

DeepMVD: A Novel Multiview Dynamic Feature Fusion Model for Accurate Protein Function Prediction

Chaolin Song, Shiwen He, Yurong Qian,* Xinhui Li, Yue Hu, Jiaying Chen, Jingfu Wang, and Lei Deng



Cite This: *J. Chem. Inf. Model.* 2025, 65, 3077–3089



Read Online

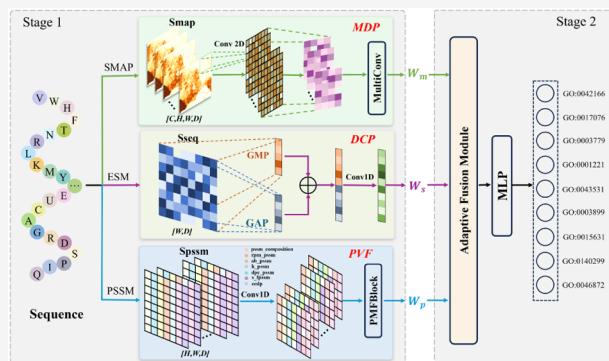
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Proteins, as the fundamental macromolecules of life, play critical roles in various biological processes. Recent advancements in intelligent protein function prediction methods leverage sequences, structures, and biomedical literature data. Among them, function prediction methods for protein sequences remain an enduring and popular research direction. Existing studies have failed to effectively utilize the multilevel attribute features reflected in protein sequences. This limitation hinders the enrichment of protein descriptions needed for high-precision prediction of protein functions. To address this, we propose DeepMVD, a novel deep learning model that enhances prediction accuracy by dynamically fusing multiview features. DeepMVD employs specialized modules to extract unique features from each view and utilizes an adaptive fusion mechanism for optimal integration. Evaluation of the CAFA4 data set shows that DeepMVD significantly outperforms existing state-of-the-art models in terms of BP, MF, and CC terminology, all obtaining the highest Fmax (0.523, 0.712, 0.740). Ablation studies confirm the model's robustness. Source code and data sets are available at <http://swanhub.co/scl/DeepMVD>.



INTRODUCTION

Protein, formed by specific amino acid sequences and complex three-dimensional folded structures, is crucial biomolecule for all living organisms.¹ Proteins play a central role in the complex activities of life, and the multiple functions determined by sequence and structure are essential for maintaining the homeostasis of living organisms, making proteins the centerpiece of life sciences and biomedicine.² Recently, the intricate link between protein function and cancer therapy mechanisms has increasingly emerged as a research hotspot. The development of protein functions can accurately pinpoint target proteins for cancer therapy and significantly aid the research of more effective anticancer drugs.³ Meanwhile, it will lay a crucial theoretical foundation for curing cancer and accelerates bioinformatics's progress.

Sequence, as the information that determines protein function, has reached an unprecedented scale and depth with the application of high-throughput technologies. This technology has expanded the UniProt database to over 356 million protein sequences.⁴ However, the high cost and time-consuming nature of experimental validation has left a large number of proteins without corresponding functional annotations.⁵ To address this situation, initial protein function prediction relied on sequence homology similarity, where known protein functional information is migrated to unknown proteins with highly similar sequences by calculating sequence similarity.⁶ With methods BLAST,⁷ Diamond,⁸ and Hhblits.⁹ However, some proteins with similar sequences exhibit

different functions due to the diversity of amino acid positions and structural constraints. Consequently, these methods are only suitable for certain proteins.

To overcome the limitations of earlier approaches, machine learning methods introduced the concept of transforming biological evidence into feature vectors for prediction. Models like PoGO¹⁰ and PANNZER2¹¹ utilized sequence similarity and structure to construct feature vectors to improve prediction accuracy. However, these feature extraction methods cannot fully capture the complexity and diversity of protein sequences due to difficulties in dealing with complex nonlinear data, which significantly limits the model's predictive ability.^{12,13} Fortunately, deep learning methods address these challenges by effectively processing intricate protein sequences. Deep_CNN_LSTM_Go¹⁴ and MSFPFP¹⁵ utilized neural network architectures to effectively enhancing prediction performance by mapping the nonlinear and complex sequential data into high-dimensional vector spaces.

Deep learning models widely use traditional one-shot encoding methods but cannot capture the deep semantic

Received: December 3, 2024

Revised: March 3, 2025

Accepted: March 3, 2025

Published: March 7, 2025



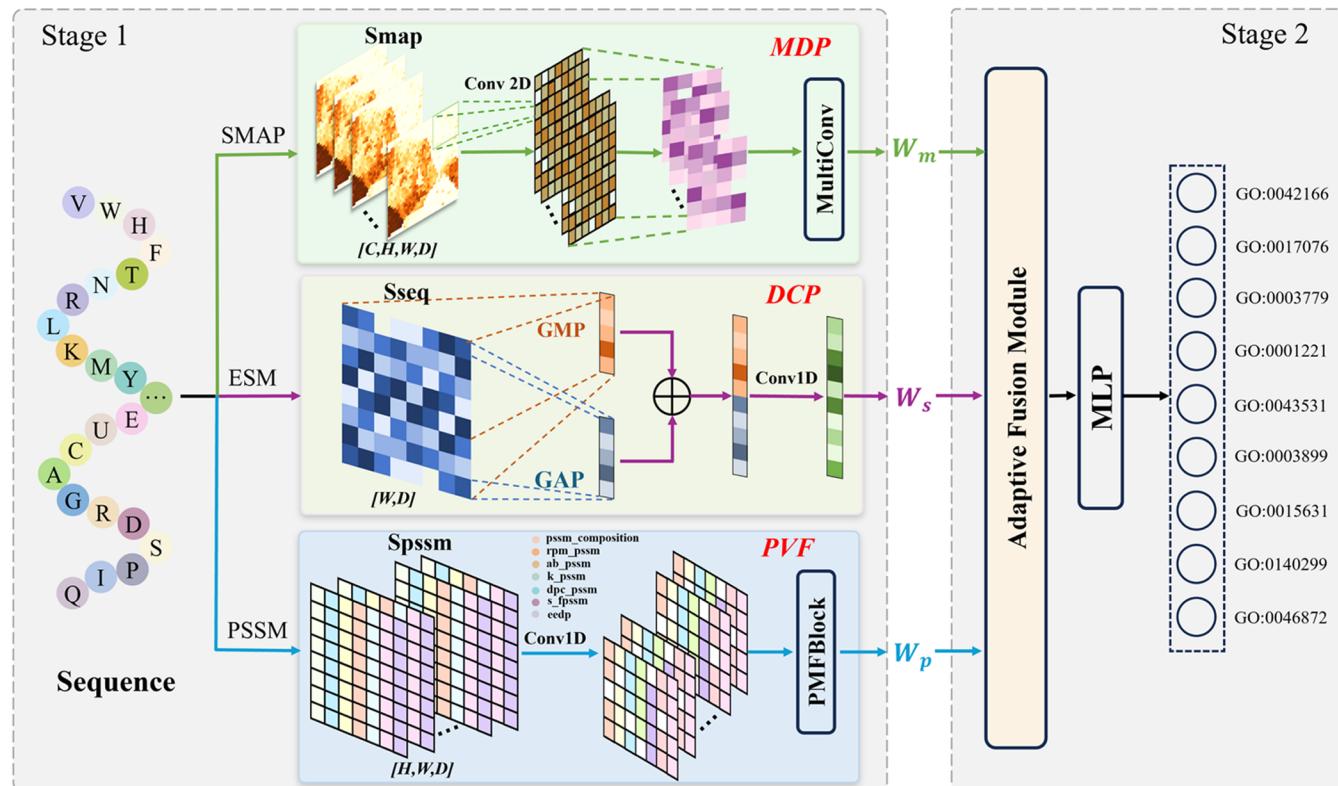


Figure 1. Overview of DeepMVD: This model begins by taking a sequence of proteins as input and constructing three distinct feature views: Smap, Sseq, and Spssm. Each view is processed through specialized modules—MDP, DCP, and PFV. The outputs are then integrated and classified within an adaptive fusion module, ensuring precise and dynamic analysis.

relationships between amino acid site.¹⁶ Recently, pretrained language models in natural language processing, such as ESM¹⁷ and ProtTrans,¹⁸ have achieved successful applications in some fields. Through in-depth comparative analysis of the experimental results, protein language models that generate rich contextual representations from large-scale data have significant advantages in improving the accuracy and depth of protein function prediction tasks, providing a new perspective on protein sequences.¹⁹ DeepAF²⁰ and ATGO²¹ use pretrained models to deeply parse protein sequences for deep semantic features and integrate different protein data to improve the accuracy of functional predictions.

It is not difficult to find that these automated methods are inevitably constrained by challenges in data acquisition, the reliance on single feature descriptors, and limitations in extracting deeper sequence features. Data such as structures or biomedical literature are often cumbersome and costly to obtain compared to sequence data, making sequence-based methods the mainstay of automated function prediction. Researchers are increasingly acknowledging that protein sequences are not merely simple arrangements of characters but complex entities rich with biological information and structural insights. Sequences are transformed into multifeatured views, including physicochemical properties, evolutionary conservation, and spatial conformational trends, and the models constructed on these feature views are rational and efficient.^{22,23} Selecting appropriate sequence feature views simultaneously is crucial for achieving complementarity and enhancing model performance.

