ECMLF: AN EXPLAINABLE CROSS-MODAL LEARNING FRAMEWORK FOR COLD-START PROTEIN-PROTEIN INTERACTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cold-start protein-protein interaction (PPI) prediction aims to identify potential interactions involving newly discovered proteins that lack sufficient experimentally observed interaction records, thereby facilitating the functional characterization of these proteins. Its central challenge is to capture the similarity between new and known proteins using information from diverse modalities. However, existing methods have three limitations in capturing such similarity: (1) only partial modalities are considered, resulting in incomplete or biased similarity modeling; (2) the similarity between new and known proteins is computed offline and independently across modalities, causing inconsistency between similarity modeling and cold-start PPI prediction; (3) current explanations are largely single-modal and feature-oriented, overlooking biologically grounded evidence for interaction mechanisms. To address these issues, we propose an explainable cross-modal learning framework (ECMLF) with three components. First, comprehensive modalities of intrinsic amino acid sequences and four external biomedical knowledge graphs are leveraged to capture similarity between new and known proteins. Second, an adaptive gated fusion module is designed to combine similarity modeling and cold-start PPI prediction within an end-to-end learning framework. Finally, a dual-pathway explanation mechanism is proposed to provide biologically grounded explanations of interaction mechanisms. Extensive experiments on two real datasets show that ECMLF achieves superior prediction accuracy and explanation quality compared with existing methods for cold-start PPI prediction. The source code and datasets are submitted as Supplementary Materials for Reproducibility.

1 INTRODUCTION

Protein-protein interaction (PPI) is fundamental to cellular functions, disease mechanisms, and therapeutic discovery. The costly and time-consuming nature of experimental PPI identification motivates the development of computational methods for efficient and large-scale PPI prediction (Rapti et al., 2026; Lambourne et al., 2026; Wang et al., 2025b). Despite advances in computational prediction, most methods focus on PPI prediction involving known proteins, relying on observed interactions or sufficient network neighborhood information.

However, biomedical research, including cellular function studies, drug development, or disease treatment, often involves newly discovered proteins that lack known PPI records. Accurately predicting their interactions can facilitate protein function annotation, reveal disease-related mechanisms, and identify potential therapeutic targets. Motivated by these important applications, predicting potential interactions involving new proteins has attracted considerable attention worldwide. This task is commonly referred to as cold-start PPI prediction. Unlike conventional PPI prediction, it requires capturing similarity between known and new proteins to uncover latent associations, thereby enabling the transfer of PPI information from known to new proteins.

Yet, existing methods still face three limitations in characterizing such similarity and explaining predicted interactions. Firstly, they rely on a limited set of modalities, overlooking complementary biological information and yielding an incomplete or biased view of similarity between known and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

new proteins (Li et al., 2025; Cihan et al., 2025; Silva et al., 2025). Secondly, similarity between known and new proteins is often precomputed (Li et al., 2026) independently for each modality, leaving similarity modeling disconnected from training and misaligned with the cold-start PPI prediction objective. Thirdly, existing explanations focus on feature contributions within individual modalities, overlooking biological evidence linking the contributing features to interaction mechanisms and offering limited insight into how predicted interactions may arise.

To address these limitations, we propose ECMLF, an explainable cross-modal learning framework for cold-start PPI prediction. Initially, it combines intrinsic amino acid sequences with four external biomedical knowledge graphs, capturing complementary biological information to characterize similarity between known and new proteins. Subsequently, it integrates representations across modalities through an adaptive gated fusion module, jointly optimizing similarity modeling and cold-start PPI prediction within an end-to-end framework. Finally, it employs a dual-pathway explanation mechanism to identify interaction-related residue regions and supporting graph evidence, providing biologically grounded insights into the mechanisms underlying predicted interactions.

We summarize our main contributions as follows:

- **ECMLF Model.** We introduce ECMLF, an explainable cross-modal learning framework for accurate cold-start PPI prediction and biologically grounded explanations of potential interaction mechanisms.
- **Cross-Modal Representation Learning.** We design a cross-modal representation learning module to capture complementary biological information from intrinsic amino acid sequences and four external biomedical knowledge graphs, enabling similarity modeling between known and new proteins.
- **Jointly Optimized Cross-Modal Fusion.** We design an adaptive gated fusion module to integrate representations across modalities, aligning similarity modeling with cold-start PPI prediction through end-to-end joint optimization.
- **Dual-Pathway Explainability Mechanism.** We introduce a dual-pathway explanation mechanism that uncovers interaction-related residue regions and biomedical bridge entities, supporting the biological interpretation of predicted PPI.

2 Related Work

2.1 Amino Acid Sequence-Based Methods

Amino acid sequences provide intrinsic information for PPI prediction, enabling models to capture residue patterns, structural cues, and potential binding regions. Recent protein foundation models improve representations of protein structure, function, and functional sites (Xiao et al., 2025; Wang et al., 2025d; Porollo et al., 2026). ISM further enriches sequence representations by distilling structure-derived tokens into a protein language model (Ouyang-Zhang et al., 2025). Moreover, other PPI models learn interaction-oriented features from paired sequences or pretrained embeddings (Liu et al., 2025; Ullanat et al., 2026; Liu et al., 2026). For instance, BaPPI (Tang et al., 2024) uses an anti-symmetric balanced learning framework to capture directional imbalance and pairwise characteristics, xCAPT5 (Dang & Vu, 2024) combines ProtT5 embeddings with multi-kernel Siamese convolutions to extract multi-scale sequence patterns, and TUnA (Ko et al., 2024) learns uncertainty-aware sequence-context representations for prediction involving new proteins. Building on these sequence-based advances, ESM2_AMP (Sun et al., 2025) integrates ESM2-derived segment features through Transformer fusion to identify interaction-relevant regions and support interpretability. Other studies explore multimodal, sampling-aware, or task-adapted sequence models for generalized and virus-host PPI prediction (Wang et al., 2025a; Li et al., 2025; Cihan et al., 2025).

Although effective at capturing local residue-level signals, these methods typically operate within a single modality. Relying solely on intrinsic sequence information, they provide an incomplete view of similarity between known and new proteins. This limitation becomes pronounced in cold-start settings, where new proteins have few or no observed PPI records. Lacking complementary biomedical context, sequence-only representations struggle to capture cross-modal associations and transfer interaction knowledge from known proteins, limiting cold-start generalization.

2.2 GRAPH-BASED METHODS

Graph-based methods model interaction topology, structural contacts, and biomedical associations for PPI prediction, while MM-Lego integrates heterogeneous biomedical encoders through representation alignment and model fusion (Wang et al., 2025c; Hemker et al., 2025). Representative models include HNSPPI (Xie et al., 2023), depending on observed PPI topology; KGF-GNN (Yang et al., 2024), capturing semantic and interaction structure; and HEENN (Yang et al., 2025), fusing local and global graph representations. More recently, GATSBI (Nayar & Altman, 2026) incorporates ESM-2 embeddings with multiple biological networks through graph attention, while biomedical knowledge graphs transfer cross-entity associations (Zhang et al., 2025; Xiong et al., 2025; Wang et al., 2025e). Structure-aware models encode geometric evidence for structure learning and binding-site identification (Li et al., 2026; Arteaga et al., 2026; Mallet et al., 2025), while geometric learning supports structure generation and docking (Stark et al., 2025; Corso et al., 2025).

Nevertheless, graph-based methods rely on observed neighborhoods, producing weakly supported embeddings for new proteins with sparse or missing links. Computing protein similarity offline and independently across modalities further disconnects similarity modeling from the prediction objective. These limitations leave complementary graph evidence underused, hindering the discovery of latent associations between known and new proteins, thereby limiting cold-start prediction.

2.3 INTERPRETABILITY FOR PPI MECHANISMS

Interpretability analysis links PPI predictions to biomedical mechanisms by connecting model decisions with biological evidence. For instance, ESM2_AMP (Sun et al., 2025) identifies attention-highlighted sequence features; InterPLM and ProteinSAE (Simon & Zou, 2025; Gujral et al., 2025) associate latent or sparse language-model features with biological properties; PLMFlaw2026 (Szymborski & Emad, 2026) detects misleading shortcuts and weak interaction-specific evidence in sequence models; KGF-GNN and HEENN (Yang et al., 2024; 2025) visualize graph neighborhoods to explain patterns underpinning predictions. Overall, these methods primarily emphasize feature attribution or salient graph neighborhoods (Sun et al., 2025; Yang et al., 2025; Amato et al., 2026).

However, existing explanations remain largely feature-oriented, highlighting sequence residues or graph neighborhoods without jointly relating them to biological interaction mechanisms. They rarely connect salient regions to plausible interfaces in protein complexes or explain how similarity between known and new proteins is formed and transferred. Consequently, such explanations remain disconnected from the biological mechanisms underlying predicted interactions and provide limited support for their biological plausibility.

3 PRELIMINARIES

Amino Acid Sequence Information. Amino acid sequences are intrinsic protein attributes that contain clues to potential interaction sites. To accurately learn sequence regions associated with potential protein-binding sites, we partition each sequence according to structural region annotations, preserving local structural context. For protein p_i, the resulting sequence is denoted as $\mathbf{S}_i = [s_{i_1}, \ldots, s_{i_n}]$, where s_{i_j} represents the j-th region. These region-based representations enable the model to learn sequence regions associated with potential protein-binding sites.

Multimodal Biomedical Knowledge Graphs. Biomedical knowledge graphs provide external biological context by representing relationships between proteins and diverse biomedical entities, complementing intrinsic sequence information. We define the graph as $\mathbf{G} = (\mathcal{P}, \mathcal{E}, \mathcal{R})$, where $\mathcal{P}$, $\mathcal{E}$, and $\mathcal{R}$ denote proteins, biomedical entities, and their relations, respectively.

Problem Formulation. Let $\mathcal{P}$ denote the protein set, partitioned into the known-protein set $\mathcal{P}_{\text{known}}$ and the new-protein set $\mathcal{P}_{\text{new}}$. The PPI matrix is defined as $\mathbf{Y} \in \{0,1\}^{N_{\mathcal{P}} \times N_{\mathcal{P}}}$, where $y_{ij} = 1$ indicates that an interaction exists between p_i and p_j, and $y_{ij} = 0$ otherwise. During training, only interactions between known proteins are used for supervision, and the prediction function is defined as $\hat{y}_{ij} = \Gamma(p_i, p_j \mid \Theta, \mathbf{Y}_{\mathcal{P}_{\text{known}}}, \mathbf{S}, \mathbf{G})$, where Θ denotes the parameters and $\Gamma(\cdot)$ integrates information from $\mathbf{Y}_{\mathcal{P}_{\text{known}}}$, $\mathbf{S}$ and $\mathbf{G}$ to predict the interaction between p_i and p_j. To predict the cold-start protein-protein interaction, we consider two settings: semi-cold-start (SCS), where $p_i \in \mathcal{P}_{\text{known}}$ and $p_j \in \mathcal{P}_{\text{new}}$; and full-cold-start (FCS), where $p_i, p_j \in \mathcal{P}_{\text{new}}$.

