Expert Systems With Applications 278 (2025) 127246

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Aligning sequence and structure representations leveraging protein domains for function prediction

Mingqing Wang [a,c], Zhiwei Nie [b,c], Yonghong He [a], Athanasios V. Vasilakos [d], Zhixiang Ren [c,*]

[a] Shenzhen International Graduate School, Tsinghua University, China
[b] School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[c] Pengcheng Laboratory, Shenzhen, China
[d] CAIR, University of Agder, Norway

ARTICLE INFO

Dataset link: https://github.com/AI-HPC-Research-Team/ProtFAD

Keywords:
Protein function prediction
Protein domain
Deep learning
Functional priors
Contrastive learning

ABSTRACT

Protein function prediction is traditionally approached through sequence or structural modeling, often neglecting the effective fusion of diverse data sources. Protein domains, as functionally independent building blocks, determine a protein's biological function, yet their potential has not been fully exploited in function prediction tasks. To address this, we introduce a modality-fused neural network leveraging function-aware domain embeddings as a bridge. We pre-train these embeddings by aligning domain semantics with Gene Ontology (GO) terms and textual descriptions. Additionally, we partition proteins into sub-views based on continuous domain regions for contrastive learning, supervised by a novel triplet InfoNCE loss. Our method outperforms state-of-the-art approaches across various benchmarks, and clearly differentiates proteins carrying distinct functions compared to the competitor.

1. Introduction

Proteins play a pivotal role in the biological processes of living organisms, contributing to cell structure, functionality, signal transduction, and enzymatic reactions (Benkovic & Hammes-Schiffer, 2003; Karplus & Kuriyan, 2005; Pawson & Nash, 2000). With the development of deep neural networks, remarkable breakthroughs have been achieved in the research of proteins, including in protein–ligand binding (Corso, Stärk, Jing, Barzilay, & Jaakkola, 2023), variant effect prediction (Cheng et al., 2023; Meier et al., 2021), *de novo* protein design (Watson et al., 2023), etc. Despite these advancements, there remains a substantial gap in our understanding of proteins, particularly in deciphering the intricate relationships between protein sequence, structure, and function. Protein function prediction has emerged as a focal point in addressing this gap, aiming to identify the specific roles and activities of proteins within biological systems (Notin, Rollins, Gal, Sander, & Marks, 2024; Yan et al., 2023).

Current computational approaches for protein function prediction frequently rely on sequence or structure data (Elnaggar et al., 2021; Fan, Wang, Yang, & Kankanhalli, 2022; Jing, Eismann, Suriana, Townshend, & Dror, 2021). Protein sequences and structures provide valuable information about the composition and spatial distribution of amino acids within proteins. Extensive research (Gligorijević, Renfrew, Kosciolek, Leman, Berenberg, Vatanen, Chandler, Taylor, Fisk, Vlamakis, et al., 2021; Gu, Luo, Chen, Deng, & Lai, 2023; Zhang et al., 2023) has demonstrated that integrating these two data modalities enhances protein representations. However, existing methods typically integrate these modalities using either serial or parallel network architectures, which fail to achieve fine-grained alignment between the different modalities. This limitation poses challenges to the generalizability and interpretability of current approaches for protein function prediction.

Protein domains are distinct structural and functional units within a protein that can exist and function independently. These domains often dictate specific protein functions, such as molecular binding or catalyzing chemical reactions, making them a function-oriented implicit modality. Recent studies have highlighted the critical role of protein domains in protein representations, demonstrating their ability to enhance predictions of protein functions and behaviors (Ibtehaz, Kagaya, & Kihara, 2023; Yao et al., 2021). Existing approaches (Cai, Wang, & Deng, 2020; Fan, Guan, & Zhang, 2020; Torres, Yang, Romero, & Paccanaro, 2021; Wang, Shuai, Zeng, Fan, & Li, 2025; You et al., 2019, 2018) incorporate protein domains as complementary sources of functional information within ensemble frameworks. However, these

* Corresponding author.
E-mail addresses: wmq23@mails.tsinghua.edu.cn (M. Wang), zhiweiNie@pku.edu.cn (Z. Nie), heyh@sz.tsinghua.edu.cn (Y. He), Thanos.vasilakos@uia.no (A.V. Vasilakos), renzhx@pcl.ac.cn (Z. Ren).

https://doi.org/10.1016/j.eswa.2025.127246
Received 27 December 2024; Received in revised form 24 February 2025; Accepted 10 March 2025
Available online 25 March 2025
0957-4174/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

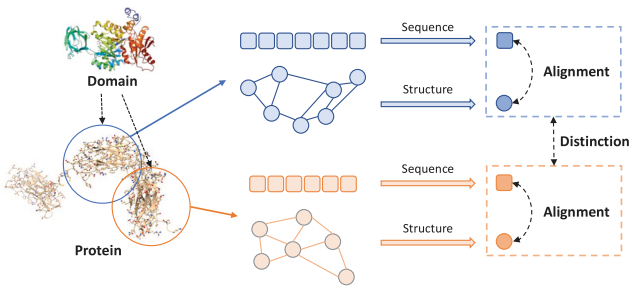

Fig. 1. The motivation and methodology of our method. Our approach leverages protein domains to facilitate fine-grained alignment between sequence and structural data, enabling more effective fusion and enhanced representation.

methods primarily treat domains as auxiliary information sources, without conducting an in-depth exploration of their relationships with protein sequences and structures.

To overcome the limitations outlined above, we propose fine-grained modalities alignment between sequences and structures by leveraging domain embeddings that incorporate functional priors, as illustrated in Fig. 1. Specifically, we partition the protein into diverse sub-views based on adjacent domains. We train the multi-modal features using the proposed domain-joint contrastive learning strategy with a novel triplet InfoNCE loss, to capture segment-level functional information. Finally, we integrate the enhanced multi-modal features and develop a comprehensive network for protein function prediction.

Given the sparsity of domain annotations in protein data, we pre-train domain embeddings to integrate functional information, thereby enhancing the generalizability of domain features. Specifically, we construct a large-scale domain knowledge dataset sourced from the UniProt (The UniProt Consortium, 2022) and InterPro (Paysan-Lafosse, Blum, Chuguransky, Grego, Pinto, Salazar, Bileschi, Bork, Bridge, Colwell, Gough, Haft, Letunić, Marchler-Bauer, Mi, Natale, Orengo, Pandurangan, Rivoire, Sigrist, Sillitoe, Thanki, Thomas, Tosatto, Wu, & Bateman, 2022) databases. This dataset consists of domain entries, corresponding textual descriptions, and associated Gene Ontology (GO) terms. By training domain vocabularies with constructed pseudo-labels and a semantically consistent loss, we derive function-aware domain (FAD) embeddings, which are subsequently employed in the protein function prediction network.

2. Related work

In this section, we review previous studies about sequence-based, structure-based, and multi-modal protein representation learning respectively. Related studies of the protein domain are also summarized and introduced. We discuss the strengths and limitations of each approach and highlight directions for our research below.

Protein representation learning. Protein representation learning plays a crucial role in protein research, such as protein function prediction (Gligorijević et al., 2021; Gu et al., 2023), protein–protein interaction prediction (Kang, Wang, Xie, Zhang, & Xie, 2023), and drug discovery (Pan, Xia, Xu, & Li, 2023; Wu et al., 2024; Zhang, Ouyang, Liu, Liao & and Gao, 2023). Several approaches have been developed to learn protein representations, leveraging different aspects of protein information. Inspired by natural language processing techniques, protein language models (PLMs) (Elnaggar et al., 2021; Lin et al., 2022; Rives et al., 2021) learn to generate meaningful embeddings that encapsulate the hierarchical structure and evolutionary relationships of proteins by training on large-scale protein sequence datasets. Rao et al. (2021), Su et al. (2023) integrate additional information (e.g. "structure-aware vocabulary" or family information) to improve the performance of PLMs. The structure of a protein directly determines its function. Therefore, more and more approaches focus on training protein representations using structural data. Fan et al. (2022), Hermosilla et al. (2021), Jing et al. (2021), Wang, Liu, Liu, Kurtin, and Ji (2023) introduces novel network operators to perform both geometric and relational reasoning on efficient representations of macromolecules. Chen, Zhou, Wang, Liu, and Dou (2023), Hermosilla and Ropinski (2022), Zhang et al. (2023, 2024) employ self-supervised learning methods to effectively capture structural information of proteins and learn meaningful protein representations. Recent efforts have explored the integration of multiple modalities of protein data to create more comprehensive representations. Gligorijević et al. (2021), Gu et al. (2023), Wang et al. (2022), Zhang, Wang, et al. (2023) introduce a joint protein representation for predicting protein functions by integrating the PLMs with graph-network-based structure encoders. In addition to deep mining of sequence and structural information, protein representation learning can also benefit from multi-modal approaches that integrate information from different sources, such as protein surface information (Lee, Yu, Lee, & Kim, 2023), gene ontology annotation (Hu et al., 2023, 2024), 3D point clouds (Nguyen & Hy, 2023), sequence homology (You et al., 2019; Zhou et al., 2022) and protein–protein interaction (Liu, Zhang, & Freddolino, 2024; Wang et al., 2023; Zhang, Zheng, Freddolino, & Zhang, 2018). By combining sequence, structure, and other data, these multi-modal representations offer a holistic view of protein characteristics, enabling more accurate predictions and deeper insights into protein functionality.

Protein domains. Protein domains are structural units within proteins that play crucial roles in determining protein function. They can act alone or in concert with other domains to carry out the biological functions of the protein. Studying protein domains helps scientists better understand the function and structure of proteins. Cai et al. (2020), Fan et al. (2020), You, Yao, Mamitsuka, and Zhu (2021) integrate protein domain information into multi-modal features and achieve accurate predictions of GO terms. Li et al. (2024), Torres et al. (2021), Wang et al. (2025), You et al. (2018) employ a computational framework to fuse multi-source data features, including domains. However, these methods only use the category label of the domain without considering its correlation with protein function. Ibtehaz et al. (2023), Melidis and Nejdl (2021), Rojano et al. (2022) learn associations between protein domains and functions combined at the protein level to derive functionally consistent representations for domains. Forslund and Sonnhammer (2008), Messih, Chitale, Bajic, Kihara, and Gao (2012) develop new methods to infer protein functions based on protein domain combinations and domain order. Inspired by these methods, we propose to integrate domain information, including their functions and combinations, into a multimodal representation.

