JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

3D PROTEIN STRUCTURE PREDICTION SYSTEM

Ms. Baddala Vijayanirmala, Spoorthi S Padukone, Sinchana B L, Swetha V, Swathi V

Associate Professor, Student, Student, Student, Student
Department of Computer Science and Engineering,
AMC Engineering College, Bangalore, India

Abstract : The Project introduces an innovative method for three-dimensional protein prediction utilizing the Streamlit framework. By incorporating machine learning techniques and interactive visualization features, we have developed a user-friendly platform that enables accurate and efficient protein structure modeling. Our results showcase the effectiveness of this approach in enhancing accessibility and facilitating data-driven decision-making in structural biology and drug discovery. The incorporation of Streamlit offers a seamless and interactive experience, making it a valuable tool for researchers and practitioners in the field of computational biology. Our system delves into an innovative approach that harnesses the capabilities of the Streamlit framework to revolutionize protein structure prediction..

I.INTRODUCTION

In the realm of molecular biology and bioinformatics, the prediction of protein structures stands as a cornerstone endeavor. Understanding the three-dimensional arrangement of proteins elucidates their functions, interactions, and opens avenues for drug discovery and disease treatment. However, this task has long been daunting due to the complexity of protein folding. Enter Streamlit, a powerful platform facilitating intuitive and interactive web applications. Leveraging Streamlit's capabilities, we embark on a transformative journey to develop a 3D Protein Structure Prediction System. Our project report chronicles the fusion of cutting-edge bioinformatics with streamlined user experience, presenting a comprehensive overview of the system's architecture, methodologies employed, and its potential impact on scientific research and beyond. The validation and benchmarking of our system are critical aspects addressed in the report. We present comparative analyses against existing methods, highlighting the accuracy, speed, and reliability of our predictions.

Through rigorous testing and validation procedures, we instill confidence in the system's capabilities and foster trust within the scientific community.Moreover, we discuss the potential implications of our 3D Protein Structure Prediction System beyond the realm of academia. From drug discovery to personalized medicine, the ability to accurately predict protein structures holds immense promise for various industries. By fostering collaborations and partnerships, we envision our system catalyzing breakthroughs in biomedical research and therapeutic development.In conclusion, our project encapsulates the convergence of advanced bioinformatics, machine learning, and user-centric design principles in the development of a 3D Protein Structure Prediction System using Streamlit. By providing a comprehensive overview of the system's architecture, methodologies, validation, and broader implications, we not only showcase a technical achievement but also pave the way for transformative advancements in structural biology and beyond.Through intuitive user interface design and clear documentation, we strive to empower users of all backgrounds to explore the fascinating world of protein structures.

II. PROPOSED SOLUTION

The proposed System entails leveraging Streamlit's user-friendly interface and integrating advanced computational algorithms, including deep learning techniques, for accurate 3D Protein Structure Prediction. By incorporating robust validation procedures and optimizing computational resources, the system aims to overcome historical inaccuracies. Additionally, collaborative efforts to curate comprehensive experimental datasets and streamline access to validation methods such as X-ray crystallography and NMR spectroscopy will enhance prediction reliability. Continuous refinement and updates to the system, informed by user feedback and emerging research, will ensure its adaptability and effectiveness in addressing the challenges of predicting complex protein structures. Through these efforts, the proposed solution seeks to establish a reliable and accessible platform for protein structure prediction, advancing research in bioinformatics, drug discovery, and molecular biology. Continuous refinement and updates will be integral to the system's evolution, incorporating user feedback and advancements in computational biology. Through these efforts, the proposed solution seeks to establish a robust and accessible platform for protein structure prediction, empowering researchers across disciplines to make significant strides in understanding protein function and facilitating advancements in drug discovery.

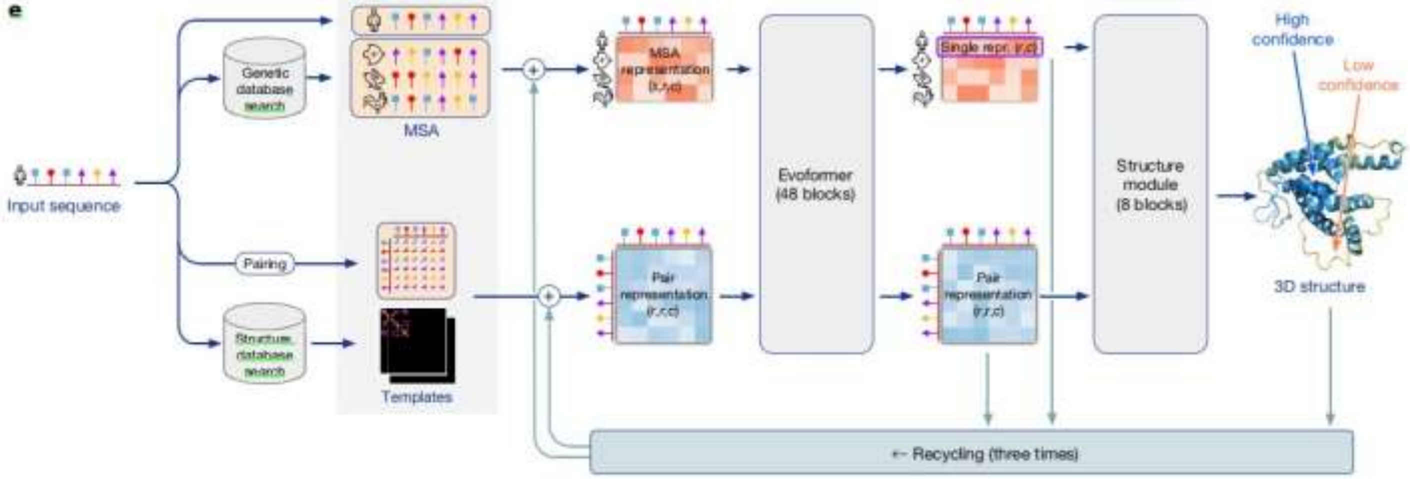

e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

fig 1. Architecture diagram of 3D protein structure prediction system

The fig 1. system begins with data collection from reputable databases like UniProt and PDB, ensuring cleanliness by removing anomalies and duplicates. Preprocessing involves feature extraction using algorithms like linear regression,catboost, generating sequence profiles, secondary structure predictions, and evolutionary conservation scores.Performance optimization focuses on scalability, speed, and accuracy, utilizing parallel computing and cloud resources. Validation against benchmark datasets and participation in community challenges ensure the system's robustness and benchmarking against state-of-the-art methods. Continuous refinement and validation mechanisms underpin the system's reliability and applicability across diverse biological contexts.

