RaFT-DM: A Residue-Aware Fusion Transformer with Domain-Wise Memory for Accurate Multi-Label Protein Function Prediction

Anonymous ACL submission

Abstract

Accurate protein function prediction via modeling protein sequences as long structured symbolic sequences is essential for understanding cellular mechanisms and guiding drug discovery. However, conventional deep learning models often use global pooling, which weakens domain-specific signals and fails to capture contextual dependencies across domains. Here, we propose **R**esidue-**a**ware **F**usion **T**ransformer with **D**omain-wise **M**emory (**RaFT-DM**), a segment-aware multimodal modeling framework that introduces explicit memory over structured segments. RaFT-DM performs residue-level cross-attention between sequence and structure embeddings, and segments the fused representation using InterProScan-derived domain annotations. A BiLSTM initialized with a global token then captures segment-level contextual semantics, preserving local discriminative features while modeling inter-segment relationships. Experiments on standard benchmarks show that RaFT-DM consistently outperforms state-of-the-art baselines. By replacing global pooling with domain-aware modeling, RaFT-DM reduces missed recalls and misclassifications, enabling more accurate and interpretable predictions. The implementation of RaFT-DM is available at `https://github.com/anonymous/RaFT-DM`.

1 Introduction

Proteins are essential molecular machines that catalyze reactions, mediate signaling, and support cell structure (Abdine et al., 2024; Alberts et al., 2015). Understanding their function is key to decoding cellular processes, disease mutations, and drug development (Eisenberg et al., 2000). While high-throughput sequencing has produced hundreds of millions of protein sequences but fewer than 1% are experimentally annotated using standard gene ontology (GO) terms across molecular functions (MF), cellular components (CC), and biological processes (BP) (The UniProt Consortium, 2023; Ashburner et al., 2000). Wet-lab validation remains the gold standard but is costly and slow (Rehm et al., 2013; Blay et al., 2020), creating a growing "annotation gap" where computational approaches are both necessary and advantageous (Derry et al., 2025; Bonetta and Valentino, 2020). From a modeling perspective, this annotation gap has motivated data-driven sequence representations, while raising concerns about how local functional structure is preserved under global aggregation.

Early function prediction relied on homology-based transfer (Altschul et al., 1990; Buchfink et al., 2015), but performance drops sharply for remote homologs. To address this, deep learning models were developed to learn from latent protein features. For instance, DeepGOPlus uses convolutional neural networks (CNNs) on sequences (Kulmanov and Hoehndorf, 2020), DeepFRI applies graph convolutional neural networks (GCNs) to structure (Gligorijević et al., 2021), and GAT-GO fuses sequence and structure via attention-based graphs (Lai and Xu, 2022). Newer methods extend this line: AnnoPRO learns multi-scale sequence encodings with long short-term memory (LSTM) decoders (Zheng et al., 2024); TransFun combines transformer embeddings with 3D-equivariant graph neural networks (GNNs) (Boadu and Cheng, 2024); and GNNGO3D uses residue-level graphs for hierarchical message passing (Zhang et al., 2024). However, these models often overlook the structural granularity and ordered dependencies of functional domains.

Domains are compact sequence–structure units (tens to hundreds of residues) that drive specific molecular tasks like binding or catalysis (Ponting and Russell, 2002; Bashton and Chothia, 2007). Each folds into a distinct 3D shape and often maps directly to specific GO terms. Domain loss or rearrangement frequently leads to functional loss. For instance, although TP53 and TP63 share over 70% sequence identity, TP63 lacks the TAD do-

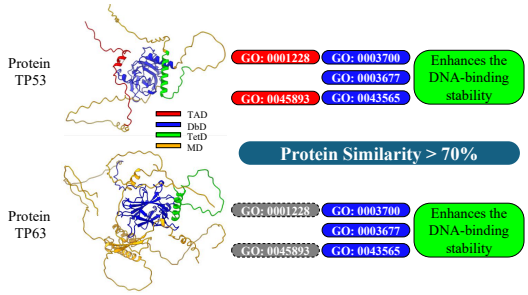

Figure 1: Functional domain–GO mapping for TP53 versus the TP63 ΔN isoform. Colored bars show TAD (red), DbD (blue), TetD (green) and MD (yellow), with same color linking each domain to its GO term. The absence of TAD in TP63 ΔN explains the loss of transcriptional activator activity (GO:0001228) and positive regulation of transcription (GO:0045893).

main of TP53 and thus cannot be annotated with GO:0001228 or GO:0045893 (Figure 1). This highlights the strong correspondence between domains and GO terms, underscoring the value of explicitly modeling domains. Yet many models treat domains as auxiliary features and use global pooling, which dilutes domain-specific signals and ignores domain order or context (Ulusoy and Doğan, 2024; Griseti et al., 2024). To address this, our work embeds domain boundaries directly into fusion and memory mechanisms, enabling the model to focus on and reason over these functional modules.

In this work, we propose Residue-aware Fusion Transformer with Domain-wise Memory (RaFT-DM), a novel multimodal framework that tightly integrates functional domain information into both representation learning and sequence modeling. Our contributions can be summarized as follows:

- We introduce a domain-aware sequence modeling framework that couples residue-level multimodal fusion with domain-wise memory, enabling structured representation learning beyond global pooling.
- We design a Transformer-based module that fuses sequence and structure embeddings at the residue level, using domain-guided attention masks to enhance sequence–structure interactions.
- We aggregate fused residue representations into a global context token, which initializes a BiLSTM to sequentially process domain-level features and capture both intra- and inter-domain dependencies.
- Extensive evaluation on the CAFA3 benchmark (Zhou et al., 2019) and large-scale UniProtKB datasets demonstrates that RaFT-DM significantly outperforms state-of-the-art methods, while providing biologically grounded interpretability.

2 Related Work

Sequence-Based Strategies Sequence-based methods are the earliest and most widely studied approach for protein function annotation. To move beyond simple homology transfer, machine learning-based models aggregate diverse sequence-derived signals into unified predictive frameworks. GOLabeler combines features like GO frequency, alignment scores, motifs, and domain counts using a learning-to-rank strategy (You et al., 2018). NetGO extends this by integrating protein–protein interaction networks, enhancing performance for proteins lacking close homologs (You et al., 2019). Despite these advances, such methods still depend on detectable homology and hand-crafted features. DeepGOPlus marks a shift to end-to-end learning, applying 1D CNNs to extract motifs directly from sequences and outperforming alignment-based baselines (Kulmanov and Hoehndorf, 2020). TALE further improves sequence-only prediction by embedding protein sequences and GO terms into a shared latent space, capturing global dependencies and label hierarchies without alignments (Cao and Shen, 2021). PFresGO introduces attention mechanisms to model GO term relationships during sequence encoding, showing strong performance on remote homologs (Pan et al., 2023).

Structure-Based Strategies As evidence grew that protein function depends on both sequence and 3D structure, researchers began incorporating structural information into prediction tasks. Initial efforts use experimental methods like X-ray crystallography and NMR, analyzing features such as active sites and fold similarity to infer function (Bartlett et al., 2002; Holm and Sander, 1996). These methods prove effective for proteins with low sequence identity, where alignment-based approaches failed. Tools like DALI and TM-align enable large-scale structure comparison but rely heavily on hand-crafted heuristics. The advent of AlphaFold2 (Jumper et al., 2021) transform the field by providing accurate predicted structures at scale, allowing deep learning models to exploit spatial information. Structure-based methods begin representing proteins as graphs, with residues as nodes and geometric relationships as edges. DeepFRI (Gligorijević et al., 2021) apply GCNs to contact maps, improving GO prediction for remote

homologs. GVP-GNN (Jing et al., 2020) introduce geometric vector perceptrons for encoding scalar and directional features in an equivariant manner. LM-GVP (Wang et al., 2021) combine pretrained sequence embeddings with GVP networks in a joint training setup, enabling effective sequence–structure fusion. TransFun (Boadu and Cheng, 2024) similarly integrates pretrained embeddings and AlphaFold2 structures using equivariant GNNs. PANDA-3D (Zhao et al., 2024) further combines GVP-GNN with a Transformer decoder, achieving improved function prediction.

Domain-Based Strategies Functional domains are evolutionarily conserved units that often correspond to distinct structural or functional modules. For decades, bioinformatics pipelines have used domain information for function annotation. Traditional approaches such as InterPro2GO (Hunter et al., 2009) and Pfam2GO (The Gene Ontology Consortium, 2025) map curated domain families to GO terms, enabling rapid annotation but failing to capture how combinations, order, or context of multiple domains influence function. With large-scale datasets, domain annotations have increasingly been incorporated into deep learning models, yielding stronger predictions for proteins with complex architectures (Ibtehaz et al., 2023; Wang et al., 2024, 2025). However, most methods still treat domains as independent units or rely on simple aggregation, without modeling sequential domain memory or inter-domain dependencies. As a result, important local discriminative features are easily diluted by global pooling, leading to information loss and reduced accuracy. To overcome these limitations, we propose RaFT-DM, a residue-aware fusion transformer with domain-wise memory that preserves both local domain features and their contextual relationships.

3 Method

In this section, we introduce RaFT-DM, a segment-aware sequence modeling framework for protein function prediction. RaFT-DM combines residue-level fusion with domain-wise memory, enabling the model to capture functional domain-specific information while overcoming the limitations of global pooling. An overview of the architecture is shown in Figure 2.

3.1 Preliminary

Proteins are heteropolymers composed of 20 canonical amino acids, whose sequence and 3D structure jointly determine biochemical function. The GO provides a structured vocabulary to describe such functions across three categories: MF, BP, and CC, typically formulated as multi-label annotations. A key unit in this sequence–function mapping is the functional domain—a contiguous residue segment that often folds independently and performs specific roles. Domains are curated in databases like Pfam (El-Gebali et al., 2021) and InterPro (Hunter et al., 2009), and exhibit a many-to-many relationship with GO terms: one domain may support multiple activities, while complex functions arise from cooperative interactions among multiple domains. As shown in Figure 3, a single domain may underlie various oxidoreductase activities, whereas more complex roles emerge from domain coordination within defined spatial contexts.

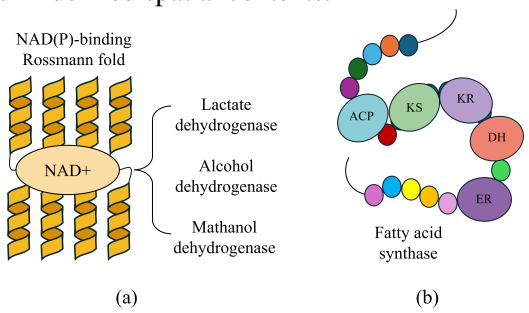

Figure 3: Functional implications of protein domain organization. **(a)** A single domain accommodates NAD^+ binding and enables diverse oxidoreductase functions. **(b)** Multiple domains function cooperatively in a defined spatial arrangement to achieve complex catalysis (Fatty acid synthase).

Rather than mapping directly from raw sequence or structure to function, we treat domains as intermediates linking molecular representations to GO terms. Our model captures both individual domain roles and their context-dependent cooperation, enhancing prediction accuracy and interpretability.

3.2 Sequence Encoding

We map the raw amino-acid sequence $s = (a_1, a_2, \ldots, a_L)$, $a_i \in \mathcal{A} = \{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}$, to continuous representations using the ESM-2 (Lin et al., 2023) model. To obtain a single, task-oriented summary while retaining residue-level detail, we prepend a dedicated <P_ABS> (**Protein Abstraction**) token to every sequence and treat its final hidden state as a global embedding.