This paper focuses on a novel multiview dynamic feature fusion model for accurate protein function prediction

(DeepMVD) to deeply capture sequence properties and address the challenges of a single-view protein description and inadequate feature extraction. The main contributions of this work are summarized as follows:

- A novel model named as DeepMVD was designed to comprehensively understand sequence characteristics by dynamically fusing multiview features.
- A multifeatured views model is constructed to characterize sequence properties including mapping of physicochemical features (Smap), deep semantic features (Sseq), and evolutionary features (Spssm), and designed unique feature extraction models.
- To achieve optimal aggregation of multiple features, an adaptive feature fusion mechanism is introduced to balance the contribution of each feature to ensure effective fusion of multiple views and improve prediction accuracy.

In conclusion, DeepMVD was constructed by pioneering the accurate extraction of features from complex sequence data in this paper. DeepMVD enriches our understanding of protein sequence-function relationships, dramatically improves the accuracy and reliability of predictions, and provides strong technical support for future proteomics development.

METHOD

DeepMVD consists of two stages for fusing multiview features of proteins, as illustrated in Figure 1. In the first stage, to enhance the feature expression capability, a novel multiview construction strategy is designed to convert a single protein sequence into three feature views that harbor different properties of the protein: Smap, Sseq, and Spssm. Meanwhile,

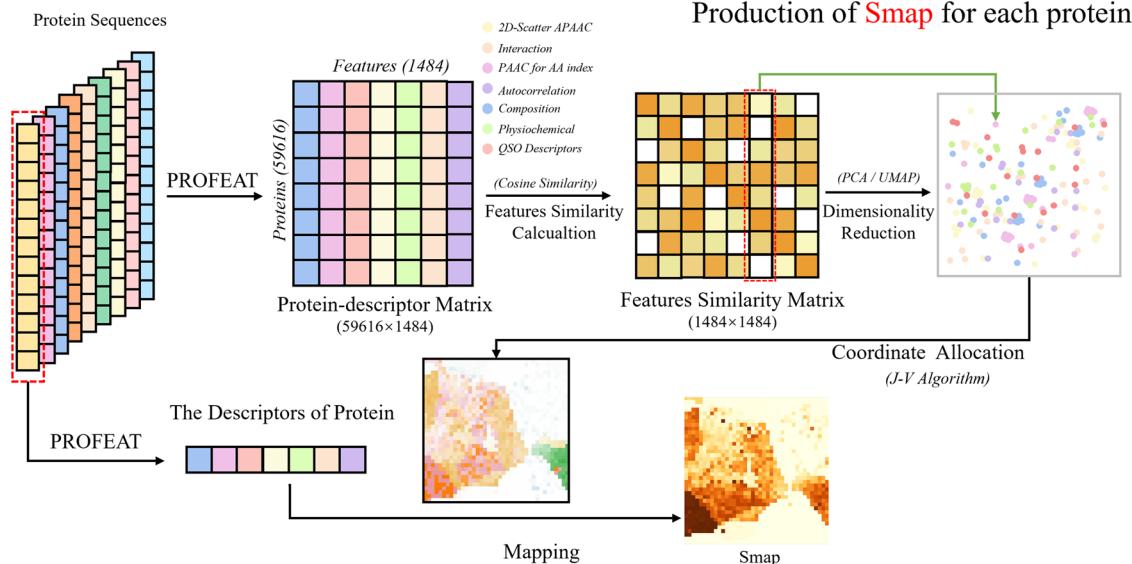


Figure 2. Process of generating Smap feature view for each protein.

unique processing module is designed respectively for each view - multidimensional perception extraction module (MDP), dual core pooling extraction module (DCP) and panoramic field-of-view extraction module (PVF). The second stage is feature fusion, which fuses three-view features to predict protein function efficiently. It connects the feature representations from the previous stage with an adaptive fusion module, which fuses them with different weights to achieve the optimal feature combination.

Multi-Dimensional Perception Extraction Module. To fully reflect the physicochemical characteristics of protein sequences, multidimensional perception extraction module (MDP) is introduced in this research. The MDP module transforms complex protein macromolecules into representative and information-intensive protein descriptors. This encoding approach not only quantifies the complex physicochemical properties, three-dimensional conformation, and biological functions of proteins, but also simplifies the inherent complexity of these molecules into intuitive and manageable parameters.²⁴

In MDP, protein descriptor technology generates similarity maps to extract essential information from protein sequences, such as physicochemical properties and structural descriptions. This module seeks to deeply mine the implicit features within sequences, addressing gaps in current feature characterization. As depicted in Figure 2, we employ the PROFEAT software²⁵ to compute the protein descriptors in our data set, which includes 1484 descriptors across seven categories: amphiphilic pseudo amino acid composition, amino acid composition, molecular interaction, amino acid autocorrelation, quasi-sequence-order, physicochemical properties, and pseudo amino acid composition. This encoding approach enabled the construction of a comprehensive protein descriptor matrix. To ensure data consistency and comparability, the descriptor matrices are normalized using the following formula

$$\chi_{ij}^{\text{norm}} = \frac{\chi_{ij} - \min r_i}{\max r_i - \min r_i} \quad (1)$$

In the formula, χ_{ij} the original descriptor data, χ_{ij}^{norm} is the normalized data, and r_i denotes the i -th feature. Additionally,

$\max r_i$ and $\min r_i$ represent the maximum and minimum values of the i -th feature across all proteins, respectively.

This module utilizes the protein descriptor matrix to generate a feature similarity matrix by calculating pairwise distances between descriptor features. The original 1484-dimensional feature space is compressed to two dimensions for visualization and analysis using PCA²⁶ and UMAP²⁷ techniques to reduce the feature dimensionality. The J-V algorithm²⁸ is subsequently employed to assign spatial locations to each feature, precisely mapping them onto a blank coordinate system and applying color rendering to produce a localized base map. The distinct features of each protein are integrated with this localization base map to create a mapping of physicochemical features (Smap) for each protein. The precise methodology for distance calculation is as follows

$$\text{distance}(r_a, r_b) = 1 - \frac{r_a \cdot r_b}{\|r_a\| \times \|r_b\|} \quad (2)$$

Taking one feature r_a as an example, it is originally represented as a 1484-dimensional vector (R_a)

$$R_a = [d_{r_a, 1}, d_{r_a, 2}, \dots, d_{r_a, b}, \dots, d_{r_a, 1484}] \quad (3)$$

where $d_{r_a, b}$ represents the pairwise distance between features r_a and r_b . Next, the feature vector R_a was transformed into a more interpretable and visually presentable 2D vector U_a by computing the interrelationships on the manifold surface using UMAP

$$U_a = [x_{r_a}, y_{r_a}] \quad (4)$$

x_{r_a} and y_{r_a} represent the coordinate values of the feature r_a on a two-dimensional plane. To allocate these 2D vectors of protein features (R_a) to a localized base map, a map (M) was defined to store the allocation results of the protein features. For instance, the feature r_a would be represented as a grid (M_a) in this localized base map

$$M_a = [m_{r_a}, m_{r_a}] \quad (5)$$

m_{r_a} and n_{r_a} are integers ranging from 0 to 38, representing the coordinates of the feature r_a . Finally, the grid locations of these features were determined by minimizing the total cost between U_a and M_a using the $J-V$ algorithm—an optimal task allocation method that achieves overall minimal cost while ensuring each value is assigned only once

$$\left[\min_{M_a} \sum_{i=1}^N d(U_a[i], M_a[i]) \right] \quad (6)$$

In the MDP module, to fully utilize the intrinsic spatial relationships and information in the similarity mapping graph, the superior ability of Convolutional Neural Networks (CNNs)²⁹ in processing image location information is considered. Therefore, the MDP module employs a multilayer CNN architecture to thoroughly extract the location and linkage features in the mapping graph. Subsequently, the maximum pooling technique is utilized to aggregate its most significant features. The calculations are detailed below

$$\text{OUT}(i, D_{\text{out}}) = \text{bias}(D_{\text{out}}) + \sum_{k=0}^{D_{\text{in}}} \text{weight}(D_{\text{out}}, k) \times X_{\text{map}}(i, k) \quad (7)$$

where D_{out} represents the output dimension of the convolution, D_{in} is the embedding dimension of X_{map} , which is the RGB feature matrix derived from the Smap, and \times denotes the convolution cross-correlation operation.

To emphasize the physicochemical properties and spatial localization features embedded within protein sequences, the MDP module has been designed to substantially improve the accuracy of protein function prediction. By rendering the similarity mapping in a fine-grained manner and incorporating a multilayer CNN architecture with MaxPooling to effectively captures deep feature representations of sequences at multiple scales. These innovative protein sequence coding strategies uniquely reveal the intrinsic correlations between descriptor features, and have the potential to be extended to existing sequence analysis models to enhance overall prediction accuracy and model performance.

Dual Core Pooling Extraction Module. In protein sequence analysis, each protein is encoded by a unique set of amino acids constituting its characteristic sequence. Nonetheless, amino acid sequences are not inherently amenable to direct computational processing. Pretrained language models based on treating amino acid sequences as “sentences” significantly improve the learning and prediction of protein sequence features.^{30–33} ESM,¹⁷ a pretrained language model based on the Transformer architecture, provides a deep understanding of the semantic information between amino acids. It is trained on evolutionary diversity data covering 250 million sequences containing 86 billion amino acids.

