Int.J.Curr.Microbiol.App.Sci (2024) *13*(08): 180-186

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 13 Number 8 (2024)**
Journal homepage: http://www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2024.1308.023

Protein Structure Prediction, Structural Bioinformatics and Deep Learning

Tejas Agarwal*

Delhi Public School R.K. Puram, New Delhi, Delhi, India

**Corresponding author*

Keywords

Protein Structure Prediction, Transformer Architecture, Deep Learning, Structural Bioinformatics

Article Info

Received:
15 June 2024
Accepted:
27 July 2024
Available Online:
10 August 2024

A B S T R A C T

Protein structure prediction is essential for understanding protein stability, and interactions. It holds immense potential for drug discovery and protein engineering. However, despite advancements in structural bioinformatics and artificial intelligence, a standardised model for structure prediction still needs to be worked out. Even prominent models like AlphaFold often undergo architectural changes. To address this gap, a comprehensive detail of recent progress and challenges in deep learning-based protein structure prediction has been presented. Additionally, a benchmark system for structure prediction and visualization, enabling analysis of user-provided protein sequences has been introduced. Looking to the need for efficient and accurate methods to decipher protein structures and their biological roles, DeepProtein has been introduced. This model leverages the potent representation learning capabilities of the Transformer architecture to directly predict secondary and tertiary structures from integer-encoded amino acid sequences. The results demonstrate DeepProtein's effectiveness in secondary structure prediction. Further refinement is necessary to enhance its performance in predicting higher-order structures. The present

Introduction

Protein structure prediction, a task in bioinformatics, aims to decipher the three-dimensional (3D) architecture of a protein solely from its amino acid sequence. This specific approach, known as end-to-end single-sequence protein structure prediction, deals with this challenge.

The model receives a protein sequence, typically represented as a string of integer-encoded amino acids, and directly outputs a predicted 3D structure in a standardized format like the Protein Data Bank (PDB). The model delves into the relationship between the protein's amino acid sequence and its nascent 3D structure (Jumper, *et al.,* 2021).

Unlike traditional methods that often rely on additional information such as homology data from similar proteins or predicted secondary structures within the same protein, single-sequence prediction requires only the amino acid sequence. This serves as a key advantage of single-sequence prediction. However, despite its straightforward nature, single-sequence prediction remains a demanding task. Achieving high accuracy in predicting complex protein structures solely from the sequence presents a major challenge. While the approach holds immense potential, its current performance may not always be sufficient for real-world applications. Consequently, single-sequence prediction is often employed in conjunction with other structure prediction methods or as a preliminary step in more intricate

workflows that leverage additional data sources. The technique is currently being put to use in multiple industries like– Drug Discovery, to identify potential binding sites for drugs; Disease Diagnosis, to better understand disease mechanisms and develop new therapies; Agriculture, to develop new crops that are more resistant to pests; etc.

The variability in protein structures poses a challenge in predicting 3D structures from amino acid sequences. Furthermore, protein structures exhibit dynamism, complicating efforts to capture their full spectrum in a single prediction. Deep learning models, including Neural Networks, capture the relationships between amino acid sequences and 3D protein structures. They excel in handling vast amounts of data necessary for accurate predictions, making them ideal for this task.

DeepProtein, a Transformer-based model, is made for end-to-end single-sequence protein structure prediction. Leveraging transformer architectures' robust representation learning capabilities, DeepProtein directly predicts protein secondary and tertiary structures from integer-encoded amino acid sequences. I delve into DeepProtein's implementation, evaluation, and limitations, exploring its potential applications in structural bioinformatics. The review highlights the efficacy of transformer-based models in protein structure prediction, opening paths for advancements in computational biology, drug discovery, and protein engineering.

Literature Review

Protein structure prediction has been revolutionized by Deep Learning resulting in understanding complex biological systems with ease. Neural Architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), etc. have been used to deal with the issues of predicting protein structures. We report the application of these architectures in structural bioinformatics and existing prediction models like AlphaFold along with their advantages and disadvantages.

Recurrent Geometric Networks

AlQuraishi introduced a novel end-to-end RGN architecture for protein structure prediction which aimed to overcome challenges of prediction by coupling local and global structures using geometric units. The model achieved success in predicting novel folds without co-evolutionary data and known folds without structural templates. The architecture is faster than existing methods, enabling new applications such as integrating structure prediction within docking and screening.

