Multimodal Mixture-of-Experts with Retrieval Augmentation for Protein Active Site Identification

Jiayang Wu[1*], Jiale Zhou[1*], Rubo Wang[2†], Xingyi Zhang[3], Xun Lin[1], Tianxu Lv[4], Leong Hou U[5], Yefeng Zheng[1†]

[1]Westlake University
[2]Chinese Academy of Sciences
[3]Mohamed bin Zayed University of Artificial Intelligence
[4]Jiangnan University
[5]University of Macau

Abstract

Accurate identification of protein active sites at the residue level is crucial for understanding protein function and advancing drug discovery. However, current methods face two critical challenges: vulnerability in single-instance prediction due to sparse training data, and inadequate modality reliability estimation that leads to performance degradation when unreliable modalities dominate fusion processes. To address these challenges, we introduce **M**ultimodal Mixture-of-**E**xperts with **R**etrieval **A**ugmentation (MERA), the first retrieval-augmented framework for protein active site identification. MERA employs hierarchical multi-expert retrieval that dynamically aggregates contextual information from chain, sequence, and active-site perspectives through residue-level mixture-of-experts gating. To prevent modality degradation, we propose a reliability-aware fusion strategy based on Dempster–Shafer evidence theory that quantifies modality trustworthiness through belief mass functions and learnable discounting coefficients, enabling principled multimodal integration. Extensive experiments on ProTAD-Gen and TS125 datasets demonstrate that MERA achieves state-of-the-art performance, with 90% AUPRC on active site prediction and significant gains on peptide-binding site identification, validating the effectiveness of retrieval-augmented multi-expert modeling and reliability-guided fusion.

Introduction

Accurate residue-level active site identification remains a critical bottleneck in mechanistic biology and drug discovery. This challenge stems fundamentally from the extreme label sparsity of catalytic/binding residues, which constitute less than 0.5% of all protein positions, compounded by significant functional divergence across homologous families (Petrova and Wu 2006). Consequently, despite decades of research, these functionally critical residues continue to elude precise prediction. This fundamental scarcity and resulting prediction inaccuracy directly impede the effectiveness of virtual screening pipelines (Gligorijević et al. 2021).

To overcome the dual bottlenecks of scarce annotations and extreme class imbalance, current research has largely pursued three complementary paradigms: (i) harnessing **sequence-only** pre-training to transfer massive unlabeled evolutionary knowledge, as demonstrated by ESM-1b (Rives et al. 2021) and ProtTrans (Elnaggar et al. 2021); (ii) incorporating sparse but precise 3D cues via **structure-aware** encoders when experimental coordinates are available, such as MIF (Yang, Zanichelli, and Yeh 2023) and PST (Chen et al. 2024); and (iii) enhancing predictive capacity via **cross-modal fusion** that integrates auxiliary modalities like protein textual descriptions through sequence-text embedding alignment, including MMSite (Ouyang et al. 2024), UniSite (Fan et al. 2025), and ProtST (Xu et al. 2023). Despite substantial progress, two persistent challenges persist in achieving reliable active-site prediction.

Challenge 1: Vulnerability of single-instance prediction. Predictions relying solely on intrinsic sequence features are inherently fragile due to sparse training data. This unreliability is particularly pronounced for rare protein sequences. While retrieval-augmented generation (RAG) (Guu et al. 2020) offers a promising solution by incorporating external contextual information, naive sequence-level homolog retrieval often introduces noise that overwhelms informative signals due to the sparsity of active site. Consequently, extracting and integrating key information from diverse perspectives becomes crucial to accommodate sequence length variability and functional diversity across homologs.

Challenge 2: Inadequate modality reliability estimation. Existing fusion approaches typically assess modality contributions through learned attention weights or MLP-based coefficients, which serve as poor proxies for true modality reliability (Han et al. 2022; Zhang et al. 2024; Yang et al. 2024). The fundamental limitation lies in conflating *contribution magnitude* with *epistemic reliability*: both cross-attention and MLP-based fusion methods optimize for signal strength rather than modality trustworthiness, making them unreliable indicators for modality confidence. When unreliable modalities dominate the fusion process, substantial performance degradation can occur. A principled approach should explicitly distinguish between modality con-

[*]Equal contribution.
[†]Corresponding author.

tribution strength and modality trustworthiness by modeling epistemic reliability at the residue level (Shen et al. 2025).

To address these challenges, we introduce **M**ultimodal Mixture-of-**E**xperts with **R**etrieval **A**ugmentation (**MERA**), a novel framework for protein active site identification. Our approach incorporates two key innovations. Multi-expert RAG (MeRAG) addresses the vulnerability of single-instance prediction by extracting meaningful information from sparse retrieval sequences through three orthogonal retrieval experts (chain, sequence, and active-site). Each expert captures distinct biological information, and their outputs are dynamically aggregated via a residue-level Mixture-of-Experts (MoE) gating mechanism. This design enables adaptive and robust feature learning even under sparse data conditions. (ii) Reliability-aware Multimodal Fusion (RMF) addresses inadequate modality reliability estimation by explicitly quantifying the trustworthiness of each modality. Inspired by Dempster–Shafer evidence theory (Amini et al. 2020; Huang et al. 2025; Deregnaucourt et al. 2025), we model the prediction of each modality as a belief mass function and apply learnable discounting coefficients to reflect modality reliability. This principled approach enables the model to appropriately attenuate less reliable modalities during the fusion process, offering more robust multi-modal integration when information quality varies across modalities (Han, Chen, and Ban 2024). Guided by these reliability scores, we perform residue-level fusion to achieve more trustworthy predictions.

Our main contributions are summarized as follows:

- We introduce MERA, the first retrieval-augmented framework for protein active site identification that employs residue-level MoE to dynamically retrieve and fuse contextual information from sequence, residue, and active-site views, yielding fine-grained enhancements tailored to individual residues.
- We propose a reliability-aware fusion strategy based on Dempster–Shafer evidence theory that quantifies modality trustworthiness through belief mass functions and learnable discounting coefficients, enabling principled multimodal integration for more robust active-site predictions.
- Comprehensive experiments validate the effectiveness of MERA on protein active site identification benchmarks, with additional protein-peptide binding site recognition experiments highlighting strong generalizability to more complex biological scenarios.