After tokenisation, the input becomes (<P_ABS>, $a_1, \ldots, a_L$). Following positional and residue embeddings, the transformer yields layer-wise activations $\mathbf{H}^{(\ell)} = \left[h^{(\ell)}_{\text{P_ABS}}, h^{(\ell)}_1, \ldots, h^{(\ell)}_L\right] \in$

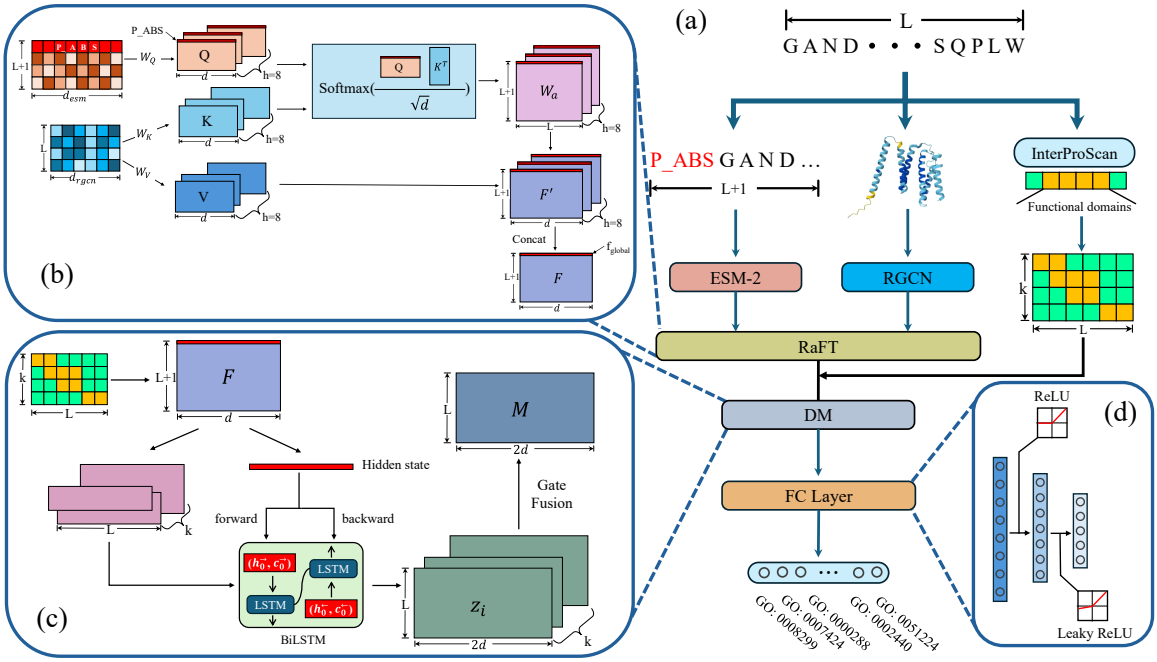

Figure 2: Model architecture of RaFT-DM. **(a)** Overview: the protein sequence and its 3D structure are separately embedded as sequence and structure embeddings; InterProScan generates binary domain masks; RaFT-DM (**(b), (c)**) fuses sequence embeddings, structure features, and domain masks for function prediction. **(b)** The details of RaFT. Sequence and graph embeddings are projected into a shared space and fused via residue-level cross-modal attention with a global token (<P_ABS>) to produce a structure-aware global vector. **(c)** The details of DM. The fused residue embeddings are segmented by domain masks and fed into a shared BiLSTM initialized with the global vector; domain outputs are gated and concatenated into the final protein representation. **(d)** Fully connected layer with ReLU and Leaky ReLU for generating final multi-label predictions.

$\mathbb{R}^{(L+1)\times 1280}, \ell = 1, \ldots, 33$. From the top layer ($\ell = 33$) we obtain $q_s = h^{(33)}_{\text{P_ABS}} \in \mathbb{R}^{1280}, x_i = h^{(33)}_i \in \mathbb{R}^{1280}, \quad i = 1, \ldots, L$, where q_s is the global sequence embedding and $\{x_i\}$ are the residue embeddings.

The ESM-2 backbone is kept *frozen*; only the <P_ABS> embedding and the projection matrices are updated during training. This strategy preserves the rich language-model prior while markedly reducing GPU memory consumption.

The resulting set $\mathbf{X}_{\text{raw}} = \{\tilde{q}_0, \tilde{x}_1, \ldots, \tilde{x}_L\}$ serves as the sequence modality input to the cross-modal fusion module detailed in Section Cross-modal Fusion.

3.3 Graph Construction and Encoding

To model protein structures, we represent each protein as a residue-level heterogeneous graph $G = (V, E)$ built from AlphaFoldDB-predicted 3D structures. Structure files are retrieved using UniProt IDs and renamed by accession numbers. Nodes V represent residues; edges E encode intra-molecular interactions.

Each node $v_i \in V$ is described by a feature vector $\mathbf{N}_i \in \mathbb{R}^d$, combining amino acid one-hot encoding, Meiler embeddings, Expasy physicochemical scales, hydrogen bond donor/acceptor flags, and secondary structure descriptors (torsion angles ϕ_i, ψ_i, ASA, RSA, and DSSP-derived class labels):

$$\mathbf{N}_i = [\, \text{AA}_{\text{onehot}},\ \text{Meiler},\ \text{Expasy},\ \text{Hbond}_{\text{donor}},\ \text{Hbond}_{\text{acceptor}},\ \phi_i,\ \psi_i,\ \text{ASA}_i,\ \text{RSA}_i,\ \text{SS}_i \,]$$

Edges are defined based on peptide bonds, hydrogen bonds, and spatial proximity (< 10), excluding non-informative contacts (< 3). Each edge is labeled by interaction type and encoded as a multi-hot vector, resulting in a heterogeneous graph capturing diverse residue-level interactions.

3.4 Functional Domain Construction

We extract functional domains from the primary amino acid sequence using InterProScan. Each functional domain is represented by a triplet (l_k, r_k, δ_k), where $1 \le l_k \le r_k \le L$ mark the start and end residues, and δ_k denotes the Pfam/InterPro identifier. Each hit returned by InterProScan carries an expectation (E)-value estimating the probability

of a random match. To ensure high confidence, hits with an E-value greater than 10^{-5} are discarded. The remaining domains form the set

$$\mathcal{D} = \{(l_k, r_k, \delta_k)\}_{k=1}^{K}.$$

In practice, domains in $\mathcal{D}$ may overlap because a residue segment can participate in multiple structural or functional modules. We therefore retain overlapping intervals, allowing residues to inherit all associated domain tags and enabling the model to capture shared semantics and multifunctional regions. This reflects the fact that domain boundaries are often ambiguous and evolutionarily coupled.

Some proteins or long regions have no InterProScan hits ($|\mathcal{D}| = 0$). To maintain a consistent memory structure, we assign two *pseudo-domains* spanning the full sequence,

$$\mathcal{D}_0 = \{[1, L]_1,\ [1, L]_2\},$$

which use separate memory parameters to provide multiple latent views of the same sequence. Their influence is down-weighted by a coefficient $\lambda_0 < 1$ to avoid overemphasizing redundant signals.

The complete set of domains, including the pseudo-domain, is expressed as $\{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_K\}$. We then collapse these onto the residue axis to obtain a binary indicator mask:

$$m_i = \begin{cases} 1, & \text{if } i \in [l_k, r_k] \text{ for any } k, \\ 0, & \text{otherwise.} \end{cases}$$

This mask and the associated domain indices are used to guide downstream aggregation, enabling RaFT-DM to focus on functionally relevant regions while respecting both domain co-occupancy and unannotated residues.

3.5 Residual-aware Cross-modal Fusion (RaFT-DM)

Building on the theoretical foundation above, we embed the information-theoretic motivation directly into our architecture through a **residue-aware cross-modal fusion framework**, termed **RaFT-DM**. The design follows the hierarchical Information Bottleneck (IB) principle: the fused representation should *retain* domain-relevant information for functional prediction ($I(Z; Y)$) while *compressing* redundant or noisy sequence and structure details ($I(Z; X, G)$). Here, X and G denote sequence and structure embeddings; $Z = z_k{}_{k=1}^{K}$ are domain-level latent variables after compression, and Y is the ground-truth multi-label annotation.

Input and Projection. RaFT-DM takes residue-level embeddings from ESM-2 and structure-aware embeddings from a heterogeneous protein graph processed by an RGCN (Schlichtkrull et al., 2018). The raw sequence representation is

$$\mathbf{X}_{\text{raw}} = [\mathbf{q}_0, \mathbf{x}_1, \ldots, \mathbf{x}_L] \in \mathbb{R}^{(L+1)\times d_{\text{esm}}},$$

where $\mathbf{q}_0$ denotes a global abstraction token. The structural embedding is

$$\mathbf{G}_{\text{raw}} = [\mathbf{g}_1, \ldots, \mathbf{g}_N] \in \mathbb{R}^{N\times d_{\text{rgcn}}}.$$

Both modalities are projected into a shared Hilbert space $\mathcal{H}$:

$$\mathbf{Q} = \Pi_{\mathcal{H}}(\mathbf{X}_{\text{raw}}; \mathbf{W}_q),$$
$$\mathbf{K} = \Pi_{\mathcal{H}}(\mathbf{G}_{\text{raw}}; \mathbf{W}_k),$$
$$\mathbf{V} = \Pi_{\mathcal{H}}(\mathbf{G}_{\text{raw}}; \mathbf{W}_v),$$

where $\Pi_{\mathcal{H}}$ denotes a linear projection. This interpretation views the attention score $\langle \mathbf{q}_i, \mathbf{k}_j \rangle_{\mathcal{H}}$ as a geometric affinity term measuring residue–structure compatibility under a unified manifold.

Residue-aware Cross-modal Attention. Each residue query attends to all structural nodes via scaled dot-product attention:

$$\mathbf{f}_i = \sum_{j=1}^{N} \alpha_{ij} \mathbf{g}_j,$$

$$\alpha_{ij} = \frac{\exp(\mathbf{x}_i^\top \mathbf{g}_j / \sqrt{d})}{\sum_{j'} \exp(\mathbf{x}_i^\top \mathbf{g}_{j'} / \sqrt{d})}.$$

This yields $\mathbf{F} = [\mathbf{f}_0, \ldots, \mathbf{f}_L] \in \mathbb{R}^{(L+1)\times d}$, where $\mathbf{f}_0$ (the fused global token) summarizes overall structural context. From an IB standpoint, this attention operation functions as a soft routing mechanism that selectively transmits structure-derived information to residues most relevant for functional prediction.

Domain-wise Memory for Hierarchical Compression. Functional domains naturally induce local information bottlenecks. Given a domain segmentation $\mathcal{D} = \{[l_k, r_k]\}_{k=1}^{K}$, RaFT-DM groups fused residue features into domain sequences:

$$\mathbf{F}_k = \mathbf{F}[l_k : r_k, :].$$

Each $\mathbf{F}_k$ is processed by a shared BiLSTM initialized with the global token $\mathbf{f}_{\text{global}} = \mathbf{f}_0$:

$$\mathbf{M}_k = \text{BiLSTM}(\mathbf{F}_k; \mathbf{f}_{\text{global}}).$$

The BiLSTM captures intra-domain residue dependencies and inter-domain contextual transitions. The domain representation is then obtained via average pooling:

$$\mathbf{z}_k = \frac{1}{r_k - l_k + 1} \sum_{i=l_k}^{r_k} \mathbf{m}_i.$$

Global–Domain Gating and Mutual Information Balancing. To regulate the contribution of global versus domain-local information, RaFT-DM employs a gating mechanism:

$$\mathbf{z}_k^{\text{fused}} = \sigma\big(\mathbf{W}_g[\mathbf{z}_k; \mathbf{f}_{\text{global}}]\big) \odot \mathbf{z}_k + \big(1 - \sigma(\cdot)\big) \odot \mathbf{f}_{\text{global}}.$$

This gate implicitly balances $I(Z;Y)$ and $I(Z;X,G)$, assigning wider bandwidth to domains with stronger functional relevance.

Finally, the protein-level representation is constructed by concatenating all fused domain embeddings:

$$\mathbf{M} = [\mathbf{z}_1^{\text{fused}}, \ldots, \mathbf{z}_K^{\text{fused}}].$$

This cross-modal fusion strategy enables residue-level interaction between sequence and structure while preserving domain-wise functional semantics, yielding a unified representation that is both biologically grounded and predictive of complex molecular functions.