III. IMPLEMENTATION

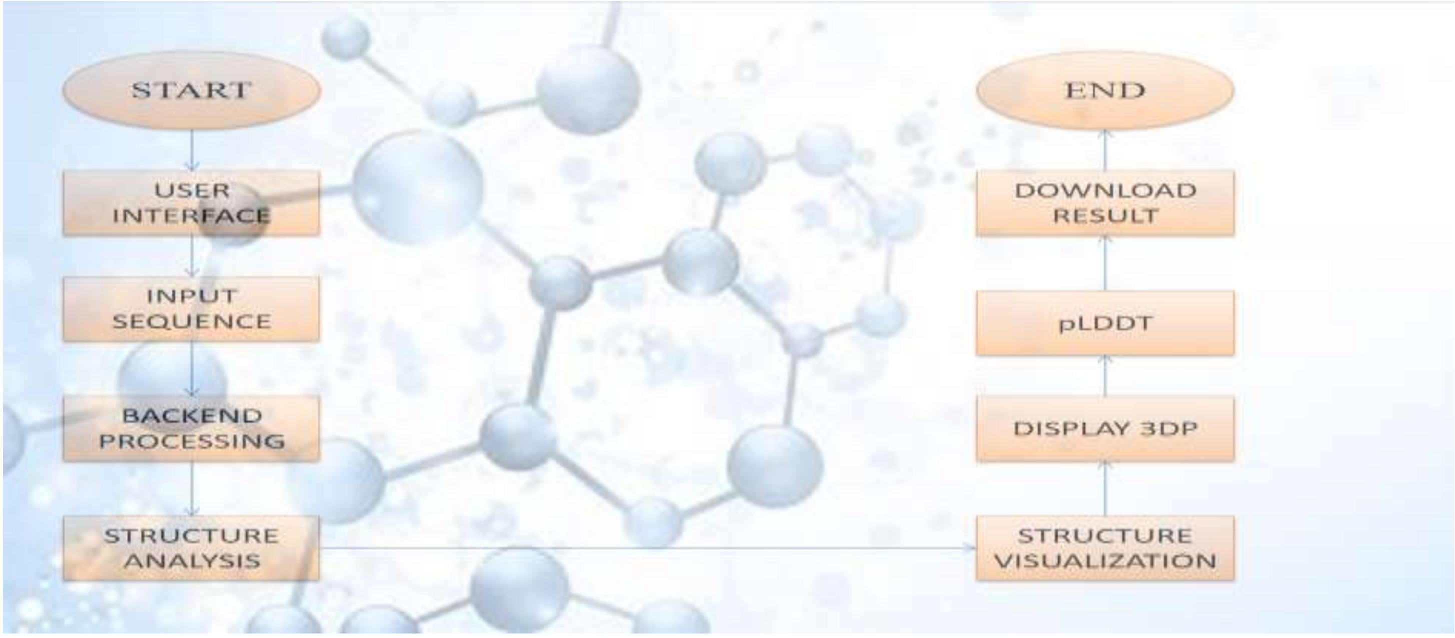


fig 2. Block diagram of 3D protein structure prediction system

The fig 2 represents the interface design for the 3D protein structure prediction system that enables the user to get the insights of how the prediction process undertakes and how to analyze the results.

Proteins, the workhorses of biological systems, perform a myriad of functions crucial for life, ranging from catalyzing biochemical reactions to providing structural support. The three-dimensional (3D) structure of a protein is intricately linked to its function, dictating how it interacts with other molecules in the cellular environment. Despite its paramount importance, experimentally determining protein structures remains challenging and resource-intensive. By harnessing the power of computational algorithms and leveraging available experimental data, such systems hold promise in revolutionizing drug discovery, enzyme design, and personalized medicine.Understanding 3D protein structure prediction is pivotal in elucidating the intricate relationship between a protein's structure and its function. Proteins, composed of amino acids, fold into complex three-dimensional structures that govern their biological activities. Predicting these structures computationally involves deciphering the folding patterns and spatial arrangements of atoms within the protein molecule.

Implementing a 3D protein structure prediction system involves a series of meticulous steps, encompassing data acquisition, preprocessing, modeling, validation, deployment, and maintenance. Here’s a detailed breakdown of the implementation process:

Data Acquisition: The implementation journey commences with acquiring relevant data, including amino acid sequences, structural templates, and experimental constraints. Data may be sourced from public repositories such as the Protein Data Bank (PDB), genome databases, or specialized protein structure prediction datasets. Additionally, experimental techniques such as X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy can provide valuable structural information for validation purposes.

Data Preprocessing: Raw data undergoes preprocessing to ensure consistency, quality, and compatibility with prediction algorithms. Preprocessing steps may include sequence alignment, feature extraction, removal of redundant or erroneous entries, and normalization of data formats. Quality control measures are implemented to filter out artifacts and ensure the integrity of input data.

Model Selection: Choosing suitable prediction models is critical for achieving accurate and reliable results. Depending on the nature of the input data and the desired level of accuracy, different modeling approaches may be employed, including. Homology Modeling Utilizes known protein structures as templates to predict the structure of a target protein with high sequence similarity. Initio Modeling Predicts protein structures solely from their amino acid sequences, often using physics-based energy functions and optimization algorithms. Threading (Fold Recognition)Aligns the target protein sequence onto a library of known protein folds to identify the most probable structural template. The selection of an appropriate modeling method is guided by factors such as data availability, computational resources, and the complexity of the protein structure being predicted.

Training and Optimization: For machine learning-based approaches, model training involves the iterative process of feeding labeled data into the model, adjusting model parameters, and evaluating performance. Techniques such as cross-validation, hyperparameter tuning, and ensemble methods are employed to optimize model performance and generalization capabilities. Training datasets may be augmented with additional experimental data or simulated structural ensembles to enhance model robustness.

Validation and Evaluation: Validation is essential for assessing the accuracy and reliability of predicted protein structures. Predicted models are compared against experimental data using validation metrics such as Root Mean Square Deviation (RMSD), Global Distance Test (GDT), or Ramachandran plot analysis. Additionally, cross-validation techniques and independent validation datasets are utilized to validate model generalization and predictive performance. Validation results inform iterative improvements to prediction algorithms and model architectures.

Maintenance and Updates: Establishing robust maintenance procedures is crucial for ensuring the long-term functionality and reliability of the prediction system. Regular updates, bug fixes, and performance optimizations are rolled out to address emerging challenges and incorporate advancements in prediction algorithms or experimental techniques. Continuous monitoring of system performance and user feedback enables proactive maintenance and enhancements to meet evolving user needs.