Related Work

Protein active site identification

Protein active sites are spatially clustered residues within proteins that are directly responsible for molecular recognition (Nussinov 2025), substrate binding (Ji et al. 2024) and catalyzing enzymatic reactions (Reisenbauer, Sicinski, and Arnold 2024). Accurate identification of these functionally critical regions is fundamental to mechanistic biology and therapeutic design, as they directly determine protein structural and functional roles. Existing methods for protein active site identification can be broadly categorized into single-modality and multi-modal approaches.

Single-modality models form the foundation of most current work and encompass both sequence-based and structure-based methods. Sequence-based models such as ESM-1b (Rives et al. 2021), ESM-1v (Meier et al. 2021), ESM-2 (Lin et al. 2022), and ProtTrans (Elnaggar et al. 2021) learn residue representations directly from amino acid sequences through masked language modeling, demonstrating strong generalization across diverse functional prediction tasks. Structure-based models utilize explicit 3D geometric information, typically employing graph neural networks or geometric encoders such as MIF (Yang, Zanichelli, and Yeh 2023) and PST (Chen et al. 2024), to improve structure-dependent site prediction. However, both approaches are inherently limited by relying on information from a single modality.

Multi-modal models address this limitation by incorporating auxiliary information from protein textual descriptions or complementary annotations. Recent approaches, including MMSite (Ouyang et al. 2024), UniSite (Fan et al. 2025), and ProtST (Xu et al. 2023), integrate protein–text pairs or fuse sequence and structure encoders to improve active site identification performance. Although these approaches align sequence and text features through cross-attention or end-to-end multimodal fusion strategies, most existing methods employ relatively shallow integration schemes that lack residue-level adaptive fusion and explicit reliability estimation. Consequently, they may not fully exploit the complementary strengths of each modality.

Retrieval-Augmented Generation

Recent works have explored retrieval-augmented and multimodal strategies to address the challenges of limited supervision and representation sparsity in protein modeling. MSM-Mut (Guo et al. 2024) enriches residue features by retrieving local structural motifs; RAPM (Wu et al. 2025) combines sequence and embedding-based retrieval for protein annotation; and ProTrek (Su et al. 2024) enables cross-modal retrieval. These approaches demonstrate that incorporating external or multi-modal information can effectively alleviate the under-determined nature of residue representations. However, existing retrieval-augmented and multimodal models typically employ static residue representations within rigid, task-specific pipelines. This design limitation restricts their ability to capture the localized, diverse, and context-dependent characteristics of active-site distributions, making it difficult to adaptively incorporate relevant information from multiple biological perspectives.

Methodology

Preliminaries

A protein sequence is represented as $\mathcal{S} = [s_1, \ldots, s_n]$ and its textual description as $\mathcal{T} = [t_1, \ldots, t_m]$, where s_i and t_j denote amino acids and text tokens, respectively. Multi-modal active-site identification is computed as $\hat{\mathbf{y}} = f_\theta(\mathcal{S}, \mathcal{T})$, where $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_n]$ is the output vector and

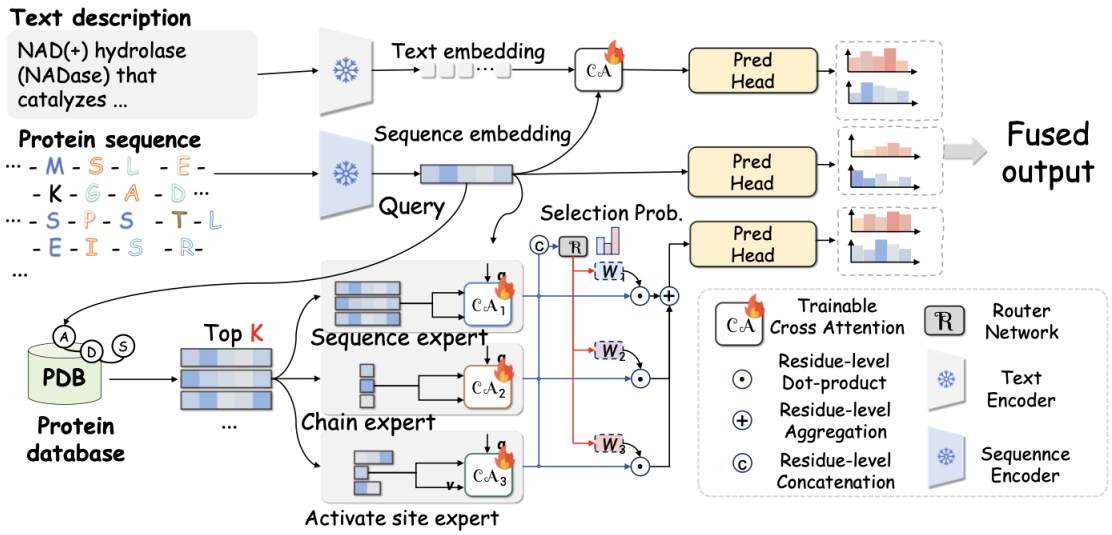


Figure 1: The framework of the proposed MERA.

each component $\hat{y}_i$ represents the likelihood that residue s_i is an active site. The function $f_\theta : \mathbb{R}^{n\times 1280} \times \mathbb{R}^{m\times 768} \to \{0,1\}^n$ is learned to perform this mapping.

Protein Vector Database. To enable retrieval-augmented modeling, we construct a protein vector database

$$\mathcal{D} = \left\{(\mathbf{h}^{\text{chain},j}, \mathbf{h}^{\text{seq},j}, \mathbf{h}^{\text{act},j})\right\}_{j=1}^{M},$$

where M denotes the total number of training proteins. For each protein j:

- $\mathbf{h}^{\text{chain},j} \in \mathbb{R}^{n_{chain}\times 1280}$ is the chain-level embedding computed as the mean of $\mathbf{h}^{\text{seq},j}$ and serves as the searchable key for database, where $n_{chain} = 1$.
- $\mathbf{h}^{\text{seq},j} \in \mathbb{R}^{n_{seq}\times 1280}$ is the residue-level embedding with n_{seq} residues.
- $\mathbf{h}^{\text{act},j} \in \mathbb{R}^{n_{act}\times 1280}$ contains n_{act} residue embeddings indicated by the ground-truth active-site mask, with $1 \le n_{act} \le n_{seq}$.