4 Experiment and results

4.1 Dataset and Preprocessing

To train the RaFT-DM model, we used data from two main sources: UniProtKB/Swiss-Prot and the CAFA3 benchmark. Protein sequences were obtained from UniProtKB/Swiss-Prot (release: November 27, 2024), comprising 572,214 reviewed entries. After excluding proteins without GO annotations, 549,763 sequences remained. We retained only GO terms occurring at least 50 times and supported by strong evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS, IC, HEP). Sequences were further filtered to lengths between 100 and 1,022 residues to meet ESM-2 input constraints. Functional domains were annotated via InterProScan, retaining hits with E-value $\leq 1 \times 10^{-5}$. For structural input, experimentally determined 3D structures from the Protein Data Bank (PDB) were used whenever available; otherwise, models predicted by AlphaFold 2 were employed. Structures were converted to residue-level graphs by linking C_α atoms within a 10 Å radius to encode spatial proximity.

For benchmarking, we used CAFA3 (Zhou et al., 2019), a widely adopted dataset for GO prediction. It includes 66,841 annotated proteins for training and held-out test sets covering 2,145 BP, 1,101 MF, and 977 CC targets, spanning 3,992 BP, 677 MF, and 551 CC terms. Sequences exceeding 1,022 residues were split into fragments, with their embeddings concatenated to form complete representations. Dataset statistics for MF, BP, and CC tasks across UniProtKB/Swiss-Prot and CAFA3 are shown in Table 1.

Dataset	Split	MF	BP	CC
UniProtKB	Train	60,555	43,535	38,172
	Val	4,365	3,872	3,783
	Test	5,359	4,788	4,657
CAFA3	Train	24,578	38,382	34,503
	Val	2,264	3,428	3,318
	Test	1,101	2,145	977

Table 1: Statistics of the splits for MF, BP, and CC tasks of two datasets.

4.2 Baselines

To assess RaFT-DM's performance, we compared it against a wide range of baselines, including sequence alignment tools (BLAST (Altschul et al., 1990), DIAMOND (Buchfink et al., 2015)), sequence-based deep learning models (DeepGOPlus (Kulmanov and Hoehndorf, 2020), SPROF-GO (Yuan et al., 2023), PO2GO (Li et al., 2024), AnnoPRO (Zheng et al., 2024)), and multimodal or structure-aware methods (DeepFRI (Gligorijević et al., 2021), HNetGO (Zhang et al., 2023), TransFun (Boadu and Cheng, 2024), PANDA-3D (Zhao et al., 2024)). BLAST and DIAMOND infer GO terms via sequence similarity; DeepGOPlus adds CNN features; SPROF-GO, PO2GO, and AnnoPRO use language models and GO-aware representations. DeepFRI utilizes structural graphs; HNetGO incorporates similarity networks and PPIs; TransFun and PANDA-3D combine transformer embeddings with AlphaFold structures. All methods were evaluated using the same data splits as RaFT-DM for fair comparison.

4.3 Evaluation metrics

We evaluated RaFT-DM on the CAFA3 dataset using two core metrics: $F_{\max}$ (Radivojac et al., 2013) and area under the precision–recall curve (AUPRC) (Davis and Goadrich, 2006). $F_{\max}$ represents the highest harmonic mean of precision and recall attained over all decision thresholds, while AUPRC more appropriately reflects performance on sparse, multi-label data than receiver operating characteristic-based measures.

4.4 Comparison with existing methods

To evaluate the effectiveness of RaFT-DM, we compared it against different methods on two widely adopted datasets—UniProtKB and CAFA3. All values are averaged over five runs with seeds $\{0, 1, 2, 3, 42\}$, and p-values are computed using paired t-tests. On the UniProtKB dataset (Table 2), RaFT-DM achieves $F_{\max}$ scores of 0.722, 0.596, and 0.743 on MF, BP, and CC, respectively. Com-

Method	F_{max} ↑						AUPRC ↑					
	MF	p-value ↓	BP	p-value ↓	CC	p-value ↓	MF	p-value ↓	BP	p-value ↓	CC	p-value ↓
BLAST	0.368	–	0.312	–	0.479	–	0.209	–	0.238	–	0.275	–
Diamond	0.323	–	0.223	–	0.335	–	0.278	–	0.276	–	0.367	–
DeepGoPlus	0.551 ± 0.006	5.60×10^{-7}	0.508 ± 0.005	6.80×10^{-6}	0.558 ± 0.004	4.80×10^{-8}	0.446 ± 0.002	4.80×10^{-9}	0.426 ± 0.006	2.30×10^{-7}	0.516 ± 0.005	4.70×10^{-8}
DeepFRI	0.613 ± 0.006	1.90×10^{-6}	0.494 ± 0.003	1.90×10^{-6}	0.630 ± 0.006	2.80×10^{-6}	0.577 ± 0.003	7.70×10^{-9}	0.487 ± 0.008	8.80×10^{-6}	0.583 ± 0.004	5.60×10^{-8}
HNetGO	0.556 ± 0.004	1.10×10^{-7}	0.475 ± 0.003	3.30×10^{-7}	0.602 ± 0.005	2.50×10^{-7}	0.547 ± 0.004	3.10×10^{-7}	0.486 ± 0.002	2.10×10^{-7}	0.598 ± 0.001	3.70×10^{-9}
SPROF-GO	0.507 ± 0.004	4.20×10^{-8}	0.312 ± 0.001	1.00×10^{-9}	0.543 ± 0.002	2.20×10^{-8}	0.496 ± 0.002	1.50×10^{-8}	0.297 ± 0.003	1.60×10^{-9}	0.586 ± 0.002	8.90×10^{-10}
TransFun	0.581 ± 0.003	1.50×10^{-7}	0.513 ± 0.004	6.60×10^{-7}	0.612 ± 0.003	8.50×10^{-8}	0.448 ± 0.003	7.80×10^{-9}	0.404 ± 0.004	5.00×10^{-8}	0.492 ± 0.005	2.70×10^{-8}
PO2GO	$\underline{0.645 \pm 0.002}$	1.90×10^{-7}	0.562 ± 0.002	2.00×10^{-5}	$\underline{0.676 \pm 0.002}$	6.40×10^{-8}	0.615 ± 0.004	2.80×10^{-6}	$\underline{0.581 \pm 0.005}$	7.90×10^{-3}	0.619 ± 0.003	2.40×10^{-7}
PANDA-3D	0.587 ± 0.005	6.50×10^{-7}	$\underline{0.567 \pm 0.005}$	4.30×10^{-4}	0.610 ± 0.006	1.30×10^{-6}	0.600 ± 0.003	1.60×10^{-8}	0.577 ± 0.006	2.60×10^{-3}	0.663 ± 0.002	2.10×10^{-7}
AnnoPRO	0.589 ± 0.004	2.70×10^{-7}	0.545 ± 0.003	1.80×10^{-5}	0.625 ± 0.003	5.60×10^{-8}	$\underline{0.660 \pm 0.004}$	3.10×10^{-5}	0.550 ± 0.006	6.60×10^{-5}	0.695 ± 0.001	1.70×10^{-7}
DPFunc	0.611 ± 0.006	3.00×10^{-6}	0.551 ± 0.003	3.10×10^{-5}	0.664 ± 0.004	9.80×10^{-7}	0.603 ± 0.003	9.00×10^{-8}	0.563 ± 0.003	7.00×10^{-5}	$\underline{0.707 \pm 0.005}$	3.30×10^{-5}
RaFT-DM	$\mathbf{0.722 \pm 0.001}$	–	$\mathbf{0.595 \pm 0.002}$	–	$\mathbf{0.743 \pm 0.001}$	–	$\mathbf{0.695 \pm 0.001}$	–	$\mathbf{0.592 \pm 0.002}$	–	$\mathbf{0.752 \pm 0.003}$	–

Table 2: Performance comparison on the UniProt dataset in terms of F_{max} and AUPRC across MF, BP, and CC. Best results are in **bold**, second-best are underlined.

Method	F_{max} ↑						AUPRC ↑					
	MF	p-value ↓	BP	p-value ↓	CC	p-value ↓	MF	p-value ↓	BP	p-value ↓	CC	p-value ↓
BLAST	0.227	–	0.245	–	0.313	–	0.150	–	0.237	–	0.251	–
Diamond	0.142	–	0.062	–	0.145	–	0.252	–	0.143	–	0.302	–
DeepGoPlus	0.420 ± 0.005	1.47×10^{-7}	0.390 ± 0.005	3.80×10^{-6}	0.582 ± 0.003	2.00×10^{-6}	0.356 ± 0.006	8.77×10^{-8}	0.320 ± 0.006	2.26×10^{-7}	0.416 ± 0.007	9.14×10^{-8}
DeepFRI	0.491 ± 0.004	3.38×10^{-7}	0.459 ± 0.002	1.71×10^{-4}	0.582 ± 0.005	5.00×10^{-6}	0.537 ± 0.003	2.50×10^{-7}	0.440 ± 0.007	3.26×10^{-6}	0.639 ± 0.003	3.77×10^{-6}
HNetGO	0.484 ± 0.005	2.50×10^{-7}	0.433 ± 0.002	4.25×10^{-6}	0.597 ± 0.002	2.00×10^{-6}	0.487 ± 0.006	4.72×10^{-7}	0.420 ± 0.009	5.69×10^{-6}	0.604 ± 0.004	1.23×10^{-6}
SPROF-GO	0.525 ± 0.011	7.20×10^{-8}	0.450 ± 0.003	6.66×10^{-5}	0.592 ± 0.007	6.10×10^{-5}	0.560 ± 0.004	8.63×10^{-7}	0.426 ± 0.003	3.49×10^{-7}	0.641 ± 0.005	5.69×10^{-5}
TransFun	0.451 ± 0.003	1.56×10^{-8}	0.396 ± 0.003	4.18×10^{-7}	$\underline{0.608 \pm 0.002}$	1.00×10^{-6}	0.447 ± 0.005	3.55×10^{-7}	0.389 ± 0.004	1.48×10^{-7}	0.600 ± 0.003	5.82×10^{-7}
PO2GO	$\underline{0.587 \pm 0.003}$	7.17×10^{-6}	0.474 ± 0.003	0.351	0.600 ± 0.003	1.10×10^{-5}	0.591 ± 0.005	3.36×10^{-4}	0.411 ± 0.006	1.41×10^{-6}	0.584 ± 0.003	9.75×10^{-8}
PANDA-3D	0.494 ± 0.003	1.44×10^{-7}	0.442 ± 0.004	6.07×10^{-5}	0.583 ± 0.005	4.00×10^{-6}	0.504 ± 0.006	1.77×10^{-6}	0.421 ± 0.005	1.11×10^{-6}	0.619 ± 0.004	3.64×10^{-6}
AnnoPRO	0.527 ± 0.002	3.37×10^{-8}	$\mathbf{0.479 \pm 0.006}$	0.208	0.602 ± 0.003	6.00×10^{-6}	0.546 ± 0.004	9.65×10^{-6}	0.459 ± 0.002	1.12×10^{-7}	0.645 ± 0.003	9.29×10^{-6}
DPFunc	0.573 ± 0.006	9.00×10^{-5}	0.471 ± 0.003	2.86×10^{-2}	0.601 ± 0.002	2.00×10^{-6}	$\underline{0.596 \pm 0.006}$	2.07×10^{-3}	$\underline{0.526 \pm 0.011}$	7.34×10^{-2}	$\underline{0.654 \pm 0.007}$	1.20×10^{-3}
RaFT-DM	$\mathbf{0.614 \pm 0.002}$	–	$\underline{0.475 \pm 0.003}$	–	$\mathbf{0.649 \pm 0.002}$	–	$\mathbf{0.618 \pm 0.002}$	–	$\mathbf{0.538 \pm 0.001}$	–	$\mathbf{0.684 \pm 0.006}$	–

Table 3: Performance comparison on the CAFA3 dataset in terms of F_{max} and AUPRC across MF, BP, and CC. Best results are in **bold**, second-best are underlined.

pared with strong deep-learning baselines, RaFT-DM improves over PO2GO by 0.076, 0.032, and 0.067 F_{max} points on the three ontologies, and substantially outperforms earlier sequence and structure based models such as DeepGoPlus and DeepFRI. Ranking performance exhibits a similar trend: RaFT-DM attains AUPRC values of 0.695 (MF), 0.592 (BP), and 0.752 (CC), exceeding AnnoPRO (0.660, 0.550, 0.695) by 0.035, 0.042, and 0.057, while also surpassing the domain-aware baseline DPFunc on most ontologies. These gains indicate that modeling functional domains explicitly, together with their residue-level sequential context, provides complementary information beyond global sequence similarity and purely structural reasoning.