3. Preliminary

3.1. Problem formulation

Protein function prediction involves determining the biological role of a protein based on its sequence, structure, and interactions with other molecules. This process is essential for understanding cellular processes and elucidating disease mechanisms. In this study, we focus on a set of well-established protein function annotation benchmarks, all of which are consistently defined by maximizing the likelihood:

$$\max_{\theta} P\left(y|x_{seq}, x_{str}, x_{dom}; \theta\right) \tag{1}$$

where $x_{seq}, x_{str}, x_{dom}$ represent the protein's sequence, structure, and domains, respectively, and y denotes the labels of the protein functions. The parameter θ corresponds to the parameters of the function prediction network.

It is important to note that the task may involve either single-label or multi-label classification, meaning that the label y can be either one-dimensional (for single-label) or n-dimensional (for multi-label). Therefore, the ultimate objective is to minimize the classification loss:

$$\min_{\theta} \mathcal{L}_{cla} = \ell\left(\mathcal{N}_{\theta}(x_{seq}, x_{str}, x_{dom}), y\right) \tag{2}$$

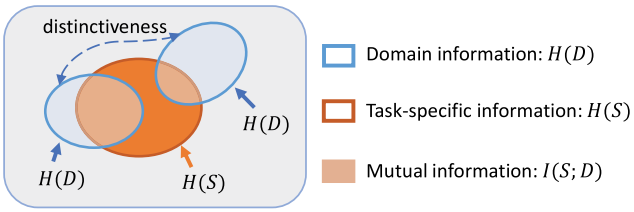

Fig. 2. Training solely on function prediction tasks may lead the model to prioritize the extraction of task-specific information, thereby constraining its representational capacity. By learning information representations from diverse domains, our approach broadens the range of extracted features, thereby enhancing the model's generalization to previously unseen data types.

where $\mathcal{N}_\theta$ denotes the function prediction network, and ℓ represents the loss function, with negative log-likelihood (NLL) used for single-label classification and binary cross-entropy (BCE) for multi-label classification.

In a protein, amino acids are linked by peptide bonds to form a linear chain. The protein sequence can be represented as $x_{\text{seq}} = (s_1, s_2, \ldots, s_n)$, where s_i denotes the residue at the ith position, and n is the total length of the sequence. For protein structures, the spatial coordinates of the backbone C_α atoms are denoted as $x_{\text{str}} = (p_1, p_2, \ldots, p_n)$, where $p_i \in \mathbb{R}^3$ represents the three-dimensional position of the ith C_α atom. These modalities are widely employed in deep learning-based protein modeling. However, existing approaches often lack fine-grained alignment between these different representations.

Therefore, we introduce a knowledge-driven, implicit modality—protein domains—to facilitate alignment between sequence and structure at the segment level. Protein domains serve as the primary determinants of protein function and are composed of specific structural regions. For example, a domain d_1 can be defined as $d_1 = (s_i, \ldots, s_j)$, where $1 \le i < j \le n$. The set of domains within a protein is denoted as $x_{\text{dom}} = (d_1, d_2, \ldots, d_t)$, where t represents the total number of domains in the protein.

3.2. Fine-grained alignment

Proteins are typically composed of multiple domains, and their biological functions often arise from the coordinated interaction of several domains. To capture this functional complexity, we partition the protein into multiple sub-views by combining adjacent domains. Within each sub-view, we aggregate the representations to enhance the alignment of features, while ensuring that representations between different sub-views remain distinct. This approach facilitates fine-grained alignment across various modalities, as illustrated in Fig. 1. Building on this framework, we define "joint domains" as the combination of adjacent domains that collectively contribute to the protein's overall function.

3.3. Functional priors

Previous studies (Fan et al., 2020; Yao et al., 2021; You et al., 2021) directly input domain indices (as shown in Fig. 3(a)) into function prediction networks. However, protein domain annotations are often sparse, which can lead to insufficient training of domain embeddings or input layers when using standard-sized function prediction datasets, typically comprising tens of thousands of protein samples.

To address this issue, we incorporate functional priors into domain embeddings through a pre-training approach. This strategy mitigates the risk of overfitting in downstream tasks, particularly when data is limited, by enabling the network to utilize richer domain representations rather than relying solely on raw domain indices. The enhanced domain embeddings, which we term 'FAD embeddings', offer a more robust representation compared to binary domain annotations. The details of the training process are outlined in Section 4.1, and the effectiveness of the FAD embeddings is demonstrated in Section 5.3.

4. Methods

4.1. Function-aware domain embeddings

To integrate functional priors into domain embeddings, we construct a dataset that includes domain indices, domain descriptions, and Gene Ontology (GO) terms. Without loss of generality, we utilize InterPro entries (Paysan-Lafosse et al., 2022) to represent a total of M protein domains. Building upon this, we create learnable vocabularies for both domain indices and GO terms, as illustrated in Fig. 3(c). Subsequently, we update both vocabularies jointly by leveraging the domain-GO probabilities and the semantic consistency between domain descriptions and text, ensuring that both vocabularies are refined in a complementary manner.

4.1.1. Domain-GO probability

GO terms are widely utilized in bioinformatics tools and databases to help interpret and analyze experimental data, enabling researchers to gain insights into the functions of proteins. We associate protein domains with corresponding GO terms to extract functional priors. Specifically, the domain indices and GO terms can be represented as sets $D = \{domain_i | domain_i \in [0,1]\}$ $(i = 1, 2, \ldots, |D|)$ and $F = \{GO_j | GO_j \in [0,1]\}$ $(j = 1, 2, \ldots, |F|)$, respectively, where $|D|, |F|$ denote the vocabulary sizes of domain indexes and GO terms, respectively. $domain_i = 1$ indicates that a protein contains the domain with index i and $GO_j = 1$ signifies that a protein is associated with the GO term indexed by j.

Each protein independently contains one or more domain indexes and GO terms. We define two types of associations: The association (D_i, F_j, P_k) indicates that protein P_k contains both domain i and the GO term j, while the association (D_i, P_k) signifies that protein P_k possesses domain i, irrespective of the presence of any associated GO terms.

We calculate the prior probability of the distribution of GO terms and utilize it to enhance the functional representation of domain vocabularies. Specifically, the conditional probability of a protein that contains domain i having the GO term j is:

$$p\left(\text{GO}_j \mid \text{domain}_i\right) = \frac{p\left(\text{domain}_i, GO_j\right)}{p\left(\text{domain}_i\right)} = \frac{\sum_{k=1}^{N} I\left(D_i, F_j, P_k\right)}{\sum_{k=1}^{N} I\left(D_i, P_k\right)} \tag{3}$$

where N denotes the total number of protein samples and the operator $I(\cdot)$ indicates the existence of a specific association. The conditional probabilities serve as pseudo-labels to train domain embeddings. We employ a simple network structure consisting of a Hadamard product operator and several feed-forward layers to facilitate the learning of functional relevance within the domain vocabulary. Finally, a mean squared error (MSE) loss function is utilized to train the entire network. The detailed implementation is presented in Appendix A.1.

4.1.2. Domain-text semantically consistent

We further enrich the domain vocabulary with the functional information embedded in the textual descriptions. Specifically, we embed the descriptions by pre-trained BiomedBERT (Gu et al., 2021) model and train the domain vocabulary through contrastive learning:

$$\mathcal{L}_{sem} = -\log \frac{\exp\left(\text{sim}\left(f(\boldsymbol{\phi}_i), f(\boldsymbol{\Phi}(\mathbf{T}_i))\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(f(\boldsymbol{\phi}_i), f(\boldsymbol{\Phi}(\mathbf{T}_j))\right)/\tau\right)} \tag{4}$$

where $f(\cdot)$ denotes a learnable projector that maps embeddings into a shared semantic space, $\boldsymbol{\Phi}$ represents the text encoder, $\phi_i \in \mathbf{R}^c$ is the embedding of domain i, and T_i is the textual description of domain i.

The combination of these two loss functions enables the domain and GO vocabularies to learn complementary information. Subsequently, we utilize the resulting domain vocabularies, referred to as FAD embeddings, to enhance the representations derived from sequence and structure-based features.

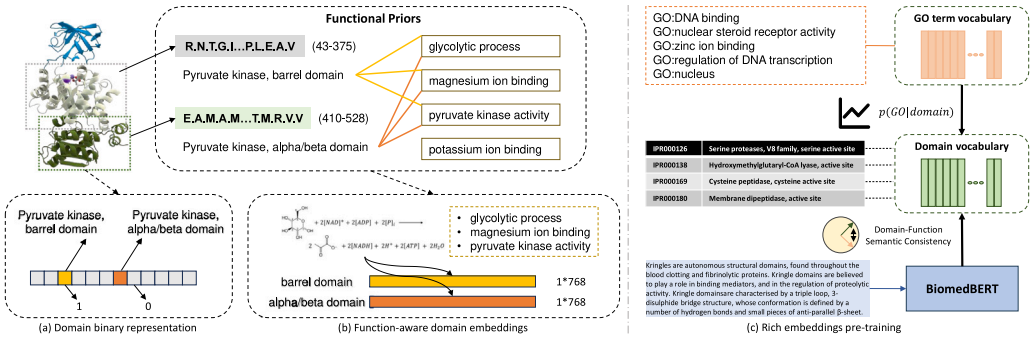

Fig. 3. Introduction to FAD embeddings. (a) Existing methods often rely on binary representations to encode protein domains, which may fail to capture the full complexity of domain relationships. (b) In contrast, we incorporate functional priors into domain embeddings, addressing the issue of domain sparsity and enhancing the representational power of protein domains. (c) We train the rich domain embeddings with domain-GO probabilities pseudo-label and the textual semantic consistency.

4.2. Function prediction network

4.2.1. Modality-specific encoder

We develop a multi-modal framework for protein function prediction, as illustrated in Fig. 4. The framework utilizes the pre-trained protein language model (ESM-2) to extract sequence features (served as $z^s \in \mathbf{R}^c$) and a graph network (CDConv) to extract structure features (served as $z^p \in \mathbf{R}^c$). In addition to these two primary modalities, we employ InterProScan (Jones et al., 2014) to retrieve protein domain information and obtain FAD embeddings from the domain vocabulary. Furthermore, a protein domain attention module, incorporating box positional encodings, is used to adaptively extract the functional representation of joint domains (served as $z^d \in \mathbf{R}^c$). Additional details regarding this process are provided in the appendix.