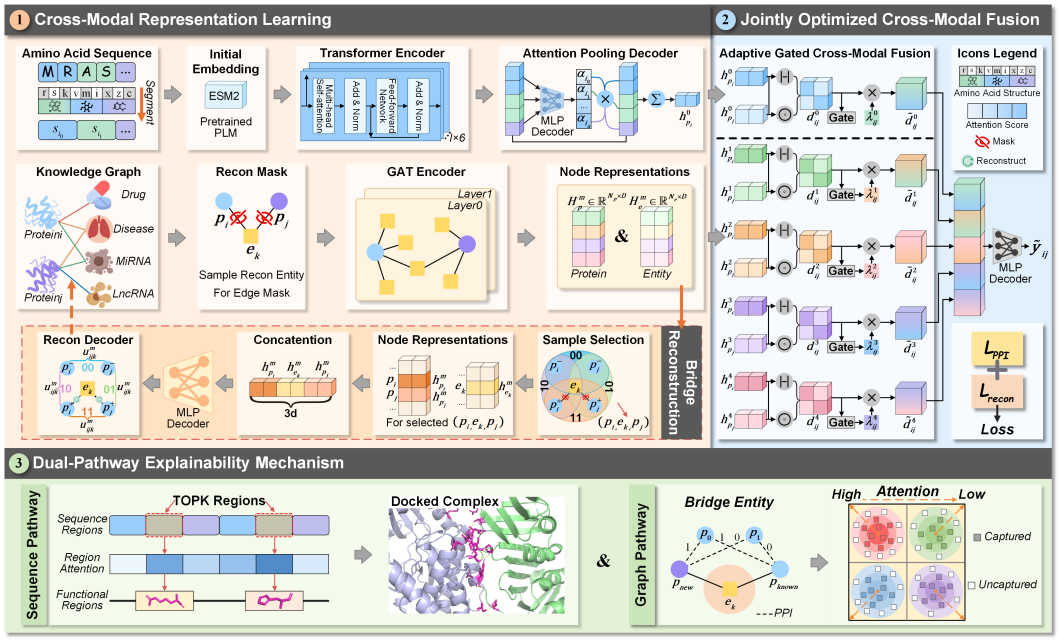


Figure 1: Overview of the proposed ECMLF structure.

4 Proposed Method

Overview. As shown in Figure 1, ECMLF consists of three components: (1) a cross-modal representation learning module, (2) a jointly optimized cross-modal fusion module, and (3) a dual-pathway explainability mechanism module. These modules together enable cold-start PPI prediction and facilitate biological interpretation of the predicted interactions within a unified framework.

4.1 Cross-Modal Representation Learning

4.1.1 Region-Aware Amino Acid Sequence Encoder

Amino acid sequences. Amino acid sequences provide the primary intrinsic description of proteins, as they determine conserved domains, functional motifs, and residue regions that may participate in protein molecular binding. For proteins without observed PPI records, sequence information is especially important because it remains available even when relational neighborhoods are missing. Therefore, the sequence encoder is designed to capture local interaction patterns while retaining the structural organization of amino-acid sequences.

For protein p_i, its sequence is partitioned into n structural regions $\mathbf{S}_i = [s_{i_1}, \ldots, s_{i_n}]$, which are encoded by a pretrained PLM into $\mathbf{X}_i = [\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}]^\top \in \mathbb{R}^{n \times D_{\text{PLM}}}$. These embeddings are projected into a D-dimensional latent space, and learnable positional embeddings are added to encode region order. A six-layer Transformer then captures contextual dependencies across regions:

$$\mathbf{H}_i = \text{Norm}\Big(\mathcal{T}^{(6)}_{\text{encoder}}(\text{Norm}(\mathbf{X}_i)\mathbf{W}_s + \mathbf{b}_s + P_i)\Big), \quad \mathbf{H}_i \in \mathbb{R}^{n \times D}, \tag{1}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are learnable projection parameters, P_i denotes the positional embedding, $\mathcal{T}^{(6)}_{\text{encoder}}$ denotes the Transformer encoder, and Norm denotes layer normalization. Each row of $\mathbf{H}_i$ corresponds to the contextualized embedding $h_{i_j} \in \mathbb{R}^D$ of region s_{i_j}. To highlight regions potentially involved in protein-protein contacts, an attention pooling layer is applied to compute normalized weights for their contextualized representations:

$$\alpha_{i_j} = \frac{\exp\big(w_{\text{seq}}^\top \phi(\mathbf{W}_{\text{seq}} h_{i_j} + \mathbf{b}_{\text{seq}})\big)}{\sum_{k=1}^{n} \exp\big(w_{\text{seq}}^\top \phi(\mathbf{W}_{\text{seq}} h_{i_k} + \mathbf{b}_{\text{seq}})\big)}, \tag{2}$$

where $\mathbf{W}_{\text{seq}}$ and $\mathbf{b}_{\text{seq}}$ denote the learnable weight matrix and bias vector, $\mathbf{w}_{\text{seq}}$ is a learnable attention vector, and $\phi(\cdot)$ denotes the $\tanh$ activation function. The attention weights quantify regional importance and guide aggregation into a protein embedding for cross-modal fusion:

$$h_i^0 = \sum_{j=1}^{n} \alpha_{i_j} h_{i_j}, \quad h_i^0 \in \mathbb{R}^D. \tag{3}$$

4.1.2 Multimodal Biomedical Knowledge Graph Encoder

Multimodal Biomedical Knowledge Graph. Unlike sequences, biomedical knowledge graphs represent proteins through links to biomedical entities, thereby encoding functional, disease, and molecular associations. To characterize proteins from complementary biomedical perspectives and provide evidence for modeling similarity among proteins, we construct four biomedical knowledge graphs covering *protein-drug*, *protein-disease*, *protein-miRNA*, and *protein-lncRNA associations*.

To capture such latent similarities, we randomly select $R\%$ entity nodes and mask a subset of protein-entity edges incident to each selected node, yielding an incomplete graph $\widetilde{\mathbf{G}}^m$ for reconstruction below. After edge masking, the relation matrix $\widetilde{\mathbf{R}}^m \in \mathbb{R}^{N_p \times N_e^m}$ is mapped to initial protein and entity embeddings for modality m. A two-layer GAT incorporates neighborhood information into these embeddings, weighting neighboring nodes according to their contributions to each update. For node u and its neighbor $v \in \tilde{\mathcal{N}}_u^m$, the normalized attention coefficient is defined as:

$$\alpha_{uv}^{m(l)} = \frac{\exp\Big(\psi\Big(\mathbf{a}^{m(l)\top}[\mathbf{W}^{m(l)} h_u^{m(l)} \| \mathbf{W}^{m(l)} h_v^{m(l)}]\Big)\Big)}{\sum_{q \in \tilde{\mathcal{N}}_u^m} \exp\Big(\psi\Big(\mathbf{a}^{m(l)\top}[\mathbf{W}^{m(l)} h_u^{m(l)} \| \mathbf{W}^{m(l)} h_q^{m(l)}]\Big)\Big)}, \tag{4}$$

where $\mathbf{W}^{m(l)}$ ($l \in \{0, 1\}$) is a learnable transformation matrix, $\mathbf{a}^{m(l)}$ is a learnable attention vector, $\psi(\cdot)$ denotes the LeakyReLU activation function, and $\|$ denotes vector concatenation. These coefficients weight the transformed neighboring features, updating the representation of node u as:

$$h_u^{m(l+1)} = \delta\left(\sum_{v \in \tilde{\mathcal{N}}_u^m} \alpha_{uv}^{m(l)} \mathbf{W}^{m(l)} h_v^{m(l)}\right), \tag{5}$$

where $\delta(\cdot)$ denotes the ELU activation function. After two layers of aggregation, a linear projection maps the resulting node embeddings to D-dimensional graph representations:

$$h_u^m = \mathbf{W}_{\text{out}}^m h_u^{m(2)} + \mathbf{b}_{\text{out}}^m, \quad h_u^m \in \mathbb{R}^D. \tag{6}$$

Bridge Reconstruction. Shared biomedical entities may reflect common functional roles, disease associations, or drug-response patterns, providing relational evidence of latent protein similarity. To learn these associations, we formulate bridge reconstruction as a four-class task that predicts the association states of two proteins. For each selected entity e_k (all of its edges have been masked), we sample two distinct neighbors p_i^+, p_j^+ and two distinct non-neighbors p_i^-, p_j^- from the original graph to form four types of reconstruction triples: (p_i^-, e_k, p_j^-), (p_i^-, e_k, p_j^+), (p_i^+, e_k, p_j^-), and (p_i^+, e_k, p_j^+). The reconstruction label is defined as $r_{ijk}^m \in \{00, 01, 10, 11\}$, with the two bits indicating whether the left and right proteins are connected to the middle entity, respectively (1 denotes an observed connection and 0 denotes no observed connection). The node representations obtained from the two-layer GAT encoder are concatenated and fed into a four-class reconstruction decoder:

$$\mathbf{u}_{ijk}^m = \mathbf{W}_r^{m(2)} \Phi\Big(\mathbf{W}_r^{m(1)}[h_{p_i}^m \| h_{e_k}^m \| h_{p_j}^m] + \mathbf{b}_r^{m(1)}\Big) + \mathbf{b}_r^{m(2)}, \tag{7}$$

where $\Phi(\cdot)$ denotes ReLU activation function, $\mathbf{u}_{ijk}^m \in \mathbb{R}^4$ denotes the logits for classes $C = \{0, 1, 2, 3\}$ (corresponding to the bit labels $\{00, 01, 10, 11\}$), and $u_{ijk,c}^m$ denotes the decoder output for class c before softmax. The reconstruction loss is defined as:

$$\mathcal{L}_{\text{recon}}^m = \frac{1}{N_{\text{edge}}^m} \sum_{t=1}^{N_{\text{edge}}^m} \left(-\log \frac{\exp\big(u_{t,r_t^m}^m\big)}{\sum_{c \in C} \exp\big(u_{t,c}^m\big)}\right), \tag{8}$$

where N_{edge}^m denotes the number of supervised reconstruction triples in the m-th modality. The matrix $\mathbf{U}^m = [u_{t,c}^m] \in \mathbb{R}^{N_{\text{edge}}^m \times 4}$ collects the reconstruction decoder logits, with $u_{t,c}^m$ denoting the logit for class c of the t-th triple, while $\mathbf{r}^m = [r_t^m]$ contains the corresponding target labels.

