipah virus threat represents a worldwide health risk with conspicuous outbreaks in Southeast Asia and substantially frightening mortality. This project, using computational biology tools such as AlphaFold2 and ProteinMPNN, aims

to develop revolutionary resource-sparing approaches for a rapid antiviral small molecular therapeutic development process. The long years of traditional drug discovery can be significantly shortened because of computational approaches.

From an industrial perspective, predicting and designing protein interactions within short periods could change the landscape of drug discovery pipelines. Pharmaceutical companies and biotechnological enterprises would gain from faster and cost-effective methods to develop targeted therapies for newly-emerged viral pathogens, and could consequently reduce their dependence and reliance on traditional labor-intensive experimental approaches. Besides, there are great opportunities and spaces that AI and machine learning would open for creating innovative biotech-class drugs while saving R and D costs and scaling options for vaccine and therapeutic development processes. Hence, this research could have far-reaching consequences from combating an outbreak of the Nipah virus to pandemic preparedness and response.

1.8 Organization of the Report

The report begins with the Introduction, outlining the background of the project, its significance, and the problem it addresses, with a focus on leveraging computational tools for protein structure prediction and complementary protein design. The Background section provides a detailed context for the project, highlighting advancements in molecular visualization and protein design. The Problem Definition articulates the challenges posed by the Nipah virus and the gaps in current therapeutic approaches. The Scope specifies the boundaries of the project, emphasizing the use of tools like AlphaFold2, ProteinMPNN, and PyMOL, while excluding experimental validations. Objectives are defined to include structure prediction, complementary protein design, stability analysis, and visualization. The Methodology details the workflow, from data input to visualization and result analysis, highlighting the algorithms and tools used. The Challenges section discusses technical and computational hurdles faced during the project. Social Relevance explains the broader impact of this work on global health and antiviral research. Finally, the Conclusion summarizes the findings, contributions to protein structure prediction, and potential future applications.

Chapter 2

Literature Survey

This chapter contains a thorough literature review on significant developments in protein structure prediction, visualization, and design. It presents reading materials on how tools such as AlphaFold2 change the landscape completely by making it possible to actually predict protein structures using their sequences; it describes PyMOL, the application, for visualizing these structures and analyzing the functional attributes of these; and discusses ProteinMPNN, which encompasses the use of such structural understanding for designing novel protein sequences. By reviewing recent research into these methodologies, the chapter would provide a solid coverage of state-of-the-art technologies in computational biology towards applications in protein engineering.

2.1 A-Prot: protein structure modeling using MSA transformer [1]

The authors of the paper presents A-Prot, a new protein 3D structure prediction approach that relies on the MSA Transformer which is a state-of-the-art protein language model. A-Prot generates distance and dihedral angle predictions by converting evolutionary information obtained from multiple sequence alignments into accurate protein models. When assessed from the perspective of CASP13 and CASP14 targets A-Prot can be claimed to be better than most of the top tier tools in long-range contact determination and is also competitive in a 3D model generation when compared to AlphaFold, trRosetta, among other tools. Protein structure prediction can be classified into before and after AlphaFold was created as it set a new and high benchmark. In contrast to AlphaFold2 which is heavily reliant on multiple GPUs, A-Prot has been developed with a single GPU in mind allowing for more diverse audience to make use of advanced protein structure prediction.

2.1.1 A-Prot's Methodology and Innovation

A-Prot uses the MSA Transformer in the task of retrieving evolutionary characteristics from multiple sequence alignments. The A-Prot approach and its implementation is a great milestone in the field of protein structure prediction as it developed an innovative technique of using the MSA Transformer which is an example of a protein language model. It's not customary with classical approaches, whereby the majority of them are dependent upon engineered features or require expensive coevolutionary analysis; A-Prot takes a step further by performing this automatically and bettered as an using the Transformer architecture. The MSA inputs are first processed by the MSA Transformer which gives two important results; row attention maps and MSA features which display how different residues relate to each other in a given set of sequences or in a different sequence that is the targeted sequence. Such outputs are a source of evolutionary and structural information and aid in better making decisions in the down stream processes.

These transformations applied by A-Prot convert the above facts into testable predictions. The MSA features are dimensionally reduced, and then combined with row attention maps to create 2D feature maps for residue-residue interactions. The feature maps then undergo a significant processing pipeline through a deep convolutional neural network consisting of a dilated ResNet with 28 residual blocks. Such an implementation refines the features and predicts important inter-residue geometries, such as distances and angles, with which 3D structures of proteins are built. Finally, it converts predicted geometries into a module for protein structure modeling called trRosetta to output the final 3D structure.

Innovation in A-Prot lies primarily in the synergistic combination of advanced machine learning with protein-specific biological insights. A-Prot uses a pre-trained MSA Transformer trained on millions of MSAs to efficiently extract coevolutionary and structural signals that would otherwise be missed, or that require vast computational resources, by traditional approaches. Further, it is modular, thus allowing flexible integration of components, be it feature extractors or structure modeling algorithms, making it a versatile tool. The use of diversity-minimized MSA subsampling improves computational efficiency at no loss in predictive accuracy. Eventually, while improving long-distance contact modeling, it also allows A-Prot to predict structures at much lower computational loads than other state-of-the-art approaches, thus embodiment of efficiency and accuracy.

2.1.2 Benchmarking Against State-of-the-Art Methods

The evaluation of A-Prot against cutting-edge solutions such as AlphaFold2 and trRosetta includes some tough targets, including CASP13 and CASP14. Indeed, with respect to the sessions regarding free modeling (FM) and template-based modeling (TBM), the CASP datasets made for evaluating methods in terms of their capability to predict the structures over protein without structural templates or with very minimal ones. A-Prot succeeded in those long-range contact predictions and achieved the highest precision in metrics like top L/5, L/2, and L at least 7–9 percent better than the existing methods like DeepDist and AlphaFold1. For example, A-Prot achieved top L/5 precision of 0.812, which was significantly better than that of other approaches in terms of residue-residue interaction capture across long distances in the protein. Characteristic in this area, the results of strong contact predictions were comparable in 3D modeling performance, particularly with respect to FM/TBM-hard targets as it managed to generate consistently high-quality models.

In the CASP14, the models of A-Prot outclassed a majority of leading server groups as far as lDDT scores are concerned when compared with them: for instance, about FEIG-S. Therefore, significantly high scores are indicative of superior local structural accuracy it provides. A parallel comparison with trRosetta using the same MSAs brought out the fact that A-Prot makes better predictions with regard to the various measures of TM-score, lDDT and dihedral angle predictions. In this context, A-Prot innovatively utilizes the MSA Transformer to garner rich evolutionary information, paving its way as one of the most competent, if not alone, efficient. The study shows A-Prot is a very powerful option for protein structure prediction, especially in very complex and resource-limited scenarios.

2.1.3 Limitations and Future Directions

Although A-Prot is no doubt accurate and fast, it is also burdened by limitations with room for improvement. It depends strongly on the quality of the input multiple sequence alignments (MSAs) and struggles when homologous sequences are limited, thus limiting its utility for proteins that are not well-conserved in sequence.However, while it can be computed much more cheaply than AlphaFold2, it is not sufficiently optimal to scale up

for larger proteins and across diverse datasets in the future. Future directions include improved sequence search techniques that provide better MSA quality, enhanced metagenomics dataset integration, and improvement in extracting features that will be used in model building. Some of the extensions that can be added to A-Prot include the prediction of properties such as binding site interactions and proteins' dynamics, making it applicable in drug discovery while increasing its relevance in protein design, thus sealing its place as a general-purpose tool in computational biology.

2.1.4 Conclusion

Driven by the advanced algorithms, the model A-Prot succeeded in building high-confidence structural models that were in line with experimental information wherever they existed. Such knowledge improves comprehension of the structural finer points for A-proteins and may thus provide insight into their biological function and potential contributions to disease processes. Very importantly, the accuracy of both AlphaFold and other comparable approaches has been demonstrated through comparisons to several known experimental structures, showing that these methods can also work to study convoluted proteins. Such work undertaken will include functional characterization of predicted protein interactions, as well as the design of potential therapeutic strategies targeting certain sites in the protein on the basis of the acquired structural data.

By taking advantage of these cutting-edge algorithms, high-confidence structural models that have compatibility with experimental data, where available, are produced by the A-Prot model. Improved understanding of A-protein structural intricacies can inform the biological functions that A-proteins might subserve and probably how they are involved in disease processes. Accurately and reliably, AlphaFold and similar tools have been proven to work in comparing them with several known experimental structures; hence, these computational approaches are viable in studying complex proteins. Future work will concern the functional characterization of protein interactions predicted, coupled with the design of possible therapy strategies focused on the specific protein sites as derived from structural information.

2.2 UCSF ChimeraX: Structure visualization for researchers, educators, and developers [2]

Molecular visualization is crucial in structural biology because it essentially enables scholars to understand protein and nucleic acid structures and functions in the biomolecular perspective. In this regard, UCSF ChimeraX, a next-generation molecular visualization tool developed by Pettersen et al. (2020), signifies a significant advancement. Whereas Chimera was based on its predecessor, ChimeraX incorporates modern visualization techniques, usability enhancement, and offering support for modern computational and hardware platforms. It has thus become crucial to researchers, educators, and developers in structural biology.

2.2.1 Key Features and Innovations

ChimeraX provides advanced visualization utilities to deal with diversified and huge-scale molecular datasets, including cryo-electron microscopy maps and X-ray crystallography data. Among other features offered by the product are real-time rendering techniques, ambient occlusion and shadow mapping, which afford realistic and detailed rendering of molecular graphics. In addition to such visualization tools, ChimeraX provides advanced structure analysis functionalities such as hydrogen bond identification, torsion angle measurement, and molecular docking evaluations. These would enable thorough protein and ligand interaction analysis and structural dynamics evaluation. One of the most important aspects of ChimeraX is the entire package of features that treats virtual reality (VR) immersion in complex molecular assemblies. With VR goggles, molecular constructs look-just like structures that occupy space within an office building-in three dimensions, thus improving spatial understanding of in-person biomolecular interactions. This would prove particularly useful in molecular docking activities and structural fitting in cryo-EM maps. In addition, ChimeraX has introduced new segmentation, sequence analysis, and volume rendering tools, which extend the resource possible in many areas.

2.2.2 Contribution to Molecular Visualization

It is the only molecular visualization tool that can deal with huge datasets as well as different data types, such as atomic structure, electron microscopy density map, and

multichannel light microscopy data-while PyMOL and VMD cannot do that. It is highly modular with its app store for plugins, conducive to customization and extension, making it versatile enough for most research needs.

The performance of ChimeraX has also been seen as it has been used in research on COVID-19 by analyzing more than 200 protein structures of SARS-CoV-2 published by public databases. Larger and better representations deliver fast analyses and, as such, ChimeraX proves a significant body in virus mechanisms and knowledge for drug discovery.

2.2.3 Relevance to Current Project

The parallels in the objectives of this project and the capabilities of ChimeraX pertain mostly to the visualization and analysis of predicted protein structure. Considering that it has good interactive exploration tools for processing large-scale cryo-EM data, it fits perfectly in the search for the structure of the Nipah glycoprotein and complementary protein design.

2.2.4 Conclusion

The latest advances developed by UCSF ChimeraX prove the importance of visualization in molecular research. The most innovative functions for structural biology combined with user-friendliness put it at the core of the discipline, linking experimental data to computation modeling.

2.3 Identification of Binding Site and Complementary Protein Design [3]

The process of identifying binding sites and designing complementing proteins entails the use of several computational tools. It captures the geometric features of proteins, including concaves and convexe surfaces, and also aids in defining the topology of the proteins and potentially identifying the binding regions. The Dictionary of Secondary Structure of Proteins (DSSP) is a valuable source of information about the protein's secondary structure and helps to locate the function and/or binding domains of the protein. In addition, MPNN advances the approach of complementing proteins' design by modeling protein structures and their interaction on a residue base, enhancing binding potency and specificity. Moreover, surface scoring methods examine the electrostatic and hydropho-

bic properties of the residues on the interface of the protein and facilitate the interface design. Therefore, by utilizing these methods, the peptide binding sites can be identified and designed specific complementing proteins suitable for biomolecular or therapeutic purposes. These strategies also make it possible to reduce the time spent on designing suitable protein interactions.

2.3.1 Alpha Shape Theory for Binding Site Detection

Alpha shape formulations fit objects to point clouds and thus have a practical application in computational geometry. Alpha shapes in the context of a molecular structure are applied for modeling the 3D outline of a registered protein by using its atomic coordinates as points within space. This involves the construction of a hierarchy of simplicial complexes which are polygons and polyhedra generalizations of solids. This hierarchy may vary in detail by controlling the alpha parameter.

The alpha shape possesses detail-oriented surface elements and surface indentation which range from low to medium alpha values while comprising minor scale surface detail complemented by medium to high alpha. Hence Concavities, Claustrophobic narrow sections and grooves, Pockets with shape indentations suitable for ligands or protein interaction, are some of the methods that can be utilized to analyze the complexes. Ligand sites or molecules that have a shape that is opposite or charge that is opposite the complementary site are often located in these concave areas.

Application to Identifying Potential Binding Sites: The theory of alpha shape application offers an elaborate and flexible interpretation of the surfaces of proteins and makes it an excellent candidate for ligand binding site identification. For example, using the alpha shape of a protein, researchers may be able to identify regions with concavity as potential natural ligand binding pockets or for binding of other proteins. Such cavities are mostly found at sites where the molecular surface is relatively more dynamic or suitable for other molecules, like the active sites of enzymes or sites of antigen-antibody complexes.

2.3.2 Molecular surface algorithm

The property mapping algorithms of molecular surfaces are very crucial in describing the protein surface, which means understanding molecular recognition sites, functional bind-

ing sites, and the design of more complementary proteins. Their relationship, such as the electro- as well as hydrophobic, has a great impact on the binding. Possession of surface contour mapping has revealed to show, at least at a gross level, which parts of the protein surface have positive or negative electrostatic potentials. For example, with charged molecules such as ions or some small drug candidates, these regions are very informative in giving insights into protein-ligand interaction. The clustering of the nonpolar amino acids can be defined by the hydrophobic interactions that, from a molecular point of view, act very much the same as those established between the protein surface and the solvent; the areas where the binding partners interact are often referred to as hydrophobic patches and include small organic molecules. Another thing would be solvent-accessible surface area (SASA), which calculates the surface area of the solvent-exposed protein part of the molecule, thus indicating potential interaction sites.

This specific surface algorithms are used for predicting the functional binding sites, which are regions on a protein that allow for molecular interaction. Surface features, such as electrostatic potential and hydrophobicity, are utilized by these algorithms to predict the binding pockets and, hence, identify sites that may come into play when determining protein-protein or protein-ligand interactions. Further, surface algorithms are considered as an instrumental tool in designing better docking interfaces for protein-ligand or protein-protein interactions and understanding how these interactions should be optimized. For example, the improvement of the binding site at which the two molecular entities bind together would significantly improve the binding affinity by optimizing either electrostatic complementarity or enhancing hydrophobic contacts. Surface algorithms are also used for analyzing druggable sites in drug discovery, which is important for the design of therapeutics. Such algorithms usually provide ways for effective identification of binding sites and refinement of molecular interfaces and optimization of interactions between proteins; this makes them indispensable to computational biology, drug design, and protein engineering.

Such surface algorithms serve to predict such functional binding sites, i.e., the regions on a protein that allow for molecular interactivity. Features of the surface such as electrostatic potential and hydrophobicity are employed by these algorithms to predict the binding pockets and identify the sites involved in protein-protein or protein-ligand interactions. Such surface algorithms are also an important resource in designing excellent docking interfaces with respect to protein-ligand or protein-protein interactions,

understanding how such interactions could be optimized. For example, optimizing electrostatic complementarity or enhancing hydrophobic contacts would generally improve binding affinity at the interface where the two molecular entities come together. Surface algorithms can also be used to explore druggable sites in drug discovery, an essential element for the development of therapeutics. In sum, these algorithms allow precise binding site identification, refinement of molecular interfaces, and optimization of protein interactions and hence are essential in computational biology, intramolecular design, and protein engineering.

2.3.3 Dictionary of Secondary Structure of Proteins

The Dictionary of Secondary Structure of Proteins (DSSP) is a key computational tool for characterizing protein secondary structures, providing a detailed classification of protein segments based on their conformational states. DSSP assigns specific structural labels to each residue in a protein, categorizing them into elements such as alpha helices, beta sheets, turns, and loops based on hydrogen bonding patterns, dihedral angles, and other structural criteria. This structural information is crucial for understanding how proteins fold and how different regions contribute to their overall function. By analyzing the secondary structure elements, researchers gain insights into the stability and flexibility of a protein, which directly impacts its interaction with other molecules.

