JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

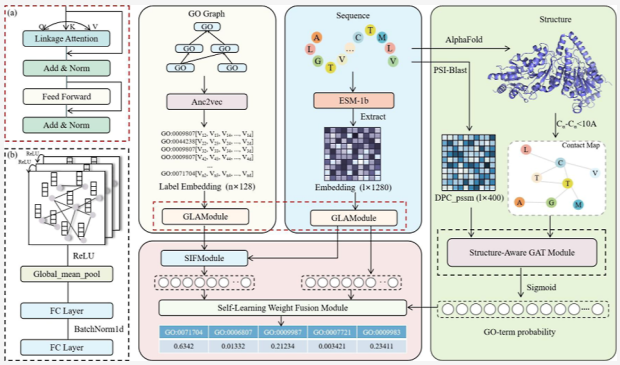
# DeepMFFGO: A Protein Function Prediction Method for Large-Scale Multifeature Fusion

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

Cite This: J. Chem. Inf. Model. 2025, 65, 3841−3853 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Protein functional studies are crucial in the fields of drug target discovery and drug design. However, the existing methods have significant bottlenecks in utilizing multisource data fusion and Gene Ontology (GO) hierarchy. To this end, this study innovatively proposes the DeepMFFGO model designed for protein function prediction under large-scale multifeature fusion. A fine-tuning strategy using intermediate-level feature selection is proposed to reduce redundancy in protein sequences and mitigate distortion of the top-level features. A hierarchical progressive fusion structure is designed to explore feature connections, optimize complementarity through dynamic weight allocation, and reduce redundant interference. On the CAFA3 data set, the $F_{max}$ values of the DeepMFFGO model on the MF, BP, and CC ontologies reach 0.702, 0.599, and 0.704, respectively, which are improved by 4.2%, 2.4%, and 0.07%, respectively, compared with state-of-the-art multisource methods.




## INTRODUCTION

Proteins, as the core molecules that construct the basic framework of organisms and perform life functions, their functional studies have irreplaceable importance in drug target discovery, revealing disease mechanisms and guiding drug design.[1,2] With the rapid progress of biotechnology, especially the popularization of high-throughput sequencing technology, protein sequence data have shown unprecedented explosive growth.[3] However, this flood of data also presents unprecedented challenges. Less than 0.1% of the proteins in the Uniprot database have been experimentally annotated,[4] making it an urgent challenge to accurately and efficiently predict protein function from massive sequence data.

Traditional protein function prediction methods mainly rely on sequence similarity analysis, such as FASTA,[5] BLAST,[6] and diamond.[7] These methods are grounded in the principle of sequence homology. They predict the functions of unknown proteins by comparing their sequences with those of proteins whose functions are already known. However, the validity of these methods relies heavily on the sequence similarity threshold, and usually, the prediction results have high confidence only when the similarity is 60% or higher.[8] In addition, prediction accuracy is often limited by the lack of sufficient known functional proteins to serve as references for newly discovered protein sequences.

In order to break through these limitations, researchers have actively explored new methods to extract deep features from protein sequences using deep learning techniques. Among them, DeepGOPlus,[9] as a classic in deep learning, employs a convolutional neural network to extract sequence features and significantly improves the performance of protein function prediction by fusing sequence features with homology information. MMSMAPlus[10] utilizes multiscale separable convolutional technology to deeply explore the deep semantic features, evolutionary features, amino acid species features, and physicochemical properties of protein sequences, which provides richer information dimensions for function prediction. PhiGnet[11] uses a two-channel graph convolutional network to learn pretrained feature embeddings and integrates them with evolutionary couplings (EVC) and residue communities (RC) to accurately infer protein functional annotations, even without structural templates or low homology sequences.

However, the above methods rely only on sequence information and fail to fully mine and fully utilize the multisource data of existing proteins. In order to compensate for the limitation of single information and provide a more comprehensive protein characterization, in recent years, researchers have begun to explore protein function prediction methods that integrate multiple types of data, including protein 3D structural information, GO hierarchies, protein−protein interactions(PPI), and biomedical literature. These multi-

Received: January 17, 2025
Revised: March 12, 2025
Accepted: March 12, 2025
Published: March 21, 2025





https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

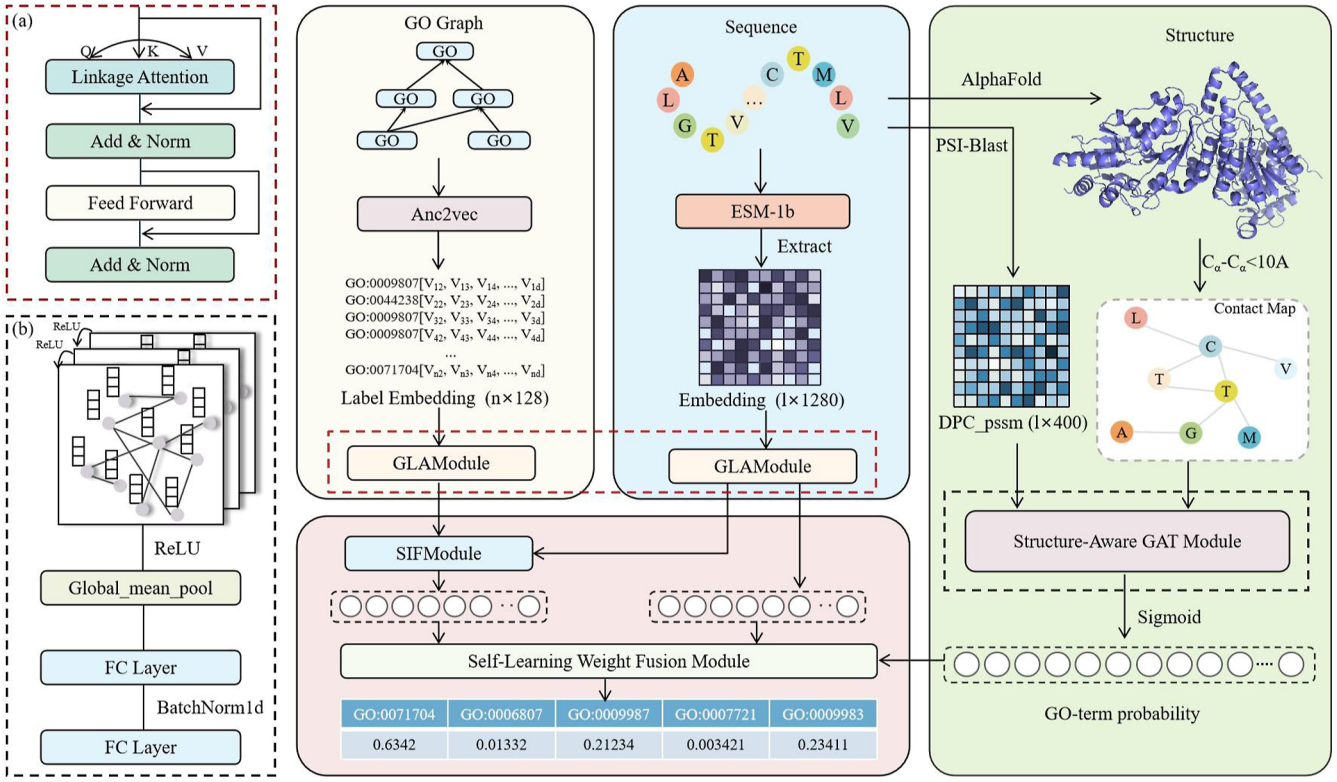



**Figure 1.** Overview of the DeepMFFGO method. First, the input features are encoded to encode the protein sequence, structure, GO hierarchy information, and PSSM matrix into a feature matrix. Second, the feature matrix is fed into a well-designed feature extractor to obtain deep semantic information (where l is the number of proteins), labeled semantic network information, and 3D structure information. Finally, the features of the three channels are passed through the sequence−label interaction fusion layer and the self-learning weighted fusion module to obtain the final prediction results. (a) Global linkage aggregation module and (b) structure-aware graph attention feature extractor.

source data integration methods significantly outperform single-source function prediction methods using only sequence information in effectiveness. For example, the GAT-GO[12] method utilizes a pretrained protein language model to extract features and dig deeper into the contact graph information between residues through graph attention networks to learn structure−function relationships. GNNGO3D[13] uses four graph convolutions to learn sequence-structure−function relationships and uses a multistage feature fusion strategy to improve the accuracy of function prediction significantly. The DeepSS2GO[14] method innovatively utilizes a multilayer perceptron to extract features from secondary structures and supplements the sequence and homology information, realizing an organic combination of the sequencing efficiency and the accuracy of partial spatial structure information. The MSF-PFP[15] method takes the sequence, structural domains, and PPI of proteins as inputs and focuses on protein sequences, structural domains, and interaction information through a specialized feature extraction module, which is ultimately fed into the multilayer fully connected network for fusion classification.

