Enhancing Multimodal Protein Function Prediction Through Dual-Branch Dynamic Selection with Reconstructive Pre-Training

Xiaoling Luo[1], **Peng Chen**[2], **Chengliang Liu**[3], **Xiaopeng Jin**[3*], **Jie Wen**[4*], **Yumeng Liu**[4] and **Junsong Wang**[4]

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2]College of Applied Technology, Shenzhen University, Shenzhen, China
[3]Laboratory for Artificial Intelligence in Design, Hong Kong
[4]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China
[5]College of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
xiaolingluoo@outlook.com, 2300411008@emal.szu.edu.cn, liucl1996@163.com, jinxiaopengit@gmail.com, wenjie@hit.edu.cn, liuyumeng@sztu.edu.cn, wangjunsong@sztu.edu.cn

Abstract

Multimodal protein features play a crucial role in protein function prediction. However, these features encompass a wide range of information, ranging from structural data and sequence features to protein attributes and interaction networks, making it challenging to decipher their complex interconnections. In this work, we propose a multimodal protein function prediction method (DSRPGO) by utilizing dynamic selection and reconstructive pre-training mechanisms. To acquire complex protein information, we introduce reconstructive pre-training to mine more fine-grained information with low semantic levels. Moreover, we put forward the Bidirectional Interaction Module (BInM) to facilitate interactive learning among multimodal features. Additionally, to address the difficulty of hierarchical multi-label classification in this task, a Dynamic Selection Module (DSM) is designed to select the feature representation that is most conducive to current protein function prediction. Our proposed DSRPGO model improves significantly in BPO, MFO, and CCO on human datasets, thereby outperforming other benchmark models.

1 Introduction

Protein function prediction has become a key challenge in biology, with the rapid development of bioinformatics [Hasselgren and Oprea, 2024]. The Gene Ontology (GO) framework [Ma *et al.*, 2025] standardizes protein functions into three categories: biological process (BPO), molecular function (MFO), and cellular component (CCO) [Aleksander *et al.*, 2023]. In recent decades, numerous deep learning methods [You *et al.*, 2021; Zhang *et al.*, 2023] have been developed to predict protein functions. However, using single-modal features often faces data limitations [Kulmanov and Hoehndorf, 2020]. Many studies [Fan *et al.*, 2020] have shown that using protein sequence information significantly improves the accuracy of MFO. Still, many proteins share functional similarities but have dissimilar sequences [Lin *et al.*, 2024]. As a result, for proteins with low sequence similarity, the accuracy of predictions may be compromised. Moreover, structure-based methods usually perform better, but the high complexity of protein structures and data acquisition costs limit their application [Paysan-Lafosse *et al.*, 2023]. Furthermore, the noise introduced during the generation of protein-protein interaction (PPI) networks [Wang *et al.*, 2022] through high-throughput techniques poses risks to the accuracy of predictions [Chen and Luo, 2024].

Therefore, integrating these different types of protein data and taking advantage of their complementary advantages in functional prediction is an important way [Zhao *et al.*, 2024] to improve the performance of protein function prediction. These methods mainly adopt two strategies: graph neural networks (GNNs) [You *et al.*, 2021] and autoencoders [Gligorijević *et al.*, 2018; Fan *et al.*, 2020; Pan *et al.*, 2023]. Graph2GO [Fan *et al.*, 2020] integrates sequence similarity and PPI networks using GNNs, treating protein sequences and structures as node features. However, those using GNNs [Zhou *et al.*, 2019] may amplify noise and face issues with over-smoothing. To address these limitations, CFAGO [Wu *et al.*, 2023] introduces Transformer-based fusion within autoencoders to enhance multimodal feature integration.

However, current multimodal approaches mainly fuse information without exploring the potential complementarity between different modalities. To address this issue, we propose a multimodal method for protein function prediction that efficiently mines the complex internal relationships among spatial structure features, such as PPI networks, subcellular locations, and protein domains, as well as sequence features, specifically the amino acid sequence. Furthermore, due to the complexity of protein information, existing models tend to ignore the detailed features inside the information, such as PPI local network topology, connection strength, amino acid frequency distribution, and key sequence fragments. We add a reconstruction pre-training step to obtain more low-semantic and fine-grained features from protein information of multi-

*Co-corresponding authors: Xiaopeng Jin, Jie Wen.

ple modes. By learning these basic features, the model provides a richer representational basis for downstream tasks.

In addition, large language models play an important role in improving protein function prediction. Inspired by large language models, the protein sequence information in our method is extracted using the pre-trained ProtT5 [Elnaggar *et al.*, 2021]. In this work, to better learn multimodal information, our proposed DSRPGO model includes a shared and an interactive learning branch. In the shared learning branch, we concatenate features from different modalities and perform joint analysis in a unified representation space. Moreover, we introduce the Bidirectional Interaction Module (BInM), where each modality both influences and receives information from others, enhancing overall understanding.

Besides, faced with thousands of protein functions, accurately predicting the protein function of a sample remains a challenging issue. Protein function prediction is essentially a complex hierarchical multi-label classification problem. In this situation, we propose the Dynamic Selection Module (DSM) to dynamically select the optimal feature combination for fitting more diverse protein functions. The code and supplementary materials have been open-sourced[1]. Our main contributions can be summarized as follows:

- We propose a multimodal feature-based approach for protein function prediction that overcomes the limitations of single-modality methods, effectively representing protein functional characteristics.
- A reconstructive pre-training phase is designed to make the model capable of learning more low-semantic fine-grained features to assist the model in understanding protein function.
- Our proposed BInM incorporates a bidirectional interaction mechanism to promote efficient fusion and information exchange between sequence and spatial features, enhancing the model's ability to capture strong protein information between different modes.
- We construct the DSM that enables the model to adaptively select channel features most relevant to specific functional labels, resulting in enhanced performance.