In dual core pooling extraction module (DCP), the ESM¹⁷ algorithm is employed to extract high-level information from protein sequences, aiming for accurate prediction of protein functions. Specifically, for the input protein sequence $\{a_1, a_2, a_3, \dots, a_L\}$ (where L denotes the sequence length). ESM¹⁷ employs a masking strategy to process the amino acids. It calculates the similarity between amino acid pairs using the self-attention module, subsequently generating the feature embedding $\{x_1, x_2, x_3, \dots, x_L\} \in \mathbb{R}^{L \times 1280}$. The standard embedding dimension for ESM¹⁷ is set at 1280. Subsequently,

each element of the sequence is scored with an average weight, and the sequence context information is computed using the normalized scores. This is expressed as $S_{\text{seq}} = \text{mean}(\{x_1, x_2, x_3, \dots, x_L\}) \in \mathbb{R}^{1280}$. The procedure for extracting features from protein sequences is as follows

$$\{x_1, x_2, x_3, \dots, x_L\} = \text{ESM}(\{a_1, a_2, a_3, \dots, a_L\}) \in \mathbb{R}^{L \times 1280} \quad (8)$$

$$S_{\text{seq}} = \text{mean}(\{x_1, x_2, x_3, \dots, x_L\}) \in \mathbb{R}^{1280} \quad (9)$$

As illustrated in Figure 1, within the DCP module, the sequence feature S_{seq} is processed and directed separately to the global average pooling and maximum pooling layers. And pooling operations are performed on the embedded dimensions to compress the features and extract the basic information from the compressed dimensions. These are subsequently concatenated and iteratively processed to construct the comprehensive feature matrix of the sequence. Furthermore, a dynamic adjustment strategy is employed to capture local positional features more efficiently, where the sensory field is proportionally tailored to the dimension. The relevant calculation process can be expressed as follows

$$P_{\text{avg}} = \text{avg pool}(S_{\text{seq}}) \quad (10)$$

$$P_{\text{max}} = \text{max pool}(S_{\text{seq}}) \quad (11)$$

Where the shape of P after pooling is represented as (batch_size, embed_dim)

$$k = \left\lceil \frac{\log_2(D) + 1}{2} \right\rceil \quad (12)$$

$$\text{OUT} = \text{bias}(D_t) + ((P_{\text{avg}} + P_{\text{max}}) \times k) \quad (13)$$

where k represents the convolution kernel, D denotes the embedding dimension, D_t is the output dimension of the one-dimensional dynamic convolution, and \times signifies the one-dimensional convolutional cross-correlation operation.

In order to enhance the stability of the feature representation, the strategy substantially improves the accuracy of feature extraction by effectively capturing the global structure and local fine-grained features of the sequence while organically combining global and local features. This comprehensive feature extraction tool not only enhances the understanding of the complexity of protein sequences but also promises to provide valuable insights for other feature extraction tasks.

Panoramic Field-of-View Extraction Model. Protein sequences are amino acid residues that vary under specific environmental conditions and possess the potential for mutation. They form the foundation of protein evolution, and therefore, evolutionary information is indispensable for predicting protein function.⁵

Panoramic field-of-view extraction module (PVF) analyzed target proteins against the SwissProt database using the PSI-BLAST to assign each amino acid residue a vector set of 20 scores. These scores measure the conservation of amino acids across evolutionary periods, offering insights into proteins' functional and structural analysis. As a result, a Position-Specific Scoring Matrix (PSSM) with dimensions of $20 \times L$ was constructed for any protein sequence of length L , facilitating further analyses using this matrix.

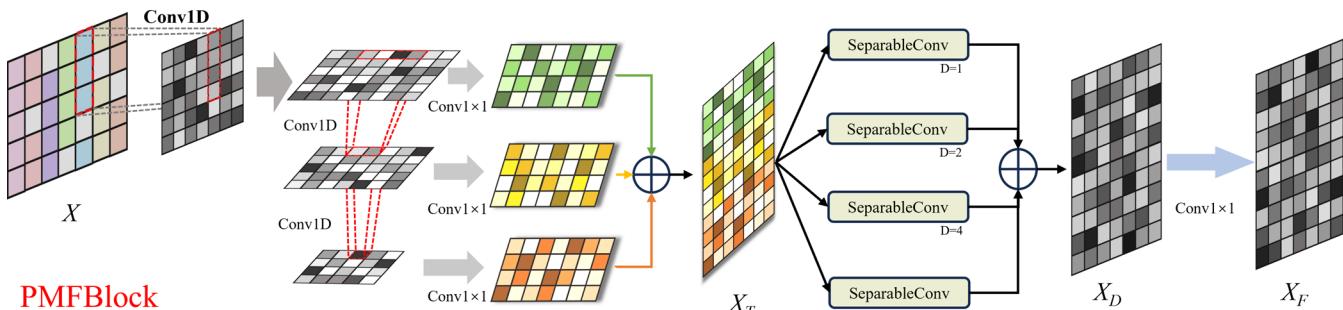


Figure 3. PMFBlock module overview.

To ensure PSSM data consistency and enhance subsequent analysis, the PVF module applies seven advanced transformation algorithms to generate refined 400-dimensional feature vectors. These algorithms capture various aspects of PSSM information: global sequence patterns (AB-PSSM³⁴), functional loci conservation (PSSM-position³⁵), reduced redundancy (RPM-PSSM³⁴), structure–function relationships (S-FPSSM³⁶), local correlations (DPC-PSSM³⁷), long-distance dependencies (k -separated-bigrams-PSSM³⁸), and evolutionary importance (EEDP³⁹). The resulting structured Spssm feature matrix preserves evolutionary information, optimizes data structure, and facilitates efficient protein function prediction and structural analysis.

After an in-depth exploration of PSSM data transformation characteristics, this module further focuses on optimizing data representation to enhance the efficiency of protein function prediction and structural analysis. Considering the variations in details embedded within the same semantic information at different scales,⁴⁰ this PVF module meticulously designs and implements a module called PMFBlock (Figure 3). The PMFBlock module aims to enhance evolutionary information closely related to function prediction through fine-grained feature processing. Specifically, the module takes as input feature data processed through one-dimensional convolution and normalization. First, 1×1 convolution is applied to upsample the feature information at different scales, ensuring that the feature matrices $\{M_1, M_2, M_3\}$ are consistent in size. Subsequently, these matrices are concatenated into a comprehensive feature matrix $X_T \in \mathbb{R}^{3d_t \times L}$. To broaden the receptive field and capture a wider range of contextual information, the module processes X_T in parallel using four separable convolutional layers with varying dilation rates. Each convolutional layer performs fine-grained feature extraction with its unique dilation rate, enriching feature diversity. The features extracted from convolutional layers with different dilation rates are sequentially concatenated to form a feature matrix $X_D \in \mathbb{R}^{4d_t \times L}$, characterized by a rich hierarchical structure. A final 1×1 convolution block is applied to transform X_D into the final multiscale deep feature $X_F \in \mathbb{R}^{d_c \times L}$. The complete calculation is presented below

$$M_k = C_k(X_{in}) \quad (14)$$

where k represents different convolution scales, C_k denotes the k -th multiscale convolution operation, and X_{in} is the input feature

$$X_T = \text{concatenate}(C_{1 \times 1}(M_1) | C_{1 \times 1}(M_2) | C_{1 \times 1}(M_3)) \quad (15)$$

$C_{1 \times 1}$ represents a 1×1 convolution used for upsampling to ensure all feature maps are of uniform size before concatenation

$$X_{D_r} = DConv_r(X_T) \quad (16)$$

where $DConv_r$ denotes a separable convolution with dilation rate r and X_{D_r} represents the corresponding output feature

$$X_F = C'_{1 \times 1}(\text{concatenate}(X_{D_{r1}} | X_{D_{r2}} | X_{D_{r3}} | X_{D_{r4}})) \quad (17)$$

The core of this module is to thoroughly explore evolutionary information in protein sequences and optimize data structure using advanced feature engineering techniques, thereby enhancing the efficacy of protein function prediction and structural analysis. The PMFBlock module was designed to refine position-specific scoring matrix (PSSM) data, thereby effectively reinforcing evolutionary information closely related to functional prediction. The module employs separable convolution to expand the receptive field of features and capture more comprehensive contextual information, thereby constructing deep multiscale features with complex hierarchical structures. This strategy not only preserves the original evolutionary information in PSSM but also reveals intrinsic correlations between features, significantly enhancing the prediction accuracy and performance of existing sequence analysis models.

Multiview Adaptive Fusion Module. Upon in-depth exploration of the characterization from the three perspectives, it is found that each perspective encapsulates unique information. Specifically, Smap reveals physicochemical properties and spatial localization, Sseq captures the deeper semantic attributes of the sequences, and Spssm reflects the evolutionary characteristics of the sequences. An adaptive multifeature view fusion network is constructed based on feature extraction to integrate these three viewpoint features, thereby complementing the sequential features and forming the final decision output. The network automatically assigns weight coefficients to different views and optimizes the weight combination through iterative training, thereby balancing feature redundancy and novelty to achieve optimal feature fusion. The following formula represents the integrated decision-making mechanism

$$F = \sum_{V=1}^M W_V Y_V \text{ where } W^T 1 = 1, W \geq 0 \quad (18)$$

Here M denotes the total number of feature views, $W_V \in \mathbb{R}$ is the weight parameter of the V -th view, Y_V represents the initial prediction of the V -th view, and F is the final prediction after fusion.

The DeepMVD approach mitigates issues related to model overcomplexity while enhancing generalization, resulting in more accurate predictions. Moreover, DeepMVD demonstrates higher computational efficiency, making it a viable choice for protein function prediction tasks. In summary, DeepMVD addresses key challenges in multifeature fusion by employing an adaptive fusion module as the core structure of the feature fusion process. This design provides a robust foundation for the final classification task, significantly enhancing the computational efficiency and accuracy of the traditional protein function prediction method.

