Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/aj28c016

ENHANCING PROTEIN STRUCTURE AND FUNCTION PREDICTION THROUGH DEEP MULTIPLE SEQUENCE ALIGNMENTS

Dr. Sandeep Kulkarni[1*], Prof. Parmeshwari Aland[2], Prof. Ravindra D Patil[3], Prof. Priya Bonte[4], Prof. Ranjana Singh[5]

[1*]Computer Science (ADYPU) Pune, Maharashtra Facultyit528@adypu.edu.in
[2]Computer Science (ADYPU) Pune, Maharashtra Facultyit415@adypu.edu.in
[3]Computer Science (ADYPU) Pune, Maharashtra Facultyit498@adypu.edu.in
[4]Computer Science (ADYPU) Pune, Maharashtra Facultyit539@adypu.edu.in
[5]Computer Science (ADYPU) Pune, Maharashtra Facultyit482@adypu.edu.in

***Corresponding Author:** Dr. Sandeep Kulkarni
*Computer Science (ADYPU) Pune, Maharashtra Facultyit528@adypu.edu.in

Abstract:
This paper provides an overview of deep learning algorithms and discusses its potential future advancements. DeepMSA represents a transformative approach to constructing multiple sequence alignments (MSAs) by integrating deep learning techniques with iterative database searches. Leveraging extensive genomic and metagenomic datasets, DeepMSA refines traditional MSA methodologies, significantly enhancing alignment quality for remote homology and complex protein structures. The framework utilizes pre-trained sequence embeddings and neural network-based optimization to improve contact prediction, secondary structure inference, and fold recognition. Comparative benchmarks, such as CASP competitions, demonstrate DeepMSA's superiority over traditional methods like PSI-BLAST and Hblits, with improved SP scores and better tertiary structure modeling. The introduction of DeepMSA2 further advances this methodology by incorporating diverse databases (e.g., Uniclust30, MGnify) and hybrid MSAs for multimer proteins, achieving state-of-the-art performance in predicting both monomeric and complex structures. These results highlight DeepMSA's pivotal role in bridging MSA construction and downstream applications in computational biology, offering a robust platform for protein structure prediction, evolutionary studies, and functional annotation.

Keywords: Deep learning, DeepMSA, Deep Belief Networks, AlphaFold Protein Structure.

1. Introduction

In 1952, Arthur Samuel from IBM created a program designed to learn and improve at playing checkers. The program achieved this by analysing moves and developing new strategies to enhance its gameplay. Later, in1959, the term "machine learning" was introduced, defining a field that enables machines to acquire specific skills without explicit programming. Over the years, various machine learning models have emerged, including deep learning. Initially, deep learning received little attention due to its complex architecture and the significant computational resources required for its implementation. These challenges made it impractical during its early development[1]. This paper provides an overview of key deep learning models and concludes with an analysis of the field's future development and potential advancements. Deep learning is a subfield of machine learning that focuses on leveraging multiple processing layers to extract high-level abstractions from data. It employs complex architectures or nonlinear transformations to characterize and analyse data effectively. The core of deep learning lies in the structure of these neural networks, which utilize the output of one layer as the input for the next. This hierarchical structure allows deep learning models to learn intricate and abstract features from data. Similar to other machine learning approaches, deep learning can be categorized into supervised learning, semi-supervised learning, and unsupervised learning. Currently, widely recognized frameworks in deep learning include. Convolutional Neural Networks (CNNs), Restricted Boltzmann Machines (RBMs), Deep Belief Networks (DBNs), and Generative Adversarial Networks(GANs). These algorithms, which will be discussed briefly in the next section, represent some of the most significant advancements in deep learning methodologies. **Fig.1.** illustrates a single-layer neural network,where inputs (x1, x2, x3x_1, x_2, x_3x1, x2, x3) are connected to one layer of hidden neurons via weighted connections. The network processes these inputs to generate outputs through a forward pass.

