*Bioinformatics*, 2023, 1–8
doi: PREPRINT
Advance Access Publication Date: 26 April 2023
Original Paper

Structural bioinformatics

# Deep Learning for Protein Structure Prediction: Advancements in Structural Bioinformatics

**Daniel Szelogowski***

Bioinformatics, UW - Green Bay, 54311, Wisconsin, USA

*Corresponding author. E-mail: dszelogowski@gmail.com

## Abstract

**Motivation:** Accurate prediction of protein structures is crucial for understanding protein function, stability, and interactions, with far-reaching implications in drug discovery and protein engineering. As the fields of structural bioinformatics and artificial intelligence continue to converge, a standardized model for protein structure prediction is still yet to be seen as even large models like AlphaFold continue to change architectures. To this end, we provide a comprehensive literature review highlighting the latest advancements and challenges in deep learning-based structure prediction, as well as a benchmark system for structure prediction and visualization of amino acid protein sequences.

**Results:** We present ProteiNN, a Transformer-based model for end-to-end single-sequence protein structure prediction, motivated by the need for accurate and efficient methods to decipher protein structures and their roles in biological processes and a system to perform prediction on user-input protein sequences. The model leverages the transformer architecture's powerful representation learning capabilities to predict protein secondary and tertiary structures directly from integer-encoded amino acid sequences. Our results demonstrate that ProteiNN is effective in predicting secondary structures, though further improvements are necessary to enhance the model's performance in predicting higher-level structures. This work thus showcases the potential of transformer-based architectures in structure prediction and lays the foundation for future research in structural bioinformatics and related fields.

**Availability and implementation:** The source code of ProteiNN is available at https://github.com/danielathome19/ProteiNN-Structure-Predictor.

**Key words:** Protein Structure Prediction, Transformer Network, Deep Learning, Computational Biology

## 1. Introduction

End-to-end single-sequence protein structure prediction is a task in Bioinformatics to predict the 3D structure of a protein from its amino acid sequence. In an end-to-end single-sequence model, the input is a protein sequence (as a string of amino acids, typically then integer encoded), and the output is a predicted 3D structure in a standard format such as **PDB (Protein Data Bank)**. The model is trained to learn the relationship between the amino acid sequence and the nascent 3D structure of the protein, often either by angles or coordinates. The main advantage of single-sequence prediction is that it does not require any additional information about the protein other than its amino acid sequence, unlike traditional methods that often rely on additional information such as homology information or predicted secondary structure. Single-sequence prediction is still challenging, however; its accuracy is not always sufficient for practical applications, so it is often used in combination with other methods or as a preliminary step in more complex structure prediction pipelines.

Protein structure prediction is currently being applied in numerous fields:

- **Drug discovery** — accurate predictions of protein structures can aid researchers in the design of new drugs that target specific proteins. By understanding the structure of a protein, researchers can identify potential binding sites for drugs, which can be helpful in the design of new drugs
- **Biotechnology** — predicted structures can be applied to design new enzymes and other biotechnology products that are more efficient and effective
- **Disease diagnosis** — understanding the structure of proteins involved in diseases can help researchers better understand

disease mechanisms and develop new therapies. For example, predictions of the structures of proteins involved in cancer can be used to design new drugs that target these proteins
- **Agriculture** — predictions can be used to improve crop yields and develop new crops that are more resistant to pests and diseases
- **Environmental monitoring** — predictions can be used to monitor ecological pollutants and understand their effects on living organisms

The task of structure prediction presents numerous challenges for researchers — although recent advances in deep learning have provided a valuable baseline for accelerating research on the topic. A significant degree of variability in protein structures makes it difficult to predict the 3D structure from the amino acid sequence. Protein structures are also dynamic, meaning they can change over time, making it challenging to capture the full range of structures in a single prediction. Neural Networks and other Deep Learning models have been able to capture the complex relationships between the amino acid sequence and the 3D structure of proteins, leading to more accurate predictions, as well as being able to handle the large amounts of data that are required for protein structure prediction — making them well-suited for this task.