While many contemporary protein function prediction methods are capable of effectively incorporating protein structure information, they frequently overlook the hierarchical structure information on GO. As a result, they do not fully harness the potential value of this hierarchy, leading to limited improvements in performance. In addition, when the protein language model is employed as an encoder, a common practice is to utilize only the top-level representation of the model. This representation is the output of the final layer. However, this approach does not fully exploit the rich semantic information that is embedded within the internal representations of the model.[16] This approach, which employs top-level characterization of the encoder, leads to overly complex coding structures that trigger information distortion and may introduce noise and feature errors, leading to feature loss and biased attention mechanisms. Given this, there is an urgent need to optimize how representations are used in protein language models to improve the models' performance and accuracy. More importantly, the current trend in protein function prediction is multifeature fusion, aiming to achieve more comprehensive and accurate prediction results. Indeed, the prevalent methodologies in current solutions often merely superimpose data sources, overlooking the intricate relationships of complementarity and redundancy among features. This oversight produces a critical deficiency in feature selection and weight distribution strategies. As data volumes swell and feature dimensions expand, the challenge of executing feature fusion efficiently becomes more pronounced, culminating in a squandering of computational resources. This brings forth a pivotal inquiry: How should we fully utilize the multisource data of proteins, delve into the complementarity between data for feature fusion, and thereby enhance the accuracy of protein function prediction? To this end, we urgently need innovative models to fully exploit the multisource information potential of proteins and effectively utilize the advantages of feature fusion in terms of information complementarity.

To systematically address the key shortcomings of the existing methods in protein function prediction, this study innovatively proposes the DeepMFFGO model to cope with


https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

protein function prediction under large-scale multifeature fusion. To address the problem of information distortion and the introduction of noise caused by using only the top-layer representation in protein language models, we propose a fine-tuning strategy based on intermediate-layer feature selection, which effectively reduces the redundant feature information on protein sequences. In order to overcome the problem of underutilization of linkages in protein data, we propose a global linkage aggregation module in this research. This module slices protein data into multiple linkages with accessible information, mines features related to their functions on each linkage, and finally performs the aggregation operation. It empowers the model to capture the deep semantic features of the sequences and labels. To address the redundancy problem caused by the simple superposition of multisource features, this study innovatively proposes a hierarchical progressive fusion structure. Through this structure, this research effectively integrates four key pieces of information, namely, sequence information, structure information, GO hierarchy, and PSSM, and realizes feature information's complementary and synergistic enhancement. By deeply mining the intrinsic connection between the data, our model can generate more accurate and reliable predictions of functional annotations. Comparative experiments show that the DeepMFFGO model significantly outperforms the existing state-of-the-art models based on multisource protein information. The ablation experiments strongly corroborate the effectiveness of the modules in DeepMFFGO, with the contributions of the three modules being 0.407, 0.179, and 0.414, respectively. These experiments fully demonstrate our substantial progress in optimizing the utilization of the protein language model representations, reducing the feature redundancy and improving the prediction accuracy.

## METHODS

**Global Linkage Aggregation Module.** In bioinformatics and computational biology, feature extraction is critical in protein structure prediction, functional analysis, and interaction studies. The existing feature extraction modules such as CNN or RNN focus on capturing local features or short-range dependencies. However, features such as protein, GO, and PPI, whose functions are often determined by long-range dependencies, involve interactions between widely separated positions in the feature vector. Due to the inherent limitations of traditional feature extraction methods in extracting long-range dependencies, they cannot adequately capture global features. The absence of global features may mean that important biological information is not utilized, thus affecting the predictive power and interpretability of the model.

To address these challenges, this research proposes a global linkage aggregation module (GLAM) based on the Encoder part of Transformer,[17] as shown in Figure 1a. The GLA module endows the model with the ability to capture local detailed features and long-range dependencies through its surrounding information capture and integration properties. The GLA module consists of linkage attention, residual connectivity, normalization, and position feedforward networks, and these components work together to achieve a deep understanding and feature extraction of complex interaction patterns of protein and GO. The linkage attention mechanism is capable of subdividing into multiple linkages in the embedding dimension and capturing link information across neighboring linkages. This ensures that the model can effectively extract and utilize deeper connections between nodes when dealing with complex network structures. In the linkage attention mechanism, each feature vector is linearly transformed to generate a multiset collection of vectors $Q$, $K$, and $V$. $Q$ and $K$ are used to determine the fusion weights, and $V$ is the fused value. The linkage attention mechanism is computed as follows

$$\text{Linkage attention}\ (Q, K, V) = \text{concat}\ (\text{linkage}_1, ..., \text{linkage}_H)W^{\text{O}} \tag{1}$$

where $H$ is the number of linkages, and $\text{linkage}_h$ is defined as

$$\text{linkage}_h = \text{attention}\ (Q\cdot W_h^Q, K\cdot W_h^K, V\cdot W_h^V) \tag{2}$$

$$\text{Attention}\ (Q, K, V) = \text{softmax}\left(\frac{Q\cdot K^T}{\sqrt{d_k}}\right)V \tag{3}$$

where $W_h^Q \in \mathbb{R}^{d\times d_k}$, $W_h^K \in \mathbb{R}^{d\times d_k}$, $W_h^V \in \mathbb{R}^{d\times d_k}$, and $W_h^O \in \mathbb{R}^{d\times d_k}$ are projection weight parameter matrices. Here, $d$ is the dimension size of hidden embedding vectors and $d_k = d/h = 1280$. After the linkage attention mechanism layer, we use a feed-forward neural network and layer normalization to learn a hierarchical feature representation of the data and stabilize the input distribution for each layer. The feed-forward neural network is computed as follows

$$\text{FFN}(h) = W_2\cdot\sigma(W_1\cdot h + b_1) + b_2 \tag{4}$$

where $\sigma$ is the nonlinear activation function, $W_1 \in \mathbb{R}^{d\times d_l}$ and $W_2 \in \mathbb{R}^{d_l\times d}$ are the weight parameter matrices of the feed-forward network, $b_1 \in \mathbb{R}^{d_l}$ and $b_2 \in \mathbb{R}^d$ are bias parameter vectors, and $d_l = 1280$.

In an in-depth exploration of the extraction process of protein sequence features, this research utilizes the pretrained protein language model ESM-1b[18] as an encoder, which was pretrained on the Uniref50 data set. To effectively reduce redundant information and improve prediction accuracy and efficiency, we designed a fine-tuning method to use the optimally performing intermediate layer (layer 29) in the ESM-1b model as the feature output layer. Proteins of length l are encoded as $E_1 \in \mathbb{R}^{l\times 1280}$ residue-level features. The deep semantic features obtained from the encoding include the physicochemical properties of amino acids, residue direct context information, and expression and regulatory feature information. To further mine the long-range dependencies and global information in these deep semantic features, we input the encoded feature matrix into GLAM. When processing sequence data, the GLAM can capture its long-range dependencies and integrate self-attention and local features to extract more compact and information-rich deep semantic features, represented as $V_s \in \mathbb{R}^{1280}$. These features not only reveal sequences in key regions that are closely related to specific functions but also capture long-range interactions within the sequence. As a result, this approach provides a more comprehensive information base for protein function prediction tasks.

The GLAM proposed in this research demonstrates excellent protein feature extraction capabilities. It is designed to be highly generalizable and equally suitable for extracting GO term hierarchy features. To further validate the generalization and effectiveness of the module, this research generates



https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

embeddings of GO terms using the pretrained model Anc2vec,[19] an unsupervised neural network model for learning GO term embeddings that embed GO terms into Euclidean space, and these embeddings preserve the terms, their ancestral terms, and ontological information uniquely. For GO terms $G_i$, they are embedded into n-dimensional labeled representation vectors $G_i \in \mathbb{R}^n$ where n is the hidden dimension, which is set to 128 dimensions to balance the complexity of the model and the representation capability in this study. Next, we input these encoded GO term features into the GLAM, which can capture the vector representations of their parent and ancestor terms when processing GO data and thus extract vectors that contain richly annotated semantic network features, represented as $V_l \in \mathbb{R}^{1280}$. These feature vectors capture the positions and roles of GO terms in the hierarchical structure, thus providing a robust feature representation for the function prediction task.

By successfully extracting deep semantic features of protein sequences and deep associations between GO terms, we demonstrate the promising and powerful potential of the GLAM proposed in this research for various applications in different bioinformatics fields.

**Hierarchical Progressive Fusion Structure.** In terms of feature fusion, this research innovatively proposes a hierarchical progressive fusion structure (HPFS) and designs specialized fusion methods at three key feature fusion stages. First, this research proposes the structure-aware graph attention (SGAM), which skillfully integrates the 3D structural features of proteins and PSSM information; then, this research designs the sequence-label interactive fusion module (SIFM), which effectively combines the deep semantic features of sequences and GO hierarchical structure information. Finally, this research innovatively proposes a self-learning weighted fusion module (SWFM), which integrates the features of the GLAM, SGAM, and SIFM, which fuses the features of three channels. Through this structural design, this research successfully integrates the four types of information, sequence information, structure information, GO hierarchy, and PSSM, with high efficiency and precision, and realizes the deep complementarity and significant enhancement of feature information.

