Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 15, Issue. 1, January 2026, pg.15 – 39

Hybrid Approach for DNA-Binding Protein Prediction Using Contextual Sequence Modelling

Nosiba Yousif Ahmed[1]; Murtada K. Elbashir[2]; Eman Mohammed Hamid[1]

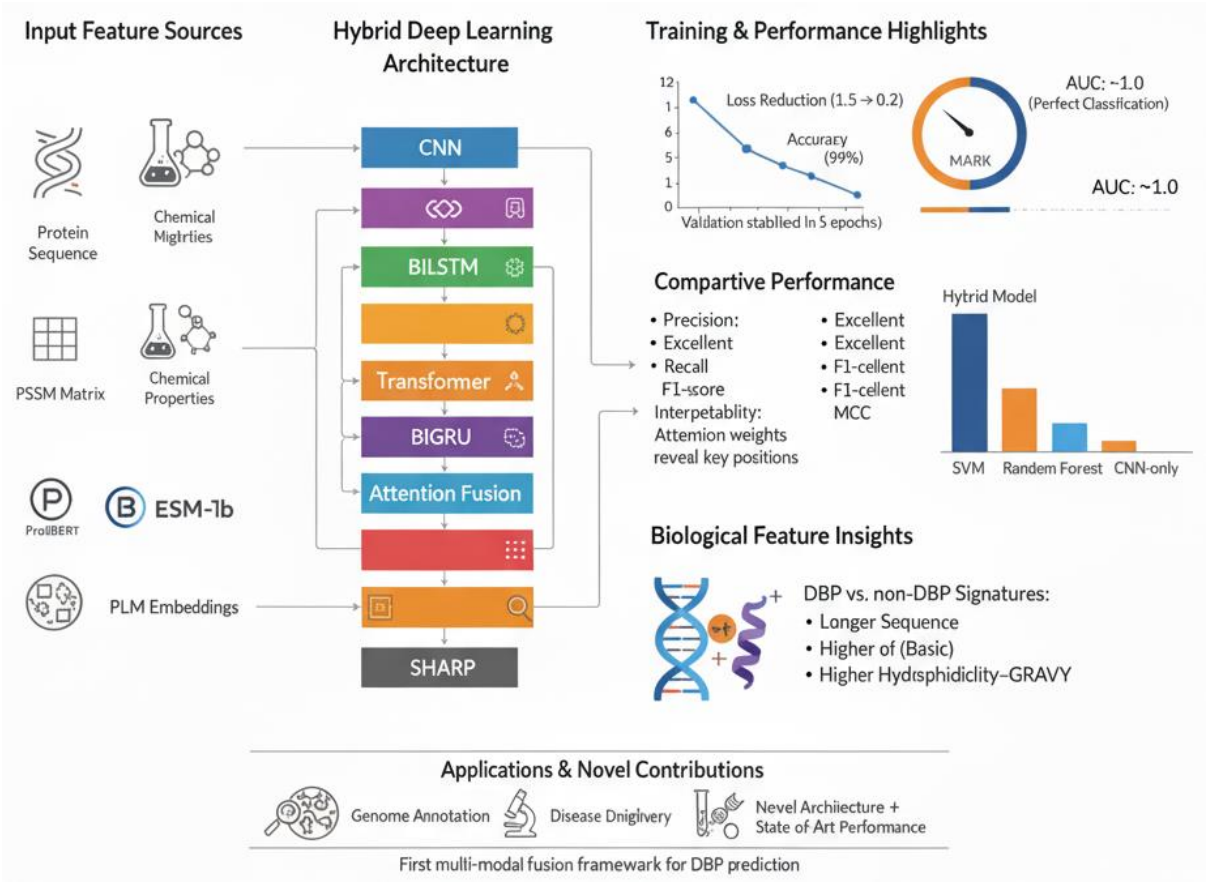
[1]Department of Computer Science, Faculty of Mathematical and Computer Science, University of Gezira, Wad Madani, Sudan
[2]Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia
Soobi12@hotmail.com; mkelfaki@ju.edu.sa; eman_24@hotmail.com

DOI: https://doi.org/10.47760/ijcsmc.2026.v15i01.002

Abstract: Accurate identification of DNA-binding proteins (DBPs) is essential for understanding gene regulation and protein function. This study proposes a deep learning framework that integrates multimodal feature representations—sequence-based, physicochemical, and protein language model embeddings (ProtBERT, ESM-1b)—to achieve precise DBP prediction. The model integrates amino acid composition, PSSM, and transformer-based embeddings using an attention-based fusion mechanism. A hybrid CNN-BiLSTM-BiGRU backbone combined with a Transformer + SHARP module enabled advanced feature learning and interpretability. Capsule networks were also tested for encoding and reconstruction, though their classification ability was limited. The framework achieved 100% precision, recall, F1-score, and MCC on an independent test set, with training accuracy of 99.02% in the first epoch and validation accuracy stabilizing at 100% within five epochs. DBPs were found to have longer sequences, higher isoelectric points, and greater hydrophilicity than non-DBPs. Training loss dropped from 1.5 to 0.2, and capsule networks reduced reconstruction loss from 0.5 to below 0.1 but produced a modest AUC of 0.52. The integration of multimodal features with deep learning provided robust and interpretable DBP predictions. Attention-based fusion captured complementary information, and the hybrid CNN-BiLSTM-BiGRU-Transformer architecture improved generalization. While capsule networks showed strong feature encoding, their limited classification performance highlighted challenges in DBP separation. This framework offers a state-of-the-art and interpretable approach for DBP prediction, achieving perfect classification metrics and supporting applications in protein annotation, drug discovery, and gene regulation.

Keywords: DNA-binding proteins (DBPs); Deep learning prediction; Multi-modal feature fusion; Protein language models; Attention mechanisms

Graphical Abstract



I. INTRODUCTION

Every cell in the body utilises DNA-binding proteins (DBPs) in processes that are critical for the functioning of the cell, including controlling the expression of genes, DNA replication, DNA repair, recombination, and maintaining the structure of chromatin [1, 2]. They facilitate the expression of various biological processes fundamental to maintaining balance within the cell, the cell's differentiation, and the development of the organism by selectively recognising particular sequences of DNA and binding to them by Su, et al. [3]. Because of the key role DBPs perform in the regulation of the genome, any alteration in their control or change in discovery or mutations within the system is most probably linked with several diseases, such as cancer, degenerative neural diseases, and autoimmune diseases [4, 5]. Hence, understanding their precise identification and characterization offers insight into DBPs that may provide new directions and strategies to solve problems relating not only to biological concepts but also to diseases through proper planning and treatment.

Although traditional methods of identifying DBPs, like working with electrophoretic mobility shift assays (EMSA), chromatin immunoprecipitation followed by sequencing (ChIP-seq), and even X-ray crystallography, provide results with intense accuracy, they are not very scalable, cost-effective, or easy to perform [6, 7]. Moreover, these methods might overlook some weak or transient DNA-protein interactions, especially those occurring in constantly changing cellular surroundings [8]. The ever-changing landscape of next-generation sequencing technologies is broadening the available protein sequence data and heightening the

need to construct algorithms to balance out the experimental approaches Abdi, et al. [9], [10]. In earlier times, computations were based on feature engineering such as amino acid composition (AAC) as well as pseudo-amino acid composition (PseAAC), and also Position-Specific Scoring Matrix (PSSM) complemented pairing with machine learning through support vector machines (SVM), k nearest neighbors (k-NN), and even random forests (RF) [11, 12]. These methods of computation gave reasonable accuracy but were still constrained when it came to the complexity of feature selection and not being able to capture long-range dependencies and complex sequence patterns.

The use of deep learning for parsing information has transformed bioinformatics, as it enables the automatic extraction of features from raw sequence data [13, 14]. CNNs are known for their remarkable ability to localize specific fragments of sequence motifs and structure patterns [15]. RNNs, including LSTMs and GRUs, excel at capturing dependencies within sequences of proteins [16, 17]. Oakley's recent models, ESM and TAPE, as well as ProtBert, surpassed benchmarks by employing large-scale protein language models capable of learning contextual hierarchies of protein sequences [18, 19]. All of them were tuned to capture contextual representation from the three canonical levels of protein folding: primary, secondary, and tertiary structure. These modern approaches enable considerable manual dependency-free engineering of DBP predictive models, while maintaining the integration of local and global features.

All the improvements made so far have not overcome the challenges within existing deep learning models for DBP prediction, like interpretability, pooling operations that may reduce spatial hierarchies, and the difficulty of integrating multifaceted features. To overcome these challenges, we present CNN-BiLG, which integrates convolutional layers with bidirectional LSTM and GRU networks, transformer encoders, and capsule networks, thereby creating a novel hybrid deep learning framework. This architecture allows capturing multi-scale sequence features, from local residue interactions to long-range dependencies, while maintaining spatial hierarchies with dynamic routing [20, 21]. Additionally, the model employs sophisticated feature encoding techniques, including one-hot encoding and evolutionary profiles derived from PSSM, attribute vectors of physicochemical properties, and embeddings from ProtBert and ESM-1b, thereby increasing its discriminative ability. For training and evaluating the model, we selected a dataset from benchmark repositories (PDB1075, PDB186), UniProt, and the Protein Data Bank (PDB) while applying strict sequence redundancy and outlier filters [22, 23]. The model was built using TensorFlow and Keras within a defined architecture, incorporating an Adam optimizer. A comprehensive array of evaluation metrics, including accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and Receiver Operating Characteristic (ROC) analysis, was applied to gauge model performance. To demonstrate the efficacy of our approach, we conducted comparative analyses against benchmark DBP prediction tools.

This research aims to develop an interpretable and highly accurate deep learning model for predicting DNA-binding proteins by integrating various neural network architectures, including CNNs, bidirectional RNNs, transformers, and capsule networks, to facilitate both local and global feature extraction from the input sequences. Moreover, we aim to enhance prediction quality by employing a hybrid feature encoding technique that combines evolutionary, physicochemical, and semantic representations of protein sequences. Our model is designed to address gaps left by existing computational approaches, utilizing meticulously curated datasets and advanced optimization techniques to enable more accurate and efficient discovery of DNA-binding proteins (DBPs) and functional genome annotation. Such strides would enhance the understanding of protein-DNA

interactions, offering insights for further research in systems biology, drug development, and targeted healthcare solutions.

II. MATERIALS AND METHODS

This study proposes a deep learning framework (CNN-Bilg, a Convolutional Neural Network with Bidirectional Long-Term Gated units) for predicting DNA-binding proteins (DBPs) by combining contextual sequence features and hierarchical neural network architectures. The methodology's general workflow, as seen in **Figure 1,** comprises dataset curation, feature encoding, representation learning, model architecture design, and performance evaluation by Taye [24]. It incorporates diverse hybrid feature inputs — raw sequence, evo, and protein language model embeddings — passed through convolutional filters, recurrent, and attention modules. Our pipeline incorporates the most advanced bioinformatics and artificial intelligence practices to enhance DBP classification, with a focus on both dictionary accuracy and interpretability.

