LiteSeqCNN + MULTI-ATTENTION BASED FRAMEWORK FOR PROTEIN FUNCTION PREDICTION

A PROJECT REPORT

Submitted by

SHAMSUDDEEN MOHAMMED MAMMAN KISLAY

SINGH RAJPUT

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

C.V. RAMAN GLOBAL UNIVERSITY

BHUBANESHWAR- ODISHA-752054

APRIL-2025

C.V. RAMAN GLOBAL UNIVERSITY

BHUBANESHWAR- ODISHA-752054

BONAFIDE CERTIFICATE

Certified that this project report **"Attention based LiteSeqCNN: A deep CNN based architecture for protein function prediction"** is a bonafide work submitted by **Shamsuddeen Mohammed Mamman and Kislay Singh Rajput** of CGU-Odisha, Bhubaneswar who carried out the project under my supervision.

Dr. Monalisa Mishra
HEAD OF THE DEPARTMENT
Department of CSE

Dr. Ashish Ranjan
SUPERVISOR
Assistant Professor
Department of CSE

C.V. RAMAN GLOBAL UNIVERSITY
BHUBANESWAR-ODISHA-752054

CERTIFICATE OF APPROVAL

This is to certify that we have examined the project entitled**"Attention based LiteSeqCNN: A deep CNN based architecture for protein sequence prediction"** is bonafide work submitted by **Shamsuddeen Mohammed Mamman and Kislay Singh Rajput** of CGU-Odisha, Bhubaneswar.

We hereby accord our approval of it as a**8th Semester** major project work carried out and presented in a manner required for its acceptance for the partial fulfilment for **Bachelor Degree of Name of the Department** for which it has been submitted. This approval does not necessarily endorse or accept every statement made, opinion expressed, or conclusions drawn as recorded in this major project, it only signifies the acceptance of the major project for the purpose it has been submitted.

Dr. Ashish Ranjan
(SUPERVISOR)
Assistant Professor
Dept. of CSE

C.V. RAMAN GLOBALUNIVERSITY

BHUBANESWAR-ODISHA-752054

DECLARATION

I declare that this project report titled "**Attention based LiteSeqCNN: A deep CNN based architecture for protein sequence prediction"**submitted in partial fulfilment of the degree of **B. Tech in (Computer Science and Engineering)** is a record of original work carried out by me under the supervision of **Dr. Ashish Ranjan, Assistant Professor, Department of Computer Science and Engineering**, and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited.

SHAMSUDDEEN M. MAMMAN KISLAY SINGH RAJPUT

PLACE: BHUBANESHWAR, ODISHA DATE:
30/04/25

ACKNOWLEDGEMENT

We extend our sincere gratitude to all those who contributed to the realization of this report for our major project. Our appreciation goes to our guide **Dr. Ashish Ranjan, Assistant Professor, Department of Computer Science and Engineering,** whose insights and expertise enriched the content of this document.

Special thanks to our institution C.V. Raman Global University that provided valuable resources, and support throughout the process. We also acknowledge the dedication and commitment of our team members who diligently worked on this project, ensuring its thoroughness and accuracy.

Shamsuddeen Mohammed Mamman

Kislay Singh Rajput

ABSTRACT

Protein function prediction is a vital task in computational biology, with applications in understanding biological processes, drug discovery, and disease research. Lightweight architectures like Lite-SeqCNN have demonstrated the ability to balance efficiency and accuracy for this task. However, these models lack mechanisms to focus dynamically on the most biologically informative sequence regions, which limits their predictive performance.

In this study, we enhance Lite-SeqCNN by integrating an attention mechanism, allowing the model to identify and prioritize key functional motifs within protein sequences. Evaluated on the Data2017 dataset, the proposed attention-enhanced Lite-SeqCNN demonstrates significant improvements over the baseline model, achieving **0.669 precision**, **0.532 F1-score for BP dataset and 0.809 precision**, **0.654 F1-score for MF dataset**. Additionally, the attention mechanism enhances interpretability, providing insights into the sequence regions most critical for function prediction.

Our findings highlight the potential of lightweight architectures augmented with attention mechanisms to advance protein function prediction and uncover biologically meaningful patterns.

TABLE OF CONTENTS

LIST OF TABLES

TABLE DESCRIPTION **PAGE NUMBER**

LIST OF FIGURES

FIGURE DESCRIPTION **PAGE NUMBER**

LIST OF ABRIVIATIONS

BP: Biological Process

MF: Molecular Function

TP: True Positive

FP: False Positive

FN: False Negative

TN: True Negative

AUPR: Area Under Precision and Recall curve

CNN: Convoluted Nural Network

LiteSeqCNN: Lite Sequential Convoluted Nural Network

Chapter 1

Introduction

Protein function prediction plays a pivotal role in advancing various biological research domains, including genome annotation, molecular biology, and drug discovery. The growing volume of protein sequence data, driven by advancements in high-throughput sequencing technologies, necessitates the development of computational models that can analyze and interpret large-scale data efficiently and accurately. Traditional sequence-alignment-based methods, while effective in some cases, often fail to scale and perform well when faced with datasets of increasing size and complexity. As a result, machine learning approaches, particularly deep learning, have emerged as transformative tools in for protein function prediction [1], [2], [3], [4], [5], [6].