**Structure-Aware Graph Attention Module.** The AlphaFold database[20] stores protein structure information in 3D atomic coordinates. In this research, we obtain the inter-residue contact map from its 3D coordinates by setting a distance threshold of 10 Å. That is, when the distance between the atoms of two amino acids $C_\alpha$ is <10 Å, this research recognizes that the two amino acids are in a contact state. In this research, the contact graph of a protein is constructed as a binary adjacency matrix, in which each amino acid is represented as a node. The edges of the adjacency matrix indicate whether two amino acids are in a contact state. Thus, the spatial structure of the protein is represented as the topological relationship of amino acids in space.

The evolutionary information on the primary structure can be efficiently expressed with the help of PSSM. DeepMFFGO uses the PSI-BLAST algorithm[21] to match the target proteins against the SwissProt database[22] to generate PSSM. These raw PSSM matrices were further processed by calculating the frequency of occurrence of 20 amino acids at each amino acid position and transforming this frequency information into feature vectors for the obtained PSSM matrices, dipeptide composition position specific scoring matrix (DPC-PSSM) combining amino acid composition eigenvectors and dipeptide composition eigenvectors were obtained. A protein sequence is embedded into a 400-dimensional DPC-PSSM vector by summing and averaging the product of the first amino acid and the first amino acid in two adjacent rows. For a length L protein, its DPC-PSSM feature vector is defined as

$$Y = (y_{1,1}, y_{1,2}, \ldots, y_{1,20}, y_{2,1}, y_{2,2}, \ldots, y_{2,20}, \ldots, y_{20,1}, y_{20,2}, \ldots, y_{20,20})^T \tag{5}$$

$$y_{i,j} = \frac{1}{L-1} \sum_{k=1}^{L-1} P_{k,i} \times P_{k+1,j} \quad 1 \le i,\ j \le 20 \tag{6}$$

We propose an SGAM feature extractor, as shown in Figure 1b. The extractor makes the first innovative use of the DPC-PSSM matrix of proteins, which encodes the evolutionary information on amino acid sequences and helps identify conserved regions. At the same time, the contact map of the protein is used to represent the spatial proximity between amino acid residues in the 3D structure of the protein, directing the model to focus on spatially neighboring residues. The combination of the two enables the model to utilize both sequence evolution and structural information to enhance the accuracy of feature extraction. By introducing a graph attention network (GAT), the dependencies between nodes can be effectively captured, and the neighbor information can be aggregated. In the feature extraction process, DPC-PSSM is used as the initial feature of nodes, and the contact graph is the initial feature of edges. At each layer of GAT, the attention mechanism considers the similarity between node features and edge features, thus adjusting the attention weights. This way, high contact probability residues will have higher attention weights to enhance the information flow between them. The GAT network will aggregate and update the node features, thus realizing the deep mining and utilization of protein structure information. First, the attention coefficient is calculated

$$\alpha_{ij} = \mathrm{LeakyReLU}(\alpha^T [W \cdot f_i \| W \cdot f_j \| e_{ij}]) \tag{7}$$

where $\|$ denotes the splicing of feature vectors, $f_i$ and $f_j$ are the features of nodes i and j, respectively, W is the weight matrix, $\alpha$ is the learnable parameter, and LeakyReLU is the activation function. The attention coefficients are then normalized

$$\alpha'_{ij} = \mathrm{softmax}_j(\alpha_{ij}) = \frac{\exp(\alpha_{ij})}{\sum_{k \in N(i)} \exp(\alpha_{ik})} \tag{8}$$

Finally, the node features are updated using the normalized attention coefficients

$$h'_i = \sigma\Big( \sum_{j \in N(i)} \alpha'_{ij} (W \cdot f_j + b) \Big) \tag{9}$$

where $N(i)$ is the set of neighboring nodes of node $i$, $\sigma$ is a nonlinear activation function, and $b$ is a bias term. In addition, to reduce the graph size and extract the key structural information, we introduce the global average pooling (gep pooling) technique, which is applied to select the subset of nodes with the most information. For each node $v$, the information content of each node $v$ is evaluated by an importance scoring function $s(v)$



https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

$$s(v) = \sigma\left(\sum_{u \in N(v)} \alpha'_{uv} h'_u\right) \tag{10}$$

Then, we select the top-k nodes to form a new set of nodes $V'$

$$V' = \{v_i | s(v_i) \in \text{top-k}(s(V))\} \tag{11}$$

where top-$k$ denotes the selection of the k nodes with the highest importance score.

**Sequence-Label Interactive Fusion Module.** We explore the best solution for GO hierarchy incorporation by fusing sequence features with labeled semantic network features through a cross-attention mechanism to form a unique feature channel. This design fully considers the hierarchical information on GO items, realizes the interactive fusion of sequences and labels, and effectively avoids the problem of poor results when label features are used alone. We propose a SIFM to fuse vectors $X_s \in \mathbb{R}^{1280}$ containing deep semantic features of sequences and vectors $X_l \in \mathbb{R}^{1280}$ containing semantic network features of labels. Different linear projections are first applied to map the two matrices ($X_s$, $X_t$) to the query ($Q_s$, $Q_t$), the keys ($K_s$, $K_t$), and the values ($V_s$, $V_t$), which can be represented as follows

$$Q_s = W_Q^s \cdot F_s, \quad K = W_K^s \cdot F_s, \quad V = W_V^s \cdot F_s \tag{12}$$

$$Q_l = W_Q^l \cdot F_l, \quad K = W_K^l \cdot F_l, \quad V = W_V^l \cdot F_l \tag{13}$$

In order to interact the information between two different features for two-way information fusion, the attention score is calculated by exchanging the $Q_s$, $Q_t$ matrix

$$\begin{aligned} &\text{Attention scores } sl \\ &= \text{softmax}\left(\frac{Q_l \cdot Ks^T}{\sqrt{d_k}}\right), \text{ attention scores } ls \\ &= \text{softmax}\left(\frac{Q_s \cdot K_l^T}{\sqrt{d_k}}\right) \end{aligned} \tag{14}$$

where $W_*^* \in \mathbb{R}^{d \times d}$, $d = d_k$ = the number of labels in the ontology. Finally, the attention output is computed and spliced to obtain a protein sequence feature containing the tagged semantic network features

$$\begin{aligned} &\text{Output } sl = \text{attention scores } sl \cdot V_s, \text{ output } ls \\ &= \text{attention scores } ls \cdot V_l \end{aligned} \tag{15}$$

$$\begin{aligned} &\text{Concatenated output} \\ &= \text{cat}(\text{output } sl, \text{ output } ls, \text{ dim} \\ &= -1) \end{aligned} \tag{16}$$

**Self-Learning Weighted Fusion Module.** Based on our previous work, this research successfully extracts three classes of features: first, sequence deep semantic features obtained by GLAM; second, 3D structural features embedded with evolutionary information captured with the help of the SGAM; and third, sequence information fused with label hierarchical features using the sequence−label interaction fusion module. Each type of information constitutes an independent and complete channel feature, one that has unique information and complements each other in order to realize the organic fusion of these three channel features while retaining the unique information contained in each feature and avoiding the redundancy caused by direct fusion; we innovatively propose a SWFM. The module adopts an advanced decision-level fusion strategy, which can dynamically adjust the weight allocation of each channel feature to achieve the optimal fusion effect. During the fusion process, the feature information on each channel is fully respected and effectively utilized, ensuring that the uniqueness and importance of the original features are retained after fusion. The following equation is defined

$$F_{\text{merged}} = \sum_{i=1}^{m} W_i \cdot F_i \tag{17}$$

$$\sum_{i=1}^{m} W_i = 1 \tag{18}$$

where $W_i \in \mathbb{R}$ is the learned weight of the $i$th channel, $W_i \cdot F_i$ is the initial prediction result of the $i$th channel, and $F_{\text{merged}}$ is the prediction result of fusing all the channels.

The SWFM can realize the organic fusion of multiple channel features, ensuring that the synergy between the features is maximized while also accurately retaining the unique information contained in the respective features, avoiding the loss or dilution of key information in the fusion process. It can show higher flexibility and adaptability when dealing with complex data. In addition, it also provides a reference and a lesson for multisource data fusion for other downstream tasks in the field of bioinformatics.