2 Methodology

Our proposed method efficiently captures multimodal information about proteins through a strategy for two-step training. In the pre-training stage, we use the encoder-decoder model to learn and inject multimodal knowledge. For spatial features including PPI, subcellular location, and protein domains, a Protein Spatial Structured Information (PSSI) encoder-decoder model using the BiMamba blocks is introduced in this stage. To mine sequence features including protein sequences, we design a Protein Sequence Information (PSeI) encoder-decoder model based on the Transformer blocks for pre-training. Then, during our DSRPGO model training phase, we integrate and learn features from multimodal information. The proposed model is primarily divided into two major branches: one is the multimodal shared learning branch (MSL-Branch), and the other is the multimodal interactive learning branch (MIL-Branch). Protein data are processed through these branches to generate several sets of features, which serve as inputs for DSM. Finally, the model dynamic selects the optimal features for the current protein, to enhance performance in protein function prediction. An illustration of our proposed method can be seen in Figure 1.

[1]https://github.com/kioedru/DSRPGO

2.1 Reconstructive Pre-training

In the reconstructive pre-training stage, to obtain feature extractors that are good at mining fine-grained features from multi-modal protein information, we utilize the PSSI and PSeI encoder-decoder model for feature reconstruction.

PSSI Encoder-Decoder Learning

The PPI network gets an $N \times N$ adjacency matrix by matrix conversion as input to the encoder. Moreover, another input to the encoder is obtained by concatenating the bag-of-words encodings of subcellular location and Protein Domain.

Mamba Preliminaries. Mamba [Gu and Dao, 2023] extends the capabilities of the State-Space Models (SSMs) [Gu *et al.*, 2023] by enabling the transformation of a continuous 1D input $x_t \in \mathbb{R}$ to $y_t \in \mathbb{R}$ via a learnable hidden state $h_t \in \mathbb{R}^{\hat{N}}$ with discrete parameters $\bar{A} \in \mathbb{R}^{\hat{N}\times\hat{N}}$, $\bar{B} \in \mathbb{R}^{1\times\hat{N}}$, and $\bar{C} \in \mathbb{R}^{1\times\hat{N}}$ as follows:

$$\begin{aligned} &h_t = \bar{A}h_{t-1} + \bar{B}x_t,\ y_t = Ch_t + Dh_t, \\ &\bar{A} = e^{\Delta A},\ \bar{B} = (\Delta A)^{-1}(e^{\Delta A} - I)\cdot \Delta B,\ \bar{C} = C. \end{aligned} \tag{1}$$

$\bar{A}$ and $\bar{B}$ are continuous A and B converted to discrete evolution parameters using a timescale parameter Δ. To process discrete-time sequences sampled at intervals of Δ, SSMs can be calculated using the recurrence formula. $\bar{C}$ represents the projection parameters. In addition, the models compute output through a global convolution as follows:

$$\bar{K} = (\bar{C}\bar{B}, \bar{C}\bar{A}\bar{B}, \ldots, \bar{C}\bar{A}^{\hat{N}-1}\bar{B}),\ y = x * \bar{K}, \tag{2}$$

where $\hat{N}$ is the length of x, and $\bar{K}$ is a convolutional kernel.

BiMamba Block. Inspired by the selective scan mechanism in Vision Mamba [Zhu *et al.*, 2024], BiMamba Block introduces a novel bidirectional selective scanning mechanism designed for protein data, capturing both the start and end of spatial structure features for enhanced detail and context. Multi-dimensional features are first converted into one-dimensional vectors. Features x_{sp} from PPI, subcellular location, and protein domains are then passed through BiMamba blocks, interleaved with linear layers and residual operations. As shown in Figure 2, forward (FSScan) and backward selective scans (BSScan) extract bidirectional matrix features via positional transformations and reconstructions. Transformed tokens are scanned using Equation 1 to produce new features, with BiMamba's output $\tilde{x}_{sp}$ expressed as:

$$\begin{cases} \tilde{x}_{sp} = FSSCan(x_{sp}) + FSSCan(Linear(F_\alpha \odot F_\sigma + F_\beta \\ \qquad \odot F_\sigma + F_\sigma)), \\ F_\alpha = FSSCan(BSSCan(SSM(Conv1d \\ \qquad (BSSCan(FSSCan(x_{sp})))))), \\ F_\beta = FSSCan(SSM(Conv1d(FSSCan(x_{sp})))), \\ F_\sigma = SiLU(FSSCan(x_{sp})), \end{cases}$$