The following Algorithms which are implemented:

Linear Regression: Linear regression is a statistical method used in 3D protein structure prediction to model the relationship between input features (e.g., amino acid properties, structural characteristics) and the target variable. It is simpler compared to more complex algorithms, it can still provide valuable insights into protein structure predictions, especially when dealing with linearly related features.It finds the best-fitting linear equation between a dependent variable and one or more independent variables. It minimizes the difference between predicted and actual values, estimating the relationship's slope and intercept.

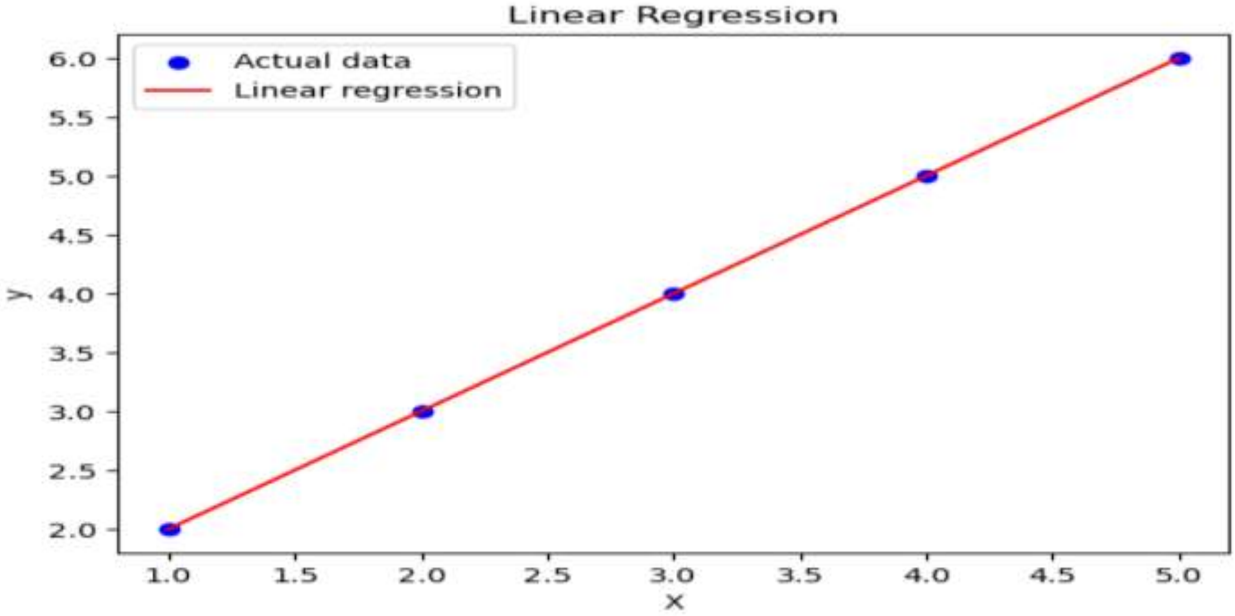


fig 3. The graph representing the training data in linear regression.

The above fig 3 representing the linear regression graph visually represents the relationship between two variables in a dataset. In this type of graph, the independent variable (often denoted as X) is plotted on the horizontal axis, while the dependent variable (often denoted as Y) is plotted on the vertical axis. The graph typically shows a straight line that best fits the data points, indicating the linear relationship between the variables. The slope of the line represents the rate of change in Y with respect to changes in X, while the intercept shows the value of Y when X is zero. The points on the graph represent the actual data points, and the line serves as a predictive model that estimates Y based on X.

RandomForest Regression: It is an ensemble learning algorithm used in 3D protein structure prediction systems to improve accuracy. It constructs multiple decision trees and combines their predictions, reducing overfitting and enhancing generalization.It is a versatile ensemble learning algorithm used for classification and regression tasks. It builds multiple decision trees during training and combines their predictions through voting (for classification) or averaging (for regression) to improve accuracy and reduce overfitting. It randomly selects subsets of features and data points for each tree, enhancing diversity and robustness.It randomness in feature selection and data sampling enhances diversity among trees, making it a popular and effective algorithm for various machine learning tasks. One key aspect of Random Forest is its use of random feature selection.

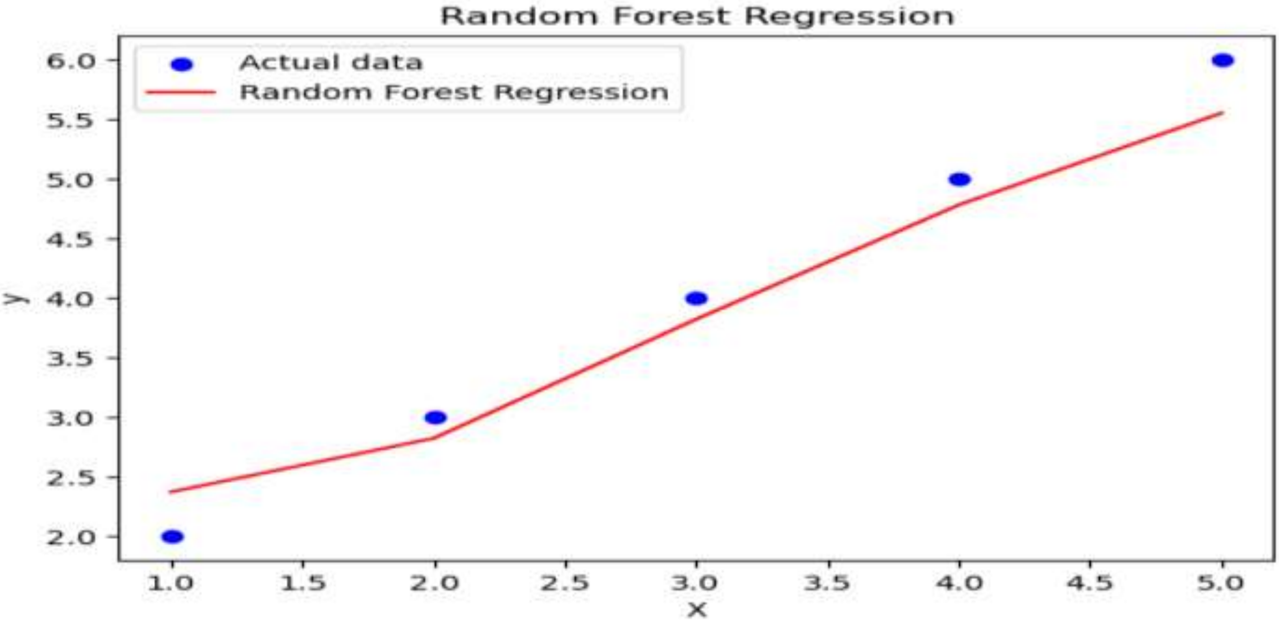


fig 4. The graph representing the training data in random forest regression

The above fig 4 representing the Random Forest Regression Graph, the horizontal axis represents the input data, while the vertical axis represents the predicted output values. The graph may show a collection of curves or a shaded area that indicates the range of predictions made by the random forest model. A random forest regression graph visualizes the predictive performance of a random forest regression model. Unlike a linear regression graph that shows a single line of best fit, a random forest regression graph is more complex and dynamic. It typically consists of multiple lines or curves that represent the predictions made by individual decision trees within the random forest ensemble. Each decision tree in the random forest makes its own prediction, and the final prediction is often an average or a weighted combination of these individual tree predictions.