4.2 JOINTLY OPTIMIZED CROSS-MODAL FUSION

Adaptive Gated Cross-Modal Fusion. Protein interactions depend on complementary information across modalities, motivating adaptive fusion to account for their varying contributions across protein pairs. The sequence and four knowledge graph modalities are indexed by $m \in \{0, 1, 2, 3, 4\}$, respectively. For each protein pair (p_i, p_j), we compute the absolute difference to capture representational discrepancies and the Hadamard product to capture correspondence, yielding a symmetric pair representation for each modality:

$$d_{ij}^m = \left[\left| h_{p_i}^m - h_{p_j}^m \right| \| \left(h_{p_i}^m \odot h_{p_j}^m \right) \right]. \tag{9}$$

An adaptive gate module is then used to estimate the contribution of each modality to the pair:

$$\lambda_{ij}^m = \frac{\exp(w_g^\top \phi(\mathbf{W}_g d_{ij}^m + \mathbf{b}_g) + \rho_m)}{\sum_{t=0}^{4} \exp(w_g^\top \phi(\mathbf{W}_g d_{ij}^t + \mathbf{b}_g) + \rho_t)}, \tag{10}$$

where ρ_m is a learnable parameter. The weighted representations are concatenated into z_{ij} and passed through two hidden layers and a sigmoid output layer to predict the interaction probability:

$$z_{ij} = [\lambda_{ij}^0 d_{ij}^0 \| \lambda_{ij}^1 d_{ij}^1 \| \lambda_{ij}^2 d_{ij}^2 \| \lambda_{ij}^3 d_{ij}^3 \| \lambda_{ij}^4 d_{ij}^4], \quad z_{ij} \in \mathbb{R}^{10D}. \tag{11}$$

$$\hat{y}_{ij} = \sigma\Big(\mathbf{W}_o^{(3)} \Phi\Big(\mathbf{W}_o^{(2)} \Phi(\mathbf{W}_o^{(1)} z_{ij} + \mathbf{b}_o^{(1)}) + \mathbf{b}_o^{(2)} \Big) + \mathbf{b}_o^{(3)} \Big). \tag{12}$$

Finally, the supervised PPI prediction loss over known protein pairs is defined as:

$$\mathcal{L}_{\text{PPI}} = -\frac{1}{|\mathcal{D}_{\text{known}}|} \sum_{(i,j) \in \mathcal{D}_{\text{known}}} \left[y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}) \right], \tag{13}$$

where $\mathcal{D}_{\text{known}}$ denotes the training set of known-protein pairs, and y_{ij} is the PPI label for (p_i, p_j).

Joint Supervision Optimization. ECMLF is trained through an end-to-end jointly optimized multi-task objective that simultaneously combines PPI prediction with bridge reconstruction across four graphs, where η controls the relative contribution of reconstruction to the overall loss:

$$\mathcal{L} = \mathcal{L}_{\text{PPI}} + \eta \sum_{m=1}^{4} \mathcal{L}_{\text{recon}}^m. \tag{14}$$

4.3 DUAL-PATHWAY EXPLAINABILITY MECHANISM

Sequence-Pathway Explanation. The sequence explanation assesses the biological relevance of highly attended regions by examining their correspondence with annotated functional sites implicated in PPI and provides sequence-level evidence for predicted interactions.

(1) *Candidate localization*: For protein p_i, the sequence segments $\{s_{i_j}\}_{j=1}^n$ are ranked by their regional attention weights $\{\alpha_{i_j}\}_{j=1}^n$, and the top-K segments are retained as candidate regions potentially relevant to PPI prediction.

(2) *Functional consistency validation*: The biological relevance of the localized sequence evidence is assessed by comparing each selected region with known functional annotations. Each candidate region s_{i_j} is compared with every annotated functional region $A_{i_k} \in \mathcal{A}_i$ of protein p_i, and its hit indicator is defined as:

$$\text{Hit}(s_{i_j}) = I\left[\max_{A_{i_k} \in \mathcal{A}_i} \frac{|s_{i_j} \cap A_{i_k}|}{|A_{i_k}|} \geq \tau \right], \tag{15}$$

where τ is the consistency threshold.

Graph-Pathway Explanation. By identifying shared entities across knowledge graphs, the graph explanation reveals how the model captures latent similarities between new proteins and associated known proteins, providing transferable relational evidence for cold-start predictions. This analysis is conducted after PPI prediction, and the annotated PPI network is used to evaluate explanations.

(1) *Associated protein retrieval*: For each new protein $p_{\text{new},i}$, a known protein $p_{\text{known},j}$ is designated as associated if the two proteins share at least one interaction partner in the annotated PPI network. All known proteins satisfying this condition form the associated-protein set $U(p_{\text{new},i}) \subseteq \mathcal{P}_{\text{known}}$.

(2) *Bridge entity retrieval*: For each new protein $p_{\text{new},i}$, a neighboring entity e_k in modality m is retained as salient if $\alpha^m_{p_{\text{new},i},e_k} \geq \bar{\alpha}^m_{p_{\text{new},i}}$, where $\bar{\alpha}^m_{p_{\text{new},i}}$ denotes the mean attention score over all neighboring entities. A salient entity also linked to at least one associated known protein $p_{\text{known},j} \in U(p_{\text{new},i})$ is identified as a bridge entity. The number of bridge entities across graph modalities is then quantified to characterize relational evidence for latent protein similarity.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate ECMLF on two datasets, each containing a PPI network, multimodal biomedical associations, and protein sequences. Dataset1 is built from STRING and four biomedical sources (Szklarczyk et al., 2017; Piñero et al., 2017; Wishart et al., 2018; Chou et al., 2018; Cheng et al., 2019), while Dataset2 uses an updated STRING network with additional resources (Szklarczyk et al., 2025; Knox et al., 2024; Zhao et al., 2023; Cui et al., 2025; Davis et al., 2025).

Evaluation Metrics. We evaluate PPI prediction performance using seven metrics: ACC, REC, PRE, F1, AUC, AUPR, and MCC. These metrics provide a comprehensive assessment of predictive performance from complementary perspectives. Detailed definitions are provided in Appendix A.3.

Baselines. We compare ECMLF with seven PPI prediction baselines: HNSPPI (Xie et al., 2023), EResCNN (Gao et al., 2023), BaPPI (Tang et al., 2024), KGF-GNN (Yang et al., 2024), HEENN (Yang et al., 2025), ESM2_AMP (Sun et al., 2025), and a cold-start adaptation of GATSBI (Nayar & Altman, 2026). Their descriptions are summarized in Appendix Table 6.

5.2 COMPARATIVE PERFORMANCE WITH BASELINES

Tables 1 and 2 (all metrics have been converted into percentages) compare ECMLF with the baselines under five-fold cross-validation in SCS and FCS, respectively. Under SCS, ECMLF ranks first on 13/14 metrics, achieves 97 wins and 1 loss, and yields the superior Friedman rank of 1.07. Its relative gains in AUC and AUPR over the strongest baseline reach 3.63% and 4.10% on Dataset1, and 22.63% and 22.95% on Dataset2. Under the more challenging FCS setting, ECMLF ranks first on 11/14 metrics, achieves 95 wins and 3 losses, and attains a Friedman rank of 1.21. The corresponding AUC and AUPR gains reach 24.53% and 25.44% on Dataset1, and 28.43% and 26.17% on Dataset2. Overall, the reported Wilcoxon signed-rank p-values are below 0.05, supporting ECMLF's advantage under the adopted testing protocol.

Table 1: Performance comparison of ECMLF and baseline models under SCS.

Dataset	Metric	HNSPPI	EResCNN	BaPPI	KGF	HEENN	ESM2_AMP	GATSBI	ECMLF
Dataset1	ACC	63.38±0.41	65.38±1.76	66.40±1.41	56.71±4.38	62.81±1.46	77.99±0.81	55.76±3.01	**81.31±1.28**
	REC	63.72±0.91	44.13±5.74	73.06±1.62	22.54±2.90	51.83±25.53	77.02±2.57	16.57±5.87	**78.41±3.10**
	PRE	63.56±0.78	77.14±2.17	64.78±2.69	79.52±0.49	70.50±7.64	78.67±1.10	76.38±3.42	**83.61±1.40**
	F1	63.39±0.40	55.92±4.32	68.62±1.06	35.08±3.57	56.18±8.95	77.80±0.96	26.88±8.51	**80.89±1.40**
	AUC	68.23±0.79	75.39±1.24	71.96±0.71	75.16±2.81	73.98±1.71	85.68±0.71	70.24±1.91	**88.79±0.98**
	AUPR	68.23±0.63	74.77±1.14	69.08±1.69	72.39±2.06	71.12±1.00	84.90±0.75	68.08±1.81	**88.38±0.54**
	MCC	27.24±0.81	34.18±2.66	33.12±2.39	24.51±1.96	30.43±3.78	56.04±1.57	18.44±5.56	**62.83±2.35**
Dataset2	ACC	57.20±0.50	55.38±0.47	62.95±3.18	62.80±0.79	66.33±3.67	58.72±0.86	54.11±1.09	**84.61±0.59**
	REC	57.18±0.77	**84.91±0.64**	62.95±2.88	42.88±3.42	48.72±12.75	67.97±10.56	10.85±2.42	77.44±1.08
	PRE	57.23±0.65	53.41±0.52	46.10±3.11	71.39±1.19	76.36±6.03	57.57±2.38	80.55±3.12	**90.43±0.60**
	F1	57.20±0.50	65.57±0.33	53.18±2.58	53.49±2.51	58.34±7.99	61.95±2.93	19.06±3.81	**83.43±0.57**
	AUC	61.39±0.79	55.34±0.50	68.67±2.00	69.88±0.87	75.24±3.50	62.71±0.76	73.35±2.24	**92.27±0.50**
	AUPR	63.42±0.51	53.19±0.56	51.78±0.90	69.01±0.86	75.41±3.68	64.26±0.50	70.64±2.04	**92.72±0.28**
	MCC	14.40±1.01	13.30±1.06	24.56±5.34	27.97±1.11	35.34±6.42	18.27±0.47	16.39±2.93	**69.95±1.06**
* Statistic	Win/Loss	0/14	1/13	0/14	0/14	0/14	0/14	0/14	97/1
	p-value	0.00195	0.00097	0.00097	0.00781	0.00097	0.00390	0.00098	-
	F-rank	6.00	4.93	5.07	5.00	4.14	3.36	6.43	**1.07**

*Underlined: second-best result. **Win/Loss**: comparison of ECMLF's performance against each baseline across metrics. **p-value**: Wilcoxon signed-rank test results. **F-rank**: average Friedman rank across metrics.

Table 2: Performance comparison of ECMLF and baseline models under FCS.