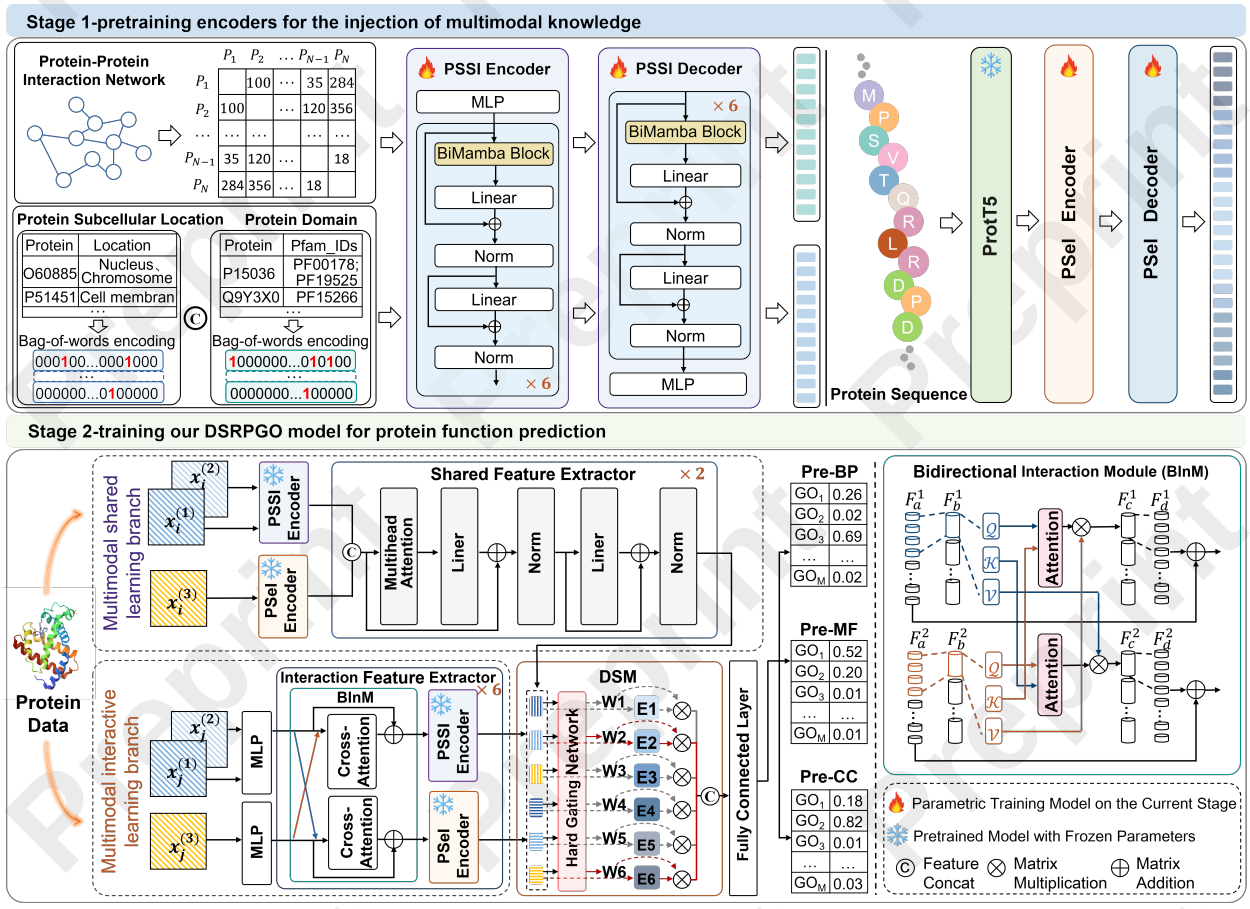

Figure 1: An illustration of our proposed method. This method is mainly divided into two stages. The first stage is to pre-train the Protein Spatial Structure Information (PSSI) encoder and Protein Sequence Information (PSeI) encoder for the injection of multimodal knowledge . The second stage is training our proposed DSRPGO model, which consists of an MSL-Branch, a MIL-Branch with the Bidirectional Interaction Module (BInM), and the Dynamic Selection Module (DSM).

where the operation $\odot$ denotes the Hadamard product.

PSSI Encoder. In this section, we propose a PSSI encoder architecture designed to effectively map high-dimensional input data into a low-dimensional latent space. The PSSI encoder consists of multilayer perceptrons (MLPs), BiMamba block, Linear and Norm layers, which work in concert to extract features from the input data and generate a compact latent representation. Assume that the input feature $x_i^{h(k)} \in \mathrm{R}^{H_i^k}$ is a high-dimensional vector of the i-th protein, where H_i^k represents the feature dimension of the k-th input source. This feature is reconstructed using MLP to output a low-dimensional representation $x_i^{d(k)} \in \mathrm{R}^D$, where D denotes the size of the MLP hidden layer. **PSSI Decoder.** The architecture of the PSSI decoder is a counterpart to that of the encoder. The PSSI decoder rebuilds the given protein spatial structure information based on the hidden representations output by the encoder. This process involves BiMamba computation and residual operations, optimizing the cross-entropy loss function to enhance the performance. After taking the output $x_i^{d(k)}$ of the PSSI encoder and passing through the BiMamba block, alternating Linear and Norm layers, we obtain the recovered high-dimensional features $\bar{x}_i^{h(k)} \in \mathrm{R}^{H_i^k}$. The overarching objective of the encoder-decoder architecture is to minimize the sample wise binary cross-entropy loss between the original and reconstructed source features, thereby enhancing the model's predictive accuracy and fidelity in representing protein data. The loss function of PSSI encoder-decoder is:

$$\mathcal{L}_{sp} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{H_i^k} -\Big[x_{ij}^{h(k)} \log \bar{x}_{ij}^{h(k)} + \left(1 - x_{ij}^{h(k)}\right) \log\left(1 - \bar{x}_{ij}^{h(k)}\right)\Big], \tag{3}$$

where N is the number of total proteins, K is the number of input sources, $x_{ij}^{h(k)}$ and $\bar{x}_{ij}^{h(k)}$ denotes the j-th dimension vector of $x_i^{h(k)}$ and $\bar{x}_i^{h(k)}$.

PSeI Encoder-Decoder Learning

In PSeI encoder-decoder, the transformer block with multi-head self-attention (MSA) mechanism [Dosovitskiy *et al.*, 2021] extracts long-distance features from protein sequences.

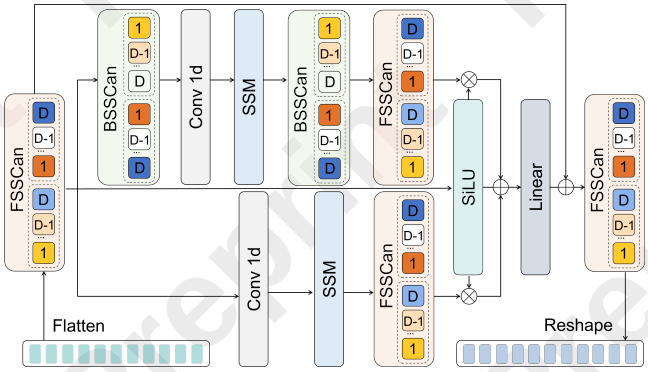

Figure 2: Structure of the BiMamba block.

Then, to further leverage these features, we use the pre-trained ProtT5 [Elnaggar *et al.*, 2021] model to parse the protein sequences. To achieve this, we froze the parameters of ProtT5 and connected it to the PSeI encoder for further pre-taining.

PSeI Encoder. The PSeI encoder consists of an MLP block and 6 self-attention blocks. The self-attention block includes an MSA computation layer, as well as alternating linear and norm layers, connected through a residual structure. Assuming the input of the self-attention block is $\tilde{s}_i^d = MLP(s_i^h)$, the output feature is $\hat{s}_i^d \in \mathrm{R}^D$:

$$\hat{s}_i^d = N(N(\tilde{s}_i^d + L(MSA(\tilde{s}_i^d))) + L(N(\tilde{s}_i^d + L(MSA(\tilde{s}_i^d))))), \quad (4)$$

where $s_i^h \in \mathrm{R}^{H_i}$ is the i-th input sequence feature of encoder, and H_i is the dimension of input feature. $L(x)$ denotes the fuction of Linear layer, and $N(x)$ denotes the Norm layer.