4.2.2. Domain-joint contrastive learning

To achieve fine-grained alignment across modalities, we leverage the shared information among modalities for the same joint domains. First, we introduce semantic-enhanced embeddings to ensure a more robust representation. Subsequently, we propose a novel contrastive strategy, namely domain-joint contrastive learning, which is coupled with a triplet InfoNCE loss to broaden the range of extracted features in protein data.

Inherent modality enhancement. The semantics in the original input data are often complex, and some information is inevitably lost when encoding it into the feature space. When connecting and aligning existing representation spaces, this loss and bias of meaning will be inherited and amplified, affecting the robustness of alignment. Inspired by Wang et al. (2023), we add zero-mean Gaussian noise into the features and project them to the unit hyper-sphere with L2 normalization:

$$\begin{aligned} \tilde{\mathbf{z}}^s &= \text{Normalize}\left(\mathbf{z}^s + \boldsymbol{\xi}_1\right); \quad \tilde{\mathbf{z}}^p = \text{Normalize}\left(\mathbf{z}^p + \boldsymbol{\xi}_2\right); \\ \tilde{\mathbf{z}}^d &= \text{Normalize}\left(\mathbf{z}^d + \boldsymbol{\xi}_3\right); \end{aligned} \tag{5}$$

where noise items $\xi_1, \xi_2, \xi_3 \in \mathbf{R}^c$ are sampled from zero-mean Gaussian distribution with variance σ^2, and they are not learnable. Hence, aligning two embeddings with noise forces the model to acquire the ability to align all the embeddings within the two circles, leading to a more comprehensive and robust semantic representation.

Domain-joint alignment. To establish the connection between two modalities, we project the semantic-enhanced embeddings (i.e. $\tilde{\mathbf{z}}^s, \tilde{\mathbf{z}}^p, \tilde{\mathbf{z}}^d$) to a new shared space (Poklukar et al., 2022; Wang, Zhao, et al., 2023) via a knowledge-shared projector $f(\cdot)$, respectively.

$$\hat{\mathbf{z}}^s = f\left(\tilde{\mathbf{z}}^s\right); \quad \hat{\mathbf{z}}^p = f\left(\tilde{\mathbf{z}}^p\right); \quad \hat{\mathbf{z}}^d = f\left(\tilde{\mathbf{z}}^d\right) \tag{6}$$

In the projected space, our objective is to ensure that embeddings with similar semantics are close to each other. The various modalities $(x_{seq}, z_{str}, d_{dom})$ derived from the same protein are naturally semantically consistent, and thus, they can be considered as positive pairs for contrastive learning. In contrast, embeddings from different proteins are typically treated as negative samples, as demonstrated in previous work (Hermosilla & Ropinski, 2022; Zhang, Xu, et al., 2023). However, as shown in Fig. 2, while protein-level alignment enhances the protein-level functional representations, it does not significantly expand the information content of the features.

Building on recent advancements in contrastive learning methods (Chen, Kornblith, Norouzi, & Hinton, 2020), we perform domain-joint cropping for proteins to generate negative samples from diverse sub-views, thereby facilitating the extraction of fine-grained functional information at the segment level. Specifically, we randomly sample different joint domains while ensuring that they do not overlap. In our experiments, we utilize two sets of joint domains, $\zeta_1(x_i)$ and $\zeta_2(x_i)$. The sequence-domain contrastive loss $\mathcal{L}_{sdc}$ and the structure-domain contrastive loss $\mathcal{L}_{pdc}$ are defined as follows (with further details provided in the Appendix):

$$\mathcal{L}_{sdc} = \sum_i^N \left[\underbrace{-\text{sim}\left(\hat{\mathbf{z}}^s(\zeta_1(x_i)), \hat{\mathbf{z}}^d(\zeta_1(x_i))\right)}_{L^s_{imc}\text{: pull positive close}} + \underbrace{\text{sim}\left(\hat{\mathbf{z}}^s(\zeta_1(x_i)), \hat{\mathbf{z}}^d(\zeta_2(x_i))\right)}_{L^s_{ivd}\text{: push negative away}} \right] \tag{7}$$

$$\mathcal{L}_{pdc} = \sum_i^N \left[\underbrace{-\text{sim}\left(\hat{\mathbf{z}}^p(\zeta_1(x_i)), \hat{\mathbf{z}}^d(\zeta_1(x_i))\right)}_{L^p_{imc}\text{: pull positive close}} + \underbrace{\text{sim}\left(\hat{\mathbf{z}}^p(\zeta_1(x_i)), \hat{\mathbf{z}}^d(\zeta_2(x_i))\right)}_{L^p_{ivd}\text{: push negative away}} \right] \tag{8}$$

where x_i represent a complete protein, ζ_1, ζ_2 represent various sub-views divided according to the joint domains and N is the number of proteins. The first term is the inter-modality consistency loss which enhances the semantic consistency between multi-modal representations. The second term is the inter-view distinctiveness loss which encourages the representation to efficiently distinguish different protein functions. However, the analysis in Liang, Zhang, Kwon, Yeung, and Zou (2022) suggests that different data modalities are embedded at arm's length in their shared representation in multi-modal models, which is termed as modality gap. It is demonstrated that contrastive learning keeps the different modalities separated by a certain distance and varying the modality gap distance has a significant impact on improving the model's downstream zero-shot classification performance and fairness.

Recent work (Wang, Zhao, et al., 2023) proposes closing the modality gap and guaranteeing that embeddings from different modalities with similar semantics are distributed in the same region of the representation space by removing the repulsive structure in the contrastive loss. However, simply deleting the repulsive structure easily leads to reducing the mutual information (MI) between modalities and cannot keep task-relevant information intact, which leads to decreasing downstream classification accuracy as discussed in Tian et al. (2020).

Inspired by Schroff, Kalenichenko, and Philbin (2015), we construct triplets using various modalities of the sub-views and propose a triplet

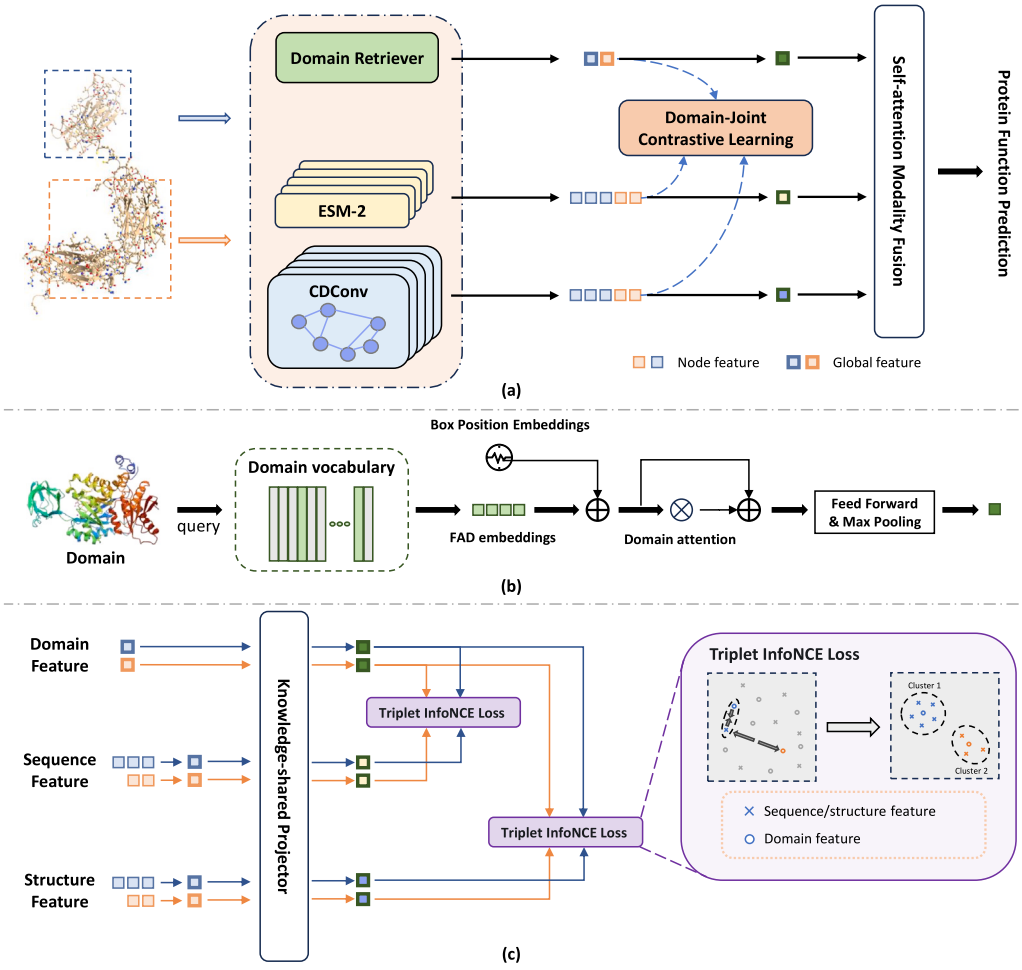

Fig. 4. Multi-modal function prediction architecture. (a) We introduce the comprehensive architecture of our method for protein function prediction. (b) FAD embeddings are derived from enriched vocabularies and aggregated using a domain attention mechanism to capture nuanced functional relevance. (c) To achieve fine-grained alignment across modalities, we employ domain-joint contrastive learning, enabling precise differentiation of diverse protein functions while harmonizing multi-modal information.

loss to replace the inter-view distinctiveness loss in Eq. (7) (similarly for Eq. (8)). Specifically, we use the structure of one sub-view as the anchor, the joint domains from that sub-view as the positive sample, and the joint domains of the other sub-view as the negative sample. The resulting inter-view distinctiveness loss can be formulated as follows:

$$\mathcal{L}^{p}_{ivd} = \sum_{i}^{N} \left[-\operatorname{sim}\left(\hat{\mathbf{z}}^{p}(\zeta_1(x_i)), \hat{\mathbf{z}}^{d}(\zeta_1(x_i))\right) + \operatorname{sim}\left(\hat{\mathbf{z}}^{p}(\zeta_1(x_i)), \hat{\mathbf{z}}^{d}(\zeta_2(x_i))\right) + \alpha \right]_{+} \tag{9}$$

where α is a hyper-parameter used to control the gap between protein representations derived from different sub-views (proofs are provided in the Appendix). We combine the inter-modality consistency loss with the newly inter-view distinctiveness loss, which is referred to as the triplet InfoNCE loss. This combined loss function aims to minimize the modality gap while preserving the distinctions between different sub-views.

$$\mathcal{L}^{p}_{triplet} = \mathcal{L}^{p}_{imc} + \lambda \mathcal{L}^{p}_{ivd} \tag{10}$$

where λ is a hyper-parameter that regulates training stability. When λ is small, it is less likely to create modality gaps, although this may lead to decreased training efficiency. Ultimately, the domain-joint contrastive learning is supervised by the combination of the sequence-domain triplet loss $L^{s}_{triplet}$ and the structure-domain triplet loss $L^{p}_{triplet}$.