DSSP plays a critical role in identifying functionally relevant binding regions on proteins. Many binding sites, whether for small molecules, other proteins, or DNA, are located within or adjacent to specific secondary structure elements. For instance, alpha helices and beta sheets often participate in molecular recognition and can be involved in binding interactions due to their spatial organization and structural rigidity. Understanding the arrangement of these elements helps identify functional domains—regions of the protein that are likely to be involved in interactions. For example, surface-exposed loops, often rich in polar or charged residues, can be particularly important for binding, as they are more flexible and accessible for ligand docking. Similarly, beta sheets in certain conformations may form interaction sites for other proteins or molecules.

2.3.4 Protein Message Passing Neural Network(ProteinMPNN)

They have formed a powerful machine learning approach in which one could revolutionize protein structure prediction and interaction design, MPNNs or protein-based message-passing neural networks. These are a class of tailored neural networks that may function on their graph representations from protein structures, where nodes are made of either atoms or residues due to their substances, with interactions such as covalent bonds, spatial proximity, etc. between them as edges. Message passing is the iterative way through which each node (residue) modifies its information by collecting messages from the adjacent nodes. This is what allows the entire network to learn the solution to complexity in dependencies among residues and to localized contexts. Such a method could effectively model and account for the complicated relationships within a protein as well as between two interacting proteins; thus, it can answer well in terms of predicting protein structures and binding interfaces.

MPNNs are developing very good computation models for predicting three-dimensional shapes of proteins that are significant in remediation to their function as well as how they can be acted upon by other molecules. About protein-protein interactions, these MPNNs can predict not only the overall folded state of that protein but also its most probable 'synthetic interface' where two proteins are likely to bind. Therefore, based on pattern learning across a large dataset of known protein structures and their interaction sites, MPNNs will provide accurate predictions about which residues at the surface of that protein will participate in binding and how these residues interact with the other protein.

2.3.5 Conclusion

It represents a powerful framework for understanding the protein structure and designing therapeutic interventions by integrating alpha shape theory, DSSP, ProteinMPNN, and molecular surface algorithms. Alpha shape theory provides highly detailed models of protein surfaces; hence, critical binding pockets may be identified for drug design. The strength of this model is enhanced through DSSP, which gives precise secondary structure annotations to identify functionally relevant regions for interaction and inhibition. ProteinMPNN is able to design complementary protein sequences that can form targeted

inhibitors with high specificity against viral or protein-protein interactions. Meanwhile, molecular surface algorithms have deepened the analysis of electrostatic and hydrophobic interactions, thus improving the identification of druggable sites and optimizing protein-ligand binding interfaces. These tools together are making drug discovery more efficient while providing rapid and resource-effective ways of developing novel therapeutics for many emerging infectious diseases.

2.4 End-to-End Differentiable Learning of Protein Structure [4]

Structural biology has struggled for a long time with the complex task of protein structure prediction. Most of the time-consuming and resource demanding, X-ray crystallography and cryo-electron microscopy are the conventional methods of analysis. The methodology presented in Al-Quraishi (2019) end to end differentiable learning, enables protein structure prediction by explicitly learning the mapping from the protein sequences to protein three-dimensional structures, using neural networks. This neural network method eliminates the sequential processes of established techniques and provides a fast, direct, and scalable solution to the problem. The novelty is to merge evolutionary data derived from multiple sequence alignments with contemporary deep learning, thus facilitating progress in structural biology.

2.4.1 Methodology and Innovation

This new model may provide valuable tools for the future of protein and peptide design with computer systems allowing for the prediction of the backbone structure from the sequence of the polypeptide chain and eliminating the need for elaborate features at the same time. By this automated and efficient approach, they are able to help transition into a whole new world of computational biology. From the sequence, it too required the outlines of predicted backbone structures in order to be able to set them directly on the atomic coordinates and ots of amino acid.

2.4.2 Contribution to Protein Structure Prediction

The overall framework, in contrast to customary procedures, represents a paradigm shift as it simplifies the entire procedure and does away with intermediate stages such as man-

ual feature engineering. This allows for better parallel efficiency and decreases the time needed for computation, making it superior to more traditional cycles. It provides a good option over available deep neural network-based predictors in that it does not sacrifice interpretability for differentiability. Additionally, the method provides outstanding results on standard databases, providing results comparable to standard techniques and reinforcing its application in computational biology.

2.4.3 Relevance to Current Project

This approach focuses on achieving the aims of predicting the Nipah virus glycoprotein structure within your project as efficiently as possible. Its easy operating paradigm directly translates to the protein backbone parts of the molecule, which can be plugged into the downstream applications such as AlphaFold2 without additional changes. The method reduces load expenses and therefore is suited for fast-paced and wide-ranging protein design projects. Also, the structural estimation that was produced is also required for visualization and stability examination with PyMOL or ChimeraX that meets your project's goal of analyzing viral proteins very well.

2.4.4 Conclusion

AlQuraishi's approach for end-to-end differentiable learning of protein structure prediction models is such a big leap in computational biology. It collapses the multi step modeling process of predicting three-dimensional structures from protein sequences into a direct problem, thanks to deep learning. Because of this, traditional methods are often inefficient You must have a fully differentiable structure for seamless learning of the architecture's ability to use large scale computations.

In the case of the current project, this method is Self-Contained and relevant in That it predicts the glycoprotein structure for Nipah viruses in a manner that complements existing workflows for protein engineering over designing. Its strength especially in head prediction is synergistic for downstream such as protein-protein interaction prediction, structure stability prediction, and visualization with PyMOL or ChimeraX.

More importantly, this approach broadens the scope of structural biology experiments, such as the analysis of protein-protein interaction, drug development, and therapeutic protein construction. Therefore, decision-making activity has become simpler, less costly,

and more tolerant to changes in technology that have emerged during the implementation of this approach. In the end, such structural biology frameworks assist in understanding how solving critical linear problems works.

2.5 Summary and Gaps Identified

Table 2.1: Summary of Literature Survey

Paper	**Advantages**	**Disadvantages**
A-Prot: protein structure modeling using MSA transformer(2022)	• Improved Accuracy with Transformers • Efficiency in Handling MSAs	• Computational Complexity • Dependence on MSA Quality • Limited Availability of MSAs
UCSF ChimeraX	• Real-time rendering with ambient occlusion and shadow mapping • Efficient handling of large datasets	• Limited support for custom scripting compared to other tools • High system requirements for VR applications
Robust deep learning–based protein sequence design using ProteinMPNN.(2022)	• High Precision • Scalability	• Dependency on Structural Data • Limited Experimental Validation

Paper	Advantages	Disadvantages
End-to-End Differentiable Learning of Protein Structure. 2019	• Fully differentiable framework • End-to-end approach	• Complexity • Generalization

Gaps Identified

1. **Limited Accuracy in Complex Protein Interactions**
   While AlphaFold2 has made significant strides in predicting the 3D structures of individual proteins, it still faces challenges in predicting protein-protein interactions with high accuracy. Many biological processes involve complex protein interactions that are not fully captured by current models, leading to gaps in the ability to predict how multiple proteins may function together.

2. **Lack of Experimental Validation in Drug Design**
   Although protein structure prediction tools like AlphaFold2 and ProteinMPNN offer promising results in simulating protein structures and designing new sequences, the lack of experimental validation in real-world settings limits their application in drug discovery. The prediction models need to be more closely integrated with experimental data to ensure their practical utility in therapeutic development.

3. **Insufficient Focus on Structural Stability in Protein Design**
   Current protein design models, such as ProteinMPNN, emphasize generating novel protein sequences but often overlook the stability of the designed proteins under physiological conditions. Ensuring that these proteins remain stable and functional in biological environments remains a challenge that has not been fully addressed by existing tools.

4. **Limited Integration of Visualization and Simulation Tools**
   While PyMOL is a powerful visualization tool for analyzing predicted protein structures, there is a lack of seamless integration between protein prediction tools and

visualization platforms. Incorporating real-time structural analysis, dynamic simulation, and visualization in a unified workflow could greatly enhance the utility of these technologies for drug discovery and protein engineering.

5. **Challenges in Handling Large-Scale Data and Complexity**

   The increasing complexity and size of biological datasets present a challenge for computational tools. Current prediction and design tools may struggle with handling large-scale protein datasets or generating accurate models for proteins with highly variable sequences, leading to inefficiencies in data processing and predictions for less-studied proteins.

Chapter 3

System Design

3.1 System Architecture

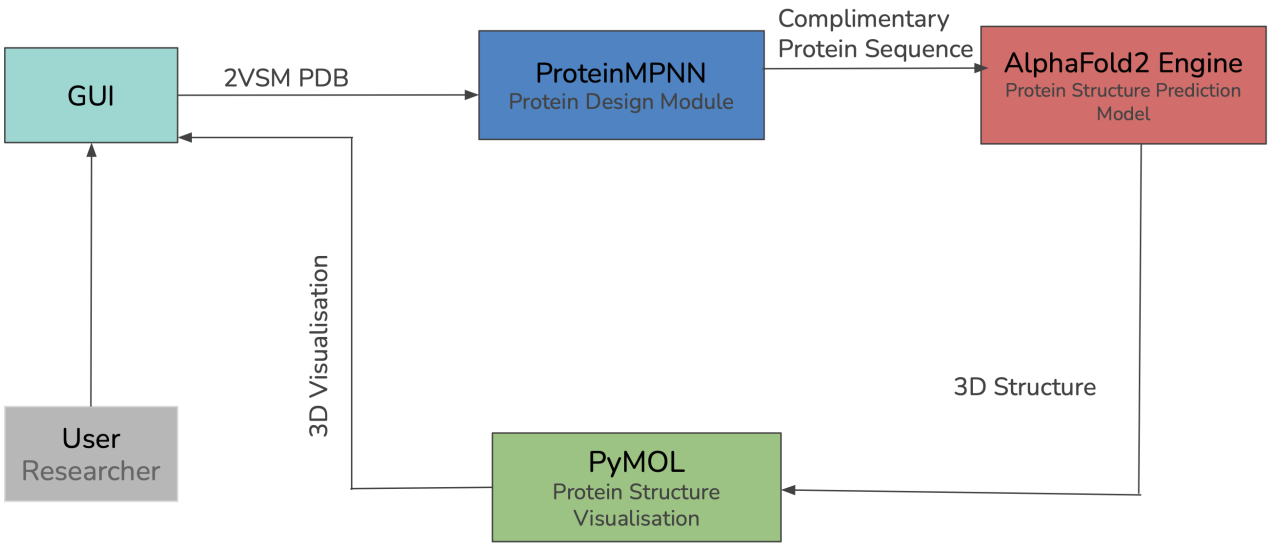


Figure 3.1: Architecture Diagram

3.2 Component Design

3.2.1 GUI

The website provides an interactive GUI that allows users to design complementary proteins, predict their 3D structures, and analyze accuracy parameters seamlessly. With an intuitive layout and real-time visualization, users can effortlessly input PDB files, generate models, and assess structural stability.

3.2.2 Alphafold2 Engine

The core model for the protein structure prediction is a program that takes in an amino acid sequence and then generates a 3D protein structure as output. Advanced AI models are used for making precise predictions of protein folding.

3.2.3 Protein MPNN

The ProteinMPNN algorithm designs complementary protein sequences according to the predicted protein structure. It is coupled with AlphaFold2 output for optimal protein design under specific functional needs and requirements.

3.2.4 PyMOL Visualization Tool

It is used for visualizing 3D protein structures. Supports research by providing insights into folding, binding sites, and interaction of proteins. Enhances the clarity of the predictions and designs produced by AlphaFold2 and ProteinMPNN.

3.3 Algorithm Design

3.3.1 Protein Structure Prediction - AlphaFold

Input:

Protein amino acid sequence in FASTA format, MSA of homologous sequences

Output:

3D protein structure in PDB format.

Step 1: Input Preprocessing

Generate per-residue features using physicochemical properties of the sequence.
Compute pairwise features .
Integrate templates to enhance structural predictions.

Step 2: Evoformer Block

```
def evoformer_block(single_repr, pair_repr, num_iterations):
```

```
    for i in range(num_iterations):
        msa_repr = msa_attention(single_repr)
        pair_repr = pair_attention(pair_repr)
        pair_repr = triangle_update(pair_repr)
    return msa_repr, pair_repr
```

Step 3: Structure Module

```
def structure_module(pair_repr, single_repr):
    predicted_structure = predict_coordinates(pair_repr, single_repr)
    return predicted_structure
```

Step 4: Iterative Refinement

Refine the predicted structure iteratively using gradient descent to minimize energy and improve structure quality.

Step 5: Confidence Scoring

Compute pLDDT scores for each residue. Output global confidence score, where high pLDDT indicates accurate predictions.

```
    def alphafold(sequence, msa, templates=None):
    single_repr, pair_repr = encode_features(sequence, msa, templates)
    msa_repr, pair_repr = evoformer_block(single_repr, pair_repr, num_iterations)
    predicted_structure = structure_module(pair_repr, msa_repr)
    refined_structure = refine_structure(predicted_structure)
    plddt_scores = compute_confidence_scores(refined_structure)
    return refined_structure, plddt_scores
```

3.3.2 Protein Design Prediction - ProteinMPNN

Input:

3D backbone structure of a protein (C, N, C coordinates for each residue).

Output:

Return: Designed protein sequence compatible with the input backbone structure.

Step 1: Input Preprocessing

Represent the backbone as a graph with:

Nodes: Residues (amino acids).

Edges: Spatial relationships between residues.

Compute:

Pairwise distances between residues.

Backbone dihedral angles (phi, psi).

Step 2: Feature Encoding

```
def encode_features(backbone_structure):
    graph = create_graph_from_backbone(backbone_structure)
    pairwise_distances = compute_distances(graph)
    dihedral_angles = compute_dihedral_angles(graph)
    node_features = encode_node_features(graph, pairwise_distances, dihedral_angles)
    edge_features = encode_edge_features(graph, pairwise_distances)
    return node_features, edge_features
```

Step 3: Iterative Node Feature Update Block for Graph Neural Network (GNN)

```
def gnn_block(node_features, edge_features, num_iterations):
    for _ in range(num_iterations):
        node_features = gnn_layer(node_features, edge_features)
    return node_features
```

Step 4: Sequence Generation

```
def sequence_generation(node_features):
    sequence_probabilities = predict_amino_acids(node_features)
    designed_sequence = sample_sequence(sequence_probabilities)
    return designed_sequence
```

Step 5: Displaying Protein Sequences Using ProteinMPNN Framework

```
def protein_mpnn(backbone_structure, num_iterations):
    node_features, edge_features = encode_features(backbone_structure)
    updated_node_features = gnn_block(node_features, edge_features, num_iterations)
    designed_sequence = sequence_generation(updated_node_features)
return designed_sequence
```

3.3.3 Stability Evaluation

Input:

Protein 3D structure (generated from AlphaFold2 or ProteinMPNN).

Output :

Stability score (low G, RMSD, and RMSF values indicate higher stability).

Step 1: Input Preprocessing

Initialize variables for energy calculation and stability metrics.

Step 2: Energy Calculation

```
def calculate_energy(protein_structure):
    energy_components = compute_energy(protein_structure)
    total_energy = sum(energy_components.values())
    return total_energy
```

Step 3: Molecular Dynamics (MD) Simulation

```
def molecular_dynamics_simulation(protein_structure):
    md_trajectory = run_md_simulation(protein_structure)
    rmsd = calculate_rmsd(md_trajectory)
    rmsf = calculate_rmsf(md_trajectory)
    return rmsd, rmsf
```

Step 4: Thermodynamic Stability

Assess free energy changes at different temperatures and pH levels.