PSeI Decoder. The PSeI decoder takes the hidden states from the encoder as input, which contains compressed information about the input sequence. To obtain the final protein sequence encoding, we designed the PSeI decoder using a combination of 6 self-attention blocks and one MLP block. Then, the output feature of the PSeI decoder is $\hat{s}_i^h \in \mathbb{R}^{H_i}$. Like the PSSI encoder-decoder, the loss function $\mathcal{L}_{se}$ for the PSeI encoder-decoder also adopts the form of cross-entropy:

$$\mathcal{L}_{se} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{H_i} -\left[s_{ij}^h \log \bar{s}_{ij}^h + \left(1 - s_{ij}^h\right)\log\left(1 - \bar{s}_{ij}^h\right)\right], \quad (5)$$

where i denotes the sequence input of the i-th protein, j is the j-th dimension vector of the feature map.

2.2 Bidirectional Interaction and Dynamic Selection for Protein Function Prediction

In this section, we apply the encoders sensitive to low semantic features obtained in the pre-training stage to high semantic tasks. Specifically, to improve the performance of protein function prediction, BInM and DSM modules are proposed to capture deep interaction information between multimodal features and dynamically screen the features most suitable for the current task.

Bidirectional Interaction Module

The proposed BInM enhances the model's ability to learn complex patterns by integrating information across modalities. Using cross-attention, it compares query (Q) vectors with key (K) vectors from the opposite branch, enabling bidirectional interaction. This approach captures interdependencies between branches more effectively, similar to MSA but focused on cross-branch connections.

Therefore, we assume that the features transformed by PPI are represented as $x_i^{(1)}$, and the features obtained from the encoding of subcellular location and protein domains are concatenated to form $x_i^{(2)}$, while the features extracted through the ProtT foundation model for protein sequences are denoted as $x_i^{(3)}$. Subsequently, $x_i^{(1)}$ and $x_i^{(2)}$ get features with the same dimension after the MLP reconstruction features, and their concatenated feature map $\widetilde{x}_i^B$ is used as the input of the first branch of BInM. Similarly, $\overline{x}_i^B$, the input to the second branch of BInM, is derived from $x_i^{(3)}$ after its transformation through the MLP. In BInM, the input embedded patches $F_a^1 \in \mathbb{R}^{L_a \times D_a}$ and $F_a^2 \in \mathbb{R}^{L_a \times D_a}$ are initially and randomly divided into multiple heads vectors $F_b^1 \in \mathbb{R}^{L_a \times D_b \times H_b}$ and $F_b^2 \in \mathbb{R}^{L_a \times D_b \times H_b}$, where H_b is the number of multiple heads.

As shown in Figure 1, F_b^1 and F_b^2 are converted into queries $\mathcal{Q}^1(F_b^1)$ and $\mathcal{Q}^2(F_b^2)$. The key $\mathcal{K}^1$ and value $\mathcal{V}^1$ of F_b^1, and the key $\mathcal{K}^2$ and value $\mathcal{V}^2$ of F_b^2 are obtained using three generators $\mathcal{Q}$, $\mathcal{K}$, and $\mathcal{V}$. Then, $F_c^1 \in \mathbb{R}^{L_a \times D_b \times H_b}$ obtained by cross-attention is defined as:

$$F_c^1 = softmax(\mathcal{Q}^1(F_b^1) \otimes \mathcal{K}^2(F_b^2)^T) \otimes \mathcal{V}^2(F_b^2), \quad (6)$$

where the operation T means matrix transpose, the operation $\otimes$ represents matrix multiplication, and the goal of $softmax$ function is to normalize the F_c^1. Finally, the cross-attention output feature $F_d^1 \in \mathbb{R}^{L_a \times D_a}$ of the first branch is obtained by feature mapping. Similarly, we can get the cross-attention output $F_d^2 \in \mathbb{R}^{L_a \times D_a}$ of the second branch. In this way, the model takes into account not only the meaning of each branch itself but also the relationships with other branch features, resulting in a more complete representation of multimodal data.

Algorithm 1 Dynamic Selection Moudle Procedure

Input: Protein vector X_{dsm} , Threshold t
Output: Fusion feature after DSM

1: Initialize expert weights $W \leftarrow 0_N$.
2: Compute expert confidence coefficients $\hat{p} \leftarrow \text{Softmax}(\text{MLP}(X_{dsm}))$.
3: Select active experts $S \leftarrow \{E_i | \hat{p}_i \geq t\}$.
4: **for** each experts E_i in S **do**
5: Normalize $\hat{p}$ to obtain weights $W_i \leftarrow \frac{\hat{p}_i}{\sum_{E_j \in S} \hat{p}_j}$.
6: **end for**
7: **return** $\text{DSM}(X_{dsm}) \leftarrow \text{Concat}(W_i \cdot E_i(X_{dsm}))$

Dynamic Selection Module

In the final feature selection stage, we introduce DSM to enhance key features and mitigate the impact of conflicting ones. As illustrated in Algorithm 1 and Figure 1, this module employs an improved Mixture-of-Experts (MoE) strategy based on Masoudnia et al [Masoudnia and Ebrahimpour, 2014]. The MSL-Branch and MIL-Branch each output a single vector with three channels, where the three channels rep-

resent PPI, sequence, and subcellular localization combined with domain features, respectively. All six-channel feature maps serve as the input $X_{dsm} = (x^1_{dsm}, x^2_{dsm}, \cdots, x^V_{dsm})$ for the DSM. The function of DSM is:

$$\text{DSM}(X_{dsm}) = \text{Concat}(\frac{\hat{p}_i}{\sum_{E_j \in S} \hat{p}_j} \cdot E_i(X_{dsm})), \quad (7)$$

where E_j is the experts belonging to the selected expert group S, $\hat{p}_i$ denotes the confidence coefficient of expert E_i.