## EXPERIMENTS AND RESULTS

**Assessment of Indicators.** We use three evaluation metrics: $F_{\max}$ (maximum $F$-score), $S_{\min}$, and AUPR (area under the precision-recall curve); $F_{\max}$ and $S_{\min}$ are used as the primary evaluation metrics in CAFA.[23] AUPR is widely used in the evaluation of multilabel classification tasks. AUPR penalizes false positives more than AUC and is therefore more frequently used when a high cost of label acquisition is required. $F_{\max}$ is a protein-centered metric that measures the accuracy of assigning GO terms to proteins

$$\text{F}_{\max} = \max_t \frac{2 \cdot \text{AvgPr}(t) \cdot \text{AvgRc}(t)}{\text{AvgPr}(t) + \text{AvgRc}(t)} \tag{19}$$

$$\text{AvgPr}(t) = \frac{1}{k(t)} \cdot \sum_{i=1}^{k(t)} \text{pr}_i(t) \tag{20}$$

$$\text{AvgRc}(t) = \frac{1}{n} \cdot \sum_{i=1}^{n} \text{rc}_i(t) \tag{21}$$

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

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

where $p$ and $G$ represent proteins and GO terms, respectively, and $n$ is the total number of proteins. The threshold t varies between 0 and 1 in steps of 0.01, and $k(t)$ is the number of proteins with at least one GO term score not lower than the



https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

queue. The function $1(S(p_i,G_j) \geq t)$ is an indicator function that returns 1 or 0 depending on whether the score $S(p_i,G_j)$ is greater than or equal to threshold $t$. The function $1(S(p_i,G_j) \geq t)$ is the number of proteins in the GO item that have scores greater than or equal to the queue value.

$S_{min}$ is a GO term-centered evaluation metric that calculates the semantic distance between the actual and predicted annotations

$$S_{min} = \min_t \sqrt{\mathrm{ru}(t)^2 + \mathrm{mi}(t)^2} \tag{24}$$

$$\mathrm{ru}(t) = \frac{1}{N_T} \sum_{i=1}^{N_T} \sum_j \mathrm{IC}(G_j) \cdot T(G_j, p_i) \cdot 1(S(p_i, G_j) < t) \tag{25}$$

$$\mathrm{mi}(t) = \frac{1}{N_T} \sum_{i=1}^{N_T} \sum_j \mathrm{IC}(G_j) \cdot (1 - T(G_j, p_i)) \cdot 1(S(p_i, G_j) \geq t) \tag{26}$$

$$\mathrm{IC}(G_j) = -\log_2 \Pr(G_j | \mathrm{parent}(G_j)) \tag{27}$$

where $\mathrm{ru}(t)$ is the residual uncertainty. $\mathrm{mi}(t)$ is the error information. $\mathrm{IC}(G_j)$ denotes the information content of the GO term $G_j$.

**Data Set and Parameter.** This study uses the PyTorch 2.0 deep learning framework to build a training environment based on the NVIDIA A40 GPU hardware platform. The model architecture is based on the ESM 1b pretraining model (https://huggingface.co/facebook/esm1b_t33_650M_UR50S) as the encoder. The binary cross entropy is used as the loss function during the training process, together with the Adam optimizer for gradient update, and the initial learning rate is set to 0.0001 to balance the retention of pretrained knowledge and the adaptation of new features. To prevent overfitting, a dropout ratio of 0.2 is set for the network, and the batch size is controlled at 16 to fit the memory capacity. A total of 20 epochs of iterative training are performed. Based on the principle of optimal performance of the validation set, we adopt the model parameters corresponding to the peak of the $F_{max}$ index as the final preferred solution.

In this study, the CAFA3 data set was used for method validation, and the data were obtained from DeepGOPlus public resources. The training set sequences and experimental annotations were released in September 2016, and the test set data were released in November 2017. We strictly screened the raw data to exclude proteins with sequence lengths of more than 1000 amino acids (9% of the total samples) and remove proteins with "fuzzy201d" amino acids (less than 2% of the total samples). The sequence length distribution of the data set is shown in Figure 2. Protein 3D structure data from the AlphaFold database show excellent performance, with a global prediction accuracy of more than 92% and an average error controlled within 1 Å. It has been shown that there is no significant difference between the AlphaFold-predicted structures and the PDB experimentally resolved structures in terms of key indicators.[24,25] The complete statistical information on the experimental data set is detailed in Table 1.

**Comparative Experiments.** The naïve method is an intuitive prediction strategy that annotates proteins based on the frequency of occurrence of GO terms in the data set. The algorithm is uniformly labeled with the same annotations for all samples in the test set and is used as a baseline method in

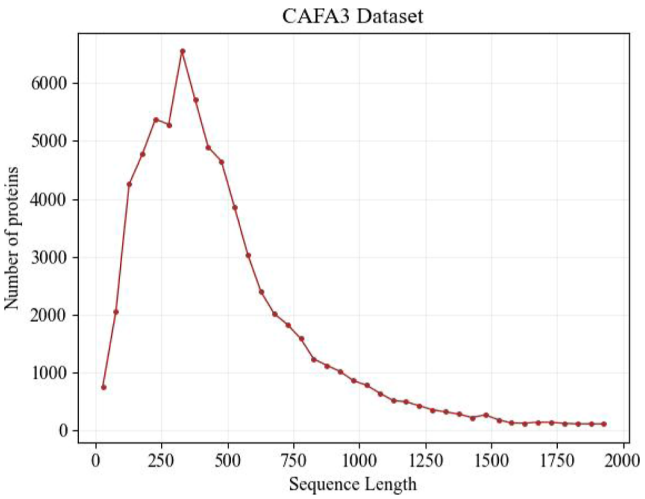



**Figure 2.** Distribution of protein sequence lengths in the CAFA3 data set.

**Table 1. Number of Proteins and Number of GO Terms on the CAFA3 Dataset**

| data set | ontology | training | Validation | test | terms |
|---|---|---|---|---|---|
| CAFA3 | MF | 28,679 | 3228 | 1035 | 677 |
| | BP | 42,250 | 4748 | 2185 | 3992 |
| | CC | 39,893 | 4510 | 1117 | 551 |

CAFA competitions. Diamond is a prediction method based on the sequence similarity of proteins, and its core idea is to assign similar functional annotations to similar proteins. For a given protein p and functional term $G$, the method evaluates the similarity between the two by calculating the set of proteins E with a confidence value greater than 0.001 in sequence similarity to the target protein p and the set of annotations Ts for protein s. The method is based on the idea of assigning similar annotations to similar proteins.

$$\text{Similarity score}\ (p, g) = \frac{\sum_{s \in E} \mathrm{bitscore}(p, s) \cdot I(f \in T_s)}{\sum_{s \in E} \mathrm{bitscore}(p, s)} \tag{28}$$

DeepGOPlus employs a convolutional neural network to extract sequence features and improves protein function prediction performance by integrating sequence features with homology information. This method performed exceptionally well in the CAFA3 challenge, demonstrating its strong prediction capability. DeepFRI[26] constructs pretrained LSTM language models to extract amino acid contact map features from PDB sequences and then further processes and integrates these features through a three-layer GCN. The PredGO[27] method extracts sequence features by utilizing a protein language model trained on many protein sequences. In addition, it utilizes a graph neural network with a geometric vector perceptron (GVP-GNN) to extract information from the protein structures predicted by AlphaFold2. It employs a multihead attention mechanism to integrate PPI features.

According to Figure 3, the diamond method based on sequence homology outperforms the statistically based plain method on the MF and BP ontologies. However, it slightly underperforms the latter in the CC ontology. In contrast, according to Table 2, our model improves the performance by 40.7%, 39.9%, and 25.94% on the MF, BP, and CC ontologies


https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

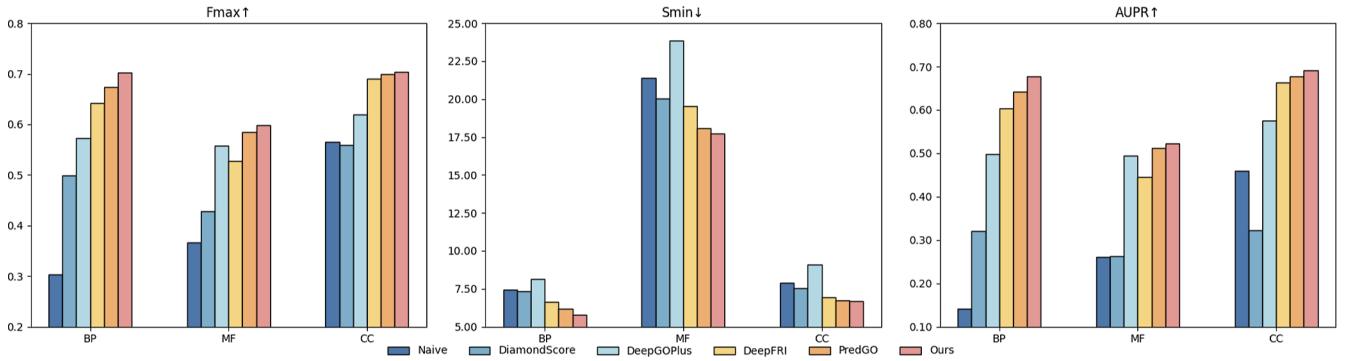



**Figure 3.** $F_{max}$, $S_{min}$, and AUPR scores of our method with other state-of-the-art methods on the CAFA3 data set.