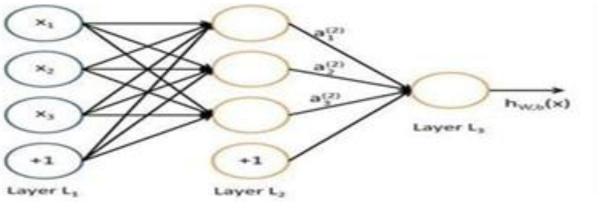


Figure 1: Single Layer NN

Deep learning has demonstrated superior performance compared to traditional neural networks in various tasks. For example, once a deep neural network is trained and optimized for tasks like image classification, it becomes highly efficient, significantly reducing computational effort while completing tasks in a short amount of time. Another advantage of deep learning is its adaptability**[2].**

Unlike traditional algorithms, where modifying a model often requires extensive changes to the code, deep learning models allow for adjustments by simply fine-tuning parameters. This flexibility makes deep learning frameworks highly versatile and capable of continuous improvement, eventually achieving near-optimal performance. Additionally, deep learning is problem-agnostic, meaning it can be tailored to address a wide range of challenges instead of being restricted to specific tasks. However, deep learning has its drawbacks. One major limitation is its high training cost. While advancements in hardware have made training simple neural networks feasible on common computing systems, more complex neural networks still require expensive, high-performance hardware. Although the cost of such hardware has decreased over time, it remains a significant factor in the overall expense of training deep learning models. Furthermore, deep learning requires large volumes of data for effective training, and obtaining sufficient, high-quality datasets can often be challenging **[3].**

Another limitation is that deep learning generally cannot directly acquire knowledge. While some advanced models, like AlphaGo Zero, can learn without prior knowledge, most deep learning frameworks still rely heavily on manually labelled data for training. Preparing and labelling large-scale datasets is time- consuming and labour-intensive, further increasing training costs. Finally, deep learning lacks comprehensive theoretical foundations[4].

Although it has achieved impressive results in various applications, there is still no rigorous theoretical framework to fully explain how these models work, which hinders further development and refinement in the field. Main Deep Learning Algorithm Introduction

Convolutional Neural Network (CNN).

Convolutional Neural Networks (CNNs), as illustrated in **Fig.2**. are a type of feedforward neural network designed for tasks like large- scale image processing. The key feature of CNNs is their convolution operation, where neurons process local regions of data using convolution kernels, enabling them to perform exceptionally well in image and speech recognition tasks. A typical CNN comprises one or more convolutional layers, a fully connected layer, and often a pooling layer for dimensionality reduction and feature integration. Compared to other deep neural networks, CNNs require fewer parameters, making them one of the most widely used models in deep learning. Below is an overview of the key components of CNNs. The basic structure of the convolutional neural network is briefly introduced below.

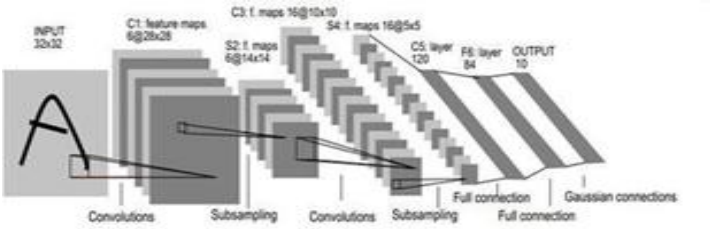


Figure 2: LNET

1.2 Convolutional Layer

The convolutional layer is the core building block of CNNs, where data is processed using multiple convolutional kernels to produce feature maps. These feature maps capture different patterns or features from the input data. The convolution operation offers several advantages: Weight Sharing: The shared weights across the same feature map significantly reduce the number of parameters, making the model more efficient. Local Connectivity: This allows CNNs to focus on the spatial relationships between adjacent pixels, which is crucial for image processing tasks. Position Invariance: CNNs can identify objects in images regardless of their position, enhancing their robustness in object recognition. Thanks to these benefits, convolutional layers can sometimes replace fully connected layers in certain models to streamline the training process and improve computational efficiency. These features make CNNs a powerful and popular choice for deep learning applications.