Dataset	Metric	HNSPPI	EResCNN	BaPPI	KGF	HEENN	ESM2_AMP	GATSBI	ECMLF
Dataset1	ACC	50.22±1.31	52.75±1.89	57.48±2.28	53.33±3.63	53.41±3.35	59.23±1.60	56.29±2.66	**74.40±1.85**
	REC	50.25±1.87	8.18±1.61	59.91±6.33	16.48±15.22	18.36±6.86	**79.36±9.69**	17.13±4.32	68.47±7.83
	PRE	49.73±2.63	**79.16±5.12**	57.24±5.00	53.97±5.44	60.46±5.24	56.24±2.98	76.34±6.05	77.43±3.79
	F1	49.92±1.05	14.78±2.69	58.22±3.12	22.86±17.60	27.46±7.96	65.59±3.66	27.69±5.89	**72.34±3.75**
	AUC	50.53±1.91	63.56±2.54	59.23±2.73	52.59±5.20	56.07±2.85	64.14±2.55	65.97±2.42	**82.15±2.75**
	AUPR	50.07±3.75	63.82±2.87	52.24±10.26	49.28±1.10	54.51±1.45	59.47±13.40	65.37±2.98	**82.00±2.61**
	MCC	0.53±2.60	13.54±1.89	14.56±4.74	2.86±3.92	8.36±4.14	20.44±2.82	18.75±2.71	**48.85±3.59**
Dataset2	ACC	51.47±0.54	54.84±0.90	59.29±2.15	53.39±1.03	55.79±0.71	58.29±2.00	56.70±1.50	**79.56±1.84**
	REC	51.48±0.69	**83.79±2.34**	51.67±2.46	49.47±12.44	15.11±4.90	66.76±9.90	16.46±1.93	67.53±3.92
	PRE	51.28±1.78	52.96±1.39	40.11±0.86	52.86±2.44	86.21±4.94	55.53±3.18	82.71±3.20	**88.78±1.04**
	F1	51.37±0.95	64.87±0.83	45.15±1.44	50.74±7.84	25.40±6.84	60.13±2.29	27.43±2.80	**76.65±2.45**
	AUC	52.66±0.67	54.69±0.83	58.14±1.72	53.28±1.95	69.29±1.34	60.82±1.82	68.67±2.95	**88.99±1.52**
	AUPR	52.42±1.14	52.89±1.31	36.74±4.08	52.19±3.42	70.66±1.24	54.76±10.08	69.16±2.83	**89.15±1.34**
	MCC	2.97±1.09	12.19±2.44	13.84±3.47	6.21±2.29	21.74±1.92	15.19±2.64	21.81±2.63	**60.90±2.82**
* Statistic	Win/Loss	0/14	2/12	0/14	0/14	0/14	1/13	0/14	95/3
	p-value	0.00097	0.00195	0.00390	0.03125	0.00390	0.00390	0.00098	-
	F-rank	6.64	4.79	4.79	6.71	4.79	3.21	3.86	**1.21**

5.3 Ablation Study

Modality Ablation. To assess the contributions of different modalities, comprehensive ablation studies were conducted across two settings. Let M_s, M_1, M_2, M_3, and M_4 denote the sequence, drug, disease, miRNA, and lncRNA modalities, respectively. As shown in Table 3, combining graph modalities generally improves performance, with the full model achieving the best overall results under both SCS and FCS. These findings support the complementarity of intrinsic sequence attributes and external biomedical associations. Additional results are provided in Appendix A.6.

Table 3: Modality ablation results on Dataset1 under SCS and FCS.

Modality	SCS								FCS							
	ACC	REC	PRE	F1	AUC	AUPR	MCC	F-rank	ACC	REC	PRE	F1	AUC	AUPR	MCC	F-rank
M1	61.32	39.67	72.78	51.35	67.66	70.12	26.63	**12.57**	61.98	70.06	69.63	49.19	70.06	69.63	25.16	**7.43**
M2	62.50	50.18	68.52	57.93	68.12	70.69	26.54	**12.71**	56.99	37.04	58.41	45.33	60.24	59.69	13.62	**13.71**
M3	57.12	29.70	69.53	41.62	60.63	65.61	19.19	**15.14**	53.99	25.19	54.84	34.52	53.52	53.91	7.14	**15.71**
M4	55.73	89.48	54.23	67.54	54.63	55.37	13.15	**13.43**	50.57	86.67	49.24	62.80	52.08	71.22	5.17	**11.43**
M1+M2	66.95	68.18	67.78	67.98	73.20	74.08	33.82	**7.86**	66.76	62.22	66.56	64.32	73.21	71.18	33.34	**5.14**
M1+M3	66.75	64.69	68.82	66.69	73.18	75.39	33.62	**8.57**	63.62	50.22	66.08	57.07	69.36	69.62	27.27	**8.29**
M1+M4	62.47	79.83	60.21	68.64	68.35	69.67	25.64	**10.86**	50.78	88.00	49.38	63.26	58.53	54.07	6.06	**12.00**
M2+M3	63.51	49.22	70.98	58.13	69.58	72.25	29.09	**10.57**	62.20	50.07	63.65	56.05	65.36	64.91	24.23	**10.00**
M2+M4	63.27	60.34	65.55	62.84	68.56	69.81	26.74	**12.29**	59.49	46.67	60.23	52.59	63.15	60.95	18.65	**12.00**
M3+M4	59.16	71.71	58.41	64.38	63.55	66.35	18.21	**13.43**	54.99	24.44	57.37	34.34	58.60	57.37	9.67	**15.00**
M1+M2+M3	71.09	71.09	70.19	70.64	78.25	79.86	42.22	**4.57**	67.33	57.19	69.55	62.76	74.13	74.69	34.68	**5.14**
M1+M2+M4	69.01	69.99	69.85	69.92	75.75	77.12	37.96	**5.71**	64.62	58.67	64.60	61.49	71.01	69.34	29.02	**7.43**
M1+M3+M4	66.82	60.76	70.65	65.33	73.02	74.67	34.21	**8.14**	61.55	42.37	65.60	51.49	62.64	64.51	23.46	**11.14**
M2+M3+M4	65.54	68.99	65.75	67.33	71.76	72.75	30.96	**9.43**	61.34	60.44	59.74	60.09	66.88	63.73	22.61	**10.14**
M1+M2+M3+M4	71.72	67.45	75.07	71.06	78.73	79.24	43.81	**4.14**	67.90	55.56	71.43	62.50	75.50	75.44	36.05	**4.43**
M_s	79.60	73.00	84.30	78.10	87.80	87.10	59.80	**2.14**	70.00	66.10	72.00	68.70	75.40	75.50	40.50	**2.57**
Full	81.31	78.41	83.61	80.89	88.79	88.38	62.83	**1.43**	74.40	68.47	77.43	72.34	82.15	82.00	48.85	**1.43**

5.4 Explainability Study

Sequence-Pathway Explainability. In this section, we compare the top-1 and top-2 attended regions identified by ECMLF and ESM2_AMP with functional annotations, counting a hit at overlap thresholds of 25%, 50%, and 75%. As shown in Table 4, ECMLF achieves higher hit rates across both datasets, including at stricter thresholds. These results suggest that ECMLF more effectively highlights sequence regions associated with annotated biological functions.

Graph-Pathway Explainability. For graph-pathway explanation, we evaluate whether the model can identify key bridge entities that support latent associations between known and new proteins. As shown in Figure 2, ECMLF captures more key bridge entities than HEENN under both settings, demonstrating its stronger ability to extract biologically informative relational evidence for downstream cold-start PPI prediction. Results on Dataset2 are provided in Appendix A.7.

Table 4: Functional annotation hit rates for top-1 and top-2 attended sequence regions.

Top-1 attended regions						*Top-2 attended regions*					
Setting	τ	*Dataset1*		*Dataset2*		Setting	τ	*Dataset1*		*Dataset2*	
		ESM2_AMP	ECMLF	ESM2_AMP	ECMLF			ESM2_AMP	ECMLF	ESM2_AMP	ECMLF
SCS	25%	40.45%	55.29%	43.39%	65.90%	SCS	25%	41.18%	42.09%	44.18%	61.50%
	50%	28.89%	54.99%	29.79%	58.79%		50%	28.10%	41.72%	29.58%	51.57%
	75%	21.35%	54.50%	21.95%	47.30%		75%	22.08%	41.18%	21.93%	42.42%
FCS	25%	46.04%	48.78%	30.37%	65.65%	FCS	25%	41.46%	42.98%	29.32%	60.74%
	50%	30.49%	48.78%	20.21%	59.00%		50%	27.74%	42.68%	18.57%	53.15%
	75%	23.17%	48.17%	13.90%	48.48%		75%	22.87%	42.37%	13.67%	44.62%

Case Study. We dock CDK4 and PCNA using HDOCK (Yan et al., 2020) to assess the structural relevance of ECMLF's sequence explanations. As shown in Figure 3, residues from the top-1 and top-2 regions occur at the predicted interface. SER-293–GLU-124 and ARG-288–GLY-127 show contact distances of approximately 2.6 Å, suggesting potential hydrogen bonds that may stabilize the interface. These contacts support the possible involvement of attention-highlighted regions in CDK4-PCNA binding, offering structural context for interpreting the prediction.

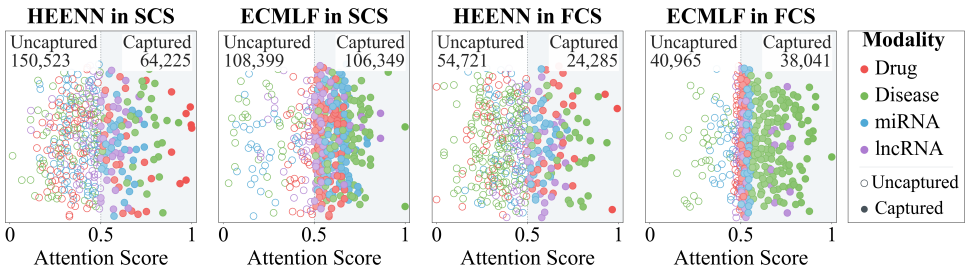


Figure 2: Graph-pathway explainability comparison between HEENN and ECMLF on Dataset1.

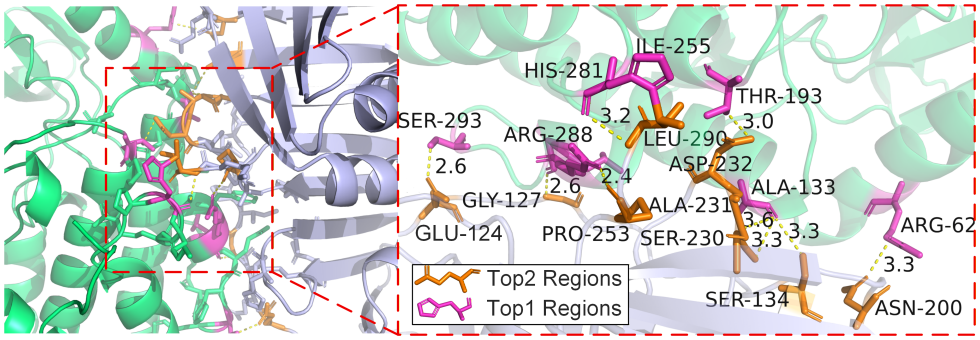


Figure 3: Case study of the docked CDK4 (green)-PCNA (purple) complex.

5.5 Hyperparameter Sensitivity Analysis

Due to space constraints, the hyperparameter sensitivity analysis is deferred to Appendix A.8.

6 Conclusion

We presented ECMLF, an explainable cross-modal learning framework for cold-start PPI prediction. Integrating intrinsic amino acid sequences with four biomedical knowledge graphs, ECMLF captures complementary cross-modal evidence to model similarity between known and new proteins. Adaptive gated fusion jointly optimizes similarity modeling and interaction prediction end to end. Dual-pathway explanations connect predictions to interaction-related residue regions and biomedical bridge entities, offering biological insights into potential interaction mechanisms. Experiments on two datasets under SCS and FCS demonstrate effective prediction and biologically relevant explanations supported by sequence- and graph-pathway analyses.