Identify potential misfolding regions or structural instability.

```
def thermodynamic_stability(protein_structure, temperature=300, pH=7.4):
    free_energy = compute_free_energy(protein_structure, conditions={"temperature":
return free_energy
```

3.4 Use Case Diagrams

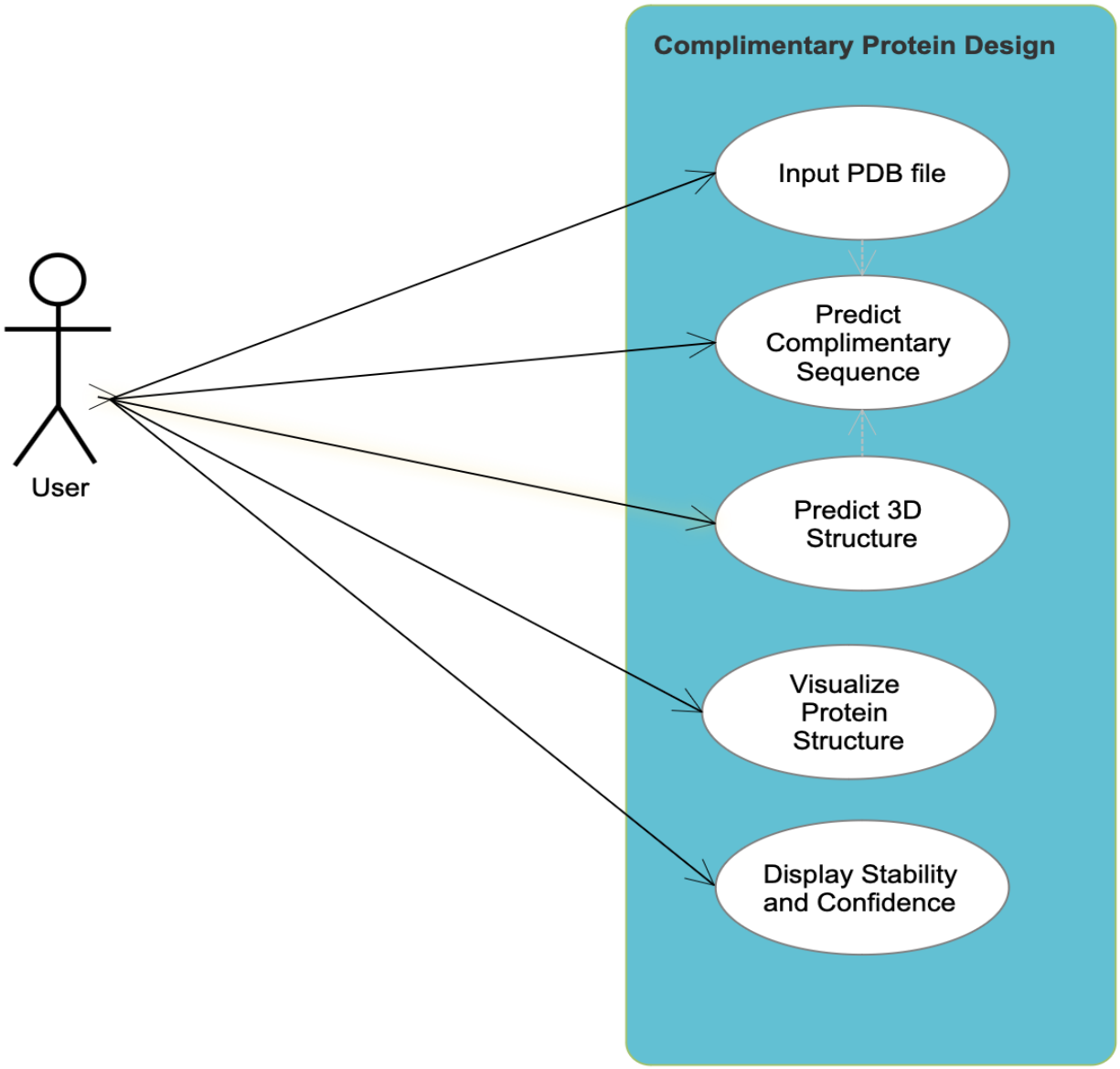


Figure 3.2: Use Case Diagram

3.5 Tools and Technologies

3.5.1 Software Requirements

- AlphaFold2: Used for protein structure prediction, leveraging deep learning models.
- PRODIGY: is a web server used to predict binding affinity (G, dissociation constant

K) of protein-protein complexes based on their 3D structure.

- ClusPro: is a molecular docking tool that predicts how two proteins interact by generating multiple binding poses.
- ProteinMPNN: Used for designing complimentary protein sequence from 3D structure of proteins.
- PyMOL: Utilized for visualizing the predicted protein structures, providing high-resolution molecular graphics.
- Python: Programming language for implementing the algorithms and handling data processing.Version 3.8
- Linux/Ubuntu or Windows: Operating systems compatible with AlphaFold2 and related software tools.
- Visual Studio Code/Jupyter Notebook: For writing and running Python code, especially for analysis and visualization tasks.

3.5.2 Hardware Requirements

- Windows PC (Minimum Specifications).
- Processor: Intel Core i5 or AMD Ryzen 5 (or higher).
- Memory: 16 GB RAM (to handle large datasets and model computations).
- Storage: 512 GB SSD (for fast data access and software execution).
- Graphics: NVIDIA GTX 1060 or higher with at least 6 GB VRAM (for efficient processing in PyMOL and deep learning tasks) .

3.6 Dataset Identified

The dataset used for this project consists of the 3D structure of the Nipah virus glycoprotein, specifically the 2VSM structure obtained from the Protein Data Bank (PDB). This PDB file contains atomic coordinates and structural details of the glycoprotein, which are essential for analyzing potential binding sites and designing complementary proteins.

The complementary protein sequences generated using ProteinMPNN are used as inputs to AlphaFold2 for predicting their 3D structures. The input files for AlphaFold2 must be in FASTA format, so the generated sequences are formatted accordingly. The dataset is preprocessed to ensure compatibility with AlphaFold2, including formatting the sequences correctly and removing any unnecessary characters or whitespaces.

```
>sp|Q9IH62|GLYCP_NIPAV Glycoprotein G OS=Nipah virus OX=3052225 GN=G PE=1 SV=1
MPAENKKVRFENTTSDKGKIPSKVIKSYYGTMDIKKINEGLLDSKILSAFNTVIALLGSI
VIIVMNIMIIQNYTRSTDNQAVIKDALQGIQQQIKGLADKIGTEIGPKVSLIDTSSTITI
PANIGLLGSKISQSTASINENVNEKCKFTLPPLKIHECNISCPNPLPFREYRPQTEGVSN
LVGLPNNICLQKTSNQILKPKLISYTLPVVGQSGTCITDPLLAMDEGYFAYSHLERIGSC
SRGVSKQRIIGVGEVLDRGDEVPSLFMTNVWTPPNPNTVYHCSAVYNNEFYYVLCAVSTV
GDPILNSTYWSGSLMMTRLAVKPKSNGGGYNQHQLALRSIEKGRYDKVMPYGPSGIKQGD
TLYFPAVGFLVRTEFKYNDSNCPITKCQYSKPENCRLSMGIRPNSHYILRSGLLKYNLSD
GENPKVVFIEISDQRLSIGSPSKIYDSLGQPVFYQASFSWDTMIKFGDVLTVNPLVVNWR
NNTVISRPGQSQCPRFNTCPEICWEGVYNDAFLIDRINWISAGVFLDSNQTAENPVFTVF
KDNEILYRAQLASEDTNAQKTITNCFLLKNKIWCISLVEIYDTGDNVIRPKLFAVKIPEQ
CT
```

Figure 3.3: Amino acid sequence of Glycoprotein of Nipah virus

3.7 Module Division

3.7.1 Complementary Protein Design Using ProteinMPNN

The module is centered on creating potential complementary proteins that can inhibit the activity of viral proteins. The process starts by providing the PDB file to proteinMPNN, which acts as the input. The ProteinMPNN algorithmthen is used to generate sequences of complementary proteins, specifically tailored to bind to and inhibit the target viral protein. The result is a set of designed protein sequences that may effectively block the function of the viral protein, setting the stage for further validation and testing.

3.7.2 3D Structure Prediction Using AlphaFold2

This module aims to predict the 3D structure of the generated complementary protein sequence using AlphaFold2. The process begins by providing the protein sequence as input to AlphaFold2, which then analyzes the sequence using deep learning techniques. AlphaFold2 employs advanced attention mechanisms to accurately predict the protein's

3D conformation. The output is a 3D structure of the complementary protein in PDB (Protein Data Bank) format, which can be further analyzed and visualized using tools such as PyMOL.

3.7.3 Molecular Visualization

The module focuses on visualizing predicted protein structures and their interactions to enhance our understanding of molecular dynamics. It takes in 3D structures produced by AlphaFold2 along with docking results. We use tools like PyMOL to generate detailed, high-resolution molecular visualizations that emphasize important structural features and interactions. The result is a set of visually engaging representations of protein structures and their interactions, which aids in further analysis and interpretation.

3.7.4 Molecular Interaction Analysis

The module is centered on evaluating protein-protein interactions using docking simulations tool ClusPro. The process entails conducting docking studies to examine how complementary proteins attach to the target viral protein. We analyze binding affinities using the PRODIGY and interaction sites to gauge the strength and specificity of these interactions. Using the results from the docking, we optimize the design of complementary proteins to improve binding efficiency and inhibitory potential. This iterative method guarantees the creation of effective inhibitors that possess high specificity and stability.

3.8 Work break down

Aparna A R

- Molecular Visualization
- Analysed Residue Flexibility,Contap Map,SASA values.

Aparna Sajeev

- Analysed Stability.
- Finding Accuracy.

Ashley K Alex

- Complementary Protein Design.
- Analyze binding affinity and perform Docking .

Athira J

- 3D Structure Prediction.
- Complementary Protein Design .

3.9 Key Deliverables

Expected Output

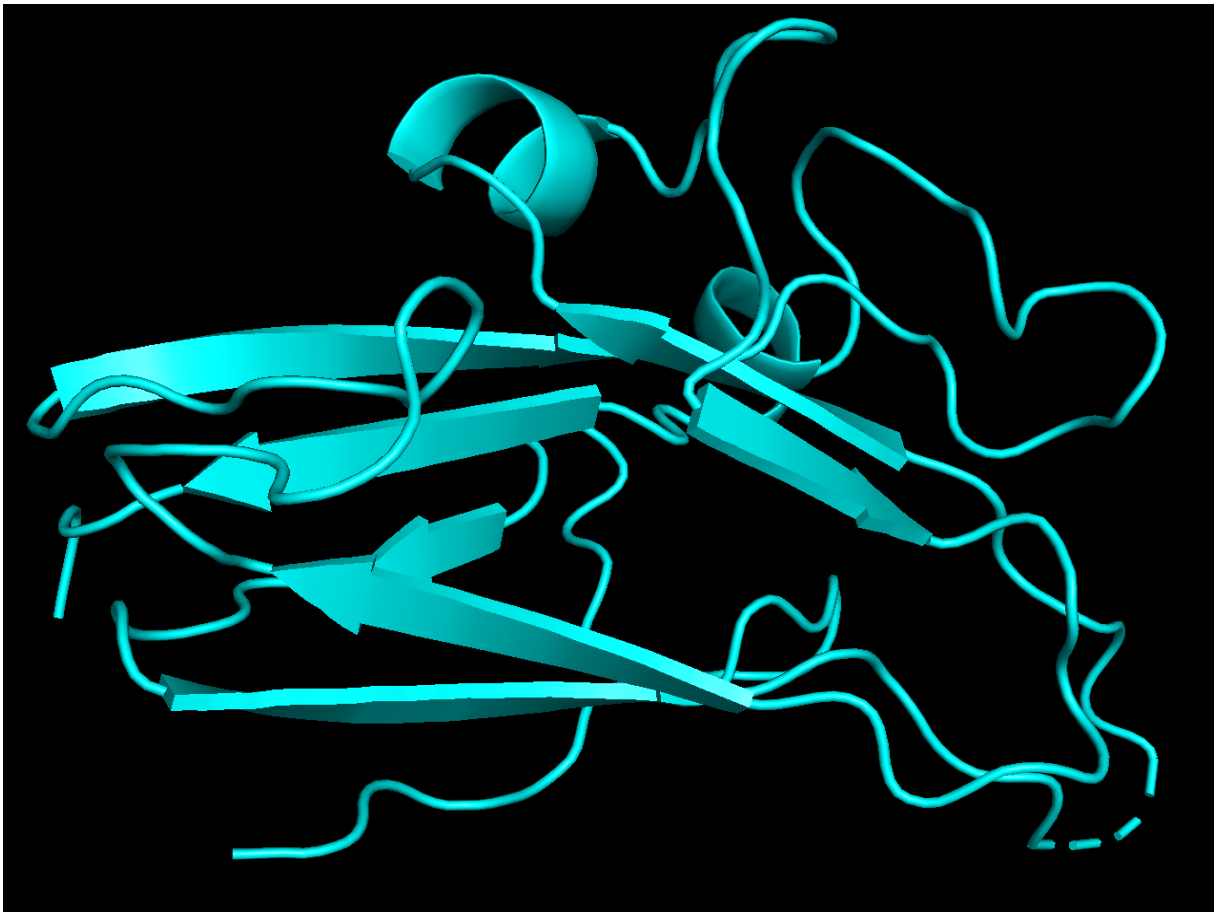


Figure 3.4: 3D predicted structure of glycoprotein of Nipah virus.

3.10 Project Timeline

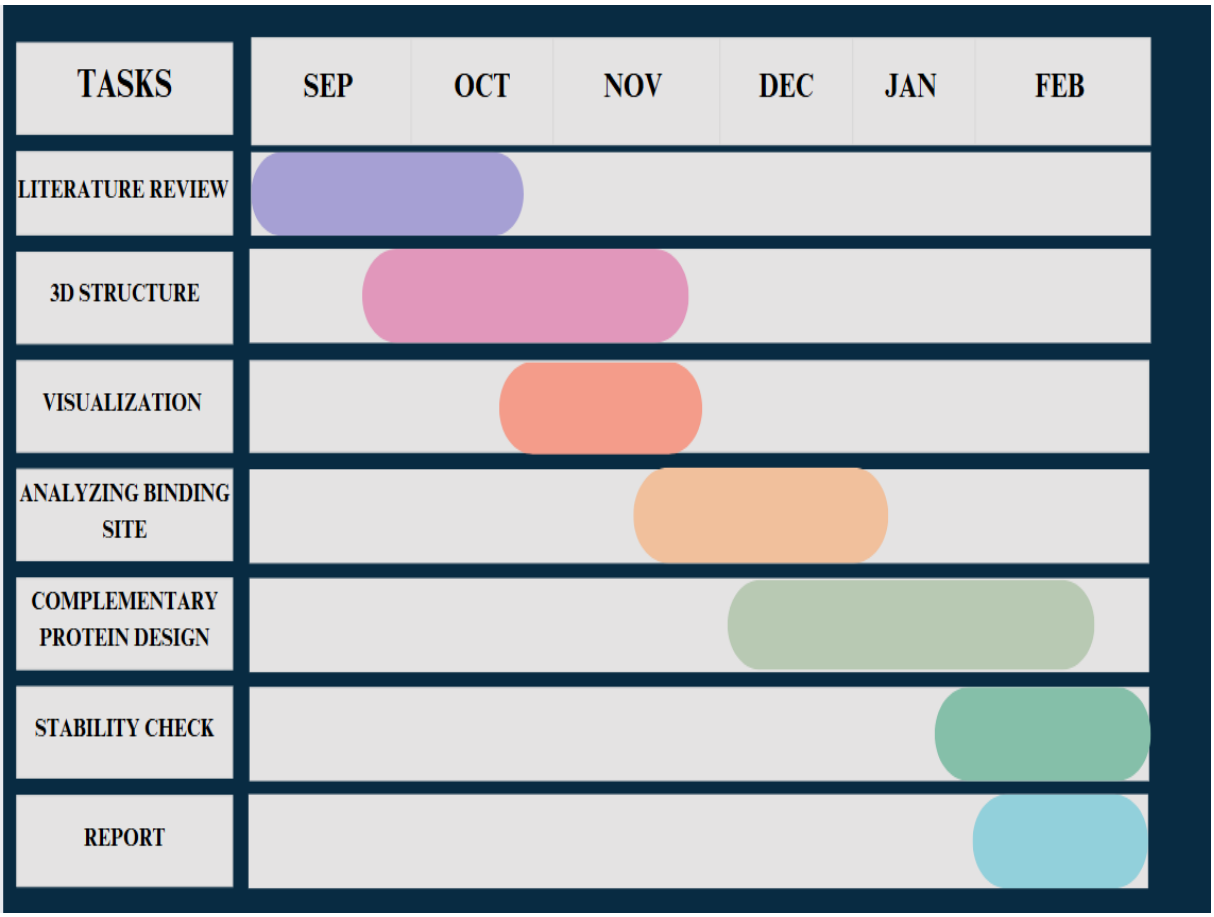


Figure 3.5: Gantt Chart.

