PPI2TEXT: CAPTIONING PROTEIN-PROTEIN INTERACTIONS WITH COORDINATE-ALIGNED PAIR-MAP DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein-protein interaction (PPI) modeling has been widely studied as a binary or multi-label classification task. While emerging multimodal large language models (LLMs) can now describe single proteins, they remain unable to generate free-form descriptions of interactions between protein pairs. Moving beyond controlled vocabulary annotations, we propose to model PPI using free-text description, enabling richer expressiveness, improved interpretability, and better integration with literature knowledge base. We present PPI2Text, a multimodal LLM for free-form PPI captioning from amino acid sequences, that encodes each protein using ESM3 encoder, constructs a pair map from the two representations to capture interactions across all residue pairs, and autoregressively generates descriptions using a Qwen3 language decoder. We further introduce PaCo-RoPE, a coordinate-aligned positional encoding that aligns each axis of the pair grid with the residue positions of the corresponding protein. In addition, we release PPI2Text-Dataset, a 351k-pair corpus of free-form PPI descriptions aggregated from ten curated biological databases and further synthesized with Gemini under evidence-tiered prompting. PPI2Text consistently outperforms strong baselines across multiple ablation settings and evaluation protocols. It not only achieves higher scores on linguistic metrics against synthesized references, but also excels on factuality metrics, where an LLM-based judge evaluates outputs against raw biological evidence. [1] [2]

1 INTRODUCTION

Protein–protein interactions (PPIs) drive a wide range of cellular processes by allowing proteins to form complexes involved in metabolism, signaling, DNA transcription and replication, and immune responses (Stites, 1997). While interaction characterization can be obtained through experimental validation, such approaches are time- and resource-intensive and do not always yield details (Rao et al., 2014). To reduce these practical constraints, existing computational PPI approaches mostly relied on classification and regression modeling to extract evolutionary signals and encode interactions through conservation patterns. However, these representations struggle to capture higher-order and non-linear dependencies and are further limited by heterogeneous data sources. Furthermore, a substantial fraction of biologically relevant PPIs remain undiscovered or uncharacterized. Consequently, continued research into robust and interpretable computational approaches for PPI characterization remains essential for advancing our understanding of complex cellular systems and improving biological discovery at scale.

Despite the availability of large-scale biological databases such as IntAct and STRING, PPI-related information remains fragmented and inconsistently structured. As a result, current approaches often reduce interaction prediction to discrete labels or fixed relational structures, which limits their ability to capture how or why a biological feature interact within context. We propose protein–protein interaction captioning, a new perspective that offers a complementary paradigm in which interactions are represented as coherent free-text descriptions, rather than being reduced to fixed categorical

[1] The source code of PPI2Text model is available at: https://anonymous.4open.science/r/PPI2Text-A74A

[2] The 351k corpus data of PPI2Text-Dataset is uploaded as supplimentary material on OpenReview.

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.

labels or numerical scores. By expressing interactions through flexible natural-language narratives, this approach enables the integration and interpretation of individual protein-specific features and emergent properties arising from their interaction, as many of them are strongly condition-dependent and emerge only through complex chains of biological events (Chindelevitch et al., 2012). Nevertheless, PPI free-text captioning remains an emerging research area, and relatively few studies have explored the logical integration of rigid interaction classifications into interpretable natural-language descriptions, despite its potential to improve biological interpretability, hypothesis generation, and contextual understanding of molecular interactions. Further details are elaborated in Appendix A.

To address these limitations, we propose a unified framework that enhances both representation and data quality by shifting from structured prediction paradigms toward a more expressive formulation of PPI understanding. Our approach is designed to better reflect the contextual and relational nature of protein interactions, enabling richer modeling of dependencies across heterogeneous biological sources. Our contributions are summarized as follows:

- **Free-text PPI modeling framework:** We introduce a novel task formulation to cast PPI modeling as human-readable description generation, enabling direct interpretation of interaction semantics beyond fixed annotation schemas.
- **Large-scale PPI description dataset:** We develop a data synthesis pipeline to augment natural language descriptions for PPI to improve expressiveness and cross-dataset generalization. A 351k-pair high-quality description corpus is released.
- **Interaction-aware protein encoding:** We propose *PPI2Text* as a unified architecture to capture residue-level relationships between proteins through pairwise interaction map modeling, with coordinate-aligned rotary positional encoding (*PaCo-RoPE*) to align multiple multimodal components and to boost understanding.
- **Strong performance on empirical evidence:** We demonstrate that the model outperforms various baselines on both linguistic metrics against synthesized text as well as factuality metrics against raw evidence.

2 Related Work

Sequence-based single-protein description: Early protein modeling relied on task-specific sequence-based methods with fixed-label supervision for function and interaction prediction. More recent large language models generate free-text descriptions from sequences (Taylor et al., 2022; Liu et al., 2024; Zhou et al., 2025b; Fei et al., 2025), translating embeddings into natural language. However, they are designed for single protein input, and are trained on single-protein data and general biochemical text without detailed grounding in interaction data, limiting their ability to capture interaction-specific semantics and relational context.

Deep learning models for PPI prediction: RNN-based models have been widely used for sequence representation learning, often within hybrid frameworks that integrate structural or biophysical features (Alakus & Turkoglu, 2021; Zhou et al., 2022). LSTM-based approaches further improve performance through regularization techniques (Tsukiyama et al., 2021; Szymborski & Emad, 2022; Deng et al., 2020). More recently, attention-based methods and transformer architectures have advanced PPI prediction by capturing long-range dependencies and complex protein relationships (Hu et al., 2024a;b; Li et al., 2022; Wu et al., 2024; Mou et al., 2023; Kang et al., 2023). However, these approaches often rely on handcrafted features or hybrid pipelines, which can introduce bias and limit scalability. Emerging tools for PPI 3D structure prediction, like the AlphaFold family (Evans et al., 2021; Abramson et al., 2024), have become widely used resources. However, these methods currently serve as orthogonal tools to PPI captioning because they produce structural representations of interactions rather than natural-language descriptions.

Sequence-based LLMs for PPI prediction: Recent LLM-based approaches model protein–protein interactions as sequence-based prediction tasks using pretrained embeddings and prompting strategies (Jin et al., 2024; Hallee & Gleghorn, 2023). Multimodal and retrieval-augmented methods incorporate sequence, text, and network data or biomedical evidence for improved prediction and reasoning (Zhuo et al., 2024; Zhou et al., 2025a; Ullanat et al., 2026; Jeon et al., 2026). However, these methods remain schema-bounded, producing labels or scores that compress biological context and limit the representation of mechanisms and evidence.

Raw Evidence:
[IntAct] `18 evidence lines, MIscore 0.94`; physical association...
[PubMed] `11 publications`
[UniProt] `ProtA:` transcriptional regulator and general inhibitor of RNA polymerase II...
[UniProt] `ProtB:` catalytic subunit of P-TEFb; phosphorylates C-terminal domain of RNA Pol II...
[CORUM] `2 complexes`; 7SK RNP, P-TEFb-7SKRNA-HEXIM1
[STRING] `0.99 hybrid confidence` [3did] `null` [SIGNOR] `null` [ComplexPortal] `null`

Tags: [Length] `long` [Verb] `firm` [Mechanism] `absent` [Silence] `none`

Synthesis: Protein A and Protein B are core components of the 7SK RNP complex and the P-TEFb-7SKRNA-HEXIM1 complex, where they physically associate to regulate RNA polymerase II transcription. Protein A is a transcriptional regulator belonging to the HEXIM family...
The catalytic activity of Protein B is distributed throughout its protein kinase domain and is regulated by a complex pattern of post-translational modifications, including autophosphorylation at the T-loop, N-terminal acetylation, and polyubiquitination by UBR5...

Figure 1: An example of the dataset generation (interaction between HEXIM1 and CDK9). The top section shows the raw data aggregated from multiple heterogeneous sources, while the middle section lists the control tags computed from the aggregated evidence profile, and the bottom section is the labeler-LLM output, synthesized from the raw evidence under the control-tag constraints.

3 Evidence-Tiered PPI Description Dataset Generation

PPI annotations capture rich biological knowledge but remain sparse, heterogeneous, and fragmented. Integration efforts such as OmniPath (Türei et al., 2016) and ConsensusPathDB (Kamburov et al., 2009) unify these data, but rely on rigid schemas that limit contextual expressivity and evidence integration. In this paper, we propose a dataset synthesis labeling pipeline that projects PPI data into a natural language space via free-form text augmentation, enabling more expressive interaction modeling and better cross-dataset generalization.

Source aggregation: We first construct a comprehensive and reliable factual base by integrating ten complementary sources capturing interaction-level evidence, molecular context, structural constraints, and higher-order functional organization: IntAct (Kerrien et al., 2012), PubMed (Canese & Weis, 2013), UniProt (Consortium, 2019), 3did (Mosca et al., 2014), Pfam (Mistry et al., 2021), STRING (Mering et al., 2003), SIGNOR (Perfetto et al., 2016), Reactome (Croft et al., 2010), CORUM (Giurgiu et al., 2019), and ComplexPortal (Meldal et al., 2015). These sources are selected to ensure coverage across experimental, structural, and curated biological evidence, with further details provided in Appendix B.1.

