成都理工大學中英合作办学
CDUT Sino-British Collaborative Education

OXFORD BROOKES UNIVERSITY

UNDERGRADUATE PROJECT REPORT

Project Title:	**Protein Structure Prediction using Attention-ProteinMeNet**
Surname:	**Jia**
First Name:	**Xinyue**
Student Number:	**202118010206**
Supervisor Name:	**Dr Grace Ugochi Nneji**
Module Code:	**CHC 6096**
Module Name:	**Project**
Date Submitted:	**May 6, 2024**

Declaration

Student Conduct Regulations:

Please ensure you are familiar with the regulations in relation to Academic Integrity. The University takes this issue very seriously and students have been expelled or had their degrees withheld for cheating in assessment. It is important that students having difficulties with their work should seek help from their tutors rather than be tempted to use unfair means to gain marks. Students should not risk losing their degree and undermining all the work they have done towards it. You are expected to have familiarised yourself with these regulations.

https://www.brookes.ac.uk/regulations/current/appeals-complaints-and-conduct/c1-1/

Guidance on the correct use of references can be found on www.brookes.ac.uk/services/library, and also in a handout in the Library.

The full regulations may be accessed on-line at https://www.brookes.ac.uk/students/sirt/student-conduct/ . If you do not understand what any of these terms mean, you should ask your Project Supervisor to clarify them for you.

I declare that I have read and understood Regulations C1.1.4 of the Regulations governing Academic Misconduct, and that the work I submit is fully in accordance with them.

Signature: JiaXinyue. Data: 06/05/2025

REGULATIONS GOVERNING THE DEPOSIT AND USE OF OXFORD BROOKES UNIVERSITY MODULAR PROGRAMME PROJECTS AND DISSERTATIONS

Copies of projects/dissertations, submitted in fulfilment of Modular Programme requirements and achieving marks of 60% or above, shall normally be kept by the Library.

| agree that this dissertation may be available for reading and photocopying in accordance with the Regulations governing use of the Library.

Signature: JiaXinyue. Data: 06/05/2025

Acknowledgment

I would really like to specific my heartfelt gratitude to all the professors who have supported and guided me at some point of my instructional adventure. My deepest appreciation goes to my supervisor, Dr. Grace Ugochi Nneji, for her invaluable guidance, encouragement, and unwavering support throughout the course of this project. Her expertise and insights have been instrumental in shaping the direction and success of my research.

I am additionally immensely thankful to Dr. Jojo, our route teacher, whose willpower and exuberance in coaching performed a key function in helping me navigate the complexities of this venture. Her recommendation and encouragement had been crucial to my development.

My honest thank you go to my fellow colleagues inside the same supervisor institution. Their considerate feedback, collaboration, and shared passion for studies have made this adventure intellectually enriching and personally rewarding. I am also grateful to my classmates, whose discussions and cooperative spirit have fostered a colourful and motivating learning environment. Special thanks go to my roommates, whose patience, expertise, and emotional assist helped me live balanced and focused throughout worrying times.

I could additionally like to increase my deepest gratitude to my mother and father and circle of relatives contributors. Their unwavering love, encouragement, and belief in me were my most powerful basis. Without their persisted assist, this adventure might now not were feasible.

Finally, I would really like to thank Oxford Brookes University for imparting a nurturing educational environment, first-rate resources, and a network that has stimulated my increase as each a scholar and a researcher.

Table of Contents

List of Figures

List of Tables

Abstract

Protein -secondary structure is an important task in prediction BIO -information science, understanding protein function, guide drug design and support disease research. Traditional experimental techniques such as X -ray crystalography and nuclear resonance spectroscopy are often forced by the requirements for the preparation of labor -intensive, expensive and complex samples. At the same time, traditional calculation methods often struggle to keep the complex sequence addiction and characteristic of structural diversity of proteins. Despite the recent progress in deep education, the current models still face challenges such as inadequate generalization, poor long -distance handling and limited interpretation. In order to remove these challenges, this study develops a new deep learning framework called the Meditation Prosecutor, developed to improve the accuracy and strength of the prediction of the protein-secondary structure. The model is trained and evaluated on two widely used benchmark datasets: RCSB-PDB and CB513, which provide high-quality annotated sequences for both three-state and eight-state secondary structure classification.

Attention-ProteinMeNet combines three components: ProteinNet for local feature extraction, Bidirectional Long Short-Term Memory (BLSTM) for modeling sequence-level dependencies, and an attention mechanism to selectively focus on key residues that influence structural outcomes. The model demonstrates strong generalization across datasets, achieving validation accuracies of 96.49% on RCSB-PDB and 94.15% on CB513, with corresponding high scores on ROC-AUC and F1 metrics.

Explainable artificial intelligence techniques, such as SHAP analysis, were employed to interpret the contribution of individual residues, thereby validating the biological relevance of the model's predictions. In addition, a user-friendly graphical interface was developed to support both single-sequence and batch predictions. Collectively, this work presents a robust, interpretable, and efficient solution for protein structure prediction, offering a valuable resource for the computational biology community.

Keywords: *Protein Structure Prediction, Deep Learning, Attention ProteinMeNet, Attention Mechanism, Bidirectional LSTM (BLSTM), ProteinNet, Secondary Structure*

Abbreviations

AI – Artificial Intelligence: The development of systems that perform tasks typically requiring human intelligence.

CNN – Convolutional Neural Network: A deep learning architecture commonly used for spatial and sequential data processing.

BLSTM – Bidirectional Long Short-Term Memory: A type of recurrent neural network that captures dependencies in both forward and backward directions.

ProteinNet – A model component that utilizes one-dimensional convolutional layers to extract local features from protein sequences.

ProteinMeNet – A hybrid deep learning model that combines ProteinNet and BLSTM for protein structure prediction.

Attention-ProteinMeNet – An extended version of ProteinMeNet that incorporates an attention mechanism to selectively focus on the most informative residues.

PDB – Protein Data Bank: A global repository of experimentally determined protein and nucleic acid structures.

RCSB-PDB – Research Collaboratory for Structural Bioinformatics Protein Data Bank.

CB513 – Cuff and Barton 513.

SST3 – Simplified Secondary Structure Three-class system: Classification into Helix, Strand, and Coil.

SST8 – Secondary Structure Eight-class system: A finer-grained classification of protein structures including alpha helix, beta strand, turn, etc.

ROC – Receiver Operating Characteristic: A graphical representation of a classifier's performance across different thresholds.

AUC – A calculation that determines a classifier's general ability to distinguish between classes.

F1-score – Harmonic Mean of Precision and Recall: A calculation that balances the trade between accurate and recall.

Precision – The ratio of true positive predictions to the total number of predicted positives.

Recall – The ratio of true positive predictions to the total number of actual positives.

TP – True Positive: The number of actual positive instances correctly predicted as positive.

TN – True Negative: The number of actual negative instances correctly predicted as negative.

FP – False Positive: The number of actual negative instances incorrectly predicted as positive.

FN – False Negative: The number of actual positive instances incorrectly predicted as negative.

SHAP – Shapley Additive Explanation: A method of explaining the model output by assigning function score for functions.

ReLU – Improved linear unit: An activation feature that gives the input for zero and positive values for negative inputs.

GPU – Graphics Processing Unit: A special processor adapted to parallel components with high decay, especially in deep learning.

GUI – Graphic user interface: A visual interface that enables user interaction with software through graphic elements.

XAI – Explanation of artificial intelligence: Technology and equipment whose goals can be understood for the behavior and decisions of the AI model for humans.

API – Application Programming Interface: A set of tools and protocols that allow different software components to communicate.

Glossary

Attention Mechanism: A deep learning technique used to focus the model's attention on specific parts of the input sequence that are more important for making predictions, improving the model's performance in capturing long-range dependencies and complex relationships.

Alpha-Helical: A common secondary structure found in proteins, characterized by a right-handed spiral or helix, formed by hydrogen bonds between peptide bonds.

Beta-Sheet: A secondary protein structure composed of beta strands linked by hydrogen bonds, often forming a pleated sheet-like structure.

BLSTM (Bidirectional Long Short-Term Memory): A type of recurrent neural network (RNN) that processes input sequences in both forward and backward directions, improving the model's ability to capture dependencies over time.

CB513 Dataset: A widely used dataset in protein structure prediction research, consisting of 513 proteins with known secondary structures, used for training and testing machine learning models.

Confusion Matrix: A performance measurement tool used for classification models, showing the number of true positives, false positives, true negatives, and false negatives. It helps evaluate the accuracy of the model.

F1 Score: A performance metric that combines precision and recall into a single value, representing the harmonic mean of precision and recall. It is particularly useful for imbalanced datasets.

ProteinNet: A protein sequence-based deep learning model for predicting protein secondary structures, which serves as a baseline for comparison with more advanced models like Attention-ProteinMeNet.

RCSB-PDB Dataset: The Research Collaboratory for Structural Bioinformatics Protein Data Bank, a repository of 3D structures of large biological molecules, including proteins, nucleic acids, and complex assemblies. It provides data for training and evaluating protein structure prediction models.

ROC Curve (Receiver Operating Characteristic Curve): A graphical representation used to evaluate the performance of a classification model by plotting the true positive rate against the false positive rate at various thresholds.

SHAP (Shapley Additive Explanations): A method for explaining machine learning model predictions by attributing the contribution of each feature to the final prediction. It helps interpret the model's decision-making process in a more understandable way.

Validation Accuracy: The accuracy of a model on a validation dataset, used to assess how well the model generalizes to unseen data.

Training Accuracy: The accuracy of a model during its training phase, indicating how well it has learned from the training data.

Chapter 1 Introduction

Protein structure prediction is crucial in bioinformatics for understanding protein functions and speeding up drug discovery. However, traditional methods often face limitations in accuracy, speed, and scalability. This project proposes a deep learning model called Attention-ProteinMeNet, combining ProteinNet and BLSTM with an attention mechanism to enhance prediction performance. Trained on RCSB-PDB and CB513 datasets, the model aims to improve accuracy, generalization, and interpretability. This chapter overviews the research background, objectives, goals, scope, and content of the project.

1.1 Background

Proteins are the fundamental molecules that perform a myriad of essential functions within living organisms. Their structures, which are intricately linked to their amino acid sequences, are crucial for understanding their functions and for various applications such as drug development, protein engineering, and disease research [1]. Predicting the three-dimensional structure of proteins from their sequences is a complex problem that has been widely studied, particularly under the HP simplified model with different types of lattices as conformational representations[1]. The traditional methods for protein structure prediction, such as homology modeling and threading, are limited when the target protein sequence does not have a close homolog with a known structure[2]. This limitation has driven the exploration of deep learning techniques, which can automatically extract features from input data and capture nonlinear correlations between inputs and outputs[3] .

This project aims to develop a Attention-ProteinMeNet model to predict protein structure using TensorFlow. This model aims to exploit the strengths of each component to improve the accuracy of protein structure prediction. This project will contribute to ongoing efforts in the field of protein structure prediction and has the potential to impact a variety of applications in bioinformatics and computational biology. This section will describe the risks and factors of the project, its challenges, project objectives, subjects, scope, and audience.

1.1.1 Risks and Factor

- Limited drug discovery and development

Protein structure is essential for drug target identification and rational drug design. If protein function is not understood accurately and without accurate structural insights, drug development will become very inefficient, leading to increased costs and longer time to develop new

therapies[4]. The absence of structural information may also lea d to ineffective or nonspecific drug candidates, reducing the success rate of clinical trials[5].

Figure 1. Protein Structure Prediction for Drug Discovery

- High Cost and Time-Consumption

Prior experimental techniques, such as X-ray crystallography and NMR spectroscopy, often require expensive laboratory equipment and significant time to collect and process data. This makes the study of protein structure limited and costly.

- Incomplete and Low-Resolution Data

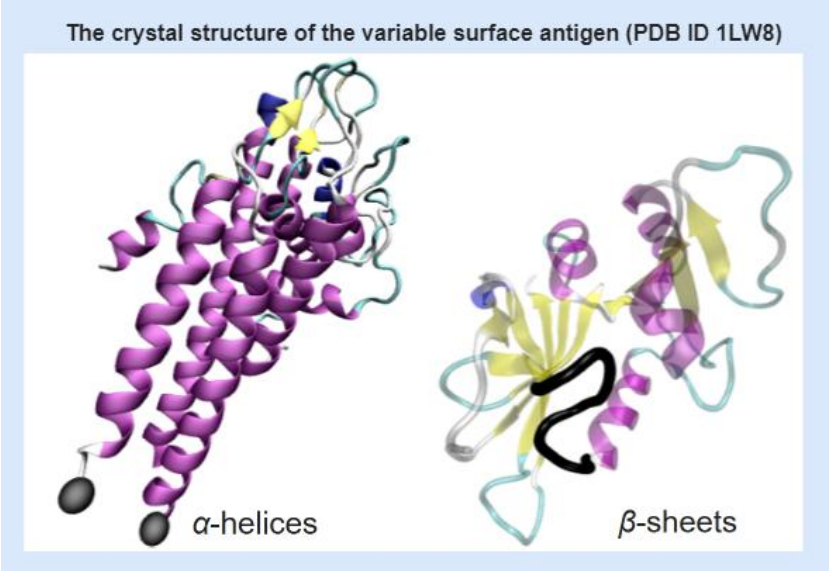

Figure 2. The Variability of Protein Structure

Many protein structures remain unresolved due to difficulties in sample preparation and data acquisition. Even if the structure is parsed out, there may be low resolution, missing regions, or

inaccuracies due to experimental conditions that differ from physiological environments. Even if structures are parsed, they may suffer from low resolution, missing regions such as flexible loops, or inaccuracies caused by experimental conditions different from the physiological environment[6]. This can lead to unreliable structural models that may mislead downstream applications for drug design and functional annotation. Figure 2 illustrates the different secondary structural elements of proteins, loop segments in protein chains tend to be highly mobile, even in generally stable proteins.

- Inaccuracy of Computational Methods

Figure 3 illustrates the flow of computational methods for protein structure prediction. Computational methods such as homology modeling rely on the availability of template proteins that are structurally similar. However, homology modeling produces erroneous predictions for proteins with low similarity to known structures. Ab initio methods, which attempt to predict structures purely from physicochemical principles, are computationally expensive and often fail to achieve experimentally relevant levels of accuracy[7].

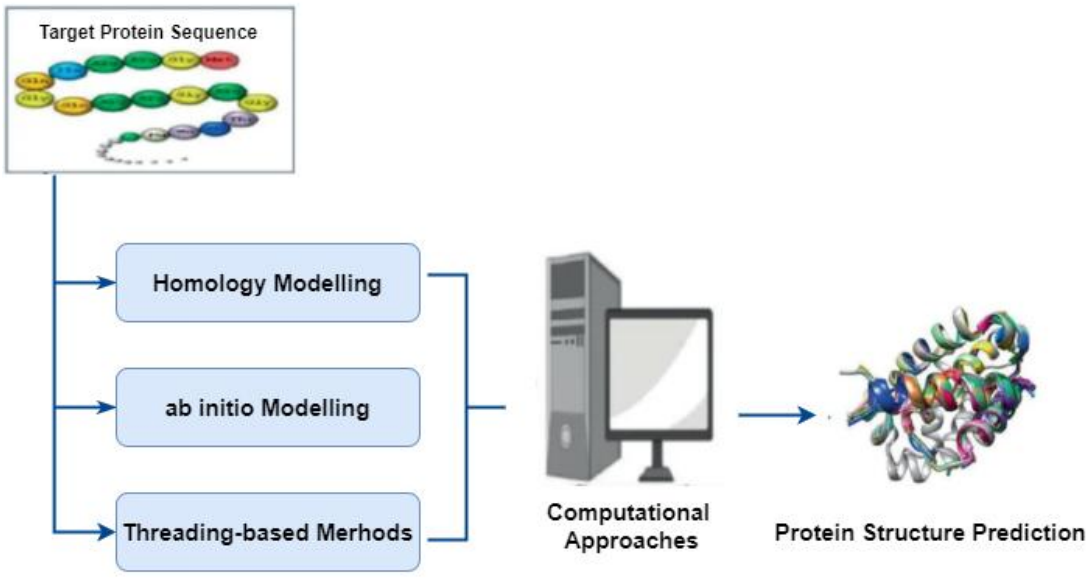

Figure 3. Protein Structure Prediction Through Computational Approaches

- Lack of Automation and Generalization

Many traditional prediction methods require expert intervention, such as manual model optimization and parameter tuning, resulting in poor agreement between different research groups[8]. In addition, these methods usually have insufficient generalization ability when predicting novel protein sequences, limiting their effectiveness in novel protein structure prediction[8]. Figure 3 shows that the bar graph (dark blue) of AlphaFold using deep learning is

significantly lower than the other methods, indicating that AlphaFold's prediction accuracy is significantly higher than the other methods in these test cases. The other methods (G009 to G216) have relatively low prediction accuracy and show little difference in performance from each other. In addition, the black lines above each bar chart indicate the margin of error or variability of the prediction accuracy. In summary, as shown in Figure 4, the use of deep learning allows AlphaFold2 to significantly outperform traditional methods in protein structure prediction.

1.1.2 Challenge

Deep learning has revolutionized the field of protein structure prediction, with its ability to capture complex patterns and dependencies in data[9]. Among the various deep learning architectures, the Convolutional Neural Network (CNN) combined with Long Short-Term Memory (LSTM) networks and Attention mechanisms has emerged as a powerful tool for sequence data analysis with long-term dependencies[10]. CNNs excel at capturing local patterns, while LSTMs process sequence data in both forward and backward directions, capturing long-range dependencies[11]. The addition of an Attention mechanism allows the model to focus on the most relevant parts of the sequence, enhancing the prediction accuracy[12]. Despite significant advancements in protein structure prediction, several challenges remain that hinder accuracy, efficiency, and generalization.

- Complexity of Protein Folding

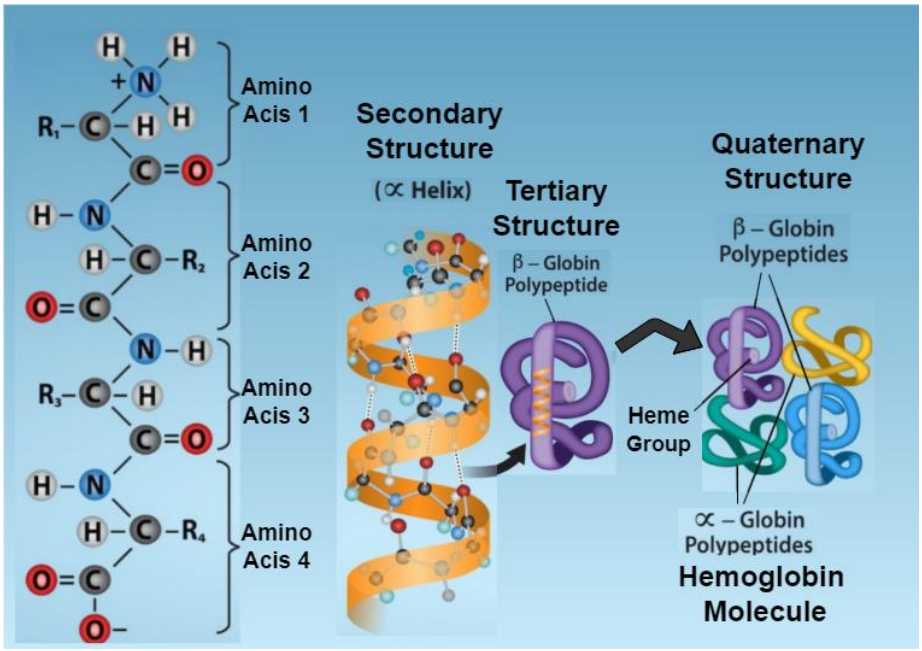

Figure 4. Complex Protein Structure

Protein structure prediction models heavily rely on experimentally determined structures as training data. However, high-quality protein structures are limited due to the high cost and time-consuming nature of experimental techniques such as X-ray crystallography and NMR spectroscopy[13]. Figure 4 illustrates the hierarchical construction process of proteins from amino acids to the final quaternary structure, from which the complexity of protein folding can be seen. This data sparsity affects the ability of deep learning models to generalize across different protein families

- Computational Complexity

Deep learning-based protein structure prediction models usually require a large amount of computational power, especially when integrating ProteinNet, BLSTM, and attention mechanisms. Training and inference on large protein datasets requires high performance Gpus and optimized architectures to maintain efficiency without compromising accuracy[14].

In recent years, there has been a significant interest in developing Attention-ProteinMeNet models for protein structure prediction. These models leverage the strengths of ProteinNet for feature extraction, the temporal capabilities of LSTMs for sequence analysis, and the selective focus of Attention mechanisms to improve the prediction of protein structures[15]. The integration of these components in a single model has shown promising results in capturing the complex relationships between amino acid sequences and their corresponding structures, offering a powerful tool for researchers in the field[16]. The use of TensorFlow as a backend for implementing such models is also gaining attraction due to its flexibility and efficiency in handling complex neural network architectures. TensorFlow's ability to work with different types of data and its support for GPU acceleration make it an ideal platform for training deep learning models like CNN-LSTM-Attention networks[17].

1.2 Aim

The aim of this paper is to explore the application of deep learning, specifically the ProteinMeNet model, in predicting protein structures and functions using evolutionary data. This model combines the strength of ProteinNet for local feature extraction with the ability of bidirectional BLSTM to capture remote interactions within sequences, so that evolutionary data from multiple sequence alignments can be used to predict protein structure.

1.3 Objectives

- ❖ Review the literature, database and existing research about the Protein Structure Prediction.

- This paper tries to use the prediction ability of ProteinMeNet model. By exploiting the evolutionary information embedded in the CB513 and RCSB-PDB, a C-BLSTM model is trained to identify complex patterns and dependencies that indicate protein structure.
- The project will use CB513, RCSB-PDB and other data sets to collect data on protein structure amino acid sequences from the web. The classification of the two helical structures was selected for the training set, validation set and the test set, where 80% of the part was used for training, 10% for validation and 10% for testing.
- Although primarily concerned with the binary classification problem, the project also treats it as a multi-classification problem and also attempts to apply the model to the triple classification in order to later identify more detailed protein structures and enhance generalization.
- In addition, the evaluation of the model will include indicators such as accuracy, loss, recall, and F1-Score will also be used to evaluate performance.

1.4 Project Overview

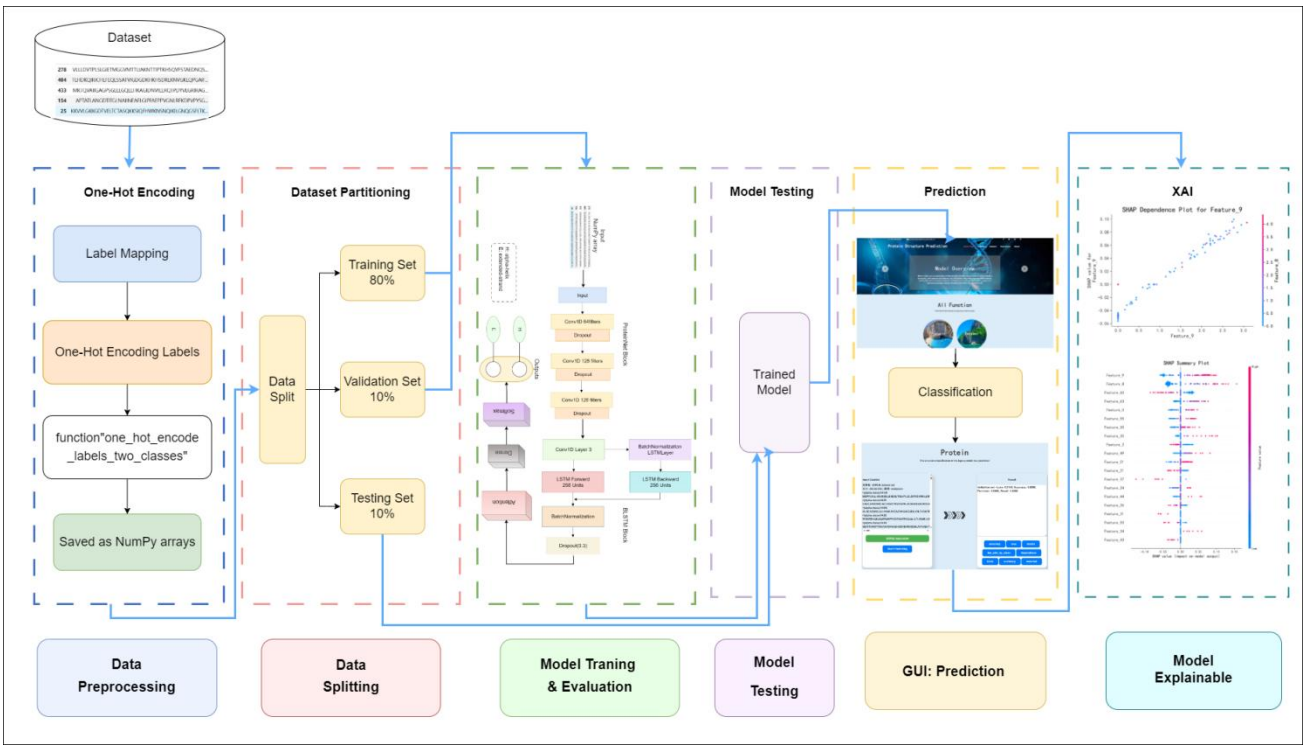

Figure 5. Architectural Overview of Attention-ProteinMeNet Model

Protein structure prediction is of great importance in bioinformatics to help relate genetic data to protein function. This study explores deep learning techniques to improve accuracy and

efficiency and address challenges in traditional and computational approaches. The goal is to advance structural bioinformatics and its applications in medicine and biotechnology. Figure 5 is the flowchart of the project, which includes modules such as data preprocessing and model framework, to illustrate the overall structure of the project.

1.4.1 Scope

The significance of this have a look at lies in its potential to revolutionize the sphere of protein shape prediction. Traditional strategies, inclusive of X-ray crystallography and NMR spectroscopy, are often time-ingesting, high-priced, and restrained by way of the availability of suitable samples[1].

This challenge attempts to deal with the lengthy-standing task in bioinformatics of as it should be predicting protein systems from series records[2], which is essential for understanding protein function and for packages in drug layout, disorder expertise, and treatment improvement.

Furthermore, the manner can be paved for the development of greater correct and green computational tools in structural biology, with profound implications for biomedical research and healthcare[18].

In essence, this look at objectives to decorate our capability to expect protein structures, that's vital for advancing clinical understanding and has practical packages in healthcare and biotechnology.

1.4.2 Audience

- For drug development scientists, accurately predicting protein structures can simplify drug target identification and improve discovery efficiency.
- Data scientists will find value in leveraging these advanced computational methods to analyze complex biological data.
- For doctors and patients alike. Doctors can give better treatment plans with more accurate data. Patients may experience better medications and more precise treatments, leading to improved outcomes and quality of life. The cost of drug development will be lower for biotech companies.