CatBoost: CatBoost, a gradient boosting algorithm, is used in 3D protein prediction to handle complex data patterns efficiently. It helps optimize prediction accuracy by leveraging categorical features without preprocessing, handles missing data intelligently, and provides robustness against overfitting, contributing significantly to accurate protein structure modeling. It is a gradient boosting algorithm designed to handle categorical features efficiently in machine learning tasks. It iteratively builds an ensemble of decision trees, combining their predictions to create a strong model. Unlike other algorithms, CatBoost handles categorical variables automatically, eliminating the need for pre-processing.

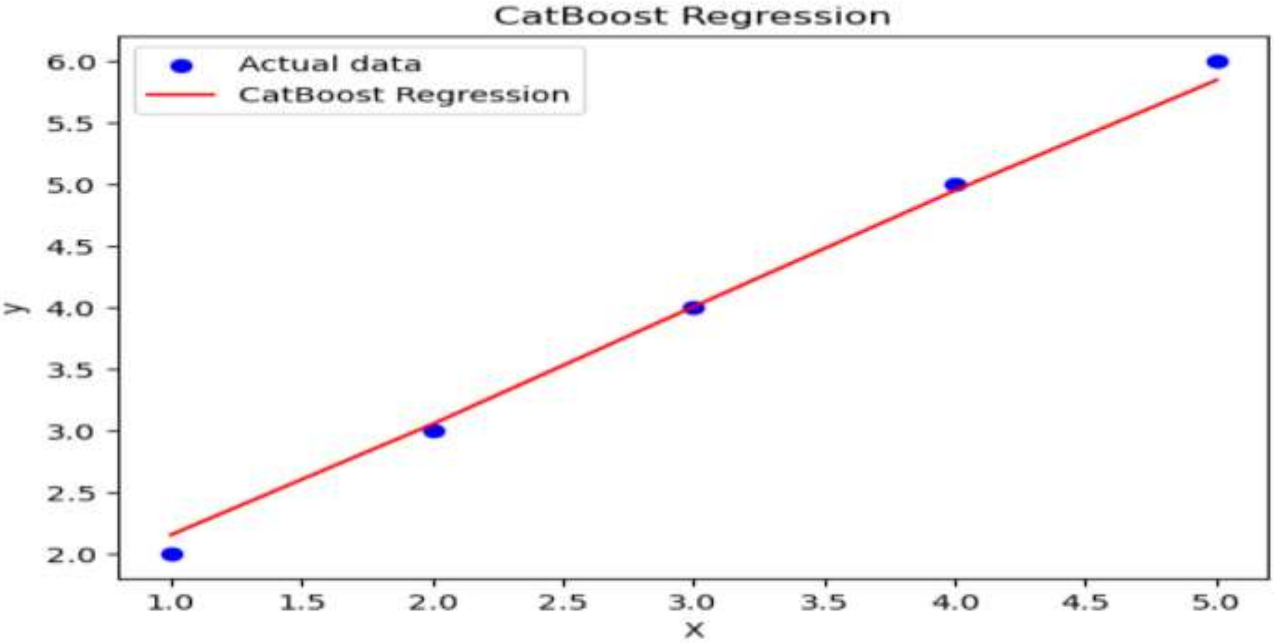


fig 5. The graph representing the training data in catboost.

The above fig 5. representing the CatBoost graph, the vertical axis usually displays the names of the input variables (features), while the horizontal axis represents the importance score of each feature. The importance score indicates how much each feature contributes to the model's predictive power. Higher scores suggest greater influence, while lower scores imply less impact. CatBoost, short for Categorical Boosting, is a machine learning algorithm specifically designed to handle categorical variables effectively. A CatBoost graph typically represents the feature importance or the impact of different variables on the model's predictions. Unlike linear regression or random forest graphs that show relationships or predictions, a CatBoost graph focuses on variable importance.

IV. RESULT AND DISCUSSION

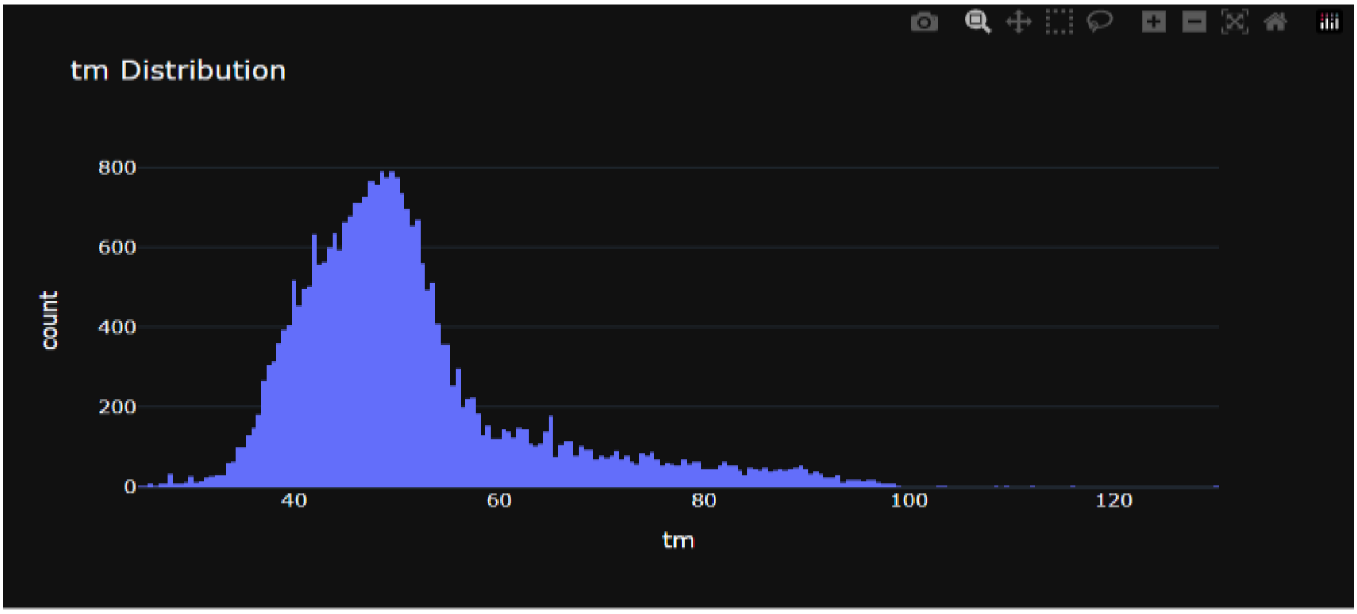


fig 6. The graph representing the distribution of TM domains within a protein sequence.