After collection, integration and de-duplication, we obtain approximately 1.08M protein pairs. However, annotation quality is highly imbalanced. Many PPIs are supported only by single high-throughput assays (e.g., yeast two-hybrid), which provide limited mechanistic resolution. This heterogeneity introduces a fundamental trade-off between descriptive expressivity and factual fidelity in downstream synthesis: while strongly supported PPIs allow reliable detailed descriptions, weakly supported interactions risk hallucinated or over-specified narratives.

Evidence-tiered quality filtering: To address this, we introduce an evidence-tiered filtering and control strategy. We define a unified *evidence score* $E(r)$ for each pair r that integrates complementary interaction-level and contextual biological signals, capturing both direct experimental support and broader biological coherence between interacting partners. We aggregate these signals while respecting their structured dependencies and heterogeneous reliability, yielding a score that reflects consistency, redundancy, and overall evidential strength. Further details can be found in Appendix B.2.

We then apply K-means clustering over the evidence scores to partition PPIs into three tiers (T1–T3), ranging from weakest to strongest support. We discard T1 due to insufficient evidence beyond single high-throughput assays, retain all T3 interactions as strongly curated, and apply a homology-aware subsampling strategy to intermediate T2 tier using MMSeq2 (Steinegger & Söding, 2017), selecting representative high-evidence pairs within homologous clusters. This process yields approximately 351k high-confidence PPIs, balancing diversity with evidential reliability and forming a robust foundation for LLM-based synthesis. Further details are provided in Appendix B.3 and B.4.

Description synthesis pipeline: Finally, we augment the retained PPIs using a labeler LLM. The full system prompt is detailed in Appendix I. To ensure faithfulness to the underlying evidence, we introduce a tier-controlled generation framework that explicitly constrains the output space according to the strength of biological support. This reduces the effective hypothesis space of the model, preventing unsupported extrapolation or over-detailed descriptions for weakly supported interactions. Four orthogonal constraints are explicitly enforced in each prompt: *descriptive granularity*, *epistemic strength*, *mechanistic attribution*, and *silence policy* (see Appendix B.5).

We release the resulting dataset of 351k protein interactions, each paired with a free-form textual description of its interaction context and mechanism, synthesized by Gemini-3-Pro-Preview (Team et al., 2023).

4 PPI Free-Text Modeling with Pair-Map

PPI takes place at the interface between the two proteins and is determined by the set of contacts between their residues. A natural representation of this information is a residue-level coupling matrix, with positions on protein A along one axis and positions on protein B along the other (Evans et al., 2021). In this paper, we propose *PPI2Text*, as shown in Figure 2, that models the interaction by this 2D representation, then bridges it to the 1D token sequence of a language decoder, enabling direct attention over pairwise residue interactions. The design avoids compressing the 2D interaction map into a global vector or relying solely on single-protein embeddings.

4.1 Single-Protein Parallel Encoding

We use pretrained ESM3 (Hayes et al., 2025) as the frozen single-protein encoding backbone. ESM3 is a multimodal foundation model jointly pretrained over multiple per-residue tracks including amino-acid sequence, AlphaFold2-predicted (Jumper et al., 2021) 3D structure coordinates, plus the derived discretized structure tokens, secondary structure (SS8), solubility factors (SASA), and structure confidence scores (pLDDT). For each protein $P \in \{A, B\}$, it returns per-residue embeddings $E_P \in \mathbb{R}^{L_P \times d_{ESM}}$, where L_P is the length of the protein (see Appendix C). We use ESM3 specifically because its multitrack pretraining already entangles evolutionary and structural signals in a single tensor.

Then, a two-layer strided 1D convolutional compressor reduces the sequence-long dimension of the embedding X_P by a factor of four, with parameters shared across the two proteins, so that $X_P \in \mathbb{R}^{L'_P \times d_{\text{hidden}}}$, where the compressed length $L'_P = \lfloor L_P/4 \rfloor$ depends only on L_P. Since per-residue ESM3 embeddings are already context-rich through the encoder's full self-attention stack, a moderate pooling preserves most of the local information. Moreover, compression is necessary because both protein sequences can be up to thousands of residues long, and uncompressed representations can lead to redundant and noisy inputs to the decoder, affecting both training and inference efficiency. The two single-protein representations X_P are projected then passed as part of the input embeddings to the decoder to provide partner context information on the interaction. However, we argue that the interaction information cannot be fully captured by single-protein information, and a 2D pair map is necessary.

4.2 Pair-Map Encoding

We use two bidirectional cross-attention blocks to exchange information between the two protein representations before any pair representation is formed, with weights also shared between the directions $A \rightarrow B$ and $B \rightarrow A$. Each block combines a cross-attention to the partner and a

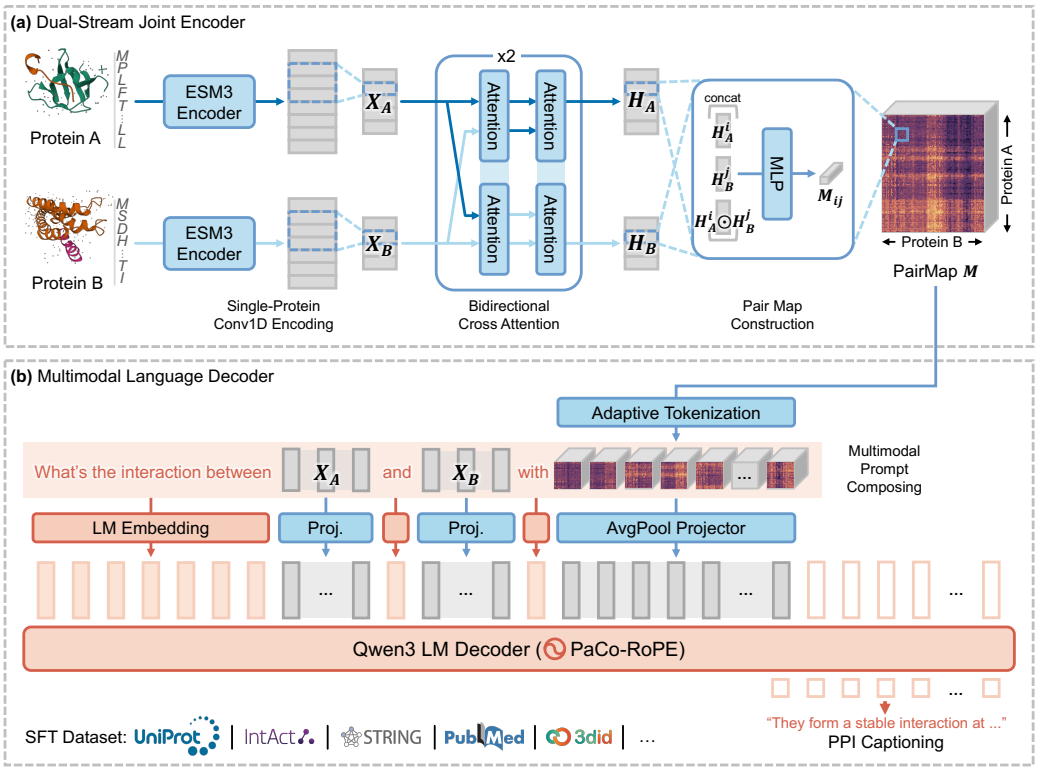

Figure 2: Model architecture of PPI2Text. **(a)** In dual-stream joint encoder, both proteins are first separately encoded and compressed in parallel then jointly encoded to construct a PairMap, modeling interactions between pairs of residues. **(b)** For multimodal language decoding, both single-protein representations and pair map tokens from the encoder are projected then concatenated with text token embeddings to compose the interleaved multimodal prompt, and finally a Qwen3 language model decoder with Pair-Coordinated RoPE (PaCo-RoPE) generates the free-text description of the interaction.

self-attention to each protein itself, and the output of the last block is marked as $H_A \in \mathbb{R}^{L'_A \times d_c}$ and $H_B \in \mathbb{R}^{L'_B \times d_c}$. As inspired by MINT (Ullanat et al., 2026), the cross attention mechanism naturally guide corresponding interactive segments to enrich then highlight their role and context in the interaction. Furthermore, we interleave self-attention layers to stabilize the interaction embedding and reinforce the integration of global contextual information for each protein.

Then, for each $(i, j) \in [L'_A] \times [L'_B]$, we form a per-position pair feature

$$M_{ij} = \phi_{\text{pair}}\Big([H^i_A \,\|\, H^j_B \,\|\, H^i_A \odot H^j_B]\Big) \in \mathbb{R}^{d_{\text{hidden}}}, \tag{1}$$

where $\|$ is concatenation, $\odot$ is the Hadamard product, and $\phi_{\text{pair}} : \mathbb{R}^{3d_{\text{hidden}}} \to \mathbb{R}^{d_{\text{hidden}}}$ is a 2-layer MLP with GeLU as the activation function of the first layer to introduce non-linearity. The Hadamard term captures coordinate-wise feature co-activation between H^i_A and H^j_B. Remarkably, we choose to project every pair of H^i_A and H^j_B into an embedding vector instead of a scalar, aiming to maximize the semantic understanding of the mechanistic details and biological context. The resulting 3D tensor $M \in \mathbb{R}^{L'_A \times L'_B \times d_{\text{hidden}}}$ is thus length-dependent in this case.