Chapter 2 Background Review

Diving into the domain of protein structure prediction reveals a field significantly shaped by technological progress. While traditional methods laid the groundwork, machine learning has transformed the field by enabling more precise and efficient handling of complex biological data. This chapter highlights the shift from conventional approaches to modern machine learning techniques.

2.1 Traditional-Based Method for Protein Structure Prediction

Protein structure prediction is a key problem in bioinformatics, whose goal is to accurately predict the three-dimensional structure of a protein based on its amino acid sequence. The traditional method of protein structure prediction originated in the 1950s and developed gradually with the discovery of three-dimensional structure of proteins.

Template-Based Modeling (TBM) is one of the earliest prediction methods, and its core technique is homology Modeling developed by Levitt and Chothia[19] in 1976. Based on the assumption of high sequence similarity, protein structure can be predicted with high accuracy. However, this approach does not perform well in cases where template proteins are missing or where sequence similarity is low.

Further development followed. It further broadened the application scope of the traditional method by matching the target sequence with known 3D structural features, but the computational complexity was relatively high. Statistical methods also play an important role in traditional protein structure prediction. In the late 1970s, Garnier et al. [20] developed the GOR method for predicting the local secondary structure of proteins. In addition, the hidden Markov model is a statistical model suitable for processing time series data.

In protein structure prediction, Eddy[21] used HMM for homology search, sequence alignment and domain identification. Figure 6 illustrates the progression of profile HMM architectures from simple BLOCKS motifs to the sophisticated HMMER2 "Plan 7" model, enhancing the prediction accuracy and versatility in protein structure analysis.

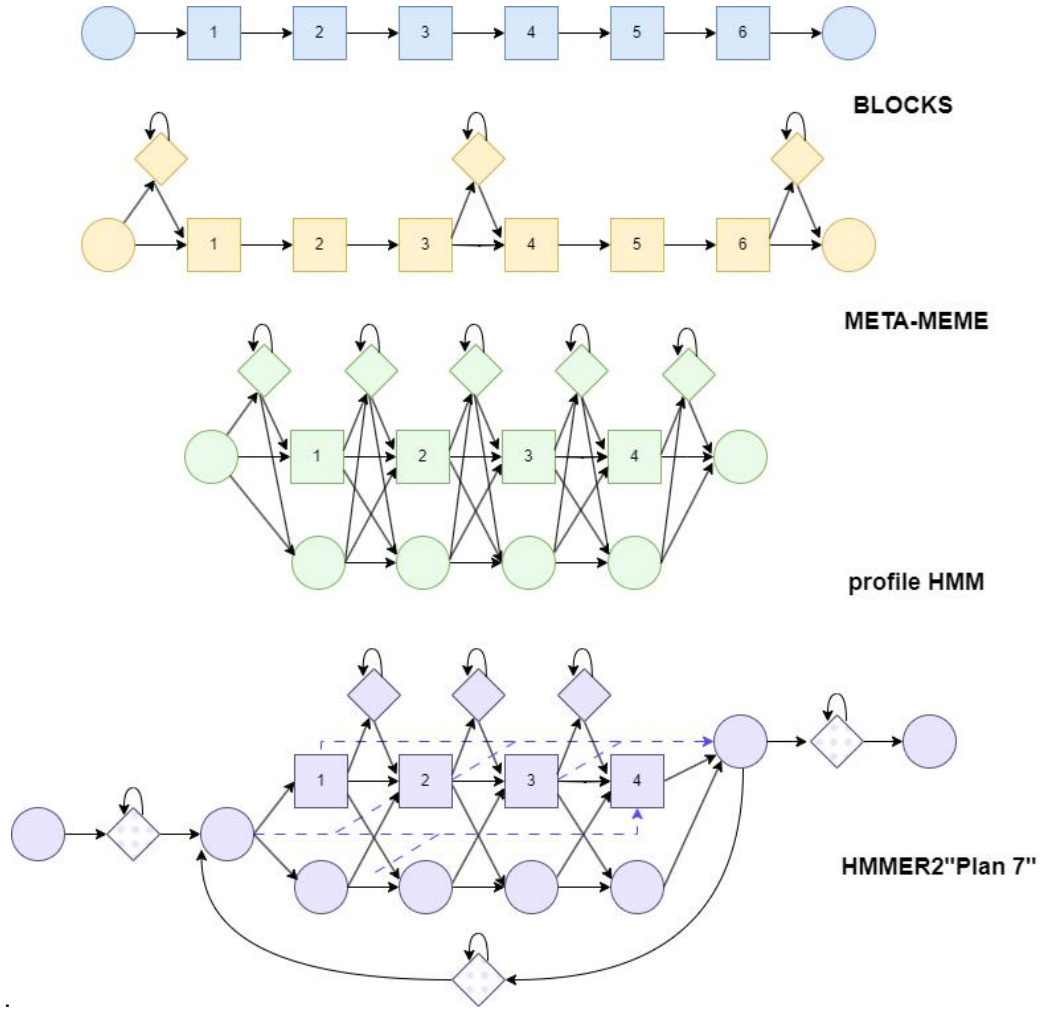

Figure 6. Different model architectures used in current methods by Eddy[21]

2.2 Machine Learning-Based Method for Protein Structure Prediction

In the field of protein structure prediction, machine learning methods have made remarkable progress on the basis of traditional methods. These methods improve the accuracy and efficiency of predictions by learning complex relationships between sequences and structures from large amounts of data. In this area of research, SVM is used to predict secondary structure, functional domains, and protein-protein interactions. In 2017, Yanfei et al. [22] combined protein structure and amino acid sequence information to construct a multi-enzyme function prediction model using SVM and the nearest neighbor algorithm.

Zhenyu et al. [23] study introduced a new multi-objective evolutionary algorithm MO4, which aims to improve the accuracy and efficiency of protein structure prediction by

integrating four different energy functions, while employing PSIPRED for secondary structure prediction. Figure 7 illustrates the MO4 model by Zhenyu et al., which utilizes an evolutionary algorithm to iteratively refine and predict protein structures through processes akin to natural selection, including mutation, crossover, and selection.

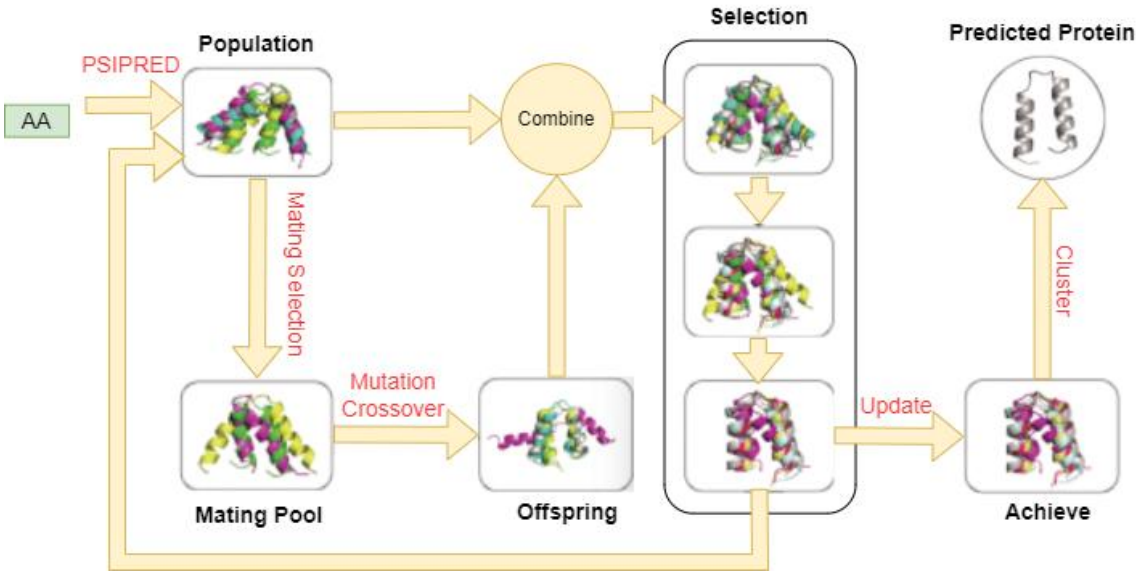

Figure 7. Illustration of the MO4 by Zhenyu et al. [23]

2.3 Deep Learning-Based Method for Protein Structure Prediction

2.3.1 Convolutional Neural Network(CNN)

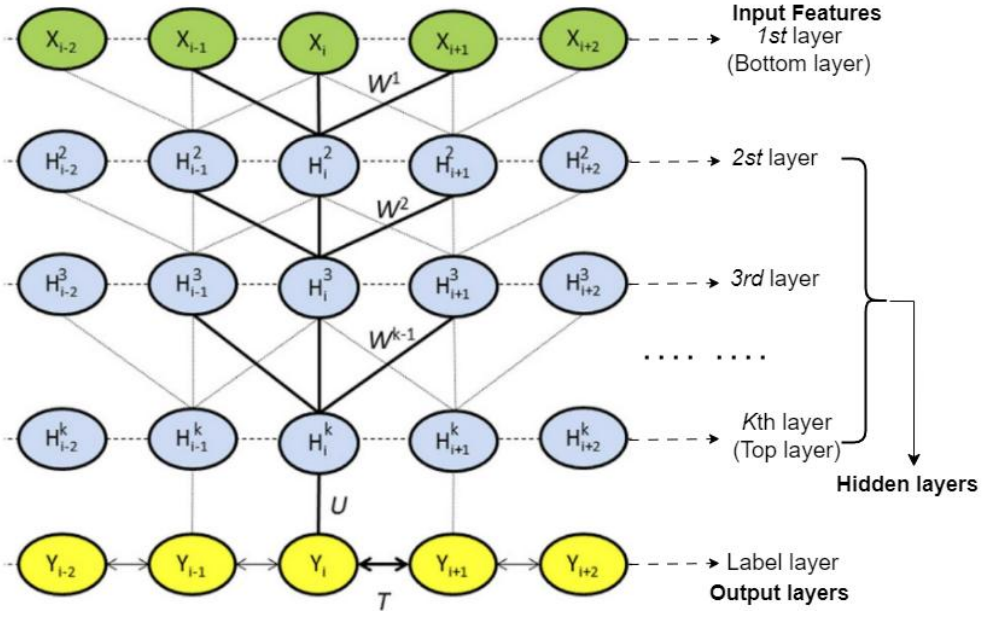

Figure 8. The Architecture of DeepCNF by Wang and Peng[24]

With advances in computational methods, protein structure prediction has evolved from traditional methods to machine learning to deep learning. The study by Wang and Peng[24] proposes the DeepCon model, which uses deep convolutional neural networks (CNNS) for protein secondary structure prediction. The study shows how local features in a sequence can be captured by convolutional layers to significantly improve the accuracy of secondary structure prediction. The DeepCon model performed well on multiple datasets, demonstrating CNNS 'strong feature learning and generalization ability when working with protein sequences. The structure is shown in Figure 8.

In addition, Yihui et al. [25] proposed a method to predict the secondary structure of proteins using two-dimensional convolutional neural networks (2D CNN). The study directly uses a representation of a two-dimensional position-specific score matrix (PSSM) to extract features through a convolutional filter, and this new representation reflects not only evolutionary information, but also sequence interactions between residues.

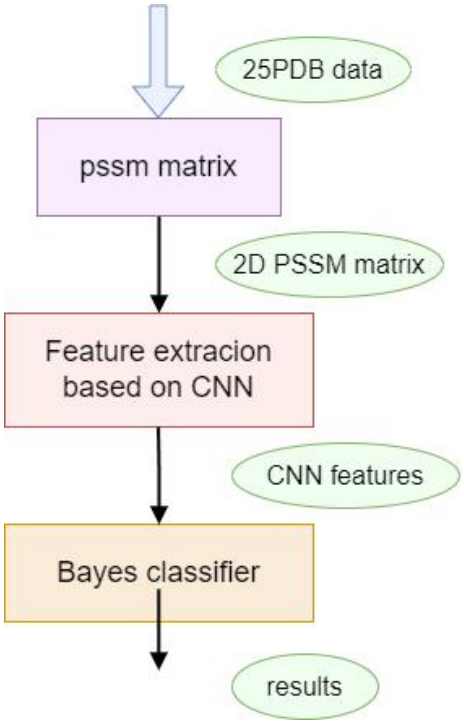

Figure 9. Method Based on 2D Convolutional Neural Network by Yihui et al. [25]

2.3.2 Long Short-Term Memory(LSTM)-Based Method for Protein Structure Prediction

Long Short-Term Memory(LSTM) is a special type of recurrent neural network that is good at capturing long-term dependencies in sequences. In protein structure prediction, LSTM can handle remote interactions in amino acid sequence to improve the accuracy of prediction.

Sonderby and Winther's[26] method successfully improved prediction accuracy by capturing long-term dependencies of protein sequences through LSTM. It was found that LSTM can handle complex dependencies in sequences and achieve significant performance improvements in protein secondary structure prediction. This study provides a new perspective and research direction for deep learning methods of protein structure prediction. Figure 10 illustrates the architecture of unidirectional and bidirectional LSTM networks, highlighting their capability to process sequence data in both d irections to enhance the prediction accuracy in tasks such as protein structure prediction.

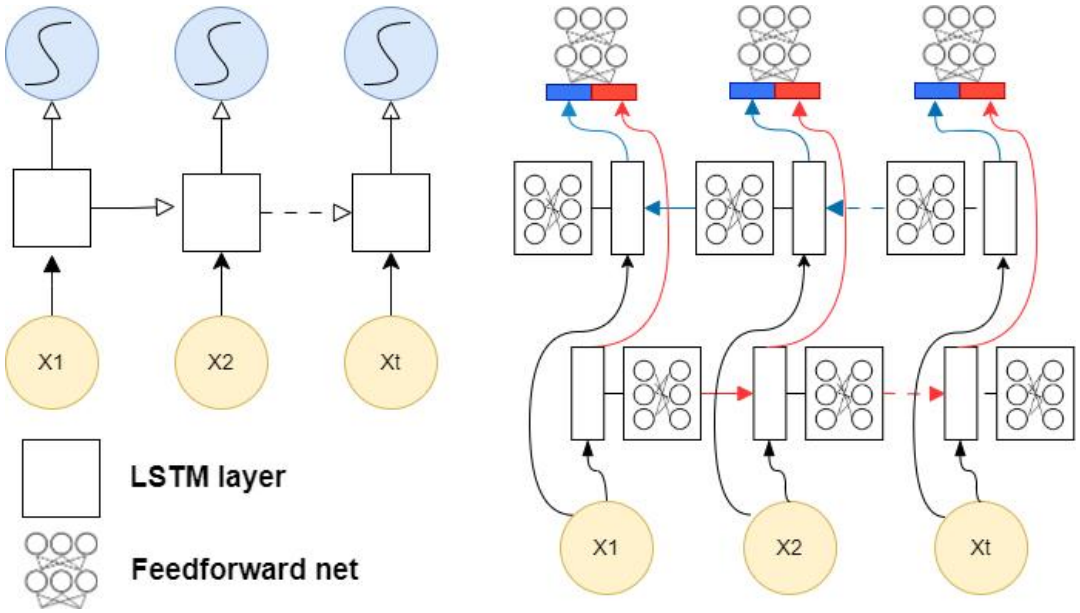

Figure 10. Unrolled Recurrent Neural Networks by Sonderby and Winther[26]

2.3.3 CNN-Attention Based Method for Protein Structure Prediction

Combining CNN with the attention mechanism can focus on key amino acid residues in the sequence while extracting local features. Liu et al. [27] introduces MMSNet, a multimodal deep learning method that integrates protein structure information predicted by AlphaFold2 with a combination of one-dimensional and two-dimensional convolutional neural networks. The model employs a residual attention mechanism to enhance feature extraction, leading to improved performance in grain protein function prediction. The framework diagram of the proposed model is presented in Figure 11.

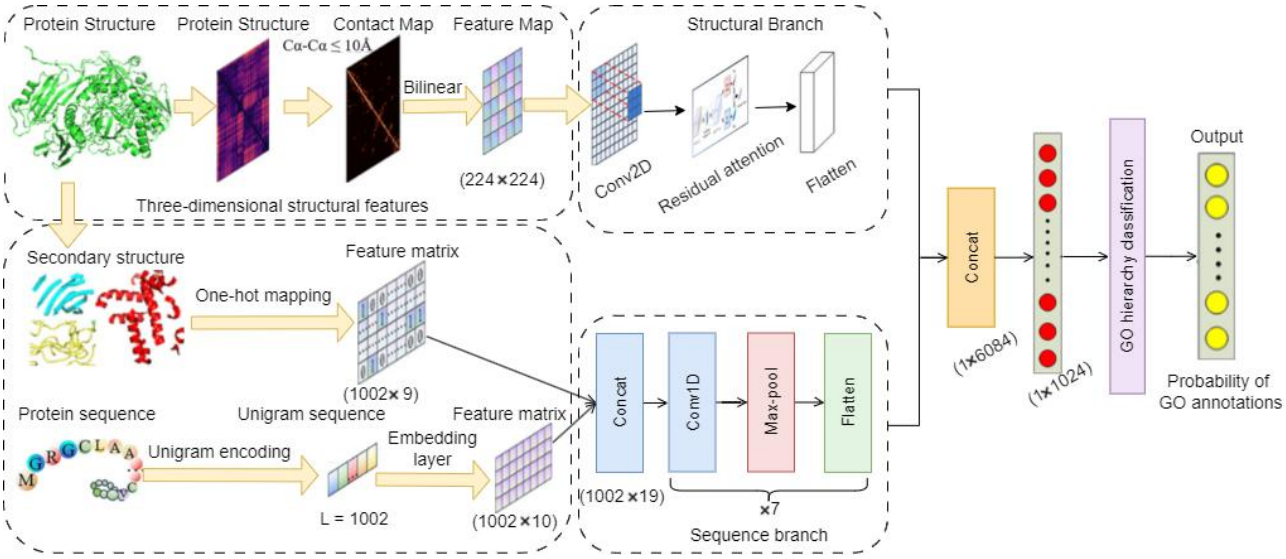

Figure 11. Overview of the flow of the MMSNet method by Liu et al. [27]

And after that, it continued to grow, Jiang and Wang [28] developed a CNN with an attention mechanism to improve protein structure prediction. The attention mechanism enhanced feature selection, leading to better performance on complex sequences and improved accuracy and generalization over traditional CNNs.

2.3.4 Hybrid CNN-LSTM-Based Method for Protein Structure Prediction

The hybrid model combines the advantages of CNN and LSTM to extract local features and capture global dependencies. In protein structure prediction, the hybrid CNN-LSTM model can understand the sequence information more comprehensively and improve the prediction accuracy.

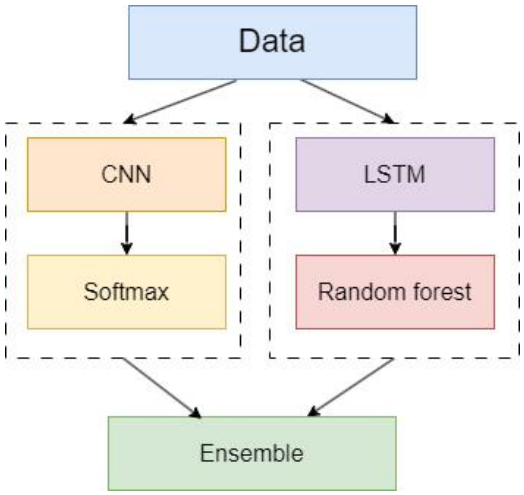

Figure 12. The structure of the prediction model proposed by Cheng et al. [28]

Cheng et al. [29] propose a hybrid model that combines CNN and attention mechanisms for protein secondary structure prediction. The model uses CNN to extract local features of the sequence and identify important feature regions through the attention mechanism, thus improving the accuracy and efficiency of the prediction. Experiments show that this model has significant advantages in processing complex protein data. The framework diagram of the proposed model is presented in Figure 12.

Khan and Mazumdar[30] developed a hybrid deep learning model based on CNN and LSTM that focuses on the modeling of local and global information in protein sequences. The research shows that the model has high accuracy and robustness in predicting secondary structure and identifying key sequence patterns, and is superior to traditional single deep learning models. Srushti et al. [31] introduced two deep learning models, ConcatenatedConvolutional1D model (CCN1D) and C-BLSTM, to predict protein secondary structure with high accuracy using evolutionary spectrum and long-range dependence capture, and achieved very high Q8 accuracy on the CullPDB5926 dataset.

A summary of the different researchers and their findings and possible results can be found in Table 1.

Table 1. Summary of Related Works

	Author	Datasets	Methods & Models	Limitation	Results
Traditional Method	Levitt, M. & Chothia, C. [19]	Protein Data Bank (PDB)	Template-Based Modeling (Homology Modeling)	Limited accuracy for sequences with low homology to templates; failure to predict novel folds.	Accurately predicted structures for sequences with high homology to templates.
	Garnier, J. et al. [20]	Custom dataset	Statistical Methods (GOR for Secondary Structure)	Struggled to predict structures for proteins with complex folds; limited to α-helices, β-sheets,	Achieved basic accuracy in predicting α-helices, β-sheets, and coils based on

				and coils.	sequence patterns.
Machine Learning	Eddy, S. R. [21]	HMMER database	Hidden Markov Model (HMM)	Struggled with predicting novel protein families; limited to sequence-based analysis rather than structure.	Enabled sequence alignment and protein domain identification with probabilistic modeling.
	Yanfei, C. et al. [22]	PDB database	Support Vector Machine (SVM)	Limited in handling complex structures with few training samples; performance drops with diverse sequences.	Predicted enzyme function with 95.5% accuracy for the first EC number classification.
	Zhenyu, L. et al. [23]	34 representative proteins from the Protein Data Bank (PDB)	a new many-objective evolutionary algorithm (MO4)	Challenged by high computational cost and limited to a small set of proteins; struggles with large-scale datasets.	average bRMSD is 5.18Å
CNN	Wang,S. & Peng, Y. [24]	Multiple protein datasets	DeepCon: Deep Convolutional Networks	Limited by the need for large labeled datasets; may struggle with proteins having non-standard folds.	Improved protein secondary structure prediction using deep convolutional layers, capturing local features, and achieving

					high accuracy across various datasets.
	Yihui, L. et al. [25]	25PDB dataset	Proposed a 2D CNN approach for protein secondary structure prediction using two convolutional layers and one max-pooling layer.	Struggled with higher-order structural predictions and complex protein topologies.	Achieved Q3 accuracy of 77.7% using CNN features, outperforming the original features with 73.8% accuracy.
LSTM	Sonderby, S.K. & Winther, O.[26]	Multiple protein datasets	Long Short-Term Memory Networks (LSTM)	Struggled with capturing long-range dependencies in sequences and computational inefficiency for large datasets.	Enhanced ability to handle complex dependencies in sequences, achieving higher precision and stability.
CNN-Attention	Jing, L.et al. [27]	AlphaFold2 Structure Data	CNN with Residual Attention Mechanism (MMSNet)	Dependent on high-quality AlphaFold2 predictions, potential overfitting to training data	Improved grain protein function prediction accuracy with enhanced feature extraction.
	Jiang, Y., &	Multiple protein	Attention-	Still struggles with efficiently selecting	Enhanced feature selection

	Wang, W.[28]	datasets	guided CNN	the most relevant features for large and diverse protein sequences.	with attention, achieving superior performance on complex sequence data.
Hybrid CNN-LSTM	Cheng, J., Liu, H., & Hu, X.[29]	Multiple protein datasets	Hybrid CNN with Attention Mechanism	Limited by the model ' s inability to process longer sequences efficiently; relies on sufficient training data for effective attention mapping.	Improved accuracy and efficiency by focusing on important sequence regions through attention mechanisms.
	Khan, S., & Mazumdar, J.[30]	Multiple protein datasets	Hybrid CNN-LSTM Model	Limited by challenges in combining CNN and LSTM strengths and may struggle with high-dimensional data.	Demonstrated high accuracy and robustness in predicting secondary structures and recognizing critical sequence patterns.
	Srushti, C. S. et al. [31]	CullPDB5926 CB513	CCN1D &C-BLSTM	Accuracy can be reduced when predicting less common secondary structure types or for less annotated proteins.	CCN1D Q8 accuracy = 71.73% C-BLSTM Q8 accuracy = 72.47%

Chapter 3 Methodology

This chapter describes the methodology used to predict protein secondary structures using the Attention-ProteinMeNet model. It covers the datasets used, including RCSB-PDB and CB513, along with the data preprocessing steps. The model architecture integrates ProteinNet for local feature extraction, Bidirectional LSTM (BLSTM) for sequence dependencies, and an attention mechanism to focus on key residues. The chapter also details the experimental setup, version management, and evaluation metrics used to assess the model's performance.

3.1 Approach

This project uses the RCSB-PDB and CB513 datasets for protein secondary structure prediction, with sequences preprocessed via one-hot encoding and normalization. The model integrates ProteinNet for feature extraction, BLSTM for sequence modeling, and an attention mechanism to highlight key residues. Performance is evaluated using accuracy, precision, recall, and loss. A permutation-based analysis assesses the impact of each sequence position by measuring performance drop when shuffled.

3.2 Dataset

Two datasets are used for protein secondary structure prediction: the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB-PDB) and CB513. These datasets provide annotated protein sequences essential for training and evaluating the model. The following sections describe these datasets and their preprocessing steps in detail.

3.2.1 Dataset 1 - Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB-PDB)

The RCSB-PDB dataset is a large-scale resource for protein secondary structure prediction, containing numerous protein sequences obtained from the Protein Data Bank (PDB). These sequences are accompanied by high-resolution structural annotations, ensuring high reliability. The dataset provides detailed amino acid sequences and corresponding secondary structure labels in both SST8 and SST3 classifications. Due to its extensive size and complexity, it is primarily used for model training, while smaller datasets like CB513 are employed for validation and testing to complement its training capabilities.

3.2.2 Dataset 2- CB513

CB513 is a classic benchmark dataset for protein secondary structure prediction, consisting of 513 protein sequences. The secondary structure labels are derived from experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR). There are 8 classes of

secondary structures in CB513: H(alpha-helix), G(3-helix), I(5-helix), E(extended-strand), B(isolated-strand), T(turn), S(bend), and coil ('_'), which are normally reduced into 3 classes. In our research, we map H, G, and I to H; E, B to E; all other states to C, which usually results in lower prediction accuracy than other definitions[24]. The data of this project adopts binary classification, and only H and E are used for test training.

3.2.3 Data Preprocessing

The RCSB-PDB and CB513 datasets are both csv files and have the same format, so the data is preprocessed in the same way.

3.2.3.1 Label Mapping (SST8 to SST3)

The RCSB-PDB and CB513 dataset originally used an 8-class secondary structure classification system (SST8: H, G, I, E, B, T, S, C). To simplify the task, these labels were mapped to a 3-class system (SST3: Helix, Strand, and Coil) as follows:

Helix (H, G, I) → H

Strand (E, B) → E

Coil (T, S, C) → C

Each sequence's SST8 label was mapped to the SST3 system, with a dedicated mapping function.