4.2.3. Prediction head

Domain-joint contrastive learning serves as a feature constraint within our architecture, and the outputs of the knowledge-shared projector are not directly used for function prediction. Instead, we utilize a self-attention layer and a two-layer MLP to aggregate the multi-modal features z_s, z_p and z_d. Finally, we apply the loss function described in 3.1 to train the prediction network:

$$\mathcal{L}_{cla} = \ell\left(\mathbf{Agg}[z_s(x_i), z_p(x_i), z_d(x_i)], y_i\right) \tag{11}$$

The overall loss function is the combination of $\mathcal{L}_{cla}$ and $L_{triplet}$.

5. Results

5.1. Experimental setups

Domain embeddings pre-training. We collected 570,830 protein entries from Swiss-Prot (The UniProt Consortium, 2022) (release 2024_1) including the InterPro IDs and GO term IDs. We discarded all proteins which had no InterPro annotations. We also collected the mappings of InterPro entries to GO terms and textual descriptions from the InterPro Database (Paysan-Lafosse et al., 2022) (release 2023_10). In summary, our dataset contained 551,756 proteins with 31,929 unique domains and 28,944 unique GO terms. We finally retained 1,454,811 domain-GO term paired samples. For domains with several textual descriptions, we random select one description in each batch.

Table 1
F_{max} of gene ontology term prediction and enzyme commission number prediction. The highest-performing results are highlighted in **bold**, while the second-best results are underlined for clarity.

Input	Method	Additional	Gene ontology			Enzyme
		Modality	BP	MF	CC	Commission
Sequence	ESM-1b (Rives et al., 2021)[b]	–	0.452	0.657	0.477	0.864
	ESM-2(Lin et al., 2022)[c]	–	0.460	0.661	0.445	0.880
	SaProt (Su et al., 2023)	–	0.356	0.678	0.414	0.884
Structure	GVP (Jing et al., 2021)[a]	–	0.326	0.426	0.420	0.489
	IEConv (Hermosilla et al., 2021)[a]	–	0.421	0.624	0.431	0.735[b]
	GearNet (Zhang, Xu, et al., 2023)[b]	–	0.356	0.503	0.414	0.730
	CDconv (Fan et al., 2022)[a]	–	0.453	0.654	0.479	0.820
	ClusteringPRL (Quan, Wang, Ma, Fan, & Yang, 2024)	–	0.474	0.675	0.483	0.866
Sequence & Structure	DeepFRI (Gligorijević et al., 2021)[b]	–	0.399	0.465	0.460	0.631
	LM-GVP (Wang et al., 2022)[b]	–	0.417	0.545	0.527	0.664
	ESM-GearNet (Zhang, Wang, et al., 2023)[c]	–	0.488	0.681	0.464	0.890
Multimodal	ProteinINR (Lee et al., 2023)	surface	0.508	0.678	0.506	0.890
	ProteinSSA (Hu et al., 2024)	GO terms	0.464	0.667	0.492	0.857
	DPfunc (Wang et al., 2025)	domain	0.483	0.667	0.537	0.823
	ProtFAD	domain	**0.518**	**0.701**	**0.551**	**0.911**

[a] Results are from Fan et al. (2022).
[b] Results are from Zhang, Xu, et al. (2023).
[c] Results are from Zhang, Wang, et al. (2023).

Table 2
Accuracy of protein fold classification and enzyme catalytic reaction classification. The highest-performing results are highlighted in **bold**, while the second-best results are underlined for clarity.

Input	Method	Additional	Fold classification		Enzyme
		Modality	Superfamily	Family	Reaction
Sequence	ESM-1b (Rives et al., 2021)[b]	–	0.601	0.978	0.831
	ESM-2(Lin et al., 2022)	–	0.789	0.992	0.894
Structure	GVP (Jing et al., 2021)[a]	–	0.225	0.838	0.655
	IEConv (Hermosilla et al., 2021)[a]	–	0.702	0.992	0.872
	GearNet (Zhang, Xu, et al., 2023)[b]	–	0.805	**0.999**	0.875
	CDconv (Fan et al., 2022)[a]	–	0.777	0.996	0.885
	ClusteringPRL (Quan et al., 2024)	–	0.812	0.996	0.896
Sequence & Structure	DeepFRI (Gligorijević et al., 2021)[b]	–	0.206	0.732	0.633
Multimodal	ProteinSSA (Hu et al., 2024)	GO terms	0.794	0.998	0.894
	ProtFAD	domain	**0.908**	0.998	**0.923**

[a] Results are from Fan et al. (2022).
[b] Results are from Zhang, Xu, et al. (2023).

We pre-train the FAD embeddings for 500 epochs with a batch size of 16,384. Note that the training data only includes domain indices with no protein sequences involved, so there are no generalization concerns of protein sequences. After training, we freeze the parameters of the FAD embeddings for downstream tasks.

Benchmark tasks. Following Fan et al. (2022), Hermosilla et al. (2021), Zhang, Xu, et al. (2023), we evaluate the proposed method on four tasks: protein fold classification (Hermosilla & Ropinski, 2022; Hermosilla et al., 2021), enzyme reaction classification (Hermosilla et al., 2021), gene ontology (GO) term prediction (Gligorijević et al., 2021) and enzyme commission (EC) number prediction (Gligorijević et al., 2021). Protein fold classification includes two evaluation scenarios: superfamily and family (the fold scenario in Fan et al. (2022) is not relevant to protein function prediction, so we removed it). GO term prediction includes three sub-tasks: biological process (BP), molecular function (MF), and cellular component (CC) ontology term prediction. All the datasets are split into training, validation, and test sets, with each set consisting of independent protein sequences. The model is trained on the training set, and the results are reported exclusively on the test set, which consists of PDB chains with a sequence identity of $\leq 95\%$ relative to the chains in the training set, for both EC and GO tasks. Protein fold and enzyme reaction classification are single-label classification tasks. Mean accuracy is used as the evaluation metric. GO term and EC number prediction are multi-label classification tasks. The F_{max} accuracy is used as the evaluation metric.

Note that we use domains instead of GO terms as input, so there is no data leakage. For each protein in the datasets, we assigned InterPro domains using InterProScan 5 (Jones et al., 2014).

Implementation. In our experiment, we utilize the pre-trained ESM-2 model (Lin et al., 2022) with its parameters frozen to minimize unnecessary computational overhead. The embedding dimension is set to 768. A simple multi-layer perceptron (MLP) is employed to project the encodings of three modalities to a unified dimension of 1280. The noise variance σ^2 in Eq. (5) is configured to 0.01. The hyper-parameters α and λ for the triplet InfoNCE loss are set to 1 and 0.1, respectively. Further details regarding the implementation and training setup are provided in Appendix C.

5.2. Comparison with state-of-the-art

We conduct a comparative analysis of our proposed method against existing approaches, including sequence-only, structure-only, and multimodal methods. The results are presented in Table 1 and Table 2. To clearly differentiate between methods with and without the inclusion of a third modality, methods utilizing only sequence and structure are labeled as "Sequence & Structure" while those incorporating additional modalities are categorized as "Multimodal" with the specific third modality indicated.

Methods that combine sequence and structure modalities generally outperform single-modality approaches, emphasizing the importance of

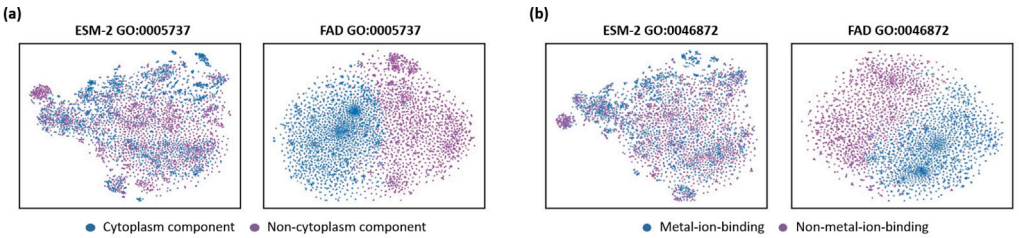

Fig. 5. The dimensionality-reduction visualizations of domain embeddings extracted by ESM-2 and our FAD embeddings. Two common types of functions, including "cytoplasm" (a) and "metal-ion-binding" (b), are selected.

Table 3
Evaluation of FAD embeddings on benchmarks. Fmax is used for gene ontology term prediction (GO) and enzyme commission number prediction (EC). Accuracy is used for protein fold classification (FC) and enzyme catalytic reaction classification (ER).

Method	FC		ER	GO			EC
	Superfamily	Family		BP	MF	CC	
ESM-2	0.789	**0.992**	**0.894**	0.460	0.661	0.445	**0.880**
domain binary representation	0.778	0.991	0.860	0.486	0.683	0.493	0.863
FAD	**0.850**	0.991	0.867	**0.511**	**0.698**	**0.533**	0.878

multimodal fusion in protein function prediction. Interestingly, methods that include knowledge modalities (such as surface features or GO terms) do not always outperform the sequence–structure-based models, which suggests that how knowledge is effectively integrated into the multi-modal model is a crucial design consideration. Despite this, ProtFAD outperforms nearly all existing methods across all benchmarks, underscoring the overall effectiveness of the network design we proposed. In subsequent experiments, we will delve deeper into the specific innovations presented in this paper and analyze their individual contributions to the model's performance.

5.3. Rich domain embeddings

To evaluate the ability of FAD embeddings to capture functional priors, we utilize t-SNE to visualize the domain representations generated by FAD and ESM-2. We focus on two common functional categories, namely "cytoplasm" and "metal-ion-binding". For each function, we sample an approximately balanced set of positive and negative domains. The domains are embedded using FAD, while the sequences corresponding to the domains are embedded using ESM-2.