Convolutional Neural Networks (CNNs) have proven particularly effective for protein function prediction, as they can learn abstract, high-level features directly from raw protein sequences without requiring extensive domain-specific feature engineering. Among these, Lite-SeqCNN [3], a lightweight CNN architecture, has gained recognition for its computational efficiency and effective performance on large-scale datasets such as Data2017. Lite-SeqCNN's ability to process sequences with fewer parameters and reduced computational overhead makes it a preferred choice for large-scale protein datasets. However, its uniform treatment of all sequence regions fails to capture the varying functional importance of different parts of the sequence, limiting its predictive accuracy and interpretability.

To address this challenge, our research enhances Lite-SeqCNN [3] by integrating an attention mechanism, which introduces a selective focus on the most informative parts of a protein sequence. Attention mechanisms, inspired by human cognitive systems, enable models to dynamically assign importance to specific sequence regions based on their relevance to the task. By embedding an attention layer into Lite-SeqCNN, the model can effectively prioritize critical functional motifs, often sparsely distributed across the protein sequence, while minimizing the influence of less relevant regions. This enhancement not only improves the model's predictive performance but also provides valuable interpretability by highlighting the specific sequence regions that influence the predictions.

The enhanced Lite-SeqCNN model was evaluated on the Data2017 dataset, a robust and widely-used benchmark for protein function prediction. This dataset includes diverse protein sequences annotated with functional properties, making it ideal for assessing the performance of computational models. Experimental results indicate that the attention-

enhanced Lite-SeqCNN significantly outperforms the baseline Lite-SeqCNN in terms of metrics such as precision, recall, F1-score, and area under the precision-recall curve (AUPR). This improvement highlights the utility of incorporating attention mechanisms for refining predictions and ensuring accurate identification of functional motifs.

Additionally, the proposed model provides an interpretable framework for protein function prediction by visualizing attention weights. These visualizations offer insights into the sequence regions that most strongly contribute to functional predictions, thereby guiding experimental validation efforts and advancing our understanding of protein functionality. The ability to identify and prioritize critical regions makes this approach highly relevant for both theoretical and applied research in protein science.

This work demonstrates that augmenting CNN-based models with attention mechanisms is a powerful strategy for addressing the challenges of protein function prediction. The improved performance and interpretability achieved by this approach set the stage for future advancements in computational protein science, potentially extending to other biological sequence analysis tasks and multi-model datasets.

Chapter 2 LITERATURE SURVEY

Protein function identification is a severe challenge pertaining to development of drugs, bio-fertilizers, bio-fuels, etc. In this regard, development of computational approaches to predict the functions are required. This chapter presents a survey of existing approaches as discussed next.

Table 2.1: Literature review

Paper name	Dataset used	Model name	Result obtained
λ-scaled attention network: : A novel fast attention mechanism for efficient modelling	UniProtKB dataset	Multi layer perceptron (MLP), Keras with	BP dataset: +10.08% improvement, MF dataset:+13.3% improvement in F1-
Lite-SeqCNN: A Light-Weight deep CNN architecture for	Data2017, CAFA3, DATA2016	Light-weight CNN model for efficient	Improved PreAvg, Fmax and AUPR over existing

MCWS-Transformers: Towards an Efficient Modeling of Protein Sequences via Multi Context-Window Based Scaled Self-	UniProtKB dataset	MCWS-Transformer.	+2.30% (BP) and +2.08% (MF) F1-score improvement; +3.38% (BP) and +2.86% (MF) F1-score improvement
Global-ProtEnc-Plus: A Global Protein encoding model. [1]	Data2017, Data2016, CAFA3.	Dual-output with multi-attention.	Improved Fmax and AUPR over existing methods.

The field of protein function prediction has seen significant advancements with the introduction of various deep learning architectures tailored to extract meaningful patterns from sequence data.

The λ-scaled attention network [2] utilizes a multilayer perceptron (MLP) architecture implemented using TensorFlow. It was applied to the UniProtKB dataset, including BP (58,310 sequences) and MF (43,218 sequences) GO terms. This model demonstrated substantial performance improvements, achieving a +10.08% improvement in F1-scores for the BP dataset and +13.3% improvement for the MF dataset, outperforming state-of-the-art models like ProtVenGen-100, ProtVenGen-Plus, and ProtVenGen-Ensemble.

The Lite-SeqCNN model [3] introduced a lightweight CNN-based architecture optimized for protein function prediction. This model emphasized computational efficiency while maintaining performance. It was tested on datasets like Data2017, CAFA3, and Data2016, demonstrating improvements in metrics such as PreAvg, Fmax, and AUPR over existing methods, thus offering a robust yet efficient solution for protein function prediction tasks.