3.2.3.2 One-Hot Encoding

Protein sequences were encoded using one-hot encoding to transform amino acid sequences into numerical formats suitable for deep learning models: Each amino acid (A, C, D, E, etc.) was mapped to a unique index, with a total of 20 amino acids plus a code for unknown residues (X). Sequences were padded to the length of the longest sequence in the dataset to ensure uniform input size. And one-hot encoding was applied to create 3D tensors, where each sequence's amino acids were represented by binary vectors.

3.2.3.3 Label Encoding for Binary Classification

For the binary classification task, the Coil class (C) was merged with the Strand class (E). The

SST3 labels were then processed to retain only two classes: First class is Helix (H). Second class is Non-Helix (E, including all Coil structures). The labels were padded to match sequence lengths and one-hot encoded to produce 3D tensors suitable for training.

3.2.3.4 Feature and Label Shapes

The encoded features and labels for each subset (training, validation, and testing) were prepared as follows:

Features (X): Represented as 3D tensors with dimensions (number of samples, sequence length, number of amino acids).

Labels (y): Represented as 3D tensors with dimensions (number of samples, sequence length, number of classes).

3.2.4 Dataset Splitting

As benchmarks for protein secondary structure prediction, the RCSB-PDB dataset and CB513 dataset are divided into three subsets: training set, validation set and test set. For the RCSB-PDB dataset, a two-stage segmentation process is used. Initially, 80% of the data is assigned to the training set and the remaining 20% is used as a temporary set. Subsequently, this temporary set was equally divided into validation and test subsets, each containing 10% of the original dataset. The CB513 dataset was split using the same strategy to ensure a balanced and repeatable split for model training and evaluation, resulting in the same final split rate of 80% for training, 10% for validation, and 10% for test.

3.3 Proposed Model Structure

The proposed model combines several components to enhance protein secondary structure prediction. These include ProteinNet for local feature extraction, BLSTM for sequence dependency modeling, and an attention mechanism for focusing on key residues. The following sections provide a detailed description of each component, starting with ProteinNet, which plays a crucial role in extracting local features from protein sequences.

3.3.1 ProteinNet

In the ProteinNet module, there are three modules included. As shown in Figure 13, the first module consists of a 1D convolutional layer with 64 filters and a Dropout layer. The other two modules are each composed of a 1D convolutional layer with 128 filters and a Dropout layer. These three modules are connected in sequence to effectively extract local features from protein sequence data. This module plays a crucial role in capturing the local correlations and important feature information of protein sequences.

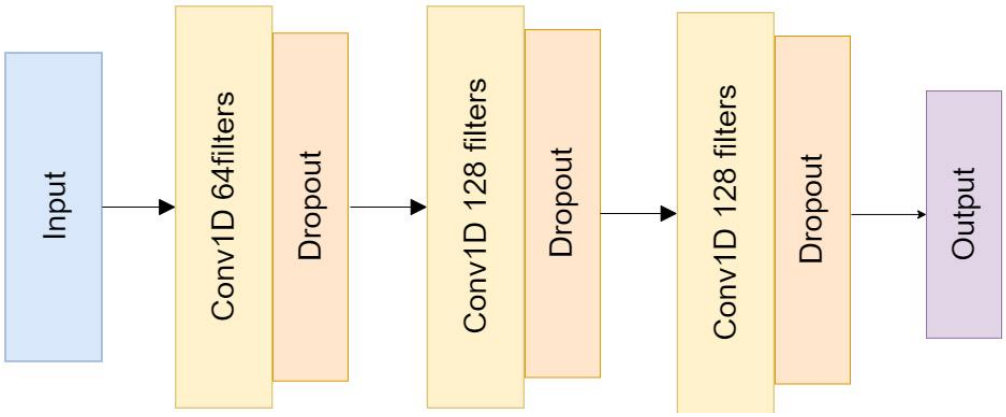

Figure 13. Architectural Overview of ProteinNet Block

3.3.2 Bidirectional Long Short-Term Memory (BLSTM) Block

The BLSTM block can perform forward and backward processing of data. This allows it to capture sequence information from both perspectives, providing a richer feature representation than the unidirectional LSTM. BLSTM efficiently processes sequential data, as shown in Figure 14. It processes sequences in both directions and combines the outputs, allowing the model to capture comprehensive context information from past and future sequence elements, enhancing the model's ability to effectively handle complex sequence tasks. It captures dependencies between sequence positions and manages remote dependencies.

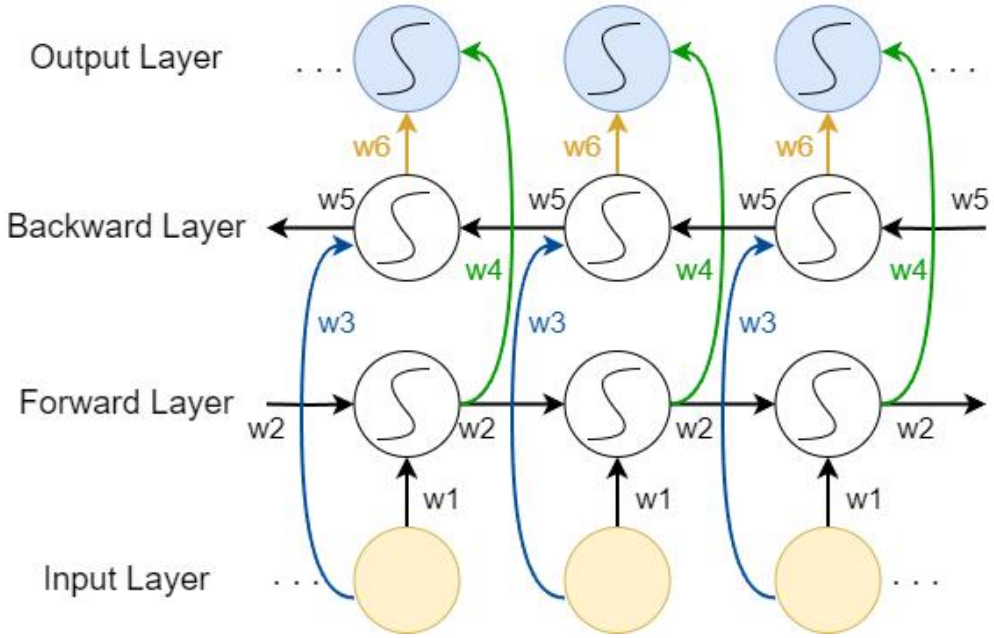

Figure 14. Network Structure of BLSTM

3.3.3 ProteinMeNet Block

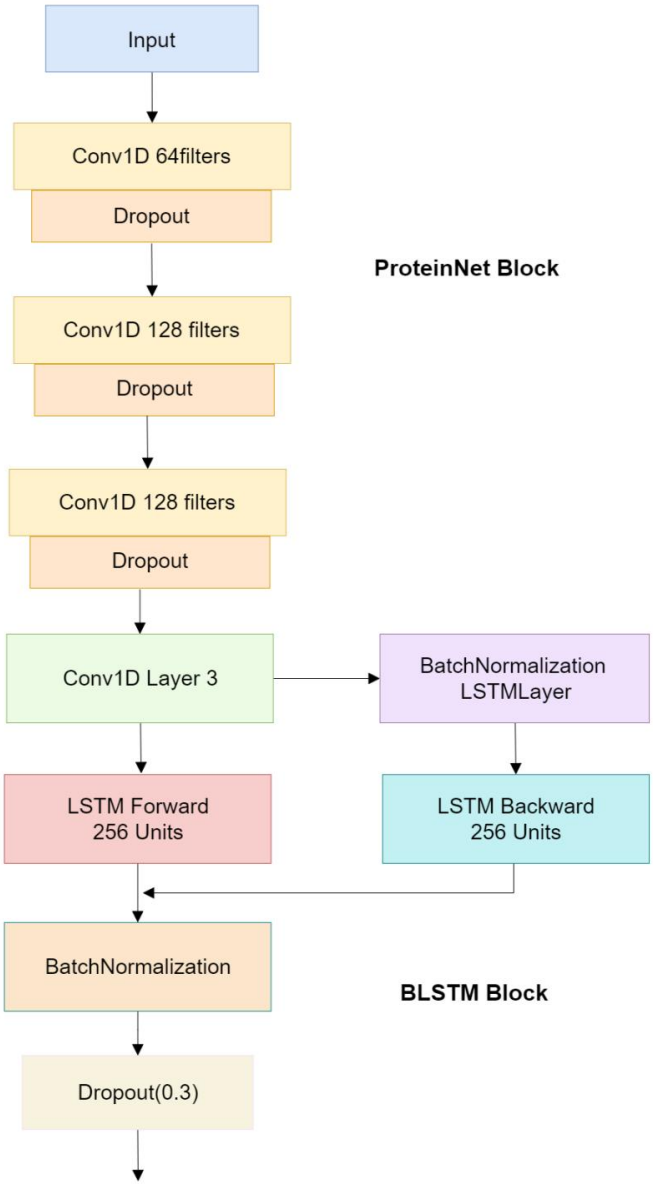

Figure 15. The Architectural Overview of ProteinMeNet Block

Figure 15 shows the ProteinMenet module, containing ProteinNet and BLSTM. Using ProteinNet alone, while local feature extraction is possible through its Conv1D layer with 64, 128, 128 filters each with Dropout, long-term dependencies of the sequences may be missed. BLSTM alone, although its forward and backward layers of 256 units capture temporal dependencies well in both directions, may not be as effective in extracting local features.

This integrated model combines both blocks to address these limitations. The ProteinNet Block first processes the input through three Conv1D layers each followed by Dropout for

regularization. The output then goes through a third Conv1D layer. The BLSTM Block takes this output, applies Batch Normalization, splits it into forward and backward LSTMs with 256 units, combines their results, and applies another Batch Normalization followed by a Dropout of 0.3. By merging local feature extraction from ProteinNet and bidirectional dependency modeling from BLSTM, the model effectively handles complex sequence data.

3.3.4 Attention Block

In protein structure prediction, the purpose of the Attention layer is to identify the most critical parts of the sequence for protein structure prediction. This helps the model focus on areas of the sequence that are critical for structural stability and functionality. As shown in Figure 16, the input consists of the query (Q), key (K), and value (V) matrices. The Q and K are multiplied to measure similarity, and then scaled to stabilize the training process. Optional masks can ignore certain elements. After applying softmax to obtain attention weights, these weights are multiplied with the value (V) matrix to generate the output. This enables the model to focus on different parts of the input sequence when processing each element. Using the self-attention mechanism allows the model to capture the dependencies between elements regardless of their distances in the sequence, thereby achieving parallel processing and usually bringing better performance in sequence modeling tasks.

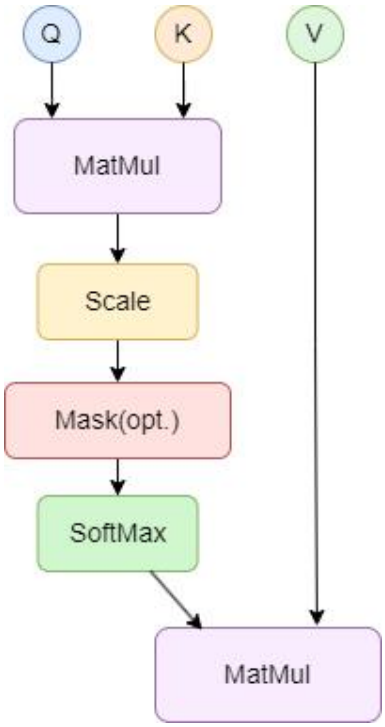

Figure 16. The Architectural Overview of Attention Block

3.3.5 Attention ProteinMeNet Model

Although the ProteinMeNet model is effective, it has certain limitations. ProteinNet struggles with capturing long-range dependencies, and while BLSTM is proficient at modeling temporal dependencies, it is less efficient at extracting local features and faces challenges in parallelizing training. To address these issues and enhance performance, the introduction of a self-attention mechanism optimizes the model. The Attention-ProteinMeNet model, as shown in Figure 17, integrates this attention mechanism to better focus on critical sequence elements, improving both feature extraction and overall model efficiency.

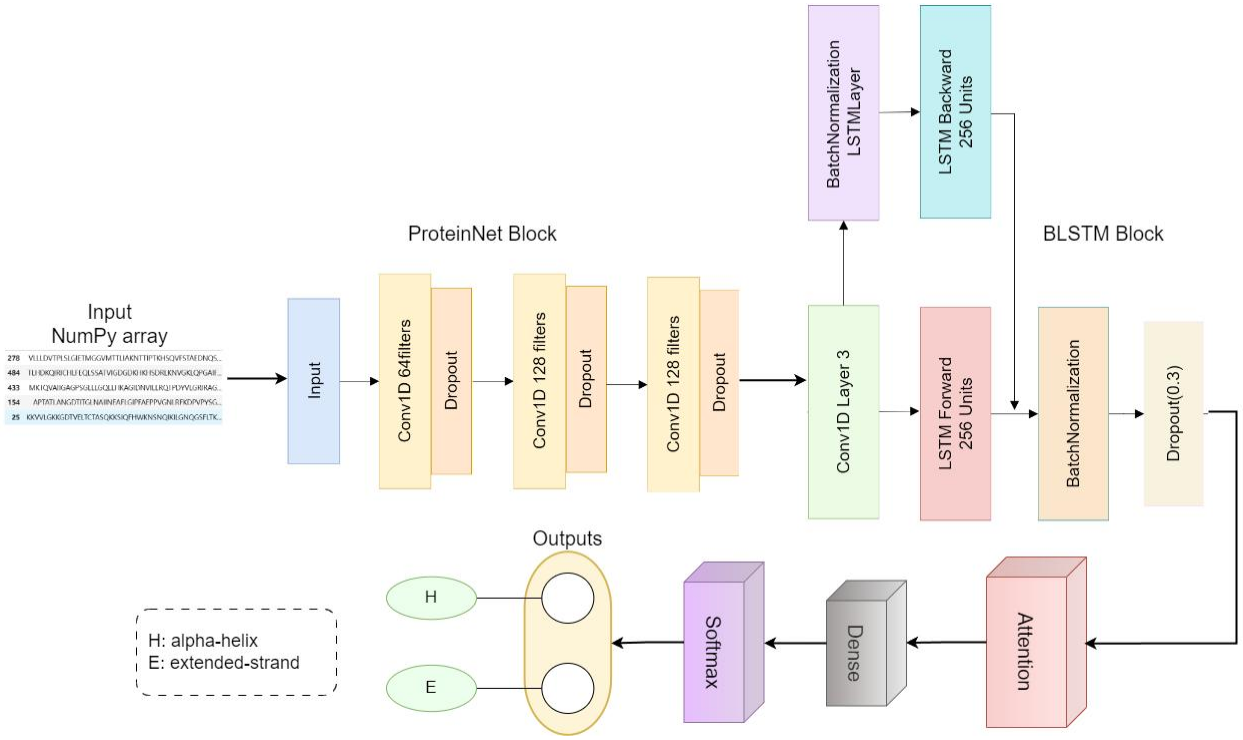

Figure 17. The Architectural Overview of Attention ProteinMeNet Model

The ProteinNet block extracts local features from the input sequences. It consists of three 1D convolutional layers. The BLSTM block is crucial for understanding sequence dependencies. It uses a bidirectional LSTM with 256 units to capture both forward and backward sequence information, ensuring comprehensive feature context is provided to the Attention layer. To prevent overfitting, L2 regularization and a 0.3 dropout rate are applied. This setup enhances the model's ability to generalize from the training data, which is vital for accurate protein structure prediction. The Attention layer receives output from the bidirectional LSTM layer. Because Bidirectional LSTM is used, the Attention layer receives LSTM output in both directions, the forward and backward hidden states. A fully connected Dense layer maps the attention-

enhanced features to the final prediction. In addition, the softmax activation function is used to produce a probability distribution over two output classes.

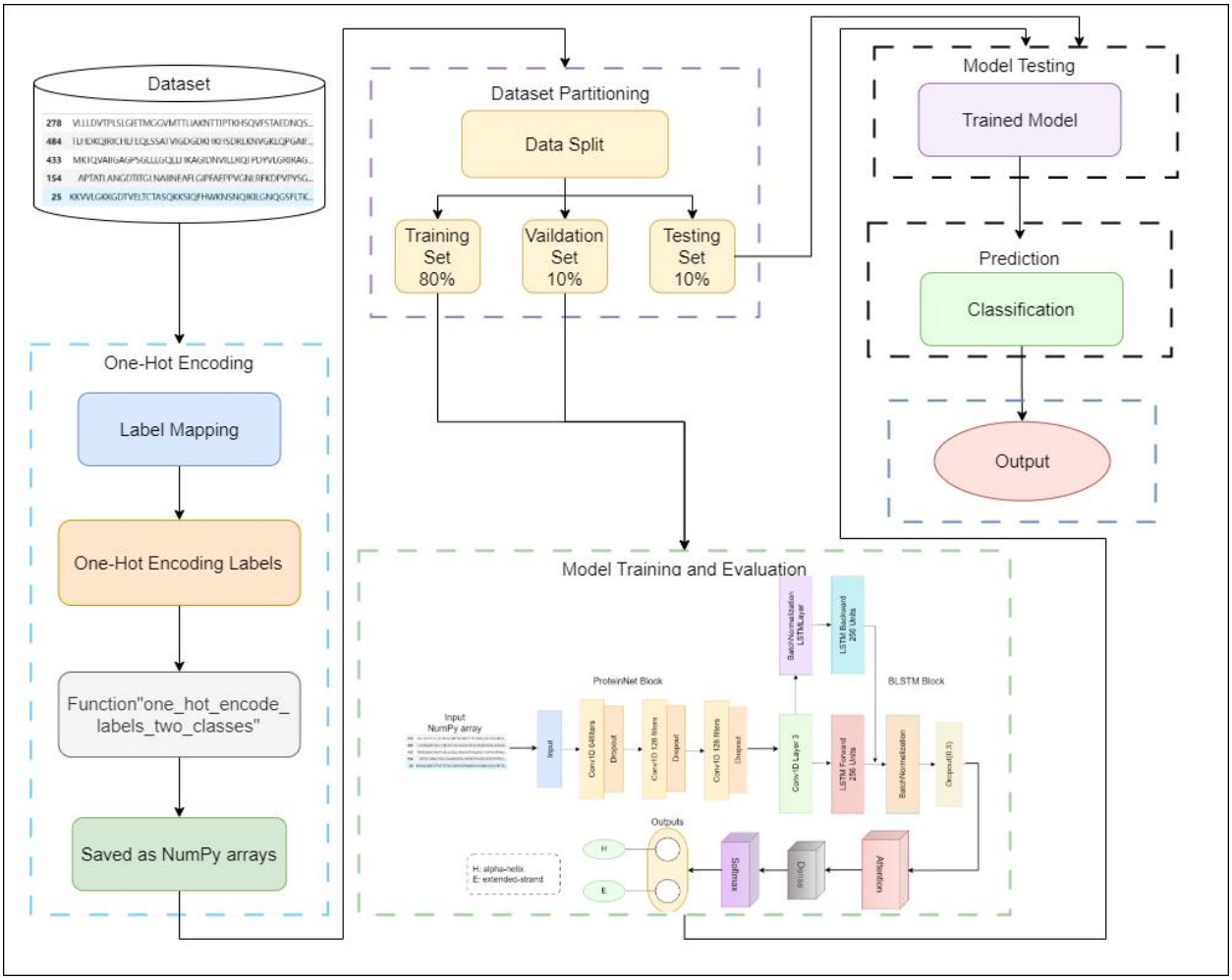

Figure 18. Flow Chart of Attention ProteinMeNet Model for Protein Structure Prediction

Figure 18 shows the overall flow diagram of the project. This design integrates robust data preprocessing, advanced hybrid model architectures and system evaluation. By combining a single model of ProteinNet, BLSTM and Attention architectures, and then integrating these models to form a combined architecture. This project efficiently handles protein sequences using one-hot encoding and label mapping techniques. The 80%-10%-10% dataset split ensures robust training, validation and testing. Regularization techniques such as dropout and batch normalization can prevent overfitting. The softmax activation function in the output layer provides clear classification results. In summary, this model is a powerful tool for protein structure prediction. This model achieves accurate and biologically meaningful predictions of protein secondary structures.

3.4 Experimental Setup & Technology

The experimental setup of this project is shown in Table 2.

Table 2. The experimental setup of Attention-ProteinMeNet

Parameter Category	parameter values
Conv1D Layers	3 layers with 64, 128, 256 filters and 0.3 Dropout each
Bidirectional LSTM	256 units with 0.3 Dropout
Output Layer	2 units with softmax activation
Optimizer	Adam with 0.0007 learning rate
Training Params	Batch size 32, max 100 epochs
Callbacks	EarlyStopping (patience 20), ReduceLROnPlateau (factor 0.5, patience 10)

Dropout of 0.3 was chosen to balance the complexity of the model with the need to prevent overfitting. This ratio provides the appropriate regularization intensity to help the model prevent overfitting during training without placing too great a restriction on the network learning useful features. This value may be optimized in several experiments, and the test results are shown in Table 3, combined with other regularization methods such as L2 regularization and Batch Normalization, and the final optimal choice is obtained.

Table 3. Dropout Test Result use Dataset CB513

Dropout	Loss	Accuracy	F1-score	Precision	Recall
0.2	0.1466	0.9404	0.9417	0.9434	0.9404
0.3	0.1396	0.9439	0.9436	0.9439	0.9433
0.4	0.1401	0.9429	0.9440	0.9408	0.9439
0.5	0.1396	0.9416	0.9363	0.9366	0.9416

The technology this project will be using is displayed in Table 4.

Table 4. Summary of Relevant Technology involved in this project

Software	Framework	Tensorflow
	Language	Python
	Libraries	Numpy, SciPy, Pandas, Matplotlib, Keras
Hardware	Central processing unit(CPU)	Intel(R) Core(TM) i9-14900HX 2.20 GHz
	Graphic Processing Unit(GPU)	NVIDIA G_SYNC GTX 4060

3.5 Project Version Management

- Keep up to date by using git to upload the weekly reference form, weekly report, and weekly code collection to GitHub
- All files, including datasets, model code, references, weekly reports and all types of files will be copied in three copies, one on the local computer, one on the hard drive and one on github.

3.6 Evaluation Metrics

Nine major evaluation metrics are used in this experiment which include Loss, Accuracy, ROC-AUC curve, Precision, Recall/Sensitivity, Specificity, Confusion Matrix, F1 score and Precision-Recall curve.

3.6.1 Loss Function

The loss function represents the discrepancy between the predicted values of the model and the true labels. The loss function is the main optimization objective in the model training process. The model adjusts its parameters by minimizing the loss function. The model is optimized to minimize the loss. The lower the loss value is, the closer the predicted results of the model are to the true values, and the better the performance of the model is. The equation for the loss function is presented in equation (1).

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{i,j} log\left(\hat{y}_{i,j}\right) \quad (1)$$

3.6.2 Confusion Matrix

The confusion matrix is a tool used to evaluate the performance of a classification model, showing the relationship between the prediction and the true label. The Confusion Matrix consists of four main parts:

- True Positives (TP): the number of beta (E) structures that correctly predicted as beta.
- False Positives (FP): the number of alpha (H) structures that the model incorrectly predicts as beta.
- True Negatives (TN): the number of alpha (H) structures that the model correctly predicts as alpha.
- False Negatives (FN): the number of beta (E) structures that the model incorrectly predicts as alpha.

Table 5. Confusion Matrix structure

	Predicted Positive	Predicted Negative
True Positive	TP	FN
True Negative	FP	TN

3.6.3 Accuracy

It indicates the ratio of the number of samples correctly predicted by the model to the total number of samples.In this project, accuracy is measured by the ratio of the number of protein samples correctly classified to the total number of evaluated samples. However, due to the potential imbalance of classes in the protein dataset, some structures are more common than others. Therefore, accuracy is combined with other indicators to conduct a more detailed performance evaluation. The equation for the accuracy is presented in equation (2).

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (2)$$

3.6.4 Precision

Precision reflects the accuracy of forward-looking predictions. It indicates how many samples are predicted to belong to a specific protein category and are actually correct. The equation for the precision is presented in equation (3).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

3.6.5 Recall/Sensitivity

Recall/Sensitivity measures the ability of a model to identify actual positive cases, that is, the number of true protein instances correctly recognized by the model. In cases where missing

positive cases may have significant consequences, such as when identifying proteins related to disease mechanisms, a high recall rate is crucial. The equation for the recall/sensitivity is presented in equation (4).

$$Recall/Sensitivity = \frac{TP}{TP+FN} = \frac{TP}{P} \quad (4)$$

3.6.6 Specificity

Specificity is complemented by recall by measuring the model's ability to correctly identify negative cases, that is, how many non-protein instances are correctly identified. The equation for the specificity is presented in equation (5).

$$Specificity\ (SP) = \frac{TN}{TN+FP} \quad (5)$$

3.6.7 F1-score

The F1 score can comprehensively measure both precision and recall, and thus is particularly useful when it is necessary to balance the two types of errors, namely false positives and false negatives. In protein classification, the losses caused by these two types of errors may be equally serious, and the F1 score can comprehensively evaluate the performance of the model. The equation for the F1-score is presented in equation (6).

$$F_1 - score = \frac{2TP}{2TP+FP+FN} \quad (6)$$

3.6.8 Receiver Operating Characteristic - Area Under the Curve(ROC-AUC)

The ROC curve represents the relationship between the true positive rate and the false positive rate at various decision thresholds. The equation for the ROC_AUC curve is presented in equation (7) and (8).

$$True\ Positive\ Rate(TPR) = \frac{TP}{TP+FN} \quad (7)$$

$$False\ Positive\ Rate(FPR) = \frac{FP}{FP+TN} \quad (8)$$

AUC measures the area under the ROC curve; the closer to 1, the better the model. An AUC of 1 indicates perfect classification, while 0.5 suggests random guessing. Higher AUC values reflect stronger ability to distinguish between positive and negative classes.

3.6.9 Precision-recall curve

The precision-recall curve shows the relationship between Precision and Recall. The AUC is the area under the Precision-Recall curve, similar to the AUC of the ROC curve.

3.6.10 Early Stopping and Reduce Learning Rate on Plateau

Early Stopping is a technique used during model training to prevent overfitting. It involves monitoring the model's performance on a validation set and halting the training process when the model's performance ceases to improve.

Key Parameters Explained:

- Monitor: Usually tracks val_loss or val_accuracy.
- Patience: The number of epochs with no improvement before stopping training.
- Min Delta: The smallest performance change needed to consider an improvement, avoiding stops due to minor changes.

"Reduce Learning Rate on Plateau" lowers the learning rate when validation performance stops improving, helping the model avoid stagnation and continue optimizing. This strategy encourages better convergence by adapting the learning rate during training.

Chapter 4 Implementation and Results

Building on the methodology outlined in the previous chapter, this chapter presents the implementation and performance evaluation of the Attention-ProteinMeNet model. This chapter presents training results on RCSB-PDB and CB513, compares them with other models, applies XAI for interpretation, and showcases a GUI for user interaction.