Chapter 4

Results and Discussions

4.1 Introduction

This chapter discusses the results from computational analysis of the Nipah virus glycoprotein (NiV-G) and its designed complementary protein. The AlphaFold2-modeled predicted 3D structure of NiV-G reveals information on its topology and receptor-binding regions, which are important for viral entry. Computationally, a complementary protein was designed using ProteinMPNN based on this structural knowledge to inhibit NiV-G function. Structural compatibility and stability of the constructed inhibitor were additionally confirmed by using AlphaFold2. Furthermore, docking simulations and PRODIGY-based binding affinity predictions assist in evaluating the interaction strength between the designed inhibitory protein and NiV-G.The B-factor heatmap, contact map, hydrophobicity score, and SASA values of amino acid residues have been analyzed to understand their flexibility and chemical properties. The results of this chapter are a basis for the rational development of antiviral drugs for Nipah virus glycoproteins.

4.2 3D Structure of Nipah virus glycoprotein

The predicted 3D structure of the Nipah virus glycoprotein (NiV-G) from AlphaFold2. This structure is involved in viral entry by promoting host cell receptor attachment and, therefore, a primary target for inhibition. The predicted model presents the overall topology, including the beta-propeller fold typical of paramyxovirus attachment glycoproteins. Structural aspects such as receptor-binding sites, probable active sites, and conformational heterogeneity Informational flexibility is emphasized, facilitating the identification of residues important for inhibitor design. Visualization is a basis for computational analysis to construct a complementary protein that can bind and inhibit NiV-G function.

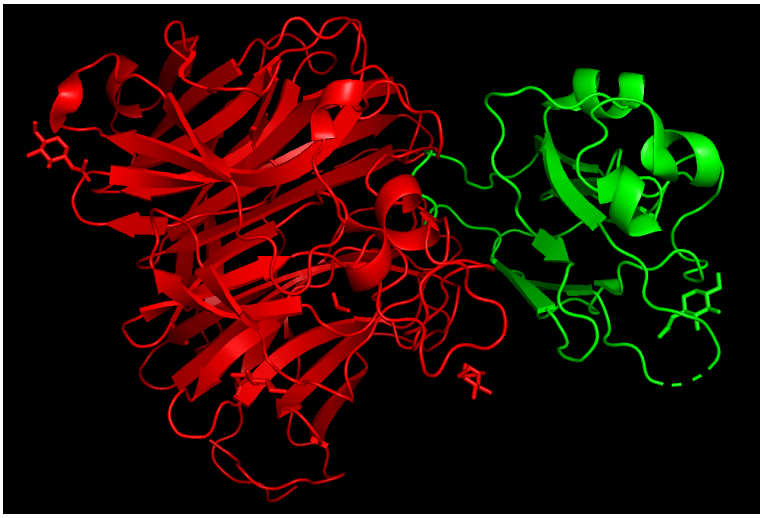


Figure 4.1: Nipah virus glycoprotein

4.3 Complementary Protein structure

To inhibit NiV-G activity, the complementary protein was computationally engineered using ProteinMPNN, from the predicted glycoprotein structure(its active sites and binding sites). The engineered protein is designed to bind to the active or receptor-binding sites of NiV-G, interfering with host cell interaction. AlphaFold2 was also employed to predict the 3D structure of the engineered inhibitory protein for structural compatibility and stability. The visualization of this inhibitory protein in complex with NiV-G sheds light on possible interactions, which can further be optimized for higher binding affinity and inhibitory efficacy.Thus, it aids in laying the basis for computational protein design strategies to create new antiviral drugs against Nipah virus glycoproteins.

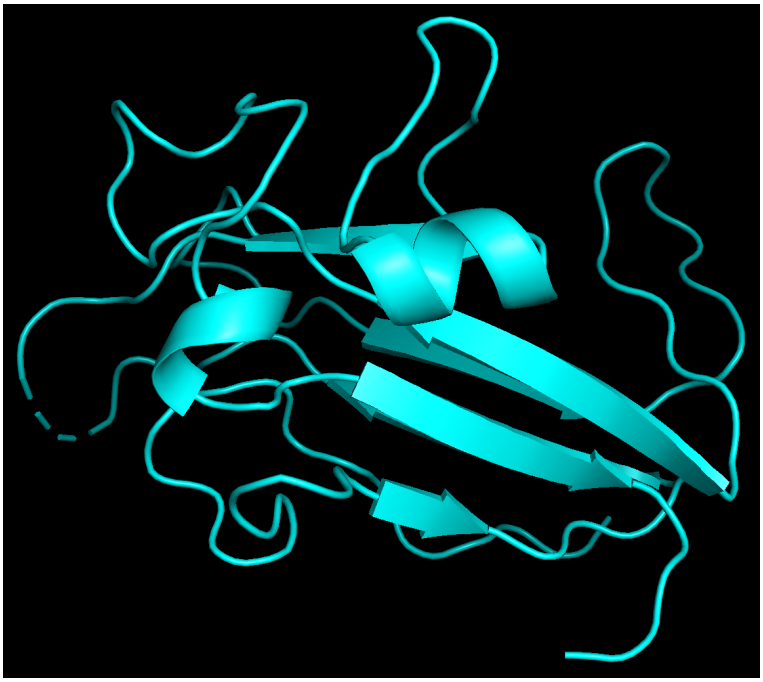


Figure 4.2: Complementary Protein

4.3.1 Binding Affinity Scores

Binding affinity and K_d prediction values from PRODIGY for the protein-protein complex predicted by ClusPro show a good interaction between the inhibitory protein designed and the Nipah virus glycoprotein (NiV-G). The free energy of binding (G) of -10.1 kcal/mol and a dissociation constant (K_d) of 3.8e-08 M reveal high binding affinity. Different types of interactions, such as charged-charged (5), charged-polar (4), charged-apolar (16), polar-polar (1), polar-apolar (9), and apolar-apolar (9), indicate the existence of electrostatic, hydrophobic, and van der Waals forces. Moreover, the non-interacting surface (NIS) values, 26.03 for charged residues and 32.92 for apolar residues, also indicate a good-balanced interaction profile. These findings indicate that the inhibitory protein built can form a stable and optimal complex with NiV-G, a lead that is ripe for optimization and experimental verification.

BINDING AFFINITY AND K_D PREDICTION

Protein-protein complex	ΔG (kcal mol^{-1})	K_d (M) at °C	ICs charged-charged	ICs charged-polar	ICs charged-apolar	ICs polar-polar	ICs polar-apolar	ICs apolar-apolar	NIS charged	NIS apolar
model_000_16	-10.1	3.8e-08	5	4	16	1	9	9	26.03	32.92

Figure 4.3: Binding Affinity result using PRODIGY

4.3.2 B-factor heatmap

A B-factor heatmap is a visual representation of the flexibility of atoms or residues in a protein structure. Higher B-factor values suggest greater flexibility or disorder, commonly observed in loops and surface-exposed regions, whereas lower values indicate rigid and well-structured areas, such as the protein core or secondary structural elements like alpha-helices and beta-sheets. By using a color gradient, a B-factor heatmap helps researchers identify flexible and stable regions within a protein, providing insights into protein dynamics, structural stability, and potential binding sites.[5]

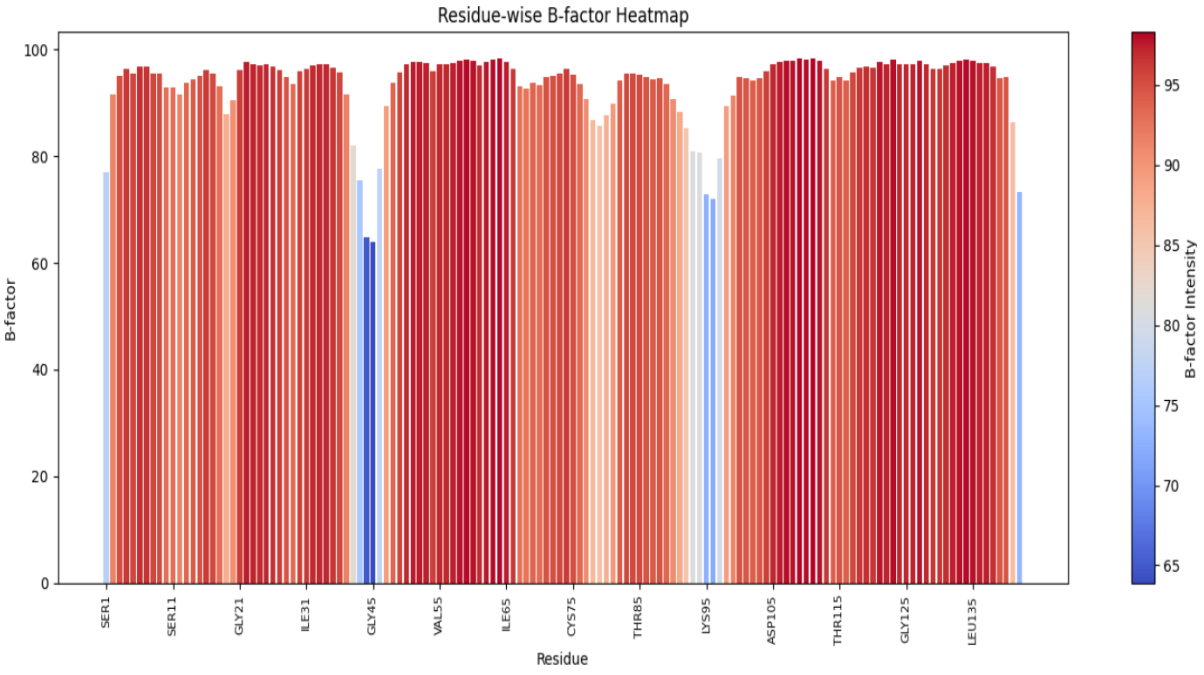


Figure 4.4: B-factor heatmap showing residue flexibility

4.3.3 Residue Contact map

A contact map is a two-dimensional representation of residue-residue interactions within a protein structure. It is a matrix where each element indicates whether two residues are in close spatial proximity, typically based on a distance threshold. Contact maps help in understanding protein folding, stability, and domain organization by revealing key interactions that maintain the structural integrity of the protein.In a contact map, dark regions represent strong or frequent contacts between amino acid residues, indicating close spatial proximity. These regions often correspond to secondary structures like alpha-helices and beta-sheets, where residues are tightly packed due to hydrogen bonding and other stabilizing interactions.Conversely, lighter regions or blank spaces indicate residues that are far apart or have weak or no interactions.[6]

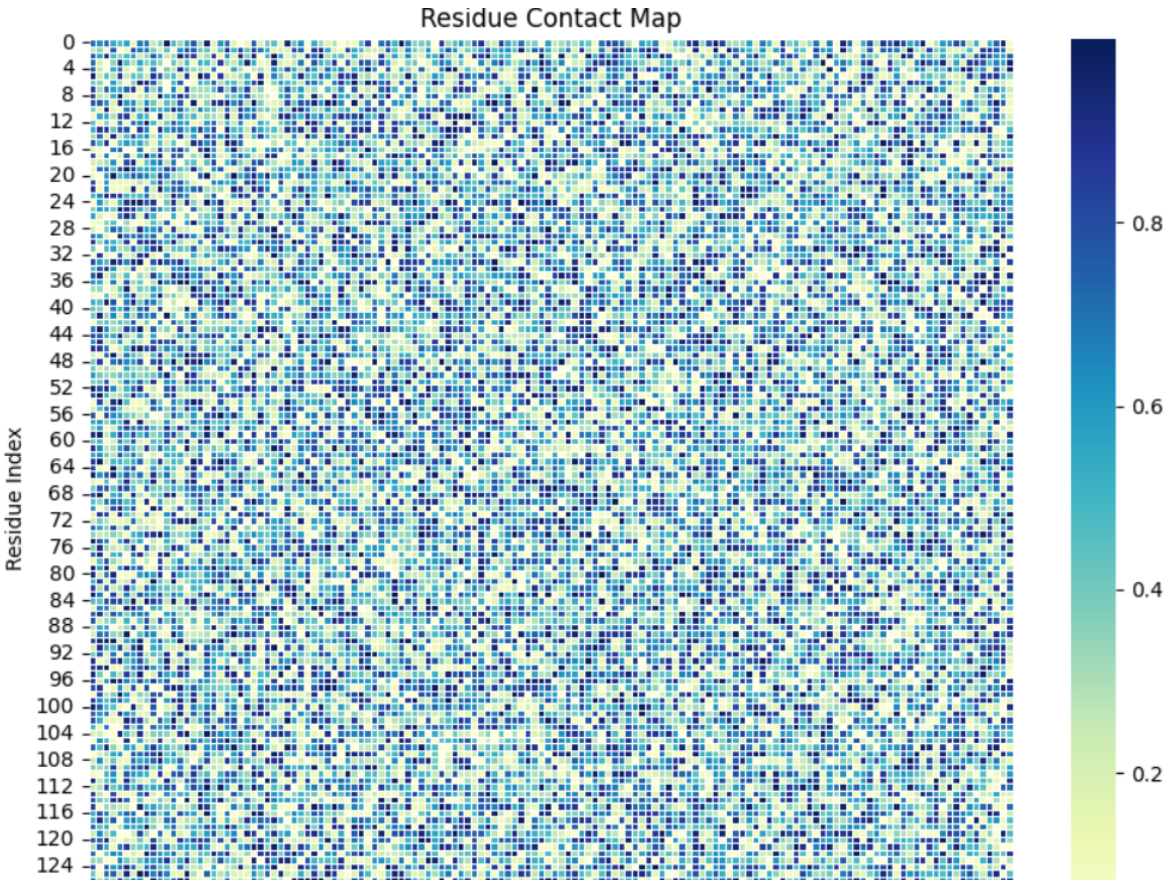


Figure 4.5: contact map

4.3.4 Hydrophobicity score

A hydrophobicity score quantifies the tendency of an amino acid residue to be hydrophobic (water-repelling) or hydrophilic (water-attracting). Each amino acid is assigned a numerical value based on its affinity for water, with higher scores indicating more hydrophobic residues (e.g., leucine, isoleucine) and lower or negative scores representing hydrophilic residues (e.g., lysine, arginine). Hydrophobicity plays a crucial role in protein folding, stability, and interactions, as hydrophobic residues are typically buried inside the protein core, while hydrophilic residues are exposed to the solvent. Hydrophobicity scales, such as the Kyte-Doolittle , are commonly used to analyze protein structures and predict membrane-spanning regions or binding sites.[7]

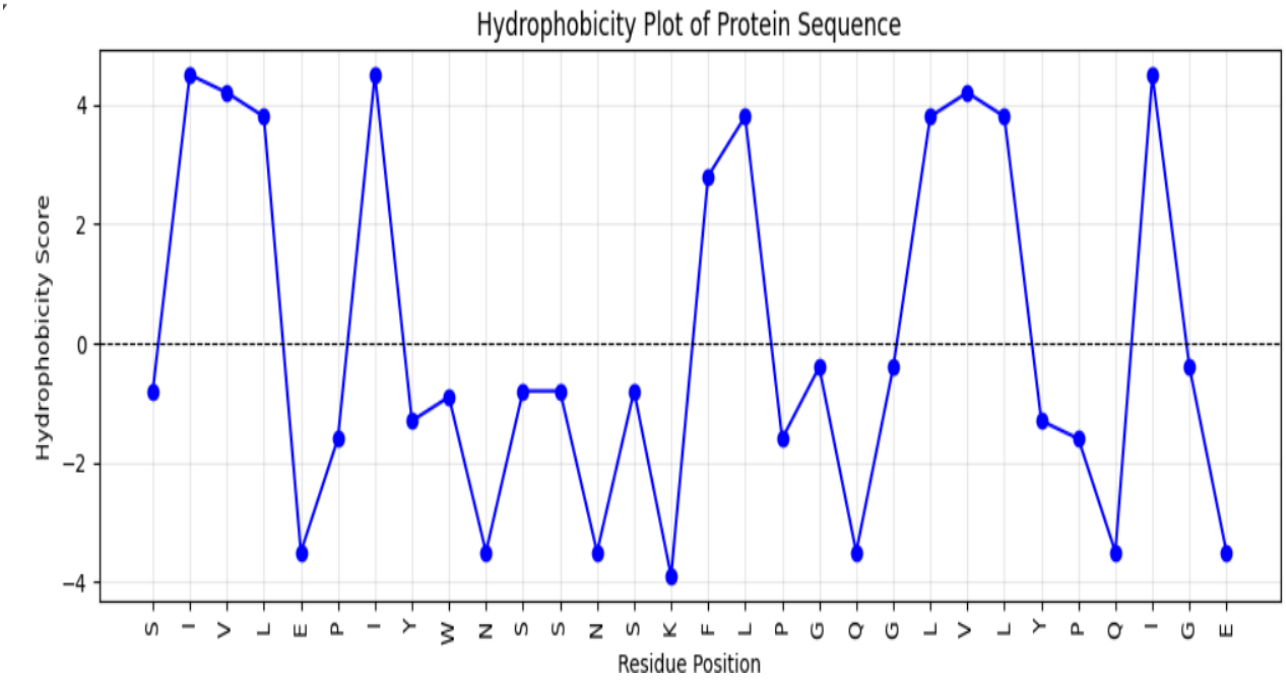


Figure 4.6: Hydrophobicity plot showing the hydrophobicity score at each residue position