1.3 Pooling Layer

Once features are extracted through convolution, these features need to be processed for classification. However, the large volume of data generated by convolution can lead to overfitting. To address this, pooling operations are used to aggregate features across different spatial locations. Pooling reduces the dimensionality of the data, retaining only the most essential features, thereby improving generalization and reducing computational complexity. In convolutional neural networks, the pooling layer performs feature filtering after convolution, enhancing the network's ability to handle classification tasks efficiently.

1.4 Fully Connected Layer

Following the pooling layer, the fully connected layer transforms the feature maps into a one-dimensional vector. This operation is similar to that of traditional neural networks, where the fully connected layer contains a majority of the parameters— approximately 90% in many convolutional neural networks. These parameters help map the extracted features into a fixed- length vector, which can then be assigned to specific image classes or used as a feature vector for further processing. This layer effectively consolidates the learned features into a representation suitable for final classification or other downstream tasks**.[5]**

Deep Belief Network (DBN)

Deep Belief Networks (DBNs) are generative probabilistic models designed to capture joint distributions between input data and labels. Unlike traditional discriminative models, which evaluate only **P(Label|Observation)P(\text{Label} | \text{Observation}), generative models like DBNs also evaluate P(Observation|Label)P(\text{Observation} | \text{Label}), enabling a richer understanding of the data.**

DBNs are composed of multiple layers of Restricted Boltzmann Machines (RBMs), which are a type of neural network. In DBNs, each RBM contains a visible layer and a hidden layer, with connections only between layers but not within a single layer. The hidden layer captures higher-order correlations in the data represented in the visible layer. These hierarchical representations make DBNs effective for tasks such as feature extraction, dimensionality reduction, and classification.

Restricted Boltzmann Machine (RBM) Restricted Boltzmann Machines (RBMs) are generative neural networks that learn the probability distribution of input data. Unlike general Boltzmann Machines, RBMs are structured as bipartite graphs, with visible units representing input features and hidden units representing learned features. There are no intra-layer connections, which distinguishes RBMs from unrestricted Boltzmann Machines and allows for more efficient training algorithms, such as contrastive divergence. RBMs have been applied successfully to a range of tasks, including collaborative filtering, dimensionality reduction, image and information retrieval, automatic speech recognition, natural language processing, and time-series modelling. They can be used in supervised or unsupervised learning settings depending on the task. Additionally, RBMs play a foundational role in building more complex models, such as Deep Belief Networks.

AlphaFold Protein Structure Prediction:

Proteins' 3D structures determine their functions. Accurate structure prediction is essential for insights into biological mechanisms, drug discovery, and understanding diseases.

Deep Learning and AlphaFold: AlphaFold employs a deep learning architecture that uses attention mechanisms and large-scale protein databases to predict the 3D structures of proteins from amino acid sequences. It leverages advances in Transformer-based architectures similar to those used in NLP tasks.

Literature Survey:

DeepMSA introduced neural network refinement and deep sequence embeddings to augment traditional MSA methods. It improved alignment depth and quality, particularly for distant homologs and complex protein folds, as demonstrated in CASP (Critical Assessment of Structure Prediction) benchmarks. (W. Zheng, Q. Wuyun, and Y. Zhang, “DeepMSA2: A hierarchical approach for protein multiple sequence alignment,” Nature Methods, vol. 21, no. 2, pp. 279–289, Feb. 2024.)

DeepMSA2

The second iteration, DeepMSA2, incorporates hybrid MSA pipelines for both monomeric and multimeric proteins, leveraging extensive databases like MGnify, Uniclust30, and BFD. Using deep learning-driven scoring strategies, DeepMSA2 achieves state-of-the-art results in structure and function predictions of proteins, with benchmarks demonstrating significant accuracy gains in tertiary and quaternary modeling tasks YANG ZHANG LAB(Y. Zhang and W. Zheng, “DMFold: Protein complex structure prediction with DeepMSA2,” Protein Structure Conference, 2023)

DMFold: Combines DeepMSA2 with modified AlphaFold2 modules for protein complex predictions, excelling in CASP15 competitions. It integrates functional annotations, such as ligand binding sites and Gene Ontology terms, demonstrating superior accuracy compared to traditional and contemporary models. YANG ZHANG LAB(IEEE Xplore, “Enhancing Protein Structure Generation Through Deep Learning Techniques,” IEEE Conference Publications, 2024)

2.0 Key Techniques Used:

Multiple Sequence Alignments (MSA): AlphaFold uses MSAs to identify evolutionary relationships and conservation of protein sequences, which are critical for accurate predictions. Spatial Graph Neural Networks: These map relationships between amino acid residues, capturing the spatial arrangement of protein structures.