**Table 2. $F_{max}$, $S_{min}$, and AUPR Scores of Our Method with other State-of-the-Art Methods on the CAFA3 Dataset**

| | $F_{max}$ ↑ | | | $S_{min}$ ↓ | | | AUPR ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | MF | BP | CC | MF | BP | CC | MF | BP | CC |
| Naïve | 0.303 | 0.366 | 0.565 | 7.416 | 21.402 | 7.876 | 0.142 | 0.261 | 0.460 |
| DiamondScore | 0.499 | 0.428 | 0.559 | 7.346 | 20.050 | 7.517 | 0.321 | 0.263 | 0.323 |
| DeepGOPlus | 0.572 | 0.558 | 0.620 | 8.154 | 23.869 | 9.083 | 0.498 | 0.495 | 0.576 |
| DeepFRI | 0.642 | 0.528 | 0.690 | 6.626 | 19.518 | 6.923 | 0.603 | 0.446 | 0.663 |
| PredGO | 0.674 | 0.585 | 0.699 | 6.194 | 18.067 | 6.717 | 0.642 | 0.512 | 0.678 |
| Ours | **0.702** | **0.599** | **0.704** | **5.766** | **17.742** | **6.679** | **0.677** | **0.522** | **0.692** |

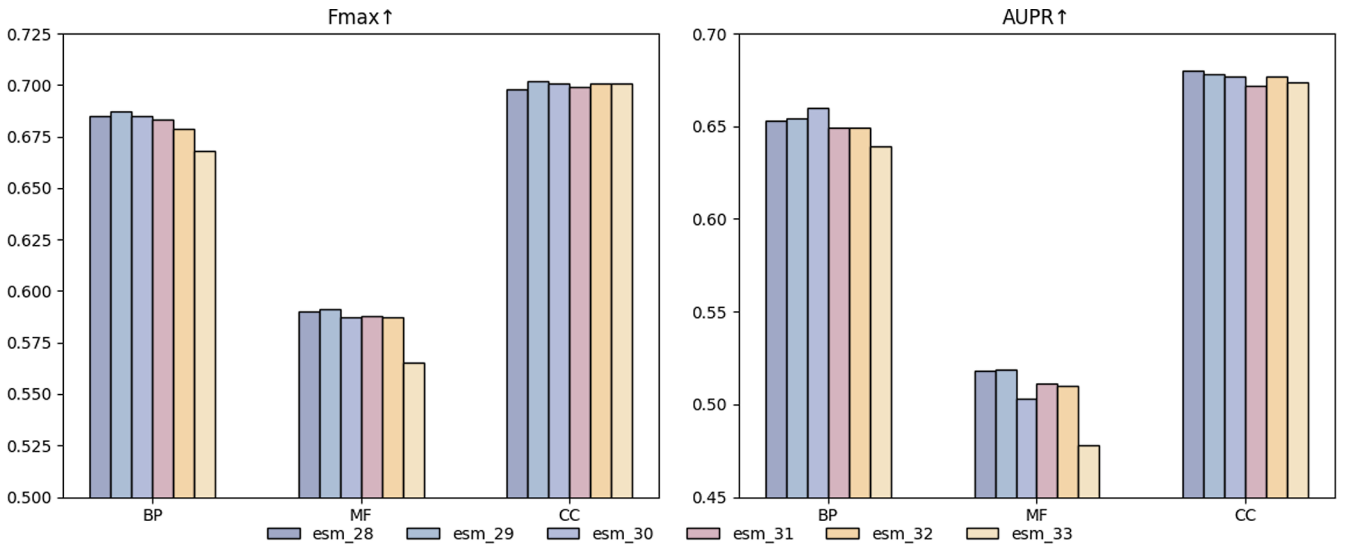



**Figure 4.** Characterization of the output from layers 28−33 of the protein language model represents the $F_{max}$ and AUPR scores of the output results via the GLAM.

compared with the diamond method, respectively. This fully demonstrates that the deep learning model can mine deeper information from massive sequences or other features, significantly improving the accuracy and reliability of function annotation prediction. This performance enhancement is mainly due to the innovative design of our model in feature representation learning, feature extraction and fusion, which enables the model to capture the complex patterns and implicit functional information in protein sequences more effectively.

In comparison with the DeepGOPlus method, which relies only on sequence information, our model achieves 22.7%, 7.3%, and 10.9% performance enhancements on the MF, BP, and CC ontologies, respectively. These results indicate that integrating multiple sources of information can effectively address the limitations of relying on a single type of data. This fusion provides a more comprehensive set of protein-level features, enhancing the thoroughness and accuracy of the functional annotation predictions.

Compared with the DeepFRI method, the $F_{max}$ metrics of our model on the MF, BP, and CC ontologies were improved by 9.30%, 13.40%, and 2.00%, respectively. In comparison with the PredGO method, our $F_{max}$ on the MF and BP ontologies improved by 4.20% and 2.40%, respectively, while $S_{min}$ decreased by 6.90% and 1.80% and AUPR improved by 5.5% and 2.0%. Although the enhancement on the CC ontology is relatively tiny, $F_{max}$ is only enhanced by 0.70%, $S_{min}$ is reduced by 0.6%, and the AUPR is enhanced by 2.1%, it reflects the advantage of our model in the fusion of multisource information. This result may be because the protein structural information introduced by our model is not as apparent as the


https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

Table 3. Characterization of the Output from Layers 28−33 of the Protein Language Model Represents the $F_{max}$, $S_{min}$, and AUPR Scores of the Output Results via the GLAM

| | $F_{max}$ ↑ | | | $S_{min}$ ↓ | | | AUPR ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | MF | BP | CC | MF | BP | CC | MF | BP | CC |
| esm1_28 | 0.685 | 0.590 | 0.698 | 5.974 | **17.790** | 6.816 | 0.653 | 0.518 | **0.680** |
| esm1_29 | **0.687** | **0.591** | **0.702** | 5.933 | 17.795 | **6.731** | 0.654 | **0.519** | 0.678 |
| esm1_30 | 0.685 | 0.587 | 0.701 | **5.902** | 18.184 | 6.801 | **0.660** | 0.503 | 0.677 |
| esm1_31 | 0.683 | 0.588 | 0.699 | 6.026 | 17.938 | 6.740 | 0.649 | 0.511 | 0.672 |
| esm1_32 | 0.679 | 0.587 | 0.701 | 6.084 | 17.796 | 6.686 | 0.649 | 0.510 | 0.677 |
| esm1_33 | 0.668 | 0.565 | 0.701 | 6.215 | 18.630 | 6.822 | 0.639 | 0.478 | 0.674 |

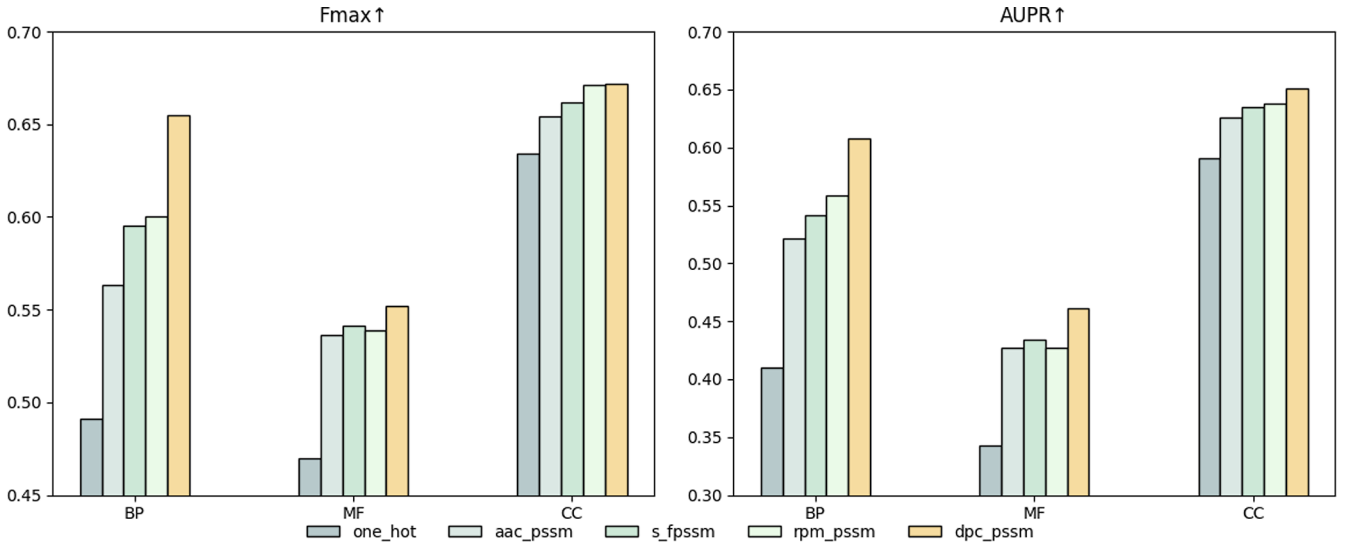



Figure 5. SGAM feature extractor combines different features of $F_{max}$ and AUPR scores.

protein interaction information employed by PredGO on the CC ontology. Nevertheless, our model still shows strong competitiveness in the functional annotation prediction.