4.1 Results of Model Training

The following results are the final results of the proposed model, which uses two datasets of RCSB-PDB and CB513. The proposed Attention-ProteinMeNet model is robust to protein structure prediction, with a validation accuracy of 96% and stable convergence in RCSB-PDB data set. The accuracy of verification at CB513 is 94%. The architecture combines three key components:

- Local features were extracted using three 1D convolution layers with 64, 128, and 256 filters, ReLU activation, and L2 regularization with a lambda of 1e-4.
- Remote dependencies were captured by a 256-cell bidirectional LSTM layer with a dropout rate of 0.3.
- The model uses a self-focusing mechanism to weigh important features dynamically. It was trained with Adam optimization at an initial learning rate of 0.0007, and converged within 100 epochs using early stopping with a patience of 20 and learning rate reduction by a factor of 0.5.

4.1.1 Attention-ProteinMeNet model(RCSB-PDB)

The evaluation is based on several key performance metrics, including accuracy, loss, and other relevant measures. These metrics provide insights into the model's ability to effectively predict protein secondary structures and assess its generalization across different sequences.

a. Accuracy

The Attention-ProteinMeNet model achieved strong performance, with training accuracy reaching 96% and validation accuracy stabilizing at 94%, as shown in Figure 19. The close alignment between the training curve and the validation curve indicates effective learning and prevents overfitting, while high precision highlights the model's ability to capture local structural patterns and remote dependencies to distinguish between α-helical and β-sheet structures. The

small gap of approximately 2.56% between the training and validation accuracies suggests the potential for further regularization improvements.

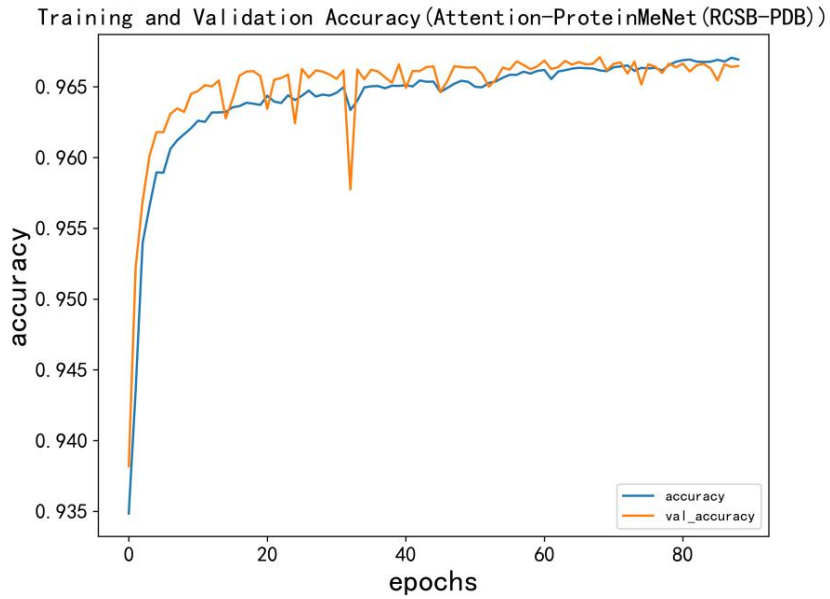

Figure 19. Accuracy_curve of Attention-ProteinMeNet using RCSB-PDB

(Train Acc = 0.9671, Val Acc = 0.9415)

b. Loss

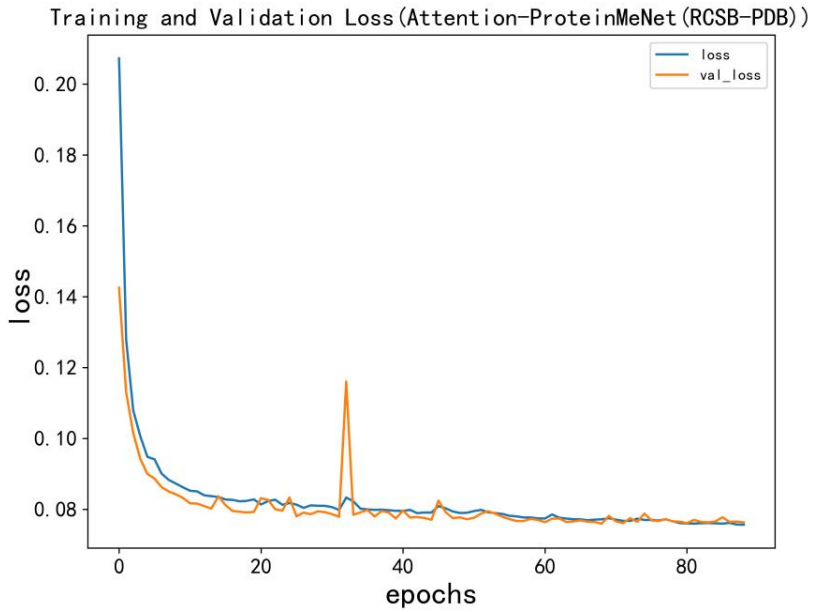

Figure 20. Loss_curve of Attention-ProteinMeNet using RCSB-PDB

(Train Loss = 0.0760, Val Loss = 0.0763)

As shown in Figure 20, the Attention-ProteinMeNet model's loss curve demonstrates stable convergence. After 80 epochs, the training loss of 0.0760 and validation loss of 0.0763 are nearly identical, indicating effective optimization and no overfitting. The curves drop sharply in the early stage and then stabilize, showing rapid initial learning and fine-tuning. The close consistency and ultra-low loss value confirm the model's strong generalization ability and high accuracy in secondary structure recognition on the RCSB-PDB dataset. Minor fluctuations in the later stage may reflect the attention mechanism's dynamic weighting of sequence features.

c. Confusion Matrix

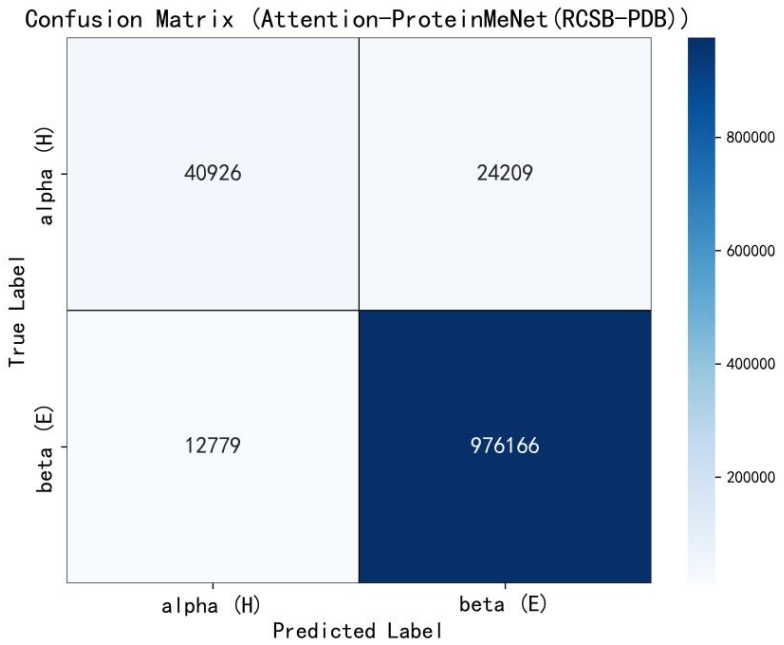

Figure 21. Confusion Matrix of Attention-ProteinMeNet using RCSB-PDB

As shown in Figure 21, the confusion matrix highlights the strong performance of Attention-ProteinMeNet on RCSB-PDB data. The model accurately predicted 97,616 β-sheets (E) and 40,926 α-helices (H). β-sheet detection achieved a 98.7% recall rate, but 12,779 misclassified α-helices indicate challenges in identifying helical fragments, possibly due to their dynamic nature. The model's overall accuracy of 94.2% confirms its effectiveness in secondary structure prediction, with β-sheets performing slightly better, likely because of their more distinct sequence patterns.

d. ROC Curve

As shown in Figure 22, the ROC curve shows excellent classification performance with an AUC

of 0.983, indicating the model's strong ability to distinguish protein secondary structures. The curve's steep rise and proximity to the top-left corner demonstrate high true positive rates while maintaining low false positives across all thresholds. This confirms the Attention-ProteinMeNet architecture's effectiveness for protein structure prediction

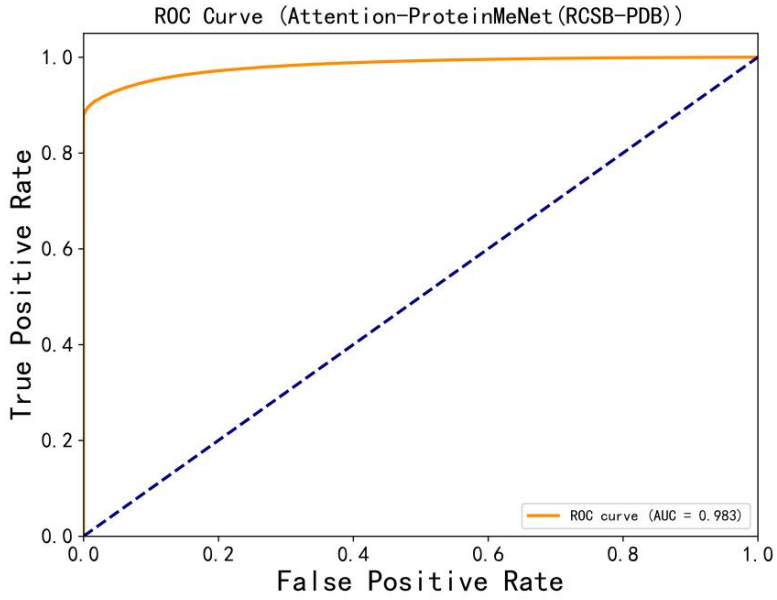

Figure 22. ROC Curve of Attention-ProteinMeNet using RCSB-PDB(AUC = 0.983)

e. Precision_Recall Curve

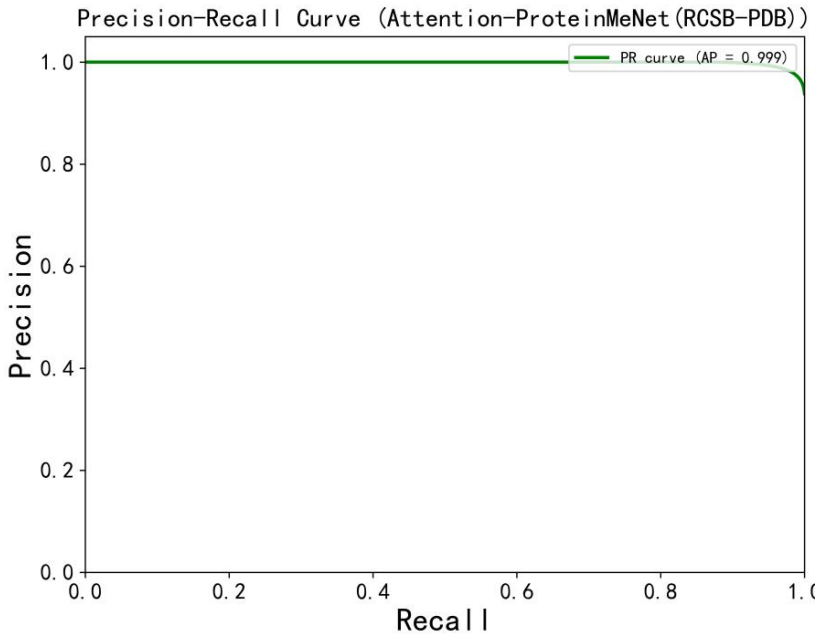

Figure 23. Precision-Recall Curve of Attention-ProteinMeNet using RCSB-PDB(AP = 0.999)

As shown in Figure 23, the Precision-Recall curve achieves near-perfect performance (AP=0.999), demonstrating the model's excellent balance between precision and recall for protein secondary structure prediction. The consistently high precision across all recall levels confirms robust classification capability on the RCSB-PDB dataset.

4.1.2 Attention-ProteinMeNet model(CB513)

a. Accuracy

During training, both accuracy and validation accuracy steadily increased, as shown in Figure 24. In the early stages (0 to 20 epochs), accuracy rose quickly—training accuracy from 87% to 92%, and validation accuracy to 90%—showing strong feature extraction. From 20 to 100 epochs, accuracy growth slowed, ending with 94% training and 92% validation accuracy. The small 2% gap indicates good generalization and no clear overfitting.

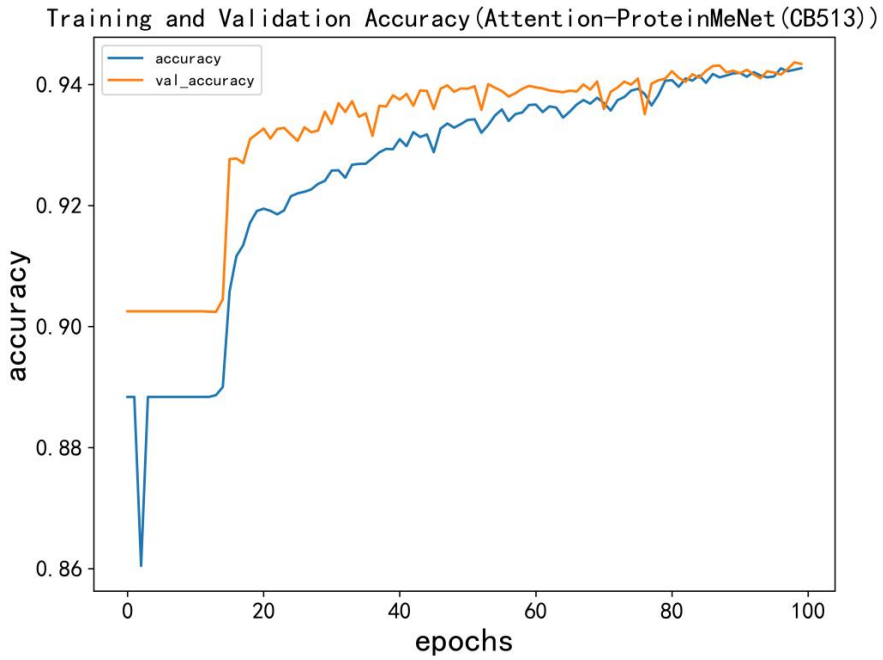

Figure 24. Accuracy_curve of Attention-ProteinMeNet using CB513

(Train Acc = 0.9427, Val Acc = 0.9434)

b. Loss

The loss function curve (Figure. 25) shows that both training loss (loss) and validation loss (val_loss) continue to decline and converge in the same trend. At the beginning of the training period (0-20 epoch), the loss value decreased rapidly (training loss decreased from 0.50 to 0.25, validation loss decreased from 0.45 to 0.22), indicating that the model optimization was in the

right direction. In the subsequent training (20-100 epoch), the loss value further slowly decreased, and finally the training loss stabilized at 0.15, and the verification loss was 0.20. The verification loss was slightly higher than the training loss, but the gap was reasonable and in line with the expected performance of the model training.

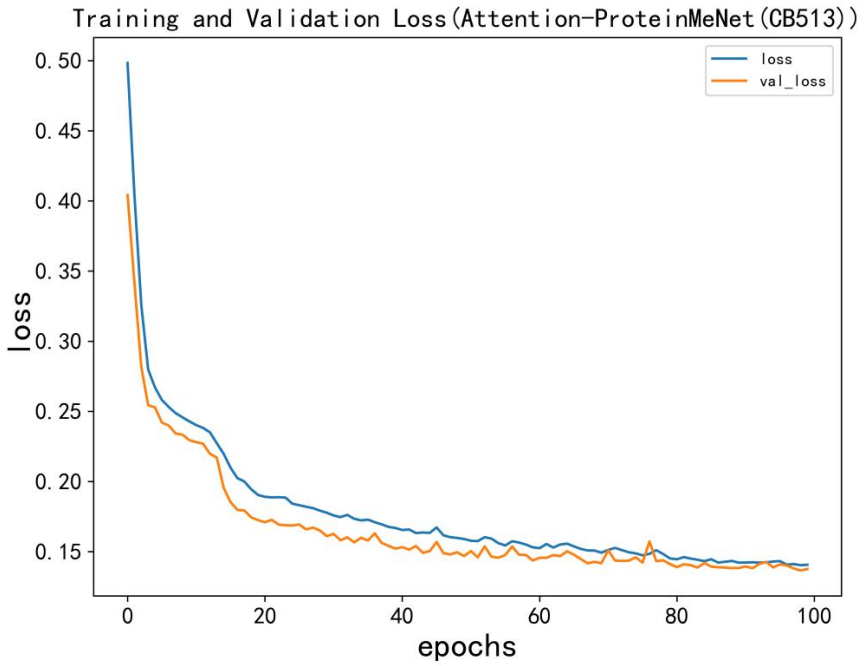

Figure 25. Loss_curve of Attention-ProteinMeNet using CB513

(Train Loss = 0.1406, Val Loss = 0.1376)

c. Confusion Matrix

Figure 26 shows that the model predicts class E (folding) with high accuracy at 97.5%, and class H (spiral) with a recall of 93.1%, though some errors occur. This may result from data imbalance and difficulty in detecting short spirals. The model combines ProteinNet, BLSTM, and attention to enhance performance, especially for class E. Future improvements for class H may involve adjusting class weights or adding evolutionary data.

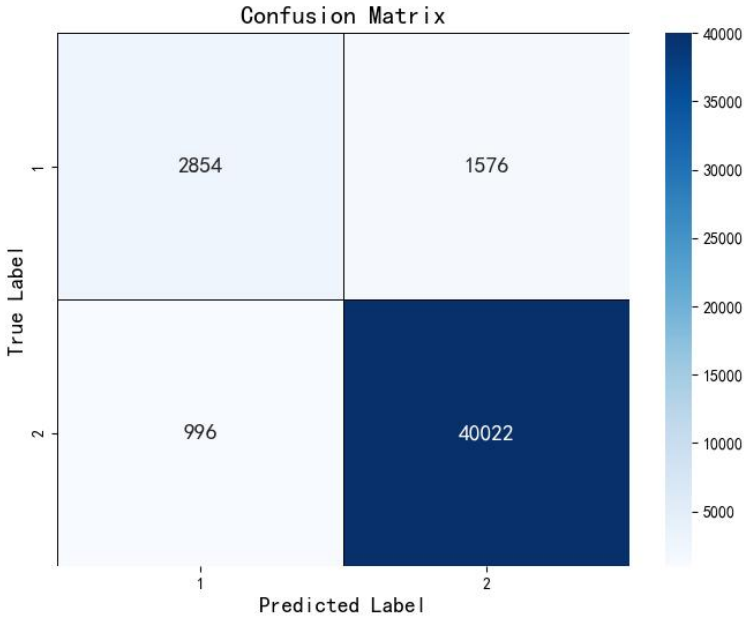

Figure 26. Confusion Matrix of Attention-ProteinMeNet using CB513

d. ROC Curve

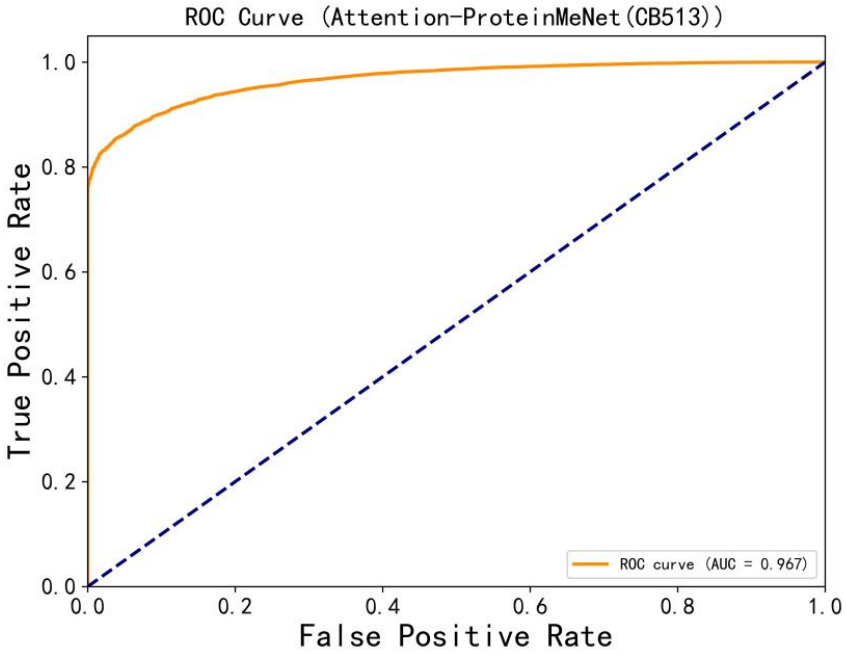

Figure 27. ROC Curve of Attention-ProteinMeNet using CB513(AUC = 0.976)

Figure 27 shows the ROC curve with an AUC of 0.967 on CB513, highlighting the model's strong ability to distinguish protein structures. The curve rises steeply and maintains a high true positive rate above 90% beyond a false positive rate of 0.2. Although performance is slightly

lower than on RCSB-PDB, likely due to CB513's greater sequence diversity, the model still shows strong generalization across datasets.

e. Precision_Recall Curve

Figure 28 shows the Precision-Recall curve on CB513 with AP=0.996, indicating high precision at all recall levels. The model has robust prediction and good cross-dataset generalization, with only a 0.3% performance drop compared to RCSB-PDB.

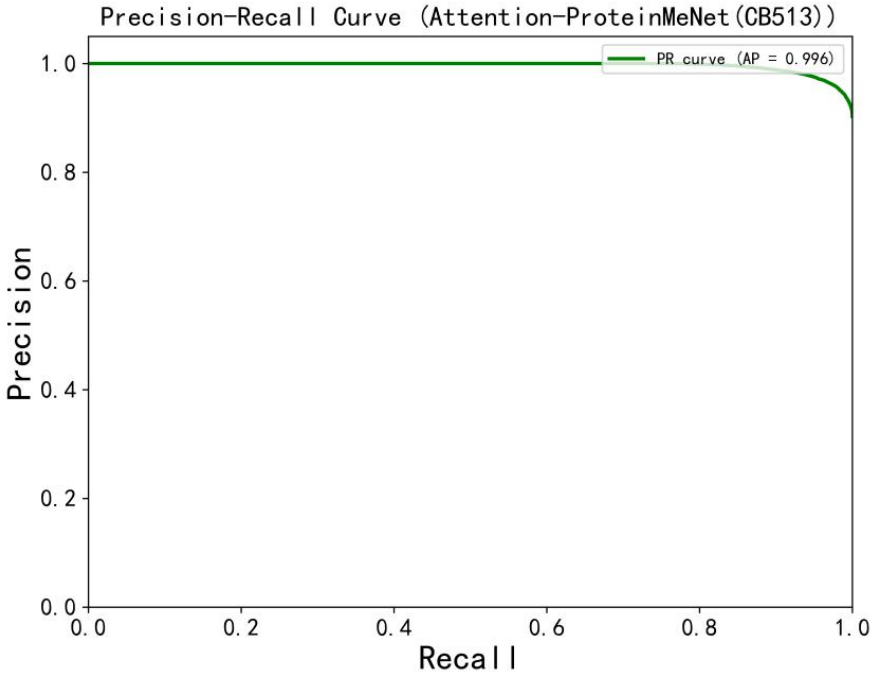

Figure 28. Precision-Recall Curve of Attention-ProteinMeNet using CB513(AP = 0.996)

4.2 Comparison with Other Models & Fine-tuning

The Attention-ProteinMeNet model is compared with other established models, highlighting their strengths and weaknesses. The section also covers fine-tuning steps used to optimize performance and improve prediction accuracy, starting with evaluation on the RCSB-PDB dataset.

4.2.1 Use Dataset RCSB-PDB

The RCSB-PDB dataset is employed to evaluate the performance of several models, including ProteinNet, BLSTM, and Attention-ProteinNet. Testing on this dataset assesses their ability to predict protein secondary structures across various sequences. Initially, the comparison is based on the RCSB-PDB dataset, followed by the CB513 dataset for further evaluation.

4.2.1.1 ProteinNet(RCSB-PDB)

Attention-ProteinMeNet integrates the feature extraction ability of ProteinNet, the sequence processing ability of BLSTM and the ability of Attention mechanism to focus on key information, thus achieving low training and verification loss and high accuracy.

a. Accuracy

ProteinNet shows steady performance with both training and validation accuracy curves converging, but it struggles with long-range dependencies. In contrast, Attention-ProteinMeNet exhibits smoother curves and better generalization, thanks to the attention mechanism that helps the model focus on relevant residues, improving its handling of complex dependencies. The accuracy curve of ProteinNet is shown in Figure 29.

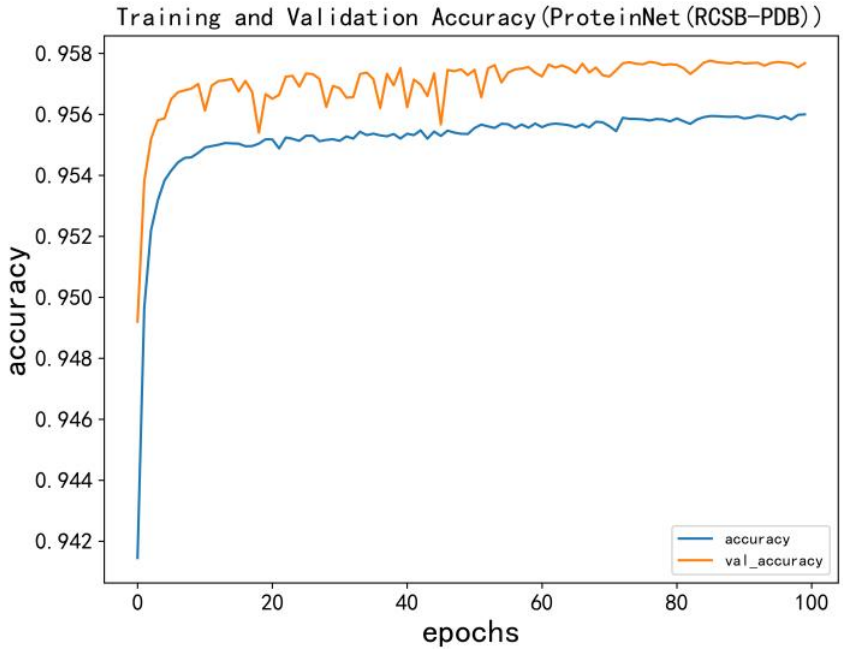

Figure 29. Accuracy_curve of ProteinNet using RCSB-PDB

(Train Acc = 0.9577, Val Acc = 0.9577)

b. Loss

Figure 30 shows that while ProteinNet converges steadily, Attention-ProteinMeNet trains faster and performs better overall. Both models avoid overfitting, but Attention-ProteinMeNet ' s combination of sequence modeling and attention helps it capture complex patterns more effectively, making it the better choice for protein structure prediction.