The above fig 6 representing the distribution of Transmembrane domains within a protein sequence, the horizontal axis represents the amino acid positions within the protein sequence, while the vertical axis indicates the likelihood or presence of transmembrane domains. The graph typically shows peaks or bars that indicate regions where transmembrane domains are predicted or detected. These peaks suggest the potential locations where the protein interacts with cellular membranes. The height or intensity of the peaks corresponds to the strength of prediction or the confidence in identifying transmembrane segments..Analyzing the distribution of transmembrane domains within a protein sequence graph is crucial for understanding the protein's structure and function. It helps identify regions responsible for membrane association, cell signaling, or transport activities. Furthermore, this information aids in studying protein topology, designing experiments to investigate protein-membrane interactions, and predicting the subcellular localization of proteins within cells.

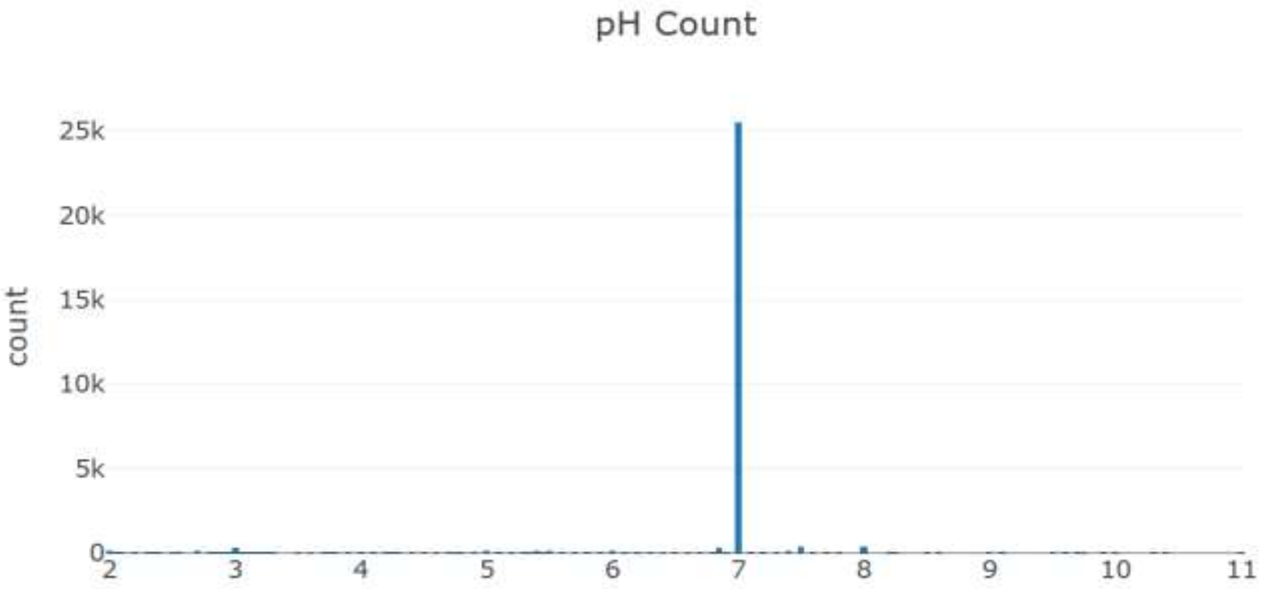


fig 7. The graph states relationship between the count of amino acids and ph levels.

The above fig 7 representing the Relationship between the count of specific amino acids and their ph levels. The horizontal axis of the graph would represent the count or frequency of a particular amino acid, while the vertical axis would denote the pH level. The graph might display data points or a line connecting the data points, showing how the pH level changes with variations in the count of specific amino acids. For example, it could show that as the count of acidic amino acids (like glutamic acid or aspartic acid) increases, the pH level tends to decrease, indicating a more acidic environment. Conversely, an increase in basic amino acids (such as lysine or arginine) might be associated with a higher pH level, indicating a more alkaline environment.This type of graph helps visualize the relationship between protein composition (in terms of amino acid types and their counts) and the biochemical properties of the environment, such as pH. It is particularly relevant in understanding protein structure-function relationships, enzyme activity under different pH conditions, and the stability of proteins in varying physiological environments. The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study,Data and Sources of Data, study’s variables and analytical framework.