Loss Functions

In this work, protein function prediction is modeled as the multi-label classification task. The predictor, constructed from fully connected layers, takes the output features of the DSM as input and produces an M-dimensional score vector of GO terms: $P_i = (p^1_i, p^2_i, \cdots, p^M_i))$. In the context of protein function prediction using GO terms, there are significantly more negative proteins than positive ones in the training set. Consequently, we employ an asymmetric loss [Wu *et al.*, 2023] as the prediction loss $\mathcal{L}$.

$$\begin{aligned}\mathcal{L} = &\frac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} -y^m_i \left(1 - p^m_i\right)^{y+} \log\left(p^m_i\right) \\ &- \left(1 - y^m_i\right) \left(p^m_i\right)^{y-} \log\left(1 - p^m_i\right), \end{aligned} \quad (8)$$

where y^m_i represents the ground truth label for the i-th protein, while p^m_i denotes the predicted score. The symbols $\{y+\}$ and $\{y-\}$ refer to the positive and negative focusing parameters respectively.

3 Experiments

In this section, we present the experimental setup, including the datasets, baseline models, training details, and evaluation metrics. Then we provide an analysis of the experimental results, supported by ablation studies and Davies-Bouldin scores to validate the effectiveness of the model.

Further experiments on the model components, structures, and parameters can be seen in Appendix Sections 1, 2, and 5.

3.1 Experimental Setup

Dataset Settings. We construct our dataset based on CFAGO [Wu *et al.*, 2023]. PPI data comes from the STRING [Szklarczyk *et al.*, 2023] database (v11.5), and protein sequences, subcellular localization, and domain data are from the UniProt [Consortium, 2022] database (v3.5.175). A total of 19,385 proteins are used for pretraining. For fine-tuning, we collect protein function annotations from the Gene Ontology [Aleksander *et al.*, 2023] database (v2022-01-13). The fine-tuning datasets for each GO branch, split by two-time points, including BPO: 3,197 training, 304 validation, 182 testing proteins (45 GO terms), MFO: 2,747 training, 503 validation, 719 testing proteins (38 GO terms), and CCO: 5,263 training, 577 validation, 119 testing proteins (35 GO terms).

More details about sequence similarity and model performance are in Appendix Sections 3 and 6.

Implementation Details. We conduct all experiments on NVIDIA GTX 4090. We set the dropout rate to 0.1 during pre-training, and the model trains for 5000 epochs, with a learning rate of 1e-5 for the first 2500 epochs and 1e-6 for the remaining 2500 epochs. During fine-tuning, we use a dropout rate of 0.3 and train for 100 epochs with the AdamW optimizer. The learning rate is set to 1e-3 for the first 50 epochs and reduced to 1e-4 for the remaining 50 epochs.

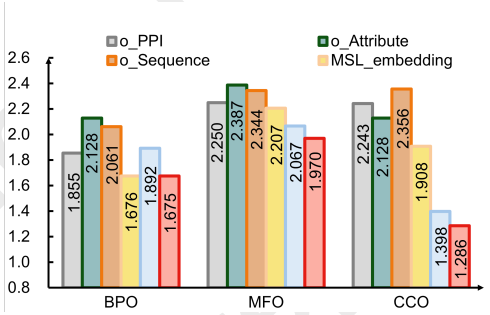

Figure 3: Davies Bouldin Score comparison of different protein features represents. o_PPI, o_Attribute, and o_Sequence represent the original embedding of PPI, subcellular localization combined with domain, and protein language model, respectively. MSL_embedding, MSI_embedding, and DSM_embedding represent the embedding from MSL-Branch, MIL-Branch, and DSM, respectively.

Compared Methods. We compare DSRPGO with nine methods, which are categorized into two groups based on their data utilization strategies. Unimodal-based methods: Naive [Radivojac *et al.*, 2013], BLAST[Altschul *et al.*, 1990], GeneMANIA[Mostafavi *et al.*, 2008], Mashup[Cho *et al.*, 2016], and deepNF[Gligorijević *et al.*, 2018]. Multimodal-based methods: Graph2GO[Fan *et al.*, 2020], NetQuilt[Barot *et al.*, 2021], DeepGraphGO[You *et al.*, 2021], and CFAGO[Wu *et al.*, 2023].

Evaluation Metrics. In this study, we evaluate predictive performance using five metrics: micro-averaged AUPR (m-AUPR) and macro-averaged AUPR (M-AUPR) [Peng *et al.*, 2021], F1-score (F1) [Wu *et al.*, 2023], accuracy (ACC), and F-max score (F_{max})[Lin *et al.*, 2024], providing a comprehensive assessment of model accuracy and effectiveness.

3.2 Comparison with Unimodal-based and Multimodal-based Methods

Comparision with Unimodal-based Methods. Most of the previous methods are based on unimodal protein features, so to verify the performance of our multimodal-based method, we compare our method with unimodal-based methods. The experimental results are shown in Table 1. DSRPGO significantly outperforms unimodal-based methods across various metrics, except for M-AUPR in MFO. Compared to unimodal methods, DSRPGO improves Fmax by at least 6.4% in BPO, 7.7% in MFO, and 15.5% in CCO. This demonstrates the advantage of integrating multimodal data for protein function prediction.