**Ablation Experiment.** *ESM-1b Fine-Tuning Strategy Based on Middle Layer Feature Selection Reduces Redundant Feature Information.* The choice of feature extractor significantly impacts prediction performance in the critical bioinformatics task of protein function prediction. In recent years, protein language models have attracted much attention due to their powerful sequence representation capabilities. However, choosing the appropriate feature representation layer has become an urgent problem when applying such models for feature extraction.

Each layer in the 33-layer architecture of ESM-1b outputs its feature embedding, and traditional approaches often tend to select the feature representation of the last layer of the model to capture the most comprehensive information. However, due to the deeply stacked structure of the protein language model, the feature representation of the last layer may contain redundant information that adversely affects the prediction accuracy. To address this problem, this study proposes an innovative ESM-1b fine-tuning strategy based on the intermediate layer feature selection. We experimentally evaluated the performance of the feature representations output from the last six layers (28−33) of ESM-1b, and Figure 4 illustrates the $F_{max}$ and AUPR metrics for each layer. According to the results in Table 3, ESM-1b_29 exhibits optimal performance with $F_{max}$ metrics of 0.687, 0.591, and 0.702 for the MF, BP, and CC ontologies, respectively. Based on this finding, we design a fine-tuning method to use the optimally performing intermediate layer (layer 29) in the ESM-1b model as the feature output layer and extract the feature representation output from this layer for downstream tasks. This method can effectively reduce redundant information and improve prediction accuracy and efficiency. Implementing this strategy provides a new direction for designing and optimizing future protein language models. It emphasizes the importance of the properties and utility of feature representations at different levels to achieve more accurate and efficient protein function prediction. We can provide a new and effective feature extraction and fine-tuning approach for protein function prediction tasks.

*Significant Performance Improvement in the Combined DPC-PSSM Matrix Validates the Predictive Value of Evolutionary Information.* In protein function prediction, the optimal choice of feature representation is a key determinant of model performance. Given the central role of evolutionary information in revealing protein function, this study systematically integrates multiple protein feature representations—one-hot coding (characterizing amino acid class information) and PSSM and its derived representations (aac-PSSM, rpm-PSSM, s-fPSSM, and DPC-PSSM)—and comprehensively evaluates them through SGAM modeling to reveal the different information types' differences in prediction efficacy. Specifically, aac-PSSM is a single-residue global statistic; rpm-PSSM employs row-averaged compression to preserve the conservative trend of global evolution; s-fPSSM focuses on global statistical features; and DPC-PSSM captures the coevolutionary pattern of local residue pairs through dipeptide composition.



https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

Table 4. SGAM Feature Extractor Combines Different Features of $F_{max}$, $S_{min}$, and AUPR Scores

| | $F_{max}$ ↑ | | | $S_{min}$ ↓ | | | AUPR ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | MF | BP | CC | MF | BP | CC | MF | BP | CC |
| one_hot | 0.491 | 0.470 | 0.634 | 8.383 | 20.460 | 7.711 | 0.410 | 0.343 | 0.591 |
| aac_pssm | 0.563 | 0.536 | 0.654 | 7.651 | 19.373 | 7.560 | 0.521 | 0.427 | 0.626 |
| s_fpssm | 0.595 | 0.541 | 0.662 | 7.177 | 19.119 | 7.389 | 0.541 | 0.434 | 0.635 |
| rpm_pssm | 0.600 | 0.539 | 0.671 | 7.206 | 19.404 | **7.197** | 0.559 | 0.427 | 0.638 |
| dpc_pssm | **0.655** | **0.552** | **0.672** | **6.474** | **19.018** | 7.228 | **0.608** | **0.461** | **0.651** |

Table 5. $F_{max}$, $S_{min}$, and AUPR Scores of Different Methods on GLAM

| | $F_{max}$ ↑ | | | $S_{min}$ ↓ | | | AUPR ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | MF | BP | CC | MF | BP | CC | MF | BP | CC |
| MLP | 0.678 | 0.587 | 0.698 | 6.113 | 18.089 | 6.835 | 0.655 | 0.515 | 0.675 |
| CNN 1d | 0.681 | **0.595** | 0.698 | 6.137 | 17.953 | 6.816 | 0.651 | 0.515 | 0.674 |
| GRU | 0.683 | 0.582 | 0.699 | 5.987 | 18.526 | 6.837 | **0.660** | 0.511 | 0.676 |
| MutilHeadAttention | **0.687** | 0.586 | 0.698 | 6.042 | 18.256 | 6.826 | 0.653 | 0.504 | 0.676 |
| GLA Module | **0.687** | 0.591 | **0.702** | **5.933** | **17.795** | **6.731** | 0.654 | **0.519** | **0.678** |

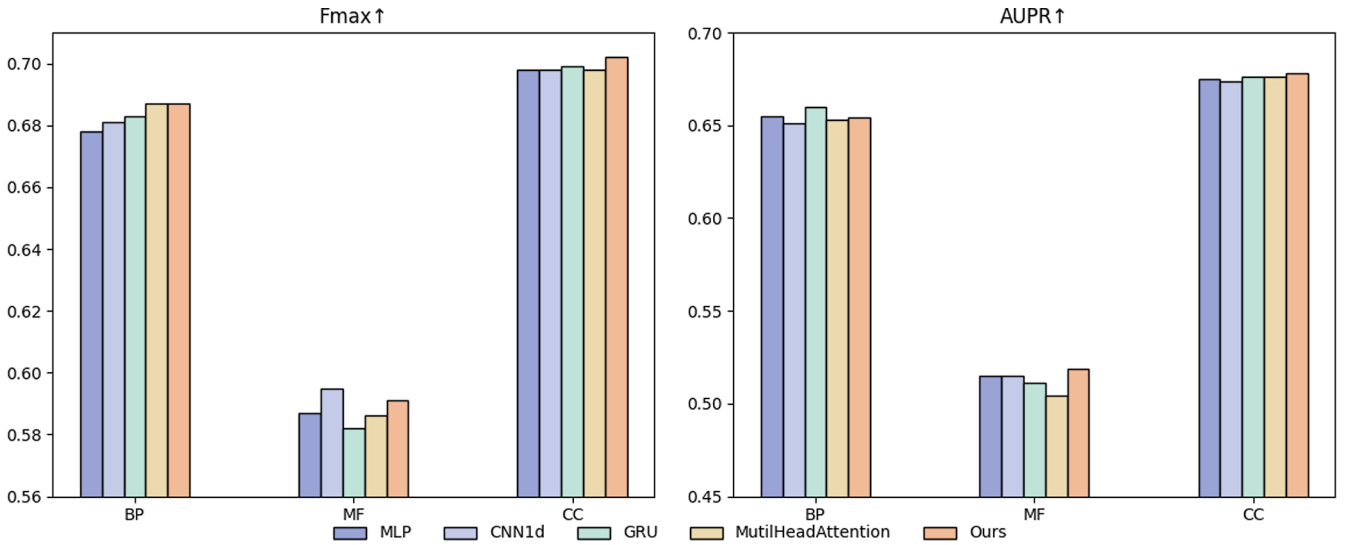



Figure 6. $F_{max}$ and AUPR scores of different methods on GLAM.

Figure 5 shows that DPC-PSSM based on the dipeptide composition exhibits significant advantages in $F_{max}$ and AUPR metrics for all three ontology classifications: MF, BP, and CC. The quantitative analysis of Table 4 shows that compared with the traditional one-hot encoding, DPC-PSSM improves the $F_{max}$ metrics of MF, BP, and CC ontologies by 0.16, 0.08, and 0.04, respectively, and this improvement fully validates the gain effect of the evolutionary information on the prediction of protein functions.

In-depth mechanistic analysis reveals that the superiority of DPC-PSSM stems from its unique local coevolutionary characterization ability. Compared with aac-PSSM, which focuses on only single residue frequencies, rpm-PSSM, which may lose positional information, and s-fPSSM, which focuses on global characterization. DPC-PSSM can efficiently model local structural domain features closely related to protein functions by capturing coevolutionary patterns of adjacent amino acid pairs. This feature is highly compatible with the biological law that protein functional sites are usually formed by short-range residue interactions, which significantly improves the function prediction accuracy. The results confirm that the modeling of local coevolutionary information has more substantial discriminative power than global statistical features in protein function prediction.