To inject the pair-map into the decoder as pseudo-tokens, we apply adaptive average pooling to a target $H_t \times W_t$ grid, flatten the grid into sequential patches, and project the result to the decoder's hidden space:

$$\bar{M} = \text{AdaptAvgPool}_{H_t \times W_t}(M), \tag{2}$$

$$U = \text{RMSNorm}(\phi_{\text{tok}}(\bar{M})) \in \mathbb{R}^{H_t W_t \times d_g}, \tag{3}$$

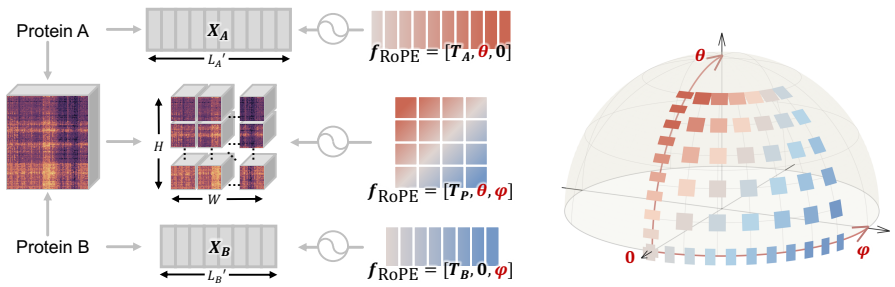

Figure 3: Illustration of the PaCo-RoPE as an extension to standard rotary positional embedding to paired-protein inputs. Each single protein representations X_A , X_B carries its own positional axis, and the two axes of the pair map are inherited from the single-protein axes correspondingly with residue-level alignment.

where ϕ_{tok} is a 2-layer MLP and d_{qwen} is the Qwen3 hidden size. We set $H_t = W_t = 32$, which adds 1024 pair-map tokens regardless of protein length. The fixed token budget in the decoder aims to maintain a rational information density and local granularity so that the decoder can focus on most significant aspects of the interaction without being distracted to noisy side-elements. The adaptive pooling also preserves the variety of information in each patch by orthogonal elements in the higher dimension representations. A Root-Mean-Square (RMS) normalization is applied in the end to align the scale of U with the text token embeddings.

4.3 Coordinate-Aligned Decoding

In the end, both single-protein representations X_P and the tokenized pair map representations U are inserted into the natural language prompt sentence with their corresponding interleaved positions. The composed multimodal prompt is passed to Qwen3 language model decoder which auto-regressively generates the description of the interaction.

However in the pretrained Qwen3 language model (Yang et al., 2025), the native 1D positional encoding over the flattened prompt would give the 1024 pair-map tokens consecutive indices, discarding the 2D grid structure of the pair-map and decoupling the pair-map tokens from the residue indices of the two proteins. We propose *PaCo-RoPE* to address both issues. It extends the multimodal interleaved RoPE of Qwen3-VL (Bai et al., 2025) to a pair-coordinate grid, where each protein $P \in \{A, B\}$ is tied to a dedicated spatial channel and the pair-map sits at the intersection of the two protein channels.

As illustrated in the Figure 3, every text or multimodal token is mapped to a unique 3D position IDs $(p^T, p^\theta, p^\varphi)$ as follows:

Text tokens: For text token at position n in the full sequence of prompt plus the generated answer, we apply:

$$(p^T, p^\theta, p^\varphi) = (n, n, n) \tag{4}$$

We collapse the 3D-RoPE to its canonical 1D form where all three channels are filled with their actual positions in the sequence. The behavior falls back to the native settings of Qwen3 model, inheriting the pretrained natural language ability.

Single-protein tokens (protein A): For single-protein representation tokens of protein A (X_A) that is at position n in the full sequence but represents continual residues centered at position i in the protein A, we apply:

$$(p^T, p^\theta, p^\varphi) = (n, \lfloor i/4 \rfloor, 0) \tag{5}$$

We specially use the second channel to indicate the location of the residue in the amino-acid sequence of protein A, while the third channel is muted. The first channel is maintained to represent the actual position in the full sequence. Such orthogonal isolation treats the protein sequence A as an independent modality to the text, enabling better understanding by the model.

Single-protein tokens (protein B): Similarly, for single-protein representation tokens of protein B (X_B), we use the third channel for j the centered location of the residue in the protein B while the

second channel is muted instead:

$$(p^T, p^\theta, p^\varphi) = (n, 0, \lfloor j/4 \rfloor) \tag{6}$$

As we isolate the position dimension of these two partners for the interaction, the decoder treats them independently without confusion on the dual multimodal inputs.

Pair-map tokens: For pair map patch of grid (s, t) and serialized into token at position n in the full sequence, we apply:

$$(p^T, p^\theta, p^\varphi) = \left(n, (s + \frac{1}{2})\frac{L'_A}{H}, (t + \frac{1}{2})\frac{L'_B}{W}\right) \tag{7}$$

so that their grid positions are tied to the position of residues on both protein A and B. Due to the adaptive average pooling that segment the entire pair map into small patches, the hop in distance on neighboring single protein embeddings is different to that on neighboring patches. With both height and width coordinates aligned between the pair map and both single proteins, the decoder can not only better understands the 2D nature of the pair map despite the serialized tokenization, but also better process these two different forms of multimodal inputs comprehensively.

After all, the encoding supplies a coordinate-aligned shortcut between the pair map and each per-protein stream at zero parameter cost, and preserves the maximal pretrained behavior of the Qwen3 decoder (see Appendix D for further details).

4.4 Training Setup

We fine-tune the PPI2Text model in an end-to-end way with cross-entropy loss computed only on response tokens under a teacher-forcing protocol. Both pretrained ESM3 and pretrained Qwen3 weights stay frozen, and we apply LoRA (Hu et al., 2022) to Qwen3, which closes the residual gap between the pretrained text distribution and the PPI description manifold. The remaining new modules are trained from scratch with a separate optimizer group, since pretrained adapters and from-scratch modules require different update scales.

5 Experiments and Results

Dataset and Test Settings: We evaluate our model on the synthesized PPI description dataset introduced in Section 3, focusing on generalization under two complementary test settings that capture different degrees of protein novelty, from realistic deployment to strict out-of-distribution (OOD) generalization. First, we introduce a *temporal holdout* split that reserves all PPIs annotated after May 1, 2025 for evaluation, approximating real-world deployment where the model predicts newly discovered interactions from previously available biological knowledge. Pair-level similarity-based decontamination is further applied to the remaining training samples to prevent leakage. Second, we adopt a more challenging *C3-hard* split with strict protein-level decontamination, ensuring no homologous proteins are shared between training and test sets. By excluding test proteins similar to those seen during training, this setting simulates interactions involving entirely novel proteins and requires generalization to unseen biological contexts. Following the increasingly adopted C3 protocol and the evaluation framework of Bernett et al. (2024), it serves as a stringent stress test of OOD generalization. The model is trained on 280k PPIs with 2,500 validation samples. The temporal holdout and C3-hard test sets contain 1,690 and 2,730 interactions, respectively. Further decontamination details are provided in Appendix F.

Baselines and Ablations: As prior works focus on structured label prediction that are not directly comparable in output space, we evaluate PPI2Text through representative learning backbone baselines and ablations. *Seq+Qwen3* uses no protein-specific encoder, but feeds raw amino-acid sequences directly to the decoder as trainable special tokens. *MINT+Qwen3* employs the pretrained state-of-the-art PPI encoder MINT with Qwen3 as the decoder. We further design four ablation variants to analyze the contributions of individual components. *SingProt-Only* passes only the single-protein representations X_A and X_B to the decoder, while *PairMap-Only* uses only the tokenized pair map U. The role of the proposed PaCo-RoPE is examined in *1D-RoPE*, where it is downgraded to the standard encoding in Qwen3. Finally, *No-Cross* removes the bidirectional attention mask module to evaluate its impact relative to the full *PPI2Text* model. For further reproducibility information on hyperparameters and computational resources, see Appendix E.

Table 1: Baseline and ablation results on temporal holdout set. Lexical metrics includes BLEU-2/4 scores (B-2/4), ROUGE-1/2/L F1 scores (R-1/2/L), with semantic BERTScore using RoBERTa (RBT) and BioBERT (BBT) embeddings. LLM-as-a-judge scores against the raw evidence cover Entities (Ent), Interaction (Int), Mechanism (Mec), and the Average (Avg).

	(against Synthesized Text)							(against Raw Evidence)			
Model	**B-2**	**B-4**	**R-1**	**R-2**	**R-L**	**RBT**	**BBT**	**Ent**	**Int**	**Mec**	**Avg**
Seq+Qwen3	37.37	20.87	52.83	23.14	28.66	88.02	80.74	1.94	7.50	1.85	3.76
MINT+Qwen3	41.54	25.36	56.22	29.17	33.67	89.03	82.48	2.68	7.80	1.63	4.04
SingProt-Only PPI2Text	43.35	27.61	58.44	31.88	36.00	89.54	83.71	4.75	8.26	2.61	5.21
PairMap-Only PPI2Text	40.98	25.36	56.96	29.79	34.02	89.23	83.09	3.22	7.95	1.03	4.07
1D-RoPE PPI2Text	39.17	23.61	54.63	27.45	32.50	88.52	81.66	3.00	7.31	2.01	4.11
No-Cross PPI2Text	45.61	29.86	60.24	34.71	38.91	89.94	84.02	4.94	7.98	4.67	5.86
PPI2Text	**48.33**	**31.86**	**62.15**	**36.56**	**39.96**	**90.79**	**86.33**	**7.72**	**8.32**	**5.59**	**7.21**

Evaluation Metrics: We first evaluate the prediction from PPI2Text against the synthesized text of test samples. We use lexical metrics including BLEU scores and ROUGE scores, then semantic metrics like BERTScore with RoBERTa and BioBERT embeddings. Moreover, to decouple any potential information bottleneck or hallucination in the synthesized text, we evaluate the prediction against the raw evidence that are gathered from sources before the synthesis. To scale up the evaluation, we use Claude-Opus-4.7 as a judge to evaluate three orthogonal aspects of the generated test. Raw evidence evaluates directly extracted interaction components: Entities (Ent), Interaction (Int), Mechanism (Mec), and the metric's overall average (Avg).