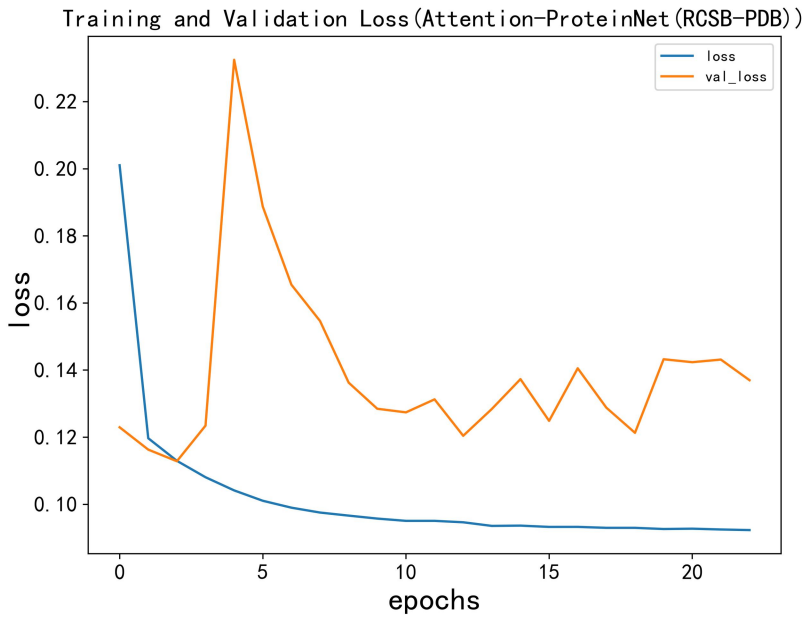

Figure 30. Loss_curve of ProteinNet using RCSB-PDB

(Train Loss = 0.0892, Val Loss = 0.0892)

c. Confusion Matrix

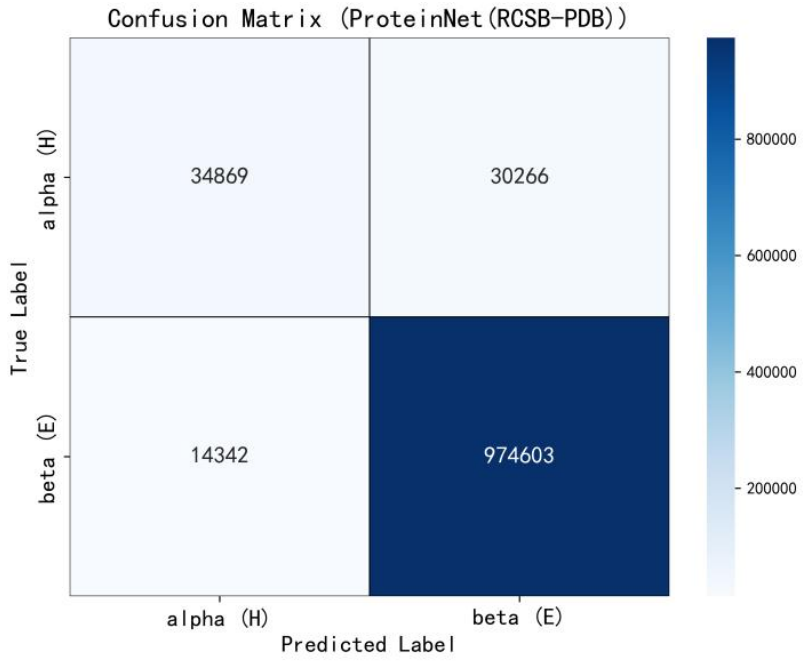

Figure 31. Confusion Matrix of ProteinNet using RCSB-PDB

ProteinNet shows good overall performance but struggles with distinguishing between alpha and beta structures, leading to higher misclassification, whereas Attention-ProteinMeNet improves this by leveraging an attention mechanism to better capture long-range dependencies and reduce errors. The confusion matrix of ProteinNet is shown in Figure 31.

d. ROC Curve

The ROC curve of ProteinNet is shown in Figure 32, which demonstrates excellent performance with an AUC value of 0.977. However, it performs poorly in distinguishing complex patterns. In contrast, Attention-ProteinMeNet improves the classification accuracy by leveraging the attention mechanism, achieving a higher AUC value of 0.983 and being better at handling long-range dependencies.

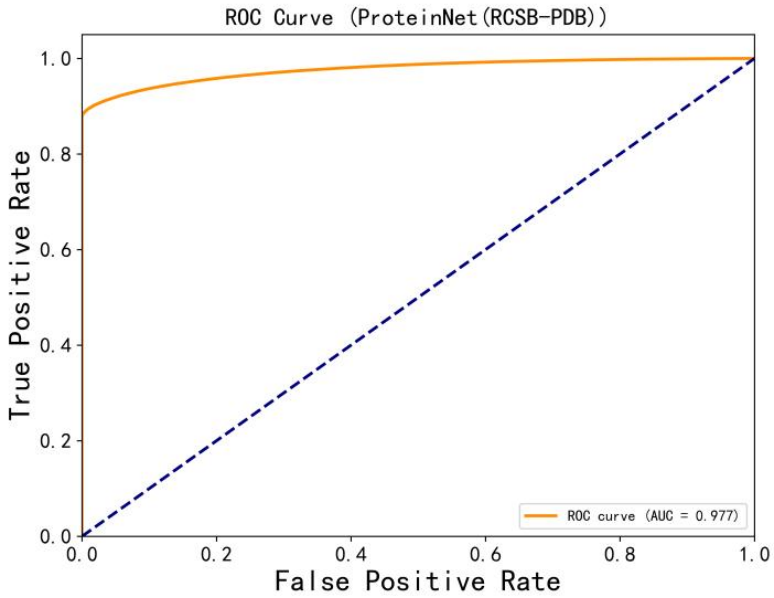

Figure 32. ROC Curve of ProteinNet using RCSB-PDB(AUC = 0.977)

e. Precision_Recall Curve

Figure 33 shows ProteinNet's high Precision-Recall score of 0.998, indicating strong precision in identifying positives. However, it struggles with subtle structural differences. In contrast, Attention-ProteinMeNet scores slightly higher at 0.999 and better captures long-range dependencies, offering improved generalization and accuracy on complex structures.

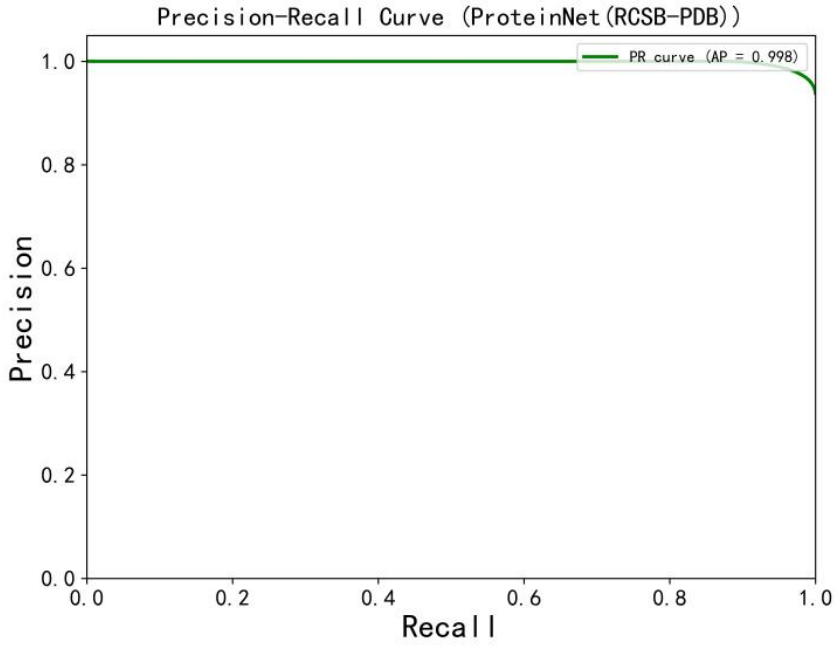

Figure33. Precision-Recall Curve of ProteinNet using RCSB-PDB(AP = 0.998)

4.2.1.2 BLSTM(RCSB-PDB)

a. Accuracy

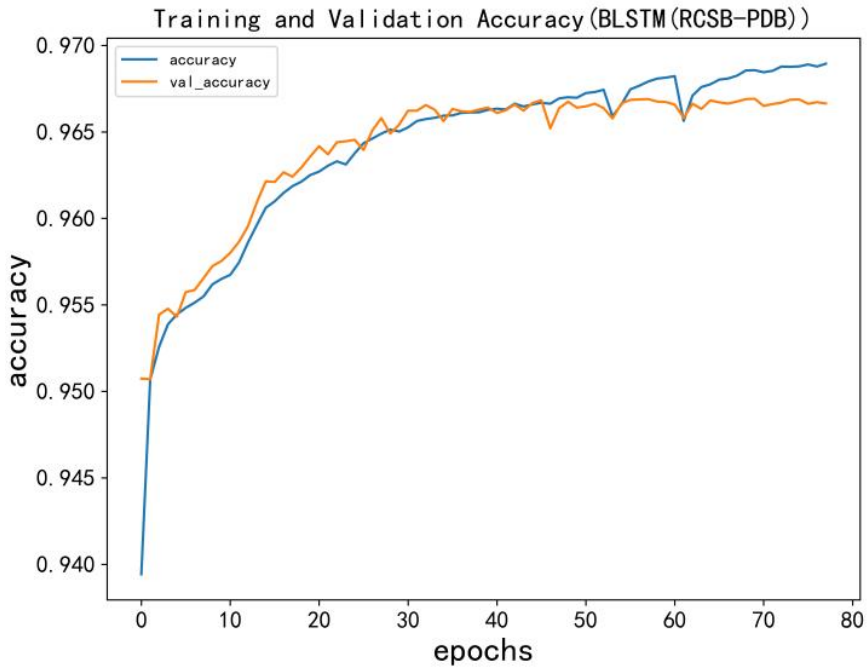

Figure 34. Accuracy_curve of BLSTM using RCSB-PDB

(Train Loss = 0.0746, Val Loss = 0.0752)

Figure 34 shows BLSTM's accuracy curve, with stable training and validation accuracy but limitations in capturing long-range dependencies. Compared to Attention-ProteinMeNet, BLSTM converges slower and has slightly lower accuracy, highlighting the benefits of attention mechanisms.

b. Loss

Figure 35 shows BLSTM's loss curve, with both training and validation losses steadily decreasing and stabilizing at 0.0746 and 0.0752, respectively, demonstrating good performance. However, it has difficulty modeling long-range dependencies, whereas Attention-ProteinMeNet improves upon this by integrating multiple components, enhancing feature extraction, generalization, and sequence modeling for better overall performance.

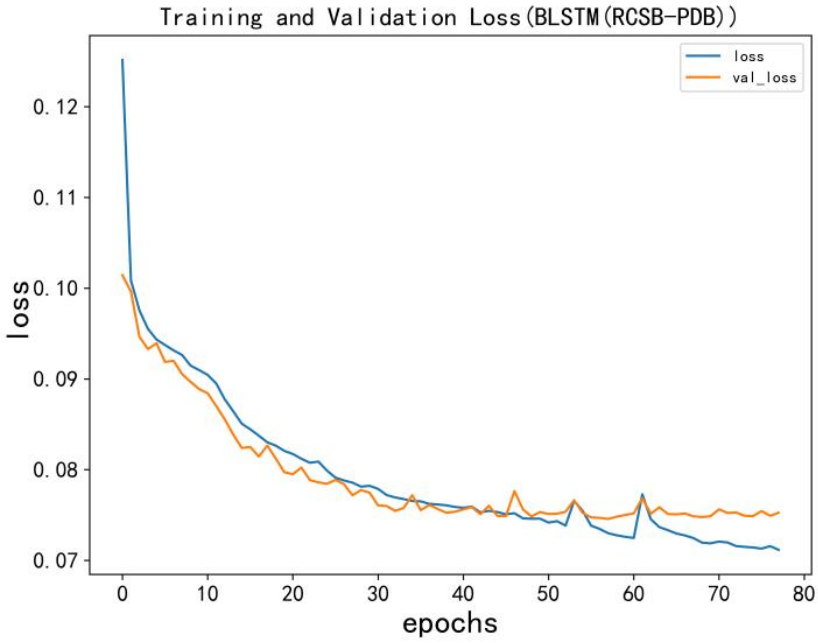

Figure 35. Loss _curve of BLSTM_model using RCSB-PDB

(Train Loss = 0.0746, Val Loss = 0.0752)

c. Confusion Matrix

Attention-ProteinMeNet shows improved performance over BLSTM, correctly predicting 976,166 beta structures with fewer misclassifications. It misclassifies only 12,779 beta structures as alpha and 24,209 alpha structures as beta. This enhancement reflects its better handling of complex sequence relationships and long-range dependencies, making it more effective in

distinguishing protein structure types compared to BLSTM. The confusion matrix of BLSTM is shown in Figure 36.

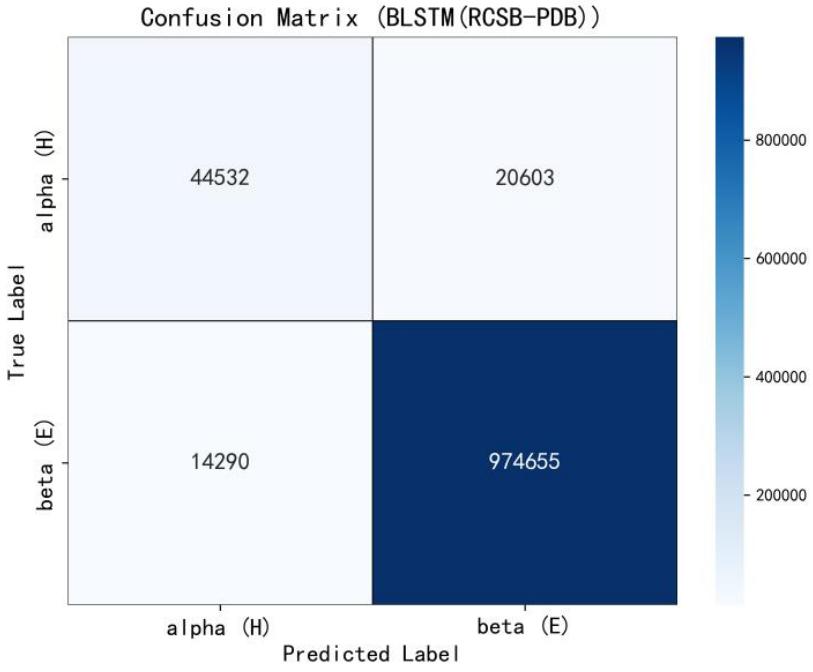

Figure 36. Confusion Matrix of BLSTM using RCSB-PDB

d. ROC Curve

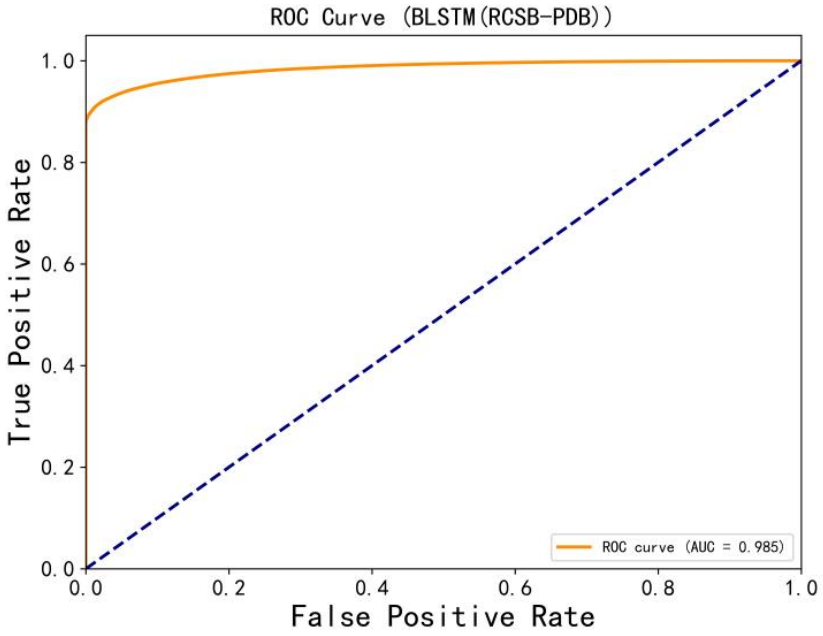

Figure 37. ROC Curve of BLSTM using RCSB-PDB(AUC = 0.985)

Figure 37 shows the ROC curve for BLSTM, with an AUC of 0.985, indicating strong performance in distinguishing protein structures. The curve rises steadily, suggesting BLSTM is effective but slightly less efficient at handling complex dependencies compared to other models. This limits its ability to capture intricate relationships in the data.

e. Precision_Recall Curve

For BLSTM and Attention-ProteinMeNet, both achieving an AP of 0.999. BLSTM shows high precision but a slightly less responsive recall rate, indicating limited adaptability. Attention-ProteinMeNet performs similarly but handles complex dependencies more effectively, giving it a slight advantage. The precision_recall of BLSTM is shown in Figure 38.

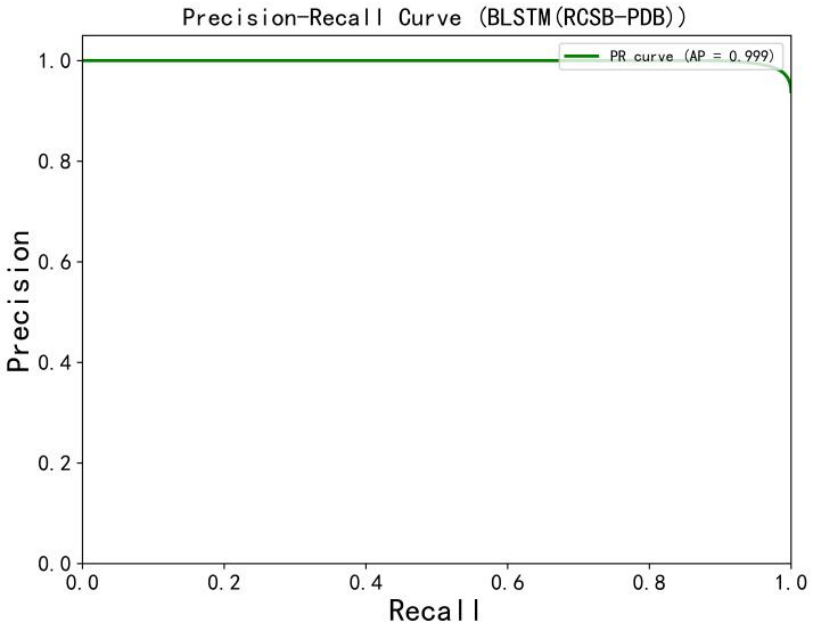

Figure 38. Precision-Recall Curve of BLSTM using RCSB-PDB(AP = 0.999)

4.2.1.3 Attention-ProteinNet(RCSB-PDB)

Attention-ProteinMeNet integrates the feature extraction ability of ProteinNet, the sequence processing ability of BLSTM and the ability of Attention mechanism to focus on key information, thus achieving low training and verification loss and high accuracy.

a. Accuracy

Figure 39 shows the accuracy curve for Attention-ProteinMeNet on the RCSB-PDB dataset. The model achieves a training accuracy of 95.37% and a validation accuracy of 95.90%. Both the training and validation accuracy curves stabilize after a few epochs, indicating effective learning

and strong generalization. The curves remain close to each other, suggesting minimal overfitting and consistent performance throughout training. Therefore, it can be concluded that the Attention-ProteinMeNet Model works best at predicting protein structures, possibly due to its ability to synthesize different types of neural network layers to capture complex patterns and relationships in the data.

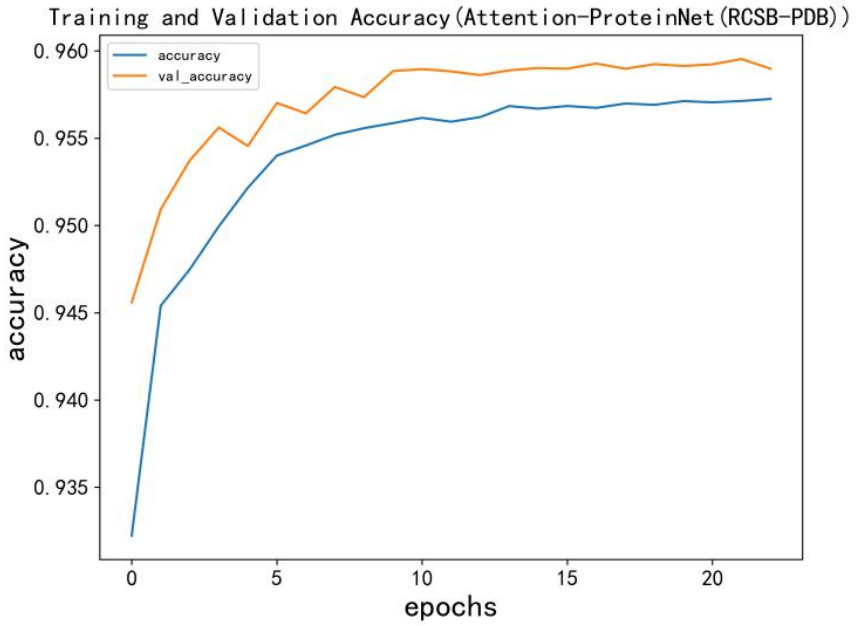

Figure39. Accuracy_curve of Attention-ProteinNet using RCSB-PDB

(Train Acc = 0.9537, Val Acc = 0.9590)

b. Loss

Figure 40 shows that Attention-ProteinMeNet has smoother loss curves and more stable accuracy than other models, including Attention-ProteinNet, which shows more fluctuation. This stability is due to Attention-ProteinMeNet's ability to capture both local features and long-range dependencies, leading to better overall performance.

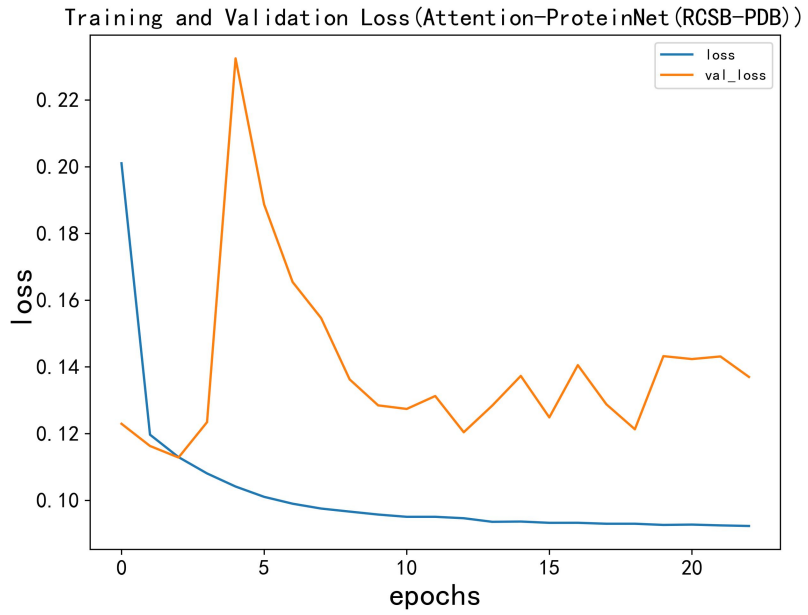

Figure 40. Loss_curve of Attention-ProteinNet using RCSB-PDB

(Train Loss = 0.1128, Val Loss = 0.1370)

c. Confusion Matrix

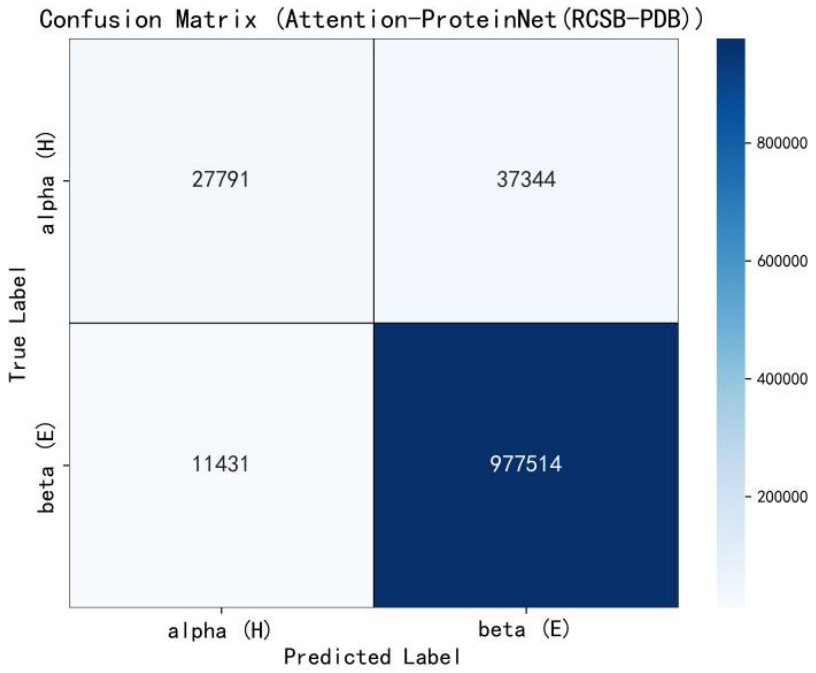

Figure 41. Confusion Matrix of Attention-ProteinNet using RCSB-PDB

In Figure 41, Attention-ProteinMeNet performs well, correctly predicting most beta structures with 977,514 true positives, while misclassifying 11,431 beta structures as alpha. There are

37,344 misclassifications of alpha structures as beta, indicating some difficulty in distinguishing between the two classes. Overall, Attention-ProteinMeNet demonstrates strong classification performance, with high true positives for both alpha and beta structures while minimizing misclassifications, making it more robust in distinguishing protein types.

d. ROC Curve

Figure 42 presents the ROC curve for Attention-ProteinMeNet on the RCSB-PDB dataset, achieving an AUC of 0.973. The curve rises sharply at the start, showing the model's strong ability to distinguish between classes with high true positive rates. While performance is excellent, the presence of some false positives suggests room for further improvement. Overall, the model demonstrates robust and reliable classification performance.

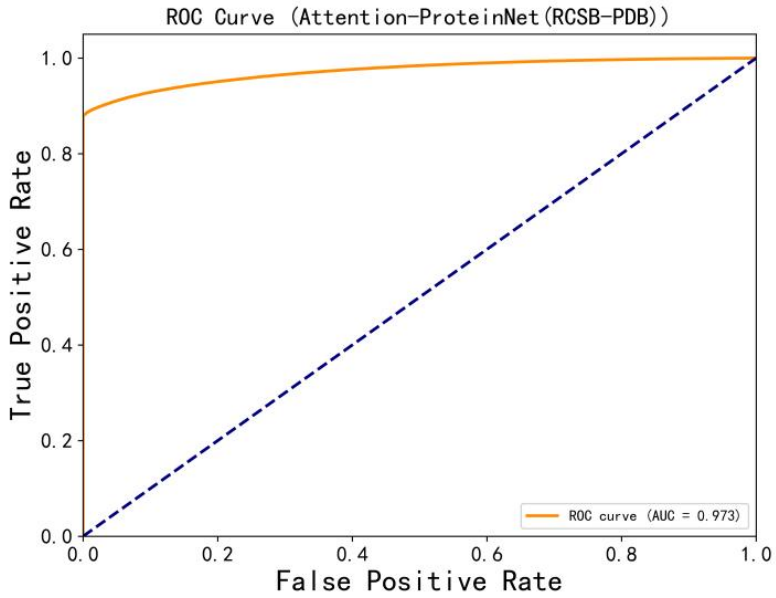

Figure 42. ROC Curve of Attention-ProteinNet using RCSB-PDB(AUC=0.973)

e. Precision_Recall Curve

Figure 43 shows the precision-recall curve for Attention-ProteinMeNet on the RCSB-PDB dataset, with an area under the curve (AP) of 0.998. The curve reaches near-perfect precision at very high recall, indicating that the model excels at correctly identifying protein structures with minimal misclassifications. Despite high precision and recall, the model may still face challenges in handling imbalanced datasets, where less frequent classes could impact its ability to generalize effectively.