RGNs learn an implicit representation of protein fold space using secondary structure as the dominant factor in shaping their representation. It can also complement existing methods, such as incorporating structure templates for learning to improve secondary structure prediction.

The author predicts that hybrid systems using deep learning and co-evolution as priors, along with physics-based approaches will soon solve the problem of accurate prediction. The model, however, has limitations, such as reliance on position-specific scoring which could potentially be addressed with architectures more efficient with data.

Convolutional Neural Networks

Yang *et al.,* (2023) explored the effectiveness of using CNNs in place of Transformer-based models for pre-trained protein sequence models. Current models are not scalable due to the way Transformers scale which restricts the maximum sequence length that can be analyzed. The authors introduced CARP (Convolutional Autoencoding Representations of Proteins), a CNN-based architecture that scales linearly with sequence length.

The study showed that CARP models were equivalent in terms of efficiency with the other Transformer models across various downstream applications, including structure prediction. The study challenged the relationship between masked language modelling and Transformers, highlighting the importance of the pre-training tasks, not the Transformer architecture, which is essential for making pre-training effective. CARP showed strong performance on sequences longer than those allowed by current Transformer models, suggesting that computational efficiency can be improved without giving away the performance.

Language and Transfer Models

Chandra *et al.,* (2023) experimented with the application of Transformer models from the field of Natural Language Processing (NLP) for predicting protein

properties. These models, also known as protein language models, are capable of learning multipurpose representations of proteins from large open repositories of protein sequences.

The architecture has shown promising results in predicting protein characteristics such as post-translational modifications. It also provides advantages over traditional deep learning models, effectively capturing long-range dependencies in protein sequences (similar to natural language).

Various protein prediction tasks have been put forward for which Transformer models have been applied, including protein structure, protein residue contact, protein-protein interactions, drug-target interactions, and homology studies. These models have demonstrated impressive results without relying on Multiple Sequence Alignments (MSAs) or structural information. Additionally, the authors highlight the interpretability of Transformer models, which allows for visualization and analysis of attention weights and subsequently provides deeper biological insights. By training models with up to 15 billion parameters, the authors discovered that the models' understanding of protein sequences correlated with structure prediction accuracy.

Lin *et al.,* (2022) demonstrated the potential of large language models for evolutionary-scale prediction. The study introduces ESMFold, a fully end-to-end single-sequence structure predictor, which achieved a speedup of up to 60x on the inference forward pass compared to state-of-the-art methods like AlphaFold and RosettaFold. They note that simplifying the neural architecture and eliminating the need for multiple sequence alignments contributed to the improvement in speed.

Recurrent Neural Networks

Torrisi *et al.,* (2019) reviewed the recent advancements in protein structure prediction with the adoption of RNNs, Long Short-Term Memory (LSTM) networks, and Bidirectional RNNs (BRNNs). These models have excelled at handling sequential data and learning long-range dependencies, making them particularly well-suited for protein sequence analysis. Various protein structure predictors have been developed in recent years that combine multiple models to improve the accuracy of map prediction. CNNs have been combined with LSTMs to create SPOT-Contact. By leveraging these recurrent architectures, researchers have been able to exploit evolutionary information, yielding more sophisticated pipelines for protein structure prediction tasks.

While the literature has seen an upsurge in methods utilizing recurrent neural models, other Deep Learning approaches such as CNNs, Feed-Forward Neural Networks (FFNNs), and Residual Networks (ResNets) have also made significant contributions to the field. These advancements have resulted in considerable improvements in contact and distance map predictions, which have directly impacted the quality of 3D protein structure predictions. As computational resources, novel techniques, and experimental data continue to grow, further progress in protein structure prediction is expected, with recurrent architectures playing a significant role in this rapidly advancing field.

AlphaFold

A revolutionary deep learning method called AlphaFold[6] has emerged as a game-changer in protein structure prediction. It can predict protein structures with state-of-the-art accuracy, even for proteins without known similar structures. This achievement represents a significant leap forward in our understanding of proteins.