Results: Through our evaluation setup, Table 1 compares PPI2Text with baselines and ablations. The raw-sequence baseline performs poorly, highlighting the limited biological signal available from amino acid tokens alone, while pretrained encoders such as MINT offer modest improvements by leveraging learned evolutionary, sequential, and structural information. The ablations further demonstrate the complementary roles of single-protein representations and pair map tokens, as removing either degrades performance. The poor performance of 1D-RoPE suggests that flattening the pair map disrupts its inherent 2D structure, motivating our PaCo-RoPE design. Similarly, the No-Cross variant confirms the importance of bidirectional cross-attention in pair map construction. These findings extend to raw evidence-based evaluation, where PPI2Text achieves the strongest overall performance, particularly in entity grounding and mechanism fidelity, demonstrating its ability to preserve biological information beyond surface-level text generation.

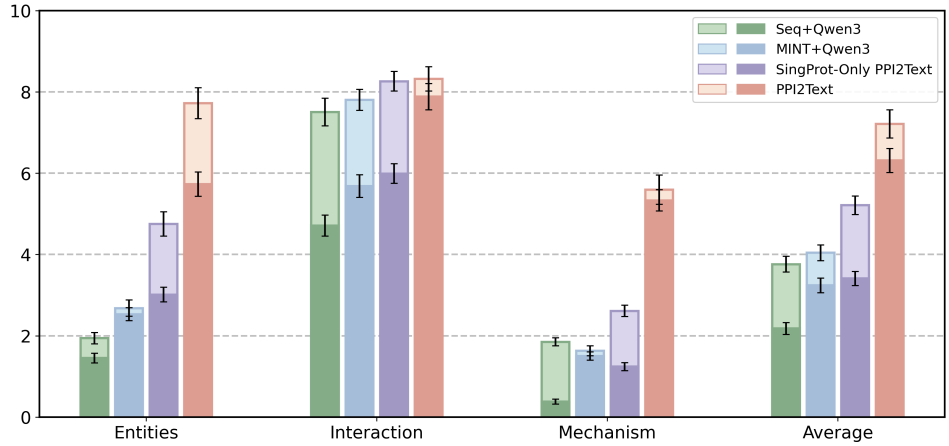

Figure 4: LLM-as-a-judge scores on baselines and ablations: evaluating the factual correctness of predicted text against raw evidence. Lighter bars denote performance on the temporal holdout split, while darker colors denote performance on the C3-hard split. PPI2Text consistently outperforms all other methods by a clear margin.

Table 2: Agreement between the LLM judge and two human experts (H1 and H2) across three scoring axes. Krippendorff's α measures inter-rater agreement, while Spearman's ρ measures pairwise rank correlation. HA (Human-Average) denotes the average of the two human expert scores.

	Krippendorff's α	Spearman's ρ			
Axis	**LLM–Experts**	**LLM–H1**	**LLM–H2**	**LLM–HA**	**H1–H2**
Entities	0.9416	0.8829	0.9248	0.9154	0.9343
Interaction	0.7441	0.7511	0.7215	0.7923	0.8190
Mechanism	0.8593	0.8299	0.8849	0.8816	0.8787
Average	**0.8984**	**0.8343**	**0.9028**	**0.8800**	**0.9179**

Figure 4 compares performance across the temporal holdout and C3-hard splits (see Appendices G and F). All models perform better on the temporal holdout due to greater training–test distribution overlap, whereas C3-hard imposes stricter OOD generalization by excluding homologous proteins across splits. Despite performance declines across all models, PPI2Text consistently outperforms baselines and ablations on both splits. Its gains in raw evidence-based metrics further suggest improved biological grounding rather than surface-level learning, demonstrating robust generalization in free-text PPI modeling under challenging OOD conditions.

Human validation of LLM-as-a-judge: To assess the reliability of our automated evaluation, we conducted a human validation study on 50 randomly sampled PPIs, independently evaluated by two human experts. For each PPI, every evaluator was presented with the complete collected raw evidence alongside the description generated by PPI2Text and asked to assess its biological fidelity along the same three axes. We quantify agreement using Krippendorff's α and pairwise Spearman's ρ, comparing the LLM-judge against individual experts and their averaged scores, as well as measuring inter-expert consistency. As shown in Table 2, the LLM-judge achieves a Krippendorff's α of 0.8984 on the per-PPI average of the three evaluation axes. Notably, inter-expert correlations are comparable in magnitude to those between the LLM-judge and individual experts, indicating that discrepancies involving the LLM-judge are broadly consistent with natural variability in human assessments. These findings support the reliability of our automated evaluation framework as a scalable proxy for expert judgment in assessing the biological fidelity of generated PPI descriptions.

6 Limitations and Future Work

A limitation of the framework is its assumption that every input pair forms a protein–protein interaction, requiring a prior PPI screening step, although binary PPI screening is a well-studied task with many strong existing models. In addition, the lack of wet-lab validation limits empirical verification of de-novo predictions. Future work will incorporate interaction-level structural information, including predicted complex structures from models such as AlphaFold3, beyond the individual protein structures currently used. Additional safeguards are discussed in Appendix H.

7 Conclusion

In this work, we introduce, to our knowledge, the first framework for generating free-form, human-readable descriptions of protein–protein interactions. By moving beyond rigid annotation schemas and task-specific classifiers, our approach models interaction semantics in natural language, enabling richer and more interpretable representations. Experiments with strong baselines and targeted ablations show that each component contributes meaningfully, and that the proposed interaction-aware encoding captures fine-grained residue-level relationships. These improvements are reflected in both standard text generation metrics and evidence-grounded evaluations, indicating greater biological consistency and mechanistic fidelity. We believe this work also establishes a foundation for future research opening directions for downstream applications such as knowledge extraction and reasoning. To support further progress, we release a dataset of 351k interaction descriptions to encourage the development of more expressive and interpretable models.

AI USE STATEMENT

In this work, the free-form reference descriptions of our synthetic PPI2Text-Dataset are synthesized by a labeler LLM (Gemini-3-Pro-Preview) from aggregated curated evidence as described in Section 3 and Appendices B.5. We further used Claude-Opus-4.7 as an automated LLM-as-a-judge to evaluate results. We have not used generative AI tools to propose or refine hypotheses, to design the research methodology or experiments, or to formulate or prove mathematical claims. Additionally, we used generative AI tools to edit the manuscript for readability.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Talha Burak Alakus and Ibrahim Turkoglu. A novel protein mapping method for predicting the protein interactions in covid-19 disease by deep learning. *Interdisciplinary Sciences: Computational Life Sciences*, 13(1):44–60, 2021.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Judith Bernett, David B Blumenthal, and Markus List. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(2):bbae076, 03 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae076. URL https://doi.org/10.1093/bib/bbae076.

Kathi Canese and Sarah Weis. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1):2013, 2013.

Leonid Chindelevitch, Daniel Ziemek, Ahmed Enayetallah, Ranjit Randhawa, Ben Sidders, Christoph Brockel, and Enoch S Huang. Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics*, 28(8):1114–1121, 2012.

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47 (D1):D506–D515, 2019.

David Croft, Gavin O'kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697, 2010.

Lei Deng, Jiaojiao Zhao, and Jingpu Zhang. Predict the protein-protein interaction between virus and host through hybrid deep neural network. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 11–16, 2020. doi: 10.1109/BIBM49941.2020.9313117.

Richard Evans, Michael O'neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Xiao Fei, Michail Chatzianastasis, Sarah Almeida Carneiro, Hadi Abdine, Lawrence P Petalidis, and Michalis Vazirgiannis. Prot2text-v2: Protein function prediction with multimodal contrastive alignment. *arXiv preprint arXiv:2505.11194*, 2025.

Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research*, 47(D1):D559–D563, 2019.

Logan Hallee and Jason P Gleghorn. Protein-protein interaction prediction is achievable with large language models. *bioRxiv*, pp. 2023–06, 2023.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/abs/10.1126/science.ads0018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jun Hu, Kai-Xin Chen, Bing Rao, Jing-Yuan Ni, Maha A Thafar, Somayah Albaradei, and Muhammad Arif. Protein-peptide binding residue prediction based on protein language models and cross-attention mechanism. *Analytical Biochemistry*, 694:115637, 2024a.

Jun Hu, Zhe Li, Bing Rao, Maha A Thafar, and Muhammad Arif. Improving protein-protein interaction prediction using protein language model and protein network features. *Analytical biochemistry*, 693:115550, 2024b.