Online Retrieval. At inference, given a query protein $\mathcal{S}$ with chain embedding $\mathbf{h}^{\text{chain}}$, we retrieve the top-$K$ most similar neighbors from the database

$$\{\mathbf{h}^{e,k}\}_{k=1}^{K} = \underset{j\in\mathcal{D}, j\neq\text{query}}{\arg\max^{K}} \frac{\mathbf{h}^{\text{chain}} \cdot \mathbf{h}^{\text{chain},j}}{\|\mathbf{h}^{\text{chain}}\|\|\mathbf{h}^{\text{chain},j}\|},$$

where cosine similarity is utilized to measure the relatedness between query and database proteins. Each retrieved neighbor is represented as $\mathbf{h}^{e,k} = (\mathbf{h}^{\text{chain},k}, \mathbf{h}^{\text{seq},k}, \mathbf{h}^{\text{act},k})$.

Framework Overview

The overall framework is illustrated in Figure 1. For a protein sequence $\mathcal{S} = [s_1, \ldots, s_n]$, we use ESM-1b to obtain sequence-level embeddings $\mathbf{h}^{\text{seq}} \in \mathbb{R}^{n\times 1280}$ and compute a sequence key $\mathbf{h}^{\text{chain}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}^{\text{seq},i}$. The corresponding textual description $\mathcal{T}$ is first encoded by BioMedBERT into a text embedding $\tilde{\mathbf{h}}^{\text{text}} \in \mathbb{R}^{m\times 768}$. We then apply a cross-attention module to generate text-guidance embeddings for each residue, producing a text-enhanced residue representation $\mathbf{h}^{\text{text}} \in \mathbb{R}^{n\times 1280}$ aligned in length with sequence embedding $\mathbf{h}^{\text{seq}}$ to facilitate subsequent fusion. Next, we query the database with $\mathbf{h}^{\text{chain}}$ and retrieve the top-K neighbors $\{\mathbf{h}^{e,k}\}_{k=1}^{K}$. These neighbors and $\mathbf{h}^{\text{seq}}$ are fed into a Multi-expert RAG (MeRAG) module to produce an enhanced representation $\mathbf{h}^{\text{rag}} \in \mathbb{R}^{n\times 1280}$ that aggregates information from three complementary perspectives: sequence, chain, and active-site experts. The Reliability-aware Multimodal Fusion (RMF) module dynamically evaluates the reliability of three modalities for each residue: the original sequence modality $\mathbf{h}^{\text{seq}}$, RAG-enhanced $\mathbf{h}^{\text{rag}}$, and textual-enhanced $\mathbf{h}^{\text{text}}$. This enables the model to assign higher importance to more confident modalities at each position, resulting in a fused residue-level representation $\mathbf{h}^{\text{fused}} \in \mathbb{R}^{n\times 1280}$ that effectively balances information from all perspectives according to their estimated trustworthiness. This representation is finally mapped to the active-site probability vector $\hat{\mathbf{y}} \in [0,1]^n$.

Multi-expert RAG (MeRAG)

This section describes our Multi-expert RAG (MeRAG) module that leverages multiple expert perspectives to enhance residue-level representations through retrieval-augmented generation. Given a query protein sequence $\mathcal{S}$ with n residues, we extract sequence embeddings $\mathbf{h}_{\text{seq}}$ and the chain-level key $\mathbf{h}_{\text{chain}}$ to retrieve the top-K neighbors $\{\mathbf{h}^{e,k}\}_{k=1}^{K}$, where each neighbor k consists of $\mathbf{h}^{\text{seq},k}$, $\mathbf{h}^{\text{chain},k}$ and $\mathbf{h}^{\text{act},k}$.

Per-expert neighbor summarization. Three experts *seq*, *chain*, and *act* operate in parallel on every residue i of the query across the neighbors:

$$\mathbf{H}^{e} = \text{Expert}_e\left(\mathbf{h}^{\text{seq}}, \{\mathbf{h}^{e,k}\}_{k=1}^{K}\right), \quad e \in \{\text{seq}, \text{chain}, \text{act}\},$$

yielding expert outputs $\mathbf{H}^e \in \mathbb{R}^{n\times 1280}$ for calculating the retrieval-augmented enhancements from three perspectives. Rather than applying cross-attention directly between the query and all neighbor residues, we adopt a hierarchical (intra/inter) aggregation scheme that first denoises and summarizes local context within each neighbor before global integration. This design, inspired by recent advances in protein modeling and graph learning (Jumper et al. 2021; Fout et al. 2017; Borgeaud et al. 2022), reduces computational cost and overfitting while yielding more robust and interpretable retrieval-augmented representations.

Intra-neighbor aggregation. For each query residue i and retrieved neighbor $k = 1, \ldots, K$, we aggregate the n_k residue embeddings of that neighbor into a single vector $\mathbf{z}_i^{e,k} \in \mathbb{R}^{1\times 1280}$. Specifically, we weight every residue j within neighbor k by its similarity to query residue i:

$$\mathbf{z}_i^{e,k} = \sum_{j=1}^{n_k} \beta_{ij}^{e,k} \mathbf{h}_j^{e,k}, \quad \beta_{ij}^{e,k} = \frac{\exp\big(\mathbf{h}^{\text{seq},i} \cdot \mathbf{h}_j^{e,k}/\tau\big)}{\sum_{j'=1}^{n_k} \exp\big(\mathbf{h}^{\text{seq},i} \cdot \mathbf{h}_{j'}^{e,k}/\tau\big)},$$

where $\tau = 0.1$ is a temperature parameter that sharpens the attention weights $\beta_{ij}^{e,k}$, j indexes residues within neighbor k, and n_k is the length of neighbor k.