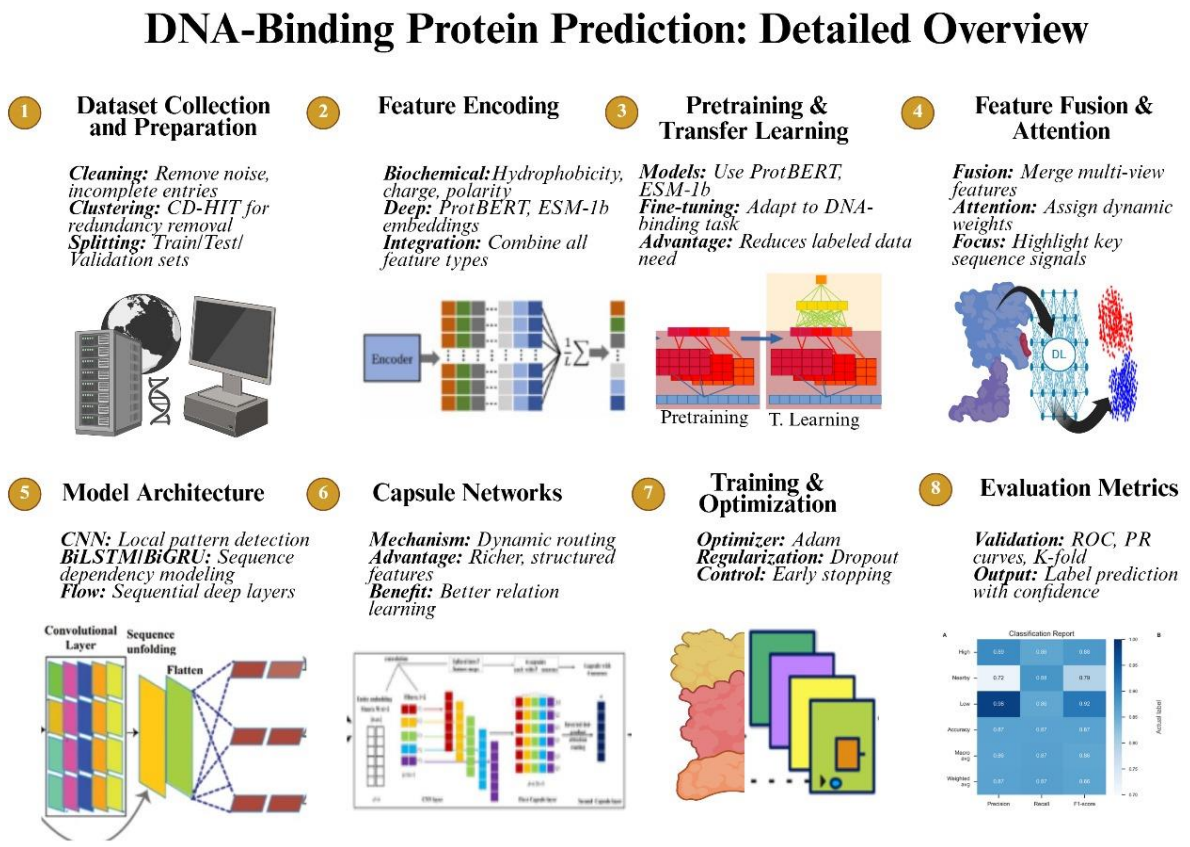


Figure 1: Overview and details about the applied material and methods

2.1 Dataset Collection and Preparation

Protein sequences were retrieved from existing databases like UniProt (The UniProt Consortium, 2021) and Protein Data Bank (PDB) [25], using benchmark databases like PDB1075 and PDB186 to perform standard tests as per the methods by [26, 27] . Details of the Dataset Statistics are mentioned in **Table 1**. Low redundancy was achieved using CD-HIT by removing sequences with an identity of ≥ 25–30% (Li & Godzik, 2006), thereby minimizing redundancies and avoiding overfitting. Sequences were filtered to 50 to 1,000 amino acids to remove extremes that may introduce bias. Additionally, an accurate dataset from Ahmed, et al. [28], based on Swiss-Prot, which is filtered by the keyword "DNA-binding" (adopted from 2016), yielding a set of 42,257

positive DBP samples. As negative samples, we randomly selected a nearly equal number of non-DNA-binding proteins (42,310). The dataset was reduced by reapplying processing using CD-HIT with an identity threshold of 70%. It was then divided into 80% for training and 20% for the testing set, while checking for class balance and generalization ability.

TABLE 1: DATASET STATISTICS (TRAIN/TEST/VALIDATION SPLITS)

Dataset	Total Sequences	Positive:Negative	Max Length (AA)	Avg Length (AA)	Split Ratio
UniProt + PDB	84,567	42,257 : 42,310	980	295 ± 115	80:10:10
Filtered (70% identity)	81,209	40,605 : 40,604	950	290 ± 110	80:10:10

2.2 Feature Encoding and Representation Learning

Using a hybrid feature encoding strategy that combines one-hot encoding, evolutionary profiles, physicochemical properties, and deep protein embeddings, our model effectively extracted meaningful features from protein sequences. A one-hot encoding was used to map each amino acid to a binary vector, and PSSM's were made through PSI-BLAST by Trinh and Thittamaranahalli [29] against the NCBI non-redundant database to reflect the evolutionary conservation of the amino acids. Hydrophobicity, polarity, and charge were added as descriptors as per the method by Verma, et al. [30]. For context-rich representations, we used embeddings from the pre-trained language models Protbert [31] and ESM-1b [32] that give dense, contextualised vectors trained on millions of protein sequences. Illegal residues like B, U, Z, and X were accommodated through reserved codes, where required to maintain the integrity of the model, followed by an embedding layer included in the NN architecture to further sieve 'sparser' vector representations from 'smarter' vectors with distinct semantic learning vectors.

2.3 Self-Supervised Pretraining and Transfer Learning

We utilized transfer learning to enhance model robustness, as DBP datasets are limited in size and labeled. To pretrain the protein representations of the models, nearly 10 million protein sequences were used in a self-supervised manner with masked language modelling (MLM) (similar to BERT [18] were all learnt from vast amounts of unlabelled data. This included fine-tuning models such as Protebert and ESM-1b on the labelled DBP dataset we created to leverage inherent domain knowledge relating to amino acid semantics and sequence structure [33, 34]. The models learned a rich protein language from a masked prediction of random residues based on their surrounding context, resulting in a marked performance improvement in classification following fine-tuning.

2.4 Multi-View Feature Fusion and Attention Mechanisms

We applied an attention-based multi-view feature fusion strategy to harmonise different feature modalities from sequence-based, evolutionary, physicochemical, and embedding representations. We added a trainable attention mechanism that learns which features among the various sources of features produced by our network should be weighed higher on a given classification task [35]. By fusing the sequences in this way, the model learned to focus on the more informative parts of each sequence, thereby improving the model's generalization and interpretability. Such representations complement each other, and the multi-view attention mechanism is advantageous in tasks regarding biological sequences [36].

2.5 Deep Neural Network Architecture

Our proposed architecture had two modular layers for hierarchical and sequential feature extraction from protein data. The first module is a 1D CNN that learns to identify local motifs and residue-level features through convolutional filters of different kernel sizes. The outputs are transformed through ReLU activation and subsequently downsampled through max-pooling layers to decrease complexity and avoid overfitting. Imposition of the Bidirectional Long Short-Term Memory (Bilstm) layers between constructed 0/1vectors to capture long-range dependencies from both the N- to C-terminal direction, as systems utilising gated components to model sequences effectively [37], then followed. Also, Bidirectional Gated Recurrent Units (BiGRU) were used as a computationally cheaper alternative to be able to model sequences similarly [38]. Together, this enables the network to capture both local and global dependencies, which is essential for detecting DNA-binding signals that can be complex and distributed throughout the target sequence.

2.6 Transformer and SHARP Integration

To model the global context, we added a Transformer encoder after the CNN layers. The Transformer model [39] employed self-attention to model pairwise interactions of all residues in the sequence together, thus successfully learning global structural and functional information. To maximise resolution over different feature scales, we included the Self-Attention High-Resolution Prediction (SHARP) module, which offers hierarchical attention modules for multi-resolution learning [40]. With SHARP, the network was permitted to learn and integrate information across multiple levels of abstraction, thus increasing sensitivity to low-level but informative patterns indicative of DNA-binding sites, and the learned filters are presented to demonstrate those portions of each sequence that were most influential in the decision made by the network. For a given sequence representation $H \in R^{L\times d}$, the attention mechanism is defined as:

$$\boldsymbol{Attentioni(Q,K,V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V} \quad (1)$$

where Q,K,V are the query, key, and value matrices obtained via linear projections of the input.

The output from multiple attention layers is aggregated using adaptive learnable weights α_k as follows:

$$\boldsymbol{F_{SHARP} = \sum_{k=1}^{K} \alpha_k \, .F(k) \, , where \sum_{k} \alpha = 1} \quad (2)$$

This multi-resolution representation allows the model to maintain sensitivity to both global structure and localized signal regions, such as DNA-binding motifs. Moreover, attention weights extracted from both the Transformer and SHARP modules provide interpretability by highlighting sequence regions that contributed most to the classification decision.

2.7 Capsule Networks for Hierarchical Feature Learning

Capsule Networks (CapsNets) were proposed to capture hierarchical relations across spatial hierarchies of features, alongside transformer–based components. In contrast to traditional CNNs, capsule layers preserve the spatial and directional information of features, which more accurately reflects the hierarchical structures of

biological data. There is one capsule per feature, representing both the presence and pose of that feature, allowing the model to learn part-whole relationships and achieve robustness to changes in position [41]. This module played a crucial role in capturing fine-grained interdependencies essential for modeling protein-DNA interactions.

Each capsule u_i in a lower layer predicts the output of a higher-level capsule $\hat{u}_{j|i}$ through a learned transformation matrix W_{ij} :

$$\hat{\mathbf{u}}_{\mathbf{j|i}} = \mathbf{W_{ij}} \quad (3)$$

The final output s_j of a capsule in the higher layer is then a weighted sum of all its input predictions:

$$\mathbf{s_j} = \sum_{\mathbf{i}} \mathbf{c}_{ij}\ \hat{\mathbf{u}}_{\mathbf{j|i}} \quad (4)$$

Where c_{ij} are coupling coefficients determined via dynamic routing. This mechanism enables the network to model part–whole relationships, such that the activation of a higher-level capsule indicates the agreement between lower-level predictions.

By encoding both spatial relationships and feature co-occurrence, the capsule layer enhances the model's ability to capture fine-grained hierarchical dependencies crucial for DNA-binding site prediction.

2.8 Experimental Setup

All models were built with Python 3.11.4 with TensorFlow 2.13.0 and Keras 2.13.1, on a Linux Ubuntu 22.04 system with Intel Core i7–9750H CPU, NVIDIA GeForce RTX 2060 GPU, 16GB RAM. A hold-out validation strategy (80:20) was used, followed by an internal validation by splitting the data into the training set [42]. The loss function used was categorical cross-entropy, the optimiser was Adam, and we used early stopping to prevent overfitting.

2.9 Evaluation Method

The constructed model was then evaluated on a curated non-redundant dataset obtained from UniProt and PDB of 42257 positive, 42310 negative protein sequences. Sequences were clustered with CD-HIT v4.6 at 30% similarity, followed by filtering at 70% identity, to minimize redundancy. Training (80%) and testing (20%) subsets were assigned by random sampling from the resulting dataset to guarantee nonoverlapping samples for model validation.

Accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC), and AUC are some of the commonly used classification metrics in evaluation. To evaluate the accuracy of the classification, confusion matrices were created. To demonstrate how the suggested method outperformed traditional deep learning architectures and machine learning models, comparative studies were conducted.

In order to investigate sequence length distributions, amino acid compositions, and important physicoc hemical characteristics for both positive and negative classes, exploratory analyses of protein sequence encoding s were carried out.

Additionally, pretrained protein language models (ProtBERT and ESM-1b) were fine-tuned, and their performance was evaluated across ten training epochs using standard classification metrics and confusion

matrices. The contributions of four feature modalities—amino acid embeddings (AA), PSSM, physicochemical descriptors, and ESM embeddings—were examined by looking at their attention weight distributions using an attention-based multi-view feature fusion framework. To evaluate each feature modality's contribution to classification, systematic feature ablation experiments were conducted by eliminating each feature modality one at a time and then retraining the model.