The secret to AlphaFold's success lies in its completely redesigned neural network architecture. This architecture incorporates knowledge of protein physics, biology, and multi-sequence alignments. During a key evaluation (14th Critical Assessment of Protein Structure Prediction), AlphaFold's performance was truly outstanding. It surpassed all other methods, achieving results comparable to – and often exceeding – those obtained through traditional, expensive experimental techniques.

Several key innovations within AlphaFold's architecture and training procedures contribute to its remarkable feat. These innovations leverage evolutionary relationships, physical constraints, and the geometric principles that govern protein structures. For instance, AlphaFold utilizes a novel architecture to effectively combine multi-sequence alignments and pairwise features. Additionally, it employs a new output representation specifically designed for accurate protein structure prediction, along with a tailored loss function for training the model. The model also benefits from a novel equivariant attention architecture and the use of intermediate losses to progressively refine its predictions. Notably, AlphaFold can even learn from unlabelled protein sequences

through a process called self-distillation and by using self-estimated accuracy measures. These advancements have collectively propelled AlphaFold to significantly improve protein structure prediction accuracy, providing valuable insights into protein function and opening new avenues for biological research.

Materials and Methods

This study utilizes the ProteinNet dataset, a comprehensive resource designed to train and evaluate models for protein structure prediction (AlQuraishi, 2019b). Protein Net integrates sequence, structure, and evolutionary information in user-friendly formats. Notably, it includes high-quality multiple sequence alignments and standardized data splits that mimic the difficulty of past CASP experiments. These splits allow the creation of validation sets distinct from official CASP sets while retaining their challenge.

Here, the CASP-12 dataset has been leveraged from SideChainNet, a ProteinNet extension offering curated protein sequences and structures (King and Koes, 2021). CASP-12 includes target proteins with experimentally determined structures and corresponding predictions from CASP12 participants (CASP12, 2016). Standard metrics like Root-Mean-Square Deviation (RMSD) are used to evaluate predicted structure quality. These metrics assess the model's ability to predict overall protein fold, topological similarity, and atomic-level accuracy. SideChainNet pre-processes data by integer-encoding sequences, computing dihedral angles, and providing missing residue masks. The model utilizes this data, ingesting encoded sequences and masks, predicting angles, and generating protein structure PDB files.

Implementation

The system is an Attention-based model for predicting the protein structure from amino acid sequences. The approach can predict 12 angles provided by the dataset which is more efficient than AlQuaraishi's (2019a) model which only used 3 angles to predict (see Figure 1).

The model outputs a shape of $L \times 12 \times 2$ with values between [-1, 1]. This allows the management of the circular nature of angles more effectively (Basu *et al.,* 2022). The model takes an amino acid sequence as integer tensors which produce angle vectors for each acid. The model predicts the sin and cos values for each acid angle finally using the $atan2$ function to recover the angles which returns an absolute value between [0, π].

The model leverages a powerful attention mechanism based on the concept of Multi-Head Self-Attention. This mechanism dissects the input sequence of amino acids into three components: queries, keys, and values. Each component is obtained by transforming the original sequence using separate matrices.

The key advantage of Multi-Head Self-Attention is its ability to learn relationships between amino acids in parallel across multiple "heads." This allows it to capture structural features within the protein sequence more effectively.

Additionally, a gating mechanism is integrated to control the flow of information between the input and output. This gating function enables the model to selectively emphasize important relationships between amino acids while discarding irrelevant information.

To calculate the attention scores, the model first scales the query vectors and then computes their dot product with the key vectors. This dot product reflects the similarity between each pair of amino acids in the sequence.

These similarities are then used to weigh the value vectors, resulting in a weighted sum that captures the structural information at each position in the sequence. Finally, the scores are normalized using a $softmax$ function to ensure they represent a valid probability distribution over the sequence positions.

Often, the protein sequences contain missing residues. The proposed model incorporates the attention mechanism to deal with this issue. A binary mask is used to identify the presence or absence of each amino acid. During the attention calculation, positions corresponding to missing residues are filled with a very negative value. This effectively excludes them from the final probability distribution, ensuring the model focuses solely on the available structural information and ignores missing residues.

The system also offers functionalities for visualizing and comparing predicted protein structures to their actual

structures (see Figure 2). These visualizations are generated using the BatchedStructureBuilder class from SideChainNet (Basu *et al.,* 2022). This class takes two tensors: one representing the batch of amino acid sequences and another containing the predicted angles for each residue.