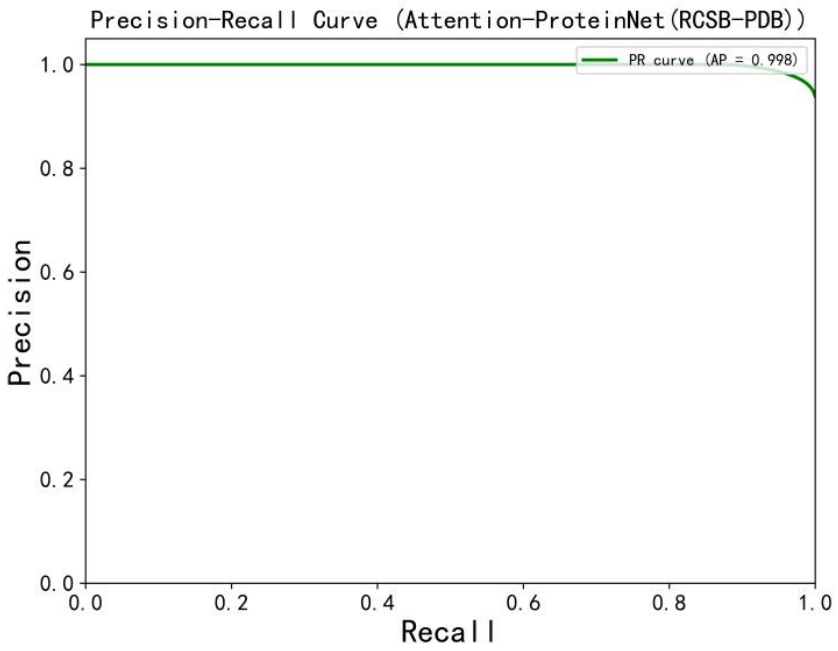

Figure 43. Precision-Recall Curve of Attention-ProteinNet using RCSB-PDB(AP = 0.998)

4.2.1.4 Summary(RCSB-PDB)

In addition, the table below provides more detailed evaluation parameters to compare performance between the four models:

Table 6. Different Model Performance Comparison using RCSB-PDB

	Models	Loss	Accuracy	Precision	Recall	F1-score
Dataset1 (RCSB-PDB)	Attention-ProteinMeNet	0.0778	0.9663	0.9663	0.9663	0.9633
	ProteinNet	0.0892	0.9577	0.9577	0.9577	0.9549
	BLSTM	0.0746	0.9669	0.9669	0.9669	0.9661
	Attention-ProteinNet	0.1128	0.9537	0.9537	0.9537	0.9483

Table 6 shows that Attention-ProteinMeNet's performance is obviously superior to other models. This indicates that Attention-ProteinMeNet achieves better performance in protein structure prediction tasks by combining the feature extraction ability of ProteinNet, the sequence processing advantage of BiLSTM, and the ability to focus key information of Attention.

4.2.2 Use Dataset CB513

The following subsections evaluate the performance of the models using the CB513 dataset, focusing on their ability to predict protein secondary structures. The results of different models are discussed, beginning with ProteinNet.

4.2.2.1 ProteinNet(CB513)

The performance of ProteinNet on the CB513 dataset is analyzed, highlighting its accuracy in predicting protein secondary structures. This provides a baseline for comparing its effectiveness against other models.

a. Accuracy

Figure 44 shows the training and validation accuracy curves for ProteinNet using the CB513 dataset, with training accuracy at 92.65% and validation accuracy at 93.01%. While the model performs well, slight fluctuations in the validation curve suggest some overfitting, indicating potential struggles with generalization on unseen data.

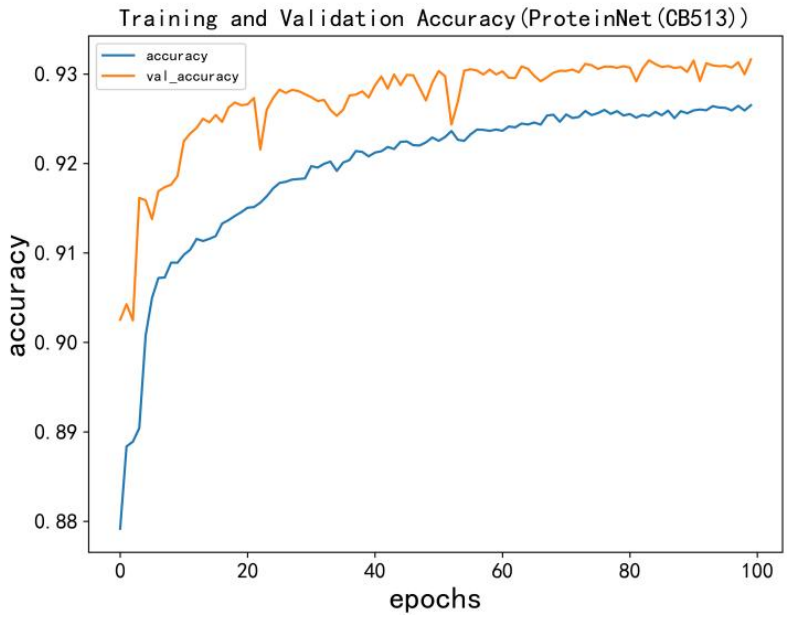

Figure 44. Accuracy_curve of ProteinNet using CB513

(Train Acc = 0.9265, Val Acc = 0.9301)

b. Loss

During the training process, the loss of ProteinNet model decreases rapidly and the accuracy rate increases rapidly, which indicates that ProteinNet performs well in feature extraction.

However, the loss and accuracy of the validation set fluctuate greatly, which may indicate that the model has insufficient generalization ability.

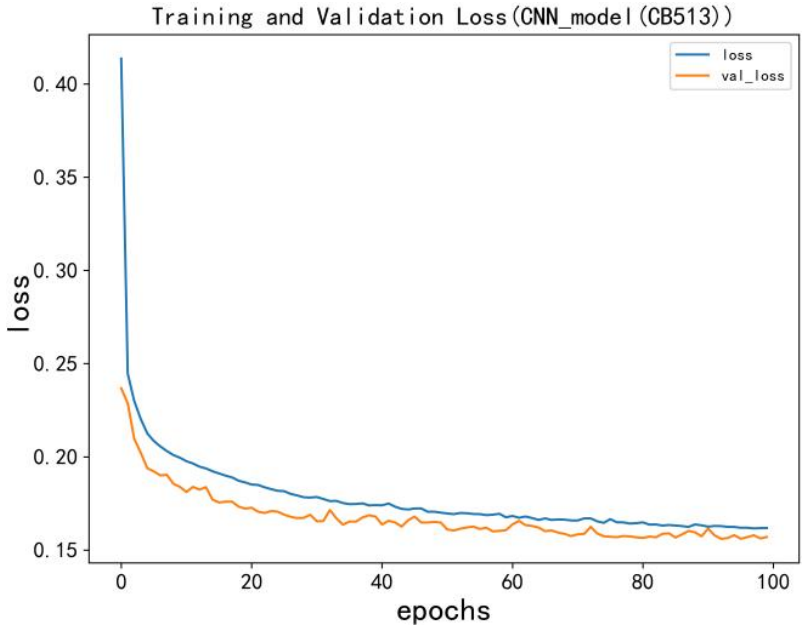

Figure 45. Loss_curve of ProteinNet using CB513

(Train Loss = 0.1618, Val Loss = 0.1570)

c. Confusion Matrix

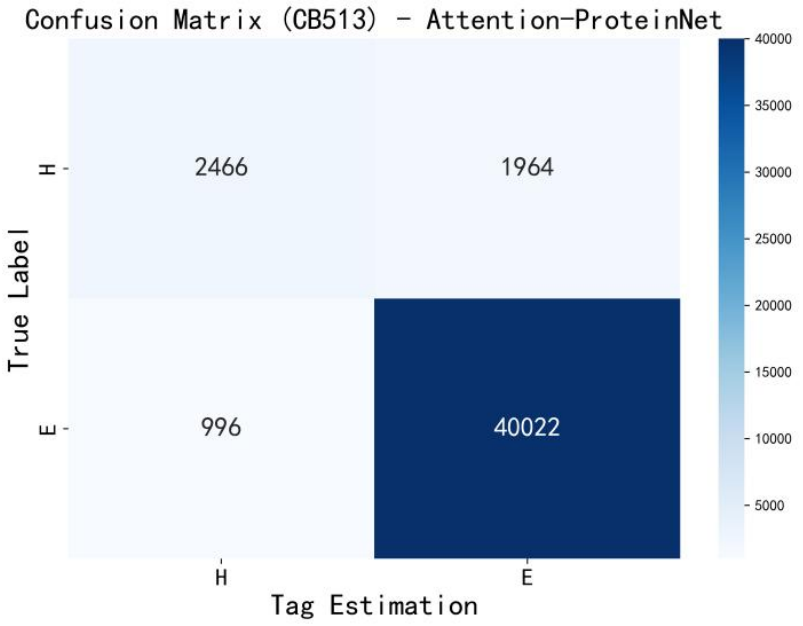

Figure 46. Confusion Matrix of ProteinNet using CB513

Figure 46 shows the confusion matrix for ProteinNet using the CB513 dataset. The model correctly predicts most beta structures, but misclassifies 996 beta as alpha and 2,466 alpha as beta, indicating some difficulty in distinguishing between the two classes. Limitation: The model struggles with correctly classifying alpha and beta structures, especially in complex sequences.

d. ROC Curve

Figure 47 shows ProteinNet's ROC curve on the CB513 dataset with an AUC of 0.956, indicating strong performance and good class separation. However, a slight gap in the curve suggests potential to reduce false positives and improve generalization.

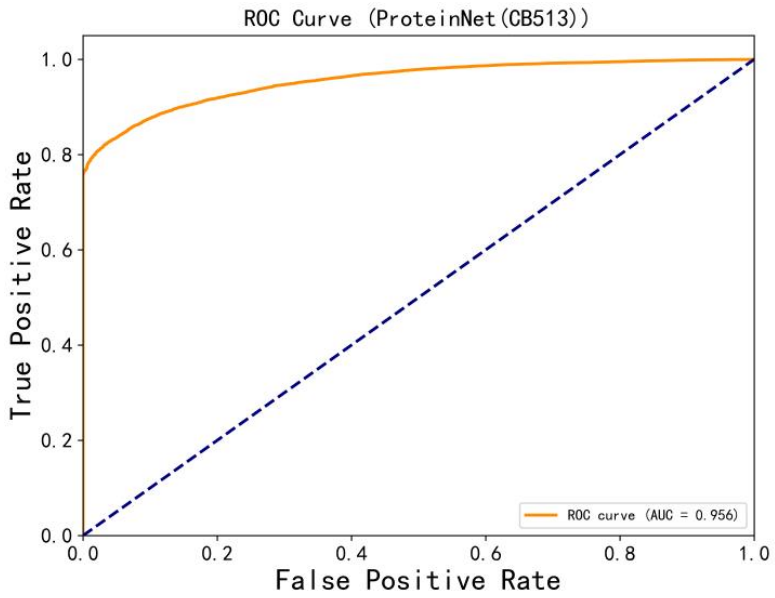

Figure 47. ROC Curve of ProteinNet using CB513(AUC = 0.965)

e. Precision_Recall Curve

Figure 48 shows that Attention-ProteinMeNet's Precision-Recall curve is closer to the upper left corner, indicating higher precision across recall levels and better performance than ProteinNet. Its smoother curve suggests more stable and reliable results across thresholds.

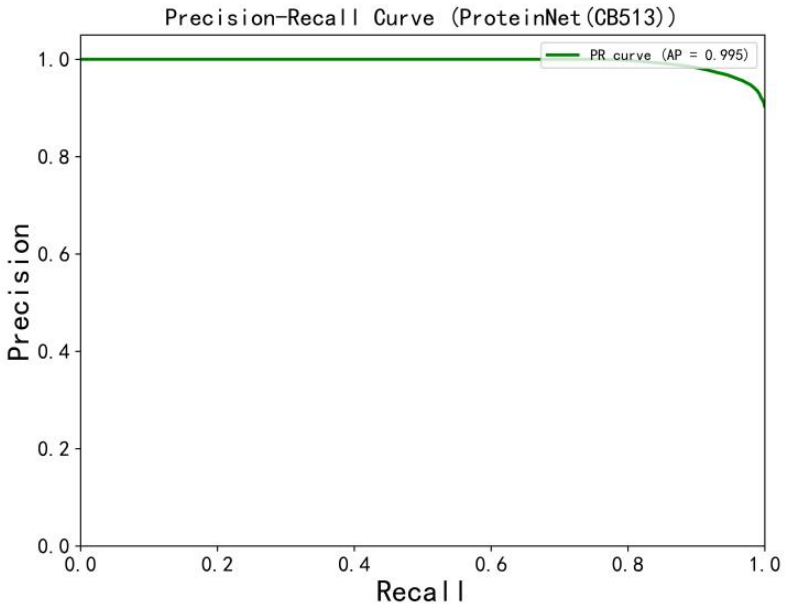

Figure 48. Precision-Recall Curve of ProteinNet using CB513(AP = 0.995)

4.2.2.2 BLSTM(CB513)

a. Accuracy

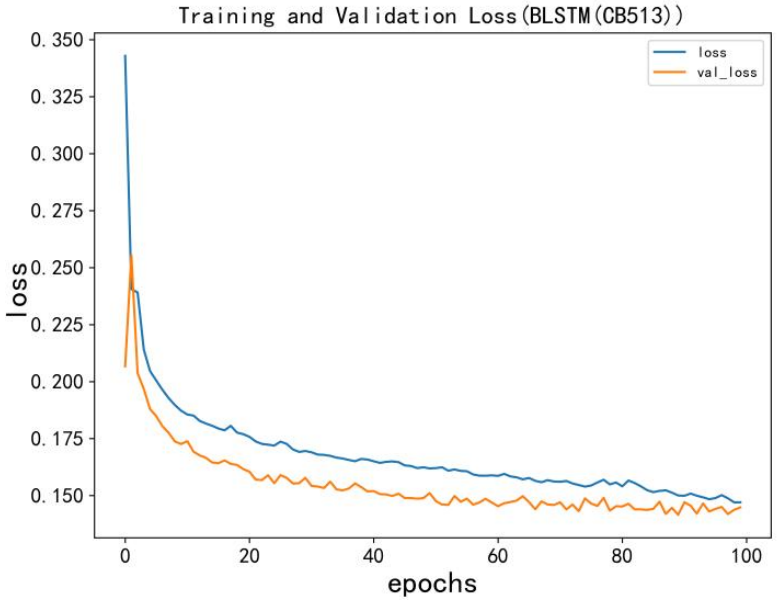

Figure 49. Accuracy_curve of BLSTM using CB513

(Train Acc = 0.9405, Val Acc = 0.9323)

Figure 49 shows the training and validation loss curves for BLSTM using the CB513 dataset, with a training accuracy of 94.05% and a validation accuracy of 93.23%. Both losses decrease

steadily, but the gap between them suggests some overfitting, indicating the model may not generalize well to new data.

b. Loss

Figure 50 shows BLSTM's training and validation accuracy on the CB513 dataset, with losses of 0.1364 and 0.1511, respectively. Accuracy stabilizes after a few epochs, but a slight gap indicates minor overfitting and limited generalization to unseen data.

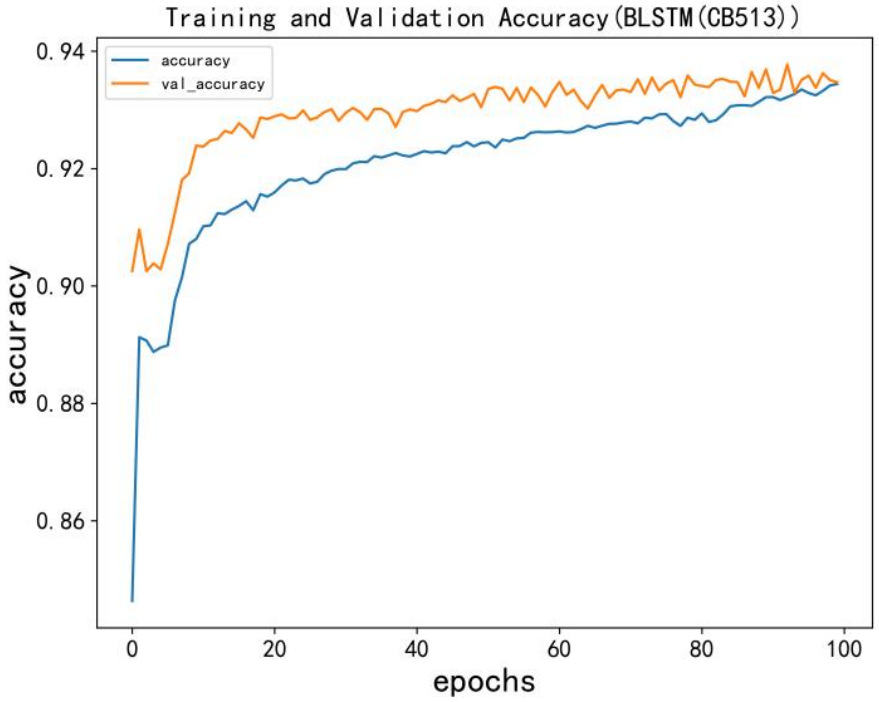

Figure 50. Loss_curve of BLSTM using CB513

(Train Loss = 0.1364, Val Loss = 0.1511)

c. Confusion Matrix

Figure 51 shows BLSTM's confusion matrix on the CB513 dataset. While it correctly predicts most beta structures, it often confuses alpha and beta, indicating difficulty in distinguishing these classes, especially in complex sequences.

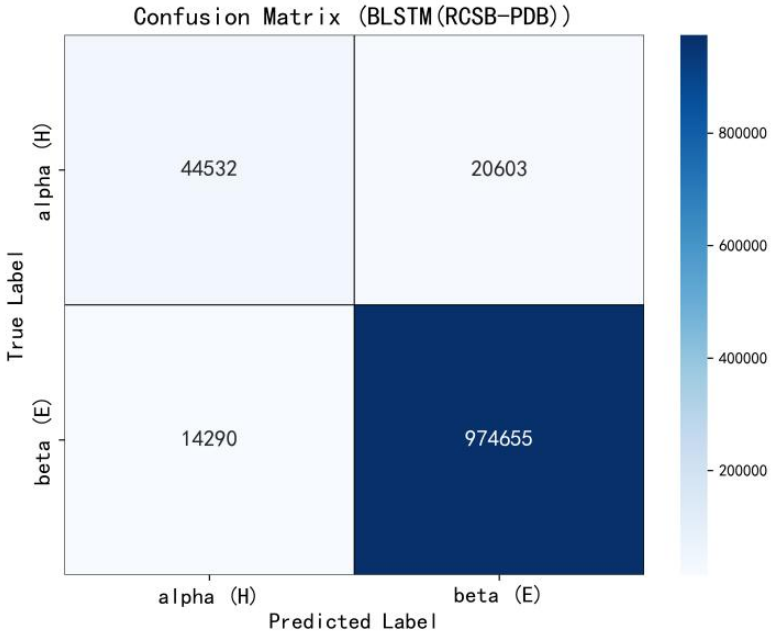

Figure 51. Confusion Matrix of BLSTM using CB513

d. ROC Curve

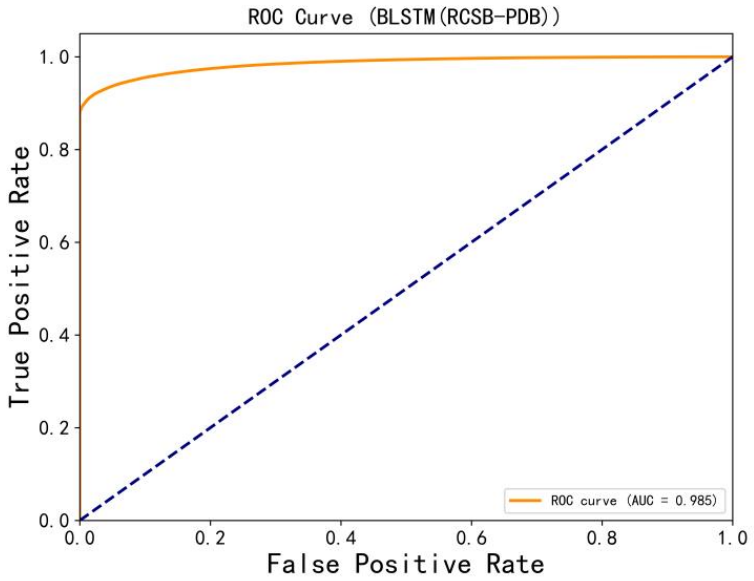

Figure 52. ROC Curve of BLSTM using CB513(AUC = 0.985)

Figure 52 shows the ROC curve for BLSTM with an AUC of 0.985, indicating strong performance. However, the slight gap suggests room for improvement in reducing false

positives. Limitation: The model could benefit from enhanced sensitivity and fewer false positives.

e. Precision_Recall Curve

Figures 53 show the precision-recall curves for BLSTM on the CB513 dataset. While BLSTM achieves an AP of 0.999, Attention-ProteinMeNet performs slightly better with an AP of 0.996. This suggests Attention-ProteinMeNet excels at capturing more complex dependencies in protein sequences, making it a more robust model compared to BLSTM.

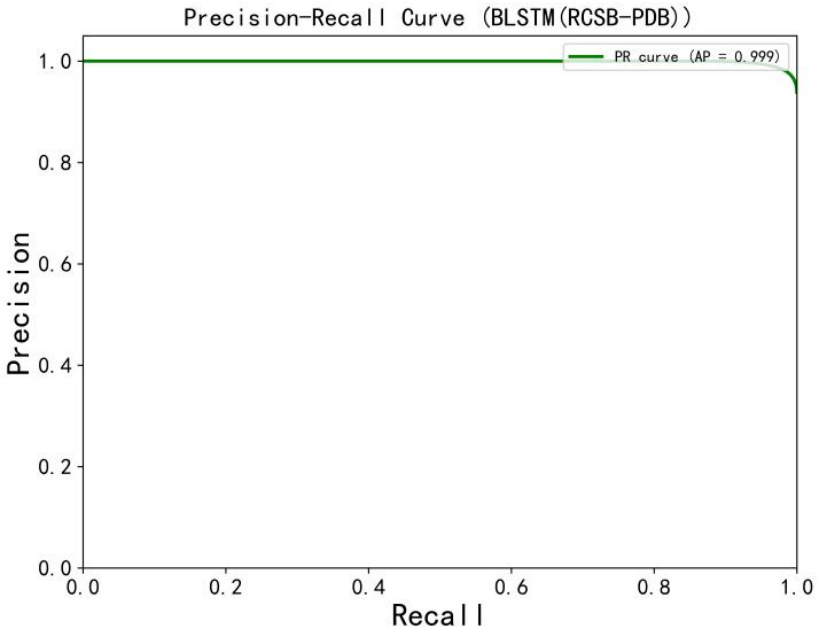

Figure 53. Precision-Recall Curve of BLSTM using CB513(AP = 0.999)

4.2.2.3 Attention-ProteinNet(CB513)

a. Accuracy

Figure 54 shows Attention-ProteinMeNet's training accuracy at 92.71% and validation accuracy at 93.61% on the CB513 dataset. Both curves rise steadily, though early fluctuations in validation accuracy suggest slight instability during initial training.

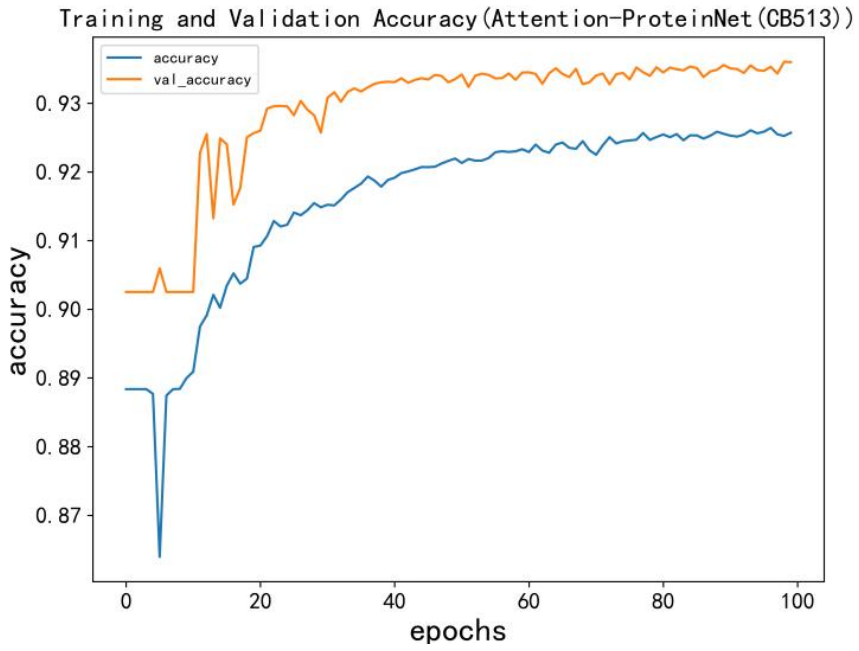

Figure 54. Accuracy_curve of Attention-ProteinNet using CB513

(Train Acc = 0.9271, Val Acc = 0.9361)

b. Loss

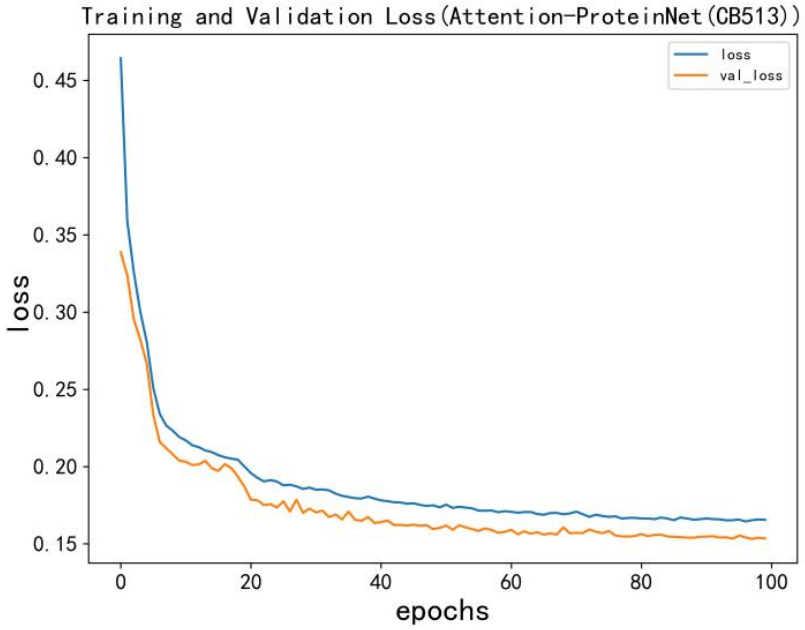

Figure 55. Loss_curve of Attention-ProteinNet using CB513

(Train Loss = 0.1638, Val Loss = 0.1536)

The loss decline trend of Attention-ProteinNet is similar to that of BLSTM, the accuracy is improved significantly, and the performance on the training and verification sets is consistent, which indicates that the attention mechanism helps the model to better focus on key features and improve generalization ability.

c. Confusion Matrix

Figure 56 shows the confusion matrix for Attention-ProteinMeNet using the CB513 dataset. The model correctly predicts 39,884 beta structures, but misclassifies 1,134 beta structures as alpha and 1,999 alpha structures as beta, indicating some difficulty in distinguishing between the two.