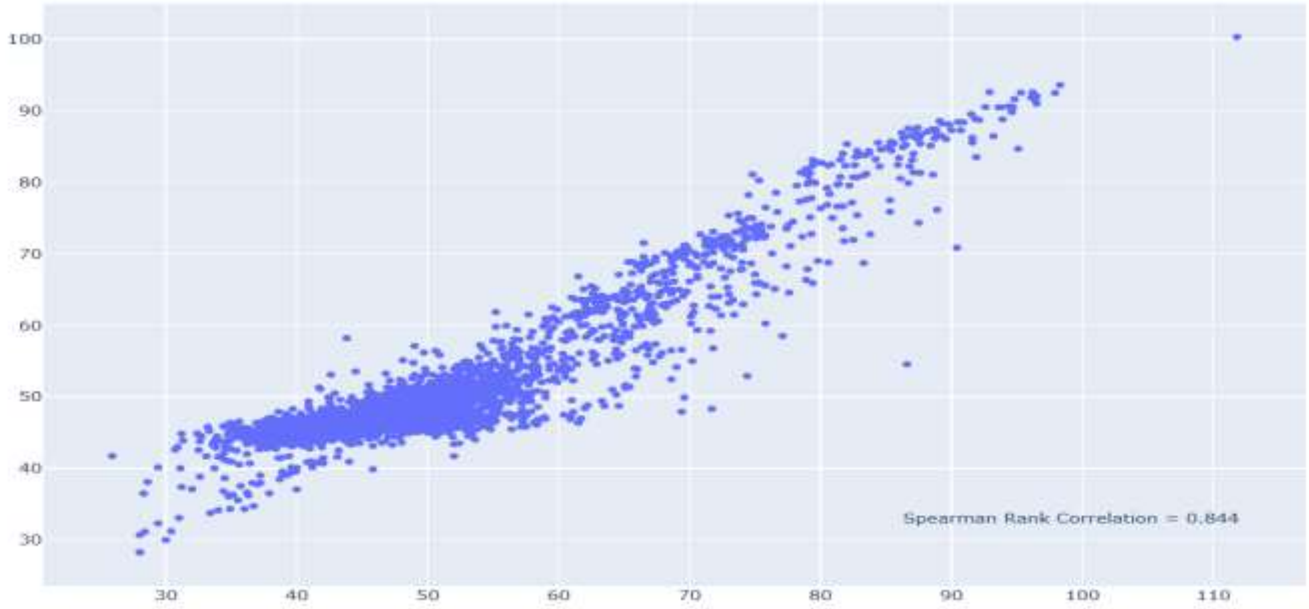
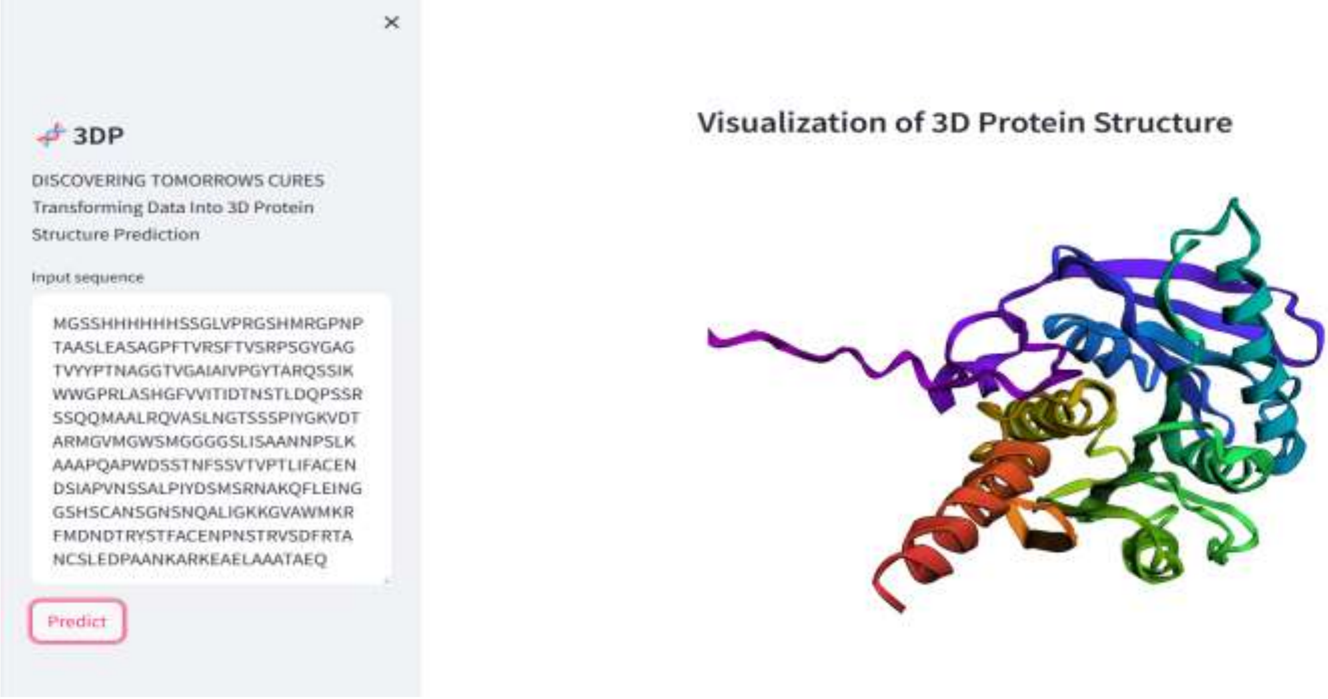


fig 8. The graph shows the relationship between two different sequences in protein data.

The above fig 8 representing the relationship between two different sequences in protein data, each sequence is typically depicted as a horizontal line, with the amino acid positions along the sequence shown on the horizontal axis. Matching or similar amino acids between the two sequences are connected by lines or shaded regions, highlighting regions of sequence similarity. Conversely, differences or gaps in the alignment are often indicated by spaces or mismatches in the lines. This type of graph helps researchers identify conserved regions, which are indicative of functional importance or evolutionary relatedness, as well as variable regions that may contribute to distinct functional characteristics. Another approach is a distance matrix graph, where pairwise similarity or dissimilarity scores between positions in the sequences are represented as a matrix. Darker colors in the matrix indicate higher similarity scores, while lighter colors signify lower similarity. These graphs are valuable for tasks such as understanding evolutionary relationships, predicting protein structure based on sequence conservation, and identifying functionally important regions within proteins.



DISCOVERING TOMORROWS CURES
Transforming Data Into 3D Protein
Structure Prediction