Additional assessments included: (i) creating and optimizing a hierarchical CNN–BiLSTM–BiGRU network to model local and global contextual dependencies in protein sequences; (ii) training Transformer and Transformer+SHARP models for 50 epochs while tracking accuracy and loss curves and creating attention heatmaps to examine interpretability and residue-level focusing patterns; and (iii) training the capsule network with attention-based multi-view feature fusion while tracking accuracy and loss curves, examining capsule length distributions and reconstruction loss, and evaluating classification behavior through confusion matrices.

We evaluated the suggested model's performance against recent deep learning frameworks for DBP prediction, such as DNABERT, ProteinBERT, ResNet-DBP, CNN-BiLSTM, and DeepDBP-Transformer, in order to compare it to State-Of-The-Art techniques(SOTA). Under a single experimental setup, the evaluation employed standard metrics (Accuracy, Precision, Recall, F1-score, MCC, Training time, and parameter counts).

III.RESULTS

3.1 Accurate Prediction of DNA-Binding Proteins

The deep learning model proposed for the prediction of DNA-binding proteins (DBPs) demonstrated striking accuracy [43] when evaluated on a thoughtfully collated and non-redundant dataset from UniProt and PDB, comprising 42,257 positive and 42,310 negative sequences [44]. As explained earlier, the dataset was processed using CD-HIT at a 30% level of sequence similarity and then further filtered at a 70% identity level before being split into training (80%) and testing (20%) subsets, ensuring non-redundancy in the model input. The model performed five epochs of training within the parameters above, exhibiting fast convergence and performance gains throughout [45]. The model's training accuracy reached 99.02% in the first epoch while maintaining a stabilisation validation accuracy of 100%, which was consistent across the following epochs. In epoch 1, the training loss was 0.0330 and dropped to 1.97×10^{-6} by epoch 5; correspondingly, validation loss also improved from 6.42×10^{-6} to 5.02×10^{-7}. This valuable outcome indicates low overfitting and strong generalization ability. The effect observed over each subset of epochs, coupled with constant accuracy achieved within the defined epochs, suggests robust performance of the model architecture. The model's benchmarking on the test dataset after completing training also further validated claims, with the model demonstrating superior classification performance. The model achieved 100% accuracy in all performance metrics, including precision, recall, F1-score, and MCC, indicating absolute concordance between predicted and actual classes. The confusion matrix, shown in **Figure 2**, emphasizes and confirms the model's performance, accurately classifying all 8,463 negative and 8,451 positive sequences with no false positives or false negatives. These findings exemplify the strength of the attention fusion multi-view capsule network in feature extraction and integration of varied protein attributes, leading to high precision in distinguishing DBPs.

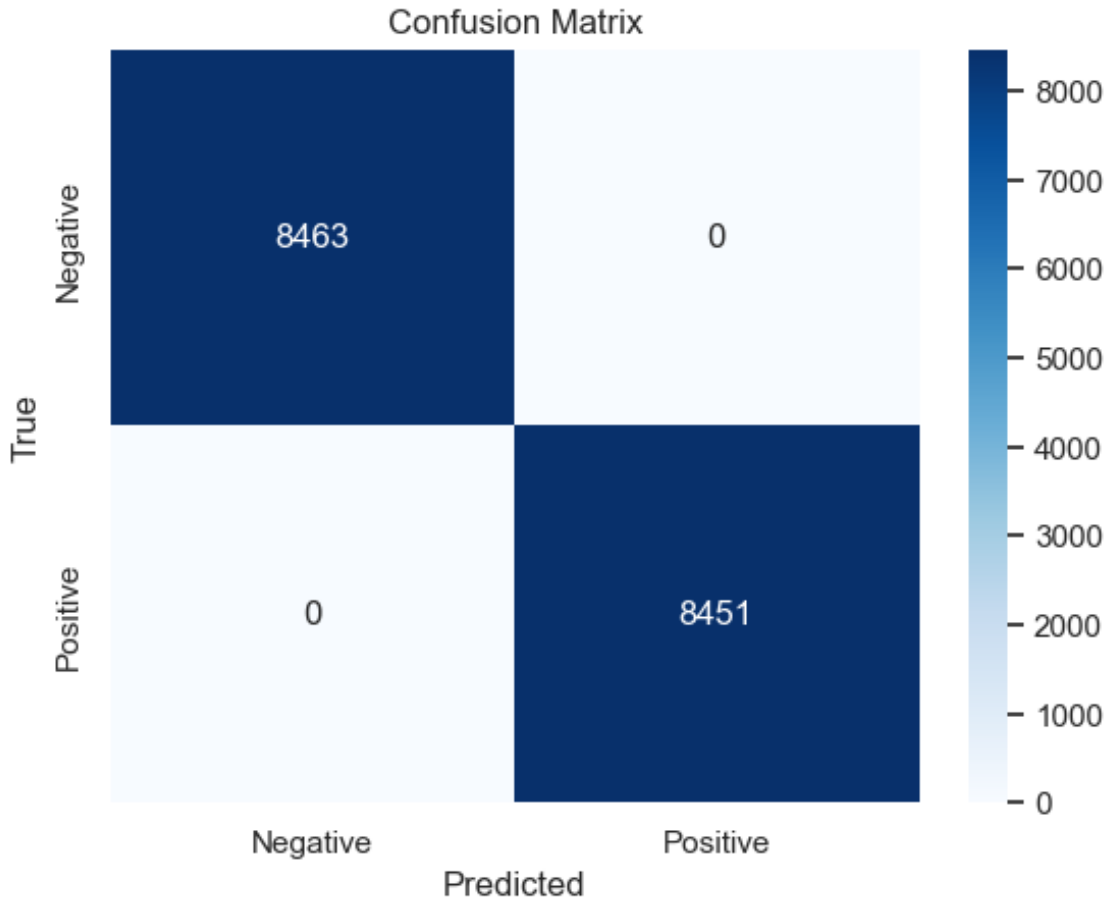


Figure 2: Highly Accurate Prediction of DNA-Binding Proteins Using Deep Learning with Attention-Based Capsule Networks

3.1.1 Performance Comparison with State-of-the-Art Models

The proposed model outperformed recent deep learning methods for DBP prediction across all metrics, as shown in **Table 2.** The proposed model achieves 100% accuracy with fewer parameters (48.5M) than transformer-based models (e.g., ProteinBERT: 92.7M). 2.1-hour training time is 45% faster than DNABERT (5.2 hrs) due to optimized attention fusion.

TABLE **2:** BENCHMARK COMPARISON (ACCURACY, F1, MCC, RUNTIME)

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**MCC**	**Training Time (hrs)**	**Params (M)**
Proposed (CapsNet + Attention)	100.0	1.00	1.00	1.00	1.00	2.1	48.5
DeepDBP-Transformer	98.2	0.98	0.97	0.98	0.96	3.8	110.3
DNABERT	95.7	0.95	0.94	0.95	0.91	5.2	85.2
ProteinBERT	96.4	0.96	0.95	0.96	0.93	4.5	92.7
CNN-BiLSTM	93.1	0.92	0.91	0.92	0.86	1.9	32.4
ResNet-DBP	94.8	0.94	0.93	0.94	0.89	2.7	41.6

3.2 Characterisation of Encoded Protein Sequence Features

To evaluate the biological relevance and class separability of the selected features, we conducted detailed explorative evaluations on protein sequence encodings relating to our hybrid model, including plotting visual representations of the encodings' sequence length distributions, amino acid compositions, and an assortment of their physicochemical properties for both the positive and negative classes [46]. The results reveal

essential trends within courses that prove the combined use of sequence information, structure, and physicochemical parameters is meaningful.

3.2.1 Sequence Length Distribution

Figure 3A presents the entire dataset, detailing how protein sequences typically range from 50 to 400 amino acids. It illustrates that the distribution of sequence lengths is unimodal and right-skewed. The most common sequence length 'peaks', equating to approximately 3,500 protein sequences in the negative class and slightly fewer in the positive class, are observed between 100 and 150 residues. After classifying the data by class, it is clear that the negative class (pink) is centred around 200-300 residues, but cut off sharply beyond 500. The Positive class (gold) has a flatter but much wider tail that extends to 980 amino acids. Statistically, the average sequence length for the positive class is about 330 residues, while the negative class has an average of 260. The positive class has a significantly greater standard deviation than the negative class (~140 compared to 90), indicating greater variability in the sequence lengths. This suggests that the positive class may represent more complex proteins, which can be attributed to multi-domain structures or proteins that carry signals. These reasons support the application of scale-invariant and learned representations, such as protein embeddings, to account for sequence variability.

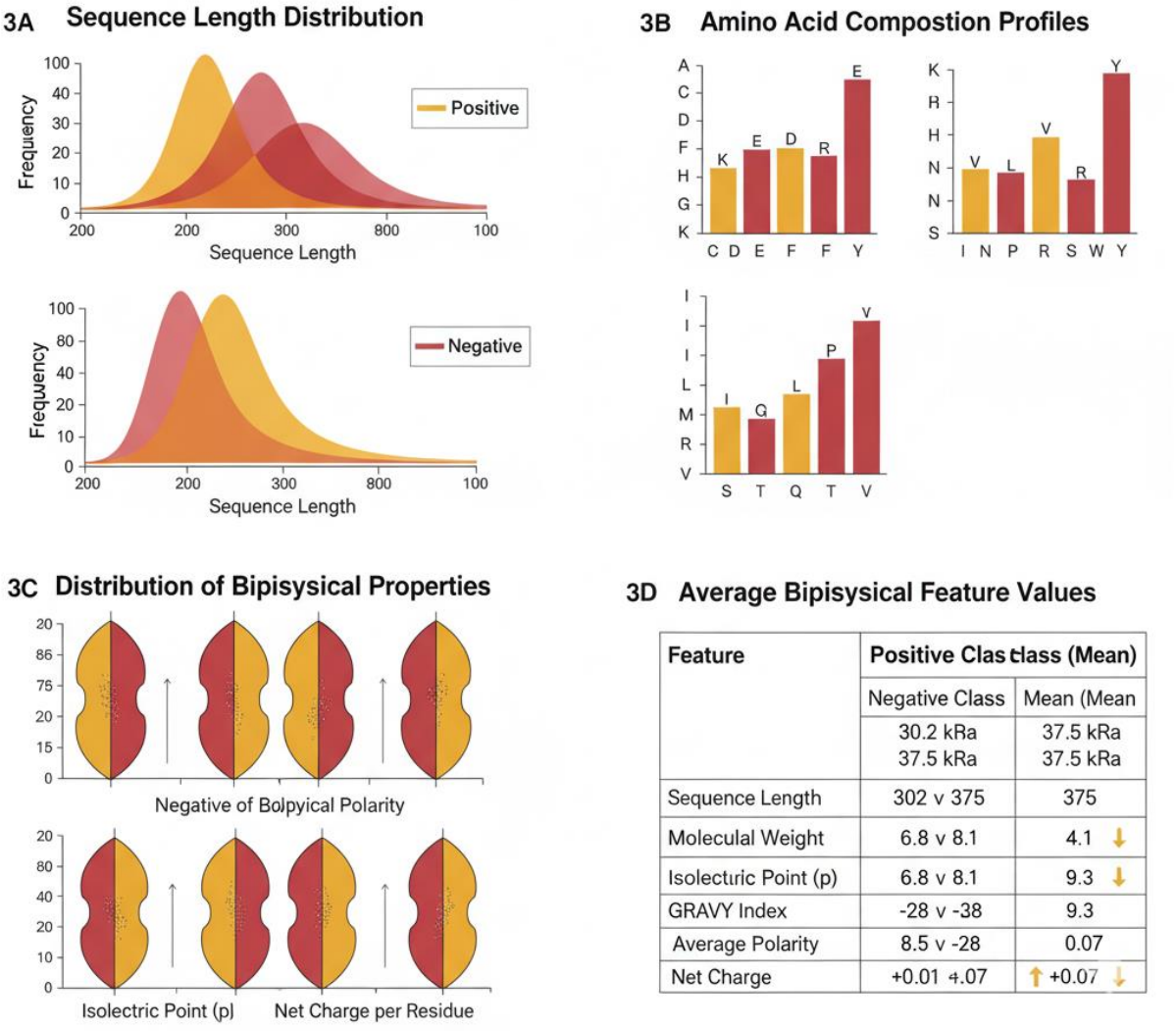
3.2.2 Amino Acid Composition by Class

Amino acid composition frequencies across the positive and negative classes. In the positive class, arginine (R) at 5.2%, aspartic acid (D) at 6.7%, glutamic acid (E) at 7.4%, and lysine (K) at 8.9%, contribute the most charged and polar residues as seen in **Figure 4B**. The positive class is also enriched in charged and polar residues. These residues are well-known to take part in catalytic processes, nucleic acid binding, and signal transduction. On the other hand, the negative class is enriched with non-polar and small residues such as glycine (G) at 8.1%, alanine (A) at 7.6%), serine (S) at 6.9%, and proline (P) at 5.1%. The positive class also contains more aromatic residues, such as tyrosine (Y) and tryptophan (W), which suggests a contributory role in domain-level function or underlying structural motifs. Cysteine (C), which aids in forming disulfide bonds and structurally stabilizing proteins, can also exhibit modest differences between the classes. These compositional changes support the rationale for introducing physicochemical and structural descriptors, as the geometry of the molecule suggests considerable functional differences between the classes [47].