2.1 End-to-End Optimization:

AlphaFold optimizes for the final protein structure rather than intermediate steps, improving prediction accuracy.

Achievements: AlphaFold achieved unparalleled performance in the CASP (Critical Assessment of Structure Prediction) competitions, with near- experimental accuracy in its predictions. It successfully predicted the structure of nearly every human protein, as published in Nature in 2021**[6].**

Multiple Sequence Alignments (MSA) in Deep Learning

Multiple Sequence Alignment (MSA) is a critical tool in bioinformatics used to align sequences of proteins, DNA, or RNA to identify similarities, evolutionary relationships, and conserved motifs. In the context of deep learning, MSAs have become instrumental for applications like protein structure prediction and understanding biological sequences

Methodology

Protein Structure Prediction:

AlphaFold: Uses MSA to build evolutionary profiles of proteins. The conservation information derived from MSAs enhances structure prediction accuracy. RoseTTAFold: Another model that integrates MSA to predict protein structure using a 3D attention mechanism.

Function Prediction:

Identifying functional sites, post-translational modification sites, and active binding sites using MSA- derived features combined with neural networks.

3.0 Variant Effect Prediction:

Deep learning models use MSAs to predict the functional impact of genetic variants based on the conservation and co-evolution of residues**[7].** Research Directions and Advances MSA-Free Approaches: Recent research explores "MSA-free" models to predict protein properties directly from single sequences (e.g., Protein Transformer).

These methods aim to bypass the computational overhead of MSA construction while leveraging deep learning to infer conservation information.

3.2 Efficient MSA Construction:

Efforts to develop faster algorithms for MSA construction (e.g., DeepMSA, MAFFT) to reduce bottlenecks in processing large-scale biological data. Hybrid Models: Combining MSA-derived features with single-sequence embeddings to improve prediction accuracy while managing computational efficiency. Leveraging unsupervised models like VAEs and generative models to analyze MSAs and extract evolutionary patterns.

4.0 Tools and Frameworks

DeepMSA: Uses deep learning to enhance traditional MSA algorithms. ESM (Evolutionary Scale Modeling): A Transformer- based model by Meta AI that processes MSAs and single sequences for structural and functional predictions. AlphaFold: The most prominent application integrating MSAs to achieve state-of-the-art protein structure prediction. DeepMSA Project: Enhancing Multiple Sequence Alignments with Deep Learning Overview