Youngseung Jeon, Ziwen Li, Thomas Li, JiaSyuan Chang, Morteza Ziyadi, and Xiang Anthony Chen. Ragppi: Retrieval-augmented generation benchmark for protein–protein interactions in drug discovery. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4345–4363, 2026.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. Prollm: protein chain-of-thoughts enhanced llm for protein-protein interaction prediction. *arXiv preprint arXiv:2405.06649*, 2024.

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

Atanas Kamburov, Christoph Wierling, Hans Lehrach, and Ralf Herwig. Consensuspathdb—a database for integrating human functional interaction networks. *Nucleic acids research*, 37 (suppl_1):D623–D628, 2009.

Yan Kang, Yulong Xu, Xinchao Wang, Bin Pu, Xuekun Yang, Yulong Rao, and Jianguo Chen. Hn-ppisp: a hybrid network based on mlp-mixer for protein–protein interaction site prediction. *Briefings in Bioinformatics*, 24(1):bbac480, 2023.

Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, et al. The intact molecular interaction database in 2012. *Nucleic acids research*, 40(D1):D841–D846, 2012.

Xue Li, Peifu Han, Gan Wang, Wenqi Chen, Shuang Wang, and Tao Song. Sdnn-ppi: self-attention with deep neural network effect on protein-protein interaction prediction. *BMC genomics*, 23(1): 474, 2022.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. Prott3: Protein-to-text generation for text-based protein understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5949–5966, 2024.

Birgit HM Meldal, Oscar Forner-Martinez, Maria C Costanzo, Jose Dana, Janos Demeter, Marine Dumousseau, Selina S Dwight, Anna Gaulton, Luana Licata, Anna N Melidoni, et al. The complex portal-an encyclopaedia of macromolecular complexes. *Nucleic acids research*, 43(D1): D479–D484, 2015.

Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

Roberto Mosca, Arnaud Céol, Amelie Stein, Roger Olivella, and Patrick Aloy. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic acids research*, 42(D1): D374–D379, 2014.

Minjie Mou, Ziqi Pan, Zhimeng Zhou, Lingyan Zheng, Hanyu Zhang, Shuiyang Shi, Fengcheng Li, Xiuna Sun, and Feng Zhu. A transformer-based ensemble framework for the prediction of protein–protein interaction sites. *Research*, 6:0240, 2023.

Livia Perfetto, Leonardo Briganti, Alberto Calderone, Andrea Cerquone Perpetuini, Marta Iannuccelli, Francesca Langone, Luana Licata, Milica Marinkovic, Anna Mattioni, Theodora Pavlidou, et al. Signor: a database of causal relationships between biological entities. *Nucleic acids research*, 44 (D1):D548–D554, 2016.

V Srinivasa Rao, K Srinivas, GN Sujini, and GN Sunand Kumar. Protein-protein interaction detection: methods and analysis. *International journal of proteomics*, 2014(1):147648, 2014.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Wesley E Stites. Protein- protein interactions: interface structure, binding thermodynamics, and mutational analysis. *Chemical reviews*, 97(5):1233–1250, 1997.

Joseph Szymborski and Amin Emad. Rapppid: towards generalizable protein interaction prediction with awd-lstm twin networks. *Bioinformatics*, 38(16):3958–3967, 2022.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Sho Tsukiyama, Md Mehedi Hasan, Satoshi Fujii, and Hiroyuki Kurata. Lstm-phv: prediction of human-virus protein–protein interactions by lstm with word2vec. *Briefings in bioinformatics*, 22 (6):bbab228, 2021.

Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, 13(12):966–967, 2016.

Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 2026.

Chuanyan Wu, Bentao Lin, Jialin Zhang, Rui Gao, Rui Song, and Zhi-Ping Liu. Attentionep: Predicting essential proteins via fusion of multiscale features by attention mechanisms. *Computational and Structural Biotechnology Journal*, 23:4315–4323, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Peng Zhou, Pengsen Ma, Jianmin Wang, Xibao Cai, Haitao Huang, Wei Liu, Longyue Wang, Lai Hou Tim, and Xiangxiang Zeng. Large language and protein assistant for protein-protein interactions prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11312–11327, 2025a.

Xiaozhou Zhou, Haoyu Song, and Jingyuan Li. Residue-frustration-based prediction of protein–protein interactions using machine learning. *The Journal of Physical Chemistry B*, 126(8):1719–1727, 2022.

Xibin Zhou, Chenchen Han, Yingqi Zhang, Huan Du, Jiayuan Tian, Jin Su, Renju Liu, Kai Zhuang, Shiyu Jiang, Anthony Gitter, et al. Decoding the molecular language of proteins with evolla. *bioRxiv*, pp. 2025–01, 2025b.

Le Zhuo, Zewen Chi, Minghao Xu, He-Yan Huang, Jianan Zhao, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. Protllm: An interleaved protein-language llm with protein-as-word pre-training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8950–8963, 2024.

APPENDIX

A WHY FREE-TEXT SUPERVISION MATTERS FOR PPI MODELING: EXAMPLE CASES

Despite the extensive body of work proposing classification and regression-based approaches for protein–protein interaction (PPI) modeling, our free-text supervision framework offers a fundamentally different perspective. Traditional methods typically rely on structured outputs, such as binary interaction labels, categorical annotations, or predefined graph triples, which constrain the learning objective to a limited representation space. However, PPI modeling is inherently a generative downstream task. As a result, enforcing structured targets introduces a mismatch between the training objective and the actual output space of the model, potentially limiting expressiveness and fidelity.

SYNJ1 ↔ AMPH (synaptic vesicle endocytosis)

Protein A and Protein B directly bind to one another within the cytoplasm and cytosol to coordinate synaptic vesicle endocytosis and clathrin-mediated membrane trafficking. Protein A is a multi-domain phosphatase belonging to the synaptojanin family that exhibits hydrolase activity toward various phosphoinositides, including phosphatidylinositol 4-phosphate, phosphatidylinositol (4,5)-bisphosphate, and phosphatidylinositol (3,4,5)-trisphosphate. Through the dephosphorylation of these lipids, Protein A regulates membrane organization, synaptic vesicle uncoating, and the rearrangement of actin filaments downstream of tyrosine kinases. Protein B is a cytoplasmic protein containing an N-terminal BAR domain and a C-terminal SH3 domain, which are essential for its roles in regulated exocytosis and the control of the membrane-associated cytoskeleton. . . .
The SH3 domain located at the C-terminal of Protein B directly contacts the disordered C-terminal region of Protein A, which contains multiple N-P-F repeats and proline-rich motifs. This coupling mechanism is central to the recruitment of Protein A to sites of endocytosis, streamlining the recycling of synaptic vesicles. . . .
. . . Together, the direct interaction between Protein A and Protein B provides a critical link between membrane curvature sensing and lipid phosphatase activity during the endocytic cycle.

Figure 5: Example of one-to-many causal effects plus cross-domain bridging. A single enzymatic activity (lipid dephosphorylation) is unfolded into three distinct downstream consequences and the closing sentence bridges two otherwise disjoint semantic domains, membrane geometry sensing and lipid phosphatase chemistry.

Our approach to modeling PPI with free-form description changes the paradigm. Instead of forcing an interaction into a predefined category, a free-form caption can describe the nature, context, mechanism and biological significance of an interaction with richer expressiveness and interpretability. Figures 5, 6, 7, 8, and 9 provide examples highlighting the power of natural language representations.

First, free-form text can capture nuances that categorical labels would ignore, including conditions for the interaction to occur, biological context, structural bias for binding, and relationships to disease and function. The ambiguity and uncertainty can also be effectively expressed in natural language. Standardized labeling in prior work leads to a significant bottleneck in extracting information from literature knowledge bases, whereas free-text synthesis enables more comprehensive and flexible summarization.

Moreover, natural language descriptions are inherently human-readable, allowing for deeper exploration and generative reasoning. A free-text caption can propose mechanisms, suggest hypotheses, and explain why an interaction occurs, which are all beyond the capacity of classification or regression approaches. The flexible schema without predefined ontology categories facilitates the exploration of emerging interactions that were previously undiscovered.

After all, the strength of PPI modeling lies not only in its ability to make predictions, ad also in how faithfully it can capture and express the complexity of the underlying biological reality.

TP53 ↔ MDM2 (ubiquitin-mediated negative-feedback loop)

Protein A and Protein B form a well-characterised transcriptional repressor complex that serves as a fundamental regulatory node in the control of cell cycle arrest, DNA repair, and apoptosis. This interaction is primarily mediated by the binding of Protein A's N-terminal transactivation motif (PF08563) to the N-terminal SWIB/MDM2 domain (PF02201) of Protein B. As an E3 ubiquitin-protein ligase, Protein B facilitates the N6-ubiquitinyl conjugation of Protein A, a critical step that promotes its nuclear export and subsequent degradation by both nuclear and cytoplasmic proteasomes. . . .
Within these contexts, the interaction is subject to complex regulation by various cofactors and post-translational modifications. For instance, phosphorylation of Protein A at its N-terminus by kinases such as CHEK2 or VRK1 can prevent its association with Protein B, thereby stabilizing the protein in response to DNA damage. Conversely, the recruitment of Protein B to Protein A can be enhanced by the presence of cofactors like YY1 or PACT, which increase their mutual affinity and streamline the ubiquitination process. . . .
. . . Small-molecule antagonists, such as Nutlins, have been shown to bind within the N-terminal pocket of Protein B, effectively disrupting its contact with Protein A and leading to the stabilization and activation of the latter. . . . In addition to its role as a ligase, Protein B can directly inhibit the transactivation capacity of Protein A within the TP53-MDM2 transcription regulation complex, ensuring tight negative feedback control over Protein A levels and activity.