AI USE STATEMENT

Generative AI tools were used for: (1) language editing and grammatical review; (2) code review and debugging; (3) enhancing the visual presentation of figures and tables. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final text, claims, figures, and other artifacts in this work.

ETHICS STATEMENT

This study analyzes publicly available protein and biomedical data without involving human participants or animal experiments. No human-subject or animal-research approval was required. Predicted interactions require experimental validation before informing biomedical decisions.

REPRODUCIBILITY STATEMENT

To support the reproducibility of our work, we have submitted the source code, datasets, implementation details, and necessary instructions as the Supplementary Materials. We believe that these resources provide sufficient information for reproducing the main experimental results in this paper.

REFERENCES

Domenico Amato, Salvatore Calderaro, Giosuè Lo Bosco, Riccardo Rizzo, and Filippo Vella. GECo: a community-based graph neural network explainer. *Computing*, 108(4):55, 2026.

David Arteaga, Nikita Chervov, and Maria Poptsova. Multimodal graph, surface, and language-based model for protein protein interaction prediction. *Scientific Reports*, 16(1):4772, 2026.

Liang Cheng, Pingping Wang, Rui Tian, Song Wang, Qinghua Guo, Meng Luo, Wenyang Zhou, Guiyou Liu, Huijie Jiang, and Qinghua Jiang. LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Research*, 47(D1):D140–D144, 2019.

Chih-Hung Chou, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, Ting-Hsuan Sun, Siang-Jyun Tu, Wei-Hsiang Lee, Men-Yee Chiew, Chun-San Tai, Ting-Yen Wei, Tzi-Ren Tsai, Hsin-Tzu Huang, Chung-Yu Wang, Hsin-Yi Wu, Shu-Yi Ho, Pin-Rong Chen, Cheng-Hsun Chuang, Pei-Jung Hsieh, Yi-Shin Wu, Wen-Liang Chen, Meng-Ju Li, Yu-Chun Wu, Xin-Yi Huang, Fung Ling Ng, Waradee Buddhakosai, Pei-Chun Huang, Kuan-Chun Lan, Chia-Yen Huang, Shun-Long Weng, Yeong-Nan Cheng, Chao Liang, Wen-Lian Hsu, and Hsien-Da Huang. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1):D296–D302, 2018.

Pınar Cihan, Zeynep Banu Ozger, and Zeynep Cakabay. Computational analysis of virus-host protein-protein interactions using gene ontology and natural language processing. *Applied Intelligence*, 55(6):393, 2025.

Gabriele Corso, Vignesh Ram Somnath, Noah Getz, Regina Barzilay, Tommi Jaakkola, and Andreas Krause. Composing unbalanced flows for flexible docking and relaxation. In *International Conference on Learning Representations (ICLR)*, 2025.

Shidong Cui, Sicong Yu, Hsi-Yuan Huang, Yang-Chi-Dung Lin, Yixian Huang, Bojian Zhang, Jihan Xiao, Huali Zuo, Jiayi Wang, Zhuoran Li, Guanghao Li, Jiajun Ma, Baiming Chen, Haoxuan Zhang, Jiehui Fu, Liang Wang, and Hsien-Da Huang. miRTarBase 2025: updates to the collection of experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 53(D1): D147–D156, 2025.

Thanh Hai Dang and Tien Anh Vu. xcapt5: protein–protein interaction prediction using deep and wide multi-kernel pooling convolutional neural networks with protein language model. *BMC bioinformatics*, 25(1):106, 2024.

Allan Peter Davis, Thomas C Wiegers, Daniela Sciaky, Fern Barkalow, Melissa Strong, Brent Wyatt, Jolene Wiegers, Roy McMorran, Sakib Abrar, and Carolyn J Mattingly. Comparative toxicogenomics database's 20th anniversary: update 2025. *Nucleic Acids Research*, 53(D1):D1328–D1334, 2025.

Hongli Gao, Cheng Chen, Shuangyi Li, Congjing Wang, Weifeng Zhou, and Bin Yu. Prediction of protein-protein interactions based on ensemble residual convolutional neural network. *Computers in Biology and Medicine*, 152:106471, 2023.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Multimodal lego: Model merging and fine-tuning across topologies and modalities in biomedicine. In *International Conference on Learning Representations (ICLR)*, 2025.

Craig Knox, Mike Wilson, Christen M. Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun (Lucy) Chin, Seth A. Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi (Rachel) Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S. Wishart. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024.

Young Su Ko, Jonathan Parkinson, Cong Liu, and Wei Wang. TUnA: an uncertainty-aware transformer model for sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(5):bbae359, 2024.

Luke Lambourne, Anupama Yadav, Yang Wang, Alice Desbuleux, Dae-Kyum Kim, Florent Laval, Kerstin Spirohn-Fitzgerald, Tiziana Cafarelli, Carles Pons, István A. Kovács, Noor Jailkhani, Sadie Schlabach, David De Ridder, Katja Luck, Vladimir V. Botchkarev, Jr., Olivia Debnath, Wenting Bian, Yun Shen, Zhipeng Yang, Miles W. Mee, Mohamed Helmy, Yves Jacob, Irma Lemmens, Thomas Rolland, Gregory G. McClain, Atina G. Coté, Marinella Gebbia, Nishka Kishore, Jennifer J. Knapp, Joseph C. Mellor, Gonen Memisoglu, Jüri Reimand, Jan Tavernier, Michael E. Cusick, Quan Zhong, Patrick Aloy, Tong Hao, Benoit Charloteaux, Frederick P. Roth, Javier De Las Rivas, Pascal Falter-Braun, David E. Hill, Michael A. Calderwood, Jean-Claude Twizere, and Marc Vidal. Experimental assessment of AI-based interactome mapping. *Nature Communications*, 17(1):4894, 2026.

Shumei Li, Yuyang Wang, Heming Zhang, Xin Xing, Ying Shao, and Zhao Qi. Dss-ppi: a self-supervised graph learning framework for protein-protein interaction prediction via multimodal sequence semantics. *BMC Genomics*, 27(1):421, 2026.

Yane Li, Chengfeng Wang, Haibo Gu, Zhentao Long, Ming Fan, and Lihua Li. Mpidnn-gppi: multi-protein language model with an improved deep neural network for generalized protein-protein interaction prediction. *BMC Genomics*, 26(1):1059, 2025.

Dan Liu, Francesca Young, Kieran D. Lamb, Adalberto Claudio Quiros, Alexandrina Pancheva, Crispin J. Miller, Craig Macdonald, David L. Robertson, and Ke Yuan. PLM-interact: extending protein language models to predict protein-protein interactions. *Nature Communications*, 16(1): 9012, 2025.

Jun Liu, Hungyu Chen, and Yang Zhang. A paired sequence language model for protein-protein interaction modeling. *Nature Communications*, 17(1):3733, 2026.

Vincent Mallet, Yangyang Miao, Souhaib Attaiki, Bruno Correia, and Maks Ovsjanikov. AtomSurf: Surface representation for learning on protein structures. In *International Conference on Learning Representations (ICLR)*, 2025.

Gowri Nayar and Russ B. Altman. GATSBI: improving context-aware protein embeddings through biologically motivated data splits. *Bioinformatics*, 42(Supplement 1):btag233, 2026.

Jeffrey Ouyang-Zhang, Chengyue Gong, Yue Zhao, Philipp Krähenbühl, Adam Klivans, and Daniel Diaz. Distilling structural representations into protein sequence models. In *International Conference on Learning Representations (ICLR)*, 2025.

Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I. Furlong. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839, 2017.

Aleksey Porollo, Om Jadhav, Aaron Alvarez, and Jichao Chen. SPPIDER-seq: sequence-based partner-aware predictor of protein-protein interaction sites. *Bioinformatics*, 42(7):btag388, 2026.

Anastasia Rapti, Bert J. C. Janssen, and Alexandre M. J. J. Bonvin. Integrating experimental and computational approaches for protein–protein interaction discovery. *Nature Reviews Bioengineering*, 4(9):921–936, 2026.

Alisson Silva, Carlos Marquez, Iury Godoy, Lucas Silva, Matheus Prado, Murilo Beppler, Natanael Avila, Bruno Travençolo, and Anderson R. Santos. Improving protein interaction prediction in genppi: a novel interaction sampling approach preserving network topology. *BMC Bioinformatics*, 26(1):296, 2025.

Elana Simon and James Zou. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, 2025.

Hannes Stark, Bowen Jing, Tomas Geffner, Jason Yim, Tommi Jaakkola, Arash Vahdat, and Karsten Kreis. Protcomposer: Compositional protein structure generation with 3d ellipsoids. In *International Conference on Learning Representations (ICLR)*, 2025.

Yawen Sun, Rui Wang, Zeyu Luo, Lejia Tan, Junhao Liu, Ruimeng Li, Dongqing Wei, and Yu-Juan Zhang. Esm2_amp: an interpretable framework for protein–protein interactions prediction and biological mechanism discovery. *Briefings in Bioinformatics*, 26(4):bbaf434, 2025.

Damian Szklarczyk, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T. Doncheva, Alexander Roth, Peer Bork, Lars Juhl Jensen, and Christian von Mering. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, 2017.

Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E. Peluso, Qingyao Huang, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2025.

Joseph Szymborski and Amin Emad. A flaw in using pretrained protein language models in protein-protein interaction inference models. *Nature Machine Intelligence*, 8(2):197–208, 2026.

Tao Tang, Tianyang Li, Weizhuo Li, Xiaofeng Cao, Yuansheng Liu, and Xiangxiang Zeng. Antisymmetric framework for balanced learning of protein–protein interactions. *Bioinformatics*, 40 (10):btae603, 2024.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025.

Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 17(1):1199, 2026.

Feng Wang, Jinming Chu, Liyan Shen, and Shan Chang. Mesm: integrating multi-source data for high-accuracy protein-protein interactions prediction through multimodal language models. *BMC Biology*, 23(1):253, 2025a.

Haoping Wang, Xiangjie Meng, and Yang Zhang. Biomolecular interaction prediction: The era of ai. *Advanced Science*, 12(38):e09501, 2025b.

Sihan Wang, Shuai Zhang, Zhuo Chen, Xuqiang Li, Jianmin Wang, Wenjie Du, and Yang Wang. TRGH-PPI: Effective and generalized prediction of protein-protein interactions through transformer and graph. *IEEE Transactions on Computational Biology and Bioinformatics*, 22(5): 2230–2240, 2025c.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *International Conference on Learning Representations (ICLR)*, 2025d.

Xun Wang, Tongyu Han, Runqiu Feng, Zhijun Xia, Hanyu Wang, Wenqian Yu, Huanhuan Dai, Haonan Song, and Tao Song. GTE-PPIS: a protein-protein interaction site predictor based on graph transformer and equivariant graph neural network. *Briefings in Bioinformatics*, 26(3):bbaf290, 2025e.

David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.

Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, Xiao Luo, Yu Zhang, James Zou, Yizhou Sun, and Wei Wang. Protein large language models: A comprehensive survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23080–23103, Suzhou, China, November 2025. Association for Computational Linguistics.