Evaluation

A rigorous evaluation framework has been utilized to assess the performance of transformer-based protein structure prediction model. This framework quantifies the accuracy of the model's predictions using the Root-Mean-Square Deviation (RMSD) metric. RMSD estimates the overall structural differences between the predicted and experimentally determined protein structures.

RMSD is calculated as the square root of the Mean Squared Error (MSE) loss. This metric effectively measures the atomic-level accuracy of the predicted protein structures.

$$RMSD = \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

During training, the model achieved a final training RMSE loss of 0.448, validation loss of 0.4418, and test loss of 0.4379. These results indicate that the model generalizes well to unseen data (see Figure 3).

One of the model's strengths lies in predicting secondary structures. These structures, such as α-helices and β-sheets, form the local backbone conformation of proteins and are crucial for understanding protein function and stability.

Results and Discussion

Evaluation of framework revealed promising results for DeepProtein, particularly in its ability to predict protein secondary structures. The model's low RMSE across training, validation, and test sets suggest good generalizability and potential for accurately predicting the local backbone conformations that influence protein function and stability.

However, initial attempts using Tens or Flow and Keras with atomic coordinates for training resulted in overfitting and high VRAM (Video RAM) requirements. The model currently faces limitations in predicting the overall 3D folding (tertiary structures) of proteins, which is crucial for understanding protein interactions and designing therapeutics.

This is likely due to insufficient training data and architectural complexity hindering the model's ability to capture long-range interactions between amino acids. Future work will focus on expanding the training data, incorporating additional information like multi-sequence alignments, and refining the model architecture to improve its grasp of these long-range interactions.

By overcoming these limitations, DeepProtein can evolve into a powerful tool for protein structure prediction, unlocking new avenues for understanding protein function and developing novel therapeutics.

DeepProtein's ability to predict secondary structures offers a valuable tool for understanding protein function and stability, however there are limitations in predicting tertiary and quaternary structures, the model's performance suggests a promising path forward in structural bioinformatics and protein folding research.

Future work needs to focus on expanding the training data with diverse protein sequences and incorporating additional information sources like Multi-Sequence Alignments (MSAs) that capture evolutionary relationships between proteins. This enriched data, combined with potential architectural refinements to capture long-range interactions, could significantly improve DeepProtein's ability to predict higher-order structures.

Successful prediction of complex protein structures would be a major milestone for computational biology. It would open doors for a deeper understanding of protein function, protein-protein interactions, to revolutionize drug discovery and protein engineering through the power of artificial intelligence. DeepProtein lays the groundwork for this future, and ongoing development holds the potential to establish a standardized neural architecture for protein structure prediction.

DeepProtein is a tool that helps predict protein structures, which is important for understanding how proteins work and their stability. Future improvements could make it even better at predicting complex structures, aiding in drug discovery and protein engineering.

Figure.1 RGN Architecture by AlQuraishi (2019a)