4.3.5 Solvent Accessible Surface Area

Solvent-Accessible Surface Area (SASA) measures the surface area of a protein that is accessible to a solvent, typically water. It is calculated by simulating a rolling probe sphere (usually 1.4 Å in radius) over the protein structure, identifying exposed and buried residues. SASA values help in understanding protein folding, stability, and interactions,

as higher SASA indicates surface-exposed residues, often involved in binding or solvent interactions, while lower SASA suggests buried residues contributing to structural stability. Factors such as residue hydrophobicity, protein conformation, and ligand binding influence SASA. [8]

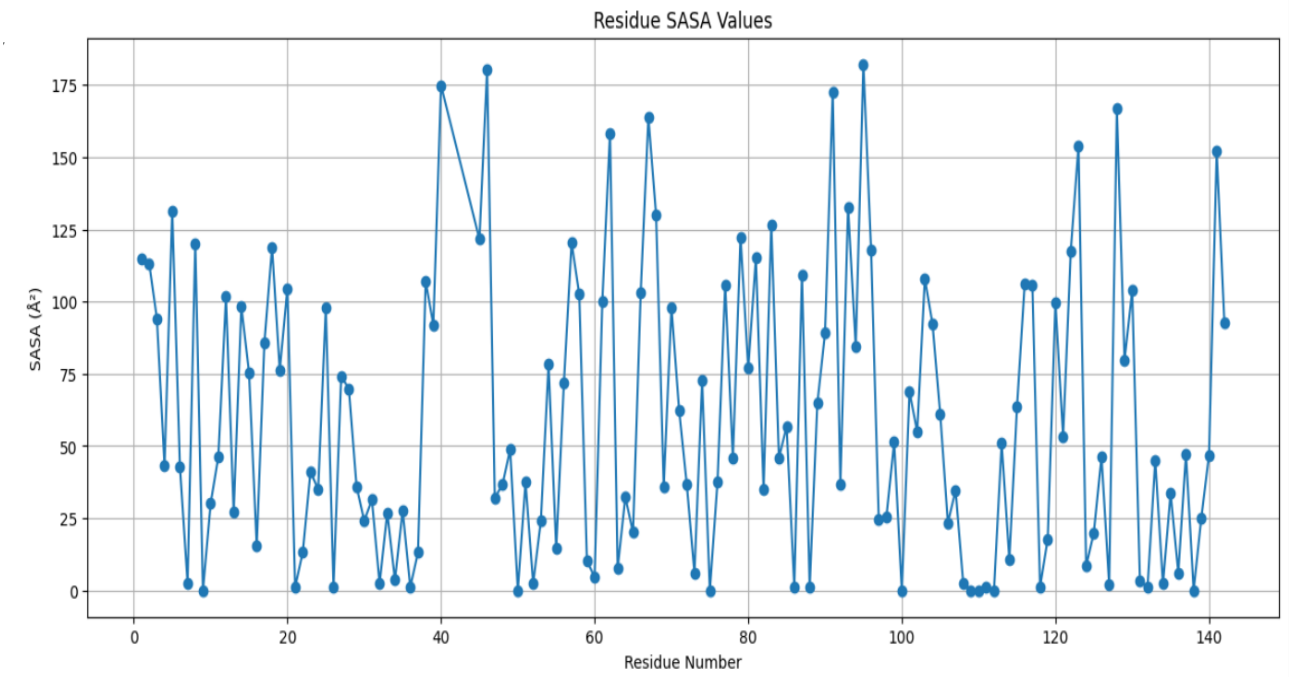


Figure 4.7: Plot showing the SASA value at each residue position

4.3.6 Nipah virus glycoprotein-Complemetary protein Complex

The 3D structure of the complementary protein-glycoprotein complex, is visualized using PyMOL. The glycoprotein structure, obtained through AlphaFold2 predictions, is shown in its stable conformation, while the complementary protein, designed using ProteinMPNN, is displayed in a contrasting color to highlight the binding interface. The visualization emphasizes the spatial interaction between the two proteins, offering insights into potential binding sites and interaction dynamics, crucial for understanding inhibitory effects and guiding further drug discovery efforts.

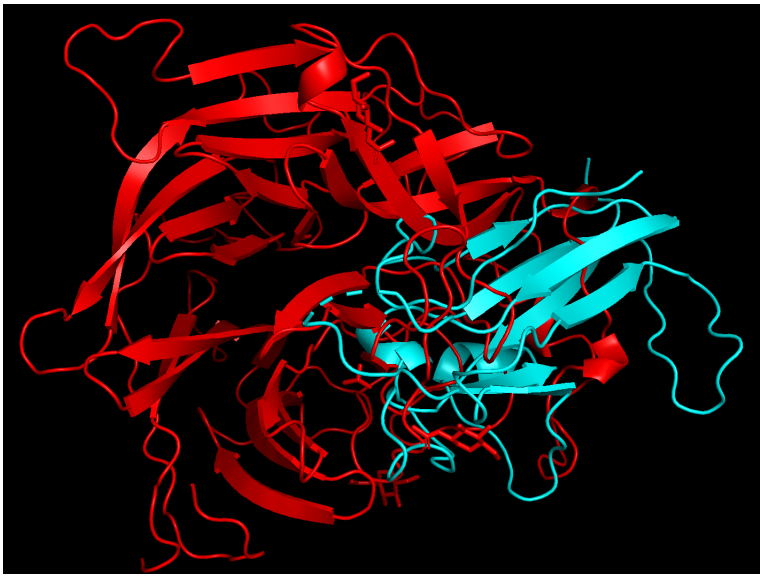


Figure 4.8: Predicted binding interaction between the designed protein and the Nipah virus glycoprotein (red chain), visualized in PyMOL.

4.4 Conclusion

This study successfully predicted the 3D structure of NiV-G and designed a complementary inhibitory protein targeted at its active sites. To understand the flexibility and properties of amino acid residues, B-factor heatmaps, contact maps, hydrophobicity scores, and SASA values were analyzed. Receptor-binding regions were specifically targeted, providing insights for inhibitor development. The B-factor heatmap revealed flexible and rigid regions, the contact map identified key residue interactions, hydrophobicity scores helped assess binding potential, and SASA values highlighted solvent-exposed residues critical for interaction. Docking and binding affinity predictions showed strong interactions, indicating therapeutic potential. Further optimization and validation are required to enhance stability and efficacy, reinforcing the role of computational protein design in developing antiviral treatments for the Nipah virus.

Chapter 5

Conclusions & Future Scope

5.1 Conclusion

In the current project, several computational algorithms have been applied to elucidate the 3D structure of the Nipah virus glycoprotein (NiV-G) and design a complementary inhibitory protein that targets the active sites of the glycoprotein. The resulting NiV-G structure prediction provided rich information for understanding the topology of NiV-G, potential receptor-binding sites, and important residues needed for viral entry. A protein designed using ProteinMPNN was found to interact within these active sites and as such could inhibit the ability for the virus to engage its host cell via the glycoprotein. Structural verification and stability predictions were made via AlphaFold2. Binding affinity and docking simulations suggested there was a strong interaction between the designed protein and NiV-G as the binding energies and dissociation constants provided meaningful notions of therapeutic efficacy. Overall, this promising evidence suggests the designed protein has high potential to be an antiviral therapeutic design with specificity amongst Nipah virus glycoproteins. While in silico evidence is promising, it is expected that optimized, experimental, and adaptive reapplication will help increase the overall stability, binding capability, and activity of the designed protein. Overall, this work furthers the newly organized field of computational protein design, creating a framework for the rational-based design of anti-viral therapeutics toward Nipah virus.

5.2 Future Scope

The study serves as the foundation for computationally engineered inhibitors of the glycoprotein of the Nipah virus (NiV-G). The binding energy and stability of the inhibitory protein can be tuned by further investigation involving molecular dynamics simulations, mutational studies, and machine learning optimization. Experimental validation in the

form of in vitro and in vivo assays will be pivotal in establishing its antiviral properties. Further, incorporation of AI-facilitated protein design tools will accelerate the optimization of best-inhibitors against upcoming viral variants. This would also be applicable to other paramyxoviruses or their associated diseases, except for Nipah virus, and play a role in broad-spectrum antiviral drug development. Computational modeling supported by experimental validation will be vital to the innovation of next-generation therapeutics against infectious pathogens.

References

[1] Y. Hong, J. Lee, and J. Ko, "A-prot: protein structure modeling using msa transformer," *BMC bioinformatics*, vol. 23, no. 1, p. 93, 2022.

[2] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. E. Ferrin *et al.*, "Ucsf chimerax: Structure visualization for researchers, educators, and developers," *Protein Science*, vol. 29, pp. 1–13, 2020.

[3] W. Tian, C. Chen, X. Lei, J. Zhao, and J. Liang, "Castp 3.0: Computed atlas of surface topography of proteins," *Nucleic Acids Research*, vol. 46, no. W1, pp. W363–W367, 2018.

[4] M. AlQuraishi, "End-to-end differentiable learning of protein structure," *Cell Systems*, vol. 8, no. 4, pp. 292–301.e3, 2019, epub 2019 Apr 17.

[5] O. Carugo, "How large b-factors can be in protein crystal structures," *BMC Bioinformatics*, vol. 19, p. 61, 2018.

[6] P.-H. Wang, Y.-H. Zhu, X. Yang, and D.-J. Yu, "Gcmapcrys: Integrating graph attention network with predicted contact map for multi-stage protein crystallization propensity prediction," *Analytical Biochemistry*, vol. 663, p. 115020, 2023.

[7] N. K. Singh, P. Bhardwaj, and M. Radhakrishna, "Hydrophobicitya single parameter for the accurate prediction of disordered regions in proteins," *Journal of Chemical Information and Modeling*, vol. 63, no. 16, pp. 5375–5383, 2023.

[8] X. Cao, M. H. Hummel, Y. Wang, C. Simmerling, and E. A. Coutsias, "Exact analytical algorithm for solvent accessible surface area and derivatives in implicit solvent molecular simulations on gpus," *Journal of Chemical Theory and Computation*, vol. 20, no. 11, pp. 4456–4468, 2024.

List of Publications

1. "Combining AlphaFold2 and ProteinMPNN for Efficient Complementary Protein Design Against Nipah Virus " submitted to the International Conference on Computing Technologies and Data Communication (ICCTDC), 2025.

Appendix A: Presentation

Protein Structure Prediction
Final Presentation

Guided By,
Ms.Sherine Sebastian
Asst. Professor

Aparna A R (U2103044)
Aparna Sajeev (U2103045)
Ashley K Alex (U2103052)
Athira J (U2103054)

Table of contents

Problem Definition

Accurately predicting protein structures for understanding biological processes and developing treatments, mandating efficient computational approaches, especially for emerging threats like the Nipah virus.

Purpose and Need

- To predict protein structures related to the Nipah virus and design complementary proteins that could aid in drug discovery.
- Meet the urgent need for antiviral treatments by using computational methods to simulate and analyze potential inhibitors that effectively neutralize viral proteins.

Objectives

- Using tools like AlphaFold2 to predict the 3D structure of proteins from their amino acid sequences.
- Utilizing software to visualize protein structures in 3D, allowing you to observe key features like binding residues.
- Engineering a protein that fits into the binding site through computational methods like docking and protein-protein interaction models.
- Ensuring that the designed protein is stable and functional under physiological conditions.

Literature Survey

Paper	Advantages	Disadvantages
1.Jumper,J.[1]Highly accurate protein structure prediction with AlphaFold. (2021).	• High Accuracy • Wide Applicability • Open-Source and Freely Available • Incorporation of Evolutionary Information	• Static Prediction • Cannot Predict Protein-Ligand Interactions • Limited Ability to Predict Protein Complexes
2. AlQuraishi M.[2] End-to-End Differentiable Learning of Protein Structure. 2019	• Fully differentiable framework • End-to-end approach • Scalability	• Complexity • Generalization

Literature Survey

Paper	Advantages	Disadvantages
3.Hong Y [3] S-Pred: protein structural property prediction using MSA transformer. (2022)	• Improved Accuracy with Transformers • Efficiency in Handling MSAs	• Computational Complexity • Dependence on MSA Quality • Limited Availability of MSAs
4.Dauparas et al. ,Robust deep learning–based protein sequence design using ProteinMPNN.(2022)	• High Precision • Scalability • Computational Efficiency • Deep Learning Utilization	• Dependency on structural data. • Limited experimental validation.

Literature Survey

Paper	Advantages	Disadvantages
5.Pettersen [5] UCSF ChimeraX: Structure visualization for researchers, educators, and developers.(2020)	• High-quality visualizations • Integration with external databases • Cross-platform compatibility	• Steep learning curve • Resource-heavy

Proposed Method

1.Protein Structure Prediction

Use advanced computational method **AlphaFold2**, which predicts protein structures with high accuracy based on amino acid sequences.

2.Molecular Visualization

Use software tool **PyMOL** to visually inspect and analyze the protein structure. These tools allow you to explore the binding residues, identify interaction points,generating the mask file and evaluate surface accessibility.

Proposed Method

3.Complementary Protein Design

Use **ProteinMPNN** to generate a complementary protein by designing a sequence or structure that forms stable, specific interactions with the binding site.

Proposed Method

4.Stability Analysis

Use tools **Prodigy,Cluspro** for binding affinity checking and **pLDDT score** for checking the accuracy of the predicted protein structure.

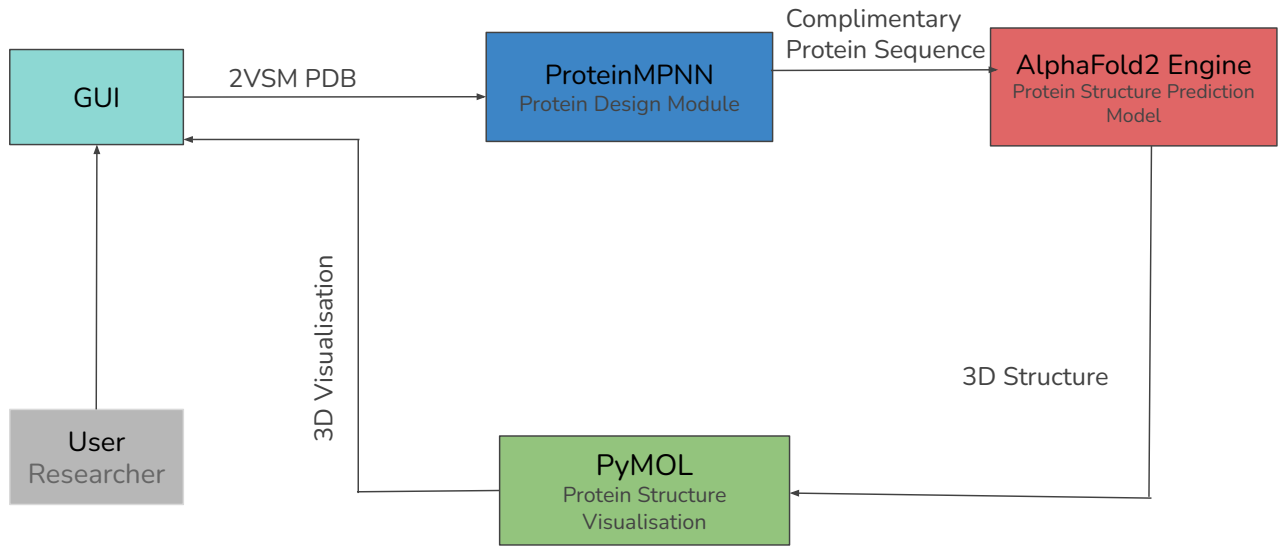
This will help you evaluate how well the protein folds, its thermodynamic stability, and its ability to maintain its structure in biological systems.