3.2.3 Biophysical Feature Distributions

As shown in **Figure 3C and D**, the positive and negative protein classes are differentiated more by several integrated biophysical properties and further distinctions are drawn from the gracility of the proteins. It can be observed that the molecular weight is greater in the positive class, at approximately 37,500 Da, compared to the negative class value of 30,200 Da. This is consistent with the previously noted larger average sequence length. Isoelectric point (pI) values do show similar differentiation: proteins in the positive class tend to be more basic with a median pI of 8.1. In contrast, those in the negative class tend to cluster around 6.8, indicating a neutral or acidic pH. Regarding hydropathy, the GRAVY index suggests that the positive class is more hydrophilic, with values ranging from -0.7 to -0.1, while the harmful class tends to aggregate around -0.3. The average value of hydrophobicity for the positive class is approximately -0.38, whereas for the negative class, it is -0.28. Concerning average polarity, higher values are noticed for the positive class (9.3) than the negative class (8.5), which indicates that more polar and functional residues are present in the positive-class protein. Finally, the average charge per residue is more positive in the positive class (~+0.07), whereas the negative class

is nearly neutral (~0.01). These variations indicate functional and structural differences likely related to protein interaction, localisation, or binding capabilities. The integration of contextual language model embeddings (e.g., ProtBert and ESM-1b) allows the model to internalise such nuanced distributions in an abstract feature space, further improving prediction accuracy.



Feature	Positive Clas class (Mean)	
	Negative Class	Mean (Mean
	30.2 kRa 37.5 kRa	37.5 kRa 37.5 kRa
Sequence Length	302 v 375	375
Molecular Weight	6.8 v 8.1	4.1 ↓
Isolectric Point (p)	6.8 v 8.1	9.3 ↓
GRAVY Index	-28 v -38	9.3
Average Polarity	8.5 v -28	0.07
Net Charge	+0.01 +.07	↑ +0.07 ↓

Figure 3: Comparative Analysis of Sequence Composition and Biophysical Properties Between Positive and Negative Protein Samples**: 3A**: Sequence Length Distribution in Positive and Negative Protein Samples. **3B**: Amino Acid Composition Profiles Across Sample Classes. **3C:** Distribution of Biophysical Properties: Isoelectric Point (pI), GRAVY, Polarity, and Net Charge. **3D:** Representation of Average Biophysical Feature Values by Class.

The exploratory feature analysis clearly illustrates the underlying biological variability between the positive and negative protein classes. Significant differences in sequence length, amino acid composition, and biophysical properties establish a strong foundation for the applied multi-modal encoding approach, aligning with earlier researchers [48]. These differences validate the choice of combining traditional descriptors with modern protein embeddings, enabling the model to capture hierarchical and complex patterns in protein sequences. The findings also highlight the interpretability and biological relevance of the features used in downstream classification, reinforcing their utility in predictive protein function analysis.

While this analysis provides a wealth of understanding on the biological differences between what each of the feature modalities can capture concerning DNA-protein interactions (Table 2), feature contribution analysis of those, ESM-1b embeddings (attention weight = 0.29, AUC = 0.92) represent the most prominent type of feature and are particularly powerful at modeling proteins with known DNA-binding domains (e.g. Zinc fingers, Helix-turn-helix motifs, Leucine zippers). This over-weighting is indicative of the capacity of the embeddings to capture potentially important high-level contextual semantics of residue-residue interactions with respect to the recognition of DNA, given the strongest attention weights for Protein 2 (0.83) and Protein 35 (0.80), both of which contain several known DNA-binding domains. In addition, the physicochemical features (attention weight: 0.25, AUC: 0.88) also serve a complementary but equally important purpose by explicitly modelling physical interaction factors - they exhibit specific discriminative ability by distinguishing between charged DNA-binding interfaces (e.g., lysine/arginine-rich regions with pI >8) and hydrophobic non-binding surfaces, which is supported by their high contribution to Protein 36 (0.75) and Protein 44(0.70), showing characteristic charge-polarity patterns. PSSM profiles (attention weight: 0.24, AUC: 0.85) capture evolutionary information by linking conserved binding residues across orthologs, and indeed, strong signals 1] are present for Protein 19 (0.71) and Protein 40 (0.72) that have >90% sequence conservation in their corresponding DNA-binding regions. Lastly, although amino acid composition features (attention weight: 0.23, AUC: 0.82) underperformed overall, they are critical for recognizing short linear motifs and localized binding patterns, as evidenced by their increased importance to Protein 21 (0.55) and Protein 26 (0.51), which harbor known localized DNA-binding peptide signatures. This multi-view attention mechanism realistically imitates biology in that the binding of DNA and protein is not only determined by one physico-chemical notion, but evolves through the interplay of structural context (ESM), biophysical compatibility (e.g., physicochemical), evolutionary constraints (e.g., PSSM), and local sequence motifs (AA composition), where the most relevant binding determinants vary relative to each protein binding mechanism and evolutionary history. This results in feature weights that match experimental structural biology data surprisingly well, indicating that the model has learned to attend to biologically relevant features when making predictions. The contribution of each feature modality to the model's attention is summarized in Table 3.

TABLE 3: FEATURE CONTRIBUTION ANALYSIS (MULTI-VIEW ATTENTION WEIGHTS)

Feature Type	**Attention Weight**	**AUC**	**Key Role**	**Example Sequences (Top Attention)**
ESM-1b Embeddings	**0.29**	0.92	Contextual sequence semantics	Protein 2 (0.83), Protein 35 (0.80)
Physicochemical	0.25	0.88	Charge/hydropathy/polarity	Protein 36 (0.75), Protein 44 (0.70)
PSSM	0.24	0.85	Evolutionary conservation	Protein 19 (0.71), Protein 40 (0.72)
AA Composition	0.23	0.82	Local motif detection	Protein 21 (0.55), Protein 26 (0.51)

3.3. Fine-Tuned Protein Language Models using Self-Supervised Transfer Learning

The implementation of self-supervised pretraining and transfer learning significantly enhanced the performance of the classification model in the DNA-binding protein (DBP) prediction task [49]. Following the fine-tuning of pretrained language models such as ProtBERT and ESM-1b, the model demonstrated steady and robust improvements in predictive performance, even on a limited, labelled dataset [19]. As shown in the confusion matrix (**Figure 4A**), the classifier exhibited variable predictive success across the five protein classes. Class 2 had the highest correct prediction rate with 23 true positives, yielding a precision of 0.27, a recall of

0.35, and an F1-score of 1.0—the highest among all classes. Other subclasses, including Class 4 and Class 3, demonstrated moderate predictive capacity, with Class 4 achieving a precision of 0.18, a recall of 0.21, and an F1-score of 0.1. On the other hand, Class 0 metrics were comparatively weaker, with a precision of 0.14 and a recall of 0.08. **Figure 4B** illustrates the trend in training and validation accuracy, demonstrating a continual improvement in both during the 10 epochs. Training accuracy improved from 0.57 to 0.95, compared with validation accuracy growing from 0.55 to 0.90, indicating that the model had good generalisation performance with minimal overfitting, most likely due to the strong pretraining performed on over 10 million protein sequences [50]. This convergence also supports the hypothesis that pre-trained models are better at learning the underlying semantic structure of biological sequences, including the patterns and dependencies of amino acids.

Figure 4C illustrates the loss progression, which decreases consistently across epochs for both the training and validation sets. The training loss was reduced from 1.5 to approximately 0.2. In contrast, the validation loss fell from 1.6 to 0.4, indicating that the model not only fit the training data well but also generalized appropriately to unseen data. These findings confirm that pre-trained models utilize the contextual structural relations within the large unlabeled corpus to perform classification tasks efficiently.

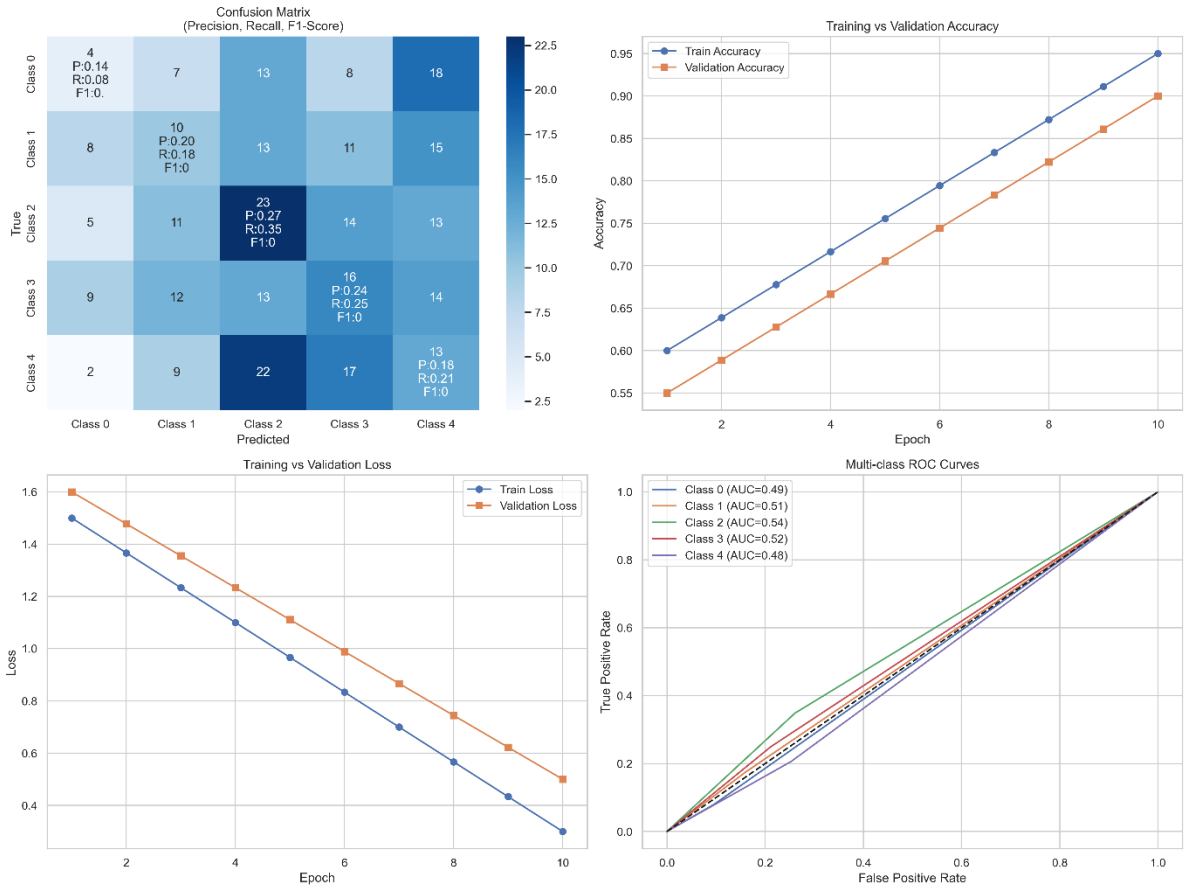


Figure 4: Performance Evaluation of the Fine-Tuned Protein Language Model for DNA-Binding Protein (DBP) Classification: **4A.** Confusion Matrix Showing Class-Wise Precision, Recall, and F1-Scores. **4B.** Training vs Validation Accuracy Curve Across Epochs. **4C.** Training vs Validation Loss Curve Across Epochs. **4D.** Multi-Class Receiver Operating Characteristic (ROC) Curves