Comparision with Multimodal-based Methods. To better evaluate our method, we also compare DSRPGO with other state-of-the-art multimodal-based methods, including CFAGO, DeepGraphGO, Graph2GO, and NetQuilt. The detailed results in Table 1 show that DSRPGO generally out-

Method		Naïve†	BLAST†	GeneMANIA†	Mashup†	deepNF†	NetQuilt	Graph2GO	DeepGraphGO	CFAGO	DSRPGO (Ours)
$\mathbf{F_{max}}$	BPO	0.051±0	0.270±0	0.000±0	0.075±0	0.394±0.006	0.164±0.014	0.335±0.010	0.327±0.028	0.439±0.007	**0.458±0.006**
	MFO	0.177±0	0.122±0	0.000±0	0.058±0	0.153±0.004	0.081±0.013	0.196±0.006	0.142±0.035	0.236±0.004	**0.254±0.022**
	CCO	0.121±0	0.196±0	0.031±0	0.000±0	0.297±0.009	0.138±0.013	0.298±0.011	0.209±0.023	0.366±0.018	**0.452±0.019**
m-AUPR	BPO	0.024±0	0.110±0	0.042±0	0.238±0	0.303±0.006	0.077±0.006	0.237±0.014	0.210±0.022	0.328±0.005	**0.330±0.006**
	MFO	0.050±0	0.044±0	0.050±0	0.053±0	0.089±0.001	0.045±0.007	0.103±0.007	0.080±0.021	0.159±0.003	**0.166±0.027**
	CCO	0.047±0	0.084±0	0.103±0	0.179±0	0.178±0.005	0.081±0.003	0.215±0.025	0.133±0.011	0.337±0.005	**0.371±0.035**
M-AUPR	BPO	0.048±0	0.093±0	0.160±0	0.146±0	0.174±0.005	0.081±0.004	0.150±0.006	0.133±0.008	**0.188±0.003**	0.182±0.003
	MFO	0.029±0	0.084±0	0.109±0	0.089±0	0.118±0.004	0.064±0.003	0.111±0.005	0.098±0.007	**0.138±0.005**	0.114±0.009
	CCO	0.060±0	0.082±0	0.150±0	0.104±0	0.155±0.009	0.063±0.004	0.159±0.021	0.133±0.006	0.210±0.007	**0.239±0.025**
F1	BPO	0.035±0	0.159±0	0.054±0	0.248±0	0.228±0.005	0.114±0.017	0.222±0.010	0.238±0.012	**0.283±0.006**	0.272±0.008
	MFO	0.004±0	0.064±0	0.008±0	0.106±0	0.117±0.004	0.070±0.016	0.167±0.009	0.165±0.056	0.234±0.005	**0.241±0.019**
	CCO	0.070±0	0.107±0	0.123±0	0.202±0	0.205±0.009	0.108±0.013	0.261±0.015	0.210±0.016	0.314±0.007	**0.357±0.033**
ACC	BPO	0.000±0	0.071±0	0.000±0	0.044±0	0.158±0.011	0.048±0.007	0.257±0.007	0.153±0.034	0.338±0.013	**0.346±0.016**
	MFO	0.000±0	0.015±0	0.000±0	0.038±0	0.034±0.002	0.017±0.002	0.114±0.015	0.048±0.007	0.100±0.003	**0.124±0.037**
	CCO	0.000±0	0.034±0	0.000±0	0.000±0	0.080±0.012	0.037±0.005	0.180±0.024	0.066±0.011	0.210±0.008	**0.262±0.017**

Table 1: Comparison results of different methods. Unimodal-based methods are marked with "†", while the rest are multimodal-based methods. The best results are highlighted in bold, and the sub-optimal results are underlined. After the ± is the standard deviation of the experimental results.

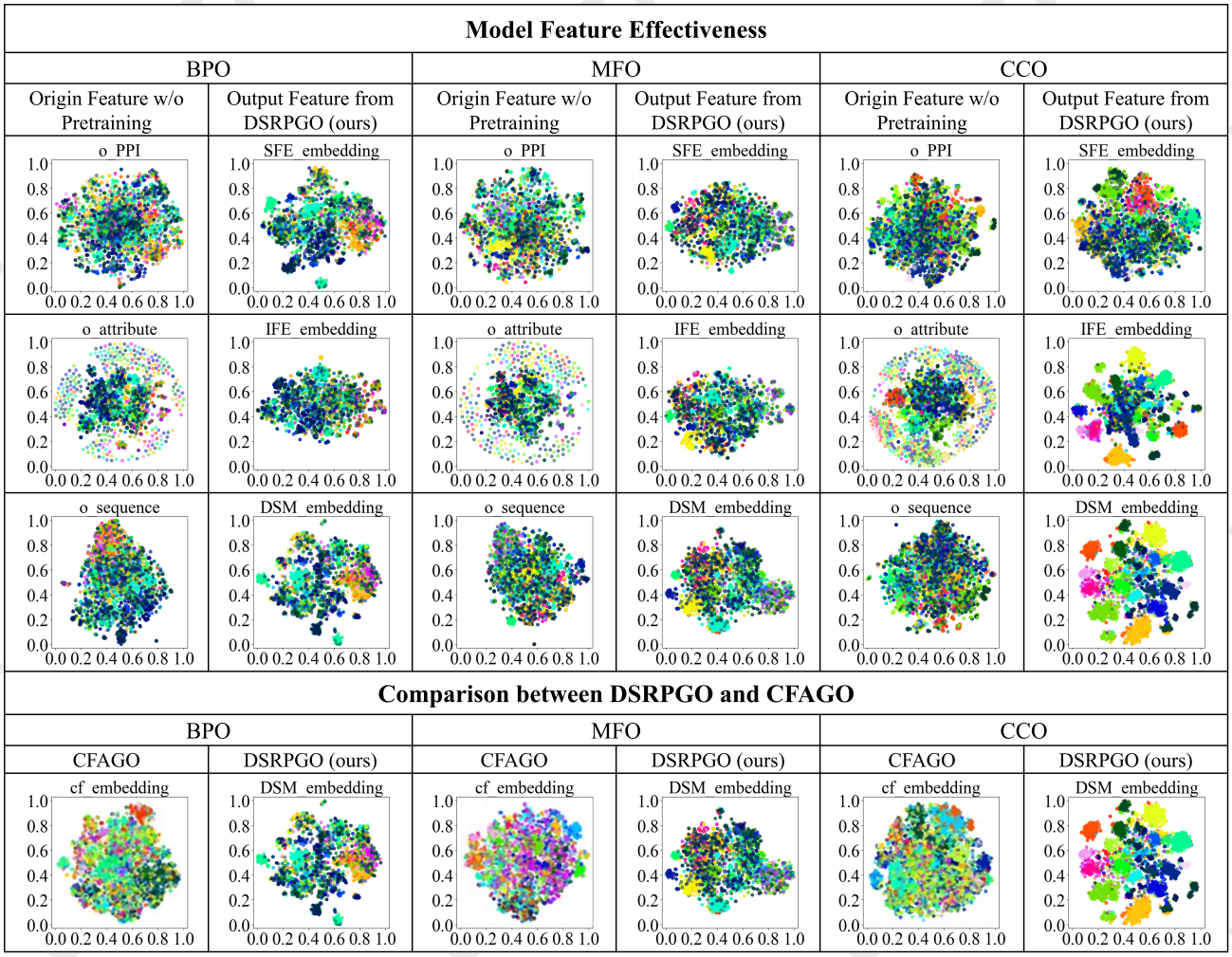

Figure 4: Visualization of different feature representations for DSRPGO, and comparison with CFAGO.