EXPERIMENTS AND RESULTS

All comparison experiment data sets are divided into training, validation, and test sets in a 7:2:1 ratio, and the models are trained on the respective BP, MF, and CC labels. F_{\max} , AUC, Recall, and Precision are used as evaluation metrics to assess the model's performance on the experimental data set. The evaluation metrics are defined as follows

$$F_{\max} = \max_t \frac{2 \cdot \text{avg } \Pr(t) \cdot \text{avg } \text{Rc}(t)}{\text{avg } \Pr(t) + \text{avg } \text{Rc}(t)} \quad (19)$$

$$\text{avg } \Pr(t) = \frac{1}{k(t)} \cdot \sum_{i=1}^{k(t)} \text{pr}_i(t) \quad (20)$$

$$\text{avg } \text{Rc}(t) = \frac{1}{n} \cdot \sum_{i=1}^n \text{rc}_i(t) \quad (21)$$

$$\text{pr}_i(t) = \frac{\sum_j T(G_j, p_i) \cdot \mathbf{1}(S(p_i, G_j) \geq t)}{\sum_j \mathbf{1}(S(p_i, G_j) \geq t)} \quad (22)$$

$$\text{rc}_i(t) = \frac{\sum_j T(G_j, p_i) \cdot \mathbf{1}(S(p_i, G_j) \geq t)}{\sum_j T(G_j, p_i)} \quad (23)$$

Here t represents the prediction threshold with a step size of 0.1, J indicates whether the protein prediction is correct (1 if true, 0 otherwise), and n is the total number of proteins.

Precision, Recall, and F_{\max} measure the accuracy of the model's predictions, while AUC evaluates the model's capability to identify relevant information. AUC is calculated based on the confusion matrix. The formula is as follows

$$\text{AUC} = \int_{-\infty}^{\infty} \text{TPR}(t)(-\text{FPR}(t))dt \quad (24)$$

$$\text{TPR}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)} \quad (25)$$

$$\text{FPR}(t) = \frac{\text{FP}(t)}{\text{FP}(t) + \text{TN}(t)} \quad (26)$$

Data Set. Three data sets were used in this experiment to evaluate the method: the CAFA4 competition data set, HUMAN, and the YEAST data set. Proteins with sequence lengths in the range [50, 1500] are considered first. For sequences longer than 1500, only the first 1500 amino acids are retained. The CAFA4 competition data set, which includes 17 MOUSE, SCHPO, and Drosophila species. Meanwhile, the HUMAN and YEAST data sets were selected as part of the HUMAN and YEAST species data. The above data sets retained entries that lacked labeling for some GO terms,

resulting in unequal amounts of data for BP, MF, and CC. The specific data are presented in Table 1.

Table 1. Experimental Data Details

| data set | ontology | train | valid | test | terms |
|----------|----------|-------|-------|------|-------|
| CAFA4 | BP | 33392 | 9477 | 4768 | 4507 |
| | MF | 23379 | 6528 | 3316 | 726 |
| | CC | 34591 | 9880 | 4945 | 628 |
| HUMAN | BP | 8321 | 2385 | 1194 | 491 |
| | MF | 8218 | 2348 | 1174 | 321 |
| | CC | 8783 | 2510 | 1254 | 240 |
| YEAST | BP | 3691 | 1131 | 566 | 373 |
| | MF | 3836 | 1166 | 603 | 171 |
| | CC | 3955 | 1203 | 609 | 151 |

Experimental Setup. We implemented the model in Pytorch and Pytorch Geometric library and trained our model with binary cross-entropy as a loss function and AdamW optimizer with a learning rate of 1×10^{-3} . We set the dropout rate to 0.2. We trained six models on the CAFA4, HUMAN and YEAST data sets for Molecular Function Ontology (MFO), Biological Process Ontology (BPO) and Cellular Component Ontology (CCO), respectively. During the training period, the model with the highest F_{\max} value on the validation set is retained as the final model. The CAFA4 data set's epoch and batch sizes of training MFO, BPO and CCO models are 70, 100, 70 and 16, 16, 16. On the HUMAN and YEAST data sets, the epoch and batch sizes of training MFO, BPO and CCO models are 50, 70, 50 and 12, 16, 12.

Comparison Experiment. ATGO²¹ employs a Transformer-based pretrained model, extracting feature embeddings from the last three Transformer layers and fusing them using a multiview approach, combined with confidence scores from a ternary network to predict protein function. ATGO effectively extracts rich feature information from protein sequences for accurate GO term prediction.

MSF-PFP¹⁵ utilizes sequences, structural domains, and protein interaction networks as inputs, extracting features through specific modules and sufficiently fusing them to achieve complementary feature effects, thereby accomplishing the prediction task.

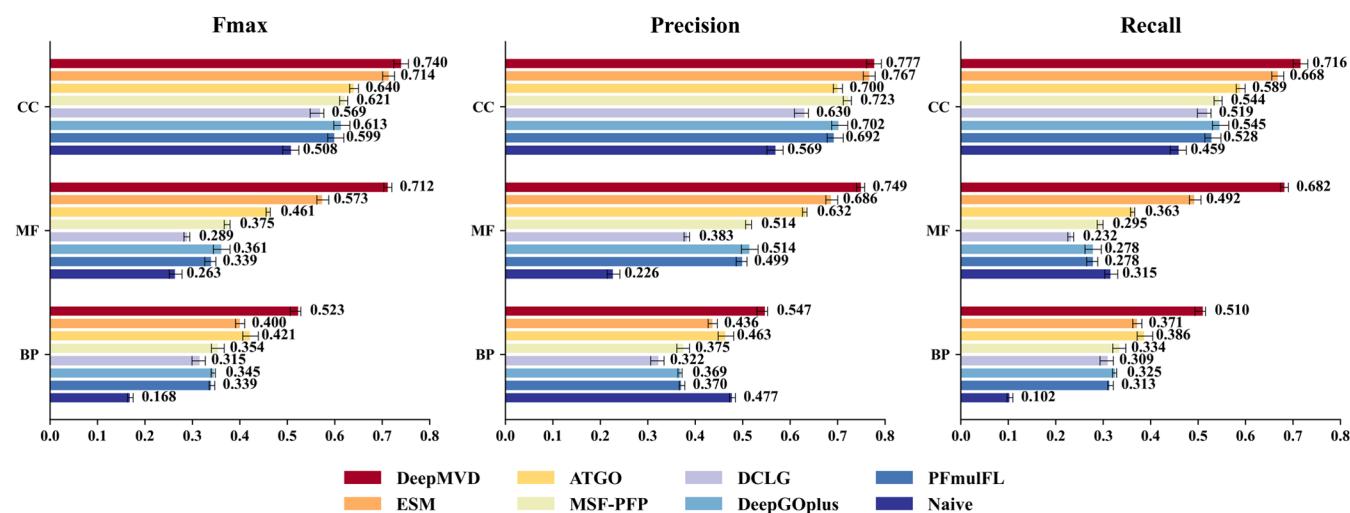
Performance Evaluation. To comprehensively evaluate the performance of the DeepMVD model, we systematically compared it with seven mainstream protein function prediction models on the CAFA4 data set. The results of the CAFA4 data set are shown in Table 2 and Figure 4. The accuracy of the BP, MF, and CC terms of the DeepMVD model were 0.547, 0.749, and 0.777, respectively, and the F_{\max} was 0.523, 0.712 and 0.74, respectively. MF usually focuses on the specific molecular functions of proteins, which are more directly and tightly linked to sequences. This model aggregates the three-view features together, making capturing the features related to MF easier. In contrast, BP usually focuses on the hierarchical and extensive nature of biological processes, which often involve multiple intertwined and complex processes, which may lead to a lower prediction accuracy relative to MF and CC. The above results indicate that DeepMVD has excellent generalization ability, validity and reliability in protein function prediction.

To explore the interspecies variability, we evaluated it on two single-species collections, human and YEAST. As can be seen from Table 3, on the single-species data set, DeepMVD shows excellent performance compared to the four models,

Table 2. Comparison of DeepMVD with Seven Models on the CAFA4 Data Set^a

| CAFA | BP | | | | MF | | | | CC | | | |
|------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | F_{\max} | recall | precision | AUC | F_{\max} | recall | precision | AUC | F_{\max} | recall | precision | AUC |
| naive | 0.168 | 0.102 | 0.477 | 0.551 | 0.263 | 0.315 | 0.226 | 0.652 | 0.508 | 0.459 | 0.569 | 0.726 |
| PFmulFL | 0.339 | 0.313 | 0.370 | 0.901 | 0.339 | 0.278 | 0.499 | 0.910 | 0.599 | 0.528 | 0.692 | 0.944 |
| DeepGOplus | 0.345 | 0.325 | 0.369 | 0.895 | 0.361 | 0.278 | 0.514 | 0.920 | 0.613 | 0.545 | 0.702 | 0.951 |
| ESM | 0.400 | 0.371 | 0.436 | 0.900 | <u>0.573</u> | <u>0.492</u> | <u>0.686</u> | 0.931 | <u>0.714</u> | <u>0.668</u> | <u>0.767</u> | 0.956 |
| DCLG | 0.315 | 0.309 | 0.322 | 0.881 | 0.289 | 0.232 | 0.383 | 0.871 | 0.569 | 0.519 | 0.630 | 0.916 |
| ATGO | <u>0.421</u> | <u>0.386</u> | <u>0.463</u> | <u>0.936</u> | 0.461 | 0.363 | 0.632 | <u>0.962</u> | 0.640 | 0.589 | 0.700 | <u>0.970</u> |
| MSF-PFP | 0.354 | 0.334 | 0.375 | 0.912 | 0.375 | 0.295 | 0.514 | 0.930 | 0.621 | 0.544 | 0.723 | 0.957 |
| DeepMVD | 0.523 | 0.510 | 0.547 | 0.952 | 0.712 | 0.682 | 0.749 | 0.969 | 0.740 | 0.716 | 0.777 | 0.978 |

^aBold indicates optimal values, and underlining indicates suboptimal values.