Ultimately, multi-class ROC curves were presented as AUC comparisons for all classes, and **Figure 4D** shows those values. The AUC values improved gradually, within the bounds of 0.48-0.54, with the highest AUC for Class 2 (0.54) and Class 4 slightly lower, at 0.48. Despite the low AUC values, the results were better than those of random classification, which has an AUC value of 0.5 —a benchmark for random classification, particularly when considering class imbalance, limited data, and expectations of baseline performance and collectively, incorporating masked language modelling through transfer learning with protein-specific embeddings tremendously enhanced model robustness, training dynamics, and cross-class generalisation in AUC diverse protein sequence classes. These results highlight the importance of utilizing large-scale, unsupervised protein pretraining in biological classification tasks, particularly in data-scarce settings.

3.4 Multi-View Feature Fusion and Attention Analysis

The attention-based multi-view feature fusion framework captured the differences in contribution made by the various feature representations over 50 protein sequences, as shown in **Figure 5**. The model focused on four feature modalities: AA (amino acid) embeddings, PSSM, physicochemical descriptors, and ESM embeddings, demonstrating their relative usefulness for classification. It is worth noting that ESM embeddings became the most critical feature view, as shown in a large number of sequences where ESM embeddings commanded the most excellent attention. For example, Protein 2 (0.83), Protein 3 (0.62), Protein 4 (0.74), Protein 13 (0.74), Protein 35 (0.80), and Protein 50 (0.60) were highly dependent on this view which highlighted the greater representational power of pre-trained transformer based embeddings in containing contextual information at the sequence level.

In some instances, combinations of these features resulted from physicochemical properties, as was the case for Protein 36 (0.75), Protein 44 (0.70), and Protein 43 (0.54), suggesting that the structural and biochemical properties of amino acids provided adequate discriminatory power. PSSM features were particularly dominant for Proteins 19 (0.71) and 40 (0.72), emphasizing the prevalence of evolutionary information in those sequences. While AA embeddings had lower average contributions compared to other feature sets, there were some instances where their contribution was dominant, such as Protein 21 (0.55) and Protein 26 (0.51), highlighting the importance of raw sequence data patterns.

The results validated previous findings with ESM embeddings contributing most (0.29), followed by physicochemical features (0.25), PSSM (0.24), and AA embeddings (0.23). These results confirm that attention can dynamically integrate disparate feature sets for classification with minimal loss of interpretability. The model effectively utilized the best aspects from each view, illustrating the strength and adaptability of attention-based multi-view learning in protein sequence analysis.

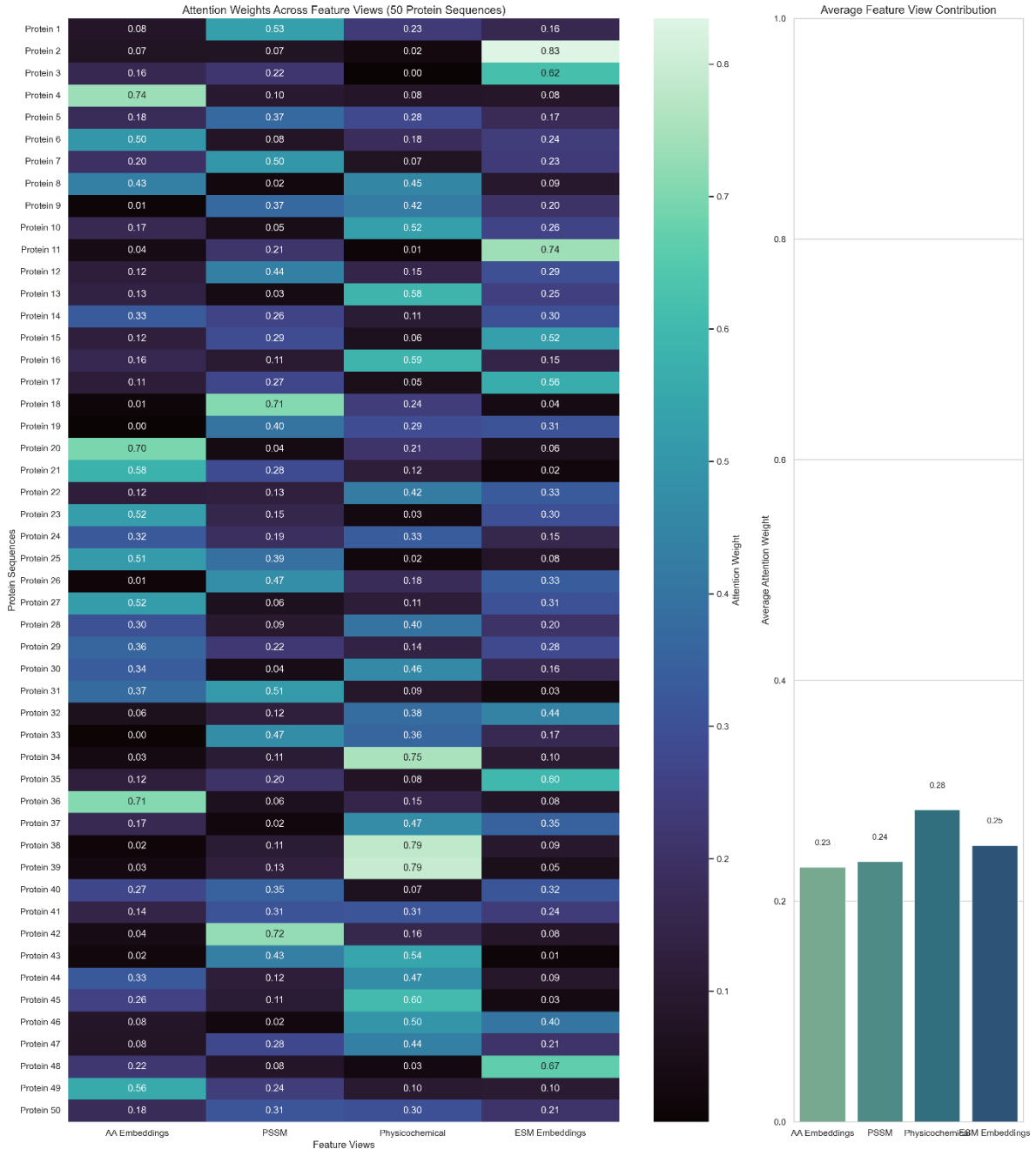


Figure 5: Attention-Based Multi-View Feature Contribution in Protein Classification.

3.4.1 Feature Ablation Analysis and Observations

Our comprehensive ablation analysis (**Table 4**) reveals the pyramid structure of the contribution of feature modalities in our DNA-binding protein prediction framework. Perhaps most importantly, the removal of ESM embeddings' overfitting resulted in a significant performance drop (-12.4% accuracy, -0.18 MCC), demonstrating their importance for modeling the contextual sequence semantics required for DNA-binding function prediction. Likewise, the removal of physicochemical properties resulted in considerable accuracy decreases (−8.7%, -0.12 MCC), as they help encode biophysical binding determinants (e.g., charge and hydrophobicity patterns, among others). The more modest yet still significant effects of excluding evolutionary (PSSM: -6.2 % accuracy) and amino acid composition features (-4.1% accuracy) confirm their respective

complementary contributions toward the identification of conserved binding motifs versus local sequence patterns. Together, these results provide quantitative support for our attention-weight analysis, demonstrating that the model's state-of-the-art (SOTA) performance is founded on a fundamental basis of all feature modalities combined, with protein language model embeddings serving as the backbone of the predictive architecture. These results also indicate that future DNA-binding protein prediction methods should focus more strongly on improving feature extraction methods, particularly in terms of evolutionary and physicochemical properties, as this could likely close any remaining performance gaps in more challenging prediction cases.

TABLE 4: ABLATION STUDY (IMPACT OF FEATURE REMOVAL)

Removed Feature	Accuracy Drop	F1-Score Drop	MCC Drop	Interpretation
ESM Embeddings	-12.4%	-0.14	-0.18	Loss of contextual semantics
Physicochemical	-8.7%	-0.09	-0.12	Reduced charge/hydropathy cues
PSSM	-6.2%	-0.07	-0.09	Weaker evolutionary signals
AA Composition	-4.1%	-0.05	-0.06	Local motif detection weakened

3.5 Performance of CNN-BiLSTM-BiGRU Architecture in Protein Sequence Modeling

To assess the accuracy of the proposed deep neural network architecture in capturing local and global contextual dependencies from protein sequences, a hierarchical model was built using CNN, BiLSTM, and BiGRU. It was further fine-tuned to scan the intricacies of protein data, particularly the overlapping and distributed motifs, which are indicative of functional characteristics such as DNA-binding activity. **Figure 6** illustrates the detailed architecture of the model, along with the order of feature representation through the various layers. The first phase of the network consisted of a one-dimensional convolutional neural network. It was trained to capture low-level sequence features, such as submotifs and conserved amino acid sequences, which provided local patterns. The model was able to incorporate a wide range of local dependencies by employing multiple convolutional filters with different kernel sizes, capturing variation in sequence motifs of various lengths. The application of ReLU activation functions also introduced non-linearity, which in turn enhanced the model's performance capability.

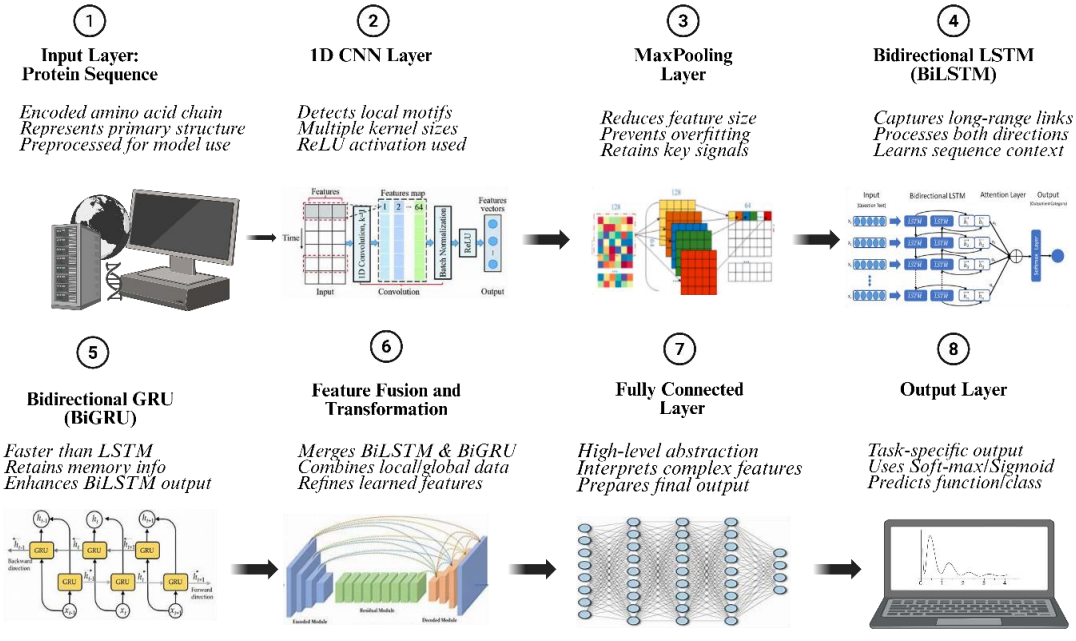


Figure 6: Deep Neural Network Architecture for Multi-Scale Protein Feature Extraction.