The MCWS-Transformer model [4] introduced a multi-context window-based scaled self-attention mechanism for efficient modeling of protein sequences. Tested on the UniProtKB dataset with 58,310 sequences and 295 classes, this approach achieved notable improvements. It showed a +2.30% (BP) and +2.08% (MF) improvement in F1-scores over standard transformer-based architectures. Additionally, the model outperformed ProtVecGen-Plus and ProtVecGen- Ensemble approaches with a +3.38% (BP) and +2.86% (MF) improvement in F1- scores.

The Global-ProtEnc-Plus model [1] incorporates dual-output functionality with multi-attention mechanisms. Tested on Data2017, Data2016, and CAFA3 datasets, this model achieved significant performance gains. For the Data2016 dataset, it demonstrated

improvements with Fmax scores of 0.45 ± 0.0031 for BP, 0.593 ± 0.0027 for MF, and 0.7 ± 0.0017 for CC, alongside corresponding improvements in AUPR metrics.

These models showcase the rapid evolution of deep learning methodologies in protein function prediction, with advancements targeting both efficiency and accuracy. Our study builds upon these approaches by integrating an attention mechanism into Lite-SeqCNN,[3] further enhancing its ability to identify biologically significant sequence motifs and improving its overall performance.

Chapter 3

Methodology

3.1 Description of the dataset used

The Data2017 dataset is widely used in bioinformatics for protein function prediction tasks. Here are some details about it:

3.1.1 Overview:

Purpose: The dataset is designed for protein function prediction and is commonly used for benchmarking machine learning and deep learning models in the field.

Source: It was curated for the Critical Assessment of Functional Annotation (CAFA) challenge, an international competition focused on evaluating computational methods for automated protein function prediction.

3.1.2 Features:

Protein Sequences: The dataset contains amino acid sequences of proteins.

Annotations: Each protein is annotated with its functional properties, typically based on the Gene Ontology (GO) terms. These GO terms are hierarchical and describe:

- **Molecular Function (MF)**: What a protein does at the biochemical level.
- **Biological Process (BP):** The biological objectives or processes the protein contributes to.
- **Cellular Component (CC):** The location in the cell where the protein is active.

Size: Depending on the task and version of the dataset, it can contain thousands to millions of sequences with corresponding annotations.

3.1.3 Challenges with the Dataset:

Imbalanced Data: Many GO terms are sparsely represented, making it challenging for models to predict rare functions.

Sequence Variability: Protein sequences can vary greatly in length, from a few amino acids to several thousand.

Complexity of GO: The hierarchical structure of GO terms adds complexity to the classification task.

3.1.4 Applications:

Protein function prediction.

Sequence-to-function learning in deep learning.

Benchmarking protein annotation methods.

3.1.5 Preprocessing Considerations:

Sequence Encoding: Convert amino acid sequences into numerical representations (e.g., one-hot encoding, integer encoding, embeddings like ProtVec, or ESM embeddings).

Handling Imbalance: Use techniques like oversampling, under-sampling, or weighted loss functions to address class imbalance.

Hierarchical Constraints: Ensure that predictions respect the hierarchical structure of GO terms

3.1.6 Number of samples in the dataset

The number of samples of each class (BP, MF and CC) in Data2017 are listed below:

Table 3.1: Overview of the dataset Data2017

Stats	BP	CC	MF
Training samples	43457	-	32280
Testing samples	3359	-	3132
Number of classes	295	-	135

3.2 Overview of the Proposed Attention-Enhanced Lite-SeqCNN Architecture

The proposed architecture enhances Lite-SeqCNN [3] by integrating an attention layer, allowing the model to focus on biologically significant regions of protein sequences. This section outlines the various components of the architecture, including the input

preprocessing, CNN blocks, attention layer, and output layer, as illustrated in Figure 1.