Proposed Method

PRODIGY is a web server used to predict the binding affinity (ΔG in kcal/mol) and dissociation constant (Kd) of protein-protein and protein-ligand complexes. It helps assess the stability and strength of molecular interactions.

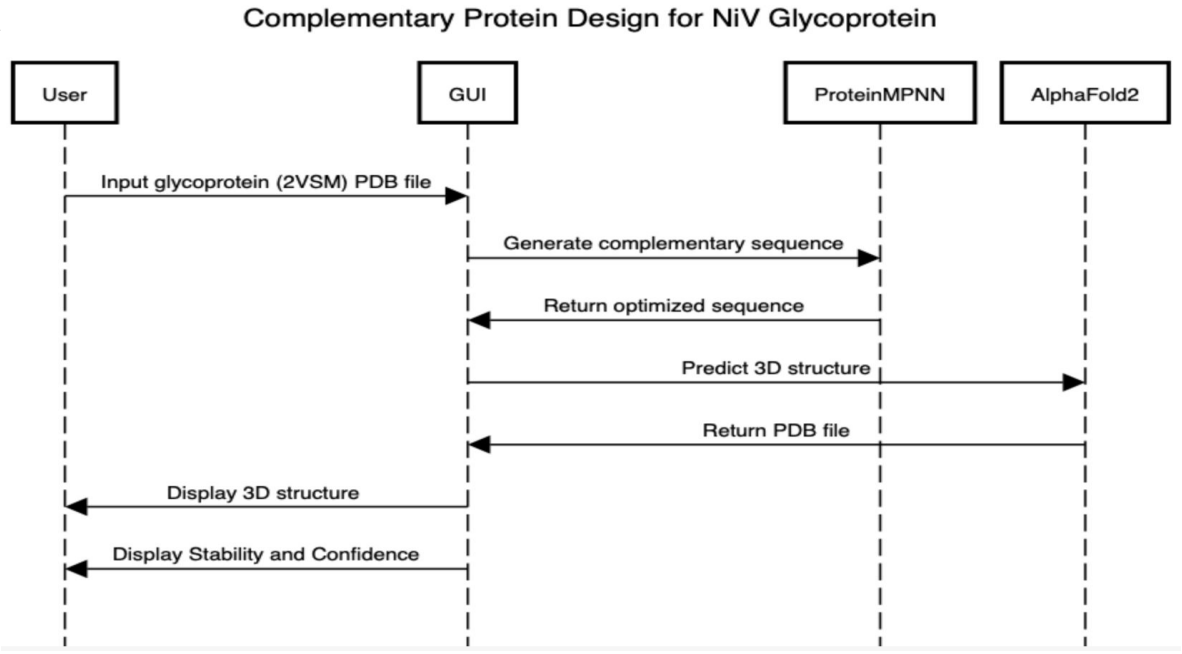
ClusPro is an automated protein-protein docking server used to predict the binding orientation and interactions between two proteins. It follows a three-step docking pipeline.

Architecture Diagram



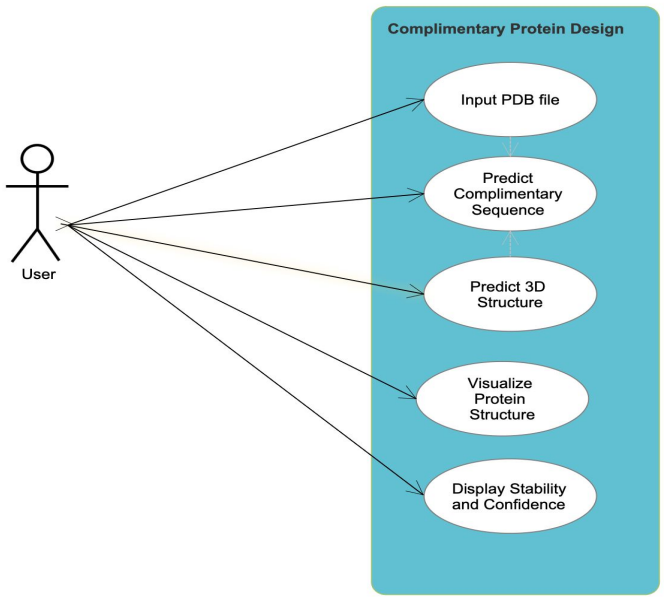
Architecture Diagram for Protein Structure Prediction System

Sequence Diagram



Sequence Diagram for Protein Structure Prediction

Use Case Diagram



Use Case Diagram for Protein Structure Prediction

Methods and Workflows

Alphafold2

- **Input:**The user provides an amino acid sequence of the complementary protein.
  - AlphaFold2 converts the sequence into a format suitable for multiple sequence alignment (MSA) searches.

- **Multiple Sequence Alignment (MSA) & Evolutionary Data Retrieval:** AlphaFold2 queries sequence databases to find homologous sequences.
  - It generates an **MSA**, which captures evolutionary relationships between similar proteins.

Methods and Workflows

Feature Extraction & Input Encoding: The model extracts:

- **MSA features** (co-evolutionary information).
- **Pairwise residue interactions** (distances, angles).
- **Template structure features** .

Structure Prediction Using Deep Learning: AlphaFold2's neural network processes the extracted features through:

- **Evoformer Module**: Captures long-range dependencies between residues.
- **Structure Module**: Converts sequence relationships into a 3D structure.

Methods and Workflows

Generation of 3D Models: AlphaFold2 outputs five PDB files, each representing a possible conformation of the input protein.

- These models are ranked based on **confidence scores** (pLDDT, PAE).

Structure Visualization & Refinement: The user loads the PDB files into **PyMOL** for visualization.

- Structural properties like **binding sites, secondary structure, and stability** are analyzed.
- Further refinements can be done using **molecular dynamics simulations** if necessary.

Methods and Workflows

Protein Design Using ProteinMPNN

- Input: The pdb file(human receptor and glycoprotein) and mask file.
- Graph representation: ProteinMPNN converts the structure into a graph with nodes (amino acids) and edges (spatial relationships).
- Message passing: The model uses message passing between nodes to capture residue interactions. Learns the context of each amino acid position within the structure.

Methods and Workflows

- **Sequence Generation:** Uses a masked language model to generate sequences that are compatible with the given structure. The mask file specifies which residues are mutable.

- **Scoring and Optimization:** Evaluates generated sequences for stability and compatibility. Uses a scoring function to prioritize sequences that maintain structural integrity.

Methods and Workflows

Mask File Creation for ProteinMPNN

- **Purpose of Mask File**:
  - Defines specific positions in the protein sequence that can be mutated.
  - Used to control which amino acids in the sequence are fixed and which can be modified during protein design.
- **Steps to Create a Mask File**:
  - **Select Residues for Mutation**: Identify the positions where mutations are allowed or desired.
  - **Mark Fixed Residues**: Indicate residues that should remain unchanged (e.g., critical for stability or function).

Methods and Workflows

- **Create Mask File Format**:
  - Typically in JSON format.
  - Includes the PDB ID and chain identifier.
  - Specifies mutable residue positions as an array of numbers.

Represents the positions in chain B of the 2VSM protein that can be mutated.

Example Mask File (for 2VSM):

```
{
  "2vsm": {
    "B": [57, 58, 60, 96, 97, 98, 99, 101, 102, 103, 105, 106, 107, 108, 109, 111,
112, 113, 114, 115, 116, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 130]
  }
}
```

Methods and Workflows

Stability Analysis

Analyzing Protein Stability (Using PyMOL & Bio.PDB)

- **Load the PDB file** containing the protein structure.
- **Identify interacting residues**:
  - Extract **hydrogen bonds** (distance: 2.5–3.5 Å).
  - Detect **salt bridges** (distance: <4.0 Å, involving charged residues).
  - Find **hydrophobic contacts** (distance: <5.0 Å, involving hydrophobic residues).

Methods and Workflows

Stability Score Calculation

- Compute stability score using weighted contributions:
  - **Hydrogen Bonds × 1.5**
  - **Salt Bridges × 2.0**
  - **Hydrophobic Contacts × 1.0**
- Categorize stability into **Low, Moderate, High, or Very High**.

Selecting the Most Stable Protein

- Compare multiple protein sets.
- The **highest stability score** indicates the best-designed protein.

Methods and Workflows

ClusPro: Protein-Protein Docking

- **Purpose**: Used for docking simulations to predict the binding affinity between the designed complementary protein and the target protein .
- **How It Works**:
  - The proteins are modeled as flexible structures, and ClusPro performs rigid-body docking followed by refinement.
  - Multiple docking solutions are generated, and the best-fitting poses are selected based on energy minimization.

Methods and Workflows

- **Role in the Project**:
  - Used to simulate the interaction between the designed complementary protein and the glycoprotein to evaluate how well they bind.
  - Helps identify potential binding interfaces and docking poses that suggest effective binding.
- **Results**:
  - **Binding Score**: A low energy score indicates a strong, favorable interaction.
  - **Docking Pose**: The resulting pose(s) show how the proteins fit together and the binding site regions.

Methods and Workflows

PRODIGY: Binding Affinity and Stability Evaluation

- **Purpose**: Used to calculate the binding affinity and estimate the stability of protein-protein complexes based on docking results.

- **How It Works**:
  - PRODIGY uses the protein complex structure to calculate free energy of binding (ΔG) based on physical and statistical models.
  - The binding affinity is quantified by calculating the **ΔG** (negative values indicate stronger binding).

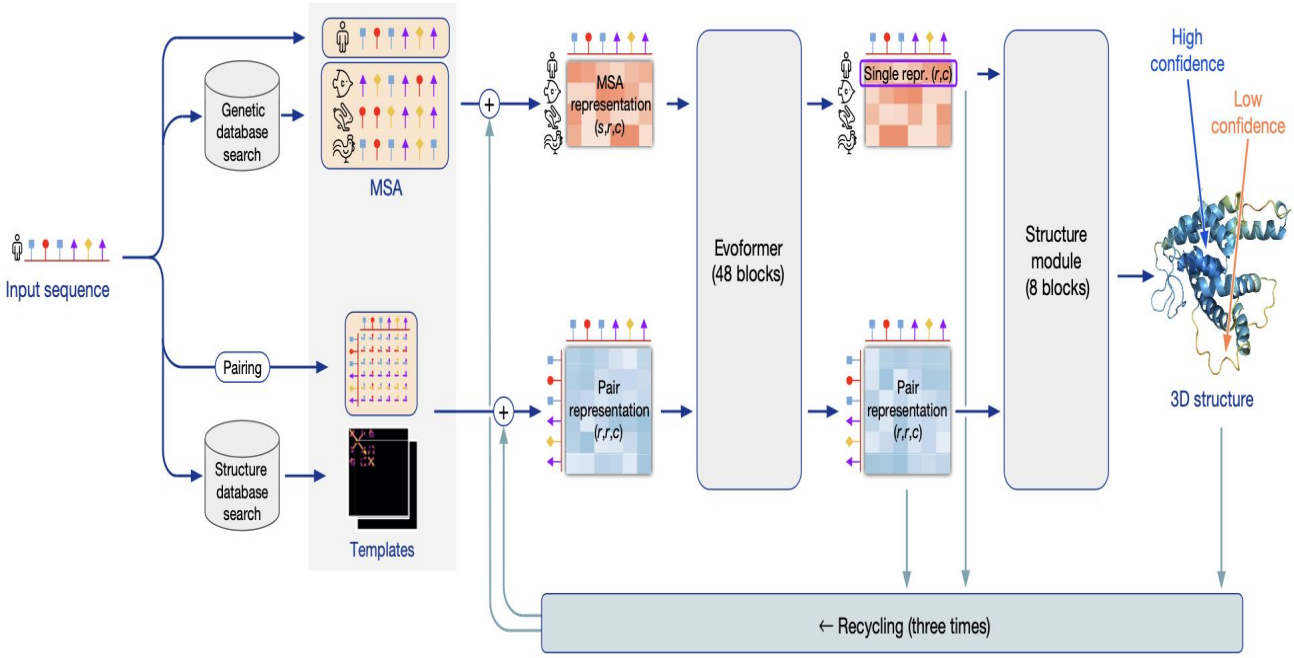
Methods and Workflows

- **Role in the Project**:
  - After obtaining docking results from ClusPro, PRODIGY evaluates the stability of the complex formed by the designed protein and the glycoprotein.
  - Helps to predict how stable the interaction is, which is crucial for designing proteins with therapeutic potential.
- **Results** :
  - **ΔG (Binding Free Energy)**: A more negative value indicates a stronger and more stable binding interaction.
  - **Stability Prediction**: A high binding affinity coupled with a low ΔG value suggests that the interaction is likely to be stable and functional.

Modules

1. Structure Prediction Using AlphaFold2

- **Purpose**: Predicts the 3D structure for sequence of amino acid residues using AlphaFold2.
- **Input**: Protein sequence .
- **Process**: AlphaFold2 leverages deep learning and attention mechanisms to predict protein folding.
- **Output**: 3D structure of the target protein.
- Retrieve the predicted 3D structure in **PDB format**.



Architecture diagram of AlphaFold2

Modules

Multiple Sequence Alignments (MSA) Representation

$$X_{i,j,a} = \begin{cases} 1, & \text{if the } j\text{-th sequence at position } i \text{ has amino acid } a \\ 0, & \text{otherwise} \end{cases}$$

i=residue index in the target sequence
j=index of homologous sequence in MSA
a=amino acid

Modules

Pair Representation

$$P_{ij} = \sum_k M_{ik} M_{jk}^T$$

Pij =relationship between amino acid i and j
Mik =presence of amino acid i and kth sequence of MSA
Mjk=presence of amino acid j in same sequence

Modules

MSA Row-wise Self-Attention (MSA Self-Attention)

- Focuses on **sequence-sequence interactions** in the MSA.
- Helps the model learn evolutionary relationships between homologous sequences.
- Each sequence in the MSA communicates with all others.

Modules

MSA Self-Attention

$$Q = W_Q X, \quad K = W_K X, \quad V = W_V X$$

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

W=weight matrix
Q,K,V query,key and value matrix
dk=dimensionality of key vectors
A=attention weighted sum of values

Modules

Pairwise Self-Attention (Axial Attention)

- Focuses on **residue-residue interactions** along the protein sequence.
- Helps capture structural constraints, such as **contacts** and **bond formations**.

For an input of shape (N,L,d) (N, L, d) (N,L,d), the **column-wise attention** (axial attention) treats each residue position **independently**, attending to all sequences at that position.

Modules

Structure Module (Building the 3D Model)

New position=Rotation x Old position +Translation

Final Structure Prediction

$$\text{pLDDT} = f_{\text{confidence}}(X, P)$$

Confidence is between score 0-100

Modules

2. Complementary Protein Design Using ProteinMPNN

- **Purpose**: Designs potential complementary proteins to inhibit viral proteins.
- **Input**: Predicted 3D structure from AlphaFold2.
- **Process**: Uses ProteinMPNN algorithm to generate sequences of complementary proteins that could bind and inhibit the viral protein.
- **Output**: Designed protein sequences that may block the viral protein's function.

Modules

3. Molecular Visualization

- **Purpose**: Visualizes the predicted structures and protein-protein interactions.
- **Input**: 3D structures from AlphaFold2.
- **Process**: Uses tools like PyMOL to generate detailed molecular visualizations.
- **Output**: High-resolution visualizations of protein structures and interactions.

Modules

4. Molecular Interaction Analysis

- Use **pLDDT scores** to check the confidence of the predicted structure.
- pLDDT (predicted Local Distance Difference Test) evaluates how well a predicted residue's local structure matches a true structure.
- It is a **per-residue confidence score**, ranging from **0 to 100**.
- Higher scores = More reliable regions

Modules

- Perform **docking** to evaluate protein-protein interactions.
- Use protein to protein docking software **ClusPro** to predict how your designed protein will interact with the target protein.
- ClusPro requires **two protein structures** as input:

1. **Receptor Protein** (Larger, typically stationary protein)
2. **Ligand Protein** (Smaller, mobile protein that binds to the receptor)

Modules

ClusPro Output Files

- **Top 10 cluster representatives** of docked complexes in **PDB format**.

- These structures show different possible binding conformations between the **receptor** and **ligand** proteins.
- Each docking model is **ranked based on cluster size** and **binding energy**.