Figure 6: Example of multi-regime contrastive reasoning. The same pair of proteins is described under three opposing regimes within one paragraph: CHK2/VRK1 phosphorylation prevents the binding, YY1/PACT cofactor recruitment enhances it, and Nutlin small molecules disrupt it.

HRAS ↔ RAF1 (nucleotide-state-dependent activation of the MAPK cascade)

Protein A directly binds Protein B to form the HRAS-RAF1 complex, a foundational coupling in mitogenic signal transduction that up-regulates Protein B activity. The binding is strictly dependent on the nucleotide-bound state of Protein A, which functions as a small GTPase; it preferentially associates with Protein B when in its active GTP-bound conformation rather than its inactive GDP-bound state. The N-terminal effector region of Protein A contacts the N-terminal Ras-binding domain (RBD) of Protein B. Additional structural evidence indicates that Protein A also contacts a phorbol-ester/diacylglycerol-binding C1 domain in the N-terminal region of Protein B. . . .
The assembly of the HRAS-RAF1 complex facilitates the translocation of Protein B to the plasma membrane, where it is activated through a complex process involving the release of autoinhibition, dimerization, and phosphorylation. . . . This pathway is subject to intricate feedback regulation; for instance, downstream kinases can phosphorylate Protein B at multiple sites in its N-terminal, central, and C-terminal regions to desensitize its activity. . . .
. . . Biochemical evidence further indicates that Protein A can be ubiquitinated by the BCR(LZTR1) E3 ubiquitin ligase complex at its C-terminal region, a modification that decreases its membrane association and inhibits downstream signaling. . . .

Figure 7: Example of state-gated activation cascade. The binding event triggers an ordered chain of physical state transitions: membrane translocation, autoinhibition release, dimerisation, and eventually phosphorylation. Each step is conditional on the previous, and the binding itself is gated by the GTP- vs. GDP-bound state of HRAS.

HSPA8 ↔ BAG1 (allosteric coupling in the Hsp70 nucleotide-exchange cycle)

Protein A and Protein B directly bind to form a heteromeric complex that facilitates nucleotide exchange within the heat shock protein 70 chaperone cycle. This interaction is mediated by the binding of the BAG domain, which forms a three-helix bundle, to the N-terminal nucleotide-binding domain of Protein A. This coupling induces a conformational switch in the ATPase region of Protein A that is incompatible with nucleotide binding, thereby triggering the release of ADP and the subsequent release of substrate proteins from the central substrate-binding domain. Protein B serves as a nucleotide exchange factor that specifically facilitates the conversion of Protein A from an ADP-bound state to an ATP-bound state, a critical step in regulating the chaperone's affinity for its polypeptides.
. . .
Protein B acts as a co-chaperone that regulates the activity of Protein A while also exhibiting independent anti-apoptotic properties. It markedly increases the anti-cell death function of BCL2 in an ATP-dependent manner and inhibits the pro-apoptotic function of PPP1R15A. . . .

Figure 8: Example of allosteric coupling across named conformational states. Binding at one site (the N-terminal nucleotide-binding domain) propagates a conformational change to a physically distant site (the central substrate-binding domain), and the synthesis explicitly names the intermediate states (ADP-bound, ATP-bound, substrate-loaded, substrate-released).

AXIN1 ↔ GSK3β (bistable switch in the Wnt destruction complex)

Protein A and Protein B directly associate as core components of the beta-catenin destruction complex, a multimeric assembly that regulates the canonical Wnt signaling pathway by controlling the stability of cytoplasmic beta-catenin. Within this context, Protein B up-regulates Protein A via phosphorylation, a direct modification that targets the central region of Protein A at specific threonine and serine residues. This phosphorylation event is critical for maintaining Protein A in an active, open conformation, which enhances its ability to bind beta-catenin and facilitates the subsequent phosphorylation of beta-catenin by Protein B. . . .
In the absence of Wnt signaling, the physical association between Protein A and Protein B promotes the hyperphosphorylation of beta-catenin, poising it for ubiquitination and proteasome-mediated degradation. Upon Wnt stimulation, the destruction complex is recruited to the membrane, leading to the inhibition of Protein B and the subsequent nuclear accumulation of beta-catenin. . . .

Figure 9: Example of context-dependent bistable switch. The identical AXIN1↔GSK3β physical association produces opposite cellular outcomes as a function of upstream pathway state: proteasomal destruction of β-catenin in the Wnt-off regime, and nuclear accumulation of β-catenin in the Wnt-on regime.

B Data Construction

B.1 Sources of Raw Evidence

The raw evidence of PPIs are gathered from ten comprehensive sources so that different aspects of the interaction are covered by different specialized datasets, as shown in Table 3 with version and license information.

Table 3: Sources for raw evidence. Feature columns are: Binary PPI (Bin), Complex/multi-subunit (Cpx), signed/causal Effect (Eff), Sequence (Seq), Functional/pathway/domain annotation (Fun), Literature (Lit), 3D Interface contact (Iface). PPI counts are pairs with non-empty data from the source, reported before and after quality-controlled filtering.

Database	Version (retrieval)	License	Bin	Cpx	Eff	Seq	Fun	Lit	Iface	Before	After
IntAct	rel. 2026-01-09[1]	CC BY 4.0	•	•	–	–	–	–	•	1,075,916	351,515
PubMed	live E-utilities (2026-04)[2]	NLM / public	–	–	–	–	–	•	–	1,067,409	350,224
UniProt	rel. 2026_01[3]	CC BY 4.0	–	–	–	•	•	–	–	997,480	345,845
3did	rel. 2024_12[4]	Academic	–	–	–	–	–	–	•	77,200	68,736
Pfam	rel. 37.0[5]	CC0 1.0	–	–	–	–	•	–	–	770,243	264,301
STRING	v 12.0 (2025-07)[6]	CC BY 4.0	•	–	–	–	–	–	–	1,075,916	351,515
SIGNOR	v 4.0 (2026-04)[7]	CC BY-SA 4.0	•	–	•	–	–	–	–	4,737	4,737
Reactome	v 96 (2026-04)[8]	CC BY 4.0	–	–	–	–	•	–	–	783,894	286,882
CORUM	v 5.3 (2026-04)[9]	CC BY 4.0	–	•	–	–	–	–	–	20,060	19,906
ComplexPortal	rel. 2026-01-09[10]	CC0 1.0	–	•	–	–	–	–	–	24,306	24,241

B.2 Evidence Scoring

We assign each pair r an **interaction** component evidence score $E_{\text{int}}(r)$ and a **context** component evidence score $E_{\text{ctx}}(r)$, and the overall scalar evidence score $E(r)$ is the gated aggregation of both:

$$E_{\text{int}}(r) = E_{\text{map}} + E_{\text{mech}} + E_{\text{lit}} + E_{\text{src}}, \qquad E_{\text{ctx}}(r) = \min\Big(\sum_{i \in \mathcal{C}} w_i \, \mathbb{1}[c_i(r)],\ 4.0\Big), \tag{8}$$

$$E(r) = \begin{cases} \min\big(E_{\text{int}}(r) + E_{\text{ctx}}(r),\ 5.0\big) & \text{if } r \text{ has no experimental detection method,} \\ E_{\text{int}}(r) + E_{\text{ctx}}(r) & \text{otherwise.} \end{cases} \tag{9}$$

Per-signal weights and trigger conditions for the four interaction axes and the context axis are listed in Table 4.The scoring schema is designed/validated by human domain experts with proper biological motivations.

[1] https://ftp.ebi.ac.uk/pub/databases/intact/2026-01-09/

[2] https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi

[3] https://ftp.uniprot.org/pub/databases/uniprot/previous_major_releases/release-2026_01/

[4] https://3did.irbbarcelona.org/download/current/3did_flat.gz

[5] https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam37.0/

[6] https://stringdb-downloads.org/download/v12.0/

[7] https://signor.uniroma2.it/releases/getLatestRelease.php

[8] https://reactome.org/download/current/UniProt2Reactome.txt

[9] https://mips.helmholtz-muenchen.de/fastapi-corum/public/file/download_current_file

[10] https://ftp.ebi.ac.uk/pub/databases/intact/complex/2026-01-09/

Table 4: Evidence-score formula composed to comprehensively evaluate the richness of raw annotations in different aspects about the interaction.