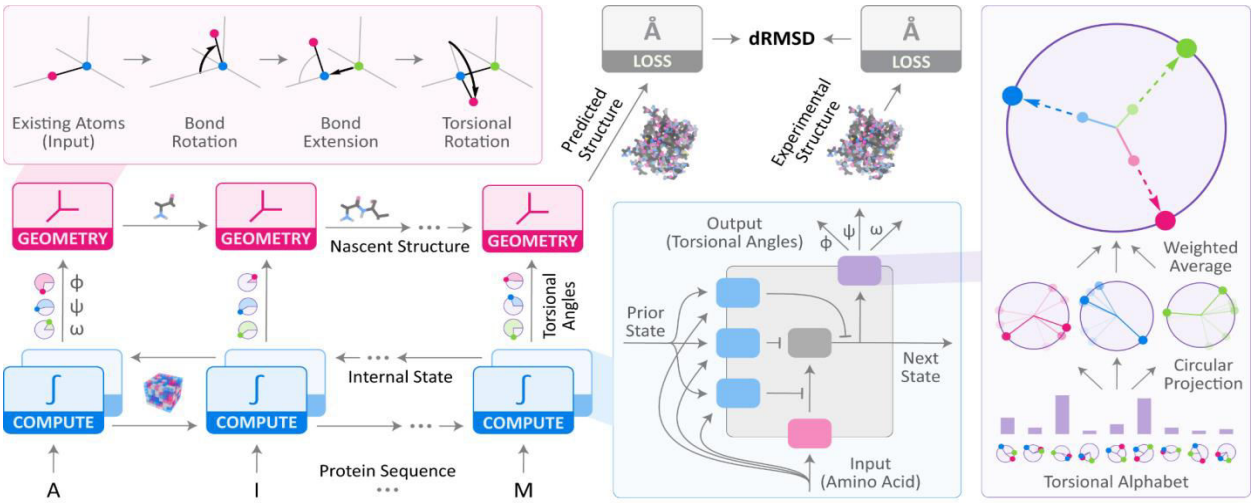


Figure.2 Example model prediction from the training set by AlQuraishi (top) and ground truth (bottom), visualized with Py3DMOL

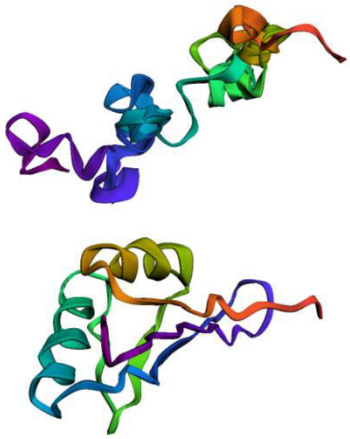
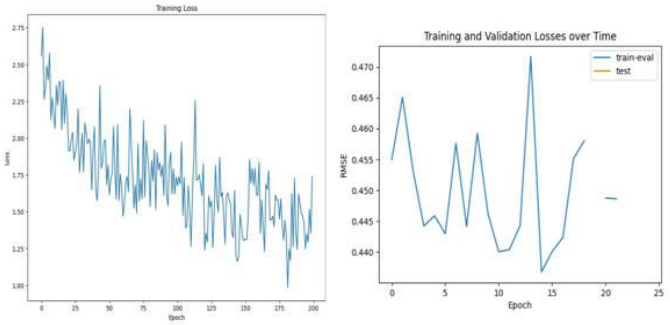


Figure.3 RMSE loss over 25 epochs with batch size 4



Author Contributions

Tejas Agarwal: Investigation, formal analysis, writing—original draft.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Conflict of Interest The authors declare no competing interests.

References

AlQuraishi, M. (2019a). End-to-end differentiable learning of protein structure. Cell Systems. https://doi.org/10.1016/j.cels.2019.03.006

AlQuraishi, M. (2019b). Proteinnet: A standardized data set for machine learning of protein structure. BMC Bioinformatics. https://doi.org/10.1186/s12859-019-2932-0

Basu, V. (2022, August). Attention based protein structure prediction. Kaggle. Retrieved from https://www.kaggle.com/code/basu369victor/attention-based-protein-structure-prediction/notebook

CASP12. (2016, April). Home. Retrieved from https://predictioncenter.org/casp12/index.cgi

Chandra, A., Tunnermann, L., Lofstedt, T., & Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. eLife. https://doi.org/10.7554/eLife.82819

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with Alphafold. Nature. https://doi.org/10.1038/s41586-021-03819-2

King, J. E., & Koes, D. R. (2021). Sidechainnet: An all-atom protein structure dataset for machine learning. Proteins: Structure, Function, and Bioinformatics. https://doi.org/10.1002/prot.26169

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. bioRxiv. https://doi.org/10.1101/2022.07.20.500902

Torrisi, M., Pollastri, G., & Le, Q. (2019). Deep learning methods in protein structure prediction. Computational and Structural Biotechnology Journal. https://doi.org/10.1016/j.csbj.2019.12.011

Yang, K. K., Fusi, N., & Lu, A. X. (2023). Convolutions are competitive with transformers for protein sequence pretraining. bioRxiv. https://doi.org/10.1101/2022.05.19.492714

How to cite this article:

Tejas Agarwal. 2024. Protein Structure Prediction, Structural Bioinformatics and Deep Learning. *Int.J.Curr.Microbiol.App.Sci.* 13(8): 180-186. **doi:** https://doi.org/10.20546/ijcmas.2024.1308.023