On the CAFA3 benchmark (Table 3), RaFT-DM shows similarly consistent improvements. In MF and CC, RaFT-DM outperforms recent deep-learning and domain-aware baselines, including AnnoPRO, PO2GO, TransFun, and DPFunc, with F_{max} increases around 4–9 percent and AUPRC gains up to 13 percent. This is in line with the nature of MF and CC annotations. Many MF terms are tightly associated with local sequence motifs and catalytic sites that can be effectively captured by language model embeddings and structural descriptors, while CC annotations reflect spatial organization and complex assembly patterns that are directly encoded by our graph-based structure module. BP remains the most challenging ontology: RaFT-DM attains only modest improvements in F_{max} and is close to PO2GO and AnnoPRO, yet still achieves a notable 0.012 AUPRC gain over AnnoPRO, which suggests better recall on rare or deeply nested BP terms. This is consistent with the fact that BP annotations often depend on protein–protein interactions and context specific networks that are only partially observable from sequence and static structure, giving an advantage to methods that integrate explicit interaction or expression information. Taken together, experiments on UniProtKB and CAFA3 demonstrate that RaFT-DM is a robust and generalizable method for protein function prediction. By consistently outperforming classical, deep-learning, and state-of-the-art models, RaFT-DM shows that explicitly modeling functional domains and their sequential context enhances domain-specific signal retention, reducing false predictions and improving recall and precision across all GO ontologies.

4.5 Ablation and Mechanistic Validation of RaFT-DM

To quantify the contribution of each architectural component, we performed an ablation study on CAFA3 using six variants: sequence-only, graph-only, sequence+DM, graph+DM, RaFT without DM, and the full RaFT-DM model (Table 4). Consistent with prior CAFA3 studies (Radivojac et al., 2013), these variants yield a clean, standardized comparison across MF, BP, and CC ontologies.

Adding DM to either modality leads to substantial gains in both F_{max} and AUPRC, with the

sequence+DM variant outperforming graph+DM. This indicates that, while structural information helps constrain representation locality, the domain-wise memory provides a more direct inductive bias for capturing functional regions. RaFT alone also improves performance over single-modality baselines, but its impact is weaker than that of DM, suggesting that without explicit domain structure, the fused representation still tends to behave like a global pooling operator, limiting its ability to isolate functionally critical residues.

To further illustrate this behavior, Figure 4 visualizes residue-level activation maps for three well-characterized proteins (PTPN11, LCK, and AKT1). Under the RaFT-only baseline, activations appear diffuse and loosely correlated with annotated domains. In contrast, RaFT-DM produces sharp, high-intensity activation bands that align closely with domain boundaries (highlighted in red). These qualitative patterns demonstrate that domain-wise memory guides the model to emphasize a compact set of residues within each functional domain, leading to more interpretable and biologically consistent predictions.

Components				F_{max} ↑			AUPRC ↑		
Seq	Graph	RaFT	DM	MF	BP	CC	MF	BP	CC
✓				0.467 ± 0.005	0.410 ± 0.006	0.528 ± 0.004	0.423 ± 0.007	0.364 ± 0.008	0.404 ± 0.009
	✓			0.477 ± 0.007	0.426 ± 0.008	0.545 ± 0.008	0.449 ± 0.009	0.407 ± 0.008	0.536 ± 0.008
✓		✓		0.532 ± 0.005	0.461 ± 0.006	0.581 ± 0.006	0.510 ± 0.008	0.481 ± 0.009	0.565 ± 0.009
	✓	✓		0.547 ± 0.008	0.454 ± 0.008	0.557 ± 0.008	0.552 ± 0.007	0.388 ± 0.009	0.501 ± 0.009
✓	✓	✓		0.556 ± 0.006	0.465 ± 0.003	0.589 ± 0.006	0.558 ± 0.009	0.412 ± 0.01	0.539 ± 0.008
✓	✓	✓	✓	$\mathbf{0.614 \pm 0.002}$	$\mathbf{0.475 \pm 0.003}$	$\mathbf{0.649 \pm 0.002}$	$\mathbf{0.618 \pm 0.002}$	$\mathbf{0.538 \pm 0.001}$	$\mathbf{0.684 \pm 0.006}$

Table 4: Ablation results of RaFT-DM on CAFA3. Checkmarks indicate included components.

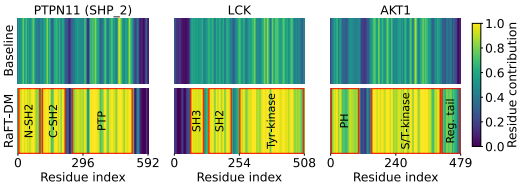

Figure 4: Residue-level contribution heatmaps for PTPN11, LCK, and AKT1 under the baseline (top) and RaFT-DM (bottom). RaFT-DM produces sharp, domain-aligned activation bands (**red boxes**), whereas the baseline responses remain diffuse. This illustrates that domain-wise memory guides the model to focus on a compact set of functionally relevant residues within each annotated domain.

4.6 RaFT-DM captures functional domain information

We further analyzed two proteins: PTPN11 (Q06124) and GRB2 (P62993), which are not included in our dataset. This case study demonstrates that RaFT-DM truly learns contextual dependencies among functional domains (as shown in Figure 5). Both proteins exhibit SH2-mediated binding (GO:0019904). Beyond this shared function, PTPN11 features tandem SH2 domains upstream of a PTP catalytic domain, which by itself offers only the dephosphorylation active site rather than any binding capability. Phosphotyrosine–peptide recognition by SH2 (GO:0019901) relieves autoinhibition and unleashes the PTP fold to execute protein tyrosine phosphatase activity (GO:0004725). In contrast, GRB2 contains a lone SH2 domain and thus exclusively engages phosphotyrosine-marked receptors or kinases (GO:0019901) without any downstream catalytic function. When applied to these proteins, RaFT-DM's cross-modal attention consistently highlights the SH2 regions for GO:0019904, yet selectively allocates attention to the PTP catalytic region only in PTPN11, accurately predicting its phosphatase activity. This behavior confirms that RaFT-DM not only recognizes individual domain signatures but also dynamically interprets their interplay—capturing inter-domain semantic dependencies to yield context-aware function predictions.

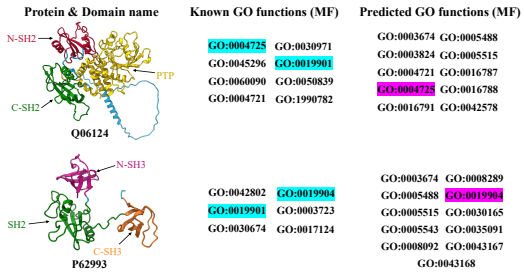

Figure 5: Context-dependent attention of RaFT-DM on two SH2-fold proteins. PTPN11 (Q06124) and GRB2 (P62993) both share SH2-mediated binding (GO:0019904), but only PTPN11 has a PTP catalytic domain conferring catalytic activity (GO:0004725). Ground-truth GO terms are shown in cyan, and RaFT-DM's predictions are highlighted in magenta, correctly assigning catalytic activity to PTPN11 and binding to GRB2.

5 Conclusion

In this work, we propose RaFT-DM, a segment-aware sequence modeling framework that integrates residue-level cross-modal fusion with domain-wise memory. By introducing explicit segment-level inductive bias, RaFT-DM preserves local discriminative signals while modeling long-range contextual dependencies. Extensive experiments demonstrate consistent improvements in predictive performance and interpretability over global pooling and flat fusion strategies.

6 Limitations

RaFT-DM relies on externally provided segment annotations to guide domain-wise memory construction. While functional domain boundaries offer a biologically grounded form of segmentation, they may be incomplete, noisy, or unavailable for certain sequences. In such cases, the model degrades to coarser segment-level representations, which may reduce its ability to localize fine-grained functional signals. Additionally, domain-wise memory introduces additional computational cost compared to flat pooling strategies due to segment-level processing. However, this overhead is modest relative to the overall model complexity and is offset by consistent gains in predictive accuracy, robustness, and interpretability. In high-value application scenarios where reliable predictions are critical, such as large-scale protein annotation or downstream target discovery, this trade-off is often acceptable. Future work may explore more efficient implementations of segment-aware memory to further improve scalability.

7 Ethical Considerations

This work focuses on methodological advances in long-sequence modeling and evaluates the proposed framework on protein sequences as a representative structured symbolic domain. The model is trained and evaluated exclusively on publicly available protein sequence and structure data, and does not involve human subjects, personal data, or user-generated content.

We acknowledge that methods which lower technical barriers in protein analysis could, in principle, be applied beyond their original intent. However, RaFT-DM is a modeling framework rather than an automated design or deployment system, and it does not directly enable the synthesis, modification, or application of biological entities.

We therefore emphasize the importance of responsible use and interpretation of model predictions. We release our implementation to promote transparency and reproducibility, and we encourage downstream users to follow established ethical guidelines and best practices when applying computational predictions in real-world scientific or engineering contexts.

References

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2024. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10757–10765.

Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2015. *Essential cell biology*. Garland Science.

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

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, and 1 others. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Gail J. Bartlett, Craig T. Porter, Neera Borkakoti, and Janet M. Thornton. 2002. Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, 324(1):105–121.

Matthew Bashton and Cyrus Chothia. 2007. The generation of new protein functions by the combination of domains. *Structure*, 15(1):85–99.

Vincent Blay, Bhairavi Tolani, Sunita P Ho, and Michelle R Arkin. 2020. High-throughput screening: today's biochemical and cell-based approaches. *Drug Discovery Today*, 25(10):1807–1821.

Frimpong Boadu and Jianlin Cheng. 2024. Improving protein function prediction by learning and integrating representations of protein sequences and function labels. *Bioinformatics Advances*, 4(1):vbae120.

Rosalin Bonetta and Gianluca Valentino. 2020. Machine learning techniques for protein function prediction. *Proteins: Structure, Function, and Bioinformatics*, 88(3):397–413.

Benjamin Buchfink and 1 others. 2015. Fast and sensitive protein alignment using diamond. *Nature Methods*, 12(1):59–60.

Yue Cao and Yang Shen. 2021. Tale: Transformer-based protein function annotation with joint sequence–label embedding. *Bioinformatics*, 37(18):2825–2833.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Alexander Derry, Alp Tartici, and Russ B. Altman. 2025. Protein functional site annotation using local structure embeddings. *bioRxiv*.

David Eisenberg, Edward M Marcotte, Ioannis Xenarios, and Todd O Yeates. 2000. Protein function in the post-genomic era. *Nature*, 405(6788):823–826.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lorna Williams, Sergey Chuguransky, Livia Paladin, Sugandha Raj, Lorna J. Richardson, Robert D. Finn, and 1 others. 2021. The pfam protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419.

Sai Teja Erukude, Suhasnadh Reddy Veluru, and Viswa Chaitanya Marella. 2025. Multimodal deep learning: A survey of models, fusion strategies, applications, and research challenges. *International Journal of Computer Applications*, 975:8887.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, and 1 others. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168.