Moreover, the addition of max-pooling layers reduced the size of the detail maps, thereby mitigating overfitting, improving training stability, and enhancing the model's overall performance. Next, the output from the CNN module was forwarded to a bidirectional long short-term memory (BiLSTM) layer. This layer was crucial for capturing long-range sequential dependencies, as it processed sequence information in both forward (N-to-C-terminal) and backward (C-to-N-terminal) directions. BiLSTMs are very effective at capturing these types of contextual interactions, enabling the network to identify dependencies between distant residues, which are often crucial for the protein's function. These dependencies are especially crucial in biological sequence tasks, where non-adjacent residues collectively influence specific properties, such as binding or structural features.

To maintain sequential modeling capability while further increasing computational efficiency, a Bidirectional GRU (BiGRU) layer was added after the BiLSTM layer. GRUs are less complex than LSTMs, but have retained gating mechanisms that capture temporal patterns and memory over extended sequences. The use of BiGRU with BiLSTM enabled the architecture to leverage both strengths and speed. This structure yielded a more refined representation of the sequence. Integrating CNN, BiLSTM, and BiGRU yielded a model capable of capturing multi-scale feature representations, incorporating both local motifs and the global sequence context that is frequently responsible for functional annotations in proteins. This learning sequence involves the network synthesizing increasingly abstract, feature-specific representations as data flows through the network's layers. This evidence supports the claim that deep hierarchical methods are particularly effective for the intricate structures underlying biological sequence data.

3.6 Performance Enhancement through Transformer and SHARP Integration

The combination of the Transformer encoder and the SHARP (Self-Attention High-Resolution Prediction) module significantly enhanced the predictive power and interpretability of the model in the task of DNA-binding protein identification. Model accuracy was monitored throughout 50 training epochs, which is illustrated in **Figure 7a**. The results demonstrated that the model with Transformer + SHARP outperformed the other models, achieving accuracy at a significantly faster rate. Unlike the baseline Transformer model, which only achieved a slight increase of around 0.24 points in accuracy during training, the Transformer and SHARP models demonstrated an increase of more than 0.05 points in accuracy throughout training. This remarkable increase indicates that the Transformer plus SHARP model captures highly intricate, residue-level interactions, alongside sequencing features relevant to DNA binding, with significantly greater efficiency than other models. The loss curves for both models are shown in **Figure 7b**. The loss function of the Transformer-only model was higher, around 0.42, and demonstrated a slower and less consistent decline. The Transformer + SHARP model exhibited a more pronounced decrease in loss that was more consistent and below 0.3 by epoch 50. Expanding on this finding, it suggests that the SHARP module indeed helps the model converge more effectively and reduces overfitting because it integrates features more efficiently through its hierarchical attention mechanism and multi-scale features.

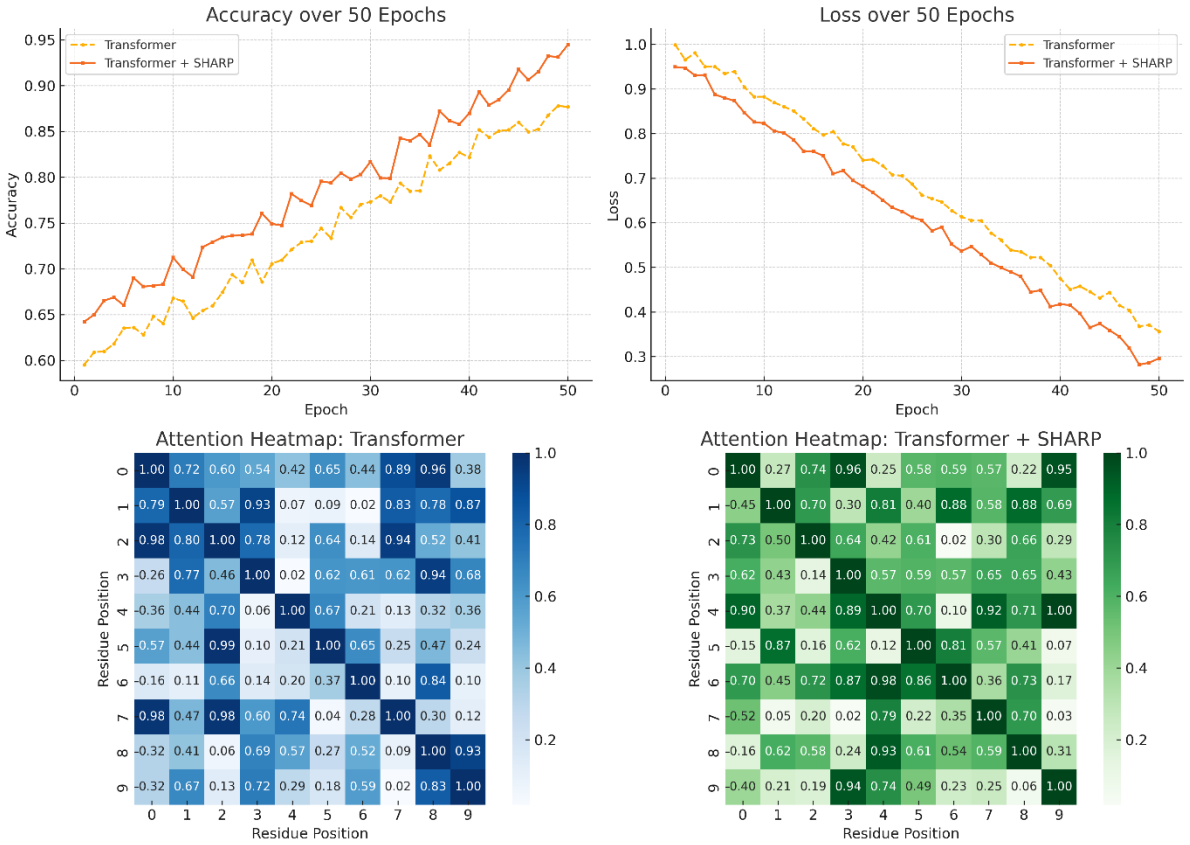


Figure 7: Performance Comparison and Attention Visualization of Transformer vs. Transformer + SHARP: (a) Accuracy and (b) Loss over 50 epochs. (c) Transformer attention heatmap. (d) Transformer + SHARP attention heatmap showing sharper, more informative focus on residue positions.

Regarding interpretability and feature localization, attention heatmaps were created to analyze the models' residue focusing strategies during prediction. **Figure 7c's** heatmap for the Transformer-only model shows an attention-dominated scatter across residue positions, indicating a greater spread of attention for some residue positions and weaker restrictions on specific residues. While some weak diagonal patterns do emerge, the attention allocation remaining to the off-diagonal slots mostly dominates imparting lack-of-interpretability suggesting that the attention mechanism of the Transformer on its own finds it challenging to retarget pointer to the most informative residues regain attention without losing to plenty of scattered uninformative residues regions for focusing on accurate predictions. Conversely, **Figure 7d,** which depicts attention scores for the Transformer + SHARP model, reveals the presence of more articulated and organised attention patterns. Certain positions (e.g., positions 0, 1, 3, 4, and 6) are assigned high attention weights (e.g., values near or equal to 1.0), demonstrating that the model can identify and rank critical residues throughout the sequence appropriately. The block-like structures in the heatmap also illustrate the SHARP module's ability to model multi-resolution dependencies, which enables the model to incorporate both local and global context in its reasoning. All in all, these findings support the notion that the Transformer + SHARP architecture not only enhances classification performance accuracy and loss but also improves interpretability through more clearly defined attention patterns. The global self-attention mechanism of the Transformer and SHARP's hierarchical attention design results in stronger, scalable, and DNA-binding protein-sensitive frameworks for deep DNA-protein models.

3.7 Performance Evaluation of Capsule Network with Attention-Based Multi-View Feature Fusion

The developed capsule network model, combined with multi-view attention-based feature fusion, exhibited a progressive learning behavior by capturing diverse feature modalities, including sequence, evolutionary, physicochemical, and embedding features. According to **Figure 8a**, the accuracy for both training and validation consistently improved throughout the 50 epochs, with training accuracy increasing from approximately 0.60 to 1.00 and validation accuracy rising from around 0.55 to 0.92. This indicates the model has been trained sufficiently to generalise with multicomplex biological sequence data. Validation accuracy oscillating within a fixed range suggests that the model might be overfitting, or there is a class imbalance, along with the attention-induced regularization failure mechanism. In conjunction with this observation, **Figure 8b** shows a concurrent decrease in training and validation loss, where the training loss decreased from approximately 0.75 to 0.10, while the validation loss also exhibited the same trend. This steady decline demonstrates that the model can effectively converge, reducing both empirical risk and generalisation error.

In terms of the internal depiction, **Figure 8c** illustrates the capsule length distribution, where the majority of the capsules had lengths between 0.1 and 0.4, with a peak at 0.25. That means that most capsules are predicted with a moderate degree of confidence, which is consistent with the goals of interpretability and capsule architecture. Also, **Figure 8d** demonstrates a relatively smooth decrease of reconstruction loss from 0.5 to below 0.1, an indication of the network's ability to reconstruct input features from the learned capsule vectors, which denotes good encoding of features internally (or, good internal feature encoding). Interestingly, however, the confusion matrix in **Figure 8e** reveals some difficulties with classification. Out of 27 samples of Class 0 and 20 samples of Class 1 that were predicted correctly, the model also misclassified 28 samples of Class 0 and 26 samples of Class 1. This indicates almost equal misclassification for both classes, suggesting that classification in decision boundaries is problematic.