We present ProteiNN, a Transformer-based model for end-to-end single-sequence protein structure prediction, designed to advance our understanding of protein folding and function. By leveraging transformer architectures' powerful representation learning capabilities, ProteiNN predicts protein secondary and tertiary structures directly from integer-encoded amino acid sequences. We discuss the implementation, evaluation, and limitations of the ProteiNN model, explore its potential applications in structural bioinformatics, and provide a comprehensive review of the literature. This work demonstrates the utility of transformer-based models in predicting protein structures and paves the way for future advancements in computational biology, drug discovery, and protein engineering.

## 2. Related Work

Deep learning has revolutionized the field of protein structure prediction, driving significant advancements in understanding complex biological systems. Various neural architectures — such as **Convolutional Neural Networks (CNNs)**, **Recurrent Neural Networks (RNNs)**, and **Recurrent Geometric Networks (RGNs)** — have been employed to tackle the challenges associated with predicting protein structures. These innovative approaches have enabled researchers to model complex protein structures more accurately, unlocking new insights into protein function, interaction, and evolution. This literature review explores the application of these deep learning architectures in the realm of structure prediction, highlighting their respective strengths, limitations, and potential for future advancements in the field.

### 2.1. Convolutional Neural Networks

Yang et al. (2023) investigate the efficiency and effectiveness of using CNNs in place of Transformer-based models for pre-trained protein sequence language models. Current protein language models are limited in scalability due to the quadratic scaling of Transformers, which restricts the maximum sequence length that can be analyzed. The authors thus introduce **CARP (Convolutional Autoencoding Representations of Proteins)**, a CNN-based architecture that scales linearly with sequence length. The study demonstrates that CARP models are competitive with the state-of-the-art Transformer model ESM-1b (Rives et al., 2021) across various downstream applications, including structure prediction, zero-shot mutation effect prediction, and out-of-domain generalization. The study challenges the association between masked language modeling and Transformers, highlighting that the pre-training task, not the Transformer architecture, is essential for making pre-training effective. Furthermore, CARP shows strong performance on sequences longer than those allowed by current Transformer models, suggesting that computational efficiency can be improved without sacrificing performance using a CNN architecture.

### 2.2. Recurrent Neural Networks

Torrisi et al. (2020) review the recent advancements in protein structure prediction that have been bolstered by the introduction of Deep Learning techniques, including 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. In recent years, various protein structure predictors have been developed, such as SPOT-Contact, which combines CNNs and 2D BRNNs with LSTM units to improve the accuracy of contact map prediction. 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. Methods such as DeepCDpred, DeepContact, DeepCov, DNCON2, MetaPSICOV, PconsC4, RaptorX-Contact, TripletRes, and AlphaFold have each employed unique architectures and input features to improve protein structure prediction. 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.

### 2.3. Recurrent Geometric Networks

AlQuraishi (2019a) introduces the novel end-to-end differentiable RGN architecture for protein structure learning. This model aims to overcome the central challenge of predicting protein structures from sequences by coupling local and global structures using geometric units. The model achieved state-of-the-art accuracy in two challenging tasks — predicting novel folds without co-evolutionary data and known folds without structural templates. The RGN architecture allows a model to implicitly encode multi-scale protein representations and predict structures by integrating information from residues upstream and downstream. Likewise, RGNs are substantially faster than existing methods, potentially enabling new applications such as integrating structure prediction within docking and virtual screening.

RGNs learn an implicit representation of protein fold space using secondary structure as the dominant factor in shaping their representation, despite not being explicitly encoded with the concept. The architecture can also complement existing methods, such as incorporating structural templates or co-evolutionary information as priors or inputs for learning to improve secondary structure prediction. As such, the author predicts that hybrid systems using deep learning and co-evolution as priors, along with physics-based approaches for refinement, will soon solve the long-standing problem of accurate and efficient structure prediction. However, the model has limitations, such as its reliance on **Position-Specific Scoring Matrices (PSSMs)**, which could potentially be addressed with more data-efficient model architectures.

### 2.4. Transformer Models and Attention Networks

Chandra et al. (2023) discuss the application of Transformer models from the field of **Natural Language Processing (NLP)** for predicting protein properties. These models — referred to 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).