Inter-neighbor fusion. After obtaining K neighbor-level summaries $\{\mathbf{z}_i^{e,k}\}_{k=1}^K$, we fuse them with the query residue representation into a single expert output $\mathbf{H}_i^e \in \mathbb{R}^{1\times 1280}$. Specifically, we treat the query embedding as the 0-th candidate and compute a weighted sum over all $K+1$ candidates:

$$\mathbf{H}_i^e = \sum_{k=0}^{K} \gamma_{ik}^e \mathbf{z}_i^{e,k}, \quad \gamma_{ik}^e = \frac{\exp\big(\mathbf{h}_i^{\text{seq}} \cdot \mathbf{z}_i^{e,k}\big)}{\sum_{k'=0}^{K} \exp\big(\mathbf{h}_i^{\text{seq}} \cdot \mathbf{z}_i^{e,k'}\big)},$$

where γ_{ik}^e quantifies the contribution of neighbor k to the final expert output $\mathbf{H}_i^e$ and index $k = 0$ corresponds to the query residue itself, with $\mathbf{z}_i^{e,0} = \mathbf{h}_i^{\text{seq}}$.

Mixture-of-Experts gating. The three expert outputs are fed into an MoE layer to obtain the residue-level RAG-enhanced representation $\mathbf{h}^{\text{rag}} \in \mathbb{R}^{n\times 1280}$ via a residue-wise soft gate:

$$\mathbf{h}_i^{\text{rag}} = \text{MoE}\big(\mathbf{H}_i^{\text{seq}}, \mathbf{H}_i^{\text{chain}}, \mathbf{H}_i^{\text{act}}\big) = \sum_e g_i^e \mathbf{H}_i^e,$$

$$\mathbf{g}_i = \text{Softmax}\Big(\text{MLP}_\theta\big(\big[\mathbf{H}_i^{\text{seq}}; \mathbf{H}_i^{\text{chain}}; \mathbf{H}_i^{\text{act}}\big]\big)\Big) \in \mathbb{R}^{3\times 1280},$$

where $[\cdot;\cdot;\cdot]$ denotes concatenation and MLP_θ is a two-layer MLP that outputs expert selection probabilities.

This residue-level gating mechanism allows the model to adaptively integrate three expert perspectives at each residue position, enabling finer-grained fusion than applying a single global MoE to the entire sequence. This fine-grained adaptation is particularly important in capturing heterogeneous contributions of different experts across the protein, which is essential for modeling the diverse local contexts present in protein sequences. This approach is consistent with previous findings in token-wise adaptive gating in large-scale neural networks (Shazeer et al. 2017; Riquelme et al. 2021).

Reliability-aware Multimodal Fusion (RMF)

We now present our Reliability-aware Multimodal Fusion (RMF) module that employs evidence-theoretic methods to quantify modality reliability and achieve principled multimodal integration. Our framework integrates three complementary modalities: the sequence representation $\mathbf{h}^{\text{seq}} \in \mathbb{R}^{n\times 1280}$, the RAG-enhanced representation $\mathbf{h}^{\text{rag}} \in \mathbb{R}^{n\times 1280}$, and the text-guided representation $\mathbf{h}^{\text{text}} \in \mathbb{R}^{n\times 1280}$, where n denotes the number of residues. Inspired by Dempster–Shafer evidence theory (Amini et al. 2020; Huang et al. 2025; Deregnaucourt et al. 2025), we model each modality's reliability through belief mass functions and apply learnable discounting coefficients to enable principled fusion at each residue position.

Modality-specific prediction head. To enable reliability-aware fusion based on evidence theory, we deploy three parallel prediction heads, each specialized for a specific modality. For each modality $s \in \{\text{seq}, \text{rag}, \text{text}\}$, we define a modality-specific prediction head Pred_s that maps the residue representation $\mathbf{h}_i^s$ to a prediction logit $\hat{y}_i^s \in [0, 1]$:

$$\hat{y}_i^s = \text{Pred}_s(\mathbf{h}_i^s), \quad s \in \{\text{seq}, \text{rag}, \text{text}\},\ i = 1, \ldots, n.$$

Each Pred_s is implemented as a two-layer MLP with independent parameterization across modalities to capture modality-specific predictive patterns.

Modality Reliability Estimation. Following Dempster–Shafer evidence theory, we model each modality's prediction as a belief mass function and quantify its reliability through learnable discounting. For each residue i, the evidence mass is computed as:

$$m_i^s = \frac{\exp(\hat{y}_i^s)}{\sum_{s'} \exp(\hat{y}_i^{s'})}, \quad s \in \{\text{seq}, \text{rag}, \text{text}\},\ i = 1, \ldots, n,$$

where m_i^s the relative evidence strength of modality s for residue i.

To assess the modality trustworthiness beyond simple prediction confidence, we compute a credibility (discounting) coefficient following (Han, Chen, and Ban 2024):

$$c_i^s = \frac{1}{2}\left(m_i^s + 1 - \max_{s'\neq s} m_i^{s'}\right),$$

where $c_i^s \in [0, 1]$ serves as a learnable discounting factor that encourages high reliability only when modality s not only achieves strong evidence but also distinguishes itself from competing modalities. This design prevents unreliable modalities from dominating the fusion, addressing the core limitation of naive logit-based weighting.

Reliability Quantification. To convert credibility coefficients into fusion weights, we measure the reliability indicator $u_i^s \in [0, 1]$ for residue i in modality s as the normalized binary entropy of its credibility coefficient:

$$u_i^s = \frac{-c_i^s \ln c_i^s - (1 - c_i^s)\ln(1 - c_i^s)}{\ln 2},$$

where lower u_i^s values indicate higher reliability for modality s at residue i.