*Global Linkage Aggregation Module Prove Its Effectiveness in Sequence Feature Extraction.* The technical approach to protein function prediction, an important application of natural language processing in bioinformatics, has advanced with the continuous development of natural language models. Natural language models have evolved remarkably from CNN to transformer, which provides new technical perspectives for protein function prediction. In order to systematically evaluate the performance of different models in protein function prediction, we conducted a detailed comparative study of the GLA module with MLP, CNN, gated recurrent units (GRU), and the multi-head attention mechanism. The experimental results are detailed in Table 5, where the GLAM achieves $F_{max}$ metrics of 0.687, 0.591, and 0.702 on the MF, BP, and CC ontologies, respectively. The results in Figure 6 show that GLAM significantly outperforms the other models in terms of the $F_{max}$ and AUPR metrics and achieves either optimal or optimal results. This excellent performance demonstrates the powerful ability of the GLAM to capture complex patterns and relationships in protein sequences. The GLAM effectively


https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

Table 6. $F_{max}$, $S_{min}$, and AUPR Scores of Different Methods on SGAM Feature Extractors

| | $F_{max}$ ↑ | | | $S_{min}$ ↓ | | | AUPR ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | MF | BP | CC | MF | BP | CC | MF | BP | CC |
| GCN + gap | 0.640 | 0.551 | 0.665 | 6.699 | 19.031 | 7.351 | 0.587 | 0.432 | 0.643 |
| GCN + gmp | 0.644 | 0.550 | 0.666 | 6.714 | 19.154 | 7.322 | 0.578 | 0.440 | 0.640 |
| GCN + gep | 0.646 | **0.556** | 0.670 | 6.606 | 19.040 | 7.322 | 0.607 | 0.440 | 0.642 |
| GAT + gap | 0.638 | 0.549 | 0.664 | 6.797 | 19.173 | 7.445 | 0.598 | 0.425 | 0.640 |
| GAT + gmp | 0.640 | 0.546 | 0.663 | 6.690 | 19.275 | 7.395 | 0.593 | 0.445 | 0.644 |
| GAT + gep | **0.660** | 0.552 | **0.672** | **6.471** | **19.018** | **7.228** | **0.611** | **0.461** | **0.651** |

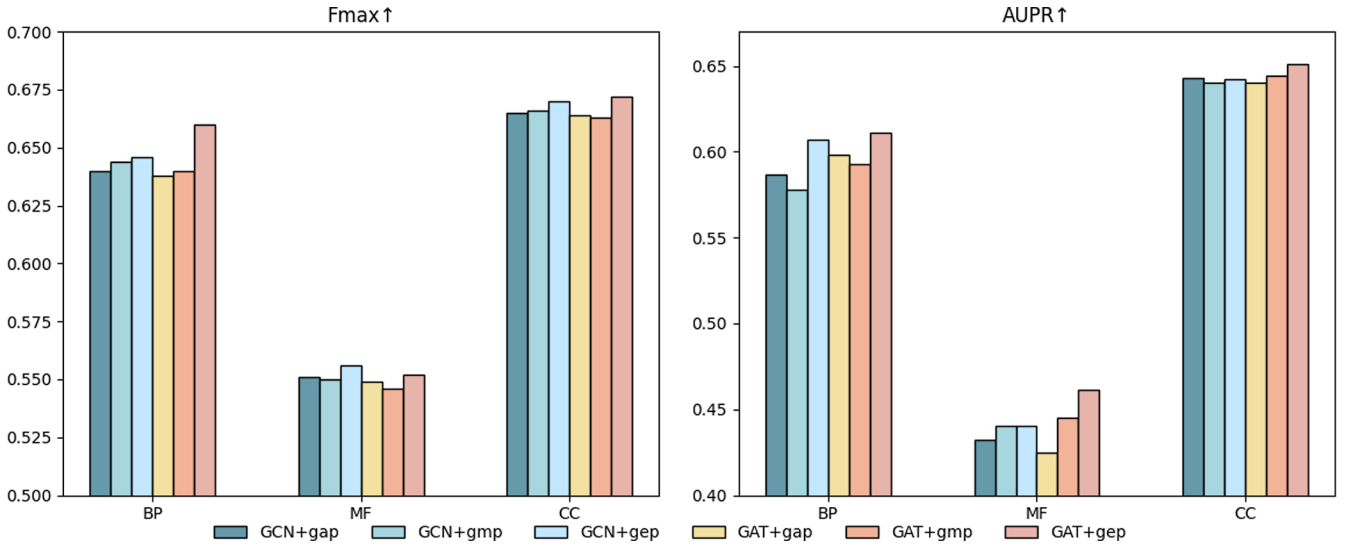



Figure 7. $F_{max}$ and AUPR scores of different methods on SGAM feature extractors.

Table 7. $F_{max}$, $S_{min}$, and AUPR Scores for 3-Channel Ablation Experiments

| HPFS | | | $F_{max}$ ↑ | | | $S_{min}$ ↓ | | | AUPR ↑ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ESM | SIFM | SGAM | MF | BP | CC | MF | BP | CC | MF | BP | CC |
| √ | | | 0.687 | 0.591 | 0.699 | 5.933 | 17.795 | 6.768 | 0.654 | 0.519 | 0.680 |
| | √ | | 0.685 | 0.567 | 0.697 | 6.001 | 18.413 | 6.783 | 0.657 | 0.473 | 0.683 |
| | | √ | 0.660 | 0.552 | 0.672 | 6.471 | 19.018 | 7.228 | 0.611 | 0.461 | 0.651 |
| √ | √ | | 0.693 | 0.598 | **0.704** | 5.876 | **17.682** | 6.705 | 0.666 | 0.517 | 0.688 |
| √ | | √ | 0.696 | 0.592 | 0.703 | 5.798 | 17.779 | 6.718 | 0.666 | **0.524** | 0.685 |
| | √ | √ | 0.681 | 0.577 | 0.700 | 6.059 | 18.231 | 6.763 | 0.654 | 0.488 | 0.685 |
| √ | √ | √ | **0.702** | **0.599** | **0.704** | **5.766** | 17.742 | **6.679** | **0.677** | 0.522 | **0.692** |

improves the model's ability to extract and represent protein sequence features by integrating the linkage attention mechanism with additional enhancement strategies. Compared with traditional MLP, CNN, and GRU models, the GLAM demonstrates significant advantages in processing long sequences and capturing long-distance dependencies. The effectiveness of a GLAM in sequence feature extraction is confirmed.

*Outstanding Performance of GAT Combined with Gep Pooling Demonstrates the Effectiveness of SGAM Feature Extractors.* We propose an SGAM feature extractor that innovatively combines GAT and global average pooling (gep) methods. In order to systematically evaluate the effectiveness of our proposed module, we designed a series of comparison experiments to compare in detail the two graph neural networks, GAT and GCN, and the three graph pooling methods gap (global maximum pooling), gep, and gmp (global summation pooling). The experimental results are detailed in Table 6, where the combination of GAT and gep pooling achieves $F_{max}$ metrics of 0.66, 0.552, and 0.672 on the MF, BP, and CC ontologies, respectively. The results in Figure 7 show that the combination of GAT and gep pooling significantly achieves optimal or optimal results for both metrics. This result indicates that GAT performs better than GCN in learning features from graph structure information to predict protein function. Since the importance of each node in the feature learning process of GAT can be dynamically adjusted, this is in contrast to the fixed assignment of node importance in GCN. Consequently, GAT is more suitable for handling scenarios where the training and test sets differ. As a result, GAT can capture the key information in the graph structure more flexibly, which enhances its feature representation capability. In addition, the gep pooling approach exhibits significantly higher performance in the experiments, attributed to its ability to effectively capture global information on the graph while preserving important local features.

*Results of Ablation Experiments Highlight the Effectiveness of Individual Channels.* We verified the critical role of



https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

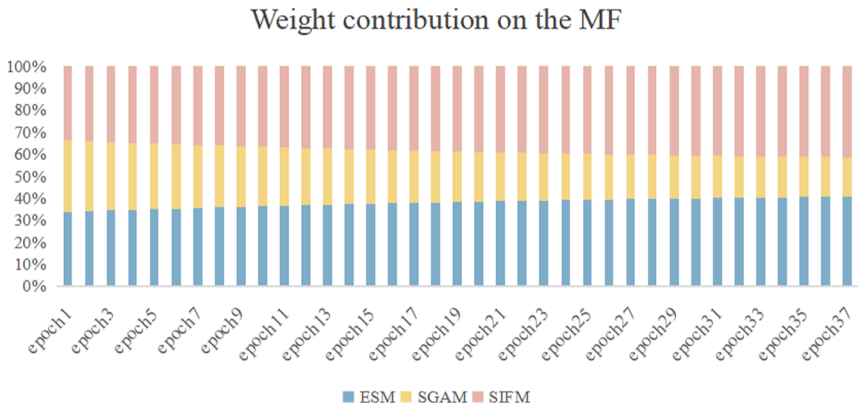



**Figure 8.** 3-Channel weight assignment of SWFM on MF with optimal configuration achieved in the 37th round of training.

multichannel features in protein function prediction by ablation experiments. As shown in Table 7, through the experimental design of blocking different channels one by one, it is observed that removing all channels triggers significant performance decay, which fully proves the effectiveness of fusion in the HPFS framework. Specifically, when the ESM channel is removed, the $F_{max}$ metrics of the model on the three ontologies MF, BP, and CC decrease by 3.1%, 3.8%, and 0.6%, respectively, with the most significant decrease across ontologies, verifying that the channel carries the richest functional discriminative information. This suggests that deep semantic characterization can effectively capture the essential features of protein function. The removal of SGAM resulted in a 2.9% and 2.3% decrease in $F_{max}$ for the MF and BP ontologies, respectively, whereas the performance of the CC ontology remained unchanged ($F_{max}$ = 0.704). This suggests that structural information mainly enhances the prediction of molecular functions and biological processes but has a limited contribution to the localization of cellular components. This may be related to the fact that the CC prediction depends more on subcellular localization features. The removal of SIFM channels caused relatively small decreases in each ontology, indicating that although the sequence-label interaction mechanism enhances prediction through hierarchical information, the magnitude of its gain is limited by the intrinsic characteristics of the structure of the GO. The optimal configuration (ESM:0.407, SGAM:0.179, and SIFM:0.414) achieved by the dynamic weight allocation mechanism during the 37th round of training, as shown in Figure 8, verifies the effectiveness of the SWFM fusion strategy. Among them, the deep semantic and interaction features occupy the dominant weights, while the structural features serve as an important complement, and the three synergize to achieve the optimal effect.