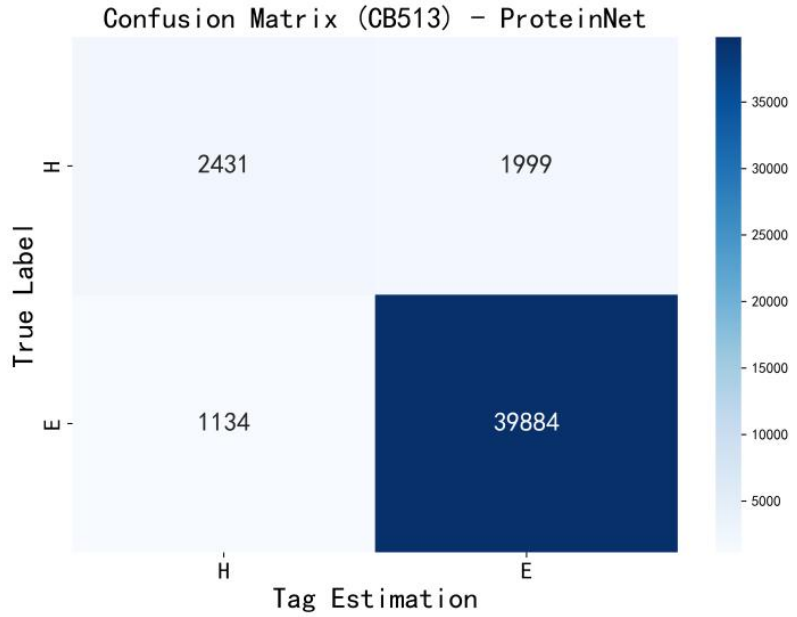

Figure 56. Confusion Matrix of Attention-ProteinNet using CB513

d. ROC Curve

Figure 57 shows that Attention-ProteinMeNet outperforms single models, with a curve closer to the upper left corner. It maintains a high true positive rate, especially at low false positive rates, due to its combined convolutional and sequential feature learning.

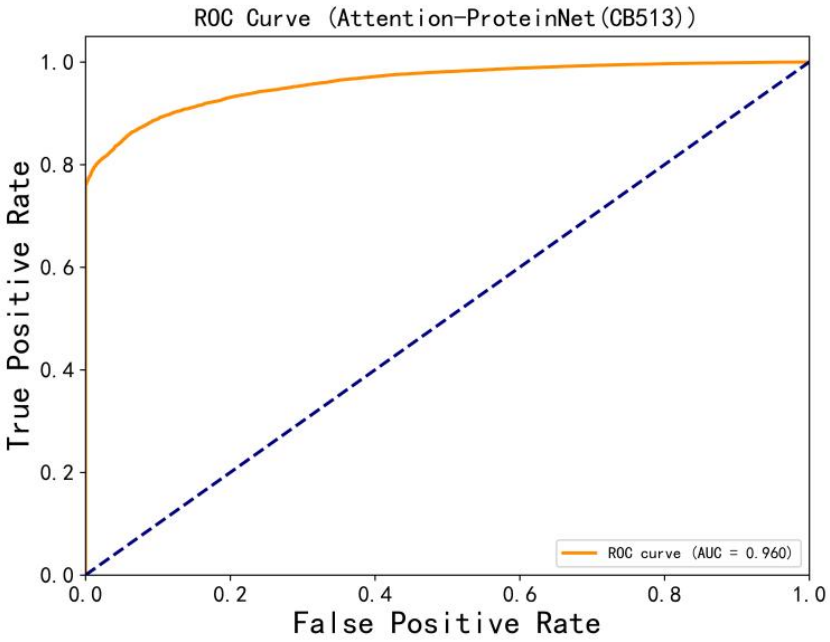

Figure 57. ROC Curve of Attention-ProteinNet using CB513(AUC = 0.960)

e. Precision_Recall Curve

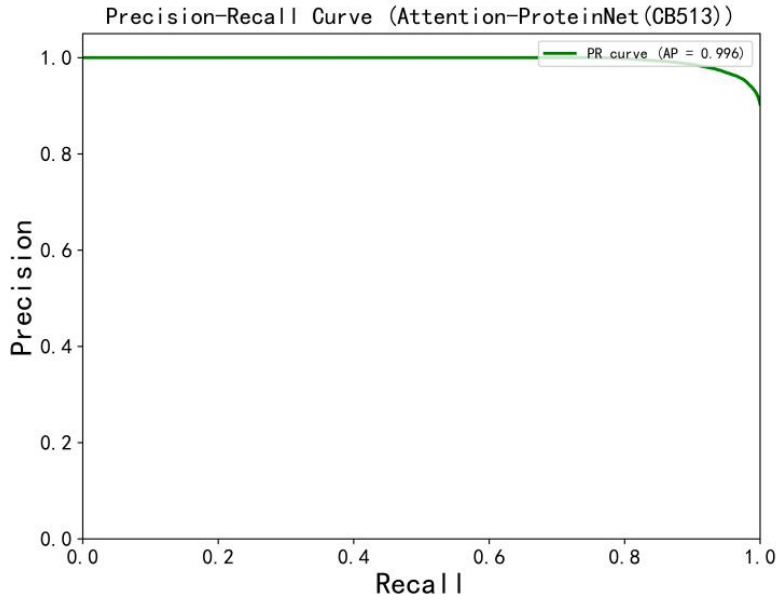

Figure 58. Precision-Recall Curve of Attention-ProteinNet using CB513(AP = 0.996)

Figure 58 shows the precision-recall curve for Attention-ProteinMeNet using the CB513 dataset, with an AP of 0.996. The curve indicates high precision and recall, demonstrating the model's strong ability to identify protein structures with minimal misclassification. Despite the high AP,

the model's performance could be further improved in handling complex or overlapping sequences.

4.2.2.3 Summary(CB513)

In addition, Table 7 provides more detailed evaluation parameters to compare performance between the four models

Table 7. Different Model Performance Comparison using CB513

	Models	Loss	Accuracy	Precision	Recall	F1-score
Dataset2 CB513	Attention-ProteinMeNet	0.1306	0.9434	0.9436	0.9378	0.9401
	ProteinNet	0.1570	0.9301	0.9269	0.9310	0.9285
	BLSTM	0.1452	0.9349	0.9319	0.9367	0.9333
	Attention-ProteinNet	0.1579	0.9440	0.9312	0.9352	0.9327

The Attention-ProteinMeNet outperforms the individual ProteinNet, BLSTM, and Attention-ProteinNet models across all performance metrics on the Dataset2 CB513.

The lower loss and higher accuracy, precision, and recall of the Attention-ProteinMeNet indicate that the integration of ProteinNet, BLSTM, and Attention mechanisms provides a more robust and accurate prediction of protein structures. The higher F1-score of the Attention-ProteinMeNet further supports its effectiveness, as it provides a better balance between precision.

4.3 Explainable Artificial Intelligence(XAI)

As shown in Figure 59, the model shows significant feature selectivity. The first 20 features contribute the main signals of the prediction. Among them, Feature_104 (SHAP=0.02) and Feature_68 (SHAP=0.01) dominate the field. The feature effects are bidirectional (for example, Feature_68 promotes β-folding and Feature_104 suppresses β-folding), which is consistent with the biological nature of protein conformation competition. It is worth noting that the long tail effect is obvious. A large number of low-contribution features, such as Feature_3 and Feature_250, have weak cumulative effects, indicating that the model can filter noise effectively.

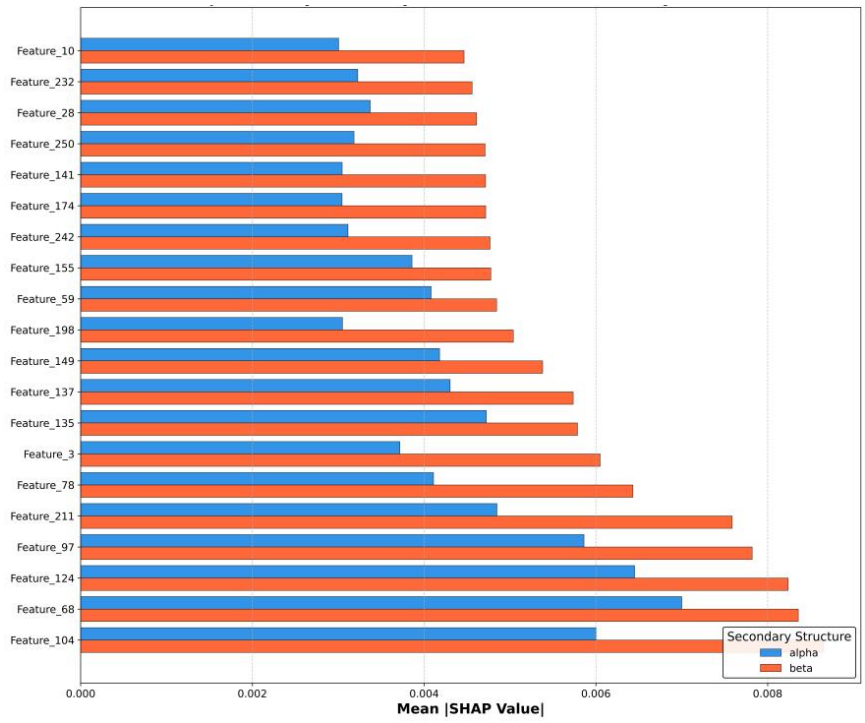

Figure 59. SHAP Bar_plot_Attention-ProteinMeNet using CB513

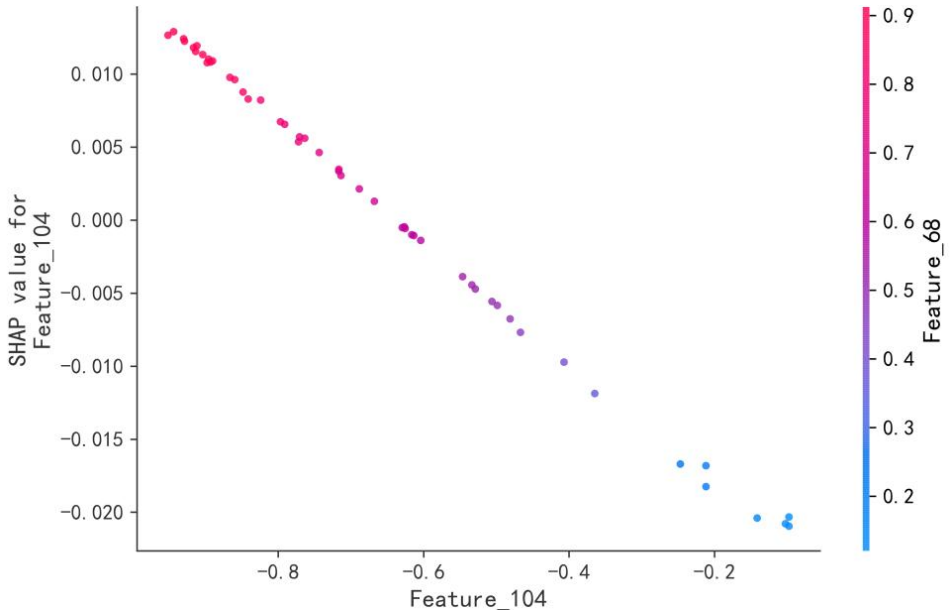

Figure 60. SHAP Dependence_plot_Attention-ProteinMeNet using CB513

The SHAp-dependence plot analysis in Figure 60 reveals the mechanism of key features in the prediction of protein secondary structure. Feature_104 shows a threshold effect in the interval [-0.8,-0.4]. A sharp increase in the SHAP value may trigger the conformational transformation of

α-helical to β-fold. The bimodal distribution of the data suggests that there are two conformational subgroups of natural state and intermediate state, which is consistent with the cooperative transformation theory of protein folding: the threshold effect corresponds to the critical point of conformational transformation, and the bimodal distribution reflects the energy barrier of the folding path, providing computational evidence for understanding the dynamic conformation of proteins.

As shown in Figure 61, the SHAP diagram confirms the model's strong performance in protein structure prediction. It accurately identifies key residues like Feature_68 and Feature_104 and captures subtle feature differences. The model balances high accuracy with interpretability, aligning well with biophysical principles.

Figure 61. SHAP Force_plot_Attention-ProteinMeNet using CB513

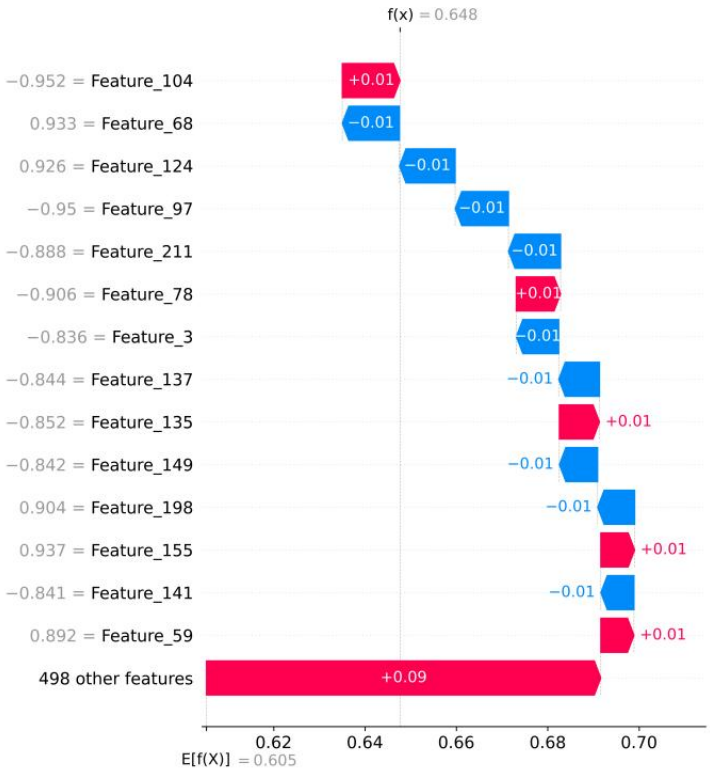

Figure 62. SHAP Waterfall_plot for Sample 47(Class:E)_Attention-ProteinMeNet using CB513

As shown in Figure 62, the category E prediction result of sample 47 (0.68) is mainly determined by the antagonistic action of Feature_104 (-0.952) and Feature_68 (+0.933). Other

features contribute little. The cumulative effect of 498 secondary features is only 0.005, indicating that the model can effectively identify key structural signals. The results reflect the competitive interactions between residues in the process of protein conformational transformation, and verify the reliability of the model prediction.

Figure 63 shows a SHAP-based interpretation of Attention-ProteinMeNet on the CB513 dataset. The summary plot highlights Feature_104 and Feature_68 as the most influential, with others like Feature_124 and Feature_97 also contributing. This analysis helps explain model decisions, identifies key features, and improves model transparency and reliability, aiding further optimization.

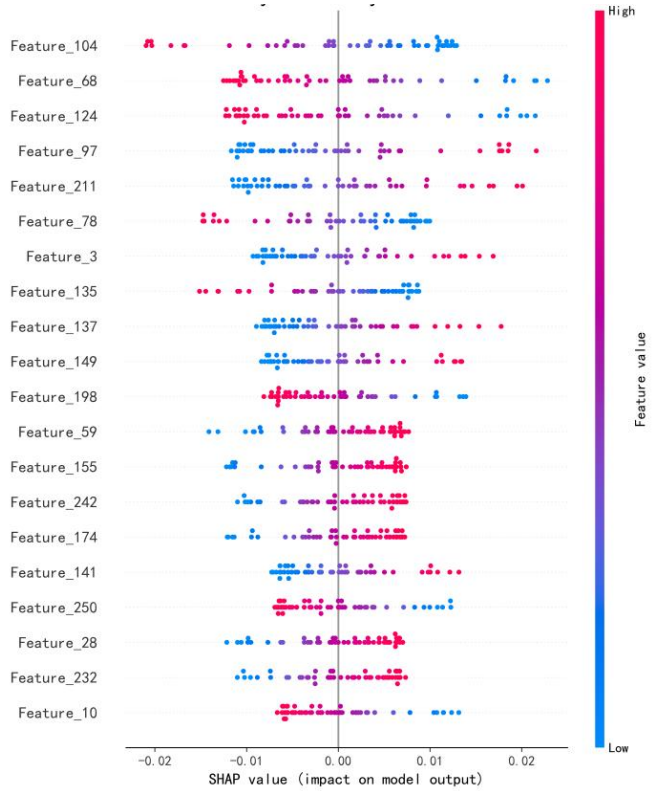

Figure 63. SHAP Summary_plot_Attention-ProteinMeNet using CB513

4.4 GUI Demonstration

This project designed and implemented an interactive website for protein structure prediction based on deep learning, aiming to improve the user experience and demonstrate the application of protein structure prediction. As shown in the Figure 64, this website offers two main functions: helical structure prediction for individual protein sequences and prediction for entire data sets.

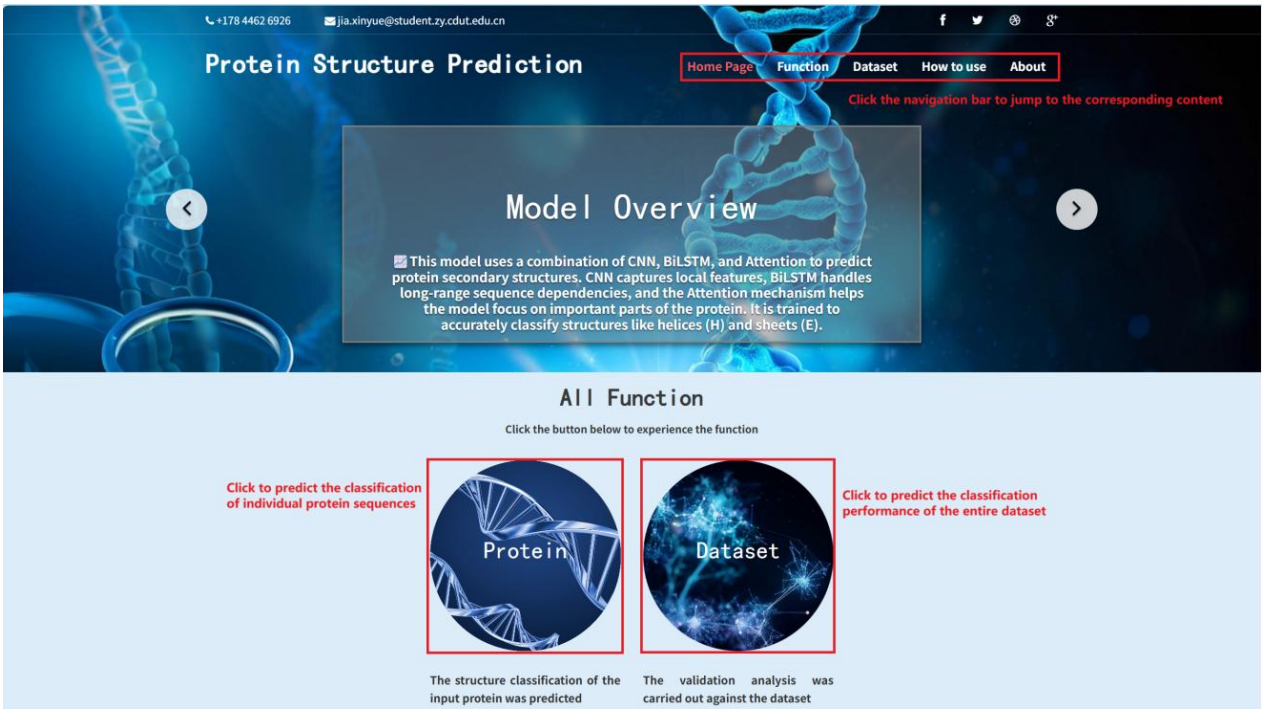

Figure 64. GUI_Home Page

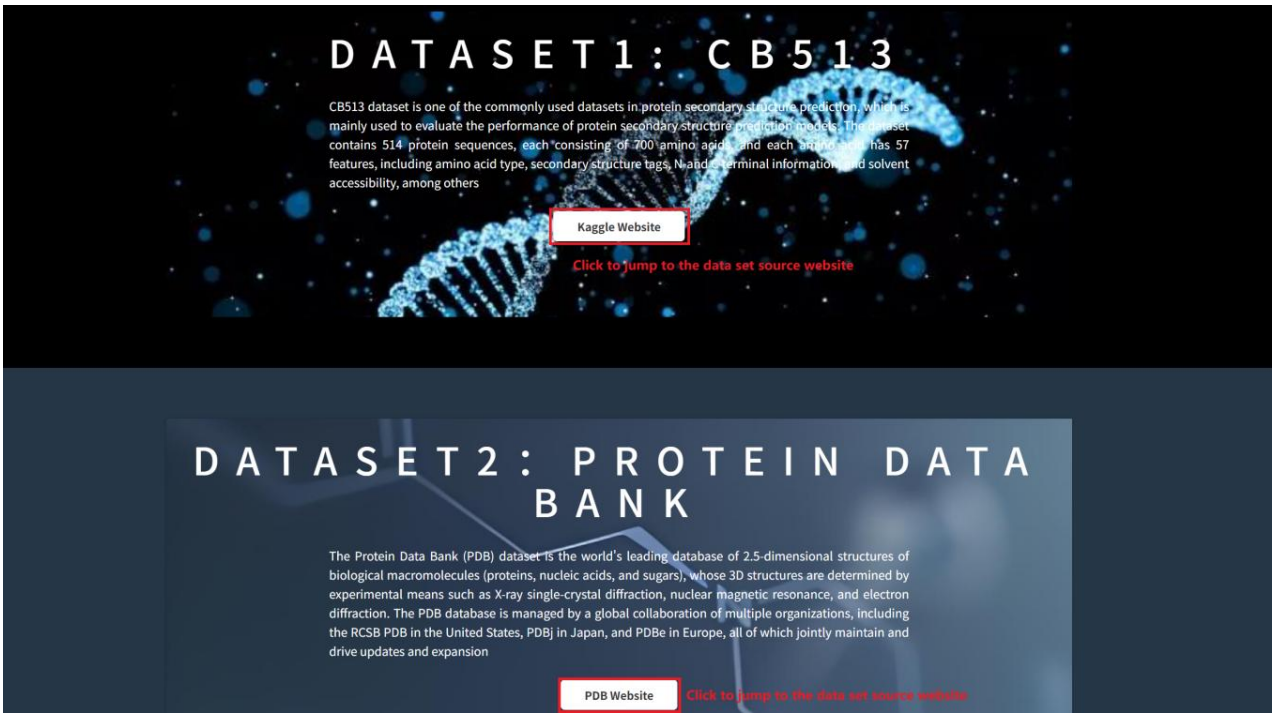

Figure 65. GUI_Dataset

On the home page, users can find basic details and links to the datasets used in this project through the navigation bar, with links to their Kaggle sources and PDB sources. This link takes

the user to the detailed data source shown in Figures 65 and 66.

Click on 'How to use' to see how to have a good interaction on the site. In addition, there are descriptions of model performance to help better understand the model.

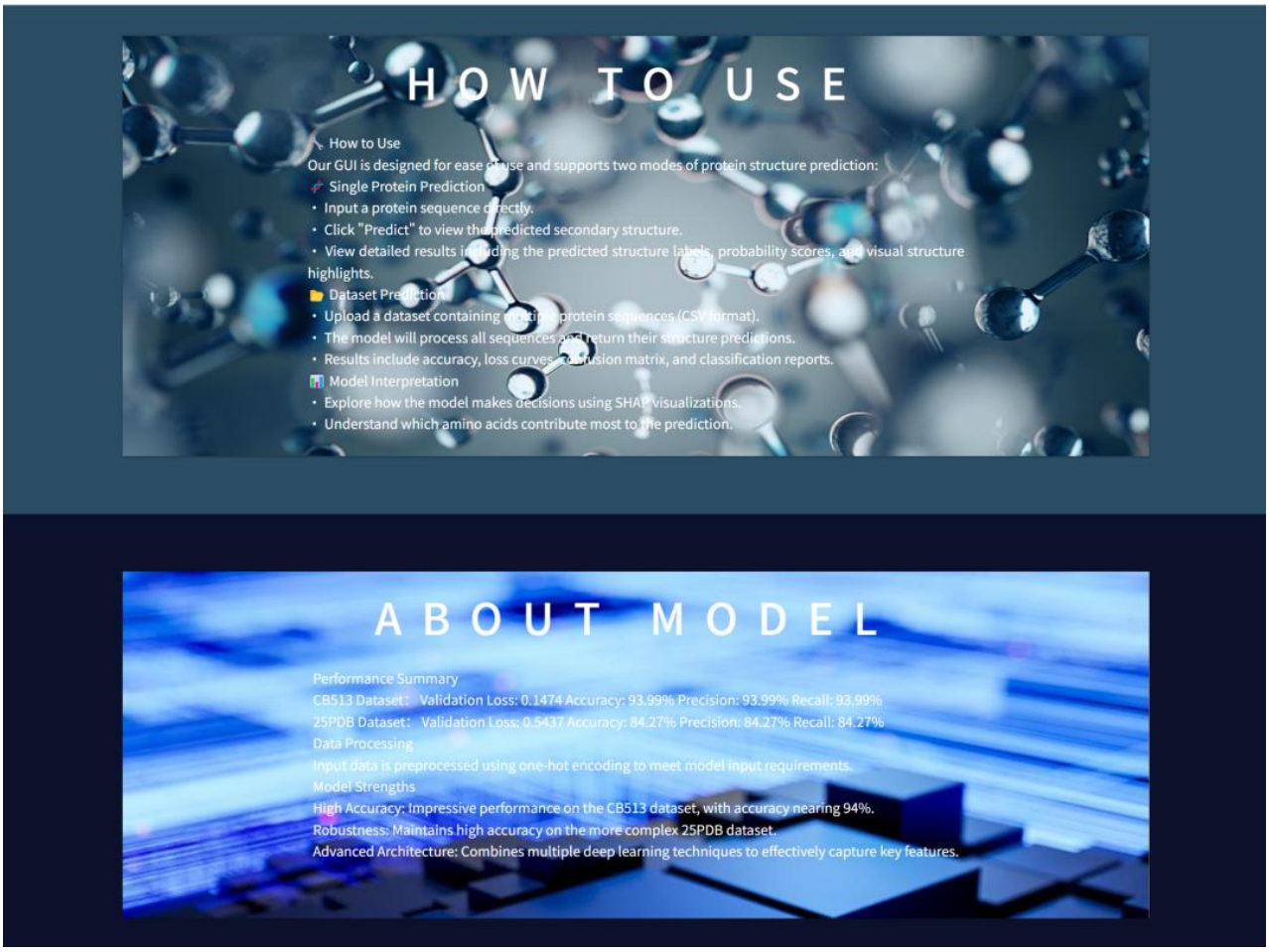

Figure 66. GUI_Hoe to use&About Model

Figure 67 shows the prediction function for a single protein sequence, where the user can enter the protein sequence and then click the start button to begin the prediction. The text box on the right shows the prediction results. Click the reset button below and the prediction will be cleared to start the next prediction.