Reliability-guided Adaptive Fusion. At each residue i, we perform adaptive fusion using normalized weights derived from their reliability indicators:

$$e_i^s = \frac{\exp(-u_i^s)}{\sum_{s' \in \{\text{seq,rag,text}\}} \exp(-u_i^{s'})},$$

where e_i^s is the evidence-based fusion weight for modality s at residue i. The final prediction is computed as a reliability-weighted combination of the logits from all modalities:

$$\hat{y}_i = \sigma\left(e_i^{\text{seq}} \hat{y}_i^{\text{seq}} + e_i^{\text{rag}} \hat{y}_i^{\text{rag}} + e_i^{\text{text}} \hat{y}_i^{\text{text}}\right), \quad i = 1, \ldots, n,$$

where $\hat{y}_i^s$ denotes the logit from modality s, e_i^s denotes the reliability-based weight, and σ represents the element-wise sigmoid function.

Training Objective

The model is trained using a binary cross-entropy loss applied to the final fused prediction $\hat{y}_i$:

$$\mathcal{L}_{\text{bce}} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right].$$

A reliability regularization term encourages each modality-specific prediction to align with the ground truth:

$$\mathcal{L}_{\text{reliability}} = \sum_{s} \sum_{i=1}^{n} |\hat{y}_i^s - y_i|^2.$$

The total training objective is

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{reliability}}.$$

Experiments

Experimental Settings

Datasets. The ProTAD dataset introduced by MMSite (Ouyang et al. 2024) provides protein–text pairs with manually curated UniProt annotations. While these human-written descriptions offer high information density, their scarcity for newly discovered proteins limits ProTAD's applicability in real-world scenarios, potentially overestimating model performance. To bridge this gap, we introduce ProTAD-Gen, a challenging extension of ProTAD where all textual descriptions are automatically generated by ESM2Text. Crucially, the residue-level active-site labels remain identical to ProTAD, enabling fair performance comparison. By replacing human curation with automated text generation, ProTAD-Gen provides a more realistic benchmark for evaluating model robustness. To prevent data leakage, we preprocess ProTAD-Gen by cleaning and filtering the raw data, followed by MMseqs2 clustering at 10% sequence identity threshold (Steinegger and Söding 2017). To further validate the generalization capability of our framework, we also evaluate on the TS125 dataset, which is annotated with peptide-binding residues (Taherzadeh et al. 2016). Compared to single-sequence identification in ProTAD-Gen, binding site identification on TS125 protein–peptide complexes poses a much greater challenge. We generate textual descriptions for all TS125 samples using ESM2Text. The preprocessing steps for both datasets follow PepCA (Huang et al. 2024), including removal of redundant sequences using BLAST (Boratyn et al. 2013) and MMseqs2 (with a 30% sequence identity threshold). For both datasets, we use an 8:1:1 train/validation/test split, resulting in 5,569/697/697 samples for ProTAD (6,963 proteins), and 1,040/75/75 samples for TS125 (1,190 proteins).

Baselines. We evaluate MERA against state-of-the-art methods on two tasks:

- **Active site identification (ProTAD-Gen):** we compare against *sequence-only models*, including ESM-1b (Rives et al. 2021), ESM-1v (Meier et al. 2021), ESM-2 (Lin et al. 2022), ProtElectra (Elnaggar et al. 2021), PETA (Tan et al. 2024), TAPE (Rao et al. 2019), and S-PLM (Wang et al. 2025); *structure-aware models* PST (Chen et al. 2024) and MIF (Yang, Zanichelli, and Yeh 2023); and *multi-modal models* MMSite (Ouyang et al. 2024), ProtST (Xu et al. 2023), and UniSite (Fan et al. 2025).
- **Peptide-binding site prediction (TS125):** We evaluate against *sequence-based deep models* IIDL-PepPI (Chen et al. 2025), PepCA (Huang et al. 2024), PepBCL (Wang et al. 2022), PepCNN (Chandra et al. 2023), and PepNN (Abdin et al. 2022); *protein language models* ESM-2 (Lin et al. 2022) and TAPE (Rao et al. 2019); *structure-aware methods:* PepBind (Zhao, Peng, and Yang 2018) and PepSite (Petsalaki et al. 2009).

Results of all baselines are reproduced using publicly available code implementations with default hyperparameters to ensure fair evaluation.

Implementation. Protein sequences are encoded using ESM-1b (1280-dimensional embeddings), while textual descriptions are processed by BioMedBERT (768-dimensional embeddings). For inference, the original MMSite framework requires UniProt IDs to generate protein descriptions via Prot2Text. To enable broader applicability, particularly for novel proteins lacking UniProt annotations, we replace this with sequence-to-text models like ESM2Text, which directly generates functional annotations solely from raw sequence data. For retrieval, we select the top-3 nearest neighbors within each cluster, achieving an effective balance between computational efficiency and retrieval performance, as validated by our ablation studies. All experiments are conducted on two 40 GB NVIDIA A100 GPUs. Models are trained for 100 epochs, requiring approximately three hours for each task. For the TS125 dataset, we extend our framework with an additional peptide expert. We use the Adam optimizer with a learning rate of 1×10^{-3}. Our model contains approximately 300M parameters.

Evaluation. The model checkpoint achieving the highest AUPRC on the validation set is selected for final evaluation. During evaluation, we report F_{max}, AUPRC, AUROC, MCC, and Hits@k (where $k = 1, 5, 10$) metrics on the test set. F_{max} denotes the maximum F1-score across varying probability thresholds, indicating the best trade-off between precision and recall. AUPRC summarizes the

precision-recall relationship and is particularly informative for imbalanced datasets. AUROC, used for TS125, measures the model's ability to distinguish between binding and non-binding residues. MCC provides a balanced assessment of prediction quality by considering all four confusion matrix categories. Hits@k reflects the proportion of true active-site or binding-site residues ranked among the top k predictions for each protein, evaluating the model's ability to prioritize true sites at the top of its output.