Shijie Xie, Xiaojun Xie, Xin Zhao, Fei Liu, Yiming Wang, Jihui Ping, and Zhiwei Ji. Hnsppi: a hybrid computational model combing network and sequence information for predicting protein–protein interaction. *Briefings in Bioinformatics*, 24(5):bbad261, 2023.

Jing Xiong, Xiaohong Wang, Fangkun Dou, Haowei Cao, Tenglong Hong, Zhiwei Li, and Xiaoli Jing. Construction and application of protein-protein interaction knowledge graph. In Huimin Lu (ed.), *Artificial Intelligence and Robotics - 9th International Symposium, ISAIR 2024, Guilin, China, September 27-30, 2024, Revised Selected Papers, Part I*, volume 2402 of *Communications in Computer and Information Science*, pp. 44–53. Springer, 2025.

Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang. The HDOCK server for integrated protein–protein docking. *Nature Protocols*, 15(5):1829–1852, 2020.

Jie Yang, Yapeng Li, Guoyin Wang, Zhong Chen, and Di Wu. An end-to-end knowledge graph fused graph neural network for accurate protein-protein interactions prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 21(6):2518–2530, 2024.

Jie Yang, Xijie Lan, Guoyin Wang, Zhong Chen, Yuwen Chen, and Di Wu. A hybrid ensemble end-to-end neural network for accurate protein-protein interactions prediction. *IEEE Transactions on Computational Biology and Bioinformatics*, 22(6):2540–2553, 2025.

Yuan Zhang, Xin Sui, Feng Pan, Kaixian Yu, Keqiao Li, Shubo Tian, Arslan Erdengasileng, Qing Han, Wanjing Wang, Jianan Wang, Jian Wang, Donghu Sun, Henry Chung, Jun Zhou, Eric Zhou, Ben Lee, Peili Zhang, Xing Qiu, Tingting Zhao, and Jinfeng Zhang. A comprehensive large-scale biomedical knowledge graph for AI-powered data-driven biomedical research. *Nature Machine Intelligence*, 7(4):602–614, 2025.

Hongying Zhao, Xiangzhe Yin, Haotian Xu, Kailai Liu, Wangyang Liu, Lixia Wang, Caiyu Zhang, Lin Bo, Xicheng Lan, Shihua Lin, Ke Feng, Shangwei Ning, Yunpeng Zhang, and Li Wang. LncTarD 2.0: an updated comprehensive database for experimentally-supported functional lncRNA-target regulations in human diseases. *Nucleic Acids Research*, 51(D1):D199–D207, 2023.

A APPENDIX

This appendix supplements the main manuscript with additional details on the experimental design, algorithmic methodology, and data analysis. It is intended to facilitate the reproduction and assessment of the study, covering topics such as notation, model implementation, and sensitivity analysis.

A.1 ALGORITHM DESIGN AND TIME COMPLEXITY ANALYSIS

Algorithm Design. Algorithm 1 outlines the training procedure of ECMLF, jointly optimizing cross-modal representation learning and adaptive gated fusion. Following the formulation in the main text, the model encodes sequence regions and multimodal biomedical graphs, combining their representations for PPI prediction. The prediction head uses two ReLU hidden layers followed by a sigmoid output layer. Here, ℓ_{BCE} denotes the mean binary cross-entropy loss over the specified set. Softmax normalization operates across sequence regions or modalities, depending on the operation.

Algorithm 1 ECMLF Training Algorithm

Input: PPI labels of known protein pairs $\mathbf{Y}_{\mathcal{P}_{\mathrm{known}}}$; amino acid sequences $\mathbf{S}$; four biomedical graphs $\{\mathbf{G}^m\}_{m=1}^4$; learnable parameters Θ.
Output: $\Gamma(p_i, p_j \mid \Theta, \mathbf{Y}_{\mathcal{P}_{\mathrm{known}}}, \mathbf{S}, \mathbf{G})$.

Segment sequences $\mathbf{S}$ into multiple regions and initialize sequence regional embeddings $\mathbf{X}$.
for $m \in \{1,2,3,4\}$ **do**
  Randomly select $R\%$ entities and mask a subset of protein-entity edges incident to each selected entity to obtain the incomplete graph $\widetilde{\mathbf{G}}^m$.
  Initialize the node representations $\mathbf{H}^{m(0)}$ from the relation matrix $\widetilde{\mathbf{R}}$ of the incomplete graph $\widetilde{\mathbf{G}^m}$.
while *ECMLF* has not converged **do**
  for all proteins p_i **do**
    $\mathbf{H}_i \leftarrow \mathrm{Norm}\Big(\mathcal{T}^{(6)}_{\mathrm{encoder}}(\mathrm{Norm}(\mathbf{X}_i)\mathbf{W}_s + \mathbf{b}_s + \mathbf{P}_i)\Big)$
    $\alpha_{i_j} \leftarrow \mathrm{softmax}_j\big(\mathbf{w}_{\mathrm{seq}}^\top \phi(\mathbf{W}_{\mathrm{seq}}\mathbf{h}_{i_j} + \mathbf{b}_{\mathrm{seq}})\big)$
    $\mathbf{h}_i^0 \leftarrow \sum_{j=1}^n \alpha_{i_j}\mathbf{h}_{i_j}$
  for $m \in \{1,2,3,4\}$ **do**
    $\mathbf{H}^{m(2)} \leftarrow \mathbf{GAT}_m(\mathbf{H}^{m(0)}, \widetilde{\mathbf{G}}^m)$
    for all masked entities e_k **do**
      Sample two distinct neighbors p_i^+, p_j^+ and two distinct non-neighbors p_i^-, p_j^- to form reconstruction triples (p_i, e_k, p_j).
    for all triples $(p_i, e_k, p_j) \in \mathcal{M}^m_{\mathrm{edge}}$ **do**
      $\mathbf{u}^m_{ijk} \leftarrow \mathbf{W}_r^{m(2)}\Phi\Big(\mathbf{W}_r^{m(1)}[\mathbf{h}^m_{p_i} \| \mathbf{h}^m_{e_k} \| \mathbf{h}^m_{p_j}] + \mathbf{b}_r^{m(1)}\Big) + \mathbf{b}_r^{m(2)}$
    $\mathcal{L}^m_{\mathrm{recon}} \leftarrow -\dfrac{1}{N^m_{\mathrm{edge}}} \displaystyle\sum_{t=1}^{N^m_{\mathrm{edge}}} \log \frac{\exp(u^m_{t,r_t^m})}{\displaystyle\sum_{c\in\mathcal{C}} \exp(u^m_{t,c})}.$
  for all known protein pairs (p_i, p_j) in $\mathbf{Y}_{\mathcal{P}_{\mathrm{known}}}$ **do**
    for $m \in \{0,1,2,3,4\}$ **do**
      $d^m_{ij} \leftarrow [|\mathbf{h}^m_{p_i} - \mathbf{h}^m_{p_j}| \| (\mathbf{h}^m_{p_i} \odot \mathbf{h}^m_{p_j})]$
    $\lambda^m_{ij} \leftarrow \mathrm{softmax}_m\big(\mathbf{w}_g^\top \phi(\mathbf{W}_g d^m_{ij} + \mathbf{b}_g) + \rho_m\big)$
    $\mathbf{z}_{ij} \leftarrow [\lambda^0_{ij} d^0_{ij} \| \lambda^1_{ij} d^1_{ij} \| \lambda^2_{ij} d^2_{ij} \| \lambda^3_{ij} d^3_{ij} \| \lambda^4_{ij} d^4_{ij}]$
    $\hat{y}_{ij} \leftarrow \sigma\Big(\mathbf{W}_o^{(3)}\Phi\Big(\mathbf{W}_o^{(2)}\Phi(\mathbf{W}_o^{(1)}\mathbf{z}_{ij} + \mathbf{b}_o^{(1)}) + \mathbf{b}_o^{(2)}\Big) + \mathbf{b}_o^{(3)}\Big)$
  $\mathcal{L}_{\mathrm{PPI}} \leftarrow \ell_{\mathrm{BCE}}(\{\hat{y}_{ij}\}, \{y_{ij}\})$
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{PPI}} + \eta \displaystyle\sum_{m=1}^{4} \mathcal{L}^m_{\mathrm{recon}}$
  Update all trainable parameters Θ using $\nabla_\Theta \mathcal{L}$
return Θ

Time Complexity Analysis. We describe ECMLF in two coupled stages: cross-modal representation learning with bridge reconstruction, followed by adaptive gated fusion for cold-start PPI prediction. Consistent with the main text, N_p, n, and D denote the number of proteins, the maximum number of regions per protein, and the representation dimension, respectively. For modality m, let N_e^m denote the number of biomedical entities, and let $V_m = N_p + N_e^m$ and E_m denote the corresponding numbers of nodes and edges. We further define $B_m = |\mathcal{M}_{\text{edge}}^m|$ as the number of sampled reconstruction triples and $N_{\text{pair}} = |\mathcal{D}_{\text{known}}|$ as the number of known-protein pairs. The sequence encoder uses $L_s = 6$ Transformer layers, while each graph encoder uses $L_g = 2$ GAT layers; hidden widths are specified as fixed multiples of D.

Reusing precomputed PLM embeddings, sequence encoding costs $O(N_p n D_{\text{PLM}} D + L_s N_p(n^2 D + nD^2))$, including projection, self-attention, and feature transformation. For each graph modality, relation-profile projection requires $O(N_p N_e^m D)$, while GAT propagation and bridge reconstruction incur $O(L_g(V_m D^2 + E_m D) + B_m D^2)$. Pooling and output projection are absorbed into these encoder terms. Adaptive gating and the two-hidden-layer prediction head cost $O(D^2)$ per protein pair. Thus, one optimization pass in Algorithm 1 requires

$$\begin{aligned}\mathcal{T}_{\text{total}} = O\Bigg(& N_p n D_{\text{PLM}} D + L_s N_p(n^2 D + nD^2) \\ & + \sum_{m=1}^{4} \left[N_p N_e^m D + L_g(V_m D^2 + E_m D) + B_m D^2 \right] \\ & + N_{\text{pair}} D^2 \Bigg).\end{aligned} \tag{16}$$

Backpropagation has the same asymptotic order. With fixed dimensions and layer counts, graph propagation, reconstruction, and prediction scale linearly with their respective graph or sample sizes. Self-attention remains quadratic in n, while relation-profile projection depends on $N_p N_e^m$. Full-graph encoding repeated across mini-batches incurs its cost at each update. During inference, cached protein representations reduce scoring Q candidate pairs to $O(QD^2)$.

A.2 DATASET

Table 5 summarizes the two datasets used in this study. Each dataset contains a PPI network, amino acid sequences, and four types of biomedical associations, including *protein-drug*, *protein-disease*, *protein-miRNA*, and *protein-lncRNA* relations. Dataset1 uses the STRING PPI network and integrates biomedical associations from DisGeNET, DrugBank, miRTarBase, and LncRNA2Target (Szklarczyk et al., 2017; Piñero et al., 2017; Wishart et al., 2018; Chou et al., 2018; Cheng et al., 2019). Dataset2 is built from an updated STRING network and additional resources, including DrugBank, LncTarD, miRTarBase, and CTD (Szklarczyk et al., 2025; Knox et al., 2024; Zhao et al., 2023).