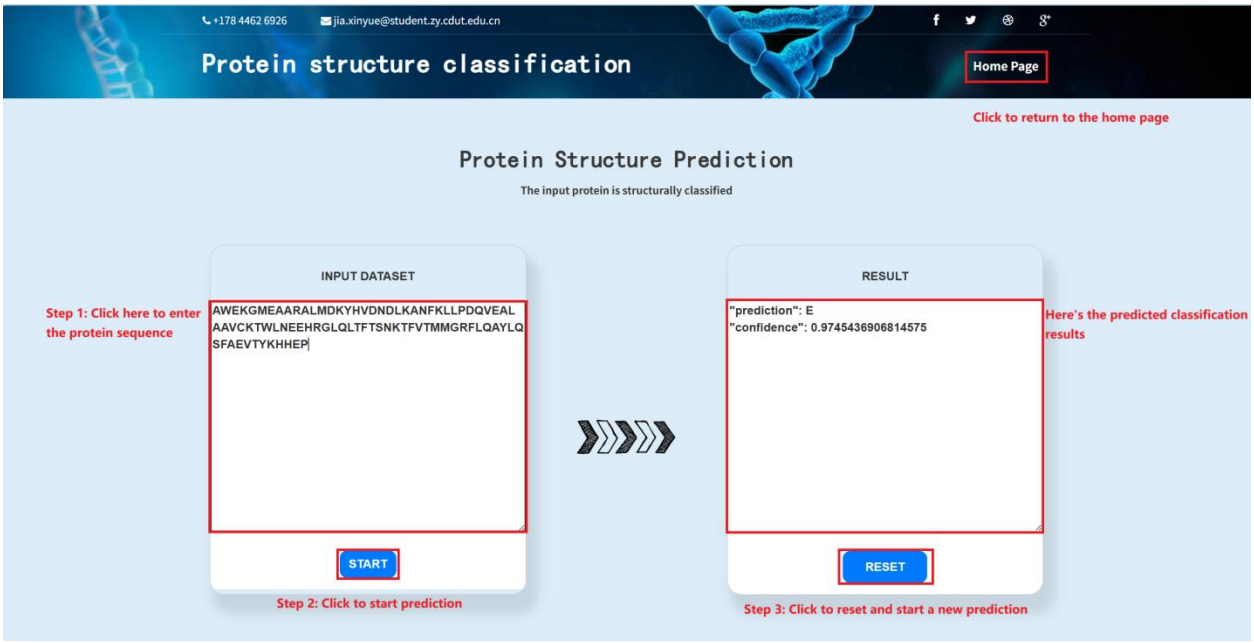

Figure 67. GUI_Protein Function

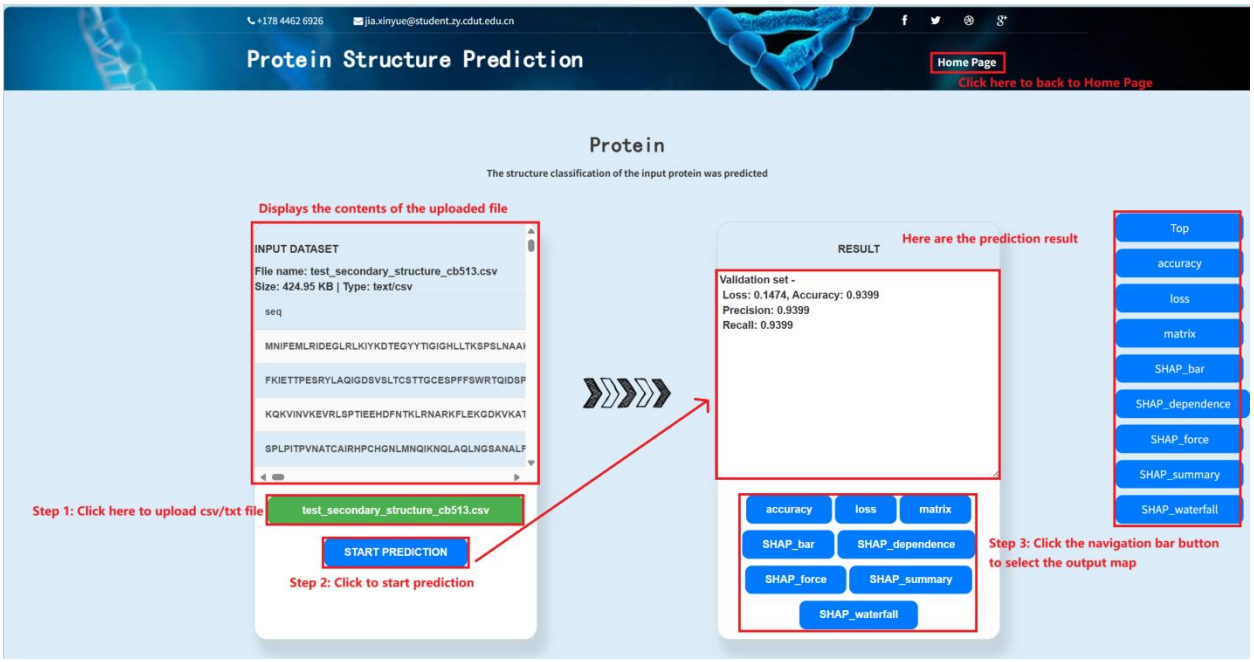

Figure 68. GUI_Dataset Function

As shown in Figure 68 and 69, the prediction function for data sets is available, where users can upload data set files in csv and txt formats. Then click the start prediction button and the system will start the prediction. The text box on the right shows the prediction results. The navigation bar below and to the right shows the resulting diagram of the model and the XAI image to help

the user further understand. Slide down the page or click the corresponding image button to see the corresponding image.

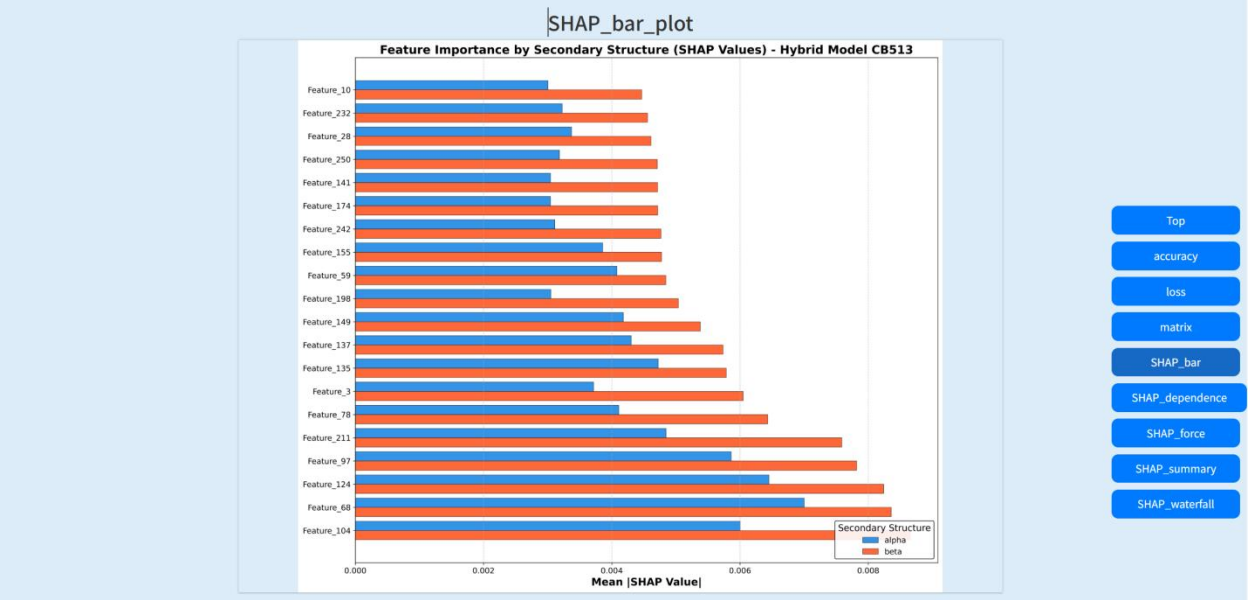

Figure 69. GUI_Result_plot

Chapter 5 Professional Issues

Building on the technical aspects discussed in previous chapters, Chapter 5 shifts focus to the professional issues encountered throughout the project. This chapter provides an overview of the project management strategies used, ethical considerations, and the broader implications of the work, ensuring that the project not only meets technical goals but also aligns with professional standards.

5.1 Project Management

Effective project management played a critical role in the success of the project. This section details the strategies implemented to ensure smooth execution, including planning, scheduling, resource management, and team coordination. The following sections will explore how these management practices were applied to achieve project goals and address challenges.

5.1.1 Activities

Table 8. Activities

Phase objective	Action
1. Preparation	--Find relevant references Read the literature, look for possible solutions, and research classification methods
2. Determine the research model	--A comparison table is used to select the appropriate model
3. Determining the datasets	--The database used to determine a suitable C-BLSTM model was cullPDB
4. Development and Implementation	--Preprocessing & Error analysis
5. Training and Evaluation Metrics	--The C-BLSTM model is trained using categorical cross-entropy loss and Adam optimizer.
6. Model and result test	--Verification of reproducibility: Record test environment Settings and parameters to ensure that the test process can be reproduced.
7. Thesis writing	--Write the paper based on the research results
8. Modify and prepare	--Review the paper carefully, revise as appropriate, and prepare the

presentation	final presentation

5.1.2 Schedule

The schedule is shown as Figure 50 below

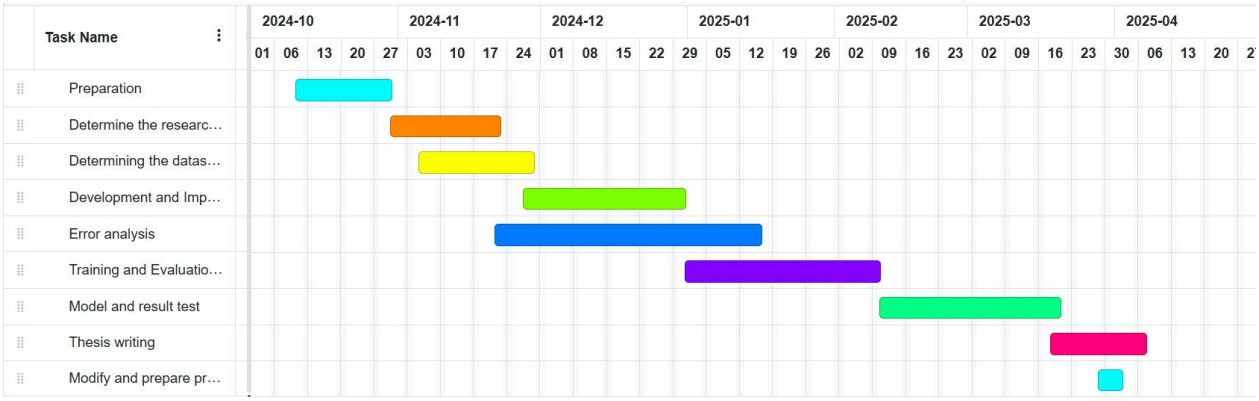

Figure 70. Gantt Chart

5.1.3 Project Data Management

- ❖ Keep up to date by using git to upload the weekly reference form, weekly report, and weekly code collection to GitHub
- ❖ All files, including datasets, model code, references, weekly reports and all types of files will be copied in three copies, one on the local computer, one on the hard drive and one on github.

5.1.4 Project Deliverables

- ❖ The project proposal
- ❖ Weekly progress project reports
- ❖ Final Project Report
- ❖ Code for the project
- ❖ Slides for personal presentation
- ❖ Project presentation

5.2 Risk Analysis

Table 9 displays the analyzed risks during the project progress:

Table 9. Risks

Potential Risk	Potential Causes	Severity	Likelihood	Mitigation
Loss of Project data	Poor version management & Physical Hardware Destruction	high	high	Uploading to at least one cloud repository, add one git reporsitory to manange and update version of codes, datasets etc.
Memory Leakage	The model has huge numbers of parameters	high	medium	First check the code, whether it has large numbers of parameters. Second, before running the model, open the monitor of GPU performance to see if the training occupies more than the memory of GPU, if it is occupying,stop the training immediately and del the model before closing the window.
Data Quality and Imbalance	Training datasets like RCSB-PDB and CB513 may contain noise, missing entries, or imbalanced labels, favoring certain protein structures.	high	high	Implement data cleaning and augmentation techniques (e.g., rotation, flipping); apply oversampling/undersampling methods; use loss weighting strategies to address class imbalance.
Difficulty in Parameter	Attention-ProteinMeNet	medium	medium	Use automated hyperparameter optimization tools (e.g., Optuna,

Tuning	involves many hyperparameters across multiple modules (ProteinNet, BLSTM, Attention), making optimization challenging.			Hyperopt); apply grid or random search methods; optimize individual modules sequentially before tuning the complete model.
Dataset Coverage Issues	Protein datasets like RCSB-PDB and CB513 may not fully represent the structural diversity of natural proteins or include experimental biases.	high	medium	Incorporate additional datasets (e.g., AlphaFold-derived data); augment datasets with synthetic protein data; leverage transfer learning to adapt the model to new datasets.
Lack of Biological Knowledge	Limited understanding of biological principles underlying protein structure may lead to poor feature selection and model design.	high	low	Use virtual environments or Docker to manage dependencies; document and share specific tool versions; rely on long-term support (LTS) versions of frameworks to ensure stability.

5.3 Professional Issues

In the realm of breast cancer detection through deep learning, a multitude of intricate considerations spanning legal, social, ethical, and environmental domains must be navigated with care. Here's a streamlined examination of these aspects:

- Legal Issues: Legal problems in calculation biology are enormous because of the sensitive nature of data used in version improvement. Protein -shaped prediction models, especially those who use organic data such as amino acid sequences, legal issues around the records are important for privatization, Hybro assets and post -possession. U.S. The mandate should treat, save and treat how non-public records with GDPR laws in the EU and HIPAA. In the case of a research, you ensure that data sets are in accordance with the legal guidelines and to protect sensitive organic facts to get stable input.

- Social Issues: The social implications of using deep learning models for protein structure prediction extend to equity and accessibility in research. Models along with Attention-ProteinMeNet and BLSTM are resource-in depth and generally require high-performance computing infrastructure. This ought to restrict access to establishments with fewer assets, exacerbating the digital divide. Further, making sure that those fashions are used to benefit worldwide fitness projects, specially in underserved areas, is an ongoing concern. Ensuring transparency in how those technologies are deployed can decorate public trust and adoption.

- Ethical Issues: Protein-size ethical concerns are the consequences of the AI-DA HAKA insight into journal use, prejudice in version improvement and biological research. There is a danger that prejudice in school journals will lead to incorrect or discriminatory conclusions, especially in applications such as drug design or scientific immunity. To ensure that the models are advanced and have been evaluated with justice in ideas. In addition, researchers should follow moral requirements by ensuring that models of models are used responsibly, by assessing capacity effects on public health and safety.

- Environmental Issues: Environmental problems Recognition of their good size energy intake related to calculation fashion. When it comes to large datasets such as deep mastery models, especially CB513 or RCSB-PDB, calculation funds are required in good size. Exercising this fashion, often on more than a GPU or TPU, can cause high carbon footprints due to the power used by information centers. Using electrically skilled algorithm, adaptation of version architecture and transition to green statistics functions is a possible

strategy to reduce these environmental effects. Future work must detect an approach to reduce environmental footprint and at the same time maintain the model's accuracy and scalability.

By addressing these key professional issues—legal, social, ethical, and environmental—this chapter underscores the need for responsible development and deployment of deep learning-based models in protein structure prediction, ensuring their societal benefits are maximized while minimizing potential harms.

Chapter 6 Conclusion

The project developed the successful attention protein model to predict protein-secondary structures, and addressed the limitations of traditional and machine learning-based methods. By integrating protein, BLSTM and attention mechanisms, the model performed improved performance in capturing local properties, long -distance sequence dependence and significant residual interactions. The model achieved verification accuracy of 94.15% on the CB513 dataset and 96.49% on RCSB-PDB dataset, improved standalone protein, BLSTM and Meditation ProteinThyto Architecture. Data was strengthened to increase the efficiency of the model of preprosecating techniques such as a hot coding, sequence padding and label mapping. ROC-AUC RCSB-PDB, decreases with AUC at 0.983 on F1 score, and assessment matrix as Confusion Matrix valued the model's credibility further. In addition, Shap provided biological insights by identifying main relicvies, which were valid for structural predictions and adapted to known biological properties. A user -friendly graphic interface was designed to allow interactive protein composition analysis, which supports both personal and batch treatment.

In addition to the strong performance of the proposed model, the project still has some limitations. One of the boundaries is that the model's accuracy can be expanded and expanded by expanding datasets with more different protein structures or synthetic data. Second, while the model better handles long -distance addiction compared to the former architecture, there is still room to improve generalization in unseen protein sequences. Third, the model's calculation complexity and its dependence on high -decommissioning hardware can limit access to it and real -time prediction skills, especially to the resource world environment. Future work can detect adaptation techniques to reduce calculation costs and improve scalability, as well as 3D structural data to expand the skills of the prediction.

Furthermore, future work can focus on increasing the architecture of the model through advanced focus mechanisms or graph nerves networks to better capture 3D structural conditions. The expansion of the dataset with alfoldelastic structures or synthetic data can improve normalization, while addressing square imbalance through techniques such as SMOT or weighted loss functions. Expanding the model to predict the finding secondary structures or tertiary structures will make their project wide. Adaptation of calculation efficiency for real-time predictions of hardware with low resources and utilizing energy-capable algorithm can reduce the environmental impact. The collaboration with practical biologists will be computational and practical applications with practical biologists to validate predictions against weight-labr results

and integrate models into pipelines. Finally, it is important to ensure moral openness and justice in biomedical distribution.

References

[1] J. Hong, Z.-H. Zhan, L. He, Z. Xu, and J. Zhang, 'Protein Structure Prediction Using A New Optimization-Based Evolutionary and Explainable Artificial Intelligence Approach', *IEEE Trans. Evol. Comput.*, pp. 1–1, 2024, doi: 10.1109/TEVC.2024.3365814.

[2] S. Prasad, N. Nandhini, R. Singh, A. Anuradha, L. Varshitha Averineni, and S. Debnath, 'Perspectives of machine learning on protein structure prediction and function', in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, May 2023, pp. 385–390. doi: 10.1109/ICACITE57410.2023.10183157.

[3] R. K. Deepak, and M. K. Praveen, 'A Review of Machine Learning Techniques and Applications for Health Care', IEEE Access, pp. 4-8, 2021, doi: 10.1109/ICATME50232.2021.9732761.

[4] X. Qiu, H. Li, G. Ver Steeg, and A. Godzik, 'Advances in AI for Protein Structure Prediction: Implications for Cancer Drug Discovery and Development', *Biomolecules*, vol. 14, no. 3, Art. no. 3, Mar. 2024, doi: 10.3390/biom14030339.

[5] A. Paiardini, 'Protein Structure Prediction in Drug Discovery', *Biomolecules*, vol. 13, no. 8, Art. no. 8, Aug. 2023, doi: 10.3390/biom13081258.

[6] A. Shehu and L. E. Kavraki, 'Modeling Structures and Motions of Loops in Protein Molecules', *Entropy*, vol. 14, no. 2, pp. 252–290, Feb. 2012, doi: 10.3390/e14020252.

[7] T. Selwate, M. A. Kamble, P. M. Sabale, D. Dhabarde, K. Dongarwar, and J. Baheti, 'Protein Structure Prediction: A Computational Approach to Unraveling Molecular Mysteries', in *Deep Learning and Computer Vision: Models and Biomedical Applications: Volume 1*, U. N. Dulhare and E. H. Houssein, Eds., Singapore: Springer Nature, 2025, pp. 63–87. doi: 10.1007/978-981-96-1285-7_4.

[8] J. Jumper *et al.*, 'Highly accurate protein structure prediction with AlphaFold', *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[9] W. Wang, J. Wang, D. Xu, and Y. Shang, 'Two New Heuristic Methods for Protein Model Quality Assessment', *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 17, no. 4, pp. 1430–1439, Jul. 2020, doi: 10.1109/TCBB.2018.2880202.

[10] ‘DNACoder: a CNN-LSTM attention-based network for genomic sequence data compression | Neural Computing and Applications’. Accessed: Dec. 18, 2024.. Available: https://link.springer.com/article/10.1007/s00521-024-10130-4

[11] X. Ma and E. Hovy, ‘End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF’, May 29, 2016, *arXiv*: arXiv:1603.01354. doi: 10.48550/arXiv.1603.01354.

[12] N. P, K. M. Sudar, V. S. Sri, Nikitha. V, V. S. S. Reddy, and V. M, ‘Enhancing Protein Structure Generation Through Deep Learning Techniques’, in *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Mar. 2024, pp. 1–6. doi: 10.1109/INCOS59338.2024.10527559.

[13] J. L. Filgueiras, D. Varela, and J. Santos, ‘Protein structure prediction with energy minimization and deep learning approaches’, *Nat. Comput.*, vol. 22, no. 4, pp. 659–670, Dec. 2023, doi: 10.1007/s11047-023-09943-4.

[14] Z. Shi and B. Li, ‘Graph neural networks and attention-based CNN-LSTM for protein classification’, Feb. 22, 2023, *arXiv*: arXiv:2204.09486. doi: 10.48550/arXiv.2204.09486.

[15] M. M. Mohamed Mufassirin, M. A. H. Newton, J. Rahman, and A. Sattar, ‘Multi-S3P: Protein Secondary Structure Prediction With Specialized Multi-Network and Self-Attention-Based Deep Learning Model’, *IEEE Access*, vol. 11, pp. 57083–57096, 2023, doi: 10.1109/ACCESS.2023.3282702.

[16] A. Golwelkar and A. Kothari, ‘A Review of Protein Sequences of COVID-19 Using Machine Learning and Deep Learning Approaches’, in *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, Dec. 2023, pp. 1–9. doi: 10.1109/ICTBIG59752.2023.10456322.

[17] ‘A New Approach Of Applying Deep Learning To Protein Model Quality Assessment | IEEE Conference Publication | IEEE Xplore’. Accessed: Oct. 29, 2024. Available: https://ieeexplore.ieee.org/document/8983005

[18] J. S. A, K. Merriliance, and N. Soundiraraj, ‘Integrating Deep Learning with Structural Bioinformatics using Next-Generation Protein Stability Prediction’, in *2024 International Conference on Inventive Computation Technologies (ICICT)*, Apr. 2024, pp. 1252–1257. doi: 10.1109/ICICT60155.2024.10544908.

[19] ‘Structural patterns in globular proteins - PubMed’. Accessed: Dec. 18, 2024. Available: https://pubmed.ncbi.nlm.nih.gov/934293/

[20] J. GAEtN, ‘Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins’.

[21] S. R. Eddy, ‘Profile hidden Markov models’, *Bioinforma. Oxf. Engl.*, vol. 14, no. 9, pp. 755–763, 1998, doi: 10.1093/bioinformatics/14.9.755.

[22] Yanfei, C. H. I. Yanfei, Chun, L. I. Chun, Xudong, and F. Xudong, ‘Advances in machine learning for protein function prediction’. Accessed: Dec. 18, 2024. Available: https://cjb.ijournals.cn/html/cjbcn/2023/6/gc23062141.htm

[23] ‘MO4: A Many-Objective Evolutionary Algorithm for Protein Structure Prediction | IEEE Journals & Magazine | IEEE Xplore’. Accessed: Oct. 29, 2024. Available: https://ieeexplore.ieee.org/document/9477421

[24] S. Wang, J. Peng, J. Ma, and J. Xu, ‘Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields’, *Sci. Rep.*, vol. 6, no. 1, p. 18962, Jan. 2016, doi: 10.1038/srep18962.

[25] Y. Liu, Y. Chen, and J. Cheng, ‘Feature extraction of protein secondary structure using 2D convolutional neural network’, in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Datong, China: IEEE, Oct. 2016, pp. 1771–1775. doi: 10.1109/CISP-BMEI.2016.7853004.

[26] S. K. Sønderby and O. Winther, ‘Protein Secondary Structure Prediction with Long Short Term Memory Networks’, Jan. 04, 2015, *arXiv*: arXiv:1412.7828. doi: 10.48550/arXiv.1412.7828.

[27] J. Liu, X. Zhang, K. Huang, Y. Wei, and X. Guan, ‘Grain Protein Function Prediction Based on CNN and Residual Attention Mechanism with AlphaFold2 Structure Data’, *Appl. Sci.*, vol. 15, no. 4, Art. no. 4, Jan. 2025, doi: 10.3390/app15041890.

[28] ‘Jiang, Y., & Wang, W. (2021). Enhanced protein structure prediction with attention-guided CNN. Bioinformatics, 37(4), 536-544. DOI: 10.1093/bioinformatics/btaa751’.

[29] ‘Protein secondary structure prediction based on integration of CNN and LSTM model’, *J. Vis. Commun. Image Represent.*, vol. 71, p. 102844, Aug. 2020, doi: 10.1016/j.jvcir.2020.102844.

[30] ‘Khan, S., & Mazumdar, J. (2022). Protein secondary structure prediction using hybrid deep learning techniques. BMC Bioinformatics, 23(1), 112. DOI: 10.1186/s12859-022-04603-4’.

[31] C. S. Srushti, P. M. Prathibhavani, and K. R. Venugopal, ‘Eight-State Accuracy Prediction of Protein Secondary Structure using Ensembled Model’, in *2023 International Conference for Advancement in Technology (ICONAT)*, Jan. 2023, pp. 1–6. doi: 10.1109/ICONAT57137.2023.10080387.