Method	F_{max}	AUPRC	MCC	Hits@1	Hits@5	Hits@10
ESM-1b	0.71	0.85	0.71	0.76	0.88	0.90
ESM-1v	0.63	0.80	0.64	0.68	0.81	0.84
ESM-2-650M	0.65	0.82	0.66	0.70	0.83	0.86
ProtElectra	0.56	0.76	0.57	0.59	0.74	0.79
PETA	0.65	0.80	0.66	0.70	0.84	0.87
S-PLM	0.73	0.86	0.73	0.78	0.89	0.91
TAPE	0.36	0.54	0.36	0.38	0.56	0.61
MIF	0.14	0.35	0.14	0.16	0.29	0.34
ProtST	0.46	0.70	0.47	0.51	0.67	0.72
PST	0.66	0.81	0.66	0.71	0.84	0.87
UniSite	0.82	0.87	0.81	0.86	0.94	0.95
MMSite	0.81	0.87	0.83	0.88	0.95	0.95
MERA	**0.88**	**0.90**	**0.88**	**0.91**	**0.97**	**0.98**

Table 1: Results on the ProTAD-Gen dataset.

Main Results

Key Observation 1: Superior and robust active-site identification across benchmarks. MERA consistently achieves the best results on both ProTAD-Gen and TS125 datasets. On ProTAD-Gen, MERA obtains an AUPRC of 0.90 and F_{max} of 0.88, representing 3% and 7% improvements over **MMSite**, respectively. On the more challenging TS125 dataset, MERA achieves the highest AUROC of 0.85, demonstrating strong cross-task generalization capability. This superior performance is attributed to the use of retrieval-augmented residue representations and adaptive multimodal fusion based on reliability, which together yield more discriminative identification of protein active sites. **Key Observation 2: Strong real-world utility and biological relevance.** MERA demonstrates substantial improvements in ranking metrics, achieving 0.98 Hits@10 score and increasing Hits@1 by 3% on ProTAD-Gen and 6% on TS125 compared to the strongest baselines. This enables rapid and confident prioritization of candidate sites for experimental validation, substantially reducing the time and cost requirements for wet-lab studies. **Key Observation 3: Reliability estimation enables trustworthy multi-modal fusion.** Our reliability estimation provides a robust measure of prediction trustworthiness across modalities. Figure 2 illustrates a consistent monotonic relationship between reliability indicators and error rates for high-confidence predictions ($\hat{y} > 0.8$ or $\hat{y} < 0.2$) across three modalities. Lower reliability consistently corresponds to higher error rates, validating that our reliability quantification offers a principled criterion for assessing prediction quality. Unlike raw predicted probabilities that may poorly reflect true model confidence, our reliability-aware approach provides calibrated measures that enable trustworthy multimodal fusion decisions.

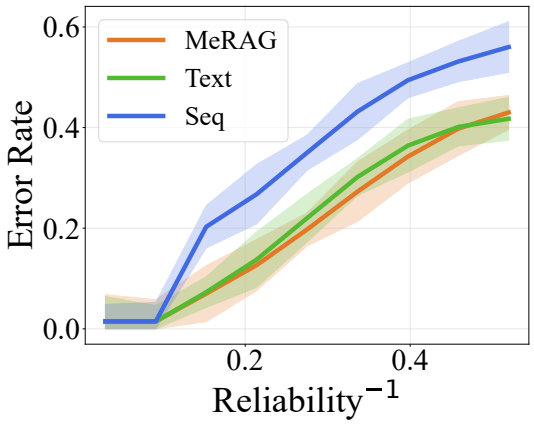
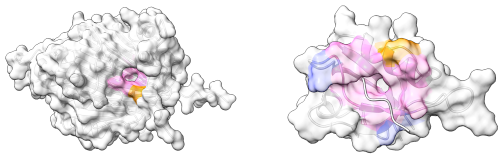


Figure 2: Comparison on ProTAD-Gen. **Left:** Precision recall curves. **Right:** Error rate vs. reliability indicator under high-confidence predictions ($\hat{y} > 0.8$ or $\hat{y} < 0.2$).



D3Z6H8 2PKU

Figure 3: Colors on the surface (residues) indicate correctly predicted sites (pink), unpredicted sites (yellow), and incorrectly predicted sites (blue).

Method	F_{max}	AUROC	MCC	Hits@1	Hits@5	Hits@10
Pepsite	0.30	0.61	0.30	0.35	0.62	0.70
PepCNN	0.33	0.68	0.33	0.40	0.67	0.75
PepBind	0.34	0.79	0.34	0.42	0.72	0.78
PepNN	0.31	0.73	0.31	0.39	0.68	0.76
TAPE	0.30	0.75	0.30	0.41	0.70	0.77
ESM2	0.32	0.77	0.32	0.43	0.72	0.80
PepBCL	0.33	0.79	0.33	0.44	0.73	0.81
PepCA	0.33	0.83	0.34	0.43	0.76	0.84
IIDL-PepPI	0.35	0.84	0.35	0.46	0.77	0.85
MERA	**0.40**	**0.85**	**0.41**	**0.52**	**0.79**	**0.86**

Table 2: Results on the TS125 dataset.

Case study

We select two representative cases from different test sets. For S-adenosylmethionine decarboxylase proenzyme 2 (UniProt: D3Z6H8), MERA successfully identified most active sites but missed one position at the edge of the binding region. This oversight is visualized in the Figure 3 A using AlphaFold 3, highlighting the structural context of the missed site. Despite this minor error, the overall prediction aligns well with the known functional residues. For the PICK1 PDZ domain (PDB: 1PKU), which involves peptide

binding. MERA correctly predicted the majority of the active sites in the primary binding region, as shown in Figure 3B. This demonstrates its effectiveness in capturing key interaction sites. However, some peripheral positions were either unpredicted or incorrectly identified, reflecting potential limitations in capturing less critical or structurally ambiguous regions. These cases provide valuable insights into the strengths and limitations of MERA, particularly in handling edge cases and distinguishing between essential and secondary binding sites. Overall, these examples underscore the robustness of our approach while identifying areas for future improvement.