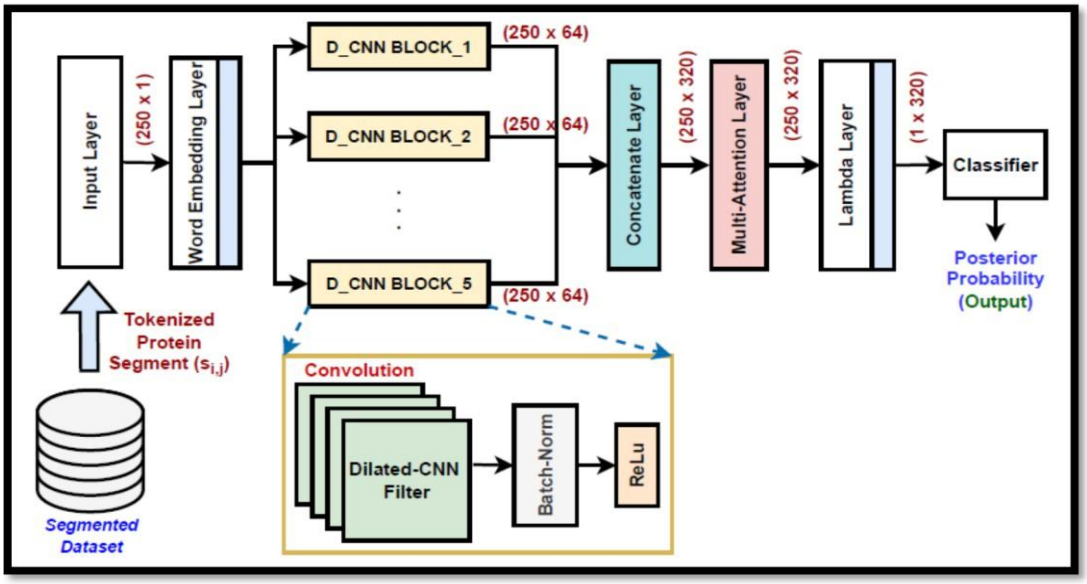

Figure 3.1: Architecture design

3.3 Working of the architecture

The process flow diagram of the model that how it works on the protein sequence is described below:

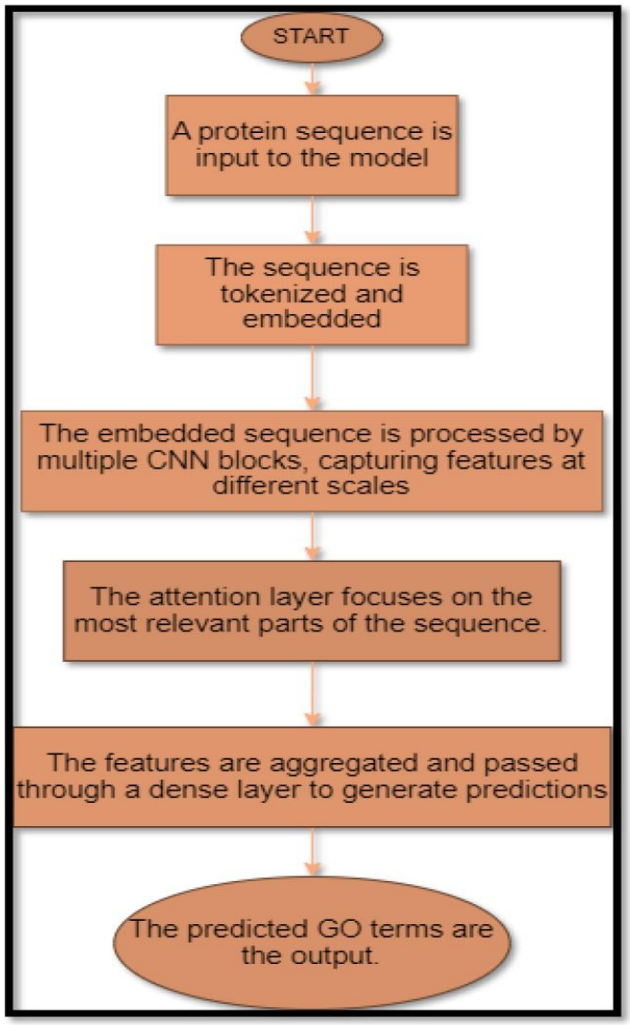

Figure 3.2: Process flow diagram

3.3.1. Input Processing: Tokenization and Word Embedding

Protein sequences are first divided into smaller segments, S_1, S_2…..S_i , to prepare for analysis. These segments are tokenized and fed into a word embedding layer, where each token is transformed into a fixed-dimensional vector representation.

These embedding captures essential information about amino acid composition and sequence structure, enabling the model to process each segment as a continuous numerical representation.

3.3.2. Convolutional Neural Network (CNN) Blocks

The core of the model consists of multiple CNN blocks [3], each designed to capture local sequence patterns and hierarchical features. Each CNN block comprises:

Dilated Convolution Layer: This layer performs convolution with a specific dilation rate, which increases the receptive field without additional computational cost. Using 256 filters, the dilated convolution layer captures diverse local and long-range patterns within the protein sequences, improving the model's ability to identify functionally significant motifs.

Batch Normalization: After convolution, batch normalization is applied to stabilize the training process and improve convergence. This layer normalizes activations, reducing internal covariate shifts.

ReLU Activation: The Rectified Linear Unit (ReLU) activation function is applied to introduce non-linearity, enabling the network to learn complex relationships in the data.

The output from each CNN block represents different hierarchical features from the input sequence, which are then concatenated to form a unified feature map.

3.3.3. Multi-Attention Layer

Following the concatenation of feature maps from each CNN block, the multi- attention layer [1] is applied to highlight the most informative regions within the sequence. The attention mechanism assigns varying weights to different regions, enabling the model to focus on critical segments relevant to protein function prediction. This selective attention not only improves predictive accuracy but also enhances the interpretability of the model by identifying sequence regions that contribute most to the prediction.

3.3.4. Global Average Pooling Layer

The feature map output from the attention layer is passed through a 1D Global Average Pooling layer, which reduces the dimensionality by computing the average of each feature channel [3]. This operation produces a fixed-size vector, effectively summarizing the key features learned across the sequence while maintaining the spatial hierarchy.

3.3.5. Dense Output Layer

The fixed-size vector from the pooling layer is fed into a dense layer, which serves as the final output layer of the model. This layer applies a softmax activation function to produce a posterior probability distribution across different classes [3], representing the predicted protein function.