Method	F_{max}			m-AUPR			M-AUPR			F1			ACC		
	BPO	MFO	CCO	BPO	MFO	CCO	BPO	MFO	CCO	BPO	MFO	CCO	BPO	MFO	CCO
MSLB	0.437	0.179	0.371	0.315	0.108	0.304	0.173	0.102	0.197	0.261	0.172	0.311	0.292	0.076	0.190
MILB	0.310	0.179	0.420	0.180	0.091	0.330	0.138	0.113	0.220	0.236	0.162	0.342	0.216	0.090	0.220
MSLB+MILB	**0.458**	**0.254**	**0.452**	**0.330**	**0.166**	**0.371**	**0.182**	**0.114**	**0.239**	**0.272**	**0.241**	**0.357**	**0.346**	0.124	**0.262**
w/o BInM	0.435	0.193	0.333	0.313	0.116	0.266	0.174	0.106	0.186	0.265	0.180	0.305	0.301	0.088	0.151
w/o DSM	0.397	0.190	0.378	0.275	0.105	0.302	0.163	0.113	0.205	0.265	0.173	0.328	0.315	0.092	0.190
w/o SP-F	0.216	0.173	0.263	0.106	0.059	0.164	0.105	0.039	0.115	0.174	0.004	0.226	0.151	0.000	0.145
w/o SE-F	0.251	0.238	0.363	0.119	0.117	0.219	0.115	0.099	0.181	0.179	0.224	0.322	0.170	**0.133**	0.193
w/o pretrain	0.297	0.167	0.356	0.196	0.093	0.284	0.129	0.095	0.200	0.205	0.162	0.286	0.200	0.085	0.197

Table 2: Results of Ablation Studies. The overall model is denoted as 'MSLB+MILB', where 'MSLB' and 'MILB' are the backbone components: MSL-Branch and MIL-Branch. w/o BInM and w/o DSM represent removing the BInM and DSM modules from the overall model. w/o SP-F refers to removing spatial structure features from the input, while w/o SE-F indicates removing sequence features. The best results are marked in bold.

performs these methods. Compared to multimodal methods, DSRPGO improves the Fmax metric by at least 1.9% in BPO, 1.8% in MFO, and 8.6% in CCO. This indicates that DSRPGO's architecture is more effective in learning deep representations among multimodal features, thereby further enhancing overall performance. At the same time, we observe that DSRPGO does not perform optimally in M-AUPR. This is because M-AUPR evaluates each class equally, including those with fewer samples, which may not reflect the model's overall performance. In contrast, m-AUPR aggregates performance across all classes, offering a more comprehensive measure of predictive capability. In addition, we discuss the Structure-based and PLM-based comparison methods, as detailed in Appendix Section 4.

3.3 Feature Effectiveness Analysis

To further evaluate the distinguishing power of the multimodal features extracted by different components of DSRPGO, we use Davies-Bouldin (DB) [Wu *et al.*, 2023] scores. In the calculation of DB scores, GO terms are set as the labels for protein clusters, meaning proteins sharing the same GO term set are grouped into the same cluster. A lower DB score indicates more compact clusters and clearer separation. As shown in Figure 3, DSRPGO components effectively capture multimodal features. Among them, DSM_embedding performs best, indicating that DSM successfully integrates inputs from the MIL and MSL branches.

To further analyze the discriminative power of protein features, we visualize them using t-SNE [Chatzimparmpas *et al.*, 2020], as shown in Figure 4. Raw input features (o_PPI, o_Attribute, o_Sequence), which are not pre-trained, show distinct patterns but lack clear clustering boundaries. In contrast, the output of the feature by various modules of DSRPGO achieves better clustering results. Additionally, compared to the output of the feature by CFAGO (cf_embedding), DSRPGO demonstrates significantly superior performance.

4 Ablation Studies

In this section, the contributions of each component in DSRPGO are evaluated, as shown in Table 2.

Analysis for Backbone Components. According to lines 1,2, and 3 of Table 2, the results of the backbone network only using MSL-Branch or MIL-Branch are not as good as those using combined branches.

Effectiveness of BInM. Considering the correlation of features among space and sequence, this method uses the BInM block to facilitate bidirectional multimodal feature interaction before DSM. As shown in rows 3 and 4 of Table 2, we verify the validity of BInM for the overall model by removing it.

Effectiveness of DSM. To enable effective feature selection and accurate prediction of protein functions, DSM is used to select channel features most relevant to specific functional labels adaptively. At the same time, it reduces the interference and conflict caused by redundant features. As shown in rows 3 and 5 of Table 2, DSM has a positive impact on protein function prediction.

Impact of Sequence and Spatial Structure Features. To verify the complementarity between sequence and spatial structure features, we perform an ablation study, retaining only spatial structure or sequence features. For the BInM module, it is removed as no interaction occurs with a single feature type. Rows 6 and 7 of Table 2 show that removing feature interaction significantly reduces model performance.

Impact of Pre-training. To evaluate the contribution of pre-training, we conduct an ablation study by removing it. As shown in the last row of Table 2, the model's performance drops significantly across all metrics without pre-training.