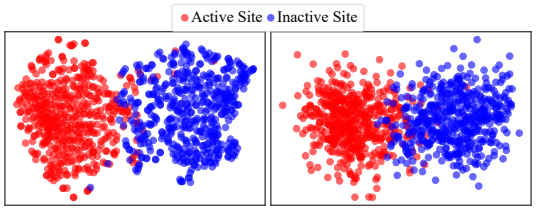
Method	F_{max}	AUPRC	MCC	Hits@1	Hits@10
MERA	**0.88**	**0.90**	**0.88**	**0.91**	**0.98**
w/o RMF	0.70	0.83	0.70	0.75	0.88
w/o Text modality	0.83	0.86	0.83	0.87	0.94
w/o RAG modality	0.79	0.86	0.80	0.86	0.92
w Seq Modality Only	0.71	0.85	0.71	0.76	0.90
w/o MeRAG	0.76	0.85	0.76	0.84	0.90
w/o Sequence Expert	0.83	0.87	0.84	0.89	0.95
w/o Chain Expert	0.84	0.86	0.83	0.87	0.93
w/o Active-site Expert	0.85	0.87	0.84	0.90	0.95

Table 3: Ablation study on the ProTAD-Gen dataset.

Ablation Study

Contribution of Each Module. Table 3 presents the ablation results for different modules on the ProTAD-Gen dataset. **Key Observation 1: Multimodal fusion is essential but requires principled integration.** Removing the RMF module entirely leads to the most severe performance degradation (AUPRC drops from 0.90 to 0.83), confirming that naive single-modality approaches are fundamentally inadequate. When modalities are integrated without proper reliability assessment, performance can degrade significantly, highlighting the importance of our reliability-aware fusion strategy. **Key Observation 2: Introducing additional modalities substantially enhances sequence-only predictions.** Both text and RAG modalities provide substantial improvements over the sequence-only baseline, demonstrating the effectiveness of incorporating auxiliary information sources. **Key Observation 3: MeRAG fusion enables effective expert integration.** Replacing our residue-level MeRAG fusion with direct expert combination leads to notable performance degradation, indicating that naive fusion of RAG experts introduces noise and fails to capture position-specific contributions. Our MeRAG mechanism is crucial for effectively utilizing retrieval-augmented information through adaptive expert weighting at each residue position. **Key Observation 4: Each expert provides complementary retrieval perspectives.** Removing any single expert consistently degrades performance, confirming the complementarity of multi-expert retrieval approach.

Visualization of Embedding Discriminability. To further visualize the distinguishability of active and inactive sites with and without the MeRAG module, we randomly sampled 500 active-site residues and 500 inactive-site residues from the test set and projected them into a 2D space for comparison. As shown in Figure 4, the full MeRAG framework produces much clearer separation and clustering between active (red) and inactive (blue) sites, whereas removing the MeRAG module leads to increased overlap between the two classes. This demonstrates the advantage of our approach in learning more discriminative residue-level representations.



(a) With MeRAG fusion (b) Without MeRAG fusion

Figure 4: Visualization of inactive (blue) and active (red) site embeddings w/ and w/o MeRAG fusion on ProTAD-Gen.

Method	F_{max}	AUROC	MCC	Hits@10
IIDL-PepPI	0.35	0.84	0.35	0.85
MERA (4 experts)	**0.40**	**0.85**	**0.41**	**0.86**
w/o Peptide Expert (3 experts)	0.37	0.84	0.38	0.83

Table 4: Ablation study of the number of expert modules on the TS125 dataset.

Evaluating Flexibility of MERA. We conduct an ablation study on the TS125 dataset (Table 4) to demonstrate the flexibility and generalization of our approach in complex peptide–protein binding scenarios. The results show that incorporating a specialized peptide expert (four experts total) further enhances the performance compared to the three-expert variant. This improvement underscores the adaptability of the MERA framework, as new experts can be flexibly integrated to handle diverse biological tasks. Importantly, even with only three experts, our model still outperforms the strongest baseline (IIDL-PepPI) across all evaluation metrics, further highlighting the robustness and strong generalization ability of our framework.

Conclusion

We introduced MERA, the first retrieval-augmented framework for protein active-site prediction that employs residue-level mixture-of-experts. By coupling hierarchical multi-expert retrieval with reliability-aware multimodal fusion, MERA addresses the dual challenges of single-instance prediction vulnerability and inadequate modality reliability estimation that plague current methods. The framework not only achieves the best performance on benchmarks, but also demonstrates unprecedented flexibility. Future work will extend MERA to incorporate additional modalities, notably 3D structural information (Kim et al. 2025), by introducing dedicated structure experts within the MeRAG module.

Acknowledgements

This work was supported by Zhejiang Leading Innovative and Entrepreneur Team Introduction Program

(2024R01007).

References

Abdin, O.; Nim, S.; Wen, H.; and Kim, P. M. 2022. PepNN: a deep attention model for the identification of peptide binding sites. *Communications Biology*, 5(1): 503.

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33: 14927–14937.

Boratyn, G. M.; Camacho, C.; Cooper, P. S.; Coulouris, G.; Fong, A.; Ma, N.; Madden, T. L.; Matten, W. T.; McGinnis, S. D.; Merezhuk, Y.; et al. 2013. BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, 41(W1): W29–W33.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2206–2240. PMLR.

Chandra, A.; Sharma, A.; Dehzangi, I.; Tsunoda, T.; and Sattar, A. 2023. PepCNN deep learning tool for predicting peptide binding residues in proteins using sequence, structural, and language model features. *Scientific Reports*, 13(1): 20882.

Chen, D.; Hartout, P.; Pellizzoni, P.; Oliver, C.; and Borgwardt, K. 2024. Endowing protein language models with structural knowledge. *arXiv preprint arXiv:2401.14819*.

Chen, S.; Yan, K.; Li, X.; and Liu, B. 2025. Protein language pragmatic analysis and progressive transfer learning for profiling peptide–protein interactions. *IEEE Transactions on Neural Networks and Learning Systems*.

Deregnaucourt, L.; Laghmara, H.; Lechervy, A.; and Ainouz, S. 2025. A Conflict-Guided Evidential Multimodal Fusion for Semantic Segmentation. In *WACV*, 1373–1382. IEEE.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7112–7127.

Fan, J.; Wu, Q.; Luo, S.; and Wang, L. 2025. UniSite: The First Cross-Structure Dataset and Learning Framework for End-to-End Ligand Binding Site Detection. *arXiv preprint arXiv:2506.03237*.