Axis	Signal	Trigger	Weight
E_{map} — *interaction type, interface mapping, named-complex grounding*			
	interaction type	"direct" **else** "physical association"	$+2.0\,/\,+1.0$
	binding-region features	one side / both sides	$+1.0\,/\,+2.0$
	3did interface	one side / both sides	$+1.0\,/\,+2.0$
	named-complex reference	non-empty	$+2.0$
	subunit mentions partner	true	$+1.0$
E_{mech} — *mechanistic detail*			
	enzymatic interaction	phospho / ubiq / cleav / acetyl / methyl / ...	$+1.0$
	STRING action modes	non-empty (+ any score ≥ 700)	$+1.0\,(+0.5)$
	biophysical parameters	non-empty	$+1.0$
	stoichiometry	either side annotated	$+0.5$
	biological roles	either side, non-default	$+0.5$
	self-interaction	flag set	$+0.5$
E_{lit} — *literature and experimental redundancy*			
	publications	threshold crossings $\{2,\ 5\}$	$+1.0$ each
	experimental method count	$n_{\text{exp-fam}} \geq 2$	$+1.0$
	IntAct miscore	threshold crossings $\{0.45,\ 0.56,\ 0.65\}$	$+0.5$ each
	evidence lines	≥ 5	$+0.5$
	abstract length (chars)	threshold crossings $\{1{,}500,\ 3{,}000,\ 5{,}000\}$	$+1.0$ each
	interaction annotations	≥ 250 chars	$+0.5$
	experimental detection	non-computational (+ not spoke-expanded)	$+1.0\,(+0.5)$
	shared pathways	threshold crossings $\{1,\ 3\}$	$+0.5$ each
	STRING combined score	threshold crossings $\{400,\ 700\}$	$+0.25$ each
E_{src} — *curator-graded complex / mechanism sources*			
	SIGNOR	non-empty (+ any entry tagged direct)	$+2.0\,(+1.0)$
	CORUM	non-empty	$+2.0$
	Complex Portal	non-empty	$+2.0$
	CORUM $\wedge$ Complex Portal	cross-validated	$+1.0$
E_{ctx} — *contextual biological coherence (capped at* 4.0*)*			
	paired UniProt fields	both sides annotated: function, domains, similarity...	$+0.25$ each
	subcellular location	shared term	$+0.5$
	shared GO term	component / process / function	$+0.25$ each
	either-side annotation	tissue, catalytic activity, PTM, modified residues, regulation, active / binding sites, motifs, Zn fingers, free-text, ...	$+0.25$ each
	disease keywords	either side (+ both sides)	$+0.5\,(+0.25)$

B.3 K-MEANS CLUSTERING

We treat the per-pair evidence score as a 1D feature and run K-means clustering on all 1.08M pairs we gathered from sources of raw evidence. As shown in Figure 10, different numbers of clusters are tested with both within-cluster sum of squares (inertia) and silhouette scores (probed with 200k subset) computed. The elbow with k=3 clusters is chosen from both curves with the kneedle criterion. The split criteria is consistent with our manual inspection, as most PPIs from T1 are mostly supported by one single high-throughput experiment and has no solid experimental evidence on its interaction details, as well as most PPIs from T3 are well supported by multiple single-sample experiments with strong evidence on interaction mechanism and biological effects.

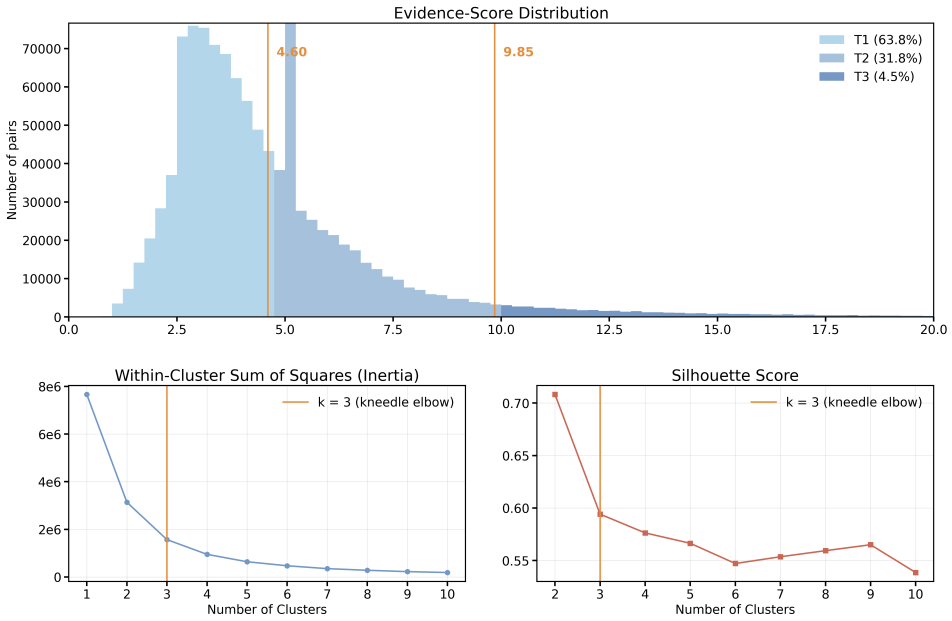

Figure 10: K-means clustering of the per-pair evidence score. Top: score distribution colored by three different clusters; bottom: inertia and silhouette scores for different numbers of clusters as well as the elbow k=3.

B.4 Manifold Coverage and Filtering Bias

With the help of K-means clustering, the dataset is filtered using an evidence-tiered quality control procedure. We analyze how this filtering affects diversity and coverage in the PPI description dataset. Specifically, we apply Principal Component Analysis (PCA) to project ESM3 protein embeddings to a lower-dimensional space. Figure 11 compares the manifold coverage of the filtered dataset against the original raw samples in both PPI space and protein space.

In the PPI space (left) where each data point represents a pair of proteins interacting, the filtered dataset closely overlaps with the high-density regions of the original distribution, indicating that the dominant interaction modes are well preserved. While a small number of peripheral clusters and low-density regions are reduced, the overall geometric structure of the manifold remains intact.

And in the protein space (right), we observe a similar trend. The filtered dataset retains nearly all high-density regions, with only sparse outliers and boundary regions being pruned. Notably, the filtering appears to preferentially remove isolated or weakly supported samples rather than systematically biasing the representation toward specific regions of the space.

We further examine whether the filtering procedure introduces bias at the semantic level by analyzing retention rates across annotation keywords. Figure 12 plots retention bias as a function of keyword frequency. The majority of keywords remain balanced, with retention rated concentrated near zero across the full frequency spectrum, indicating that the filtering process does not systematically favor or suppress most functional annotations.

A small fraction of keywords (4%) are enriched, primarily corresponding to well-supported or frequently studies biological contexts, while 9% are depleted, often associated with more specific or sparsely represented mechanisms especially for virus. Notably, the depletion mostly happens to lower-frequency keywords, suggesting that the filtering preferentially removes under-supported or noisier annotations.

Overall, these results indicate that the evidence-tiered filtering preserves the global semantic distribution of the dataset while selectively pruning less reliable or weakly supported annotations, without introducing substantial bias.

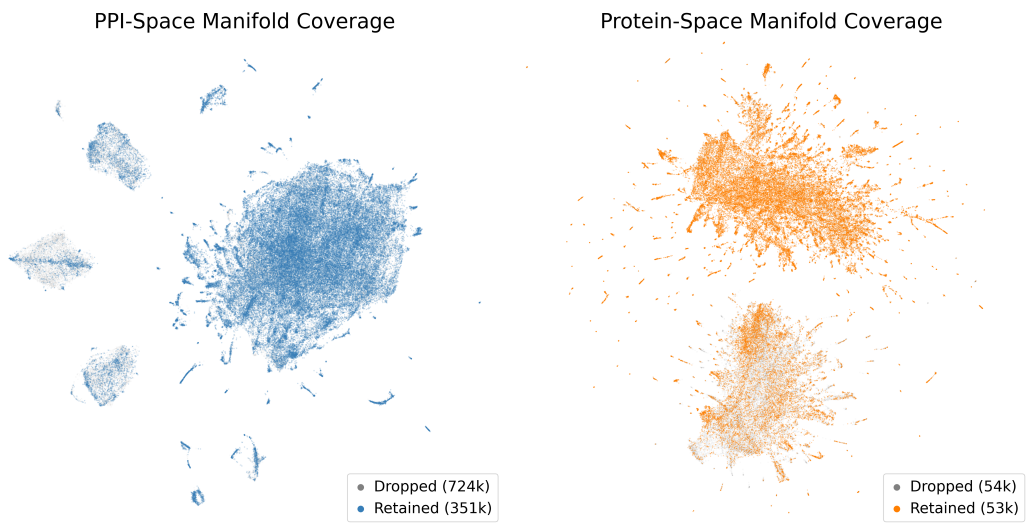

Figure 11: UMAP (Uniform manifold approximation and projection) visualization of manifold coverage of the filtered dataset against the original raw samples in both PPI space and protein space. The most of the original coverage is preserved.

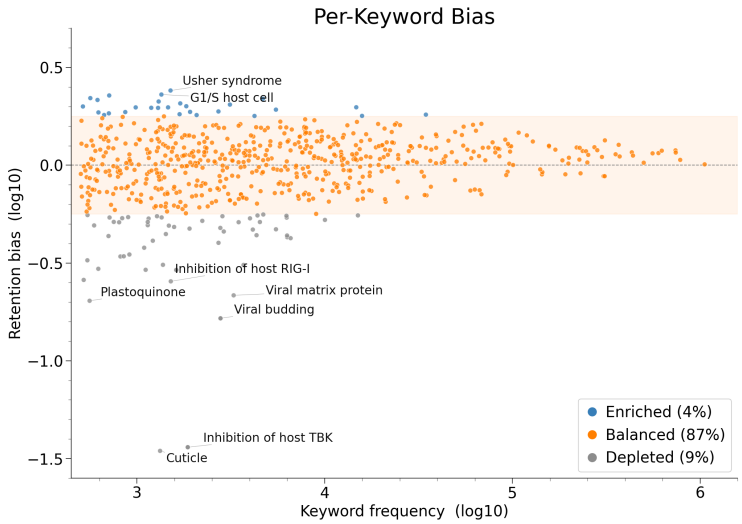

Figure 12: Retention bias as a function of keyword frequency. Most high-frequent keywords in the raw dataset are preserved in the filtering.