Modules

- Analyze binding affinities of the complementary protein using **Prodigy**.
- Required Input Files-**PDB File** (Protein complex structure)
- This must contain **both binding partners** in a **docked conformation**.
- The output contains Predicted Binding Affinity (ΔG), Dissociation Constant (Kd), Interface Residues & Contacts.

Assumptions

Accurate Input Data:

- The amino acid sequences used are reliable and accurately represent the target proteins. This assumption is crucial as the quality of input data directly affects the reliability of the predicted structures.

Complementary Protein Design Feasibility:

- It is assumed that the binding sites identified can accommodate complementary protein designs that are structurally feasible and biochemically relevant.

Software and Hardware Requirements

Software:

- **ColabFold:** A faster and more accessible implementation of AlphaFold2 used for protein structure prediction, leveraging deep learning models.
- **ProteinMPNN:** Used for designing complimentary protein sequence from 3D structure of proteins.
- **PyMOL**: Utilized for visualizing the predicted protein structures, providing high-resolution molecular graphics.
- **ClusPro:** A protein-protein docking server used to predict the binding orientation between the designed protein and the target glycoprotein.
- **PRODIGY:** A tool for calculating the binding affinity and interaction strength between the protein complex.

Software and Hardware Requirements

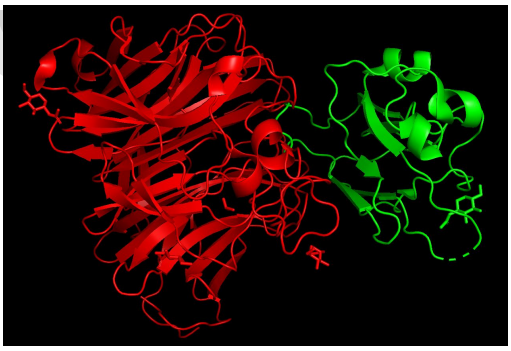
- **Python**: Programming language for implementing the algorithms and handling data processing.Version 3.8
- **Linux/Ubuntu or Windows**: Operating systems compatible with AlphaFold2 and related software tools.
- **Visual Studio Code**: For writing and running Python code, especially for analysis and visualization tasks.

Software and Hardware Requirements

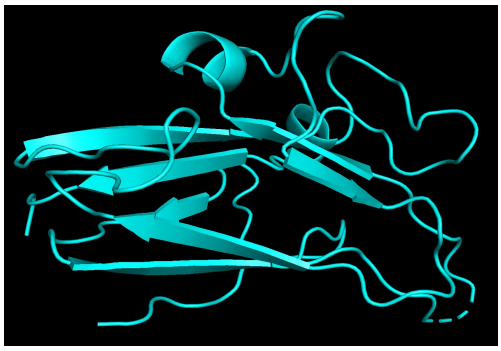
Hardware:

- **Windows PC** (Minimum Specifications)
- **Processor**: Intel Core i5 or AMD Ryzen 5 (or higher)
- **Memory**: 16 GB RAM (to handle large datasets and model computations)
- **Storage**: 512 GB SSD (for fast data access and software execution)
- **Graphics**: NVIDIA GTX 1060 or higher with at least 6 GB VRAM (for efficient processing in PyMOL and deep learning tasks)

Output

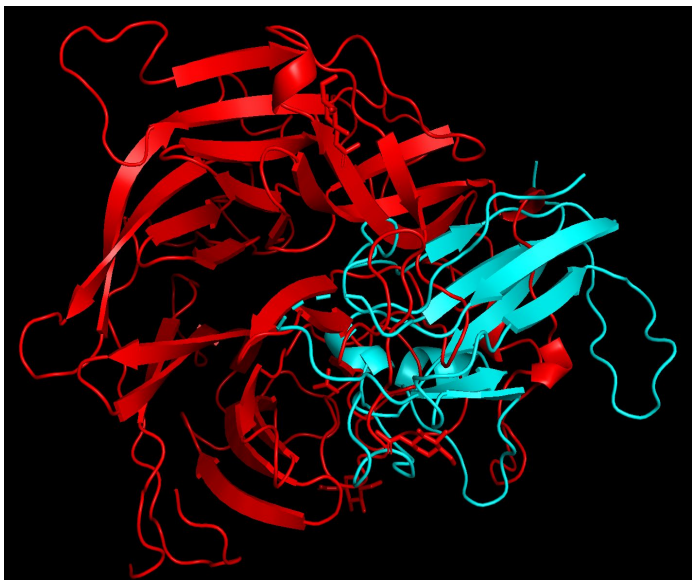


Input: Nipah virus attachment glycoprotein in complex with human cell surface receptor ephrinB2



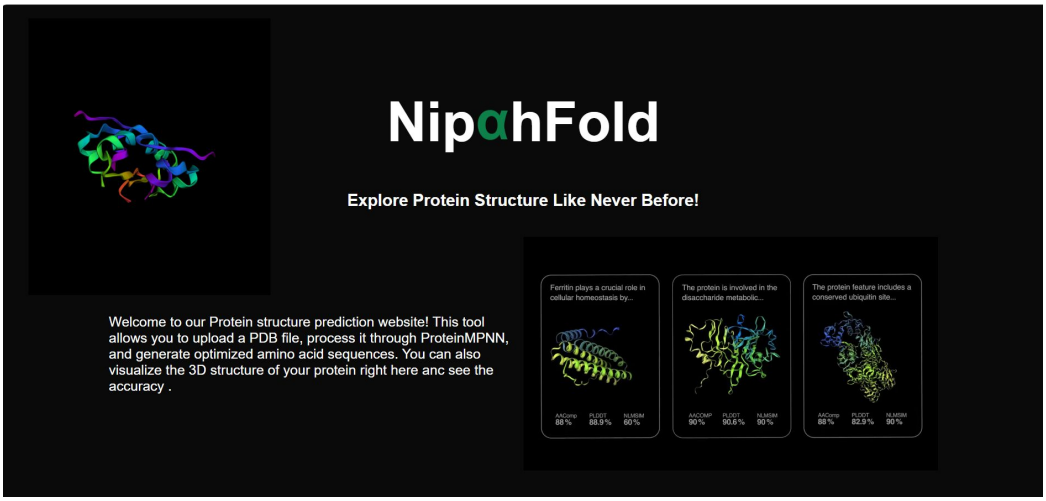
Output: 3D Structure of Predicted Protein

Output



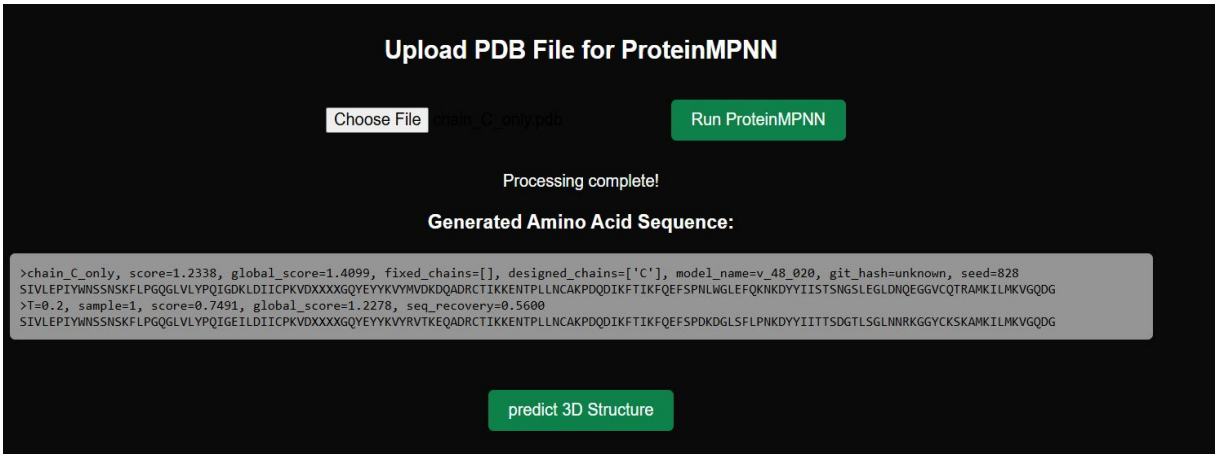
Predicted binding interaction between the designed protein and the Nipah virus glycoprotein (red chain), visualized in PyMOL.

Output



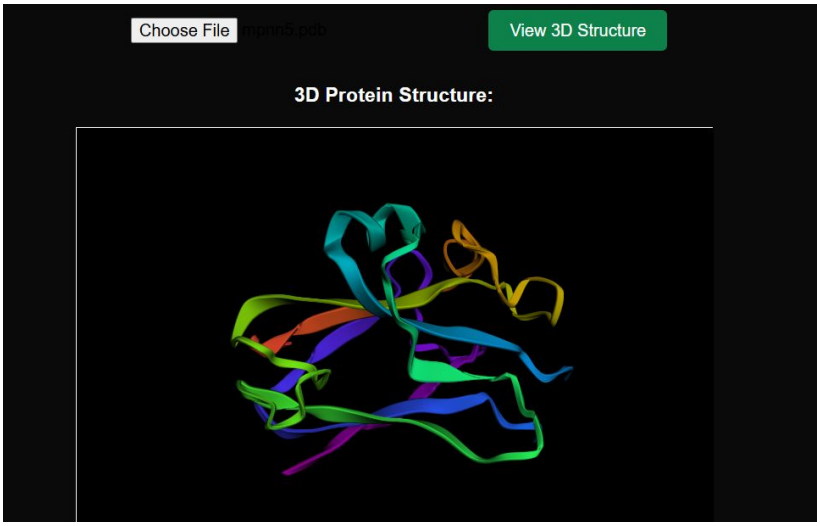
User Interface

Output



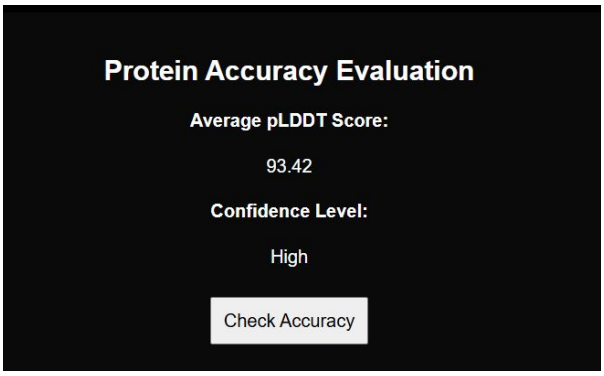
Sequence generated using ProteinMPNN

Output



3D structure of Protein

Output



Accuracy of Predicted Protein

Confidence Estimation Using PLDDT Score

pLDDT is a confidence score that estimates how well AlphaFold2 has predicted the position of individual residues in a protein structure. It is based on the uncertainty in predicted inter-residue distances rather than a direct comparison with an experimental structure

$$pLDDT_i = \frac{1}{M}\sum_{j=1}^{M}\sigma\left(1 - \frac{|\hat{d}_{ij} - d_{ij}|}{\delta}\right)$$

where:

- *d^ij* = Predicted distance between atoms
- *dij* = True distance between atoms
- M = Number of atom pairs considered.
- δ = Distance threshold (usually 15Å).

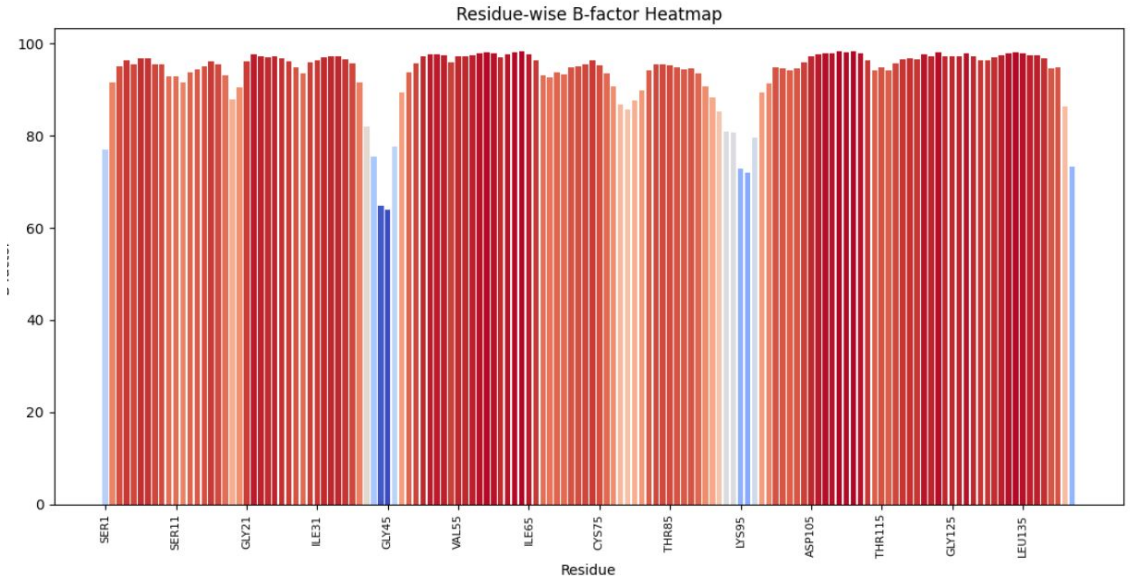
σ(x) = A step function that sets negative values to 0 and scales positive values between 0-1.

PRODIGY Binding Results

BINDING AFFINITY AND K_D PREDICTION

Protein-protein complex	ΔG (kcal mol^{-1})	K_d (M) at °C	ICs charged-charged	ICs charged-polar	ICs charged-apolar	ICs polar-polar	ICs polar-apolar	ICs apolar-apolar	NIS charged	NIS apolar
model_000_16	-10.1	3.8e-08	5	4	16	1	9	9	26.03	32.92

Residue Flexibility

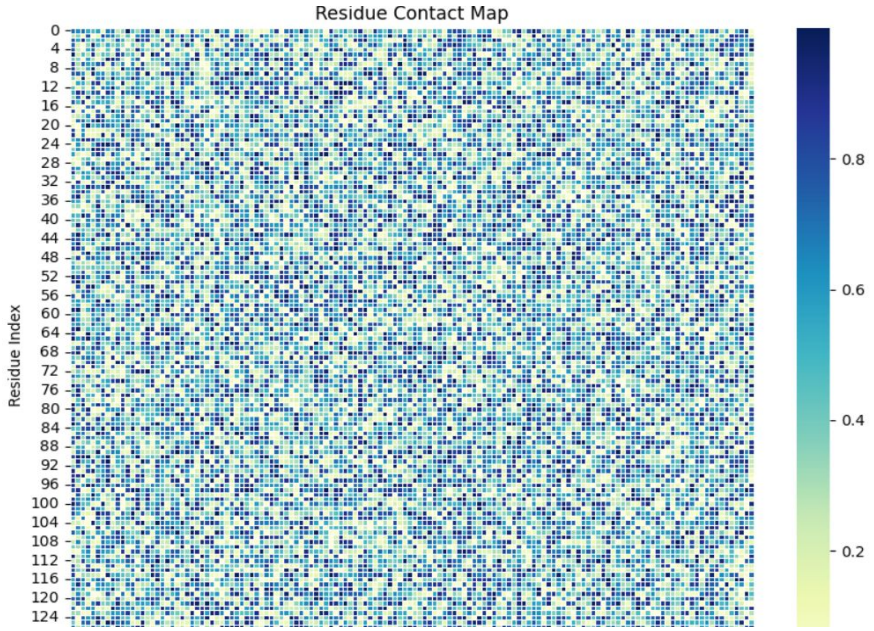


Residue Flexibility

B-factor heatmap visually represents the **flexibility or disorder** of each residue in your protein structure.

- High B-factor (Red Regions - Flexible)-High flexibility can indicate **binding sites** or **regions prone to conformational changes**.
- **Low B-factor (Blue Regions - Rigid)-**Low flexibility suggests **strong intra-protein interactions** or structural importance.