The authors review various protein prediction tasks 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.

While they have shown significant improvements over RNNs and other models, they possess limitations such as fixed-length input sequences and quadratic growth in memory requirements. Nevertheless, advancements are being made to address these issues, such as incorporating hidden states from previous fragments and implementing sparse attention mechanisms to reduce computational complexity. Although Transformers have outperformed other models in many tasks, they may not always be the best choice for all protein prediction tasks, and a combination of methods may be necessary. As the applications of Transformers in computational biology and bioinformatics are still in their infancy, further improvements and special-purpose models can be expected. The proof-of-principle example in this review displays the potential of using Transformers as general feature extractors to improve results compared to traditional protein features, like MSAs, but the authors note that more research is needed to determine their overall superiority. Ultimately, the future of computational biology and bioinformatics will likely involve advancements in Transformer models and their integration with other methods for protein analysis.

### 2.5. Language Models

Lin et al. (2022) demonstrate the potential of large language models for evolutionary-scale prediction of atomic-level protein structures. By training models with up to 15 billion parameters, the authors discovered that the models' understanding of protein sequences correlated with structure prediction accuracy. 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.

Similarly, the authors also present the ESM Metagenomic Atlas, which offers the first large-scale structural characterization of metagenomic proteins comprised of more than 617 million structures. This atlas provides an unprecedented view into the vast diversity of some of the least understood proteins on Earth. The combined results of ESMFold and the ESM Metagenomic Atlas indicate that as language models scale, their ability to predict protein structures improves, especially for proteins with low evolutionary depth. This advancement in speed and accuracy could accelerate the discovery of new protein structures and functions, leading to potential breakthroughs in medicine and biotechnology.

### 2.6. AlphaFold

Jumper et al. (2021) present a groundbreaking computational method capable of predicting protein structures with atomic accuracy even when no similar structure is known. **AlphaFold**, the neural network-based model at the core of this work, has been completely redesigned to incorporate physical and biological knowledge about protein structure and multi-sequence alignments. In the 14th **Critical Assessment of [Protein] Structure Prediction (CASP14)**, AlphaFold demonstrated remarkable accuracy, outperforming other methods and achieving results competitive with experimental structures in a majority of cases.

AlphaFold's success can be attributed to its novel neural architectures and training procedures, which are based on evolutionary, physical, and geometric constraints of protein structures. Key innovations include the use of a new architecture to jointly embed MSAs and pairwise features, a new output representation and associated loss for accurate end-to-end structure prediction, a new equivariant attention architecture, the use of intermediate losses for iterative refinement of predictions, masked MSA loss for joint training with the structure, and learning from unlabelled protein sequences through self-distillation and self-estimates of accuracy. These advances have enabled AlphaFold to significantly improve the accuracy of protein structure prediction and provide valuable insights into protein function and biology.

The end-to-end structure prediction in AlphaFold relies on a structure module that operates on a 3D backbone structure using the pair representation and the original sequence row from the MSA representation in the trunk. The structure module updates the global frame (residue gas) representation iteratively in two stages. First, it uses **Invariant Point Attention (IPA)** to update a set of neural activations without changing the 3D positions. Next, the updated activations perform an equivariant update operation on the residue gas. The final loss calculation — called the **Frame-Aligned Point Error (FAPE)** — compares the predicted atom positions to the true positions under multiple alignments. Predictions of the structure's side-chain angles and per-residue accuracy are computed with small per-residue networks at the end of the process.

AlphaFold's architecture is trained with both labeled and unlabeled data, enhancing accuracy using an approach similar to noisy student self-distillation. A separate structure module is trained for each of the 48 Evoformer blocks in the network, providing a trajectory of 192 intermediate structures that represent the network's belief of the most likely structure at each block. The authors note that the accuracy of AlphaFold decreases when the median alignment depth is less than 30 sequences; however, the model is still effective for proteins with few intra-chain or homotypic contacts compared to the number of heterotypic contacts. The architectural concepts in AlphaFold are expected to apply to predicting full hetero-complexes in future systems, overcoming the difficulty with protein chains with many hetero-contacts.