**Figure 4.** Comparison of DeepMVD with Seven Correlation Models on the CAFA4 Data set.**Table 3.** Comparison of DeepMVD with Seven Models on the HUMAN and YEAST Data Set^a

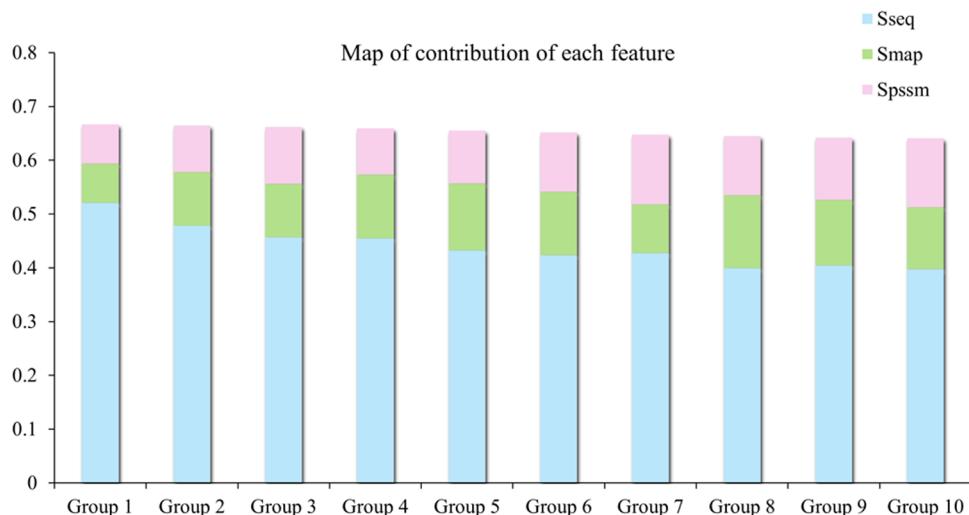
| method | F_{\max} | | | AUC | | | recall | | |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | BP | MF | CC | BP | MF | CC | BP | MF | CC |
| HUMAN | | | | | | | | | |
| naive | 0.166 | 0.300 | 0.503 | 0.548 | 0.645 | 0.701 | 0.097 | 0.299 | 0.408 |
| DeepGOplus | 0.394 | 0.463 | <u>0.765</u> | 0.905 | 0.936 | 0.973 | 0.360 | 0.363 | 0.686 |
| MSF-PFP | 0.399 | 0.472 | 0.684 | 0.930 | 0.960 | 0.975 | 0.359 | 0.388 | 0.664 |
| ESM | <u>0.551</u> | <u>0.715</u> | 0.730 | <u>0.962</u> | <u>0.981</u> | <u>0.983</u> | <u>0.540</u> | <u>0.691</u> | <u>0.722</u> |
| DeepMVD | 0.588 | 0.801 | 0.769 | 0.971 | 0.986 | 0.987 | 0.563 | 0.796 | 0.772 |
| YEAST | | | | | | | | | |
| naive | 0.211 | 0.291 | 0.578 | 0.565 | 0.673 | 0.738 | 0.131 | 0.355 | 0.481 |
| DeepGOplus | 0.399 | 0.435 | <u>0.799</u> | 0.912 | 0.928 | 0.967 | 0.376 | 0.337 | 0.722 |
| MSF-PFP | 0.393 | 0.418 | 0.722 | 0.936 | 0.954 | 0.978 | 0.354 | 0.346 | 0.671 |
| ESM | <u>0.608</u> | <u>0.777</u> | 0.788 | <u>0.981</u> | <u>0.989</u> | <u>0.990</u> | <u>0.586</u> | <u>0.745</u> | <u>0.773</u> |
| DeepMVD | 0.667 | 0.867 | 0.828 | 0.987 | 0.992 | 0.993 | 0.652 | 0.849 | 0.823 |

^aBold indicates optimal values, and underlining indicates suboptimal values.

naïve, DeepGOplus, MSF-PFP, and ESM, by optimally aggregating the three view features and thus achieving complementarity among the features. There are some differences in the performance of different species in the results, which may be because yeast is higher than humans in terms of the ratio of the number of annotations to the total number of labels in the samples, and yeast features are more straightforward, which in turn leads to the superior performance of the model on the yeast data set.

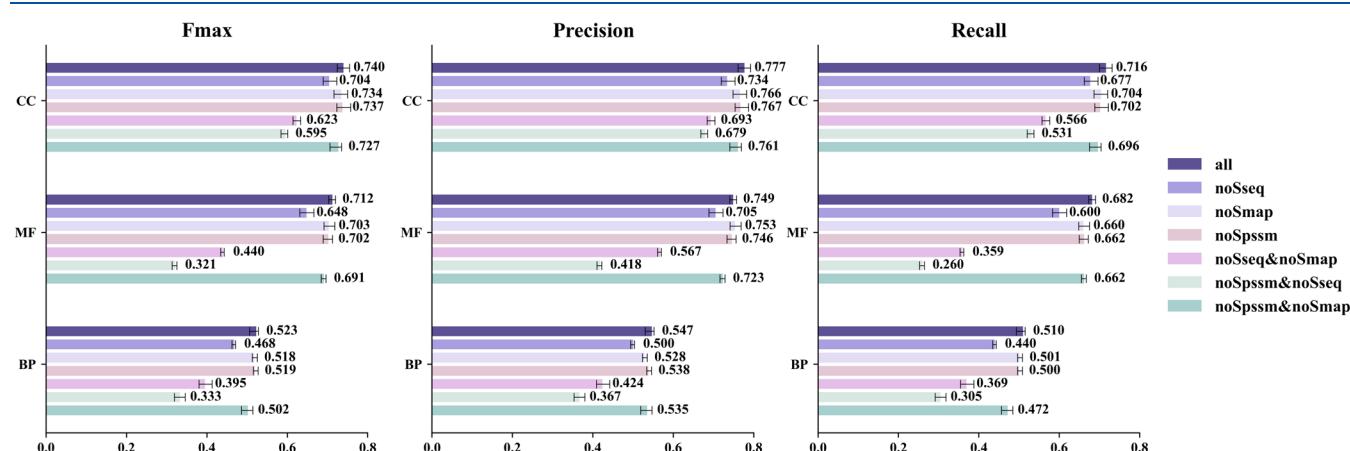
At the same time, to more intuitively reflect the contribution of each input feature to the prediction results, some visual

analysis of the degree of contribution of each input feature to the prediction results was done. The model adopts an adaptive fusion mechanism, automatically assigning weights to each feature to achieve optimal results. We selected the parameters of the top ten optimal combinations and plotted the contribution of each feature to make a comparison of the contribution of each feature (as shown in Figure 5). The performance of the combination on each data set also shows that the ESM for predicting protein functions using pretrained language models essentially achieves suboptimal performance. Ultimately, it can be concluded that deep semantic features

**Figure 5.** Map of contribution of each feature.**Table 4. Ablation Experiment with Multiview Features^a**

| S _{seq} | S _{map} | S _{pssm} | BP | | | MF | | | CC | | |
|------------------|------------------|-------------------|------------------|--------------|--------------|--------------|------------------|--------------|--------------|--------------|------------------|
| | | | F _{max} | recall | precision | AUC | F _{max} | recall | precision | AUC | F _{max} |
| ✓ | | | 0.502 | 0.472 | 0.535 | 0.941 | 0.691 | <u>0.662</u> | 0.723 | 0.969 | 0.727 |
| | ✓ | | 0.333 | 0.305 | 0.367 | 0.897 | 0.321 | 0.260 | 0.418 | 0.893 | 0.595 |
| | | ✓ | 0.395 | 0.369 | 0.424 | 0.923 | 0.440 | 0.359 | 0.567 | 0.951 | 0.623 |
| ✓ | ✓ | | <u>0.519</u> | 0.500 | <u>0.538</u> | 0.952 | 0.702 | <u>0.662</u> | 0.746 | 0.966 | <u>0.737</u> |
| ✓ | | ✓ | 0.518 | <u>0.501</u> | 0.528 | <u>0.951</u> | 0.703 | 0.660 | 0.753 | <u>0.968</u> | 0.734 |
| | ✓ | ✓ | 0.468 | 0.440 | 0.500 | 0.933 | 0.648 | 0.600 | 0.705 | 0.952 | 0.704 |
| ✓ | ✓ | ✓ | 0.523 | 0.510 | 0.547 | 0.952 | 0.712 | 0.682 | <u>0.749</u> | 0.969 | 0.740 |

^aBold indicates optimal values, and underlining indicates suboptimal values.

**Figure 6.** Ablation Experiment with Multiview Features.

play a crucial role in the performance of protein function prediction.