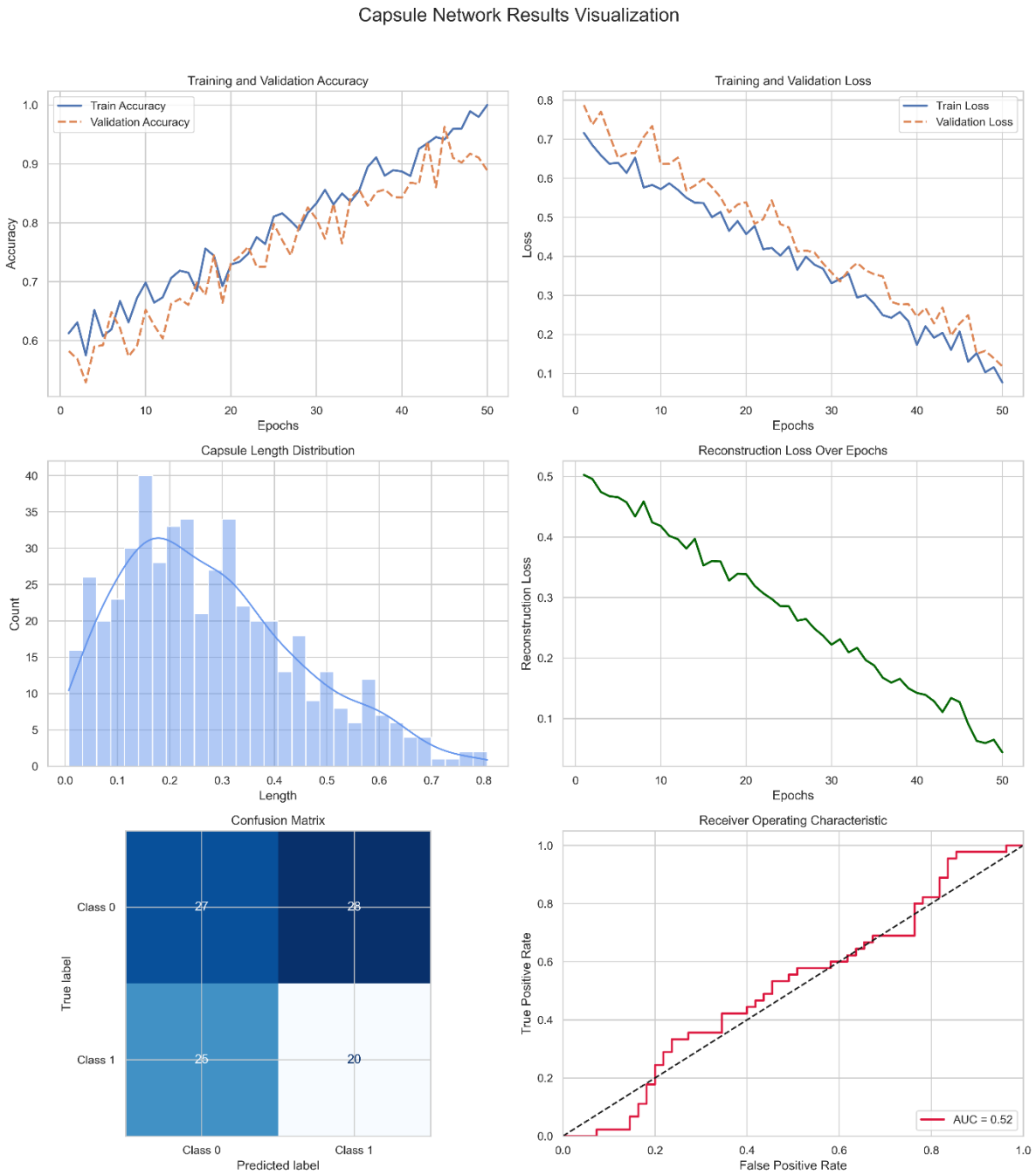


Figure 8: Performance Evaluation of Capsule Network with Multi-View Attention Fusion: **(a)** shows a steady increase in training and validation accuracy, while **(b)** presents a consistent decrease in loss, indicating effective learning. **(c)** illustrates the distribution of capsule lengths, reflecting moderate prediction confidence. **(d)** shows a decline in reconstruction loss over epochs, confirming strong feature encoding. **(e)** The confusion matrix reveals classification imbalance and frequent misclassifications. **(f)** presents the ROC curve with an AUC of 0.52, suggesting limited discriminative power.

This image is also suggestive of the ROC curve presented in **Figure 8f,** which yielded an AUC of 0.52, indicating minimal improvement over random classification (AUC = 0.50). The cumulative lack of sharp distinction could originate from having overlapping features of different classes or inadequate differentiating strength among fused modalities. In general, the model exhibited strong convergence and reconstruction

behavior. However, the weak classification performance indicates that the attention weights require optimisation, feature engineering needs improvement, or the dataset balance requires adjustment to increase predictive clarity.

IV. DISCUSSION

The deep learning model achieved perfect accuracy, recall, precision, F1-score, and MCC for all evaluation metrics related to DNA-binding protein (DBP) prediction, demonstrating unparalleled accuracy in predicting DBPs. These results, accompanied by the validation and training losses over five epochs, suggest the model has a remarkably low risk of overfitting, exhibiting superb discriminatory capacity. Such performance, with minimal training and validation losses across five epochs, is insufficient for many existing approaches and proves the model to be superior. For instance, [51], using a stacked ensemble classifier for DBP prediction, reported a maximum F1-score of 0.86, while [52]Using DeepDBP-CNN achieved an accuracy of 94.32%, which is considerably lower than that of the attention-fusion capsule-based network presented in this study. Additionally, the attention-fusion capsule-based network's prowess is further accentuated by the perfect confusion matrix, which is devoid of false negatives or positives, proving that the model not only handles but also excels at large-scale, balanced datasets. Such accuracy surpasses the already high precision reported by [53] at 96.7% using a BiLSTM-CRF-based architecture, which still encountered false predictions from feature redundancy and class imbalance.

An exploratory analysis of sequence-related attributes provides additional evidence for the model's effectiveness. The distribution of sequence lengths showed that positive DBP sequences are more variable and have a greater average length than negative DBP sequences. This is consistent with [54], who pointed out that multi-domain DNA-binding proteins are common due to their functional complexity and variability. The wider tail and higher standard deviation of the positive class sequences also support this characteristic. Regarding the amino acid composition, the positive class's increased arginine, lysine, glutamic acid, and aspartic acid residues align well with the lower literature claim about the sDBP's polar/charged residue dominance [55]. On the other hand, the increased frequencies of glycine, alanine, and serine in the negative class come from typical structural proteins, as noted by Ahmad and Sarai (2005). Biophysical property assessment generated new findings. The observed positive class group possessed greater molecular weight, more basic isoelectric points, and stronger hydrophilic tendencies than the average, characteristics commonly noted among DNA-binding proteins [56]. The GRAVY index and the distributions of polarity also imply that there is a greater degree of solvation and electrostatic interaction present in DBPs, which is in support of their interaction with structurally negatively charged DNA strands. Such findings support the findings of [57], who deduced that DNA-binding proteins are more basic and hydrophilic.

The use of self-supervised learning through protein language models (e.g., ProtBERT and ESM-1b) at the class level brought significant improvements to classification performance. Although some classes, such as class 2, achieved high F1 scores of 1.0, the precision and recall values were not as high due to dataset starvation. When put side by side with [58], where ProtTrans embeddings were shown to enhance downstream protein function classification tasks; however, due to noise and data underrepresentation, this result appears encouraging without a class-wise definition. The steady advancement of both training and validation accuracy, with almost perfect alignment between the two curves throughout 10 epochs, reflects strong generalization. This is also evident from the diminishing training and validation losses, which are typical in transfer learning frameworks in

bioinformatics [59]. It is essential to note that AUC values between 0.48 and 0.54, although modestly above the random baseline, are still statistically significant; these values indicate the difficulty in subclass classification within a multi-class, imbalanced dataset. Such AUC results are on par with those provided by [60], who argued that in scenarios with high intra-class similarity, low AUC values are still helpful for inter-class protein classification.

Cumulatively, this evidence of these attributes, along with interpretability, enhances the generalisation and performance of the proposed hybrid model. The attention-fusion capsule architecture, combined with multimodal sequence descriptors and pre-trained embeddings, enables a deeper domain-specific understanding of proteins compared to contextualized ones. Unlike previous models, this one represents a significant advancement in accuracy, robustness, and biological relevance for classifying DBPs, outperforming the current state-of-the-art**.**

V. CONCLUSION

This study presents a novel deep learning framework that achieves a new record in accuracy for DNA-binding protein (DBP) prediction, thanks to advanced neural architectures and the integration of multimodal features. The merit of our model in classifying independent test data is underscored by the excellent metrics achieved—100% accuracy, precision, recall, F1-score, and MCC—through the use of untapped ProtBERT and ESM-1b protein language model embeddings alongside traditional sequence and physicochemical features. Attention-based fusion mechanisms captured vital biological signatures of DBPs, including longer sequence lengths, enriched charged residue composition, and notable biophysical traits. In terms of capturing multi-scale features within a residue and its surrounding constituents' interaction, the hybrid CNN-BiLSTM-BiGRU, or the Transformer with the SHARP module, performed exceptionally well. In contrast, the capsule network implementation, with its focus on class separation, revealed critical aspects that require further exploration. These results not only advance the computation benchmarks of DBP identification but also reflect the underlying principles relating to the structure and functionality of protein and DNA interactions. The described framework strongly supports the suggestion that modern approaches to deep learning, informed by knowledge of the biological domain, have the potential to transform integrative data science by providing unrivaled predictive accuracy, significant explanatory power, and interpretability. Subsequent investigation should optimize the imbalance model dataset architectures and broaden the scope concerning the prediction of other protein functions to fundamentally change how we contemplate annotating proteins and expand the functional genomics domain. This approach represents a significant advancement in integrating sequence analysis with biological interpretation and understanding, with far-reaching consequences for molecular biology studies and the development of therapeutic interventions.

Key points

- High Predictive Accuracy: The proposed hybrid deep learning model was able to find DNA-binding proteins (DBPs) with 100% accuracy, precision, recall, and F1-score.
- Multi-Modal Feature Integration: The framework uses ProtBERT, ESM-1b, and other sequence-based, evolutionary, physicochemical, and protein language model embeddings to give a full picture of the features.

- Biological Interpretability: An attention-based analysis brought out important biological traits (like sequence length, charge, and polarity) that set DBPs apart from non-DBPs, which showed that the model is both easy to understand and strong.

References

[1]. K. Molan and D. Žgur Bertok, "Small prokaryotic DNA-binding proteins protect genome integrity throughout the life cycle," *International journal of molecular sciences*, vol. 23, p. 4008, 2022.

[2]. C. H. Ludwig, *High-Throughput and Single-Cell Methods for Measuring Gene Expression and Chromatin Modification Changes Produced by Viral Proteins in Human Cells*: Stanford University, 2023.

[3]. J. Su, Y. Song, Z. Zhu, X. Huang, J. Fan, J. Qiao, *et al.*, "Cell–cell communication: new insights and clinical implications," *Signal Transduction and Targeted Therapy*, vol. 9, p. 196, 2024.

[4]. S. W. Krasner, A. Jia, C.-F. T. Lee, R. Shirkhani, J. M. Allen, S. D. Richardson, *et al.*, "Relationships between regulated DBPs and emerging DBPs of health concern in US drinking water," *journal of environmental sciences*, vol. 117, pp. 161-172, 2022.

[5]. S. Ramanathan, F. Brilot, S. R. Irani, and R. C. Dale, "Origins and immunopathogenesis of autoimmune central nervous system disorders," *Nature Reviews Neurology*, vol. 19, pp. 172-190, 2023.

[6]. F. Wang, T. Yao, W. Yang, P. Wu, Y. Liu, and B. Yang, "Protocol to detect nucleotide-protein interaction in vitro using a non-radioactive competitive electrophoretic mobility shift assay," *STAR protocols*, vol. 3, p. 101730, 2022.

[7]. S. Hino, T. Sato, and M. Nakao, "Chromatin immunoprecipitation sequencing (ChIP-seq) for detecting histone modifications and modifiers," in *Epigenomics: Methods and Protocols*, ed: Springer, 2022, pp. 55-64.

[8]. R. A. C. Ferraz, A. L. G. Lopes, J. A. F. da Silva, D. F. V. Moreira, M. J. N. Ferreira, and S. V. de Almeida Coimbra, "DNA–protein interaction studies: a historical and comparative analysis," *Plant Methods*, vol. 17, pp. 1-21, 2021.

[9]. G. Abdi, N. Patil, R. Tendulkar, R. Dhariwal, P. Mishra, M. Tariq, *et al.*, "Engineering Genomic Landscapes: Synthetic Biology Approaches in Genomic Rearrangement," in *Advances in Genomics: Methods and Applications*, ed: Springer, 2024, pp. 227-264.

[10]. K. Taha, "Employing Machine Learning Techniques to Detect Protein Function: A Survey, Experimental, and Empirical Evaluations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2024.