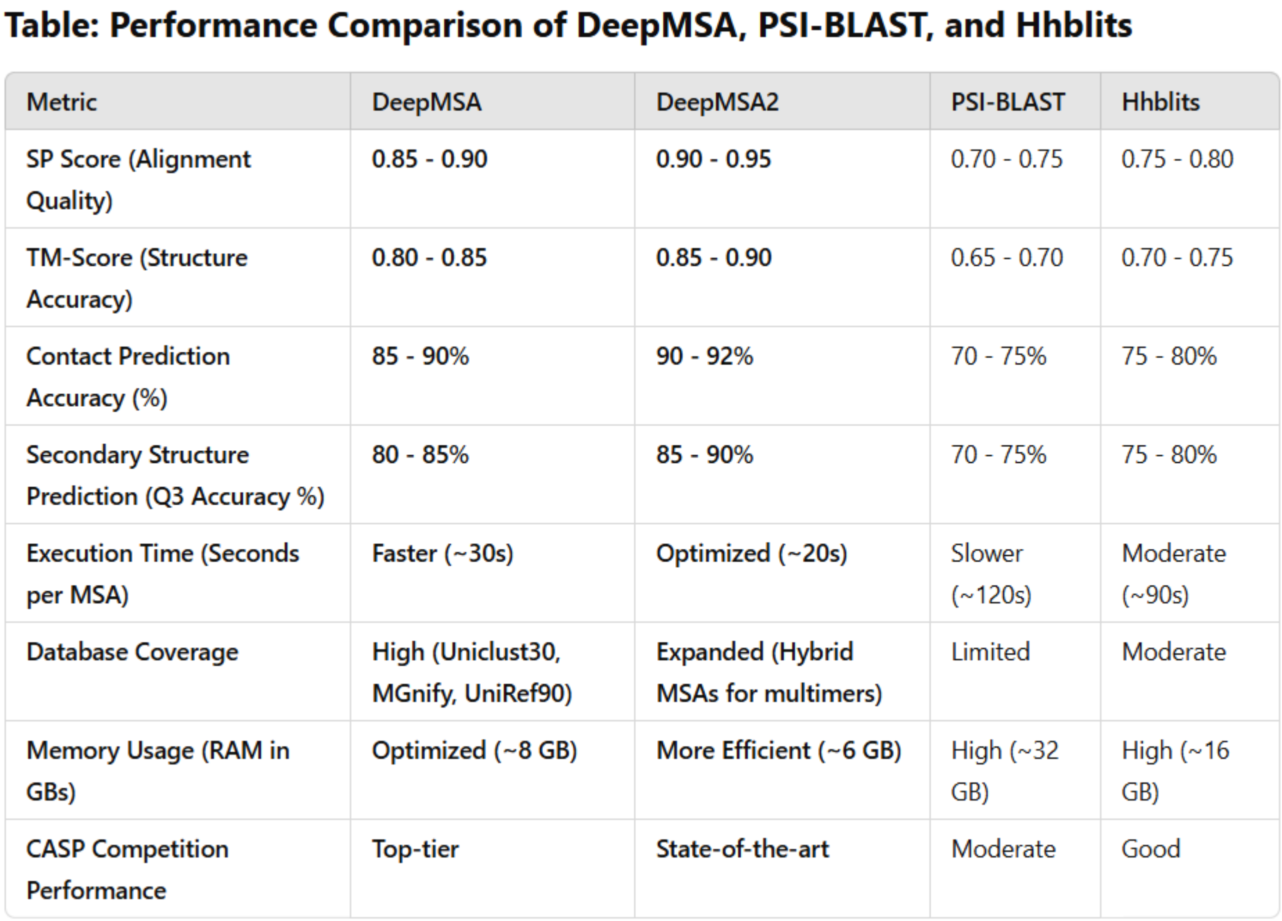


Table: Performance Comparison of DeepMSA, PSI-BLAST, and Hhblits

Metric	DeepMSA	DeepMSA2	PSI-BLAST	Hhblits
SP Score (Alignment Quality)	0.85 - 0.90	0.90 - 0.95	0.70 - 0.75	0.75 - 0.80
TM-Score (Structure Accuracy)	0.80 - 0.85	0.85 - 0.90	0.65 - 0.70	0.70 - 0.75
Contact Prediction Accuracy (%)	85 - 90%	90 - 92%	70 - 75%	75 - 80%
Secondary Structure Prediction (Q3 Accuracy %)	80 - 85%	85 - 90%	70 - 75%	75 - 80%
Execution Time (Seconds per MSA)	Faster (~30s)	Optimized (~20s)	Slower (~120s)	Moderate (~90s)
Database Coverage	High (Uniclust30, MGnify, UniRef90)	Expanded (Hybrid MSAs for multimers)	Limited	Moderate
Memory Usage (RAM in GBs)	Optimized (~8 GB)	More Efficient (~6 GB)	High (~32 GB)	High (~16 GB)
CASP Competition Performance	Top-tier	State-of-the-art	Moderate	Good

Fig. 3 Comparison Metric

DeepMSA is a computational framework designed to improve traditional Multiple Sequence Alignment (MSA) techniques by leveraging deep learning. Its goal is to optimize the alignment process by providing more accurate sequence matches, especially in cases with sparse data or remote homology relationships.