3.4. Training and Evaluation

3.4.1. Dataset: Data2017

The proposed model is trained and evaluated on the Data2017 dataset, a comprehensive dataset specifically curated for protein function prediction tasks. The dataset includes diverse protein sequences labeled according to their functional categories. Standard preprocessing steps such as tokenization and padding are applied to ensure uniform input length across sequences.

3.4.2. Training Process

The model is trained using a supervised learning approach, with a categorical cross-entropy loss function for multi-class classification. Key hyperparameters, such as learning rate, batch size, and the number of training epochs, are optimized to achieve the best performance. The Adam optimizer is used to adjust the model weights iteratively, balancing convergence speed and accuracy.

3.4.3. Evaluation Metrics

To assess the performance of our attention-enhanced Lite-SeqCNN model for protein function prediction, we use a range of evaluation metrics typically employed in classification tasks: **accuracy**, **precision**, **recall**, and **F1-score**. These

metrics provide a holistic evaluation of the model's performance, accounting for different aspects of prediction quality. Additionally, **attention weight visualization** aids in interpreting the biological relevance of the model's predictions.

In this paper we have used the below given evaluation metrices:

3.4.3.1 Accuracy

Accuracy measures the proportion of correctly predicted instances over the total number of predictions. It is given by:

$$\frac{\text{Accc TP+ TN}}{\text{Accc TP+ TN+ FP+ FN}}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

3.4.3.2 Precision

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive. It is defined as:

$$\frac{\text{recn= TP}}{\text{recn= TP+ FP}}$$

Precision highlights the model's ability to minimize false positives, which is especially critical in applications requiring high specificity.

3.4.3.3 Recall

Recall (Sensitivity or True Positive Rate), recall evaluates the proportion of actual positives correctly identified by the model. It is given by:

$$\frac{\text{Recall= TP}}{\text{TP+ FN}}$$

High recall ensures that the model does not miss relevant predictions (false negatives).

3.4.3.4 F1-Score

The F1-score is the harmonic mean of precision and recall, balancing the trade-off between these metrics. It is particularly useful when the dataset is imbalanced. The formula is:

$$\text{Accuracy} = 2 * \text{Precision} * \text{Recall}$$

$$\text{Recall+ Precision}$$

A higher F1-score indicates a well-balanced model performance between precision and recall.

3.4.3.5 AUPR Curve

Area Under the Precision-Recall Curve (AUPR) in addition to standard metrics, the **AUPR** is used for detailed evaluation in imbalanced datasets. It integrates precision and recall across various thresholds, highlighting model performance in distinguishing between true positives and false positives.

$$\text{AUPR} = \int_0^1 \text{Precision}(t) \cdot \frac{\partial \text{Recall}(t)}{\partial t} \, dt$$

Here, *t* represnts the decision threshold.

3.4.3.6 Why These Metrics Matter:

Accuracy gives an overall performance snapshot.

Precision and **Recall** are crucial in applications with skewed datasets (e.g., rare protein functions).

F1-Score provides a balanced measure in cases where false positives and false negatives carry similar importance.

AUPR ensures robustness in evaluating models on datasets with class imbalance, which is common in protein function prediction.

These metrics collectively ensure a comprehensive evaluation, making the results reliable for both computational performance and biological relevance.

.

Chapter 4

RESULTS AND DISCUSSIONS

In this study, we evaluated the performance of the LiteSeq-CNN model [3] with an added attention layer on the Data2017 dataset for protein function prediction. The results are presented for two protein function categories: Biological Process (BP) and Molecular Function (MF). The evaluation metrics used were Recall, Precision, F1 Score, and AUPR (Area Under the Precision-Recall Curve), measured across different segment sizes (200, 300, 400, and 500).

4.1. Biological Process (BP) Dataset Performance

The performance of the model on the BP dataset across various segment sizes is shown in

the bar chart (BP dataset performance metrics by segment size). The following observations were made:

Precision consistently exhibited the highest scores across all segment sizes, indicating the model's capability to accurately identify true positive instances while minimizing false positives.

Recall showed a relatively lower performance compared to other metrics, especially at smaller segment sizes, suggesting the model could benefit from improvements in capturing all relevant protein functions.

F1 Score and **AUPR** displayed moderate scores, reflecting the balance between precision and recall. These metrics were more stable across segment sizes, with a slight increase as the segment size increased.