Residue Contact Map

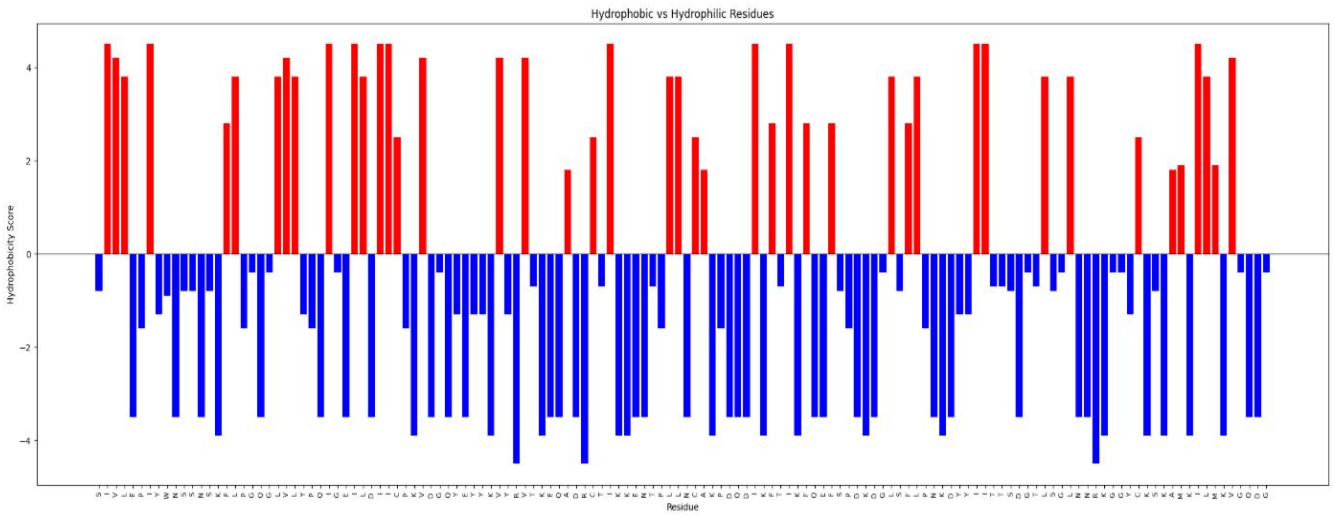


Residue Contact Map

A **contact map** is a 2D representation of the **spatial proximity between residues** in a protein. It helps in understanding **protein folding, interactions, and stability**.

- Dark/Intense Spots represents Close Residues
- Lighter or No Contact represents Distant Residues

Hydrophobicity Score

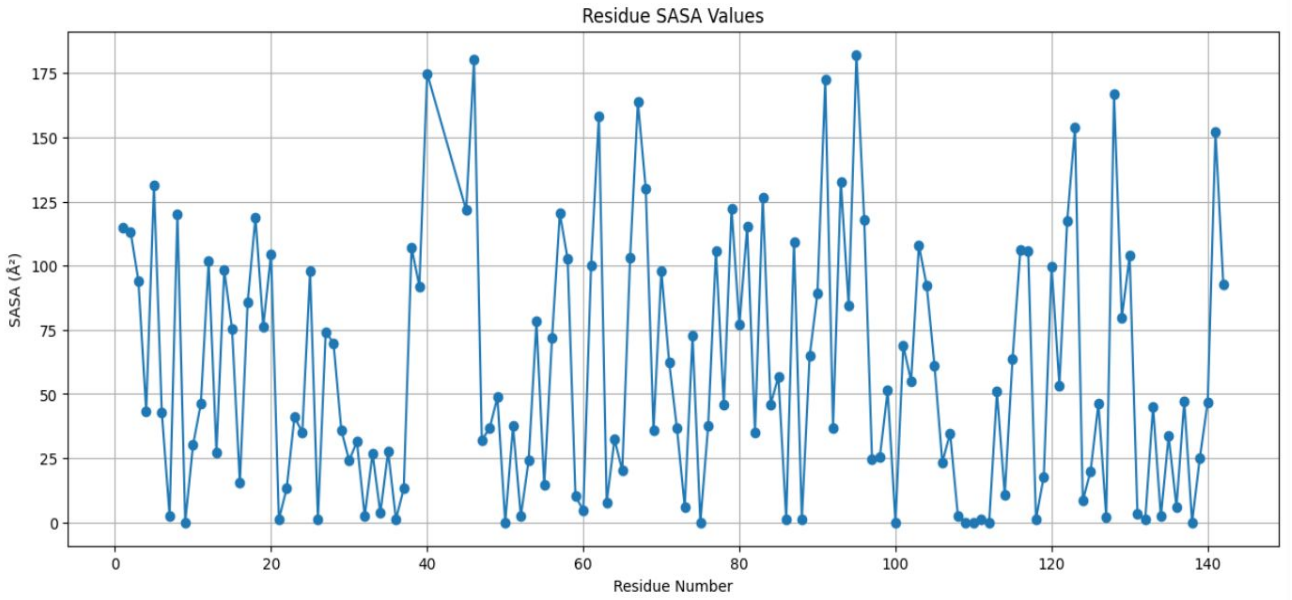


Hydrophobicity Score

Hydrophobicity scores are numerical values used to indicate the tendency of an amino acid residue to be **hydrophobic or hydrophilic**.

- higher positive values are considered hydrophobic
- negative values are considered hydrophilic.

SASA (Solvent Accessible Surface Area)

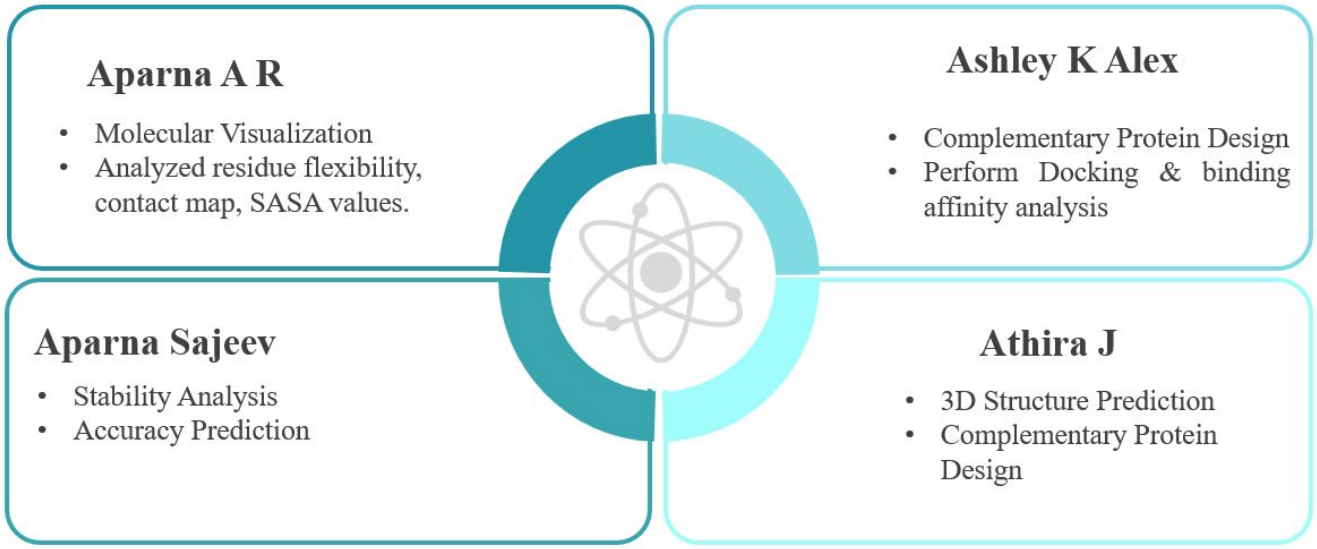


SASA (Solvent Accessible Surface Area)

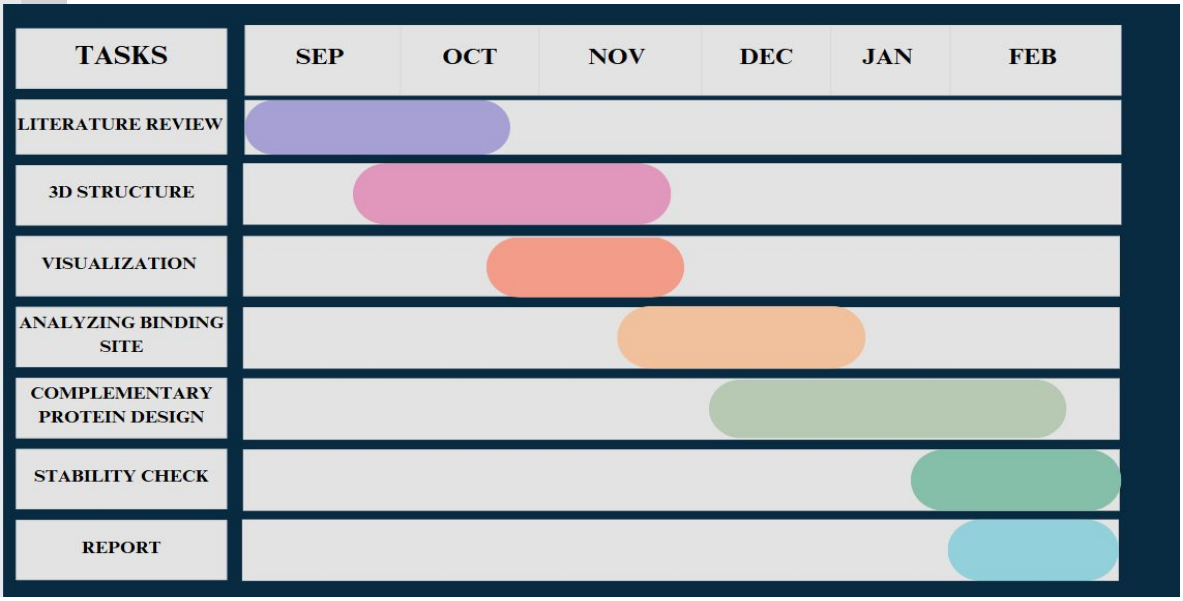
SASA (Solvent Accessible Surface Area) refers to the area of a protein that is accessible to solvent molecules, typically water.

- **High SASA value**: If a region of a protein has a high SASA value, it indicates that the region is exposed to the solvent and is likely on the protein's surface.
- **Low SASA value**: If a region has a low SASA value, it means that the region is buried inside the protein's core and is shielded from the solvent.

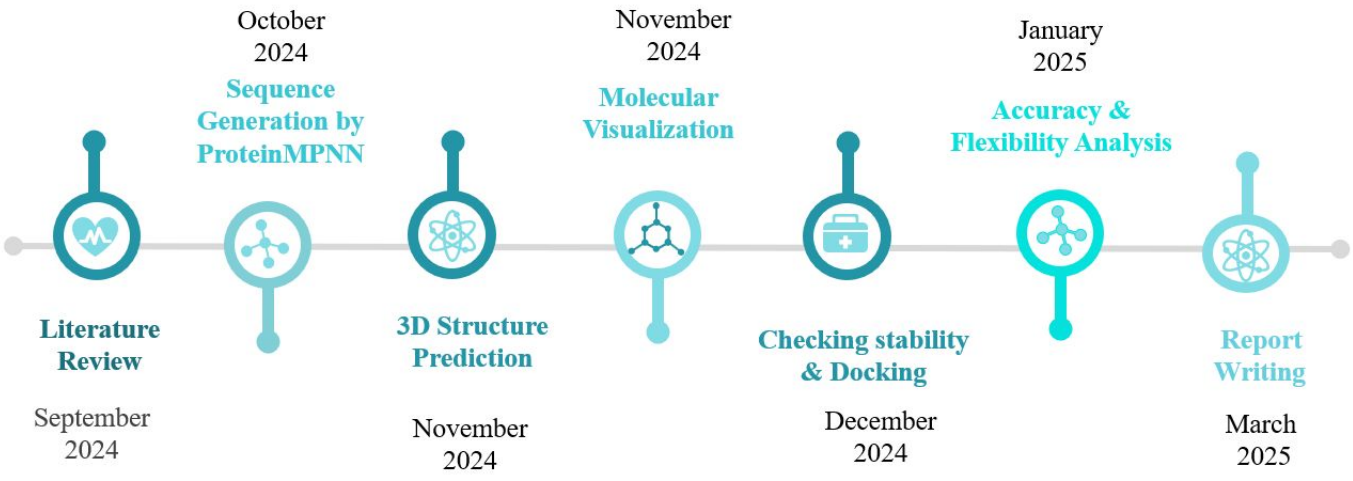
Work breakdown & Responsibilities



Gantt Chart



Timeline



Budget

Sl No.	Items	Amount
1	Google Colab Pro	2000
2	Publication charges	2000

Risk and Challenges

Stability and Docking Predictions

Predicting the stability of designed proteins and their interactions with targets can be complex and often requires experimental validation.

Ethical Considerations

Ethical considerations must be taken into account, especially when designing proteins for therapeutic use.Failing to address ethical concerns can lead to societal implications and impact the acceptance of your findings.

Expected Outcome

Protein Structure Prediction from amino acid that incorporates the following features:

1. Graphical User Interface
2. Complementary Protein Structure
3. 3D protein Structure

Conclusion

- AlphaFold2 delivers highly accurate 3D protein structure predictions using advanced deep learning techniques.
- The project designs a complementary protein that binds to the target protein's active site with high predicted affinity and specificity.
- This work demonstrates the effectiveness of structure-based design in developing protein inhibitors and lays the groundwork for future research in drug discovery.

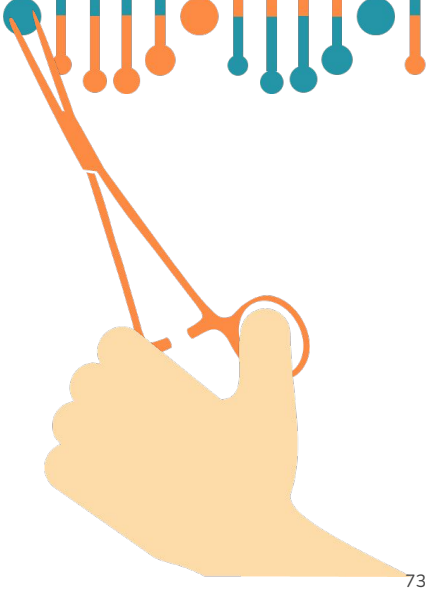
References

1. Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
2. AlQuraishi M. End-to-End Differentiable Learning of Protein Structure. Cell Syst. 2019 Apr 24;8(4):292-301.e3. doi: 10.1016/j.cels.2019.03.006. Epub 2019 Apr 17. PMID: 31005579; PMCID: PMC6513320.
3. Hong Y, Song J, Ko J, Lee J, Shin WH. S-Pred: protein structural property prediction using MSA transformer. Sci Rep. 2022 Aug 16;12(1):13891. doi: 10.1038/s41598-022-18205-9. PMID: 35974061; PMCID: PMC9381718.

References

4. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, Leung PJY, Huddy TF, Pellock S, Tischer D, Chan F, Koepnick B, Nguyen H, Kang A, Sankaran B, Bera AK, King NP, Baker D. Robust deep learning-based protein sequence design using ProteinMPNN. Science. 2022 Oct 7;378(6615):49-56. doi: 10.1126/science.add2187. Epub 2022 Sep 15. PMID: 36108050; PMCID: PMC9997061.
5. Pettersen, Eric & Goddard, Thomas & Huang, Conrad & Meng, Elaine & Couch, Greg & Croll, Tristan & Morris, John & Ferrin, Thomas. (2020). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Science. 30. 10.1002/pro.3943.

Thank You



Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills
The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.
PSO2: Programming and Software Development Skills
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.
PSO3: Professional Skills
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes (CO)

Course Outcome 1: Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).

Course Outcome 2: Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).

Course Outcome 3: Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).

Course Outcome 4: Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).

Course Outcome 5: Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).

Course Outcome 6: Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

Appendix C:
CO-PO-PSO Mapping

Course Outcomes

After completion of the course the student will be able to:

SL.NO	Description	Bloom's Taxonomy Level
CO1	Model and solve real-world problems by applying knowledge across domains.	Level 3: Apply
CO2	Develop products, processes, or technologies for sustainable and socially relevant applications.	Level 3: Apply
CO3	Function effectively as an individual and as a leader in diverse teams to comprehend and execute designated tasks.	Level 3: Apply
CO4	Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms.	Level 3: Apply
CO5	Identify technology/research gaps and propose innovative/creative solutions.	Level 4: Analyze
CO6	Organize and communicate technical and scientific findings effectively in written and oral forms.	Level 3: Apply

CO-PO Mapping

CO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12
1	2	2	1	1	-	2	1	-	-	-	-	3
2	3	3	2	3	-	2	1	-	-	-	-	3
3	3	2	-	-	3	-	-	1	-	2	-	3
4	3	-	-	-	2	-	-	1	-	3	-	3
5	3	3	3	3	2	2	-	2	-	3	-	3

CO-PSO Mapping

CO	PSO 1	PSO 2	PSO 3
1	3	1	2
2	3	2	2
3	2	2	-
4	3	-	3
5	3	-	-