## 3. Approach

The ProteinNet dataset is a comprehensive and standardized resource designed to facilitate the training and evaluation of data-driven models for protein sequence-structure relationships, including structure prediction and design (AlQuraishi, 2019b). The dataset integrates protein sequence, structure, and evolutionary information in machine learning-friendly file formats to provide accessibility for researchers. One of the critical components of ProteinNet is the high-quality multiple sequence alignments of all structurally characterized proteins, which were generated using substantial high-performance computing resources. Additionally, the dataset includes standardized splits of data (e.g., train, train-eval, test, and validation) that emulate the difficulty of past CASP experiments (which aim to assess the state-of-the-art in protein structure prediction) by resetting protein sequence and structure space to the historical states preceding six prior CASPs. These data splits (constructed using sensitive evolution-based distance metrics to segregate distantly related proteins) allow for the creation of validation sets distinct from the official CASP sets while still faithfully mimicking their difficulty. Hence, ProteinNet serves as a valuable resource for developing and assessing machine learning models in protein structure prediction.

In this study, we utilize the CASP-12 dataset provided by the SideChainNet python library, which offers a comprehensive and curated collection of protein sequences and structures as an extension of ProteinNet (King and Koes, 2021). CASP-12 consists of target proteins whose structures were determined experimentally, along with the predictions submitted by participating research groups (CASP12, 2016). To evaluate the quality of the predicted structures, we perform evaluations using standard metrics such as the **Root-Mean-Square Deviation (RMSD)** between the predicted and experimental structures. These metrics provide a comprehensive view of the model's ability to predict the overall fold, topological similarity, and atomic-level accuracy of protein structures. SideChainNet preprocesses the raw protein data (integer-encodes amino acid sequences), computes the dihedral angles for each residue, and provides the missing residue masks, which help the model handle incomplete sequence information. Our model leverages this dataset by ingesting the integer-encoded amino acid sequences and missing residue masks, predicting the angles, and generating the PDB files representing the three-dimensional structures of the proteins.

## 4. Implementation

We present a Transformer model for end-to-end single-sequence protein structure prediction implemented in Python 3.9 with the PyTorch library, trained using an RTX 3060 Ti with 8GB of VRAM. The model accepts integer-encoded amino acid sequences and missing residue masks as input and predicts the angles to generate a PDB file as output. Our approach employs Attention mechanisms to effectively capture the dependencies among amino acids and their spatial arrangement, which is critical for understanding protein structure and function.

### 4.1. ProteiNN

Our system features an Attention-based model — **ProteiNN** — for predicting protein structure from amino acid sequences. The model was trained and evaluated using the SideChainNet dataset, which provides the basis for complete model training. Our approach predicts the 12 angles provided by the dataset, differing from AlQuraishi's (2019a) models that only utilized 3 angles (see Figure 1). Furthermore, the model's output is in the shape of $L \times 12 \times 2$ with values between [-1, 1], allowing us to handle the circular nature of angles more effectively (Basu, 2022). The final model takes as input the amino acid sequences represented as integer tensors ($\mathbb{R}^1$), which are then processed through an attention mechanism to produce angle vectors for each amino acid. Our model then predicts the sin and cos values for each angle. Finally, we use the **atan2** function to recover the angles after they are predicted, which provides the model with the knowledge that angles $\pi$ and $-\pi$ are the same (i.e., the sign of the angle does not modify the radians of the angle).

The Attention module in our architecture is based on the **Multi-Head Self-Attention** mechanism, comprised of several components, including the query, key, and value transformation matrices. The input to the module is a sequence of feature vectors mapped to these component spaces via linear transformations. The multi-head mechanism enables the model to learn relationships between amino acids in parallel, improving its ability to capture complex structural features. We also incorporate a gating mechanism that modulates the information flow between the input and output, allowing the model to emphasize specific relationships and discard irrelevant information selectively (see Figure 5).