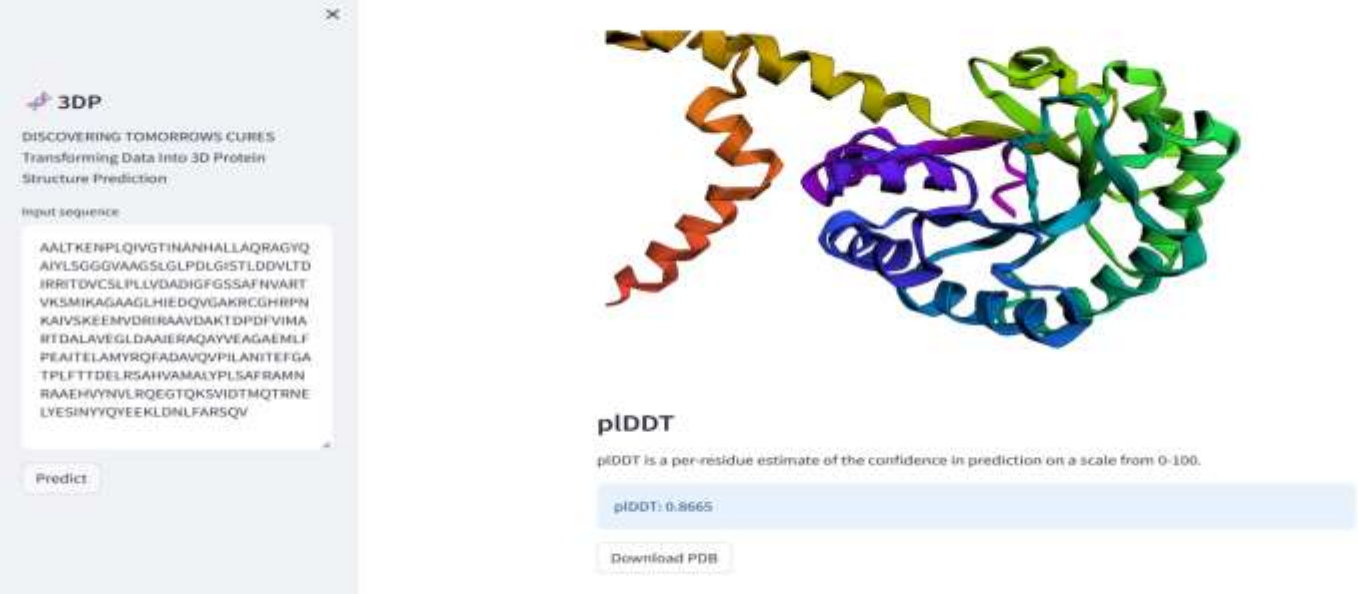
Input sequence

MGSSHHHHHHSSGLVPRGSHMRGPNP
TAASLEASAGPFTVRSFTVSRPSGYGAG
TVYYPTNAGGTVGAIAIVPGYTARQSSIK
WWGPRLASHGFVVITIDTNSTLDQPSSR
SSQQMAALRQVASLNGTSSSPIYGKVDT
ARMGVMGWSMGGGGSLISAANNPSLK
AAAPQAPWDSSTNFSSVTVPTLIFACEN
DSIAPVNSSALPIYDSMSRNAKQFLEING
GSHSCANSGNSNQALIGKKGVAWMKR
FMDNDTRYSTFACENPNSTRVSDFRTA
NCSLEDPAANKARKEAELAAATAEQ

Visualization of 3D Protein Structure

fig 9. The Image Represents the 3-Dimensional Protein Structure for the given Sequence.

The above fig 9 representing the a 3-dimensional protein structure for a given sequence provides a detailed visual representation of how the protein folds and arranges itself in three-dimensional space based on its amino acid sequence. In the image, you would see the protein's structures, such as alpha helices and beta sheets, as well as its tertiary structure, which shows how different parts of the protein fold and interact with each other. These structural elements play key roles in determining the protein's stability, enzymatic activity, and binding specificity. The 3D structure also highlights important features like active sites, where biochemical reactions occur, and binding sites, where the protein interacts with ligands or other molecules.Analyzing the 3D protein structure image provides insights into the protein's function and potential biological roles. It helps researchers understand how mutations or modifications in the sequence can impact the protein's structure and function, aiding in drug design, protein engineering, and unraveling molecular mechanisms underlying diseases. Overall, the image of the 3D protein structure is a powerful tool for studying proteins at the molecular level and advancing our knowledge of biochemistry and biology.



3DP

DISCOVERING TOMORROWS CURES
Transforming Data Into 3D Protein
Structure Prediction

Input sequence

```
AALTKENPLQIVGTINANHALLAQRAGYQ
AIYLSGGGVAAGSLGLPDLGISTLDDVLTD
IRRITDVCSLPLLVDADIGFGSSAFNVART
VKSMIKAGAAGLHIEDQVGAKRCGHRPN
KAIVSKEEMVDRIRAAVDAKTDPDFVIMA
RTDALAVEGLDAAIERAQAYVEAGAEMLF
PEAITELAMYRQFADAVQVPILANITEFGA
TPLFTTDELRSAHVAMALYPLSAFRAMN
RAAEHVYNVLRQEGTQKSVIDTMQTRNE
LYESINYYQYEEKLDNLFARSQV
```

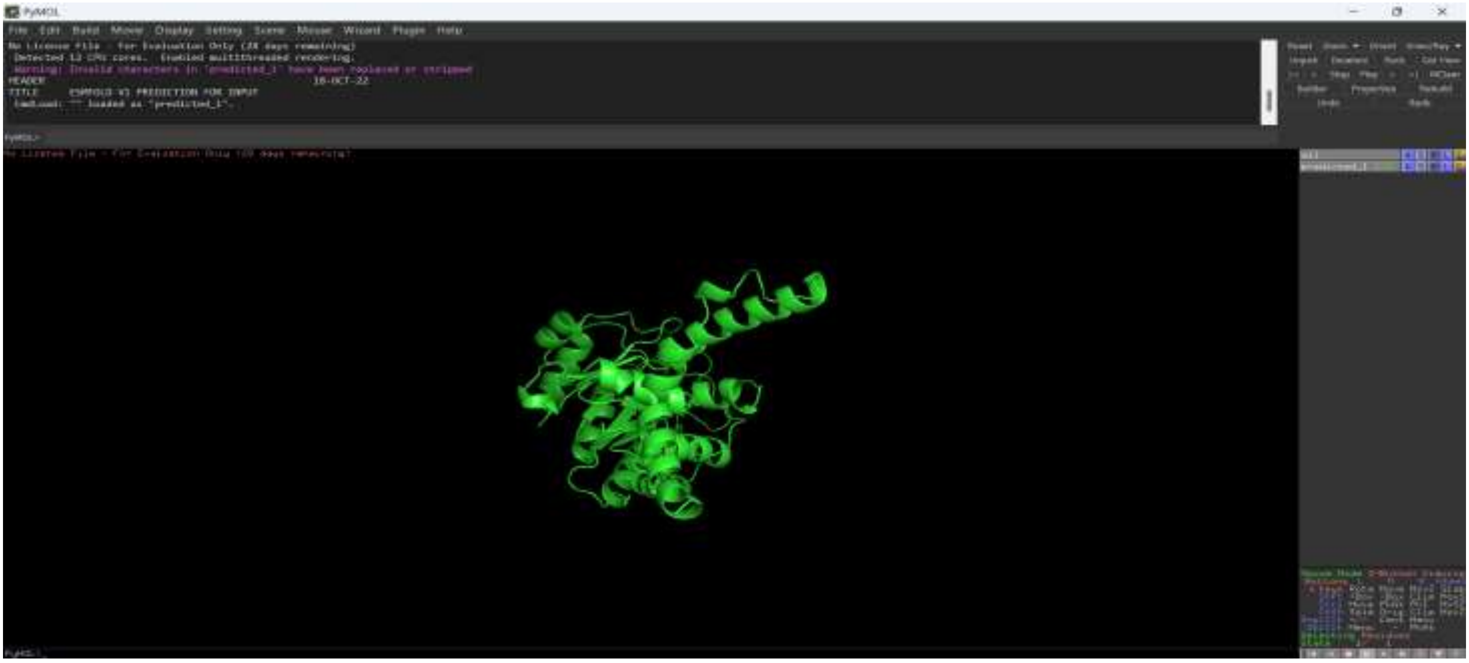
plDDT

plDDT is a per-residue estimate of the confidence in prediction on a scale from 0-100.

plDDT: 0.8665

fig 10. The image shows the evaluation of plDDT value of the 3-Dimensional Protein Structure.