Elena Griseti, Abdoul Akim Bello, Eric Bieth, Bayane Sabbagh, Jason S. Iacovoni, Joëlle Bigay, Henrik Laurell, Alenka Čopič, and 1 others. 2024. Molecular mechanisms of perilipin protein function in lipid droplet metabolism. *FEBS Letters*, 598(10):1170–1198.

Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, and 1 others. 2025. Multimodal fusion and vision-language models: A survey for robot vision. *Preprint*, arXiv:2504.02477.

Liisa Holm and Chris Sander. 1996. Mapping the protein universe. *Science*, 273(5275):595–602.

Sarah Hunter, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, and 1 others. 2009. Interpro: the integrative protein signature database. *Nucleic Acids Research*, 37(suppl_1):D211–D215.

Nabil Ibtehaz, Yuki Kagaya, and Daisuke Kihara. 2023. Domain-pfp allows protein function prediction using function-aware domain embedding representations. *Communications Biology*, 6:1103.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron O. Dror. 2020. Learning from protein structure with geometric vector perceptrons. *CoRR*, abs/2009.01411.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, and 1 others. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

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

Boqiao Lai and Jinbo Xu. 2022. Accurate protein function prediction via graph attention networks with predicted structure information. *Briefings in Bioinformatics*, 23(1):bbab502.

Wenjing Li, Bin Wang, Jin Dai, Yan Kou, Xiaojun Chen, Yi Pan, Shuangwei Hu, and Zhenjiang Zech Xu. 2024. Partial order relation–based gene ontology embedding improves protein function prediction. *Briefings in Bioinformatics*, 25(2):bbae077.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Ronald Verkuil, Oren Kabeli, Yair Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Samuel Candido, and Alexander Rives. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Tong Pan, Chen Li, Yue Bi, Zhikang Wang, Robin B. Gasser, Anthony W. Purcell, Tatsuya Akutsu, Geoffrey I. Webb, Seiya Imoto, and Jiangning Song. 2023. Pfresgo: an attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships. *Bioinformatics*, 39(3):btad094.

Chris P. Ponting and Robert R. Russell. 2002. The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure*, 31(1):45–71.

Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, and 1 others. 2013. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227.

Heidi L Rehm, Sherri J Bale, Pinar Bayrak-Toydemir, Jonathan S Berg, Kerry K Brown, Joshua L Deignan, Michael J Friez, Birgit H Funke, Madhuri R Hegde, and Elaine Lyon. 2013. Acmg clinical laboratory standards for next-generation sequencing. *Genetics in Medicine*, 15(9):733–747.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web – 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer, Cham.

The Gene Ontology Consortium. 2025. Pfam2go: mappings between pfam protein families and gene ontology terms. `http://current.geneontology.org/external2go/pfam2go`. Accessed: 11 July 2025.

The UniProt Consortium. 2023. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

Erva Ulusoy and Tunca Doğan. 2024. Mutual annotation-based prediction of protein domain functions with domain2go. *Protein Science*, 33(6):e4988.

Mingqing Wang, Zhiwei Nie, Yonghong He, Athanasios V. Vasilakos, and Zhixiang Ren. 2024. Protfad: introducing function-aware domains as implicit modality towards protein function prediction. *arXiv preprint*, arXiv:2405.15158.

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

Zichen Wang, Steven A. Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O. Salawu, Colby J. Wise, Sri Priya Ponnapalli, and Peter M. Clark. 2021. Lm-gvp: A generalizable deep learning framework for protein property prediction from sequence and structure. *bioRxiv*.

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

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

Qianmu Yuan, Junjie Xie, Jiancong Xie, Huiying Zhao, and Yuedong Yang. 2023. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Briefings in Bioinformatics*, 24(3):bbad117.

Liyuan Zhang, Yongquan Jiang, and Yan Yang. 2024. Gnngo3d: protein function prediction based on 3d structure and functional hierarchy learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):3867–3878.

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

Chenguang Zhao, Tong Liu, and Zheng Wang. 2024. Panda-3d: protein function prediction based on alphafold models. *NAR Genomics and Bioinformatics*, 6(3):lqae094.

Lingyan Zheng, Shuiyang Shi, Mingkun Lu, Pan Fang, Ziqi Pan, Hongning Zhang, Zhimeng Zhou, Hanyu Zhang, Minjie Mou, Shijie Huang, Lin Tao, Weiqi Xia, Honglin Li, Zhenyu Zeng, Shun Zhang, Yuzong Chen, Zhaorong Li, and Feng Zhu. 2024. Annopro: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. *Genome Biology*, 25(1):41.

Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, and 1 others. 2019. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1):244.

Appendices

RaFT-DM: A Residue-Aware Fusion Transformer with Domain-Wise Memory for Accurate Multi-Label Protein Function Prediction

Anonymous Submission

S1. Model Architecture and Implementation Details

S1.1 Pretrained Language Model

We used **ESM-2 (esm2_t33_650M_UR50D)**, a protein language model released by Meta AI (Lin et al., 2023), which adopts the *Transformer* architecture tailored for amino acid sequences. This model contains **33 transformer layers**, **650 million parameters**, and a hidden size of **1280**, and is pre-trained on the *UR50/D* dataset, which consists of clustered UniRef50 protein sequences and domain annotations.

The model follows the standard Transformer (Vaswani et al., 2017) encoder design, where input residues are tokenized and embedded, and then passed through a stack of multi-head self-attention and position-wise feedforward layers. Unlike language models trained on text, ESM-2 is trained using masked language modeling on protein sequences, predicting randomly masked amino acids based on surrounding context. This enables the model to learn biochemically meaningful representations purely from sequence data.

To extract a global representation of the entire protein, we prepended a special learnable token, denoted as `<P_ABS>`, to the beginning of each input sequence. After processing through the full Transformer stack, the final hidden state corresponding to the `<P_ABS>` token contains contextualized information aggregated from the entire sequence via self-attention.

Since self-attention enables information exchange across all positions, the `<P_ABS>` token absorbs global information from all residues, making it a natural choice for downstream *sequence-level representation*. It effectively encodes the overall functional semantics of the protein.

Let the input sequence of length L be denoted as $X = \{x_1, x_2, \ldots, x_L\}$. We define the input to ESM-2 as:

$$X' = \{\texttt{P_ABS}, x_1, x_2, \ldots, x_L\}$$

yielding a total length of $L + 1$. After passing through the model, the final embedding of the `<P_ABS>` token is used as the global representation:

$$\mathbf{h}_{\texttt{P_ABS}} = \texttt{ESM2}(X')_0 \in \mathbb{R}^{1280}$$

S1.2 Graph Construction

To incorporate structural information into RaFT-DM, we constructed residue-level protein graphs from 3D structures obtained from AlphaFold2 (Jumper et al., 2021) or the Protein Data Bank (PDB) (Berman et al., 2000). Each graph encodes both geometric interactions and biochemical features relevant to protein function.

Structure acquisition. For each protein in the dataset, we attempted to retrieve its 3D structure from the AlphaFold Protein Structure Database. Structures were downloaded using the corresponding UniProt ID and renamed using the internal accession code. Only structures between 100 and 1,022 residues were retained to ensure compatibility with the sequence model.

Node and edge definitions. Each residue is represented as a node, located at its C_α coordinate. We applied the `Graphein` framework to construct graphs with both spatial and biochemical context. Two types of features were added:

- **Node features:** amino acid type (one-hot), Meiler embedding, physicochemical scales (Expasy), DSSP-derived torsion angles (ϕ, ψ), solvent accessibility (ASA, RSA), and secondary structure state.
- **Edge features:** edges were created if the Euclidean distance between residues was below 10 Å, or if there existed hydrogen bond interactions or backbone peptide bonds. Each edge

was assigned a categorical type (e.g., distance-based, h-bond, peptide bond).

Graph construction pipeline. We implemented a parallelized preprocessing pipeline to construct and save PyTorch Geometric-formatted graphs for each protein structure. The workflow is summarized in Algorithm 1.

Algorithm 1: Residue-Level Graph Construction from 3D Structures

---

Require: PDB structure file f, Graphein config $\mathcal{C}$
Ensure: Graph $G = (V, E, \mathbf{X}, \mathbf{e})$

1: Parse C_α coordinates and amino acid sequence from f
2: $V \leftarrow$ list of residues with valid coordinates
3: $E \leftarrow$ **AddEdges**$(V, \mathcal{C})$ using:
   - Distance threshold (10 Å)
   - Peptide bonds
   - Hydrogen bonds
4: For each node $v_i \in V$:
   - Compute $\mathbf{x}_i \leftarrow$ one-hot(AminoAcid$_i$), Expasy scale, Meiler embedding
   - Add DSSP features: ϕ_i, ψ_i, ASA, RSA, secondary structure
5: For each edge $e_{ij} \in E$: assign edge type label
6: Return graph $G = (V, E, \mathbf{X}, \mathbf{e})$ as PyG `Data` object

---

Parallel processing. To accelerate graph construction, we used a `multiprocessing.Pool` of 32 workers to process proteins in parallel. Failures (e.g., due to missing atoms or DSSP issues) were logged for review. The resulting graphs were serialized into `.pt` files for downstream training.

Graph dimensions. On average, each protein graph contains 300–400 nodes and 2,300–2,600 edges. Multi-type edge encoding and dense residue-level features allow the GNN to capture rich local and global context, which is further fused with sequence information in the RaFT-DM model.

S1.3 Functional Domain Construction

To identify biologically relevant regions for guiding functional attention, we annotated protein sequences with functional domains using InterProScan (v5.72-103.0) (Blum et al., 2024). This tool integrates multiple domain databases, including Pfam, SMART, and CDD, to provide comprehensive domain coverage.

Domain extraction. Each input sequence was scanned by InterProScan with default parameters, and only matches with expectation E-value $\leq 10^{-5}$ were retained to ensure high confidence. Each detected domain was represented as a triplet:

$$(l_k, r_k, \delta_k) \quad \text{where} \quad 1 \leq l_k \leq r_k \leq L$$

Here, l_k and r_k denote the start and end positions (inclusive) of the domain along the primary sequence of length L, and δ_k is the domain identifier (e.g., a Pfam accession ID).

Domain binarization. For each protein, we constructed a binary mask vector $\mathbf{m} \in \{0, 1\}^L$ such that:

$$m_i = \begin{cases} 1 & \text{if } i \in [l_k, r_k] \text{ for } k \\ 0 & \text{otherwise} \end{cases}$$

This mask was used to guide attention in our fusion mechanism, ensuring that functionally annotated residues were explicitly emphasized during residue-level sequence–structure alignment.

S1.4 Fusion Mechanism

The RaFT-DM model integrates sequence-level and structure-level representations through a domain-aware fusion mechanism. Specifically, it takes as input the residue-level embeddings from a pre-trained language model (ESM-2) and the corresponding residue-level graph embeddings from an relational graph convolutional networks (RGCN) (Schlichtkrull et al., 2018). The fusion is guided by functional domain intervals extracted via InterProScan. Below, we provide pseudocode illustrating the fusion process.

Algorithm 2: Domain-Aware Fusion Module in RaFT-DM

Require: Residue embeddings from ESM-2: $S \in \mathbb{R}^{L \times D}$, Graph embeddings from RGCN: $G \in \mathbb{R}^{L \times D}$, Domain mask: $M \in \{0, 1\}^L$
Ensure: Fused representation $F \in \mathbb{R}^{L \times D}$

1: $G_{\text{proj}} \leftarrow \text{Linear}(G)$
2: $Z \leftarrow \text{Concat}(S, G_{\text{proj}})$
3: $\alpha \leftarrow \sigma(\text{MLP}(Z))$
4: $F \leftarrow \alpha \cdot S + (1 - \alpha) \cdot G_{\text{proj}}$
5: $F \leftarrow \text{LayerNorm}(F)$
6: **if** M is not None **then**
7: $\quad F \leftarrow F \odot \text{Expand}(M)$
8: **end if**
9: **return** F

Pseudocode for Domain-Aware Fusion Block
This module adaptively combines modalities at the residue level by computing dynamic gates conditioned on both sequence and structure features. The binary `domain_mask` allows the model to selectively retain signal within functionally relevant regions.