To compute the attention scores, we first scale the query vectors by the inverse square root of the key dimension. This normalization step helps stabilize the gradients during training and facilitates faster convergence. Next, we compute the dot product between the query and key vectors, which captures the pairwise similarities between the amino acids. These similarities are then used to weigh the value vectors, generating a weighted sum representing the structural information at each position in the sequence. Finally, the scores are normalized using the **softmax** function to ensure a valid probability distribution over the sequence positions.

To handle missing residue information, we introduce masking in the attention mechanism. The model employs a binary mask that indicates the presence or absence of each amino acid in the input sequence. During the attention calculation, the masked positions are filled with a large negative value to effectively exclude them from the softmax normalization step, ensuring that the model ignores the missing residues and focuses on the available structural information.

Our system also provides methods for visualizing and comparing the predicted protein structures to the ground truth

(see Figure 2). The **BatchedStructureBuilder** class from SideChainNet is used to generate these structures (i.e., an entire batch), taking a tensor (batch $\times$ $L$) of integers representing the amino acid sequences in the batch and a tensor (batch $\times$ $L$ $\times$ NUM_ANGLES) of floating-point numbers in the range $[-\pi, \pi]$ representing the predicted angles for each residue of each amino acid in the protein (Basu, 2022).

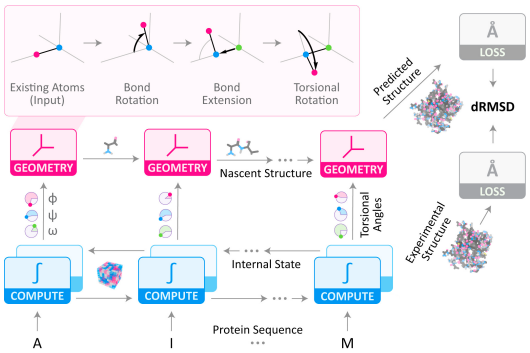



**Fig. 1**: RGN architecture, modified for our system (AlQuraishi, 2019a)

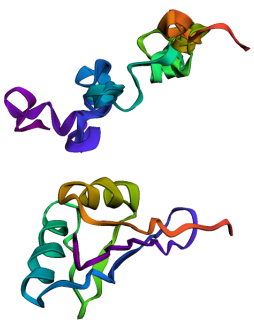


**Fig. 2**: Example model prediction from the training set (top) and ground truth (bottom), visualized with Py3DMOL

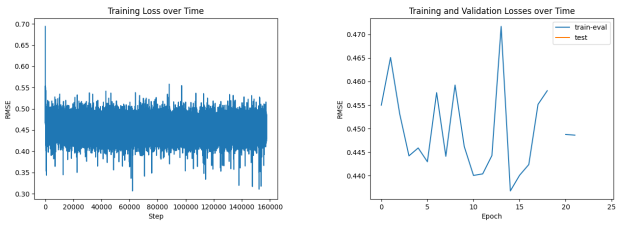
## 5. Evaluation

We employ a rigorous evaluation framework to assess the performance of our transformer-based protein structure prediction model, quantifying the accuracy of the model's predictions using the RMSD metric (which estimates of the overall structural differences between the predicted and experimental protein structures). RMSD, calculated as the square root of the **Mean Squared Error (MSE)** loss, effectively measures the atomic-level accuracy of the predicted protein structures.

Our model achieved a final training RMSE loss of 0.448, validation loss of 0.4418, and test loss of 0.4379, indicating that it generalizes well to unseen data (see Figure 3). The model excels in predicting secondary structures, such as $\alpha$-helices and $\beta$-sheets, which form the local backbone conformation of proteins and are thus crucial for understanding protein function and stability.

In addition to the quantitative evaluation using RMSE values, we also qualitatively analyze the model's performance by visually comparing the predicted protein structures to their corresponding experimentally determined structures. After training, we use the model to predict the angles for a given input sequence and subsequently construct 3D atomic coordinates for the predicted structure. We then compare the predicted structures to their true structures using the 3D molecular visualization tools Jalview and PyMOL (see Figure 4).

This visualization allows us to observe the local and global conformations of the predicted protein structures and identify regions where the model's predictions closely resemble or diverge from the true structures. By examining the overall quality of the predicted structures, we can gain valuable insights into the model's capabilities and limitations in capturing both secondary and tertiary structural features of proteins. Likewise, the combination of quantitative and qualitative evaluation methods provides a more comprehensive understanding of our model's performance in structure prediction and informs potential improvements for future iterations.