We conducted an in-depth examination of the SIFM's performance. Through meticulously designed ablation experiments, the SIFM applied to the MF ontology yielded impressive outcomes: the $F_{max}$ value reached 0.685, the $S_{min}$ value reached 6.001, and the AUPR value was 0.657. This level of performance significantly outpaced the conventional layer normalization, feature multiplication, and feature concatenation fusion techniques. It resulted in respective improvements of 2.2%, 1%, and 0.3% in the $F_{max}$. These results conclusively validate that our method achieves the interactive fusion of sequences and labels, effectively capturing and leveraging the complementary information between feature sets. This demonstrates the superiority and efficacy of our module.

In addition, for the feature fusion strategy of the three channels, we adopt SWFM and make a careful comparison with the feature-level self-learning weighted fusion, the cross-attention mechanism, and the feature splicing method. The results show that the SWFM improves 5.9%, 4.9%, and 4.2% in terms of the $F_{max}$ value of the MF ontology compared to the other three methods, respectively. By dynamically adjusting the weights of each channel's features, the feature information on each channel can maintain the uniqueness of the original features. Consequently, our module can achieve the optimal effect of feature fusion.

## DISCUSSION

In this study, the DeepMFFGO model is innovatively proposed to address the challenges of multisource data fusion and GO hierarchy utilization in protein function prediction. A fine-tuning strategy based on intermediate layer feature selection was adopted to effectively eliminate redundant features of protein sequences effectively. Meanwhile, GLAM was designed to equip the model to capture sequences and label deep semantic features. In addition, the DeepMFFGO model uniquely proposes HPFS, which effectively integrates four key information sources, sequence information, structural information, GO hierarchy, and PSSM, and realizes complementary and synergistic enhancement of feature information. This enables the model to generate more accurate and reliable function annotation predictions.

In the experimental validation on the CAFA3 data set, the performance of the DeepMFFGO model on the MF, BP, and CC ontologies is outstanding, comprehensively outperforming the existing state-of-the-art models based on multisource protein information and achieving $F_{max}$ values of 0.702, 0.599, and 0.704, respectively. The ablation experiments fully validate the effectiveness of each module in the model and its significant contribution to the overall performance.

Looking ahead, we plan to optimize the model further and use partially resolved protein structures to replace the structures predicted by AlphaFold, thus improving the accuracy of the predictions. We will focus on solving the sequence-structure mismatch problem caused by protein sequence truncation to ensure the accuracy and consistency of the model output. In addition, we propose to model the entire protein structure in an all-encompassing way, replacing the previous contact maps that focused only on the distances of $C_\alpha$ atoms in the 3D structure of proteins. By doing this, we aim to capture the structural features of proteins more


https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853

comprehensively. Regarding multisource feature fusion, we will actively explore including more information sources to enhance the model's reliability. However, the vast data volume and complex model architecture from the multisource fusion algorithm will also demand our computational resources. For this reason, we will consider lightweight prediction algorithms to achieve a balance between high performance and low power consumption. We expect DeepMFFGO to achieve more excellent results in protein function prediction and provide more powerful tools for bioinformatics and functional genomics research.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

The source code and data set for DeepMFFGO are available at 10.5281/zenodo.14676295.

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c00062.

Comparative experimental evaluation metrics for SIFM modules on the MF ontology, comparative experimental evaluation metrics for SWFM modules on the MF ontology, and the allocation of the respective weights of ESM, SGAM, and SIFM on the CC ontology (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**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;* Email: chenjy@xju.edu.cn

### Authors

**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

**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;* orcid.org/0009-0006-4141-4906

**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

**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;* orcid.org/0009-0009-7990-6828

**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*

**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.5c00062


### Author Contributions

J.W.: Software and writing—original draft. J.C.: Conceptualization and writing—review and editing. Y.H.: Conceptualization and project administration. C.S.: Conceptualization and project administration. X.L.: Methodology and validation. Y.Q.: Methodology and validation. L.D.: Methodology and validation.

### Funding

The authors are grateful to the anonymous referees for their insightful suggestions and comments. This research was supported by The Key Research and Development Project in Xinjiang Uygul Autonomous Region (no. 2022B01006, no. 2023B01029, no. 2023B01033, and no. 2023B02034-2), Natural Science Foundation of Xinjiang Uygur Autonomous Region of China (no. 2022D01C692), Basic Research Foundation of Universities in the Xinjiang Uygur Autonomous Region of China (no. XJEDU2023P012), and Tianshan Innovation Team Program of Xinjiang Uygur Autonomous Region of China (no. 2023D14012).

### Notes

The authors declare no competing financial interest.

## ■ REFERENCES


(1) Xuan, P.; Sun, C.; Zhang, T.; Ye, Y.; Shen, T.; Dong, Y. Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. *Front. Genet.* **2019**, *10*, 459.

(2) Sharma, A. K.; Srivastava, R. Protein secondary structure prediction using character bi-gram embedding and Bi-LSTM. *Curr. Bioinf.* **2021**, *16*, 333−338.

(3) Kihara, D. *Kihara Protein Function Prediction*; Springer, 2017.

(4) Bateman, A.; Martin, M. J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; et al. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523−D531.

(5) Pearson, W. R.; Lipman, D. Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 2444−2448.

(6) Ye, J.; McGinnis, S.; Madden, T. L. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* **2006**, *34*, W6−W9.

(7) Buchfink, B.; Xie, C.; Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, *12*, 59−60.

(8) Cruz, L. M.; Trefflich, S.; Weiss, V. A.; Castro, M. A. A. Protein function prediction. *Functional Genomics: Methods and Protocols* **2017**, *1654*, 55−75.

(9) Kulmanov, M.; Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **2020**, *36*, 422−429.

(10) 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*, bbad201.

(11) 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*, 6601.

(12) Lai, B.; Xu, J. Accurate protein function prediction via graph attention networks with predicted structure information. *Briefings Bioinf.* **2022**, *23*, bbab502.

(13) Zhang, L.; Jiang, Y.; Yang, Y. Gnngo3d: Protein function prediction based on 3d structure and functional hierarchy learning. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3867−3878.
(14) 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*, bbae196.
(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) 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.
(17) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*; MIT Press, 2017; Vol. *30*.
(18) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. 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.
(19) Edera, A. A.; Milone, D. H.; Stegmayer, G. Anc2vec: embedding gene ontology terms by preserving ancestors relationships. *Briefings Bioinf.* **2022**, *23*, bbac003.
(20) Varadi, M.; Bertoni, D.; Magana, P.; Paramval, U.; Pidruchna, I.; Radhakrishnan, M.; Tsenkov, M.; Nair, S.; Mirdita, M.; Yeo, J.; et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **2024**, *52*, D368−D375.
(21) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(22) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **1999**, *27*, 49−54.
(23) Zhou, N.; Jiang, Y.; Bergquist, T. R.; Lee, A. J.; Kacsoh, B. Z.; Crocker, A. W.; Lewis, K. A.; Georghiou, G.; Nguyen, H. N.; Hamid, M. N. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **2019**, *20*, 244.
(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(25) Ma, W.; Zhang, S.; Li, Z.; Jiang, M.; Wang, S.; Lu, W.; Bi, X.; Jiang, H.; Zhang, H.; Wei, Z. Enhancing protein function prediction performance by utilizing AlphaFold-predicted protein structures. *J. Chem. Inf. Model.* **2022**, *62*, 4008−4017.
(26) Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **2021**, *12*, 3168.
(27) Zheng, R.; Huang, Z.; Deng, L. Large-scale predicting protein functions through heterogeneous feature fusion. *Briefings Bioinf.* **2023**, *24*, bbad243.


https://doi.org/10.1021/acs.jcim.5c00062
J. Chem. Inf. Model. 2025, 65, 3841−3853