Below are the tables depicting the results of the research on this dataset:

Table 4.1: Ablation analysis of BP dataset, segment size 200, offset 100

S. no.	Approach	Avg.	Avg. Recall	F-Max	AUPR
1.	Without	0.665	0.415	0.513	0.39
2.	Single Attention	0.674	0.419	0.516	0.39
3.	Multi Attention	0.669	0.441	0.532	0.385

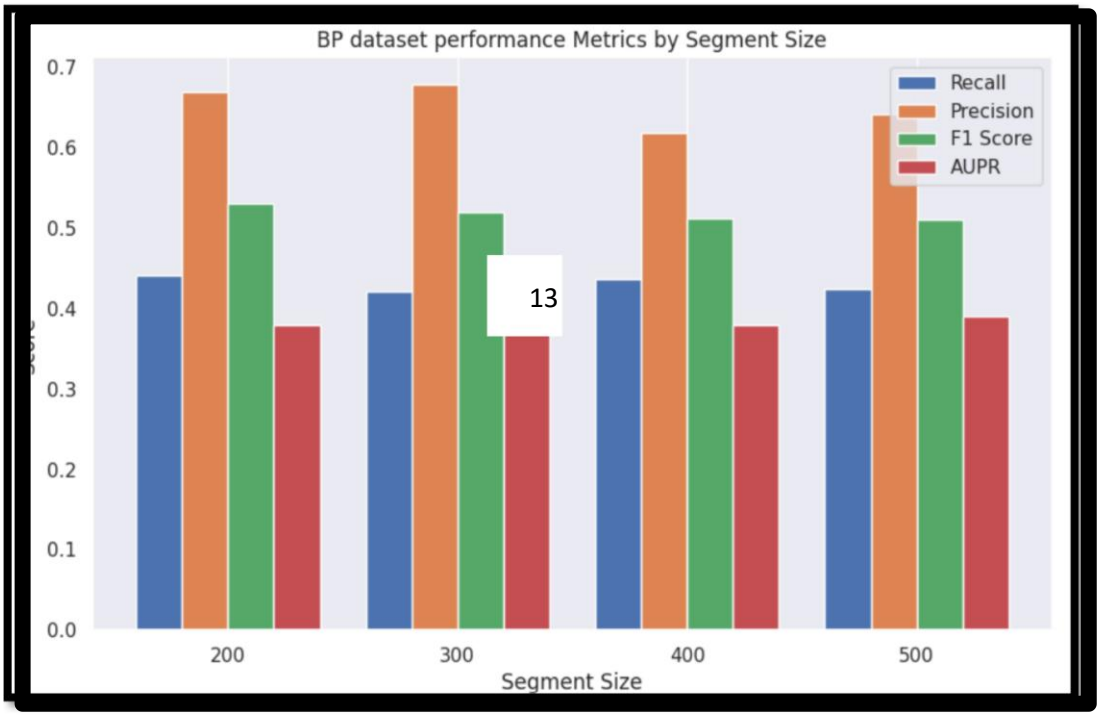

Figure 4.1: BP performance metrics by segment size

Table 4.2: State of the art comparison for Biological process dataset, segment size 200, offset 100

S. No.	Approach	Avg.	Avg.	F-Max	AUPR
1.	MLDA	0.568	-	0.400	0.303
2.	ProtVecGen-120	0.427	-	0.385	-
3.	ProtVecGen-Plus	0.449	-	0.422	-
4.	ProtVecGen-Plus + MLDA	0.694	-	0.501	0.496
5.	Global-ProtEnc-80	0.514	-	0.490	0.407
6.	Global-ProtEnc-Plus	0.680	-	0.542	0.485
7.	Lite-SeqCNN-200	0.507	-	0.512	0.320
8.	Proposed	0.669	0.441	0.532	0.385

4.2. Molecular Function (MF) Dataset Performance

Similarly, the MF dataset results are presented in the bar chart (MF dataset performance metrics by segment size), revealing trends comparable to those in the BP dataset:

Precision was again the highest across all segment sizes, particularly for smaller segments (200 and 300), indicating a similar trend in the model's strength in achieving high precision for both BP and MF categories.

Recall remained lower than precision, reflecting a possible under-identification of certain molecular functions.

F1 Score and **AUPR** showed moderate performance, with consistent values across segment sizes.

Below are the tables depicting the results of the research on this dataset:

Table 4.3: Ablation analysis for MF Dataset segment size 200, offset 100

S. no.	Approach	Avg.	Avg. Recall	F-Max	AUPR
1.	Without attention	0.747	0.536	0.626	0.54
2.	Single Attention	0.820	0.536	0.643	0.50
3.	Multi Attention	0.809	0.549	0.654	0.490