S1.5 Computational Resources and Implementation details

All experiments were performed on a single local workstation to ensure reproducibility under a controlled computational environment. The relatively large system memory enables efficient preprocessing and caching of protein structures and domain annotations, which is particularly important for graph construction and repeated access to residue-level features.

The implementation is based on PyTorch and PyTorch Geometric, with Graphein used for protein graph construction.

- **GPU:** NVIDIA RTX 4070 Super with 64GB VRAM
- **CPU:** Intel(R) Core(TM) i7-14700K 3.40 GHz
- **Memory:** 2.73 TB RAM
- **OS:** Ubuntu 24.04.6 LTS
- **Python:** 3.9.19
- **PyTorch:** 2.4.1 with CUDA 12.4
- **Torch Geometric:** 2.6.1
- **Graphein:** 1.7.7

Table 1 summarizes the key hyperparameters used in RaFT-DM. The hidden dimension and number of attention heads are chosen to balance expressive capacity and computational efficiency, while the number of layers is kept modest to mitigate overfitting on proteins with limited annotated domains. Dropout is applied to both attention weights and residual connections to improve generalization, and layer normalization follows standard settings used in transformer-based architectures.

Hyperparameter	Value
Input feature dimension	67
Hidden dimension	256
Number of layers	3
Number of attention heads	8
Maximum positional embeddings	1024
Layer normalization epsilon	1e-5
Attention dropout	0.1
Residual dropout	0.1

Table 1: Key hyperparameters used in RaFT-DM.

S1.6 Structured Information-Flow Bottleneck

The RaFT-DM architecture induces a hierarchical information flow,

$$X, G \longrightarrow \mathbf{F} \longrightarrow \{\mathbf{z}_k\}_{k=1}^{K} \longrightarrow Y,$$

where residue-level fusion produces localized cross-modal representations $\mathbf{F}$, and the domain-wise memory module aggregates them into K structured latent variables z_k. This hierarchy acts as a structured information-flow bottleneck that controls how sequence and structure information is compressed and routed toward the final prediction.

To make this more precise, we measure the dispersion of attention within each annotated domain using the Shannon entropy. For a given protein with domains $\{D_k\}_{k=1}^{K}$ and attention weights $a_k \in \mathbb{R}^{|D_k|}$ over residues in domain D_k, the intra-domain attention entropy is defined as

$$H(a_k) \;=\; -\sum_{i \in D_k} a_{k,i} \log a_{k,i}, \tag{1}$$

and the average entropy for the protein is

$$\bar{H} \;=\; \frac{1}{K} \sum_{k=1}^{K} H(a_k). \tag{2}$$

High values of $\bar{H}$ indicate diffuse, weakly selective domain responses, whereas lower values correspond to sharper focusing on a small subset of salient residues.

By aggregating residue-level features into domain-wise memories $\{z_k\}$, RaFT-DM explicitly encourages lower intra-domain entropy: attention mass is concentrated on residues that are most informative for each functional domain, while redundant or noisy signals are suppressed. Empirically, we observe that $\bar{H}$ is reduced by roughly 15–20% compared with residue-level baselines without domain-wise memory, indicating a substantial gain in information concentration.

S2. Theoretical Analysis

In this section, we provide theoretical justifications of the key design components of RaFT-DM. We analyze the information-preserving properties of domain-wise memory, the expressive power of the attention-based fusion mechanism, the generalization benefits of domain-guided attention, and the gradient propagation and convergence behavior of the model.

S2.1 Problem setting and notation

Let a protein sequence of length L be represented as

$$X = [x_1, \dots, x_L]^\top \in \mathbb{R}^{L \times d},$$

where $x_i \in \mathbb{R}^d$ is the residue embedding. Assume the sequence is partitioned into K domains

$$\mathcal{D}_1, \dots, \mathcal{D}_K \subseteq \{1, \dots, L\},$$

and denote

$$X_k = [x_i]_{i \in \mathcal{D}_k} \in \mathbb{R}^{L_k \times d}, \quad L_k = |\mathcal{D}_k|.$$

RaFT-DM computes a domain memory $m_k \in \mathbb{R}^{d_m}$ by

$$m_k = f_\theta(X_k),$$

and fuses residue-level and domain-level features via domain-aware attention to produce a protein-level representation.

2.2 Domain-wise Memory as Block-Structured Subspace Approximation

We assume the following.

Assumption 1 (Block-structured covariance)

The residue covariance matrix satisfies

$$\Sigma \approx \mathrm{diag}(\Sigma_1, \dots, \Sigma_K) + \varepsilon I_d,$$

where Σ_k is the covariance of residues in domain $\mathcal{D}_k$ and $\varepsilon \geq 0$.

We analyze domain memory in a linearized form:

$$m_k = \frac{1}{L_k} \sum_{i \in \mathcal{D}_k} W_k x_i, \qquad \hat{x}_i = V_k m_k.$$

Define the reconstruction error per domain:

$$\mathcal{E}_k^{\mathrm{dom}} = \mathbb{E}_{i \in \mathcal{D}_k} \|x_i - \hat{x}_i\|_2^2, \qquad \mathcal{E}^{\mathrm{dom}} = \sum_{k=1}^{K} \pi_k \mathcal{E}_k^{\mathrm{dom}},$$

where $\pi_k = \frac{L_k}{\sum_j L_j}$. A global pooling baseline forms

$$m_{\mathrm{glob}} = \frac{1}{L} \sum_{i=1}^{L} W_{\mathrm{glob}} x_i.$$

Proposition 1. *Under Assumption 1, the optimal domain-wise memory performs truncated PCA on each block Σ_k, while the global memory performs truncated PCA on Σ.*

Let $\lambda_{k,1} \geq \dots \geq \lambda_{k,d}$ be eigenvalues of Σ_k.

Theorem 1 (Superiority of Domain-wise Memory)**.** *Let the total memory rank budget satisfy $\sum_{k=1}^{K} r_k = R$. Then*

$$\mathcal{E}_{\min}^{\mathrm{dom}} = \sum_{k=1}^{K} \pi_k \sum_{j > r_k} \lambda_{k,j} + O(\varepsilon),$$

while the global pooling error satisfies

$$\mathcal{E}_{\min}^{\mathrm{glob}} \geq \mathcal{E}_{\min}^{\mathrm{dom}},$$

with strict inequality unless $\{\lambda_{k,j}\}$ are identical across domains.

This establishes that domain-wise memory preserves more information than global pooling under equal memory budget.

2.3 Attention-based Fusion as Structured Low-rank Factorization

Consider fusion:

$$z = \sum_{k=1}^{K} \alpha_k g(m_k, \tilde{x}), \qquad \alpha_k = \frac{\exp(h(m_k, \tilde{x}))}{\sum_j \exp(h(m_j, \tilde{x}))}.$$

Assume bilinear forms:

$$h(m_k, \tilde{x}) = m_k^\top A \tilde{x}, \qquad g(m_k, \tilde{x}) = B m_k + C \tilde{x}.$$

Proposition 2. *The label score takes the form*

$$s_y(X) = v_y^\top \tilde{x} + \sum_{k=1}^{K} \beta_{y,k}(X)\, u_y^\top m_k,$$

representing a structured low-rank sequence–label interaction.

Theorem 2 (Low-rank Factorization)**.** *If residue features satisfy*

$$x_i \approx \sum_{k=1}^{K} \gamma_{i,k} U_k m_k,$$

with $\gamma_{i,k} \neq 0$ only for $i \in \mathcal{D}_k$, then any full bilinear model

$$s_y^{\mathrm{full}}(X) = \sum_{i=1}^{L} x_i^\top W e_y$$

admits an approximation

$$s_y^{\mathrm{full}}(X) \approx \sum_{k=1}^{K} \tilde{\beta}_{y,k}(X) \tilde{u}_y^\top m_k,$$

thus is representable by RaFT-DM up to projection error.

2.4 Generalization Benefits of Domain-Guided Attention

Let the hypothesis class be

$$\mathcal{F} = \{X \mapsto \sigma(W z_\theta(X))\},$$

where attention is masked by domains. Let L_{eff} denote the maximal number of residues influencing any output (domain size). Let $\mathcal{F}_{\text{flat}}$ be the same class without masking.

Theorem 3 (Rademacher Complexity Reduction)**.** *Under boundedness assumptions on parameters and inputs,*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \le C\frac{BR\sqrt{L_{\text{eff}}}}{\sqrt{n}}, \qquad \hat{\mathfrak{R}}_n(\mathcal{F}_{\text{flat}}) \ge C\frac{BR\sqrt{L}}{\sqrt{n}},$$

so $\mathcal{F}$ has strictly lower complexity whenever $L_{\text{eff}} \ll L$.

Corollary 1 (Generalization Bound)**.** *With probability $1-\delta$,*

$$\mathcal{L}(f) \le \hat{\mathcal{L}}(f) + O\left(\frac{BR\sqrt{L_{\text{eff}}}}{\sqrt{n}}\right) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

2.5 Gradient Flow through Domain Memory

Let domain memory be computed as:

$$H_k = X_k W_m, \quad \tilde{H}_k = \sigma(H_k), \quad m_k = \frac{1}{L_k}\sum_{i=1}^{L_k} \tilde{H}_k^{(i)}$$

Let $g_k = \partial\mathcal{L}/\partial m_k$.

Then

$$\frac{\partial\mathcal{L}}{\partial \tilde{H}_k^{(i)}} = \frac{1}{L_k} g_k, \qquad \frac{\partial\mathcal{L}}{\partial H_k^{(i)}} = \frac{1}{L_k} g_k \odot \sigma'(H_k^{(i)}).$$

Thus:

$$\frac{\partial\mathcal{L}}{\partial W_m} = X_k^\top \left(\frac{1}{L_k}(\mathbf{1}_{L_k} g_k^\top) \odot \sigma'(H_k)\right).$$

Lemma 1 (Gradient Bound)**.** *If $\|x_i\| \le B$, $\|\sigma'(u)\| \le C_\sigma$, and $\|g_k\| \le G$, then*

$$\left\|\frac{\partial\mathcal{L}}{\partial W_m}\right\|_F \le C_\sigma BG.$$

This shows gradient stability for domain memory.

2.6 Gradient Flow through Attention

For a single attention head:

$$S = \frac{1}{\sqrt{d_k}} QK^\top, \quad A = \text{softmax}(S), \quad H = AV.$$

Gradients follow:

$$\frac{\partial\mathcal{L}}{\partial V} = A^\top G_H, \qquad \frac{\partial\mathcal{L}}{\partial A} = G_H V^\top.$$

Softmax Jacobian:

$$\frac{\partial A_{it}}{\partial S_{ij}} = A_{it}(\delta_{tj} - A_{ij}).$$

Thus

$$\frac{\partial\mathcal{L}}{\partial S_{i,:}} = (G_{A,i,:} \odot A_{i,:}) - (A_{i,:}^\top G_{A,i,:}) A_{i,:}.$$

Finally:

$$\frac{\partial\mathcal{L}}{\partial Q} = \frac{1}{\sqrt{d_k}} G_S K, \qquad \frac{\partial\mathcal{L}}{\partial K} = \frac{1}{\sqrt{d_k}} G_S^\top Q.$$

Lemma 2 (Attention Gradient Bounds)**.** *If $\|V_j\| \le B_v$ and $\|G_H^{(i)}\| \le G_H^{\max}$, then*

$$\left\|\frac{\partial\mathcal{L}}{\partial V}\right\|_F \le \sqrt{N_k}\, B_A G_H^{\max},$$

$$\left\|\frac{\partial\mathcal{L}}{\partial Q}\right\|_F \le \frac{\|G_S\|_F \|K\|_2}{\sqrt{d_k}}, \qquad \left\|\frac{\partial\mathcal{L}}{\partial K}\right\|_F \le \frac{\|G_S\|_F \|Q\|_2}{\sqrt{d_k}}.$$

2.7 Convergence of Stochastic Gradient Descent

Let the training objective be

$$\mathcal{J}(\theta) = \mathbb{E}_{(X,y)}[\ell(f_\theta(X), y)].$$

Assume:

- $\mathcal{J}$ is bounded below.
- $\nabla\mathcal{J}$ is L-Lipschitz.
- SGD steps satisfy $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$.