The above fig 10 representing the plDDT value, that is a critical step in assessing the accuracy and reliability of computational models. The plDDT metric, which stands for per-residue normalized Distance Difference Test, is used to compare predicted protein structures with experimentally determined reference structures. The plDDT value is calculated by performing a distance difference test (DDT) between corresponding atoms in the predicted model and the reference structure. This test measures the discrepancies in atomic positions, reflecting how closely the predicted model matches the actual structure. Normalizing the DDT score on a per-residue basis in plDDT allows for a more nuanced evaluation, as it considers variations in accuracy across different regions of the protein.A higher plDDT value indicates better agreement between the predicted and reference structures, suggesting higher accuracy and reliability of the computational model. Conversely, a lower plDDT value signifies larger discrepancies or errors in the predicted model compared to the reference structure.This evaluation is crucial for validating computational models, identifying potential areas of improvement or refinement in the structure, and gaining confidence in using these models for further analysis or applications in areas such as drug discovery, protein engineering, and understanding molecular mechanisms.



PyMOL
File Edit Build Movie Display Setting Scene Mouse Wizard Plugin Help

```
No License File - For Evaluation Only (28 days remaining)
 Detected 12 CPU cores.  Enabled multithreaded rendering.
 Warning: Invalid characters in 'predicted_1' have been replaced or stripped
HEADER                                            18-OCT-22
TITLE     ESMFOLD V1 PREDICTION FOR INPUT
 CmdLoad: "" loaded as "predicted_1".
PyMOL>
No License File - For Evaluation Only (28 days remaining)
```

fig 11. The downloaded image of the Protein Structure.

The above fig 11 representing a downloaded 3D protein structure using PyMOL typically provides a detailed and visually informative depiction of the protein's molecular architecture. PyMOL, being a powerful molecular visualization tool, can display various aspects of the protein's structure and properties.The protein rendered in a three-dimensional space, often using spheres or sticks to represent atoms and bonds. The visualization may highlight key structural elements such as alpha helices, beta sheets, loops, and turns, giving insights into the protein's folding pattern and overall tertiary structure. These elements are usually color-coded to differentiate between different regions or types of amino acids.Additionally, PyMOL can visualize functional features of the protein, such as active sites, binding pockets, and interaction surfaces. These areas are often highlighted using specific colors or surface representations, making it easier to understand how the protein interacts with other molecules or substrates.Molecular interactions, such as hydrogen bonds, disulfide bridges, and non-covalent interactions, can also be depicted in the image. PyMOL can visualize these interactions spatially, providing a clearer picture of the molecular forces at play within the protein structure.the image generated using PyMOL serves as a valuable tool for studying protein structure-function relationships, facilitating research in fieldsuch as biochemistry, structural biology.

V. CONCLUSION

The Streamlit-based 3D protein prediction project has successfully demonstrated the power of interactive web applications in computational biology. By integrating machine learning algorithms, visualization tools, and user-friendly controls, we have created a platform that facilitates accurate and efficient protein structure modeling. This project not only enhances accessibility for researchers but also contributes to advancements in structural biology, drug discovery, and personalized medicine. Moving forward, continued improvements in algorithms and user interfaces will further elevate the impact and utility of such tools in the scientific community. The 3D protein structure prediction system utilizing Streamlit spans across diverse fields within bioinformatics and molecular biology, offering transformative solutions and accelerating research endeavors. In drug discovery, the system plays a crucial role in identifying potential drug targets by predicting protein-ligand interactions with high accuracy. This capability expedites the process of lead compound discovery and optimization, ultimately contributing to the development of novel therapeutics. In structural genomics, the system facilitates the annotation and analysis of protein structures at scale, enabling researchers to unravel complex biological mechanisms and pathways. Educational institutions leverage the system's intuitive interface and interactive visualization tools to provide students with hands-on experience in molecular modeling and bioinformatics, fostering a deeper understanding of protein structure-function relationships. Additionally, the system finds applications in protein engineering, where it aids in designing and optimizing proteins for specific industrial or biomedical purposes. Overall, the Streamlit-based 3D protein structure prediction system serves as a versatile and invaluable tool, empowering researchers, educators, and industry professionals to advance scientific knowledge and innovation in the field of molecular biology.

VI. ACKNOWLEDGEMENT

First and foremost we would like to thank GOD,the Almighty for being so merciful on us.
We have a great pleasure in expressing our deep sense of gratitude to founder Chairman Dr.K.R.Paramahamsa and Executive Vice President Mr. Rahul Kalluri for having provided us with a great infrastructure and well-furnished labs for successful completionof our Project.
We are grateful to our guide Ms.Baddala Vijayanirmala Asst.Prof.,Department of CSE,AMC Engineering College,Bengaluru for her/his constant motivation& timely help, encouragement and suggestion.
We would like to acknowledge and give our warmest thanks to The Advisory and Organizing Committee of City Engineering College for this great opportunity to facilitate the exchange of ideas, research findings, and best practices among engineering professionals and experts.

VII. REFERENCES

[1] Thompson, M. C., Yeates, T. O. & Rodriguez, J. A. Advances in methods for atomicresolution macromolecular structure determination. F1000Res. 9, 667 (2023)

[2] Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706 –710 (2023).

[3] Zheng, W. et al. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins: Struct. Funct.Bioinf. 87, 1149–1164 (2023).

[4] Brini, E., Simmerling, C. & Dill, K. Protein storytelling through physics. Science 370,(2023).

[5] Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One 6, e28766 (2023).

[6] Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. Proteins:Struct. Funct. Bioinf. 57, 702–710 (2022).

[7] Mirabello, C. & Wallner, B. rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. PLoS One14, e0220182 (2022).

[8] Huang, Z. et al. CCNet: Criss-Cross Attention for Semantic Segmentation. in Proceedings of the IEEE/CVFInternational Conference on Computer Vision 603–612 (2022).

[9] Fariselli, P., Olmea, O., Valencia, A. & Casadio, R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng. 14, 835–843 (2021).

[10] Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U.S. A. 117, 1496–1503 (2021).