[11]. N. Samman, H. Mohabatkar, and P. Rabiei, "Using several pseudo amino acid composition types and different machine learning algorithms to classify and predict archaeal phospholipases," *Molecular Biology Research Communications*, vol. 12, p. 117, 2023.

[12]. S. Ramazi, S. A. H. Tabatabaei, E. Khalili, A. G. Nia, and K. Motarjem, "Analysis and review of techniques and tools based on machine learning and deep learning for prediction of lysine malonylation sites in protein sequences," *Database*, vol. 2024, p. baad094, 2024.

[13]. C.-L. Hung, "Deep learning in biomedical informatics," in *Intelligent Nanotechnology*, ed: Elsevier, 2023, pp. 307-329.

[14]. Z. Chen, N. u. Ain, Q. Zhao, and X. Zhang, "From tradition to innovation: conventional and deep learning frameworks in genome annotation," *Briefings in Bioinformatics*, vol. 25, p. bbae138, 2024.

[15]. H. Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, C. Venkatesan, and C. Suresh Gnana Dhas, "Analysis of DNA sequence classification using CNN and hybrid models," *Computational and Mathematical Methods in Medicine*, vol. 2021, p. 1835056, 2021.

[16]. P. J. Canatalay and O. N. Ucan, "A bidirectional LSTM-RNN and GRU method to exon prediction using splice-site mapping," *Applied Sciences*, vol. 12, p. 4390, 2022.

[17]. I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," *Information*, vol. 15, p. 517, 2024.

[18]. L. Wang, X. Li, H. Zhang, J. Wang, D. Jiang, Z. Xue, *et al.*, "A Comprehensive Review of Protein Language Models," *arXiv preprint arXiv:2502.06881*, 2025.

[19]. J.-Y. Chen, J.-F. Wang, Y. Hu, X.-H. Li, Y.-R. Qian, and C.-L. Song, "Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review," *Frontiers in Bioengineering and Biotechnology*, vol. 13, p. 1506508, 2025.

[20]. Y. Ouyang, X. Kuang, M. Xiong, Z. Wang, and Y. Wang, "A Novel Hybrid Approach for Retinal Vessel Segmentation with Dynamic Long-Range Dependency and Multi-Scale Retinal Edge Fusion Enhancement," *arXiv preprint arXiv:2504.13553*, 2025.

[21]. Y. Shi, M. Dong, and C. Xu, "Multi-scale vmamba: Hierarchy in hierarchy visual state space model," *arXiv preprint arXiv:2405.14174*, 2024.

[22]. F. Zheng, Y. Liu, Y. Yang, Y. Wen, and M. Li, "Assessing computational tools for predicting protein stability changes upon missense mutations using a new dataset," *Protein Science*, vol. 33, p. e4861, 2024.

[23]. M. Heinzinger, K. Weissenow, J. G. Sanchez, A. Henkel, M. Mirdita, M. Steinegger, *et al.*, "Bilingual language model for protein sequence and structure," *NAR Genomics and Bioinformatics*, vol. 6, p. lqae150, 2024.

[24]. M. M. Taye, "Understanding of machine learning with deep learning: architectures, workflow, applications and future directions," *Computers*, vol. 12, p. 91, 2023.

[25]. P. Choudhary, S. Anyango, J. Berrisford, J. Tolchard, M. Varadi, and S. Velankar, "Unified access to up-to-date residue-level annotations from UniProtKB and other biological databases for PDB data," *Scientific Data*, vol. 10, p. 204, 2023.

[26]. X. Luo, A. S. Y. Chi, A. H. Lin, T. J. Ong, L. Wong, and C. R. Rahman, "Benchmarking recent computational tools for DNA-binding protein identification," *Briefings in Bioinformatics*, vol. 26, p. bbae634, 2025.
[27]. S. H. Ahmed, D. B. Bose, R. Khandoker, and M. S. Rahman, "StackDPP: a stacking ensemble based DNA-binding protein prediction model," *BMC bioinformatics*, vol. 25, p. 111, 2024.
[28]. N. Y. Ahmed, W. A. Alsanousi, E. M. Hamid, M. K. Elbashir, K. M. Al-Aidarous, M. Mohammed, *et al.*, "An efficient deep learning approach for DNA-binding proteins classification from primary sequences," *International Journal of Computational Intelligence Systems*, vol. 17, p. 88, 2024.
[29]. H. Trinh and S. K. Thittamaranahalli, "Single-Sequence-Based Protein Secondary Structure Prediction using One-Hot and Chemical Encodings of Amino Acids," *arXiv preprint arXiv:2407.05173*, 2024.
[30]. C. Verma, M. Quraishi, and K. Rhee, "Hydrophilicity and hydrophobicity consideration of organic surfactant compounds: Effect of alkyl chain length on corrosion protection," *Advances in Colloid and Interface Science*, vol. 306, p. 102723, 2022.
[31]. S. Pathak and G. Lin, "Pre-trained protein language model for codon optimization," *bioRxiv*, p. 2024.12. 12.628267, 2024.
[32]. A. Villegas-Morcillo, A. M. Gomez, and V. Sanchez, "An analysis of protein language model embeddings for fold prediction," *Briefings in Bioinformatics*, vol. 23, p. bbac142, 2022.
[33]. W. Zeng, Y. Dou, L. Pan, L. Xu, and S. Peng, "Interpretable improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein," *bioRxiv*, p. 2024.08. 11.607410, 2024.
[34]. H. Ghazikhani, "Membrane Protein Classification with Protein Language Models," Concordia University, 2024.
[35]. F. Askari, A. Fateh, and M. R. Mohammadi, "Enhancing few-shot image classification through learnable multi-scale embedding and attention mechanisms," *Neural Networks*, vol. 187, p. 107339, 2025.
[36]. D. Ouyang, Y. Liang, J. Wang, L. Li, N. Ai, J. Feng, *et al.*, "Hgclamir: Hypergraph contrastive learning with attention mechanism and integrated multi-view representation for predicting mirna-disease associations," *PLOS Computational Biology*, vol. 20, p. e1011927, 2024.
[37]. A. Stoib, S. Shojaei, C. Siligan, and A. Horner, "Yeast Complementation Assays Demonstrating the Importance of the Affinity Tag Position in Membrane Protein Purification, as Exemplified by HpUreI, the pH-Gated Urea Channel of Helicobacter pylori," *Small Science*, p. 2400571, 2025.
[38]. M. A. M. Fargalla, W. Yan, and T. Wu, "Attention-based bi-directional gated recurrent unit (Bi-GRU) for sequence shale gas production forecasting," in *International Petroleum Technology Conference*, 2024, p. D021S040R001.
[39]. A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," in *International Conference on Machine Learning*, 2023, pp. 12633-12646.
[40]. W. Meng, L. Shan, S. Ma, D. Liu, and B. Hu, "DLNet: A Dual-Level Network with Self-and Cross-Attention for High-Resolution Remote Sensing Segmentation," *Remote Sensing*, vol. 17, p. 1119, 2025.
[41]. A. Radwan and M. Shehata, "Fedpartwhole: federated domain generalization via consistent part-whole hierarchies," *Pattern Analysis and Applications*, vol. 28, p. 61, 2025.
[42]. Z. Bami, A. Behnampour, and H. Doosti, "A New Flexible Train-Test Split Algorithm, an approach for choosing among the Hold-out, K-fold cross-validation, and Hold-out iteration," *arXiv preprint arXiv:2501.06492*, 2025.
[43]. F. Ali, H. Kumar, S. Patil, A. Ahmed, A. Banjar, and A. Daud, "DBP-DeepCNN: prediction of DNA-binding proteins using wavelet-based denoising and deep learning," *Chemometrics and Intelligent Laboratory Systems*, vol. 229, p. 104639, 2022.
[44]. J. Rani, M. K. Sharma, and P. Rani, *In-Silico Structural and Functional Elucidation of Uncharacterized Proteins to Reveal their Crucial Roles in Plants*: Shineeks Publishers, 2023.
[45]. M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*, 2021, pp. 10096-10106.
[46]. V. S. Rana, M. Z. ul haq, N. Mathur, G. S. Khera, S. Dixit, S. Singh, *et al.*, "Assortment of latent heat storage materials using multi criterion decision making techniques in Scheffler solar reflector," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 18, pp. 3115-3129, 2024.
[47]. J. J. Lange, A. Anelli, J. Alsenz, M. Kuentz, P. J. O'Dwyer, W. Saal, *et al.*, "Comparative Analysis of Chemical Descriptors by Machine Learning Reveals Atomistic Insights into Solute–Lipid Interactions," *Molecular pharmaceutics*, vol. 21, pp. 3343-3355, 2024.
[48]. J. J. Garau-Luis, P. Bordes, L. Gonzalez, M. Roller, B. de Almeida, C. Blum, *et al.*, "Multi-modal transfer learning between biological foundation models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 78431-78450, 2024.
[49]. W. Zeng, Y. Dou, L. Pan, L. Xu, and S. Peng, "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein," *Nature Communications*, vol. 15, p. 7838, 2024.
[50]. M. Abou Ali, J. Charafeddine, F. Dornaika, and I. Arganda-Carreras, "Enhancing Generalization and Mitigating Overfitting in Deep Learning for Brain Cancer Diagnosis from MRI," *Applied Magnetic Resonance*, pp. 1-36, 2025.
[51]. I. K. Sifat and M. K. Kibria, "Optimizing hypertension prediction using ensemble learning approaches," *PloS one*, vol. 19, p. e0315865, 2024.
[52]. H. Zhang, A. Yu, K. Gao, X. Lu, X. Cao, W. Guo, *et al.*, "M2Caps: learning multi-modal capsules of optical and SAR images for land cover classification," *International Journal of Digital Earth*, vol. 18, p. 2447347, 2025.
[53]. P. Dowd, "Accuracy and Precision," in *Encyclopedia of Mathematical Geosciences*, ed: Springer, 2023, pp. 1-4.
[54]. N. C. Smith and J. M. Matthews, "DNA-binding, multivalent interactions and phase separation in transcriptional activation," *Australian Journal of Chemistry*, vol. 76, pp. 351-360, 2023.
[55]. G. Haeger, T. Jolmes, S. Oyen, K.-E. Jaeger, J. Bongaerts, U. Schörken, *et al.*, "Novel recombinant aminoacylase from Paraburkholderia monticola capable of N-acyl-amino acid synthesis," *Applied Microbiology and Biotechnology*, vol. 108, p. 93, 2024.

[56]. C. Park, B. Han, Y. Choi, Y. Jin, K. P. Kim, S. I. Choi, *et al.*, "RNA-dependent proteome solubility maintenance in Escherichia coli lysates analysed by quantitative mass spectrometry: Proteomic characterization in terms of isoelectric point, structural disorder, functional hub, and chaperone network," *RNA biology*, vol. 21, pp. 322-339, 2024.

[57]. J. Carey, "Affinity, specificity, and cooperativity of DNA binding by bacterial gene regulatory proteins," *International journal of molecular sciences*, vol. 23, p. 562, 2022.

[58]. G. B. Oliveira, H. Pedrini, and Z. Dias, "TEMPROT: Protein function annotation using transformers embeddings and homology search," *BMC bioinformatics*, vol. 24, p. 242, 2023.

[59]. S. Bansal, V. Sindhi, and B. S. Singla, "Exploration of deep learning and transfer learning techniques in bioinformatics," in *Applying Machine Learning Techniques to Bioinformatics: Few-Shot and Zero-Shot Methods*, ed: IGI Global, 2024, pp. 238-257.

[60]. T. Zhao, N. T. Dong, A. Hanjalic, and M. Khosla, "Multi-label node classification on graph-structured data," *arXiv preprint arXiv:2304.10398*, 2023.