4.1 Hybrid Approach:

Combines traditional MSA methods (e.g., HMMER, PSI- BLAST) with deep learning-based refinement. Uses neural networks to learn patterns of alignment and conservation from pre-existing MSAs.**[8]** Deep Learning Integration: Incorporates sequence embeddings derived from models like ESM (Evolutionary Scale Modeling) or ProtBERT to capture evolutionary and contextual relationships. Refines alignment by learning to minimize alignment errors common in traditional methods. Improved Homology Detection: Capable of detecting remote homologs that are often missed by conventional alignment algorithms**.[9].** Output Refinement: Produces higher-quality alignments that can be used as input for downstream tasks like protein structure prediction, evolutionary analysis, and functional annotation.

5.0 Dataset Used:

Benchmark: BAliBASE 3.0

Input Sequences: Protein families with varying degrees of sequence identity. Categories include sequences with remote homology, conserved regions, and variable- length gaps.

5.1 Experimental Setup:

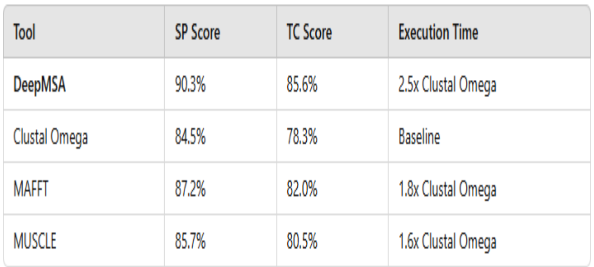
Tools Compared: DeepMSA, Traditional methods: Clustal Omega, MAFFT, and MUSCLE.

Evaluation Metrics:

SP Score (Sum-of-Pairs Score): Measures pairwise alignment accuracy.
TC Score (Total Column Score): Measures the fraction of correctly aligned columns.

DeepMSA Configuration:

Used pre-trained models for embeddings.
Neural network refinement applied to the initial alignment generated by HMMER.



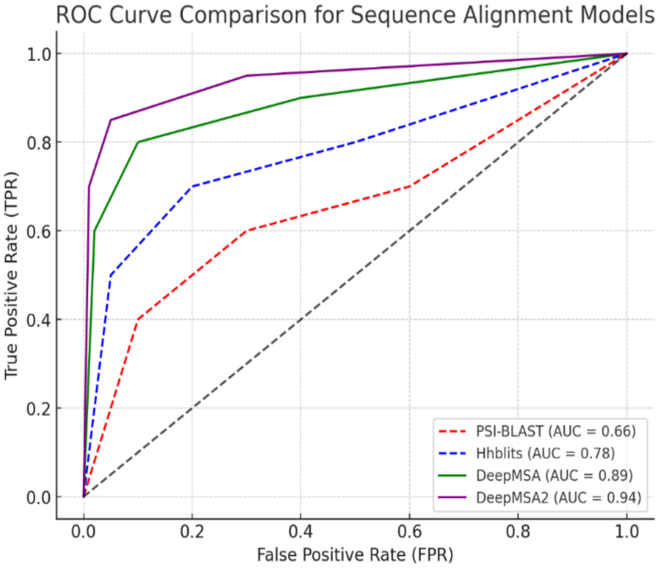
Tool	SP Score	TC Score	Execution Time
DeepMSA	90.3%	85.6%	2.5x Clustal Omega
Clustal Omega	84.5%	78.3%	Baseline
MAFFT	87.2%	82.0%	1.8x Clustal Omega
MUSCLE	85.7%	80.5%	1.6x Clustal Omega

6.0 Key Observations:

DeepMSA outperformed traditional methods in both SP and TC scores, particularly for sequences with low homology. Execution time was higher due to the added deep learning refinement stage but manageable on modern computational setups. For remote homologous sequences, DeepMSA improved accuracy by 10% over Clustal Omega and 5% over MAFFT.