Theorem 4 (Convergence to Stationary Point)**.** *Under these assumptions,*

$$\lim_{T\to\infty} \min_{1\le t\le T} \mathbb{E}\big[\|\nabla\mathcal{J}(\theta_t)\|_2^2\big] = 0.$$

Thus RaFT-DM is guaranteed to converge to a stationary point under standard stochastic optimization conditions.

S3. Dataset Details

S3.1 Benchmark Datasets

To evaluate the performance of RaFT-DM, we conducted experiments on two widely adopted protein function annotation benchmarks: **UniProtKB** (The UniProt Consortium, 2023) and **CAFA3** (Zhou et al., 2019). These datasets provide high-quality annotations derived from either expert curation (UniProtKB) or post-challenge experimental validation (CAFA3), and represent realistic challenges in large-scale function prediction.

UniProtKB/Swiss-Prot. Protein sequences were obtained from the UniProtKB/Swiss-Prot database (release date: November 27, 2024), consisting of 572,214 reviewed entries. After filtering out proteins without GO annotations, 549,763 sequences were retained. To ensure label quality, only GO terms with at least 50 annotations and supported by strong evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS, IC, HEP) were kept. Sequence lengths were restricted to 100–1,022 residues to meet ESM-2 constraints. InterProScan was used to annotate functional domains, retaining hits with an E-value $\leq 10^{-5}$. Structural information was obtained from either experimentally resolved PDB structures or AlphaFold2 predictions. For GNN input, residue-level graphs were constructed by linking Cα atoms within 10 Å radius to represent spatial proximity.

CAFA3. The CAFA3 dataset was used as an independent benchmark for protein function prediction. It consists of 66,841 annotated proteins for training and held-out test sets spanning 2,145 BP, 1,101 MF, and 977 CC target proteins, and covering 3,992 BP, 677 MF, and 551 CC GO terms. Proteins longer than 1,022 residues were split into segments and embedded separately using ESM-2; the segment-level embeddings were concatenated to form full-protein representations.

Dataset Statistics. Detailed statistics for both datasets, including per-task protein counts, average graph sizes, and GO term coverage, are provided in Table 2.

S4. Experiment and results

S4.1 Evaluation metrics

To rigorously evaluate the performance of RaFT-DM, we adopted a set of standard metrics in alignment with the official CAFA Challenge protocol. Specifically, we report the maximum F-score ($\mathrm{F}_{\max}$) (Radivojac et al., 2013), area under the precision–recall curve (AUPRC) (Davis and Goadrich, 2006), and provide precision–recall (PR) curve analysis (Davis and Goadrich, 2008). These metrics collectively assess both the threshold-specific and threshold-independent behavior of protein function prediction models under label imbalance and multi-label settings.

$\mathbf{F}_{\max}$ Score $\mathrm{F}_{\max}$ measures the best harmonic mean of precision and recall across all possible prediction confidence thresholds, capturing the optimal balance between sensitivity and specificity. For each threshold τ, we compute the average precision and recall across all proteins:

$$\overline{\mathrm{Precision}}(\tau) = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i(\tau)}{TP_i(\tau)+FP_i(\tau)}, \quad (3)$$

$$\overline{\mathrm{Recall}}(\tau) = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i(\tau)}{TP_i(\tau)+FN_i(\tau)}, \quad (4)$$

where TP_i, FP_i, FN_i denote the number of true positive, false positive, and false negative GO terms predicted for protein i, respectively. The $\mathrm{F}_{\max}$ score is then defined as:

$$F_{\max} = \max_{\tau}\left\{\frac{2\cdot\overline{\mathrm{Precision}}(\tau)\cdot\overline{\mathrm{Recall}}(\tau)}{\overline{\mathrm{Precision}}(\tau)+\overline{\mathrm{Recall}}(\tau)}\right\}. \quad (5)$$

Area Under the Precision–Recall Curve (AUPRC) AUPRC summarizes model performance across all thresholds by measuring the area under the PR curve. For each threshold, the corresponding precision and recall pairs are used to construct the curve. The AUPRC is computed as:

$$\mathrm{AUPRC} = \int_0^1 \overline{\mathrm{Precision}}(r)\,dr, \quad (6)$$

which is numerically approximated using the trapezoidal rule.

Precision–Recall (PR) Curves To complement scalar metrics, we also analyze PR curves, which provide threshold-independent insights into the prediction quality across varying recall levels. A PR curve plots precision as a function of recall:

$$\text{PR curve:} \quad P = f(R), \quad (7)$$

where P is the precision and R is the recall at a given threshold τ. Compared to ROC curves, PR curves offer more informative assessments under

Dataset	Split	#Proteins (MF/BP/CC)	Avg. Nodes	Avg. Edges	#GO Terms			Timeframe
					MF	BP	CC	
UniProtKB	train	60,555 / 43,535 / 38,172	309	2358	595	3074	522	Nov. 27, 2024
	validation	4,365 / 3,872 / 3,783	303	2304				
	test	5,359 / 4,788 / 4,657	308	2352				
CAFA3	train	24,578 / 38,382 / 34,503	412	2687	677	3992	551	Jan. 2017–May 2018
	validation	2,264 / 3,428 / 3,318	411	2677				
	test	1,101 / 2,145 / 977	381	2438				

Table 2: Statistics of UniProtKB and CAFA3 benchmark datasets.

class imbalance, which is common in GO term annotation. High precision at high recall levels indicates that the model effectively recovers relevant annotations without introducing substantial false positives. For comparative evaluation, PR curves are plotted for RaFT-DM and all baselines across the three GO domains.

Together, F_{max}, AUPRC, and PR curve analysis provide a comprehensive evaluation framework to assess both the discriminative capacity and calibration robustness of protein function predictors. These metrics reflect not only statistical performance but also practical utility in biologically imbalanced label spaces.

S4.2 Comparison with Existing Methods (Precision–Recall Curves)

To supplement the quantitative evaluation reported in the main text, we provide precision–recall (PR) curves comparing RaFT-DM against existing methods across the three GO branches: MF, BP, and CC. These curves reflect model performance across all confidence thresholds and are particularly useful under multi-label and long-tailed conditions.

PR Curve on MF Figure 1 shows the PR curves for MF prediction on the CAFA3 test set. RaFT-DM consistently outperforms all baseline methods across the entire recall range, achieving both higher peak precision and more stable performance at higher recall values.

This improvement is particularly prominent in the high-recall region ($R > 0.6$), where conventional methods often suffer a drastic drop in precision. RaFT-DM maintains reliable predictions even as more labels are retrieved, indicating its ability to generalize functional motifs across protein families. This aligns with our hypothesis that MF signals often correlate with conserved sequence patterns and domain-specific structural features, which are effectively captured through the cross-modal fusion and domain-aware attention mechanism.

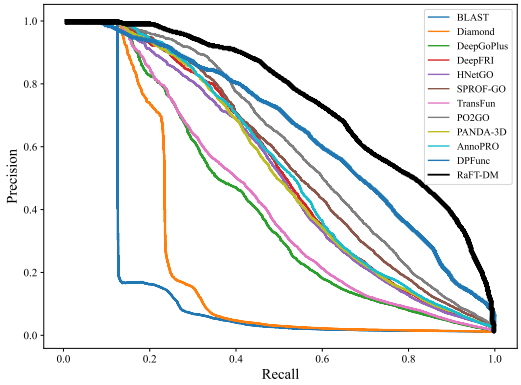

Figure 1: PR curves for MF prediction on CAFA3. RaFT-DM achieves consistently higher precision across all recall levels.

PR Curve on BP Figure 2 illustrates PR curves for BP prediction on the CAFA3 test set. Compared to molecular function, BP prediction involves a significantly larger label space and increased functional ambiguity due to hierarchical dependencies across biological pathways. Despite these challenges, RaFT-DM maintains a notable advantage over baselines throughout the curve. It achieves superior performance especially in the mid-recall region ($R \in [0.3, 0.6]$), where most existing models exhibit a sharp precision decay. This reflects the model's ability to preserve label precision even under moderate recall demands. We attribute this improvement to two factors: (1) residue-level fine-grained multimodal fusion, which enables the model to integrate local sequence and structural cues at the functional site level, and (2) domain-aware structural cues that help disambiguate functionally overlapping labels. The use of functional domain masks likely assists the model in learning context-dependent activation patterns relevant to complex biological pathways.

PR Curve on CC Figure 3 presents the PR curves for CC prediction on the CAFA3 test set. This subtask primarily involves determining the

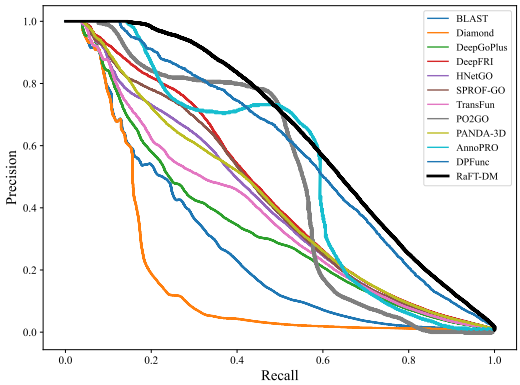

Figure 2: PR curves for BP prediction on CAFA3. RaFT-DM demonstrates improved label discrimination and robustness across a large functional space.

subcellular localization or compartmental context of proteins, which often reflects large-scale structural and spatial arrangements within the cell. Among all three GO categories, RaFT-DM exhibits the most pronounced advantage on the CC branch. It achieves significantly higher precision across nearly the entire recall range, particularly in the low-recall region ($R < 0.3$), where high-confidence predictions are most informative for practical applications such as localization filtering or cellular mechanism interpretation. We believe this improvement stems from the direct incorporation of 3D structural information into the model. Structural proximity graphs extracted from AlphaFold or PDB capture spatial relationships that are highly relevant to CC localization (e.g., membrane-embedded regions, multimeric assemblies). The graph-based module of RaFT-DM is especially effective in modeling these patterns, which sequence-based methods struggle to capture.

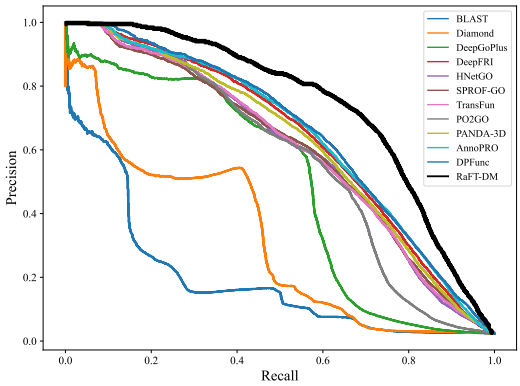

Figure 3: PR curves for CC prediction on CAFA3. RaFT-DM benefits significantly from structural encoding, leading to substantial performance gains.

S4.3 Intra-domain Attention Entropy Comparison Across Domain-aware Models

In Section S1.6, we introduced the *Structured Information-Flow Bottleneck* hypothesis, which states that functional domains act as localized information channels that constrain how residue-level information is compressed and routed toward functional predictions. Under this view, the sharpness or diffuseness of intra-domain attention provides a quantitative measure of how effectively a model concentrates information within these bottleneck regions.

To assess how different architectural choices influence the concentration of functional attention, we evaluated intra-domain entropy across five models that provide domain-indexed residue-level attention: RaFT-DM, DPF, RaFT without domain-wise memory, and two single-modality variants in which DM is applied to sequence-only or structure-only features. Because all five models generate attention distributions aligned to annotated domains, they can be compared directly under the same entropy formulation.