As illustrated in Fig. 5, the representations generated by ESM-2 exhibit significant overlap between positive and negative samples, whereas the FAD-generated representations display clear separation. This result underscores the ability of FAD embeddings to effectively capture fine-grained functional characteristics of proteins.

Moreover, to further validate the effectiveness of FAD embeddings in comparison to binary domain representations, we combine the embeddings with an MLP prediction head and evaluate their performance on benchmark datasets. As shown in Table 3, FAD embeddings demonstrate superior robustness compared to binary domain representations. Even when only the domain modality is utilized, FAD achieves competitive performance relative to pre-trained protein language models, while incurring significantly lower computational overhead.

5.4. Fine-grained alignment

We conduct a series of experiments to evaluate the effectiveness of the proposed domain-joint contrastive learning method. Specifically, we compare two models with identical architectures, one incorporating domain-joint contrastive learning and the other without the contrastive learning mechanism. For both models, we extract sequence and structural features from the protein samples and compute the average Euclidean distance between the corresponding sequence–structure feature pairs. The resulting average distances are 16.24 for the model with contrastive learning and 20.68 for the model without contrastive learning. Additionally, we split each protein sample into two fragments, deviating from the joint domain-based splitting used during training. The average structural feature distances between the two fragments for all proteins are 8.33 for the contrastive learning model and 3.42 for the non-contrastive model. The results indicate that the domain-joint contrastive learning method is more effective at aligning different modalities (i.e., achieving smaller distances between sequence–structure feature pairs) while also better distinguishing finer-grained protein features (i.e., larger distances between different fragments).

The experiments are conducted using the validation set of the cellular component ontology term prediction dataset, ensuring no risk of data leakage. We further visualize the results of these experiments using principal component analysis (PCA), as shown in Fig. 6, to illustrate the distribution of features learned by our method.

5.5. Time cost

To evaluate the computational efficiency of our approach, we conduct experiments comparing the training time and inference speed of our method with existing models. The domain pre-training phase is performed on a single Tesla V100-SXM2-32 GB GPU, with a training time of 8 h. In comparison, while the training time for the ESM-2 model is not explicitly reported, it is significantly longer due to the more complex protein sequence data used in ESM-2, whereas our domain pre-training relies on simpler data (domain indices, GO terms, and textual descriptions) that do not require training new sequence models, resulting in a more efficient feature extraction process. Regarding inference speed, the ESM-2 model takes approximately 4.1 s to process a single protein, while our method, FAD embeddings, only requires a simple embedding lookup, incurring negligible time overhead. For the functional prediction task, we compare our method with the backbone model CDConv and the multi-modal model ESM-GearNet. The training times per epoch for these models are as follows: CDConv (2.11 min), ESM-GearNet (3.62 min), and ProtFAD (ours, 2.28 min). Feature extraction for ESM-2 is conducted during the data preprocessing stage, not included in the training time. Although our method incurs slightly higher training costs than CDConv, this is due to the inclusion of the domain attention mechanism and feature fusion network, which significantly improve model performance. Furthermore, when compared to the multi-modal competitor ESM-GearNet, our method achieves superior performance with lower computational cost.

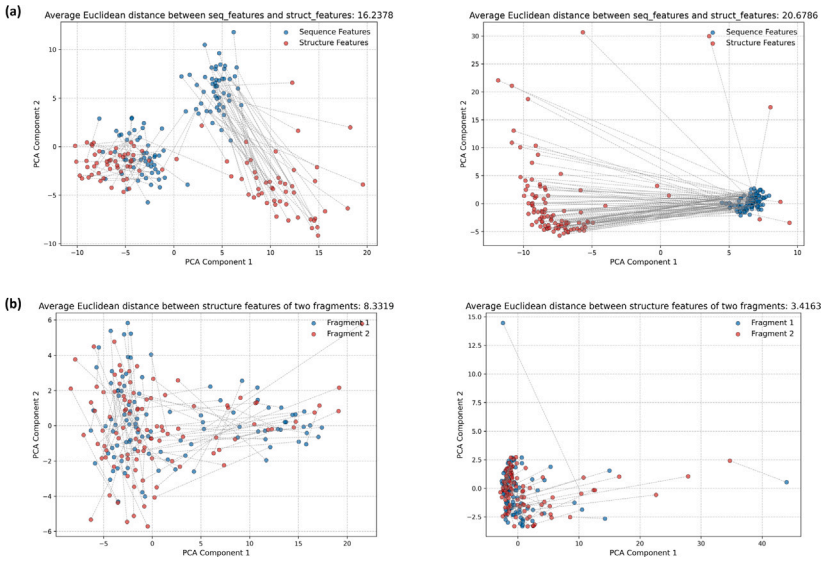

Fig. 6. The dimensionality-reduction visualizations of sequence/structure features extracted by ProtFAD (left) and its variants without domain-joint contrastive learning (right). 100 protein samples are randomly selected from the GO-CC dataset. The average distance is calculated on complete GO-CC validation set.

Table 4
Ablation study of the proposed modules. Fmax is used for gene ontology term prediction (GO) and enzyme commission number prediction (EC). Accuracy is used for protein fold classification (FC) and enzyme catalytic reaction classification (ER).

Method	FC		ER	GO			EC
	Superfamily	Family		BP	MF	CC	
ProtFAD	**0.908**	**0.998**	**0.923**	**0.518**	**0.701**	**0.551**	**0.911**
w/o domain	0.830	0.995	0.906	0.485	0.672	0.519	0.875
w/o domain pre-train	0.881	0.996	0.909	0.495	0.683	0.532	0.868
w/o domain attention	0.871	0.997	0.909	0.508	0.697	0.545	0.899
w/o contrastive loss	0.902	0.997	0.910	0.514	0.699	0.528	0.904
w/ vanilla contrastive loss	0.891	0.997	0.911	0.512	0.698	0.520	0.905

5.6. Ablation study

To assess the impact of various components, we conduct an ablation study across four tasks, with results presented in Table 4.

Overall Contributions. Initially, we aim to highlight the effectiveness of the key innovations in our approach, including function-aware domain pre-training and domain-joint contrastive learning. In this setting, we perform protein function prediction using only sequence and structure features, omitting the contrastive loss, and refer to this as "w/o domain". The substantial decline in performance across all benchmark tasks underscores the critical contributions of our proposed method.

Function-Aware Domain Pre-training. Subsequently, we investigate a degenerate version of the FAD embeddings, labeled "w/o domain pre-train", where binary features are used to represent domains instead of utilizing the pre-trained embeddings. Our findings demonstrate that domain pre-training significantly enhances the final performance of feature prediction. This improvement is likely attributed to the pre-training process, which strengthens the generalization of domain embeddings and mitigates the challenges posed by the limited data available for downstream tasks.

Domain Attention Mechanism. Next, we remove the domain attention mechanism and replace it with the mean of domain embeddings, resulting in a consistent performance degradation. The domain attention mechanism is crucial for effectively integrating interaction and positional information between domains, allowing the overall domain representation to better serve downstream functional prediction tasks. By using the mean embeddings, the model loses the ability to capture the intricate inter-domain relationships, leading to suboptimal performance.

Domain-Joint Contrastive Learning. Furthermore, we investigate the impact of excluding the protein domain-joint contrastive learning from our model architecture. The observed performance decline further emphasizes the contributions of the proposed module. Specifically, the domain-joint contrastive learning approach outperforms methods that either do not use contrastive learning or rely on vanilla sample-level contrastive learning. Our method is particularly suited to the protein multi-modal representation fusion scenario, where domain-specific interactions are crucial for accurate prediction.

6. Conclusion

In this work, we propose ProtFAD, a priors-guided multi-modal protein representation learning approach, to bridge the gap between sequence or structure modality and protein functions. By leveraging function-aware domain embeddings and a novel domain-joint contrastive learning, we extract rich functional information from protein domains and enhance existing protein representation. During the fusion of multi-modal features, we further incorporate domain positional information and the synergistic effects between domains with a domain attention mechanism. Extensive experiments on diverse protein function prediction benchmarks verify the superior performance of ProtFAD. However, due to architectural constraints, the current approach is limited to low-level functional prediction tasks. It may not

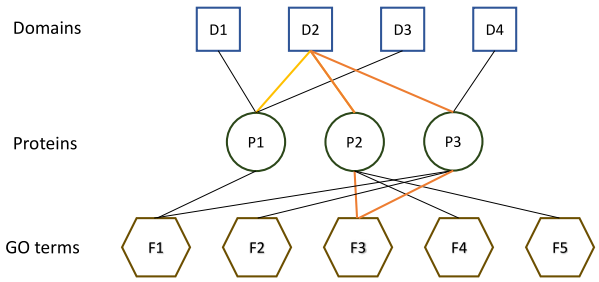

Fig. A.7. Establish associations between domains and GO terms through proteins. Define two distinct meta-paths and utilize them to summarize the connection between domains and GO terms, including domain-protein-function (red line) and domain-protein (yellow line).

perform as well on high-level tasks, such as site-specific prediction. Future work will focus on exploring a multi-scale knowledge fusion strategy for protein representation, aiming to overcome the limitations of knowledge-enhanced methods across tasks of different granularities.

CRediT authorship contribution statement

Mingqing Wang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft. **Zhiwei Nie:** Conceptualization, Visualization, Writing – original draft. **Yonghong He:** Resources, Writing – review & editing. **Athanasios V. Vasilakos:** Supervision, Writing – review & editing. **Zhixiang Ren:** Formal analysis, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by Peng Cheng Cloud-Brain.

Appendix A. Additional modules

A.1. Domain-protein-GO term association

Here, we establish the association between domains and GO terms. We collect the proteins from Swiss-Prot (The UniProt Consortium, 2022) including the InterPro IDs and GO term IDs, and establish a relationship network among them as shown in Fig. A.7. When a protein k containing domain i has the GO j, we define a meta-path (D_i, F_j, P_k). When a protein k contains the domain i, we define another meta-path (D_i, P_k). Two meta-paths may overlap for some proteins. The meta-paths are referred to as associations in Section 4.1.