Table 5: Statistics of cross-modal protein information in Dataset1 and Dataset2.

Dataset	Subpart	Proteins	Entities	Triplets
Dataset1	Protein-Drug	613	969	11,107
	Protein-Disease	832	692	25,087
	Protein-miRNA	489	271	4,944
	Protein-lncRNA	164	12	690
	Protein-Sequence	1,648	-	-
	All Proteins	**1,648**	**1,944**	**41,828**
Dataset2	Protein-Drug	2,226	7,079	19,057
	Protein-Disease	3,899	4,457	25,573
	Protein-miRNA	1,965	952	7,303
	Protein-lncRNA	1,304	1,108	4,826
	Protein-Sequence	4,283	-	-
	All Proteins	**4,283**	**13,596**	**56,759**

A.3 Evaluation Metrics

We evaluate PPI prediction using seven metrics: ACC, PRE, REC, F1, AUC, AUPR, and MCC. Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Higher values of these seven metrics indicate better predictive performance. We additionally report Wilcoxon signed-rank p-values and average Friedman ranks to assess paired performance differences and summarize relative rankings.

Accuracy (ACC). Accuracy measures the proportion of correctly classified protein pairs:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{17}$$

Precision (PRE) and Recall (REC). Precision measures the fraction of predicted interactions that are correct, while recall quantifies the fraction of actual interactions recovered:

$$PRE = \frac{TP}{TP + FP}, \quad REC = \frac{TP}{TP + FN}. \tag{18}$$

F1-score (F1). F1 combines precision and recall through their harmonic mean, accounting for both false positives and false negatives:

$$F1 = \frac{2\,PRE\,REC}{PRE + REC} = \frac{2\,TP}{2\,TP + FP + FN}. \tag{19}$$

AUC and AUPR. AUC measures the area under the ROC curve, tracing the true-positive rate (TPR) against the false-positive rate (FPR) as the decision threshold varies:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}. \tag{20}$$

AUPR measures the area under the precision-recall curve, summarizing precision across recall levels. The two areas are expressed as

$$AUC = \int_0^1 TPR(u)\,\mathrm{d}u, \quad AUPR = \int_0^1 PRE(r)\,\mathrm{d}r, \tag{21}$$

where u and r denote the false-positive rate and recall, respectively.

Matthews Correlation Coefficient (MCC). MCC measures the correlation between predicted and observed labels, incorporating all four confusion-matrix entries:

$$MCC = \frac{TP\,TN - FP\,FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{22}$$

Wilcoxon Signed-Rank Test (p-value). We use a one-sided Wilcoxon signed-rank test to compare ECMLF with each baseline. Within each cold-start setting, we average the seven metrics for each fold, pairing scores from the same dataset and fold. Let Δ_i denote the paired score difference between ECMLF and the baseline. After removing zero differences, we rank $|\Delta_i|$, assigning average ranks to ties. The signed-rank statistic is

$$W = \sum_{i=1}^{N} \mathrm{sgn}(\Delta_i) R_i, \tag{23}$$

where N is the number of nonzero differences, R_i is the rank of $|\Delta_i|$, and $\mathrm{sgn}(\cdot)$ denotes the sign function. Under the null hypothesis, paired differences are symmetrically distributed around zero. The one-sided p-value is obtained from the upper tail of the null distribution, testing whether ECMLF achieves higher scores. We use a significance threshold of 0.05.

Average Friedman Rank (F-rank). We rank all models within each dataset-metric combination, assigning rank 1 to the highest score and average ranks to ties. The average rank of model j is

$$R_j = \frac{1}{B} \sum_{i=1}^{B} r_{i,j}, \tag{24}$$

where $r_{i,j}$ denotes the rank of model j in comparison block i. Each main comparison table contains $B = 14$ blocks, covering seven metrics on two datasets under one cold-start setting. Lower F-rank values indicate better average relative performance.

A.4 BASELINE METHODS

Table 6 summarizes the baseline methods compared with ECMLF, including sequence-based, graph-based, and hybrid approaches to PPI prediction. These methods provide complementary perspectives: sequence-based approaches assess the predictive value of protein sequence features, graph-based approaches examine the contribution of relational information, and hybrid approaches combine these sources of evidence. Comparing ECMLF with these baselines across two datasets and both semi-cold-start and full-cold-start settings enables a broad evaluation of its performance.

Table 6: Summary of baseline models for PPI prediction.

Method	Description
HNSPPI (Xie et al., 2023)	A hybrid supervised model that characterizes protein pairs by integrating amino acid sequence information with connection properties derived from the PPI network for interaction prediction. (*Briefings in Bioinformatics 2023*)
EResCNN (Gao et al., 2023)	An ensemble model that fuses handcrafted sequence descriptors, including PseAAC, AC, PsePSSM, EBGW, MMI, and CT, and learns high-level features with residual CNNs. (*Computers in Biology and Medicine 2023*)
BaPPI (Tang et al., 2024)	A balanced PPI prediction model that combines attention-free sequence modeling, anti-symmetric graph convolution, and an asymmetric loss function to address imbalanced characteristics in PPI data. (*Bioinformatics 2024*)
KGF-GNN (Yang et al., 2024)	An end-to-end GNN integrating knowledge graphs to construct protein-associated networks and jointly learn protein representations from biomedical associations and observed PPI networks. (*TCBB 2024*)
HEENN (Yang et al., 2025)	A hybrid ensemble model that uses GAT to capture local topological features and an autoencoder to learn global representations, integrating both through attention-based fusion for PPI prediction. (*TCBB 2025*)
ESM2_AMP (Sun et al., 2025)	An interpretable sequence-based framework that extracts segment representations from ESM2 and uses Transformer fusion to reveal functional regions related to PPIs. (*Briefings in Bioinformatics 2025*)
GATSBI (Nayar & Altman, 2026)	A graph-attention framework initialized with ESM-2 protein sequence embeddings and adapted to our cold-start splits using observed positive training edges and supplied negative pairs. (*Bioinformatics 2026*)

A.5 IMPLEMENTATION DETAILS

For each dataset, proteins are randomly partitioned into five approximately equal folds. In each run, one fold is held out to form the new-protein set, and the remaining folds are treated as known proteins. In SCS, known-known PPI pairs are randomly split into training and validation sets (9:1), while known-new pairs are used for testing and new-new pairs are excluded. In FCS, known-known pairs are used for training and validation (9:1), new-new pairs for testing, and known-new pairs are excluded. The model hyperparameters are selected to maximize the mean of the seven validation metrics across folds. ECMLF is trained for up to 100 epochs with a batch size of 1024. Protein sequences are segmented according to UniProt structural-region annotations (The UniProt Consortium, 2025), and regional embeddings are extracted using ESM-2 *(esm2_t33_650M_UR50D)*.

A.6 ADDITIONAL ABLATION STUDY

Modality Ablation. Table 7 reports the modality ablation results on Dataset2 under both cold-start settings. Among graph-only variants, performance varies with the selected modalities, while several multi-graph combinations improve AUC and AUPR over individual graph modalities, suggesting that different biomedical graphs provide complementary relational evidence. Adding graph modalities to sequence features further improves both metrics: compared with the sequence-only model, the Full model increases AUC and AUPR by 3.07 and 2.42 percentage points under SCS, and by 6.09 and 4.35 points under FCS. These consistent gains show that integrating sequence features with biomedical graph evidence improves PPI prediction across both cold-start settings.

Table 7: Modality ablation results on Dataset2.

Modality	SCS								FCS							
	ACC	REC	PRE	F1	AUC	AUPR	MCC	F-rank	ACC	REC	PRE	F1	AUC	AUPR	MCC	F-rank
M1	55.16	96.26	52.78	68.17	68.64	65.01	18.32	**9.14**	58.19	40.58	58.37	47.88	57.52	58.47	15.50	**11.71**
M2	55.10	78.74	53.39	63.63	57.66	55.81	11.67	**12.86**	57.04	56.71	54.41	55.54	58.98	53.10	14.03	**12.14**
M3	52.36	74.27	51.57	60.88	58.44	61.20	5.36	**15.00**	49.95	73.92	48.12	58.29	55.78	52.79	2.63	**14.93**
M4	53.89	57.83	53.51	55.58	56.91	58.46	7.81	**15.29**	53.07	59.09	50.35	54.37	56.13	54.98	6.79	**14.29**
M1+M2	60.04	47.47	63.28	54.25	60.26	61.45	20.70	**11.00**	54.33	94.37	50.93	66.16	66.11	64.35	19.34	**8.14**
M1+M3	60.14	69.01	58.53	63.34	65.33	66.73	20.64	**9.71**	56.35	76.73	52.66	62.45	64.22	62.07	15.95	**9.71**
M1+M4	56.13	93.07	53.47	67.92	70.19	70.74	18.37	**6.86**	49.54	98.27	48.36	64.82	66.28	64.71	10.49	**9.57**
M2+M3	58.60	77.71	56.15	65.20	66.31	67.61	18.69	**9.14**	49.95	73.92	48.12	58.29	59.61	56.80	2.63	**13.93**
M2+M4	53.42	87.86	51.97	65.31	60.44	60.41	9.62	**12.29**	50.00	96.97	48.58	64.73	63.62	59.60	10.45	**10.57**
M3+M4	53.94	63.16	53.24	57.78	58.54	59.07	8.06	**14.57**	54.43	56.82	51.67	54.12	59.80	60.50	9.10	**12.71**
M1+M2+M3	59.88	77.59	57.22	65.87	66.35	65.80	21.19	**8.43**	58.35	79.00	54.09	64.22	66.57	63.36	20.31	**7.43**
M1+M2+M4	58.40	89.39	55.13	68.20	69.96	70.47	21.55	**6.14**	57.14	92.75	52.67	67.19	71.41	68.61	23.99	**5.29**
M1+M3+M4	64.33	76.71	61.42	68.22	69.63	69.48	29.63	**5.00**	61.67	70.89	57.73	63.64	69.07	68.28	24.58	**6.00**
M2+M3+M4	62.11	76.41	59.34	66.80	68.77	68.75	25.33	**6.86**	60.70	77.16	56.16	65.01	68.12	65.85	24.16	**5.57**
M1+M2+M3+M4	62.28	78.95	59.14	67.63	70.09	70.51	26.12	**4.86**	59.27	86.69	54.36	66.82	72.24	70.36	24.75	**4.14**
M_s	80.00	66.40	91.80	77.10	89.20	90.30	62.60	**3.57**	73.00	53.10	89.00	66.50	82.90	84.80	50.50	**4.14**
Full	84.61	77.44	90.43	83.43	92.27	92.72	69.95	**2.29**	79.56	67.53	88.78	76.65	88.99	89.15	60.90	**2.71**

Reconstruction Ablation. To assess the contribution of bridge reconstruction, we compare ECMLF with NO-Recon (removing the reconstruction module). As shown in Figure 4, ECMLF consistently achieves higher AUC and AUPR than NO-Recon across both datasets and cold-start settings. By recovering masked protein-entity associations, bridge reconstruction explicitly restores missing cross-modal links, uncovering latent similarity between known and new proteins and facilitating interaction-knowledge transfer. These results confirm its role in similarity modeling.