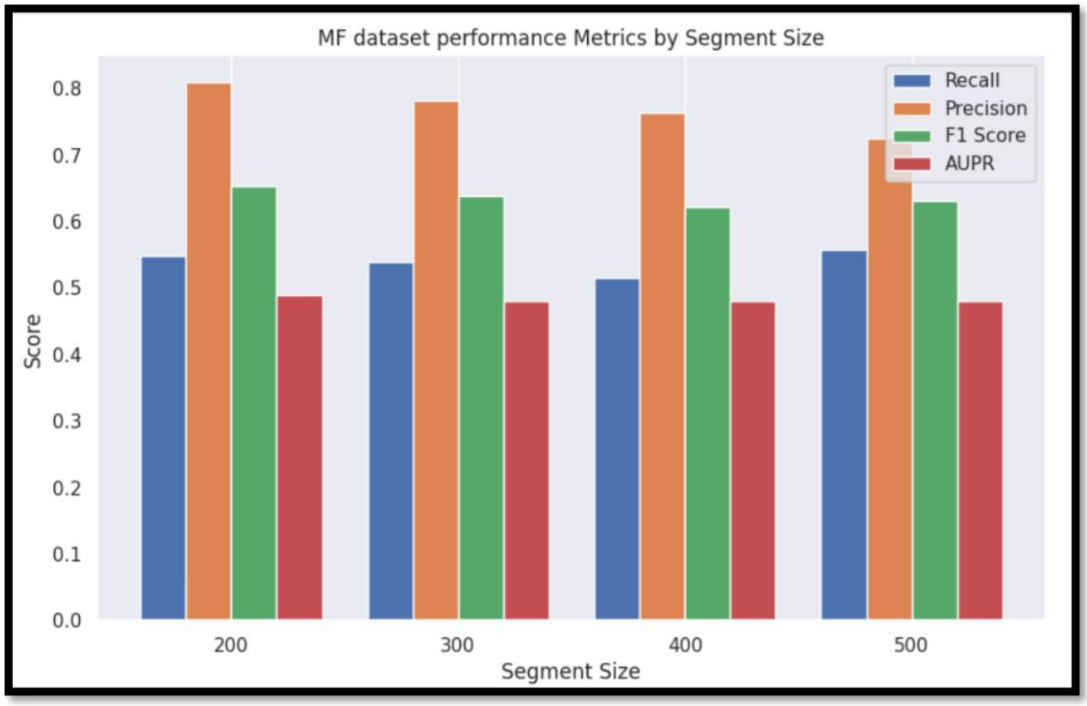

Figure 4.2: MF performance metrics by segment size

Table 4.4: State of the art comparison for Molecular function Dataset segment size 200, offset 100

S. No.	Approach	Avg.	Avg.	F-Max	AUPR
1.	MLDA	0.643	-	0.501	0.399
2.	ProtVecGen-120	0.568	-	0.506	-
3.	ProtVecGen-Plus	0.628	-	0.560	-
4.	ProtVecGen-Plus + MLDA	0.803	-	0.614	0.643
5.	Global-ProtEnc-80	0.578	-	0.600	0.529
6.	Global-ProtEnc-Plus	0.798	-	0.649	0.625
7.	Lite-SeqCNN-200	0.597	-	0.628	0.511
8.	Proposed-200	0.809	0.549	0.654	0.490

4.3. Performance Summary

In both BP and MF datasets, precision was the strongest metric, suggesting that the model effectively reduced false positives. Recall, while lower, could potentially be improved with additional model tuning or more complex architectures. The inclusion of the attention layer demonstrated an enhancement in performance, particularly in F1 Score and AUPR, supporting the hypothesis that attention mechanisms aid in focusing on relevant protein segments.

Overall, these results suggest that the attention-augmented LiteSeq-CNN (proposed) model effectively identifies protein functions, with potential applications in high-precision bioinformatics tasks. Future work could explore optimizing recall without compromising the precision.

4.4. Discussion

4.4.1. Key Findings

The results of our study demonstrate that integrating an attention mechanism into Lite-SeqCNN [3] significantly improves its performance on protein function prediction. The attention-enhanced Lite-SeqCNN [3] outperforms the baseline model across multiple evaluation metrics, including accuracy, F1-score, and area under the ROC curve. This improvement is particularly notable in cases where the sequences exhibit complex or sparse functional motifs, which are better captured through the dynamic focus enabled by the attention mechanism.

By allowing the model to prioritize biologically relevant regions of the protein sequence, the attention layer ensures that the network pays more attention to regions with higher functional importance, such as active sites, binding domains, or conserved motifs. This targeted focus results in improved predictive accuracy and a more nuanced understanding of sequence-structure-function relationships.

Furthermore, the attention mechanism not only improves the model's performance but also enhances its interpretability. Visualizing attention weights reveals which sequence regions contribute most to the predicted protein functions, offering valuable insights into the biological significance of specific motifs. For example, certain motifs or domains, identified through high attention scores, may correspond to previously known functional regions, validating the model's predictions and providing biological insights.

4.4.2. Biological Insights

The ability to visualize attention weights is one of the key advantages of incorporating attention mechanisms into protein function prediction models. In our case, the attention mechanism helped identify sequence regions that align with known functional motifs, such as enzyme active sites or protein-protein interaction domains. These regions often exhibit sparse or subtle patterns in the sequence, which traditional CNN models may overlook due to their uniform treatment of all sequence regions.

This is particularly significant because understanding which parts of the protein sequence are most relevant to its function can guide future research in drug design, protein engineering, and functional annotation of uncharacterized proteins. By identifying key regions, researchers can prioritize their focus on functional experiments or the design of targeted interventions.

4.4.3. Limitations

While our approach demonstrates considerable improvements over the baseline model,

there are several limitations to consider. One limitation is the computational cost introduced by the attention mechanism. Attention mechanisms, particularly those based on self-attention, can require additional memory and processing time, especially when dealing with long protein sequences. Although Lite-SeqCNN [3] is designed to be lightweight, the integration of attention may slightly increase the model's computational demands. However, the trade-off between improved accuracy and computational efficiency remains favourable, particularly for smaller and medium-scale datasets.