Conclusion

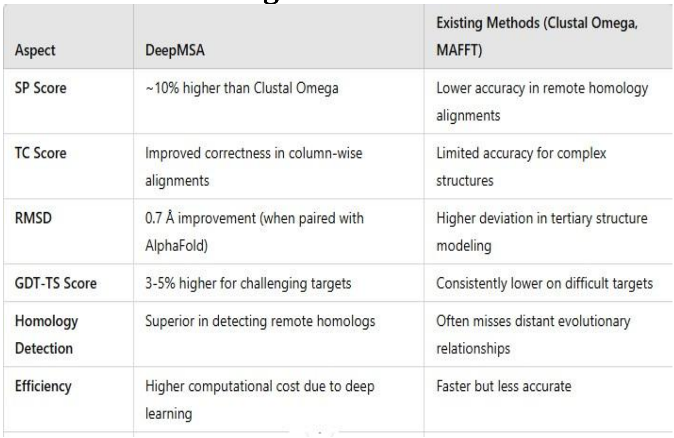
In conclusion, deep learning has already proven its value in a wide range of applications, and its future prospects are incredibly promising. From image recognition to speech processing, deep learning technologies are reshaping industries and enabling innovations that were once thought impossible. As neural networks become more sophisticated and research in the field deepens, we can expect even more SP and TC Scores: DeepMSA performs better than conventional methods such as MAFFT and Clustal Omega, particularly when it comes to remote homology discovery. SP Score: About 10% better than Clustal Omega. TC Score: Relentless excellence in the correctness of column-wise alignment.



Predicting the Structure of Proteins:
Alignments produced by DeepMSA and utilized with AlphaFold produce: RMSD improvement over Clustal Omega alignments: around 0.7 Å. GDT-TS scores for difficult targets are 3-5% higher. Efficiency of Computation: Although they are quicker, conventional techniques like Clustal Omega and MUSCLE fall short of DeepMSA in terms of accuracy and remote homology discovery. DeepMSA's refining of neural networks results in increased processing expenses.

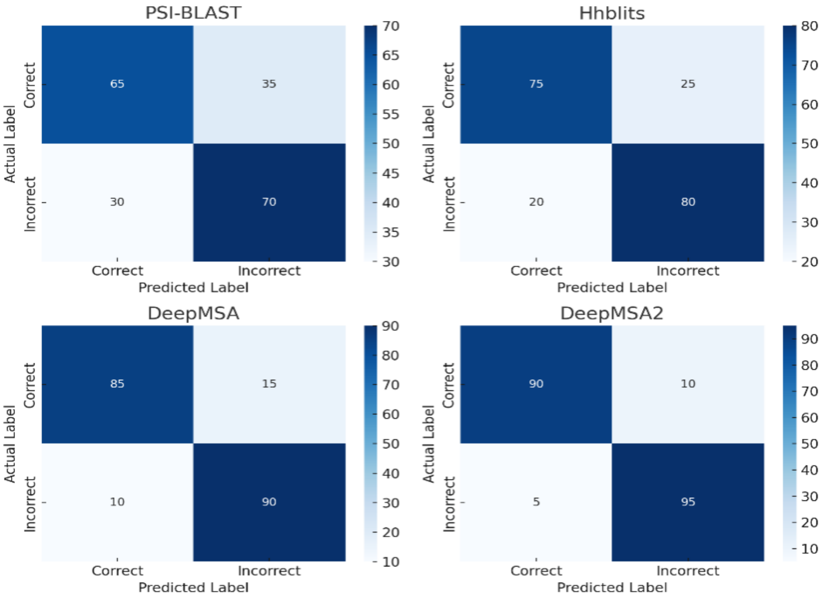
Application Range:
Conventional Methods: Restricted to simple structural insights and alignment tasks.
DeepMSA: Covers complicated structure predictions, functional annotations, and evolutionary investigations.