In conclusion, the superior performance of the DeepMVD method is attributed to its exceptional multiview feature fusion capability. The model transforms the protein sequence into three feature views to fully highlight the embedded information and employs a fusion strategy to achieve complementary roles among the features. Compared with single-view encoding approaches, DeepMVD's multiview encoding comprehensively represents each sequence's deep semantic, evolutionary, materialization, and positional information, employing corresponding modules to extract rich features from each feature

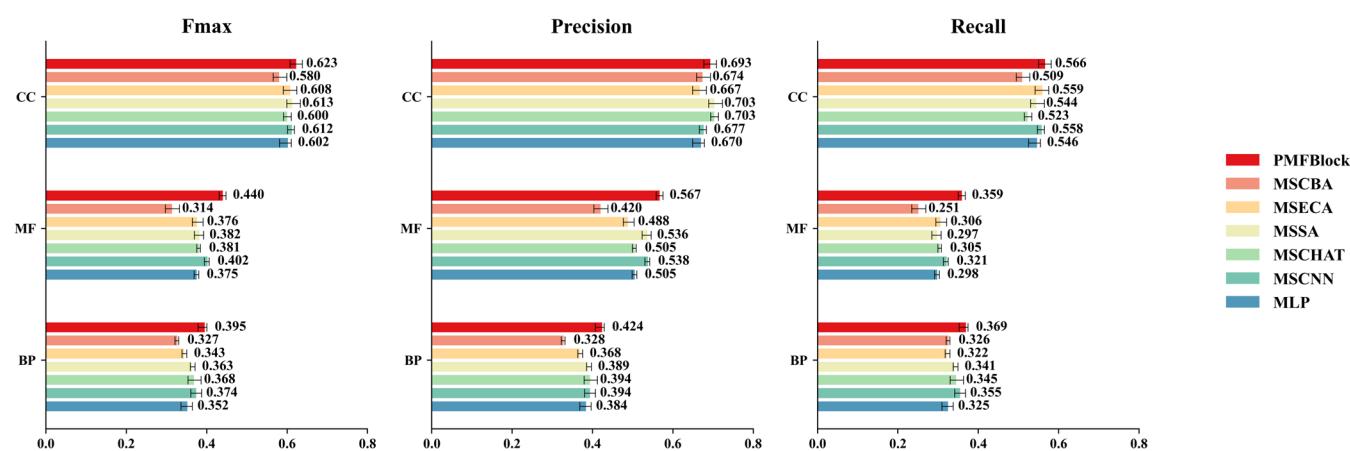
view. Introducing the PMFBlock module efficiently mines abundant evolutionary features, surpassing traditional CNN-based feature extraction methods and providing a more accurate foundation for protein function prediction.

Ablation Experiment. In this research, we propose DeepMVD, a deep learning model employing a multiview, multifeature fusion strategy to transform a one-dimensional protein sequence into a multidimensional feature representation. A series of meticulously designed ablation experiments were conducted to thoroughly explore the complementarities between feature views, verify the indispensability of the

Table 5. Ablation Experiment of PMFBlock Module^a

| method | BP | | | | MF | | | | CC | | | |
|----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | F_{\max} | recall | precision | AUC | F_{\max} | recall | precision | AUC | F_{\max} | recall | precision | AUC |
| MLP | 0.352 | 0.325 | 0.384 | 0.890 | 0.375 | 0.298 | 0.505 | 0.921 | 0.602 | 0.546 | 0.670 | 0.938 |
| MSCNN | <u>0.374</u> | <u>0.355</u> | <u>0.394</u> | 0.910 | <u>0.402</u> | <u>0.321</u> | <u>0.538</u> | <u>0.945</u> | 0.612 | 0.558 | 0.677 | <u>0.954</u> |
| MSCHAT | 0.368 | 0.345 | <u>0.394</u> | <u>0.918</u> | 0.381 | 0.305 | 0.505 | 0.942 | 0.600 | 0.523 | 0.703 | 0.948 |
| MSSA | 0.363 | 0.341 | 0.389 | 0.917 | 0.382 | 0.297 | 0.536 | 0.944 | <u>0.613</u> | 0.544 | 0.703 | 0.959 |
| MSECA | 0.343 | 0.322 | 0.368 | 0.902 | 0.376 | 0.306 | 0.488 | 0.934 | 0.608 | <u>0.559</u> | 0.667 | 0.953 |
| MSCBA | 0.327 | 0.326 | 0.328 | 0.880 | 0.314 | 0.251 | 0.420 | 0.874 | 0.580 | 0.509 | 0.674 | 0.925 |
| PMFBlock | 0.395 | 0.369 | 0.424 | 0.923 | 0.440 | 0.359 | 0.567 | 0.951 | 0.623 | 0.566 | <u>0.693</u> | 0.959 |

^aBold indicates optimal values, and underlining indicates suboptimal values.

**Figure 7. Ablation Experiment of PMFBlock Module.****Table 6. Ablation Experiments with Encoding Approaches^{a,b}**

| method | BP | | | | MF | | | | CC | | | |
|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | F_{\max} | recall | precision | AUC | F_{\max} | recall | precision | AUC | F_{\max} | recall | precision | AUC |
| A | <u>0.377</u> | <u>0.358</u> | 0.398 | 0.915 | <u>0.414</u> | <u>0.329</u> | 0.557 | 0.942 | <u>0.615</u> | <u>0.559</u> | 0.684 | 0.953 |
| A_D | 0.368 | 0.334 | <u>0.409</u> | 0.913 | 0.395 | 0.307 | 0.554 | 0.942 | 0.609 | 0.537 | 0.703 | 0.951 |
| A_D_E | 0.371 | 0.348 | 0.398 | 0.914 | 0.393 | 0.302 | 0.565 | 0.941 | 0.612 | 0.555 | 0.683 | 0.954 |
| A_D_E_K | 0.375 | 0.352 | 0.401 | 0.915 | 0.406 | 0.316 | 0.565 | 0.945 | 0.614 | 0.546 | 0.701 | <u>0.955</u> |
| A_D_E_K_P | 0.374 | 0.347 | 0.405 | 0.914 | 0.411 | 0.318 | 0.580 | <u>0.947</u> | 0.612 | 0.549 | 0.693 | 0.954 |
| A_D_E_K_P_R | 0.375 | 0.353 | 0.399 | <u>0.916</u> | 0.410 | 0.319 | <u>0.574</u> | 0.946 | 0.614 | 0.544 | 0.705 | 0.954 |
| all type | 0.395 | 0.369 | 0.424 | 0.923 | 0.440 | 0.359 | 0.567 | 0.951 | 0.623 | 0.566 | 0.693 | 0.959 |

^aBold indicates optimal values, and underlining indicates suboptimal values. ^bA-(AB-PSSM),D-(DPC-PSSM),E-(EEDP),K-(k-separated-bigrams-PSSM),P-(PSSM-position),R-(RPM-PSSM).

PMFBlock module in the PFV framework, and assess the practical utility of the adaptive fusion module.

A comprehensive multiview encoding approach is adopted to convert protein sequences into information-rich multidimensional feature views, capturing different levels of information through diverse technical means. Adaptively fusing features extracted by different modules to achieve feature complementarity and integration significantly enhances the accuracy of protein function prediction. Through well-designed experiments, the performance of each feature extraction component was individually verified and analyzed in combination to comprehensively assess the advantages of the DeepMVD model. The experimental results are presented in Table 4 and Figure 6.

As shown in Figure 6, the DeepMVD model aggregates each feature view and demonstrates strong performance across all evaluation metrics for BP, MF, and CC. The model was most effective in classifying the BP and CC ontologies, but the

accuracy of the MF ontology decreased slightly, suggesting that the link between the physical and chemical properties and the molecular compositional function may be less strong.

The traditional CNN architecture struggles to effectively capture the intrinsic correlation between protein function and evolutionary information in the feature extraction module. DeepMVD utilizes the number of PSSM encoding channels as input and incorporates the PMFBlock module to fully explore evolutionary features embedded in each encoding method, deeply analyzing the complex correlation between function and evolutionary information. In the second part of the ablation experiments, DeepMVD compares Spssm across different feature extractors, demonstrating the superior capabilities of the PMFBlock module for this task.

Table 5 and Figure 7 presents the results of this experiment. Compare PMFBlock with modules such as Multiscale Efficient Attention (MSECA), Multiscale Channel Attention (MSCHAT), and Multiscale Spatial Attention (MSSA). The

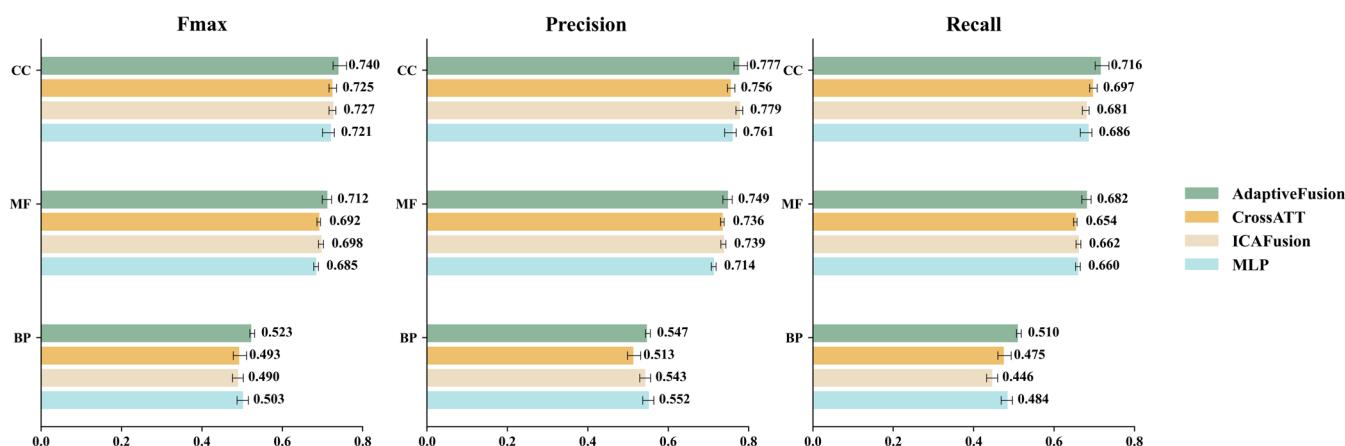


Figure 8. Ablation Experiments with Different Fusion Strategies.