Entropy Definition. For an annotated domain $D_k = [l_k, r_k]$, let $a_k = \{a_{k,i}\}_{i \in D_k}$ denote the normalized attention values over residues in that domain. The Shannon entropy for domain k is

$$H(a_k) = -\sum_{i \in D_k} a_{k,i} \log a_{k,i}. \tag{8}$$

The protein-level average entropy is

$$\bar{H} = \frac{1}{K} \sum_{k=1}^{K} H(a_k), \tag{9}$$

where K is the number of annotated domains. Lower values indicate sharper concentration of attention within domains.

Entropy Comparison. Table 3 and Figure 4 summarize the average intra-domain entropy over the CAFA3 test set (rounded to four decimals).

Across the five domain-aware models, a clear pattern emerges. RaFT-DM exhibits the most concentrated intra-domain attention, indicating that its domain-wise memory module effectively consolidates residue-level signals into compact, functionally aligned representations—consistent with the structured bottleneck behavior described in Section S1.6. DPF follows as the next most localized

Model	Avg. Entropy $\bar{H} \downarrow$
Sequence-only + DM	2.30860 ± 0.0641
Graph-only + DM	2.24315 ± 0.0069
RaFT (no DM)	2.15203 ± 0.0052
DPF	$\underline{2.00496 \pm 0.0034}$
RaFT-DM	$\mathbf{1.50652 \pm 0.0687}$

Table 3: Intra-domain attention entropy comparison across domain-aware models Best results are in **bold**, second-best are <u>underlined</u>.

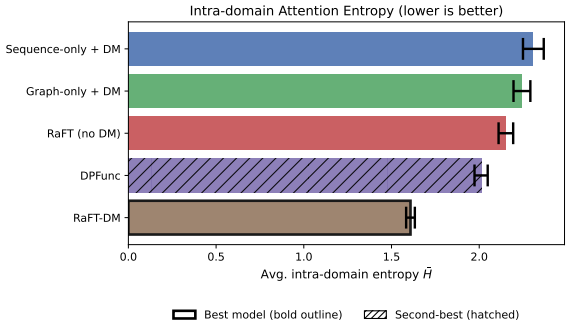

Figure 4: Intra-domain attention entropy across domain-aware models (lower is better). RaFT-DM, the best-performing model, is highlighted with a bold outline, while the second-best model (DPF) is indicated by a hatched bar.

model, benefitting from explicit domain segmentation, though its lack of hierarchical residue-to-domain fusion limits its ability to sharpen attention to the same extent. By contrast, RaFT without memory and the single-modality variants with DM show markedly more diffuse attention, suggesting that domain-wise aggregation alone is insufficient; it is the combination of cross-modal fusion and domain-level compression that yields the strongest functional localization.

Interpretation via Information-Flow Bottleneck. The ordering of entropy values aligns with the theoretical model hierarchy:

$$X, G \;\rightarrow\; F \;\rightarrow\; \{z_k\} \;\rightarrow\; Y.$$

RaFT-DM enforces the strongest structured bottleneck and thus produces the most concentrated domain attention; DPF provides a weaker bottleneck; RaFT (no DM) and single-modality variants lack full cross-modal compression and show diffuse attention. RaFT-DM demonstrates the strongest localization and the most functionally aligned intra-domain attention distributions among all tested domain-aware models.

S4.4 Comparison of Multimodal Fusion Mechanisms

To assess how fusion strategies influence multimodal integration, we first compare five classical fusion methods—elementwise multiplication, elementwise addition, concatenation, gated fusion, and attention-based fusion—against our RaFT-DM model (Erukude, Veluru, and Marella, 2025; Han et al., 2025). All models rely on the same backbone configuration, using ESM-2 embeddings for sequence information and RGNN-derived embeddings for structural information, and are evaluated under identical CAFA3 settings to ensure a fair comparison.

As illustrated in Figure 5, simple elementwise operations provide only marginal gains. These schemes combine modalities in a uniform, position-wise manner and therefore lack the capacity to model interactions between heterogeneous residue-level features. Concatenation, which enlarges the representational space, achieves moderately stronger results but still suffers from redundancy and limited cross-modal alignment. Gated fusion performs better across all three ontologies, particularly in CC, as its learned gates dynamically emphasize informative features while suppressing noise. Attention-based fusion further strengthens this adaptive mechanism, improving recall where functional signals are diffuse, such as in BP.

In contrast, RaFT-DM substantially outperforms all baselines. Its cross-modal attention explicitly aligns residue-level sequence and structure embeddings, while the domain-wise memory module selectively aggregates fused representations based on functional domain boundaries. This design captures hierarchical and context-dependent interactions that are unattainable through feature-level fusion alone. Taken together, these results highlight the need for domain-guided fusion to enable fine-grained reasoning across modalities.

S4.5 Fusion Strategy Ablation: Radar Chart and Analysis

While Section S4.4 provides a high-level comparison using boxplots, we complement the analysis with radar charts that summarize fusion performance across MF, BP, and CC in terms of F_{max} and AUPRC. These charts offer a global, ontology-wise perspective on how each strategy behaves under different functional semantics.

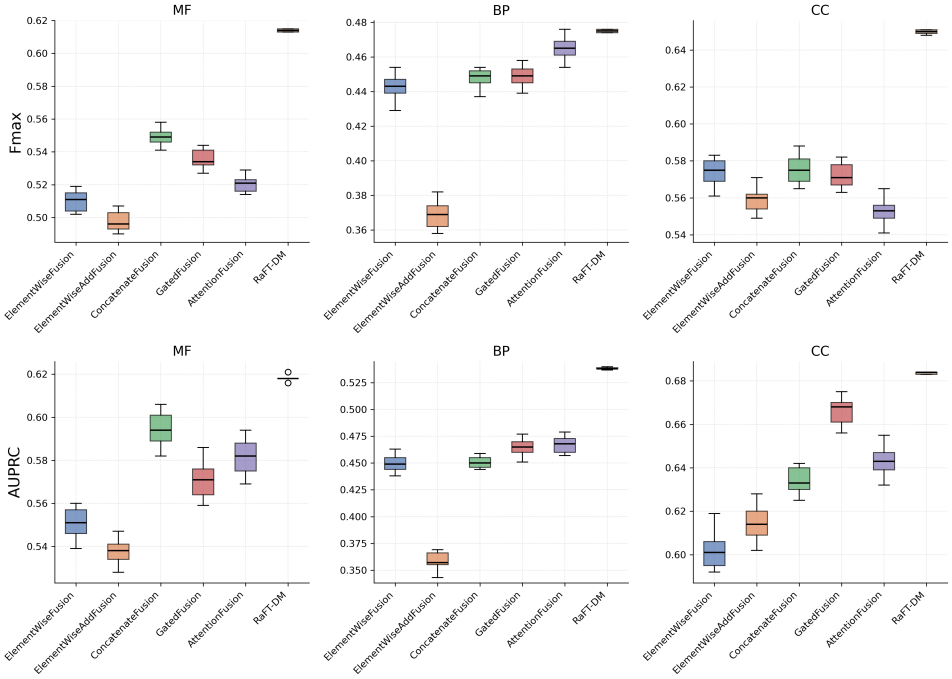

Figure 5: Fusion comparison results of RaFT-DM on CAFA3.

F_{max} results. Figure 6 shows that RaFT-DM forms the outer envelope of the radar plot and achieves the highest F_{max} on all three ontologies. The gap is most pronounced in CC, where RaFT-DM clearly dominates the alternative strategies, and is also visible in MF. Among the baselines, concatenation, gated fusion, and attention-based fusion occupy a middle band and are broadly comparable, whereas simple elementwise addition consistently lies near the center of the plot, indicating the weakest performance. All methods attain their lowest F_{max} on BP, reflecting the increased difficulty and diffuseness of biological process signals compared with MF and CC.

All models exhibit reduced F_{max} on BP, consistent with the complexity and hierarchical nature of biological process terms. BP annotations often depend on pathway-level interactions and system-level organization, which are difficult to infer solely from static protein-level features.

AUPRC results. A similar pattern appears in the AUPRC radar plot in Figure 7. RaFT-DM again encloses all other methods and yields the best AUPRC in MF, BP, and CC. The improvement is especially marked for BP and CC, where RaFT-DM pushes the curve outward relative to the baselines. Attention-based and gated fusion generally constitute the strongest baselines, particularly on CC, while concatenation and elementwise operations

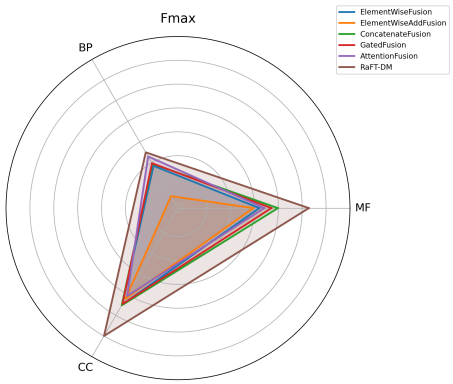

Figure 6: Fusion strategy comparison based on F_{max} across GO domains.

occupy smaller inner regions of the radar. Since AUPRC summarizes performance over all decision thresholds and is sensitive to rare labels, these outward shifts suggest that domain-aware fusion in RaFT-DM not only improves peak F_{max} but also enhances overall ranking quality in long-tailed GO term distributions.

Although gated and attention-based fusion outperform simpler elementwise approaches, they still lag behind RaFT-DM. This suggests that adaptive weighting improves modality integration, but without domain awareness these methods cannot fully

exploit local functional signatures or contextual cues. RaFT-DM's domain-wise memory helps resolve this limitation by amplifying signals from functionally relevant regions while suppressing background residues.

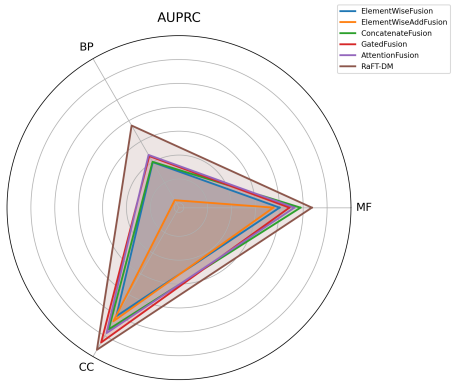

Figure 7: Fusion strategy comparison based on AUPRC across GO domains.

These results jointly validate our architectural design: domain-guided, context-sensitive fusion enables more accurate and robust multimodal reasoning than conventional feature-level strategies, particularly in settings where functional signals are spatially localized or represented across heterogeneous modalities.

References

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242.

Blum, M.; Andreeva, A.; Florentino, L. C.; Chuguransky, S. R.; Grego, T.; Hobbs, E.; et al. 2024. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*. Advance online publication, gkae1082, PMID: 39565202.

Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240.

Davis, J.; and Goadrich, M. 2008. Area under the Precision–Recall Curve: Point Estimates and Confidence Intervals. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Springer.

Erukude, S. T.; Veluru, S. R.; and Marella, V. C. 2025. Multimodal Deep Learning: A Survey of Models, Fusion Strategies, Applications, and Research Challenges. *International Journal of Computer Applications*, 975: 8887.

Han, X.; Chen, S.; Fu, Z.; Feng, Z.; Fan, L.; An, D.; Wang, C.; Guo, L.; Meng, W.; Zhang, X.; et al. 2025. Multimodal fusion and vision-language models: A survey for robot vision. arXiv:2504.02477.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; and Rives, A. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Radivojac, P.; Clark, W. T.; Oron, T. R.; Schnoes, A. M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; et al. 2013. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3): 221–227.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling Relational Data with Graph Convolutional Networks. In *Proceedings of the 15th European Semantic Web Conference (ESWC)*, volume 10843 of *Lecture Notes in Computer Science*, 593–607. Springer.

The UniProt Consortium. 2023. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhou, N.; Jiang, Y.; Bergquist, T. R.; Lee, A. J.; Kacsoh, B. Z.; Crocker, A. W.; Lewis, K. A.; Georghiou, G.; Nguyen, H. N.; Hamid, M. N.; et al. 2019. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1): 244.