Furthermore, we show the network architecture for training FAD embeddings as shown in Fig. A.8. The projectors in the network are composed of two-layer MLP with hidden layer activation function. For two matrices A and B of dimensions $m \times n$, the Hadamard product is defined as:

$$\mathbf{C} = \mathbf{A} \circ \mathbf{B}, \quad (\mathbf{A} \circ \mathbf{B})_{ij} = A_{ij} B_{ij} \tag{A.1}$$

Table A.5
Comparative study of various positional encodings (PE) on enzyme commission number prediction (EC). The methods are evaluated in the ProtFAD framework without domain-joint contrastive learning.

positional encoding	w/o PE	BERT PE	Field PE	MLP PE	Box PE
enzyme commission	0.9024	0.8998	0.9015	0.9045	**0.9108**

A.2. Domain attention module

Protein function may be determined by several domains, and different domains may contribute to different functions. Therefore, we employ a protein domain attention module to adaptively extract the functional representation of joint domains. Specifically, we use $\mathbf{e} = [e_1, e_2, \ldots, e_t]$ to represent the FAD embeddings of joint domains $d_{[1,t]}$. We adopt a self-attention layer to calculate the importance of each domain, as shown below:

$$\hat{d} = \mathbf{Agg}(\mathbf{e} + \Omega(\mathbf{e})) \tag{A.2}$$

where $\Omega(\cdot)$ is the self-attention layer and $\mathbf{Agg}(\cdot)$ is an aggregation operator (which is average pooling in our experiment).

Considering that the position of the domain may affect the function of the protein, we incorporate a positional encoding for the domain. The sequence length and relative position of each domain within a protein are different. For simplicity, we take the position of the amino acid in the middle of the domain sequence as the position of the domain and normalize it to (0, 1) by dividing it by the length of the domain sequence. Common discrete positional encodings cannot be used for the continuous position values described above. Therefore, we employ box positional encodings. Specifically, we group continuous position p within the interval (0, 1) into bins and learn a unique position embedding ψ for each bin. Finally, the calculation of domain attention can be represented as:

$$\hat{e}_i = e_i + \psi_{\lfloor b \times p_i \rfloor}, \quad i = 1, 2, \ldots, t \tag{A.3}$$

$$\hat{d} = \mathbf{Agg}(\hat{\mathbf{e}} + \Omega(\hat{\mathbf{e}})) \tag{A.4}$$

where b is the number of bins, and ψ_j represents the position embedding of j_{th} bin.

This attention module considers the relationships between different domains and the positional information of each domain and generates a single domain-joint representation $\hat{d} \in \mathbf{R}^c$, enabling the model to effectively capture the complex interplay between domains and their contributions to protein function.

In addition, we explore the performance of different positional encodings (PE):

(1) BERT positional encoding: learn a position embedding for each element in joint domains (*i.e.* $i = 1, 2, \ldots, t$). This approach considers the relative positional relationship between domains without receiving their real positions as input.

(2) Field positional encoding: learn a position embedding for all domains, and the positional encoding is the product of the embedding and the domain position p. This approach integrates the positional information with a linear function.

(3) MLP positional encoding: employ one MLP projecting the domain position p to a position vector for each domain.

The results are shown in Table A.5. The box position encoding achieves the best performance. Note that BERT PE and Field PE are less effective than not using the positional encoding, proving that the continuous position p cannot be modeled with common discrete encodings.

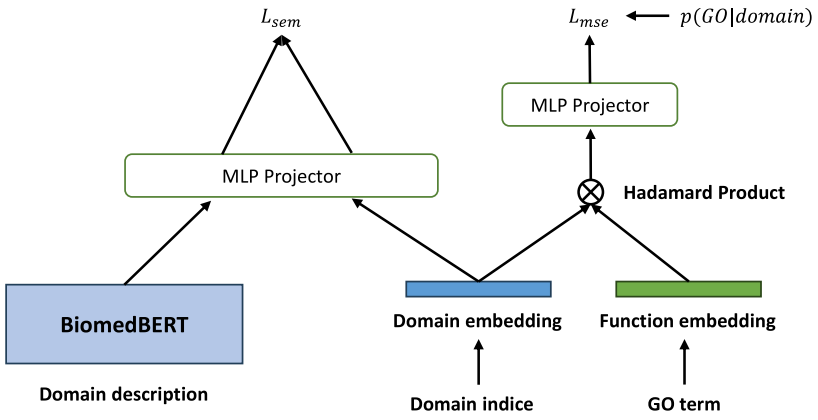

Fig. A.8. The network architecture for training FAD embeddings.

Appendix B. Proof

B.1. Derivation of Eq. (8)

We perform domain-joint cropping on proteins and construct various sub-views, i.e. $\zeta_1(x), \zeta_2(x), \ldots, \zeta_K(x)$. We consider the multiple modalities of the same sub-view as positive samples, and the different sub-views as negative samples. We utilize two modalities for contrastive learning. For structure and domain, the contrastive loss is:

$$L_{pdc} = -\log \frac{\exp\left(\mathrm{sim}\left(\hat{\mathbf{z}}^p(\zeta_1(x)), \hat{\mathbf{z}}^d(\zeta_1(x))\right)\right)}{\sum_{k\neq 1}^{K} \exp\left(\mathrm{sim}\left(\hat{\mathbf{z}}^p(\zeta_1(x)), \hat{\mathbf{z}}^d(\zeta_k(x))\right)\right)} \tag{B.1}$$

where K is the number of sub-views, ζ_k represents the kth protein sub-view. Let K=2, we get

$$L_{pdc} = -\mathrm{sim}\left(\hat{\mathbf{z}}^p(\zeta_1(x)), \hat{\mathbf{z}}^d(\zeta_1(x))\right) + \mathrm{sim}\left(\hat{\mathbf{z}}^p(\zeta_1(x)), \hat{\mathbf{z}}^d(\zeta_2(x))\right) \tag{B.2}$$

B.2. Mutual information

Here, we present the domain-joint contrastive learning approach from an information-theoretic perspective. Let the protein be denoted as X, which is partitioned into two sub-views, X_1 and X_2, such that $X = X_1 + X_2$. The function label is denoted as Y, and the feature representation derived from our encoder is Z. Our objective is to maximize the mutual information $I(Z, X)$ and $I(Z, Y)$. The function prediction loss is employed to maximize $I(Z, Y)$, while domain-joint contrastive learning is utilized to maximize $I(Z, X)$, thereby preserving more information for robust generalization to unseen data. The optimization problem can be expressed as:

$$maximum\ I(Z, X) = I(Z, X_1) + I(Z, X_2) - I(Z, X_1; X_2) \tag{B.3}$$

The model facilitates the representation Z to capture the sequence–structure correlation within X_1 by comparing its sequence and structural features. That is, by minimizing the conditional entropy $H(X_1|Z)$, the model enhances its ability to represent X_1. The optimization objective implicitly maximizes $I(Z, X_1) = H(X_1) - H(X_1|Z)$. Consequently, the positive pairwise alignment within the contrastive learning framework enhances both $I(Z, X_1)$ and $I(Z, X_2)$, thereby improving the model's representation of both sub-views.

The negative pair discrimination objective in domain-joint contrastive learning minimizes the redundancy in the representation of Z by increasing the distance between the substructures X_1 and X_2, thereby reducing the overlap of information between them:

$$I(Z, X_1; X_2) = H(Z|X_1) - H(Z|X_1, X_2) \tag{B.4}$$

Therefore, the mutual information $I(Z, X)$ is maximized through the domain-joint contrastive learning approach.

B.3. Triplet InfoNCE loss

The analysis in Liang et al. (2022) suggests that contrastive learning keeps the different modalities separated by a certain distance. Here, we demonstrate how our proposed triplet InfoNCE loss closes the modality gap while keeping the sample distinctiveness. We denote two different samples as z_1, z_2, and use superscripts p, d to indicate two different modalities (e.g. structure and domain). The contrastive loss of two sub-views can be expressed as:

$$L = \underbrace{d\left(z_1^p, z_1^d\right)}_{\text{attractive structure}} - \underbrace{d\left(z_1^p, z_2^d\right)}_{\text{repulsive structure}} \tag{B.5}$$

where $d(A, B) = 1 - \mathrm{sim}(A, B)$ represent the cosine distance. Euclidean distance and cosine distance are monotonically related (A, B are L2-normalized), that is $d_{Euclidean} = \sqrt{2d}$. So we use Euclidean distance to demonstrate the following process on a two-dimensional plane.

For example, the repulsive structure in the contrastive loss keeps the modality gap when similar samples are aligned as shown in Fig. B.9(a). Simply deleting the repulsive structure like Wang, Zhao, et al. (2023) easily leads to aggregating embeddings with different semantics as shown in Fig. B.9(b), which reduces the mutual information (MI) between modalities.

For the triplet InfoNCE loss, when the loss is not converged, we get:

$$\begin{aligned} L_{triplet} &= d\left(z_1^p, z_1^d\right) - 1 + \lambda\left[d\left(z_1^p, z_1^d\right) - d\left(z_1^p, z_2^d\right) + \alpha\right]_+ \\ &= (1+\lambda)d\left(z_1^p, z_1^d\right) - \lambda d\left(z_1^p, z_2^d\right) - 1 + \lambda\alpha \\ &= \lambda\left[(1+\frac{1}{\lambda})d\left(z_1^p, z_1^d\right) - d\left(z_1^p, z_2^d\right) + \alpha - \frac{1}{\lambda}\right] \end{aligned}$$

where λ is a hyper-parameter that adjusts training stability, and α is a hyper-parameter employed to control the gap between different samples. When $\lambda \to \infty$, $L_{triplet}$ will degenerate into a vanilla contrastive loss. And when λ is small, it is less likely to create modality gaps as the repulsive structure accounts for less in the loss function, as shown in Fig. B.9(c).