Performance Comparison of DeepMVD under Different Dimensionality Reduction Methods

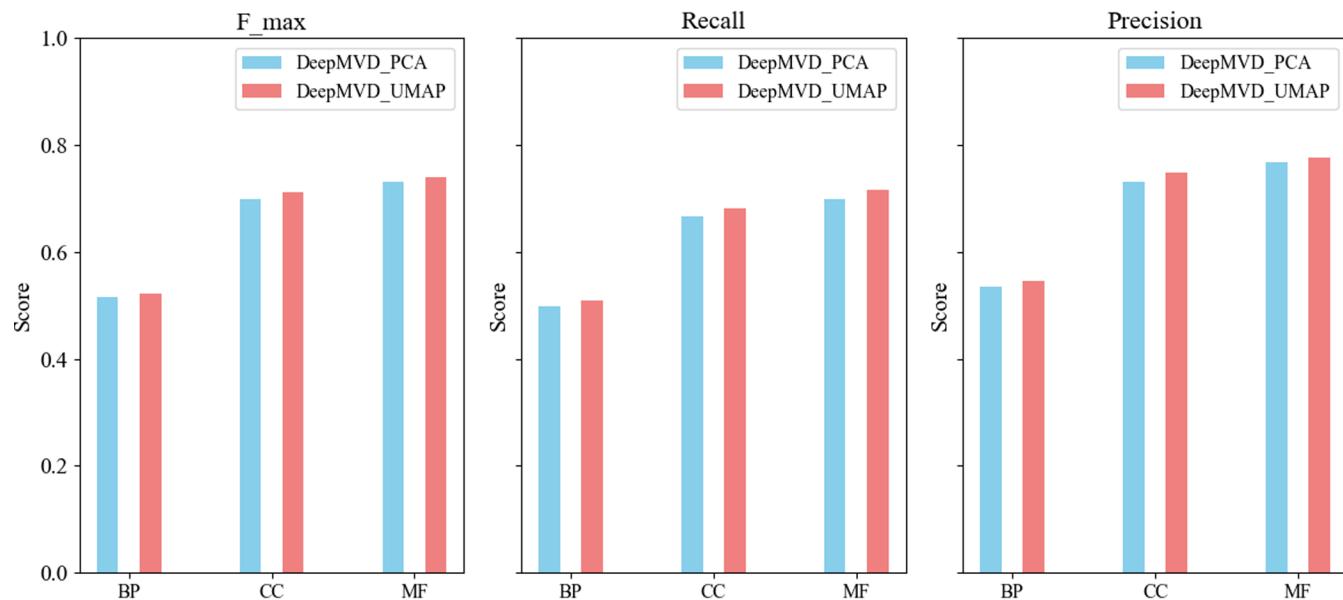


Figure 9. Performance Comparison of DeepMVD under Different Dimensionality Reduction Methods.

results indicate that PMFBLOCK excels in evolutionary feature extraction, proving that it can effectively capture fine-grained features of evolutionary information.

To verify the significance of each encoding type during PSSM matrix encoding, the seven encodings—AB-PSSM, PSSM-position, RPM-PSSM, S-FPSSM, DPC-PSSM, *k*-separated-bigrams-PSSM, and EEDP—were analyzed through sequential combinatorial ablation. The results demonstrated that each encoding reveals relevant PSSM information from different dimensions, enhancing protein sequences' feature representation. However, all methods directly or indirectly rely on the conservation and evolutionary information in PSSM. Therefore, when using all types of encoding for functional prediction, there may be some repetitive features, which in turn leads to overfitting and suboptimal performance (e.g., both DPC-PSSM and *k*-separated-bigrams-PSSM capture local amino acid relationships). The experimental results are presented in Table 6.

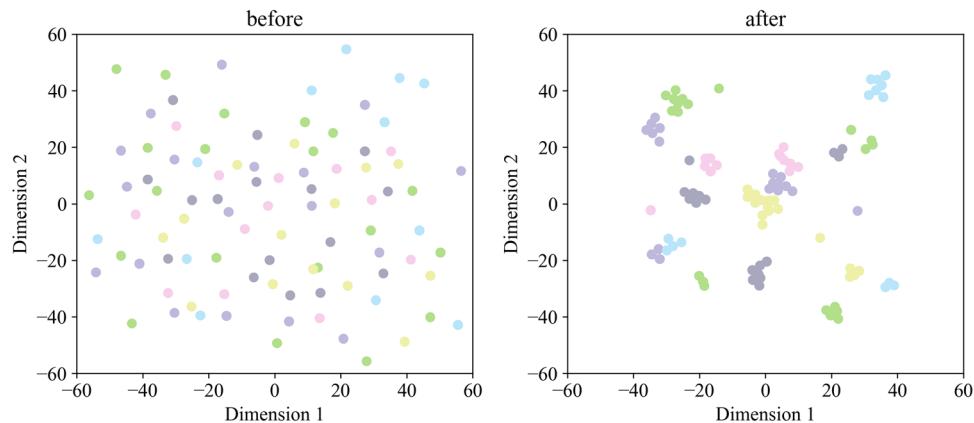
The core objective of DeepMVD is to integrate multiview feature information from protein sequences to accurately

predict protein functions, making effective aggregation of each feature an essential aspect of this study. This experiment compares the adaptive fusion module with traditional feature fusion methods, such as concatenation and cross-fusion. The experimental results are presented in Figure 8. Adaptive fusion strategies effectively compensate for the limitations of individual features, while multifeature fusion fully leverages existing protein sequence data to achieve accurate protein function prediction.

■ CORRELATION ANALYSIS

Comparison among the performances of DeepMVD using different dimensionality reduction methods (PCA and UMAP). The performances were represented using Fmax, Recall and Precision values and the performances of DeepMVDPCA, and DeepMVDUMAP were highlighted in light blue and red, respectively. As shown in the Figure 9, these two models perform roughly the same in the three GO categories (BP, CC, and MF). In particular, the predictive performance of DeepMVDUMAP is slightly higher than that of Deep-

The t-SNE visualization of features before and after fusion

**Figure 10.** T-SNE visualization of features before and after fusion.

MVDPCA (F_{\max} of 0.5 – 0.7%; Recall of 1.1 – 1.3%; Precision of 1 – 1.45%).

To fully demonstrate the relationship between the performance of DeepMVD and the features, this experiment was performed by randomly selecting six functionally annotated labels and the corresponding part of the protein data and performing t-SNE visualization before and after feature fusion (As shown in the Figure 10). As can be seen from the figure, the original features lack sufficient representation ability to distinguish different functional annotation tags well, while after fusion, the local and global information on the features are integrated, and the samples with different functional annotation tags form a more precise grouping.

DISCUSSION AND CONCLUSIONS

Numerous studies have highlighted the importance of resolving protein function to understand life processes and biological mechanisms. With advancements in high-throughput sequencing, automated protein function prediction has become a key challenge in the postgenomic era. Sequence-based methods offer improved efficiency and reduced costs for protein function annotation; however, existing approaches often fail to fully utilize protein sequence data for in-depth physiological analysis.

To address these challenges, we propose a novel approach named DeepMVD, which integrates natural language processing and computer vision techniques to capture multiview features of protein sequences. DeepMVD comprises three channels—MDP, DCP, and PFV—to construct feature views of different sequence characteristics, employing specialized modules to extract relevant information. After adaptive fusion of multiview features, DeepMVD effectively predicts protein function.

Experimental results demonstrate that DeepMVD outperforms state-of-the-art models, with ablation studies validating the contribution of each feature view, the necessity of each module, and the strength of the fusion strategy. In conclusion, DeepMVD represents a cutting-edge solution for protein function prediction using deep learning.

In future studies, we plan to incorporate new protein-related data (e.g., protein structure, biomedical literature) to further optimize our model and enhance the accuracy of protein function prediction. Additionally, we plan to adapt more

advanced large-scale language models to the bioinformatics field to further advance science and technology in this domain.

ASSOCIATED CONTENT

Data Availability Statement

The code and data sets of DeepMVD are available at [10.5281/zenodo.1492486](https://zenodo.1492486/).

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.4c02216>.