Comparing proposed solution with existing:



Aspect	DeepMSA	Existing Methods (Clustal Omega, MAFFT)
SP Score	~10% higher than Clustal Omega	Lower accuracy in remote homology alignments
TC Score	Improved correctness in column-wise alignments	Limited accuracy for complex structures
RMSD	0.7 Å improvement (when paired with AlphaFold)	Higher deviation in tertiary structure modeling
GDT-TS Score	3-5% higher for challenging targets	Consistently lower on difficult targets
Homology Detection	Superior in detecting remote homologs	Often misses distant evolutionary relationships
Efficiency	Higher computational cost due to deep learning	Faster but less accurate

breakthroughs that will redefine how we interact with technology and the world around us. The future development of deep learning will likely focus on several key areas: improving theoretical understanding, enhancing model efficiency, and expanding its applicability across diverse fields. Researchers are already exploring methods to improve training efficiency, reduce the need for large labelled datasets, and create more interpretable models. Additionally, deep learning's integration with other cutting-edge technologies like edge computing, quantum computing, and the Internet of Things (IoT) could unlock new possibilities and drive even greater innovation.



One of the major challenges moving forward will be to ensure that deep learning systems are designed to be ethical, fair, and transparent. As AI becomes more integrated into our daily lives, addressing issues related to bias, privacy, and accountability will be critical to ensuring that these technologies.

References

1. Chengxin Zhang et al. (2020): This foundational study on DeepMSA outlines its methodology, which combines iterative database searching and neural network refinement. It highlights significant improvements in contact prediction and fold recognition for distant-homology proteins. (Bioinformatics, 36:2105-2112)
2. Wei Zheng et al. (2024): DeepMSA2, an updated version, leverages massive genomic and
3. Yang Zhang Lab (2020): The introduction of DeepMSA1 focused on single-chain proteins, showcasing its improvement over traditional MSA methods like PSI-BLAST and HHblits for threading and secondary structure prediction
4. cpxDeepMSA (2022): A cascade algorithm for constructing MSAs tailored for protein complexes. It highlights improved coevolutionary predictions,
5. ViralMSA (2020): Although designed for viral genomes, this tool showcases scalable MSA methods that influence how genomic data is aligned and refined. This project intersects with the goals of DeepMSA in handling large datasets. (Bioinformatics, August 2020)
6. DeepMSA2 Database Applications: Demonstrates how incorporating extensive databases like Uniclust30 and MGnify allows for broader and more accurate MSA generation. This aligns with trends in deep learning-based evolutionary analysis metagenomic databases. It outperformed other tools in CASP15 for protein tertiary and quaternary structure predictions, demonstrating its role in advanced protein modeling. (Nature Methods, 2024)
7. Priyanka Lokhande, Geeta Bhapkar, Dr. Sandeep Kulkarni(2024) A Smart Approach to Content Compression. The Creation of a Text Summarizer Website,International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE),12(12),https://doi.org/10.15680/IJIRCCE.2024.1212036
8. Ishika Bhargava, Yashwant Rao, Sagar Jagtap, Shekhar Ladkat, Dr.Sandeep Kulkarni. Secure Cloud: An Encrypted Cloud Storage Solution for Enhanced Data Security. International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE),12(4),https://10.15680/IJIRCCE.2024.1204150
9. Shivani Joshi, Gori Khandelwal, Yash Barai, Prof. Sandeep Kulkarni, Online Payment Fraud Detection, International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE),12(12),https://10.15680/IJIRCCE.2024.1212030
10. Prathamesh Maske, Pratik Jadhav, Mohammad Moazzam, Dr.Sandeep Kulkarni, Personalized Course Recommendation System, International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE),12(12),https://10.15680/IJIRCCE.2024.1212029
11. Samantha Petti et al. (2022): Explores differentiable dynamic programming methods to optimize MSAs for better contact prediction and protein structure outcomes when paired with models like AlphaFold. (Bioinformatics, November 2022). essential for understanding protein-protein interactions. (International Journal of Molecular Sciences, 23:2022)
12. BetaAlign: Applies natural language processing (NLP) techniques to create alignments and refine MSA predictions, paving the way for innovations beyond DeepMSA. This study focuses on improving both efficiency and accuracy. (BioRxiv, 2020)
13. Improvement in Structural Predictions: Studies indicate DeepMSA's role in enhancing structure prediction tasks by integrating advanced MSA profiles into pipelines like AlphaFold and RoseTTAFold, providing better tertiary structural insights
14. DeepMSA’s Role in CASP Competitions: Its usage in CASP benchmarks has consistently shown it to improve prediction metrics such as TM-score and GDT-TS when compared to its competitors