Fout, A.; Byrd, J.; Shariat, B.; and Ben-Hur, A. 2017. Protein interface prediction using graph convolutional networks. *Advances in Neural Information Processing Systems*, 30: 6533–6542.

Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; et al. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1): 3168.

Guo, R.; Wang, R.; Wu, R.; Ren, Z.; Li, J.; Luo, S.; Wu, Z.; Liu, Q.; Peng, J.; and Ma, J. 2024. Enhancing protein mutation effect prediction through a retrieval-augmented framework. *Advances in Neural Information Processing Systems*, 37: 49130–49153.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, 3929–3938. PMLR.

Han, X.; Chen, F.; and Ban, J. 2024. FMFN: A fuzzy multimodal fusion network for emotion recognition in ensemble conducting. *IEEE Transactions on Fuzzy Systems*, 33(1): 168–179.

Han, Z.; Yang, F.; Huang, J.; Zhang, C.; and Yao, J. 2022. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 20707–20717.

Huang, J.; Li, W.; Xiao, B.; Zhao, C.; Zheng, H.; Li, Y.; and Wang, J. 2024. PepCA: Unveiling protein-peptide interaction sites with a multi-input neural network model. *iScience*, 27(10): 110850.

Huang, L.; Ruan, S.; Decazes, P.; and Denœux, T. 2025. Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation. *Information Fusion*, 113: 102648.

Ji, W.; Miao, A.; Liang, K.; Liu, J.; Qi, Y.; Zhou, Y.; Duan, X.; Sun, J.; Lai, L.; and Wu, J.-X. 2024. Substrate binding and inhibition mechanism of norepinephrine transporter. *Nature*, 633(8029): 473–479.

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.

Kim, G.; Lee, S.; Levy Karin, E.; Kim, H.; Moriwaki, Y.; Ovchinnikov, S.; Steinegger, M.; and Mirdita, M. 2025. Easy and accurate protein structure prediction using ColabFold. *Nature Protocols*, 20(3): 620–642.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022: 500902.

Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; and Rives, A. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34: 29287–29303.

Nussinov, R. 2025. Pioneer in Molecular Biology: Conformational Ensembles in Molecular Recognition, Allostery, and Cell Function. *Journal of Molecular Biology*, 169044.

Ouyang, S.; Cai, H.; Luo, Y.; Su, K.; Zhang, L.; and Du, B. 2024. MMSite: A Multi-modal Framework for the Identification of Active Sites in Proteins. *Advances in Neural Information Processing Systems*, 37: 45819–45849.

Petrova, N. V.; and Wu, C. H. 2006. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC bioinformatics*, 7(1): 312.

Petsalaki, E.; Stark, A.; García-Urdiales, E.; and Russell, R. B. 2009. Accurate prediction of peptide binding sites on protein surfaces. *PLoS computational biology*, 5(3): e1000335.

Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; and Song, Y. 2019. Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems*, 32: 9689–9701.

Reisenbauer, J. C.; Sicinski, K. M.; and Arnold, F. H. 2024. Catalyzing the future: recent advances in chemical synthesis using enzymes. *Current Opinion in Chemical Biology*, 83: 102536.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shen, J.; Wu, J.; Zhang, Y.; Zhu, K.; Wang, K.; Hu, W.; Hou, K.; Qian, K.; Zhang, X.; and Hu, B. 2025. MF-Net: Exploring a Meta-Fuzzy Multimodal Fusion Network for Depression Recognition. *IEEE Transactions on Fuzzy Systems*.

Steinegger, M.; and Söding, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11): 1026–1028.

Su, J.; Zhou, X.; Zhang, X.; and Yuan, F. 2024. ProTrek: Navigating the protein universe through tri-modal contrastive learning. *bioRxiv*, 2024–05.

Taherzadeh, G.; Yang, Y.; Zhang, T.; Liew, A. W.-C.; and Zhou, Y. 2016. Sequence-based prediction of protein–peptide binding sites using support vector machine. *Journal of Computational Chemistry*, 37(13): 1223–1229.

Tan, Y.; Li, M.; Zhou, Z.; Tan, P.; Yu, H.; Fan, G.; and Hong, L. 2024. PETA: evaluating the impact of protein transfer learning with sub-word tokenization on downstream applications. *Journal of Cheminformatics*, 16(1): 92.

Wang, D.; Pourmirzaei, M.; Abbas, U. L.; Zeng, S.; Manshour, N.; Esmaili, F.; Poudel, B.; Jiang, Y.; Shao, Q.; Chen, J.; et al. 2025. S-PLM: Structure-Aware Protein Language Model via Contrastive Learning Between Sequence and Structure. *Advanced Science*, 12(5): 2404212.

Wang, R.; Jin, J.; Zou, Q.; Nakai, K.; and Wei, L. 2022. Predicting protein–peptide binding residues via interpretable deep learning. *Bioinformatics*, 38(13): 3351–3360.

Wu, J.; Liu, Z.; Cao, H.; Li, H.; Feng, B.; Shu, Z.; Yu, K.; Yuan, L.; and Li, Y. 2025. Rethinking Text-based Protein Understanding: Retrieval or LLM? *arXiv preprint arXiv:2505.20354*.

Xu, M.; Yuan, X.; Miret, S.; and Tang, J. 2023. ProtST: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, 38749–38767. PMLR.

Yang, K. K.; Zanichelli, N.; and Yeh, H. 2023. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36: gzad015.

Yang, M.; Li, Y.; Zhang, C.; Hu, P.; and Peng, X. 2024. Test-time adaptation against multi-modal reliability bias. In *The International Conference on Learning Representations*.

Zhang, Q.; Wei, Y.; Han, Z.; Fu, H.; Peng, X.; Deng, C.; Hu, Q.; Xu, C.; Wen, J.; Hu, D.; et al. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Zhao, Z.; Peng, Z.; and Yang, J. 2018. Improving sequence-based prediction of protein–peptide binding residues by introducing intrinsic disorder and a consensus method. *Journal of Chemical Information and Modeling*, 58(7): 1459–1468.