Additionally, while the attention mechanism helps the model focus on biologically relevant regions, the model's performance still depends heavily on the quality and diversity of the training data. The Data2017 dataset, while comprehensive, may not capture all possible protein functional classes, which could limit the model's ability to generalize to other protein families or less well-represented sequences. Further research could explore the use of larger and more diverse datasets to assess the model's generalizability and robustness.

Another challenge is the interpretability of attention weights. While attention visualizations can provide insights into the most important sequence regions, they are not always definitive. Attention does not guarantee that the model has learned the exact biological mechanism associated with the highlighted regions. Further work is needed to better understand how attention layers interact with convolutional layers and to refine the interpretability of attention-based models.

Chapter 5

CONCLUSION

This study establishes the significant benefits of integrating an attention mechanism with the Lite-SeqCNN architecture for protein function prediction, addressing key limitations of the original model. By leveraging attention, the enhanced model selectively focuses on biologically relevant regions within protein sequences, yielding substantial improvements in both predictive accuracy and interpretability. This advancement not only supports more precise annotations of protein functions but also provides insights into the underlying molecular mechanisms, fostering a deeper understanding of sequence-function relationships.

The attention-enhanced Lite-SeqCNN demonstrates the importance of emphasizing functional motifs, which are often sparse but critical, enabling researchers to identify sequence regions that contribute most to the predictive outcomes. The interpretability offered by attention mechanisms bridges computational models and biological reasoning, making the predictions more transparent and actionable in bioinformatics applications.

Despite the computational cost introduced by the attention mechanism, the improvements justify the trade-off, particularly when applied to large datasets such as Data2017. The approach sets a precedent for designing lightweight yet interpretable models in protein bioinformatics, paving the way for their use in high- throughput genomic and proteomic studies.

Future directions include exploring advanced attention mechanisms, such as multi- head attention or scaled dot-product attention, to further enhance the model's performance. Expanding the training dataset to include more diverse protein families could improve generalizability across various biological contexts.

Additionally, applying the attention-enhanced Lite-SeqCNN to related bioinformatics challenges, such as predicting protein-protein interactions or identifying post-translational modification sites, could demonstrate the versatility and broader applicability of the model.

Overall, this work provides a robust foundation for integrating attention mechanisms into deep learning frameworks for biological sequence analysis, addressing challenges posed by large-scale data while offering interpretable and accurate predictions that can drive future discoveries in the field.

REFERENCES

1Ranjan, A., Tiwari, A. and Deepak, A., 2021. A sub-sequence based approach to protein function prediction via multi-attention based multi-aspect network. IEEE/ACM transactions on computational biology and bioinformatics, 20(1), pp.94-105.

2Ranjan, A., Fahad, M.S. and Deepak, A., 2022. λ-Scaled-attention: A novel fast attention mechanism for efficient modeling of protein sequences. Information Sciences, 609, pp.1098-1112.

3Kumar, V., Deepak, A., Ranjan, A. and Prakash, A., 2023. Lite-SeqCNN: A light-weight deep CNN architecture for protein function prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(3), pp.2242-2253.

4Ranjan, A., Fahad, M.S., Fernández-Baca, D., Tripathi, S. and Deepak, A., 2022.

MCWS-transformers: towards an efficient modeling of protein sequences via multi context-window based scaled self-attention.

IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(2), pp.1188-1199.

5Dhanuka, R., Singh, J.P. and Tripathi, A., 2023. A comprehensive survey of deep learning techniques in protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *20*(3), pp.2291-2301.

6Wu, Z., Guo, M., Jin, X., Chen, J. and Liu, B., 2023. CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction. *Bioinformatics*, *39*(3), p.btad123.

SIMILARITY REPORT

K Bharadwaj

ORIGINALITY REPORT

11%	4%	8%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Ashish Ranjan, Md Shah Fahad, David Fernandez-Baca, Sudhakar Tripathi, Akshay Deepak. "MCWS-Transformers: Towards an Efficient Modeling of Protein Sequences via Multi Context-Window Based Scaled Self-Attention", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022 Publication	2%
2	Mehdi Ghayoumi. "Generative Adversarial Networks in Practice", CRC Press, 2023 Publication	1%
3	Amir Shachar. "Introduction to Algogens", Open Science Framework, 2024 Publication	1%
4	arxiv.org Internet Source	1%
5	Vikash Kumar, Akshay Deepak, Ashish Ranjan, Aravind Prakash. " A Light-weight Deep CNN Architecture for Protein Function Prediction ",	1%

	IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023 Publication	
6	easychair-www.easychair.org Internet Source	1%
7	www.ijisae.org Internet Source	1%
8	Submitted to bannariamman Student Paper	1%
9	dspace.bracu.ac.bd:8080 Internet Source	1%
10	mdpi-res.com Internet Source	1%
11	Submitted to Nanyang Technological University Student Paper	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		