(a) Loss of each batch over time

(b) Loss of each epoch over time; note the overfit epoch (19) where RMSE = $\infty$

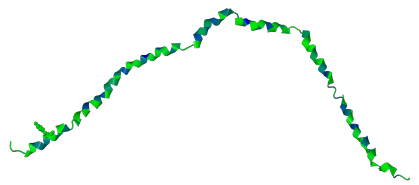
**Fig. 3**: Training history (RMSE loss) over 25 epochs with batch size 4

## 6. Discussion

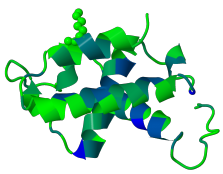
In light of the results obtained from our evaluation, we observe that the model demonstrates promising performance in predicting secondary structures. The model's RMSE values indicate good generalization to unseen data and highlight its potential for accurately predicting local backbone conformations of proteins. While we originally attempted to implement the model with Tensorflow and Keras with the atomic coordinates as training data, the results indicated drastic overfitting. This initial approach required additional preprocessing, leading to an increased requirement of VRAM to store the data during training for equally poor results.

Our model currently faces limitations in predicting tertiary structures, which represent the overall folding and three-dimensional arrangement of the protein's secondary structural elements. The accurate prediction of these structures is essential for deciphering protein-protein interactions, protein-nucleic acid interactions, and the design of novel therapeutics. This limitation can be attributed to insufficient training data and architectural complexity, which hampers the model's ability to capture the complex relationships between amino acids and their spatial

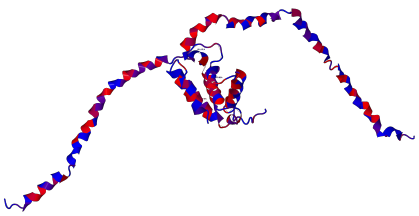
arrangement. To address this issue, future work should focus on expanding the training dataset, incorporating additional sources of information such as MSAs or evolutionary coupling data, and refining the model architecture to improve its ability to capture long-range interactions.



(a) Predicted



(b) Actual



(c) Aligned with Jalview

**Fig. 4**: Sample structure prediction on unseen data in comparison to the actual structure (from ProteinNet) and their alignment

## 7. Conclusion

To address the limitations in tertiary structure prediction, we suggest future research focusing on expanding the training dataset, incorporating additional features, and refining the neural model to improve its ability to capture long-range interactions. As well, we neglected to provide additional feature space in the architecture to support training on the coordinates or PSSMs; however, this may be better suited for use in an ensemble model, with one network for each feature set and a final concatenation layer for the complete prediction. These improvements can potentially enhance the model's performance in predicting high-level structures, ultimately contributing to a more comprehensive understanding of protein folding and function.

Our Transformer-based model for protein structure prediction has demonstrated its effectiveness in predicting secondary structures, providing a valuable tool for understanding protein function and stability. Although the model currently faces challenges in predicting tertiary and quaternary structures (especially the folded structure), the results suggest that the proposed approach holds promise for advancing our knowledge in structural bioinformatics and protein folding. As we continue to improve the model by addressing its current limitations, we anticipate that this work will contribute to the development of a standardized neural architecture. The successful prediction of higher-level protein structures is a significant milestone for computational biology and a stepping stone toward harnessing the full potential of artificial intelligence in drug discovery and protein engineering.

## References


M. AlQuraishi. End-to-end differentiable learning of protein structure. *Cell Systems*, 8(4):292–301.e3, Apr 2019a. doi: 10.1016/j.cels.2019.03.006.

M. AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20(1):311, Jun 2019b. doi: 10.1186/s12859-019-2932-0.

V. Basu. Attention based protein structure prediction. *Kaggle*, Aug 2022. URL `https://www.kaggle.com/code/basu369victor/attention-based-protein-structure-prediction/notebook`.

CASP12. Home, Apr 2016. URL `https://predictioncenter.org/casp12/index.cgi`.