Details of the seven classes of protein descriptors generated using PROFEAT; contribution analysis data for the three profiles Sseq, Smap and Spssm ([PDF](#))

AUTHOR INFORMATION

Corresponding Author

Yurong Qian — School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China; Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Xinjiang University, Urumqi, Xinjiang 830046, China; Key Laboratory of Software Engineering and Xinjiang Engineering Research Center of Big Data and Intelligent Software, School of Software, Xinjiang University, Urumqi 830091, China; Email: qyr@xju.edu.cn

Authors

Chaolin Song — School of Software, Xinjiang University, Urumqi 830091, China; Xinjiang Engineering Research Center of Big Data and Intelligent Software, School of Software and Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China; orcid.org/0009-0008-1680-515X

Shiwen He — School of Software, Xinjiang University, Urumqi 830091, China; School of Computer Science and Engineering, Central South University, Changsha 410083, China

Xinhui Li — School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China; Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Xinjiang University, Urumqi, Xinjiang 830046, China

Yue Hu — School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China; Joint International Research Laboratory of Silk Road Multilingual

Cognitive Computing, Xinjiang University, Urumqi, Xinjiang 830046, China

Jiaying Chen – School of Software, Xinjiang University, Urumqi 830091, China; Xinjiang Engineering Research Center of Big Data and Intelligent Software, School of Software and Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China

Jingfu Wang – School of Software, Xinjiang University, Urumqi 830091, China; Xinjiang Engineering Research Center of Big Data and Intelligent Software, School of Software and Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China; orcid.org/0009-0005-3834-4997

Lei Deng – School of Software, Xinjiang University, Urumqi 830091, China; School of Computer Science and Engineering, Central South University, Changsha 410083, China

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jcim.4c02216>

Author Contributions

C.S.: Data curation, methodology, validation, writing—original draft. S.H.: Formal analysis, funding acquisition, resources, writing—review and editing. Y.Q.: Conceptualization, funding acquisition, resources, writing—review and editing. X.L.: Data curation, formal analysis, investigation, writing—review and editing. Y.H.: Investigation, methodology, writing—review and editing. J.C.: Conceptualization, data curation, resources, writing—review and editing. J.W.: Data curation, writing—review and editing. L.D.: Methodology, writing—review and editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 62171474, No. 62266043); the Tianshan Innovation Team Program of Xinjiang Uygur Autonomous Region of China (No. 2023D14012); the Excellent Youth Foundation of Xinjiang Uygur Autonomous Region of China (No. 2023D01E01); the Outstanding Young Talent Foundation of Xinjiang Uygur Autonomous Region of China (No. 2023TSYCCX0043); the Finance Science and Technology Project of Xinjiang Uygur Autonomous Region (No. 2023B01029-1, No. 2023B01029-2, No. 2022B01006, No. 2023B02034-2); Basic Research Foundation of Universities in the Xinjiang Uygur Autonomous Region of China (XJEDU2023P012).

REFERENCES

- (1) Chi, Y.; Li, C.; Feng, X. Advances in machine learning for predicting protein functions. *Chin. J. Biotechnol.* **2023**, *39*, 2141–2157.
- (2) Xinhui, L.; Yurong, Q.; Haitao, Y.; Yue, H.; Jiaying, C.; Hongyong, L.; Mengnan, M. Survey of Bioinformatics-Based Protein Function Prediction. *Comput. Eng. Appl.* **2023**, *59*, 50–62.
- (3) Jang, Y. J.; Qin, Q.-Q.; Huang, S.-Y.; Peter, A. T. J.; Ding, X.-M.; Kornmann, B. Accurate prediction of protein function using statistics-informed graph networks. *Nat. Commun.* **2024**, *15*, No. 6601.
- (4) Consortium, U. UniProt: A hub for protein information. *Nucleic Acids Res.* **2015**, *43*, D204–D212.
- (5) Wang, Z.; Deng, Z.; Zhang, W.; Lou, Q.; Choi, K.-S.; Wei, Z.; Wang, L.; Wu, J. MMSMAPlus: A multi-view multi-scale multi-attention embedding model for protein function prediction. *Briefings Bioinf.* **2023**, *24*, No. bbad201.
- (6) Dhanuka, R.; Singh, J. P.; Tripathi, A. A comprehensive survey of deep learning techniques in protein function prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2023**, *20*, 2291–2301.
- (7) Ye, J.; McGinnis, S.; Madden, T. L. BLAST: Improvements for better sequence analysis. *Nucleic Acids Res.* **2006**, *34*, W6–W9.
- (8) Buchfink, B.; Reuter, K.; Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **2021**, *18*, 366–368.
- (9) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2012**, *9*, 173–175.
- (10) Jung, J.; Yi, G.; Sukno, S. A.; Thon, M. R. PoGO: Prediction of gene ontology terms for fungal proteins. *BMC Bioinf.* **2010**, *11*, 1–9.
- (11) Törönen, P.; Medlar, A.; Holm, L. PANNZER2: A rapid functional annotation web server. *Nucleic Acids Res.* **2018**, *46*, W84–W88.
- (12) Bernardes, J.; Pedreira, C. A review of protein function prediction under machine learning perspective. *Recent Pat. Biotechnol.* **2013**, *7*, 122–141.
- (13) Aggarwal, D.; Hasija, Y. A review of deep learning techniques for protein function prediction. 2022. arXiv:2211.09705. arXiv.org e-Printarchive. <https://doi.org/10.48550/arXiv.2211.09705>.
- (14) Elhaj-Abdou, M. E.; El-Dib, H.; El-Helw, A.; El-Habrouk, M. Deep_CNN_LSTM_GO: Protein function prediction from amino-acid sequences. *Comput. Biol. Chem.* **2021**, *95*, No. 107584.
- (15) Li, X.; Qian, Y.; Hu, Y.; Chen, J.; Yue, H.; Deng, L. MSF-PFP: A novel multisource feature fusion model for protein function prediction. *J. Chem. Inf. Model.* **2024**, *64*, 1502–1511.
- (16) Kulmanov, M.; Hoehndorf, R. DeepGOPlus: Improved protein function prediction from sequence. *Bioinformatics* **2020**, *36*, 422–429.
- (17) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2016239118.
- (18) Ahmed, E.; Heinzinger, M.; Dallago, C.; Rihawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Martin, S.; et al. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44* (10), 7112–7127, DOI: [10.1109/tpami.2021.3095381](https://doi.org/10.1109/tpami.2021.3095381).
- (19) Vitale, R.; Bugnon, L. A.; Fenoy, E. L.; Milone, D. H.; Stegmayer, G. Evaluating large language models for annotating proteins. *Briefings Bioinf.* **2024**, *25*, No. bbae177.
- (20) Zhao, Y.; Yang, Z.; Hong, Y.; Wang, L.; Zhang, Y.; Lin, H.; Wang, J. Improving protein function prediction by adaptively fusing information from protein sequences and biomedical literature. *IEEE J. Biomed. Health Inf.* **2023**, *27*, 1140–1148.
- (21) Zhu, Y.-H.; Zhang, C.; Yu, D.-J.; Zhang, Y. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction. *PLoS Comput. Biol.* **2022**, *18*, No. e1010793.
- (22) Zhang, Z.; Zhu, Q.; Xie, G.-S.; Chen, Y.; Li, Z.; Wang, S. Discriminative margin-sensitive autoencoder for collective multi-view disease analysis. *Neural Networks* **2020**, *123*, 94–107.
- (23) Liu, C.-L.; Xiao, B.; Hsiao, W.-H.; Tseng, V. S. Epileptic seizure prediction with multi-view convolutional neural networks. *IEEE Access* **2019**, *7*, 170352–170361.
- (24) Zheng, L.; Shi, S.; Lu, M.; Fang, P.; Pan, Z.; Zhang, H.; Zhou, Z.; Zhang, H.; Mou, M.; Huang, S.; et al. AnnoPRO: A strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. *Genome Biol.* **2024**, *25*, No. 41.
- (25) Rao, H. B.; Zhu, F.; Yang, G.; Li, Z.; Chen, Y. Z. Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **2011**, *39*, W385–W390.
- (26) Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304.

- (27) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. arXiv:1802.03426. arXiv.org e-Printarchive. <https://doi.org/10.48550/arXiv.1802.03426>.
- (28) Jonker, R.; Volgenant, T. *A Shortest Augmenting Path Algorithm for Dense and Sparse Linear Assignment Problems*; DGOR/NSOR: Papers of the 16th Annual Meeting of DGOR in Cooperation with NSOR/Vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR; Springer: Berlin, Heidelberg, 1988; p 622.
- (29) Song, F. V.; Su, J.; Huang, S.; Zhang, N.; Li, K.; Ni, M.; Liao, M. DeepSS2GO: Protein function prediction from secondary structure. *Briefings Bioinf.* **2024**, *25*, No. bbae196.
- (30) Heinzinger, M.; Weissenow, K.; Sanchez, J. G.; Henkel, A.; Mirdita, M.; Steinegger, M.; Rost, B. Bilingual language model for protein sequence and structure. *NAR Genomics Bioinf.* **2023**, *6* (4), No. lqae150.
- (31) Ofer, D.; Brandes, N.; Linial, M. The language of proteins: NLP, machine learning & protein sequences. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1750–1758.
- (32) Madani, A.; McCann, B.; Naik, N.; Keskar, N. S.; Anand, N.; Eguchi, R. R.; Huang, P.-S.; Socher, R. Progen: Language Modeling for Protein Generation. 2020. arXiv:2004.03497. arXiv.org e-Printarchive. <https://doi.org/10.48550/arXiv.2004.03497>.
- (33) Strodtthoff, N.; Wagner, P.; Wenzel, M.; Samek, W. UDSMProt: Universal deep sequence models for protein classification. *Bioinformatics* **2020**, *36*, 2401–2409.
- (34) Jeong, J. C.; Lin, X.; Chen, X.-W. On position-specific scoring matrix for protein function prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2010**, *8*, 308–315.
- (35) Zou, L.; Nan, C.; Hu, F. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* **2013**, *29*, 3135–3142.
- (36) Zahiri, J.; Yaghoubi, O.; Mohammad-Noori, M.; Ebrahimpour, R.; Masoudi-Nejad, A. PPIevo: Protein–protein interaction prediction from PSSM based evolutionary information. *Genomics* **2013**, *102*, 237–242.
- (37) Liu, T.; Zheng, X.; Wang, J. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. *Biochimie* **2010**, *92*, 1330–1334.
- (38) Saini, H.; Raicar, G.; Lal, S. P.; Dehzangi, A.; Imoto, S.; Sharma, A. Protein fold recognition using genetic algorithm optimized voting scheme and profile bigram. *J. Software* **2016**, *11*, 756–767.
- (39) Zhang, L.; Zhao, X.; Kong, L. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2014**, *355*, 105–110.
- (40) Daneshvar, M. B. Scale invariant feature transform plus hue feature. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2017**, *XLII-2/W6*, 27–32.