When the loss is converged, we get:

$$d\left(z_1^p, z_1^d\right) - d\left(z_1^p, z_2^d\right) + \alpha \leq 0 \tag{B.6}$$

When the modalities are aligned, we expect $d\left(z_1^p, z_1^d\right) = 0$. That is

$$0 = d\left(z_1^p, z_1^d\right) \leq d\left(z_1^p, z_2^d\right) - \alpha \tag{B.7}$$

z_1, z_2 are interchangeable, so we have

$$d\left(z_1^p, z_2^d\right) \geq \alpha; \quad d\left(z_2^p, z_1^d\right) \geq \alpha \tag{B.8}$$

$$d\left(z_1^d, z_2^d\right) \geq \alpha; \quad d\left(z_2^p, z_1^p\right) \geq \alpha \tag{B.9}$$

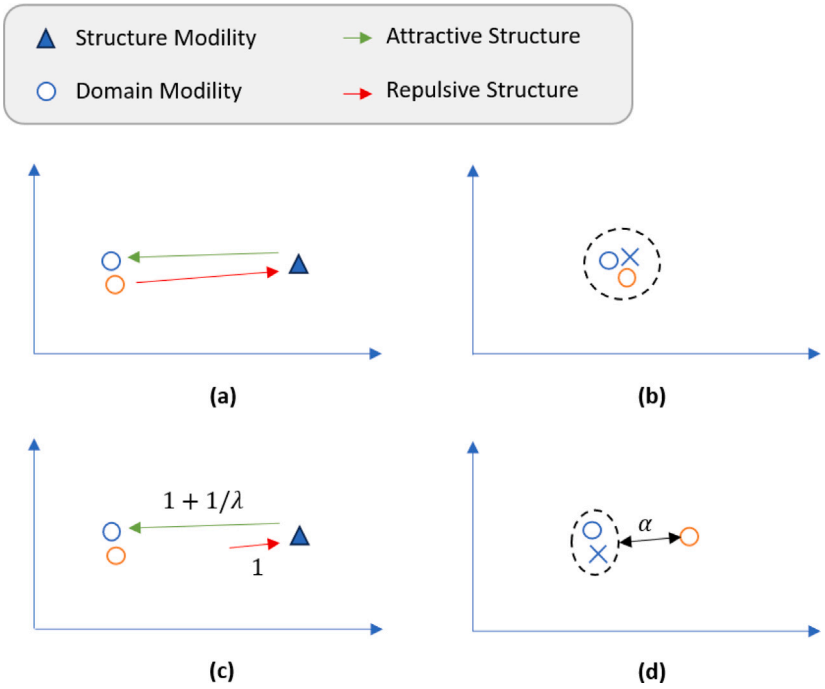

Fig. B.9. (a) Attractive structure and repulsive structure jointly keep the modality gap. (b) After deleting the repulsive structure, the different samples cannot be distinguished (the dashed line represents the aggregated embeddings). (c) The triplet InfoNCE loss eliminates the modality gap by increasing the weight of the attractive structure. (d) The triplet InfoNCE loss enlarges the mutual information by maintaining the lower bound of sample distance.

Table C.6
More details of implementation and training setup.

Hyper-parameter	EC	GO-BP	GO-MF	GO-CC	Fold	Enzyme reaction
Batch size	24	32	32	64	8	8

The α is the lower bound of the distance between different samples as shown in Fig. B.9(d), which keeps the mutual information between different modalities for better alignment (Tian et al., 2020; Wang & Isola, 2020).

Appendix C. More experiment details

We use Tesla V100-SXM2-32 GB GPU for single-card training. More details of implementation and training setup are provided in Table C.6.

Appendix D. Discussion

D.1. Domain embedding dimension

To explore the density of information contained in the domain, we compare the impact of different domain embedding dimensions on model performance. Specifically, we only employ the domain modality for enzyme commission number prediction, comparing the results of FAD and domain embeddings without pre-training under different dimension settings. The results are shown in Fig. D.10.

As the dimension of FAD embeddings increases, the model performance maintains an improving trend, indicating that FAD does integrate effective functional information. In addition, it implies that FAD has greater potential for function perception as the dimension increases. For domain embeddings that are not pre-trained, the model performance reaches saturation when the embedding dimension is 1024. This may be caused by insufficient data in the function prediction task to train a stronger domain representation, i.e., the model is overfitting. This further illustrates the importance of pre-training our functional representations.

Table D.7
Absolute study of ProtFAD without sequence modality on gene ontology term prediction and enzyme commission number prediction.

Method	Gene ontology			Enzyme commission
	BP	MF	CC	
ProtFAD	**0.518**	**0.701**	**0.551**	**0.911**
w/o sequence	0.500	0.694	0.496	0.909
Δ	−1.8%	−0.7%	−5.5%	−0.2%
w/o structure	0.515	0.701	0.548	0.906
Δ	−0.3%	0%	−0.3%	−0.5%

D.2. Is sequence or structure necessary?

We further evaluate the performance of ProtFAD without sequence modality or without structure modality, and provide the results in Table D.7. The sequence modality contributes to the multi-modal representations for most benchmarks, especially the cellular component ontology term prediction (an improvement of nearly 10% compared to the degenerate model). Surprisingly, adding sequence modality reduces the performance of enzyme commission number prediction. This may be caused as the information in structure and domain is sufficient for the task. However, for structure modality, the performance degradation is not obvious, which introduces that the information in domains is sufficient for the protein function prediction or the information mining in previous structure-based work is inadequate. It proves the correctness of using domains as an implicit modality to connect structure and function.

D.3. How to crop the protein

In the experiment, we divide the domains of a protein (*i.e.* $x_{dom} = (d_1, d_2, \dots, d_t)$) into two subsets $\zeta_1(x_{dom})$ and $\zeta_2(x_{dom})$. Specifically, we random select $k \in (1, t]$, let $\zeta_1(x_{dom}) = (d_1, \dots, d_{k-1})$, $\zeta_2(x_{dom}) = (d_k, \dots, d_t)$. Then we search the corresponding sequence and structure of the divided domains for the two modalities.

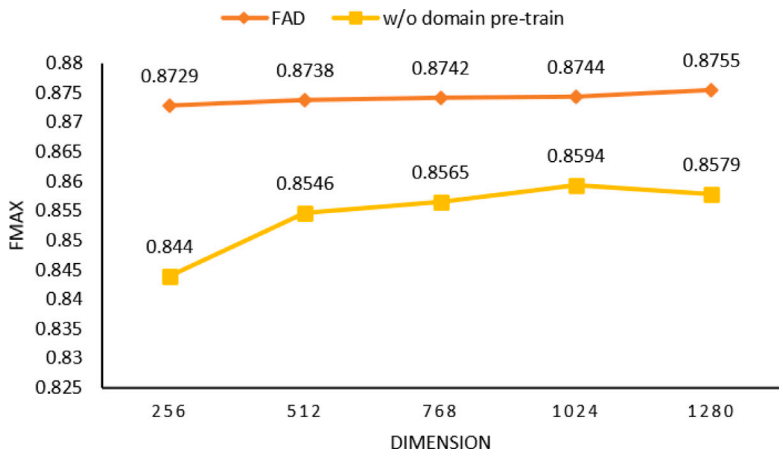

Fig. D.10. Comparative study of various domain embedding dimensions on enzyme commission number prediction.

This division cannot separate the sequence and structure of the two sub-views, introducing noise into the contrastive learning. Reducing such noise can further enhance the effect of contrastive learning. However, we have not conducted an in-depth exploration of how domains are divided, which may be a meaningful future work.

D.4. Limitation

In the context of protein function prediction, there are some limitations when using domain-based approaches. First, since domains are structural units composed of multiple atoms or residues,enhancing function prediction based on protein domains in more granular tasks (such as binding site prediction) requires increasingly complex network architectures. This added complexity can introduce challenges in model design and optimization. Second, due to the inductive bias inherent to protein domains, domain-based function prediction may suffer from severe overfitting in cases where data is limited, despite improvements from function-prior pretraining. While such pretraining can mitigate overfitting to some extent, the fixed nature of domain-specific patterns may still cause models to overly rely on these biases, reducing their generalizability to unseen data. These challenges highlight the trade-offs involved in domain-centric approaches, particularly when scaling to more granular or data-scarce tasks.

Data availability

Our implementation is available at https://github.com/AI-HPC-Research-Team/ProtFAD.

References

Benkovic, Stephen J., & Hammes-Schiffer, Sharon (2003). A perspective on enzyme catalysis. *Science*, *301*(5637), 1196–1202.

Cai, Yideng, Wang, Jiacheng, & Deng, Lei (2020). SDN2go: an integrated deep learning model for protein function prediction. *Frontiers in Bioengineering and Biotechnology*, *8*, 391.

Chen, Ting, Kornblith, Simon, Norouzi, Mohammad, & Hinton, Geoffrey (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PMLR.

Chen, Can, Zhou, Jingbo, Wang, Fan, Liu, Xue, & Dou, Dejing (2023). Structure-aware protein self-supervised learning. *Bioinformatics*, *39*(4), btad189.

Cheng, Jun, Novati, Guido, Pan, Joshua, Bycroft, Clare, Žemgulytė, Akvilė, Applebaum, Taylor, et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, *381*(6664), eadg7492.

Corso, Gabriele, Stärk, Hannes, Jing, Bowen, Barzilay, Regina, & Jaakkola, Tommi (2023). DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International conference on learning representations*.

Elnaggar, Ahmed, Heinzinger, Michael, Dallago, Christian, Rehawi, Ghalia, Wang, Yu, Jones, Llion, et al. (2021). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127.

Fan, Kunjie, Guan, Yuanfang, & Zhang, Yan (2020). Graph2GO: a multi-modal attributed network embedding method for inferring protein functions. *GigaScience*, *9*(8), giaa081.

Fan, Hehe, Wang, Zhangyang, Yang, Yi, & Kankanhalli, Mohan (2022). Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The eleventh international conference on learning representations*.

Forslund, Kristoffer, & Sonnhammer, Erik L. L. (2008). Predicting protein function from domain content. *Bioinformatics*, *24*(15), 1681–1687.

Gligorijević, Vladimir, Renfrew, P. Douglas, Kosciolek, Tomasz, Leman, Julia Koehler, Berenberg, Daniel, Vatanen, Tommi, et al. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, *12*(1), 3168.

Gu, Zhonghui, Luo, Xiao, Chen, Jiaxiao, Deng, Minghua, & Lai, Luhua (2023). Hierarchical graph transformer with contrastive learning for protein function prediction. *Bioinformatics*, *39*(7), btad410.

Gu, Yu, Tinn, Robert, Cheng, Hao, Lucas, Michael, Usuyama, Naoto, Liu, Xiaodong, et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1–23.

Hermosilla, Pedro, & Ropinski, Timo (2022). Contrastive representation learning for 3d protein structures. arXiv preprint arXiv:2205.15675.

Hermosilla, Pedro, Schäfer, Marco, Lang, Matěj, Fackelmann, Gloria, Vázquez, Pere Pau, Kozlíková, Barbora, et al. (2021). Intrinsic-extrinsic convolution and pooling for learning on 3D protein structures. *International Conference on Learning Representations*.