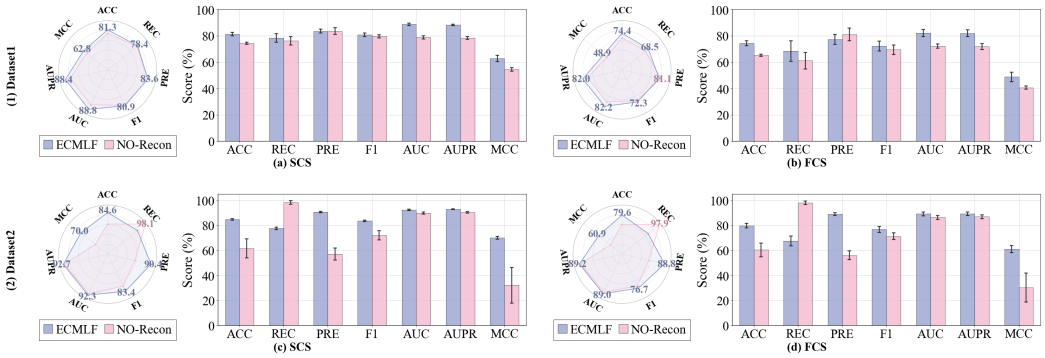


Figure 4: Bridge reconstruction ablation results across two datasets.

Sequence Segmentation Ablation. To evaluate structure-aware sequence partitioning, we replace it with ELS (equal-length segmentation). As shown in Figure 5, ECMLF generally achieves higher AUC and AUPR than ELS across both datasets and cold-start settings. By preserving structural organization and modeling region-level dependencies, structure-aware segmentation captures interaction-relevant sequence patterns more effectively, strengthening similarity modeling between known and new proteins. These results support its contribution to cold-start PPI prediction.

Graph Encoder Ablation. To assess the contribution of the graph-attention network (GAT) encoder, we replace it with MLP, GCN, and RGCN encoders. As shown in Figure 6, ECMLF generally achieves stronger AUC and AUPR than the alternative graph encoders across both datasets and cold-start settings. By assigning adaptive weights to neighboring proteins and biomedical entities, the GAT encoder captures informative relational evidence while reducing the influence of sparse or noisy connections. These results further demonstrate that attention-based aggregation effectively models cross-modal similarity for reliable cold-start PPI prediction.

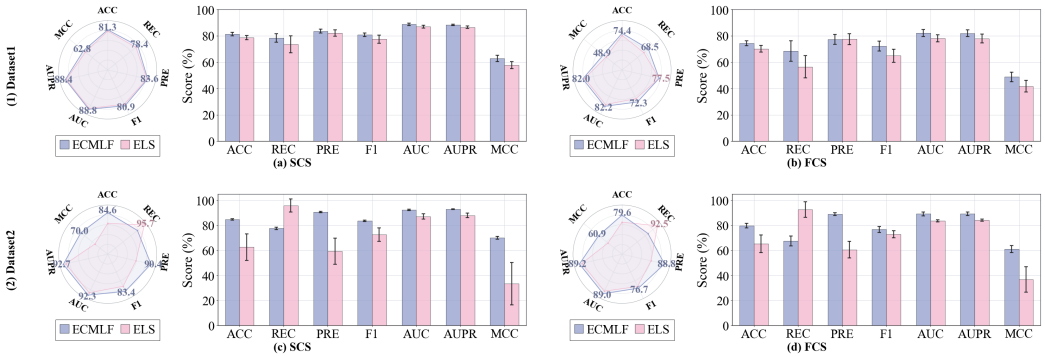


Figure 5: Sequence segmentation ablation results across two datasets.

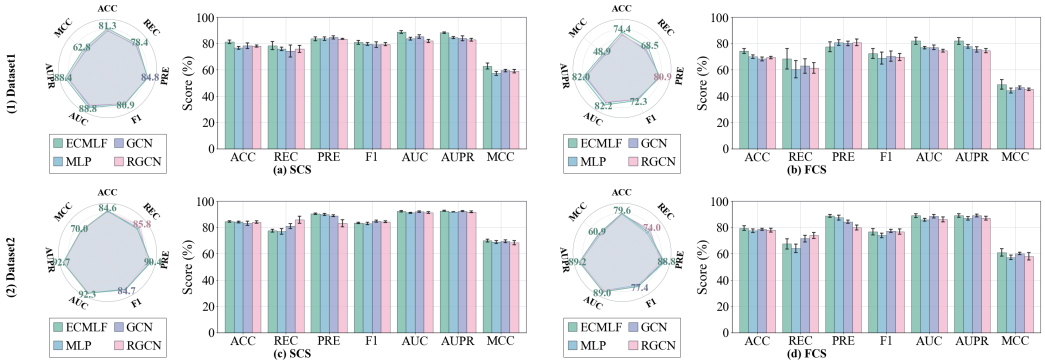


Figure 6: Graph encoder ablation results across two datasets.

Fusion Ablation. To examine the contribution of the adaptive fusion model, we compare ECMLF with variants using GatedABConcat (excluding absolute-difference and Hadamard-product features), attention pooling, and mean pooling. As shown in Figure 7, ECMLF generally achieves the strongest AUC and AUPR across both datasets and cold-start settings. By assigning pair-specific weights to multimodal pair representations, the adaptive fusion module aligns similarity modeling with the PPI prediction objective and emphasizes the most informative modalities for each pair. The results support its role in improving cross-modal knowledge transfer for cold-start PPI prediction.

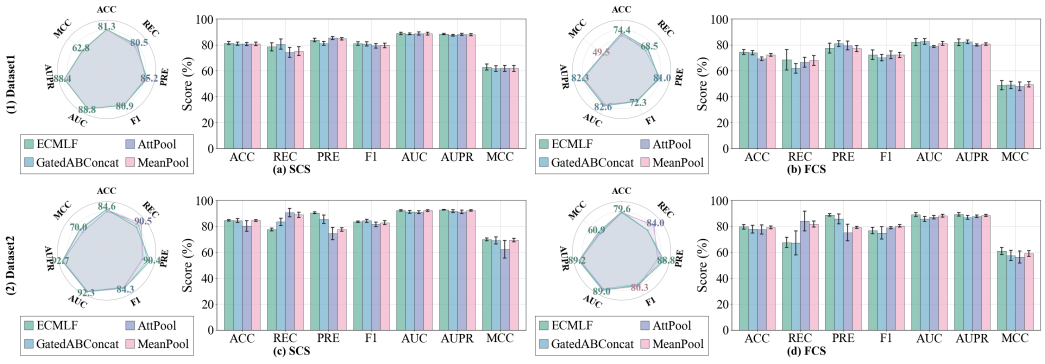


Figure 7: Fusion-strategy ablation results across two datasets.

A.7 Additional Explainability Results

Figure 8 reports the graph-pathway explainability results on Dataset2 under both SCS and FCS settings. ECMLF identifies more key bridge entities than HEENN, recovering shared biomedical associations that connect new proteins with associated known proteins. These entities provide relational evidence for interpreting the similarities captured by the model. Consistent with the findings on Dataset1, the results support the graph pathway's ability to recover biologically relevant connections across datasets and cold-start settings.

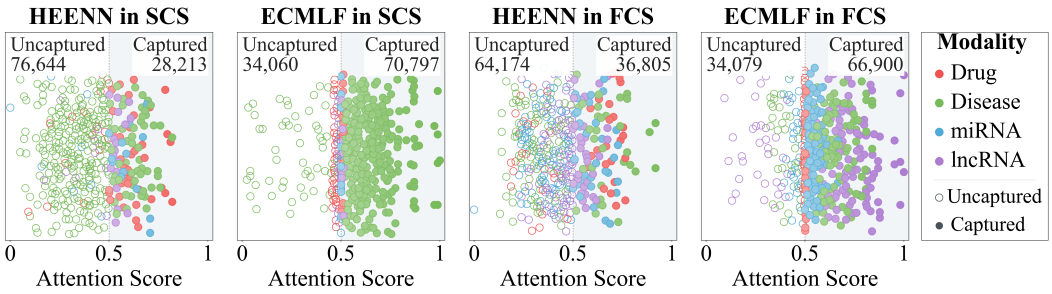


Figure 8: Graph-pathway explainability results on Dataset2.

A.8 Hyperparameter Sensitivity Analysis

We investigate three key hyperparameters of ECMLF, varying the learning rate over $\{1, 3, 5, 7\} \times 10^{-4}$, the embedding dimension over $\{32, 64, 128, 256\}$, and the reconstruction mask ratio over $\{0.1, 0.3, 0.5, 0.7\}$. Figure 9 summarizes the results, focusing on ACC and MCC. Both metrics show relatively small fluctuations under SCS, while their responses to hyperparameter changes are more pronounced under FCS. Under FCS, an embedding dimension of 128 yields the highest ACC and MCC on Dataset1, whereas 64 performs best on Dataset2. On Dataset2, a learning rate of 3×10^{-4} and a mask ratio of 0.3 also achieve the highest values for both metrics in their respective sweeps. Further increases in these hyperparameters do not consistently improve performance.

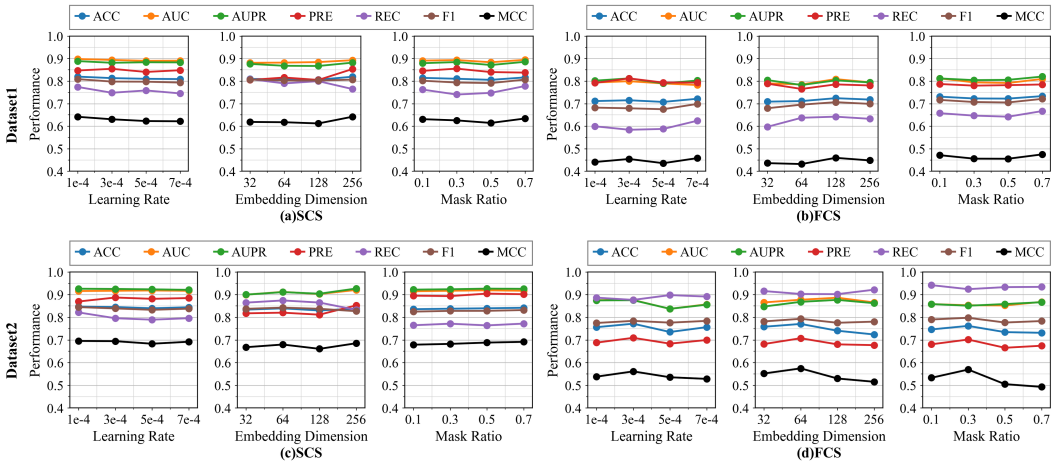


Figure 9: Hyperparameter sensitivity of ECMLF under SCS and FCS.