A. Chandra, L. Tünnermann, T. Löfstedt, and R. Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, 12:e82819, jan 2023. ISSN 2050-084X. doi: 10.7554/eLife.82819.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873): 583–589, Aug 2021. doi: 10.1038/s41586-021-03819-2.

J. E. King and D. R. Koes. Sidechainnet: An all-atom protein structure dataset for machine learning. *Proteins: Structure, Function, and Bioinformatics*, 89(11):1489–1496, 2021. doi: https://doi.org/10.1002/prot.26169.

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

A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2016239118`.

D. Szelogowski. ProteiNN-Structure-Predictor, may 2023. URL `https://github.com/danielathome19/ProteiNN-Structure-Predictor`.

M. Torrisi, G. Pollastri, and Q. Le. Deep learning methods in protein structure prediction. *Computational and Structural Biotechnology Journal*, 18:1301–1310, 2020. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2019.12.011.

K. K. Yang, N. Fusi, and A. X. Lu. Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv*, 2023. doi: 10.1101/2022.05.19.492714.

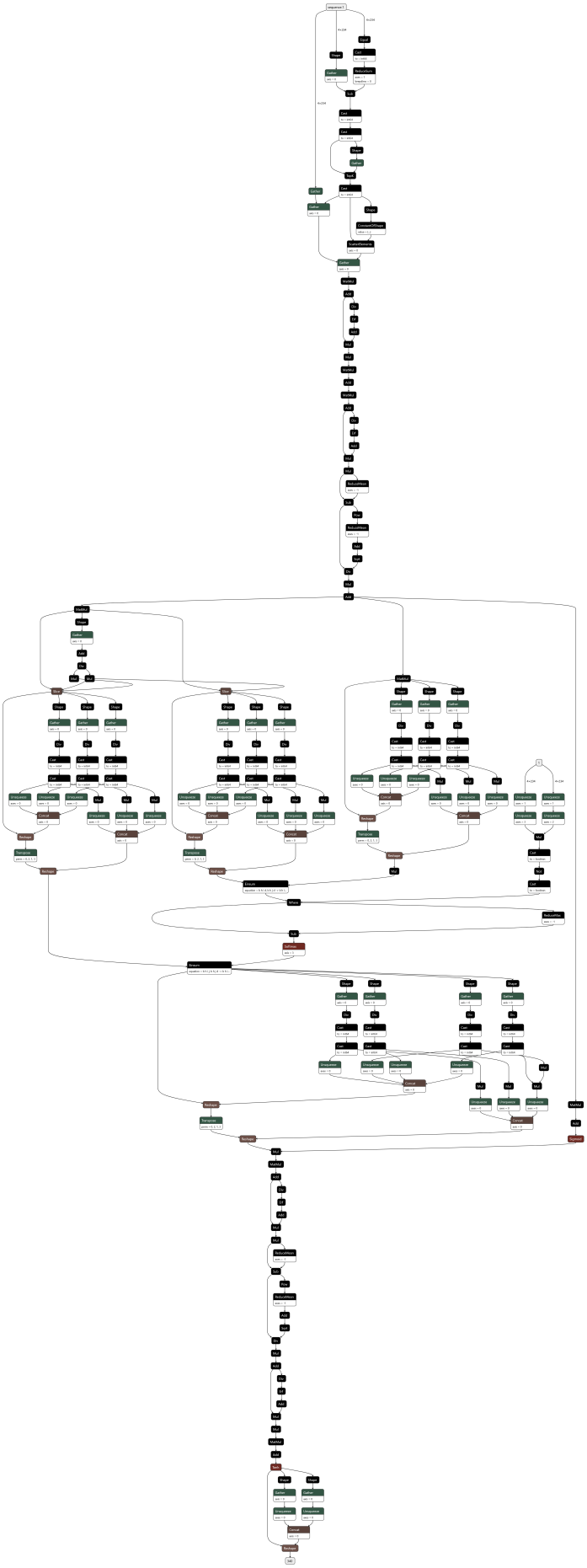


**Fig. 5**: System Architecture Diagram (Szelogowski, 2023)