Hu, Fan, Hu, Yishen, Zhang, Weihong, Huang, Huazhen, Pan, Yi, & Yin, Peng (2023). A multimodal protein representation framework for quantifying transferability across biochemical downstream tasks. *Advanced Science*, Article 2301223.

Hu, Bozhen, Tan, Cheng, Gao, Bin, Gao, Zhangyang, Wu, Lirong, Xia, Jun, et al. (2024). Multimodal distillation of protein sequence, structure, and function. URL https://openreview.net/forum?id=O0dW800ukz.

Ibtehaz, Nabil, Kagaya, Yuki, & Kihara, Daisuke (2023). Domain-PFP allows protein function prediction using function-aware domain embedding representations. *Communications Biology*, *6*(1), 1103.

Jing, Bowen, Eismann, Stephan, Suriana, Patricia, Townshend, Raphael John Lamarre, & Dror, Ron (2021). Learning from protein structure with geometric vector perceptrons. In *International conference on learning representations*.

Jones, Philip, Binns, David, Chang, Hsin-Yu, Fraser, Matthew, Li, Weizhong, McAnulla, Craig, et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240.

Kang, Yan, Wang, Xinchao, Xie, Cheng, Zhang, Huadong, & Xie, Wentao (2023). BBLN: A bilateral-branch learning network for unknown protein–protein interaction prediction. *Computers in Biology and Medicine*, *167*, Article 107588.

Karplus, Martin, & Kuriyan, John (2005). Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, *102*(19), 6679–6685.

Lee, Youhan, Yu, Hasun, Lee, Jaemyung, & Kim, Jaehoon (2023). Pre-training sequence, structure, and surface features for comprehensive protein representation learning. In *The twelfth international conference on learning representations*.

Li, Xinhui, Qian, Yurong, Hu, Yue, Chen, Jiaying, Yue, Haitao, & Deng, Lei (2024). MSF-pfp: A novel multisource feature fusion model for protein function prediction. *Journal of Chemical Information and Modeling*.

Liang, Victor Weixin, Zhang, Yuhui, Kwon, Yongchan, Yeung, Serena, & Zou, James Y. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, *35*, 17612–17625.

Lin, Zeming, Akin, Halil, Rao, Roshan, Hie, Brian, Zhu, Zhongkai, Lu, Wenting, et al. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. BioRxiv.

Liu, Quancheng, Zhang, Chengxin, & Freddolino, Lydia (2024). InterLabelGO+: unraveling label correlations in protein function prediction. *Bioinformatics*, *40*(11), btae655.

Meier, Joshua, Rao, Roshan, Verkuil, Robert, Liu, Jason, Sercu, Tom, & Rives, Alex (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, *34*, 29287–29303.

Melidis, Damianos P., & Nejdl, Wolfgang (2021). Capturing protein domain structure and function using self-supervision on domain architectures. *Algorithms*, *14*(1), 28.

Messih, Mario Abdel, Chitale, Meghana, Bajic, Vladimir B., Kihara, Daisuke, & Gao, Xin (2012). Protein domain recurrence and order can enhance prediction of protein functions. *Bioinformatics*, *28*(18), i444–i450.

Nguyen, Viet Thanh Duy, & Hy, Truong Son (2023). Multimodal pretraining for unsupervised protein representation learning. (pp. 2023–11), BioRxiv.

Notin, Pascal, Rollins, Nathan, Gal, Yarin, Sander, Chris, & Marks, Debora (2024). Machine learning for functional protein design. *Nature Biotechnology*, *42*(2), 216–228.

Pan, Shourun, Xia, Leiming, Xu, Lei, & Li, Zhen (2023). Submdta: drug target affinity prediction based on substructure extraction and multi-scale features. *BMC Bioinformatics*, *24*(1), 334.

Pawson, Tony, & Nash, Piers (2000). Protein–protein interactions define specificity in signal transduction. *Genes & Development*, *14*(9), 1027–1047.

Paysan-Lafosse, Typhaine, Blum, Matthias, Chuguransky, Sara, Grego, Tiago, Pinto, Beatriz Lázaro, Salazar, Gustavo A., et al. (2022). InterPro in 2022. *Nucleic Acids Research*, *51*(D1), D418–D427.

Poklukar, Petra, Vasco, Miguel, Yin, Hang, Melo, Francisco S., Paiva, Ana, & Kragic, Danica (2022). Geometric multimodal contrastive representation learning. In *International conference on machine learning* (pp. 17782–17800). PMLR.

Quan, Ruijie, Wang, Wenguan, Ma, Fan, Fan, Hehe, & Yang, Yi (2024). Clustering for protein representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 319–329).

Rao, Roshan M., Liu, Jason, Verkuil, Robert, Meier, Joshua, Canny, John, Abbeel, Pieter, et al. (2021). MSA transformer. In *International conference on machine learning* (pp. 8844–8856). PMLR.

Rives, Alexander, Meier, Joshua, Sercu, Tom, Goyal, Siddharth, Lin, Zeming, Liu, Jason, et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), Article e2016239118.

Rojano, Elena, Jabato, Fernando M., Perkins, James R., Córdoba-Caballero, José, García-Criado, Federico, Sillitoe, Ian, et al. (2022). Assigning protein function from domain-function associations using DomFun. *BMC Bioinformatics*, *23*(1), 1–19.

Schroff, Florian, Kalenichenko, Dmitry, & Philbin, James (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Su, Jin, Han, Chenchen, Zhou, Yuyang, Shan, Junjie, Zhou, Xibin, & Yuan, Fajie (2023). SaProt: Protein language modeling with structure-aware vocabulary. (pp. 2023–10), BioRxiv.

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

Tian, Yonglong, Sun, Chen, Poole, Ben, Krishnan, Dilip, Schmid, Cordelia, & Isola, Phillip (2020). What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, *33*, 6827–6839.

Torres, Mateo, Yang, Haixuan, Romero, Alfonso E., & Paccanaro, Alberto (2021). Protein function prediction for newly sequenced organisms. *Nature Machine Intelligence*, *3*(12), 1050–1060.

Wang, Zichen, Combs, Steven A., Brand, Ryan, Calvo, Miguel Romero, Xu, Panpan, Price, George, et al. (2022). Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific Reports*, *12*(1), 6832.

Wang, Tongzhou, & Isola, Phillip (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning* (pp. 9929–9939). PMLR.

Wang, Limei, Liu, Haoran, Liu, Yi, Kurtin, Jerry, & Ji, Shuiwang (2023). Learning hierarchical protein representations via complete 3d graph networks. In *International conference on learning representations*.

Wang, Xun, Qu, Peng, Meng, Xiangyu, Yang, Qing, Qiao, Lian, Zhang, Chaogang, et al. (2023). MulAxialGO: Multi-modal feature-enhanced deep learning model for protein function prediction. In *2023 IEEE international conference on bioinformatics and biomedicine* (pp. 132–137). IEEE.

Wang, Wenkang, Shuai, Yunyan, Zeng, Min, Fan, Wei, & Li, Min (2025). Dpfunc: accurately predicting protein function via deep learning with domain-guided structure information. *Nature Communications*, *16*(1), 70.

Wang, Zehan, Zhao, Yang, Huang, Haifeng, Liu, Jiageng, Yin, Aoxiong, Tang, Li, et al. (2023). Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, *36*, 22099–22114.

Watson, Joseph L., Juergens, David, Bennett, Nathaniel R., Trippe, Brian L., Yim, Jason, Eisenach, Helen E., et al. (2023). De novo design of protein structure and function with rfdiffusion. *Nature*, *620*(7976), 1089–1100.

Wu, Hongjie, Liu, Junkai, Jiang, Tengsheng, Zou, Quan, Qi, Shujie, Cui, Zhiming, et al. (2024). AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism. *Neural Networks*, *169*, 623–636.

Yan, Tian-Ci, Yue, Zi-Xuan, Xu, Hong-Quan, Liu, Yu-Hong, Hong, Yan-Feng, Chen, Gong-Xing, et al. (2023). A systematic review of state-of-the-art strategies for machine learning-based protein function prediction. *Computers in Biology and Medicine*, *154*, Article 106446.

Yao, Shuwei, You, Ronghui, Wang, Shaojun, Xiong, Yi, Huang, Xiaodi, & Zhu, Shanfeng (2021). Netgo 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Research*, *49*(W1), W469–W475.

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

You, Ronghui, Yao, Shuwei, Xiong, Yi, Huang, Xiaodi, Sun, Fengzhu, Mamitsuka, Hiroshi, et al. (2019). Netgo: improving large-scale protein function prediction with massive network information. *Nucleic Acids Research*, *47*(W1), W379–W387.

You, Ronghui, Zhang, Zihan, Xiong, Yi, Sun, Fengzhu, Mamitsuka, Hiroshi, & Zhu, Shanfeng (2018). Golabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, *34*(14), 2465–2473.

Zhang, Linlin, Ouyang, Chunping, Liu, Yongbin, Liao, Yiming, & Gao, Zheng (2023). Multimodal contrastive representation learning for drug-target binding affinity prediction. *Methods*, *220*, 126–133.

Zhang, Zuobai, Wang, Chuanrui, Xu, Minghao, Chenthamarakshan, Vijil, Lozano, Aurelie, Das, Payel, et al. (2023). A systematic study of joint representation learning on protein sequences and structures. arXiv preprint arXiv:2303.06275.

Zhang, Zuobai, Xu, Minghao, Jamasb, Arian, Chenthamarakshan, Vijil, Lozano, Aurelie, Das, Payel, et al. (2023). Protein representation learning by geometric structure pretraining. In *International conference on learning representations*.

Zhang, Zuobai, Xu, Minghao, Lozano, Aurelie C., Chenthamarakshan, Vijil, Das, Payel, & Tang, Jian (2024). Pre-training protein encoder via siamese sequence-structure diffusion trajectory prediction. *Advances in Neural Information Processing Systems*, *36*.

Zhang, Chengxin, Zheng, Wei, Freddolino, Peter L., & Zhang, Yang (2018). MetaGO: Predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein–protein network mapping. *Journal of Molecular Biology*, *430*(15), 2256–2265.

Zhou, Xiaogen, Zheng, Wei, Li, Yang, Pearce, Robin, Zhang, Chengxin, Bell, Eric W., et al. (2022). I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. *Nature Protocols*, *17*(10), 2326–2353.