5 Conclusion

This paper proposes a dual-branched multimodal method for protein function prediction with reconstructive pre-training. The proposed method enhances the model's ability to integrate multimodal features through two key components: the BInM and the DSM, leading to significant performance gains. Experimental results show that the DSRPGO outperforms current state-of-the-art unimodal and multimodal methods across multiple metrics. These results underscore the importance of integrating multimodal data to enhance protein function prediction, and validate the superiority of the BInM and the DSM in multimodal protein data integration.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62302317, the Natural Science Foundation of Guangdong Province under Grant 2025A1515010184, the project of Shenzhen Science and Technology Innovation Committee under Grant JCYJ20240813141424032 and JCYJ20240813112420027, and the Foundation for Young innovative talents in ordinary universities of Guangdong under Grant 2024KQNCX042, the Stable Support Projects for Shenzhen Higher Education Institutions under grant 20231122005530001 and 20220715183602001, and Guangdong Basic and Applied Basic Research Foundation grant 2024A1515220079.

Contribution Statement

Xiaoling Luo and Peng Chen contributed equally to this work.

References

[Aleksander *et al.*, 2023] Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

[Altschul *et al.*, 1990] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[Barot *et al.*, 2021] Meet Barot, Vladimir Gligorijević, Kyunghyun Cho, and Richard Bonneau. Netquilt: deep multispecies network-based protein function prediction using homology-informed network similarity. *Bioinformatics*, 37(16):2414–2422, 2021.

[Chatzimparmpas *et al.*, 2020] Angelos Chatzimparmpas, Rafael M Martins, and Andreas Kerren. t-visne: Interactive assessment and interpretation of t-sne projections. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2696–2714, 2020.

[Chen and Luo, 2024] Zhuoyang Chen and Qiong Luo. Dualnetgo: a dual network model for protein function prediction via effective feature selection. *Bioinformatics*, 40(7), 2024.

[Cho *et al.*, 2016] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Compact integration of multi-network topology for functional analysis of genes. *Cell Systems*, 3(6):540–548, 2016.

[Consortium, 2022] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2022.

[Dosovitskiy *et al.*, 2021] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S Gelly, J. Uszkoreit, and N Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Elnaggar *et al.*, 2021] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[Fan *et al.*, 2020] Kunjie Fan, Yuanfang Guan, and Yan Zhang. Graph2go: a multi-modal attributed network embedding method for inferring protein functions. *GigaScience*, 9(8):giaa081, 2020.

[Gligorijević *et al.*, 2018] Vladimir Gligorijević, Meet Barot, and Richard Bonneau. deepnf: deep network fusion for protein function prediction. *Bioinformatics*, 34(22):3873–3881, 2018.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Gu *et al.*, 2023] Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. How to train your hippo: State space models with generalized orthogonal basis projections. In *Proceedings of the International Conference on Learning Representations*, 2023.

[Hasselgren and Oprea, 2024] Catrin Hasselgren and Tudor I Oprea. Artificial intelligence for drug discovery: Are we there yet? *Annual Review of Pharmacology and Toxicology*, 64(1):527–550, 2024.

[Kulmanov and Hoehndorf, 2020] Maxat Kulmanov and Robert Hoehndorf. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.

[Lin *et al.*, 2024] Baohui Lin, Xiaoling Luo, Yumeng Liu, and Xiaopeng Jin. A comprehensive review and comparison of existing computational methods for protein function prediction. *Briefings in Bioinformatics*, 25(4):bbae289, 2024.

[Ma *et al.*, 2025] Ruixin Ma, Longfei Wang, Huinan Wu, Buyun Gao, Xiaoru Wang, and Liang Zhao. Historical trends and normalizing flow for one-shot temporal knowledge graph reasoning. *Expert Systems with Applications*, 260:125366, 2025.

[Masoudnia and Ebrahimpour, 2014] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293, 2014.

[Mostafavi *et al.*, 2008] Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9:1–15, 2008.

[Pan *et al.*, 2023] Tong Pan, Chen Li, Yue Bi, Zhikang Wang, Robin B Gasser, Anthony W Purcell, Tatsuya Akutsu, Geoffrey I Webb, Seiya Imoto, and Jiangning

Song. Pfresgo: an attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships. *Bioinformatics*, 39(3):btad094, 2023.

[Paysan-Lafosse *et al.*, 2023] Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic Acids Research*, 51(D1):D418–D427, 2023.

[Peng *et al.*, 2021] Jiajie Peng, Hansheng Xue, Zhongyu Wei, Idil Tuncali, Jianye Hao, and Xuequn Shang. Integrating multi-network topology for gene function prediction using deep neural networks. *Briefings in bioinformatics*, 22(2):2096–2105, 2021.

[Radivojac *et al.*, 2013] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013.

[Szklarczyk *et al.*, 2023] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.

[Wang *et al.*, 2022] Jun Wang, Long Zhang, An Zeng, Dawen Xia, Jiantao Yu, and Guoxian Yu. Deepiii: Predicting isoform-isoform interactions by deep neural networks and data fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4):2177–2187, 2022.

[Wu *et al.*, 2023] Zhourun Wu, Mingyue Guo, Xiaopeng Jin, Junjie Chen, and Bin Liu. Cfago: cross-fusion of network and attributes based on attention mechanism for protein function prediction. *Bioinformatics*, 39(3):btad123, 2023.

[You *et al.*, 2021] Ronghui You, Shuwei Yao, Hiroshi Mamitsuka, and Shanfeng Zhu. Deepgraphgo: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics*, 37(Supplement_1):i262–i271, 2021.

[Zhang *et al.*, 2023] Xiaoshuai Zhang, Huannan Guo, Fan Zhang, Xuan Wang, Kaitao Wu, Shizheng Qiu, Bo Liu, Yadong Wang, Yang Hu, and Junyi Li. Hnetgo: protein function prediction via heterogeneous network transformer. *Briefings in Bioinformatics*, 24(6):bbab556, 2023.

[Zhao *et al.*, 2024] Liang Zhao, Jian Zhang, Bo Xu, Yi Yang, Yangqianhui Zhang, and Ruixin Ma. Multimodal contrastive learning with neuroimaging and cognitive tests for alzheimer's disease diagnosis. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, volume 1, pages 2971–2976, 2024.

[Zhou *et al.*, 2019] Guangjie Zhou, Jun Wang, Xiangliang Zhang, and Guoxian Yu. Deepgoa: Predicting gene ontology annotations of proteins via graph convolutional network. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, volume 1, pages 1836–1841, 2019.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.