B.5 Evidence-tiered Prompting

For the augment of the PPIs with an LLM process, we introduce a tier-controlled generation framework that explicitly constrains the output space according to the strength of biological support. Four orthogonal constraints are explicitly enforced in each prompt: *descriptive granularity*, *epistemic strength*, *mechanistic attribution*, and *silence policy* further described as follows:

- First, we regulate **descriptive granularity** through tier-specific length constraints, so that low-confidence interactions are restricted to concise descriptions while high-confidence PPIs are allowed with more detailed interpretation.
- Second, we control the **epistemic strength** of language via tier-conditioned verb selections. Interactions with strong mechanistic support are instructed to use assertive biological terms,

whereas weakly supported interactions are required to use hedged expressions to reflect observational evidence.

- Third, we regulate the **mechanistic attribution** by conditioning generation on the availability of structural or curated mechanistic annotations. When such evidence is absent, the teacher model is explicitly constrained from introducing molecular-level mechanistic explanations to prevent unsupported speculation.
- Fourth, we enforce a **silence policy** on under-annotated entities to restrain over-description of proteins beyond available evidence, avoiding implicit hallucination through prior-driven completion.

We show some statistics related to the dataset components after the augmentations to each of the constraints in Figure 13.

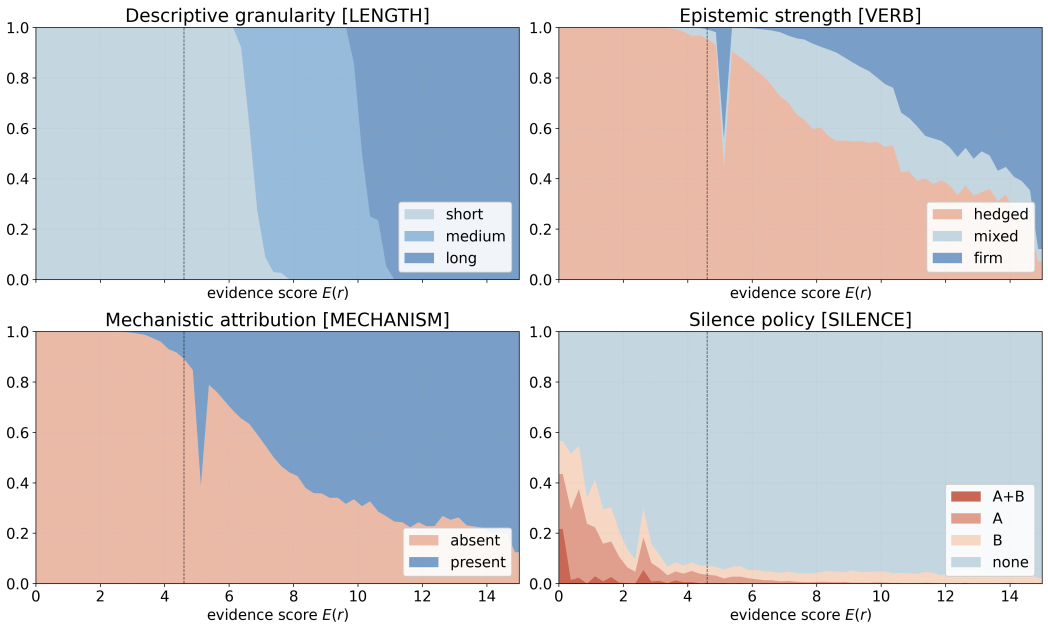

Figure 13: Statistics of the augmented PPI dataset samples with the proposed evidence-tiered constraints. A substantial variability can be observed across PPIs, making tier-controlled dynamic prompting necessary for synthesis.

C ESM3 AS A SINGLE-PROTEIN ENCODER

The main paper assumes a sequence-only setting, where PPI2Text requires only amino acid sequences as the only input. However, ESM3 is not inherently a sequence-only encoder, but rather a generative multimodal model that expects multiple per-residue input tracks, including structural information. This mismatch by design is addressed by reconstructing all required ESM3 inputs from the protein sequence alone using external predictors, and then perform a frozen forward pass to obtain per-residue embeddings.

For a protein of length L, we assemble the full set of per-residue inputs expected by ESM3. All inputs are ultimately derived from the amino acid sequence, ensuring consistency with the sequence-only assumption.

- **Sequence tokens:** The amino acid sequence is retrieved from UniProt and tokenized using the ESM3 vocabulary, including BOS and EOS tokens, resulting in a sequence of length $L + 2$.

- **Backbone coordinates:** We obtain the sequence-based predicted 3D structure from AlphaFold2. From this structure, we extract the backbone atom coordinates (N, C_α, C) for each residue, which are required by ESM3's geometric attention module.
- **Discrete structure tokens:** The backbone coordinates are passed through the frozen ESM3 structure encoder (a Vector Quantized Variational Autoencoder (VQ-VAE) released with ESM3) to produce one discrete structure token per residue.
- **Secondary structure (SS8):** We compute 8-class secondary structure assignments using the mkdssp library applied to the AlphaFold structure. To align with the ESM3 token set, DSSP symbols − (coil) and `P` (polyproline-II) are both mapped to `C` (cysteine).
- **Solvent-accessible surface area (SASA):** Residue-level SASA values are computed using the Shrake–Rupley implementation in the BioPython library, and then discretized using the ESM3 SASA tokenizer.
- **Confidence scores (pLDDT):** Per-residue confidence scores are read from the AlphaFold PDB B-factor column and normalized to $[0, 1]$ by dividing by 100. We also compute the mean pLDDT across residues as a global confidence estimate.
- **Function and residue annotation tracks:** ESM3 expects additional tracks corresponding to functional annotations (e.g., InterPro) and free-form residue annotations. To avoid potential information leakage affecting the evaluation process, both fields are not provided but filled with their respective padding tokens instead.

Although several inputs depend on structural information, all such features are derived from auxiliary predictions, which are generated from sequence. Thus, the overall pipeline adheres to the sequence-only constraint.

The assembled inputs are fed into ESM3 using a single chain identifier. The core computation of ESM3 are 48 transformer layers stacked with geometric attention. We retain only the final hidden representation corresponding to the post-LayerNorm activations of the last transformer block. All other outputs, including decoder heads and structure/function predictions, are all discarded.

D PaCo-RoPE

D.1 Formulation

We formalize PaCo-RoPE as an extension of rotary positional encoding (RoPE) to 3D position indices with channel interleaving to encode single protein representations and the interaction pair map simultane.

For a query/key vector $x \in \mathbb{R}^{d_h}$, a standard 1D RoPE applies a rotation to each frequency pair $(2i, 2i+1)$:

$$\mathrm{RoPE}(x,p)_{\{2i,2i+1\}} = \begin{pmatrix} \cos\theta_i(p) & -\sin\theta_i(p) \\ \sin\theta_i(p) & \cos\theta_i(p) \end{pmatrix} x_{\{2i,2i+1\}}, \quad \theta_i(p) = \omega_i\, p, \tag{10}$$

where $\omega_i = \omega_0^{2i/d_h}$ is the frequency.

Each token is assigned a triplet $\mathbf{p} = (p^T, p^\theta, p^\varphi) \in \mathbb{R}^3$ depending on its modality and semantic role:

- **Text tokens.** For a token at sequence position n,
$$\mathbf{p} = (n, n, n). \tag{11}$$
- **Single-protein tokens (protein A).** For a token at sequence position n, representing residues centered at index i in protein A,
$$\mathbf{p} = \big(n, \lfloor i/s \rfloor, 0\big). \tag{12}$$
- **Single-protein tokens (protein B).** For a token at sequence position n, representing residues centered at index j in protein B,
$$\mathbf{p} = \big(n, 0, \lfloor j/s \rfloor\big). \tag{13}$$

- **Pair-map tokens.** Let a pair-map be partitioned into a grid of size $H \times W$. For a patch at grid location (m, n) (zero-indexed), serialized at sequence position k, we define

$$\mathbf{p} = \left(k,\ \left(m + \tfrac{1}{2}\right) \frac{L'_A}{H},\ \left(n + \tfrac{1}{2}\right) \frac{L'_B}{W} \right), \tag{14}$$

  where $L'_A = \lfloor \frac{L_A}{4} \rfloor$ and $L'_B = \lfloor \frac{L_B}{4} \rfloor$ denote the effective lengths of proteins A and B after 1D compression, and the $+\frac{1}{2}$ term centers each patch within its spatial bin.

Here s is the stride used to compress residue indices to token indices. All coordinates are real-valued, enabling seamless integration of discretized sequence tokens and pooled pair-map patches.

Let $i \in \{0, \ldots, d_h/2 - 1\}$ index frequency pairs. We define a channel selector

$$c(i) = \begin{cases} T, & i \bmod 3 = 0, \\ \theta, & i \bmod 3 = 1, \\ \varphi, & i \bmod 3 = 2. \end{cases} \tag{15}$$

Each frequency pair is thus assigned to exactly one positional channel. Unlike the block partitioning M-RoPE, the interleaving scheme keeps any single channel from dominating a contiguous frequency band.

So that for each pair $(2i, 2i+1)$, we apply PaCo-RoPE using the selected coordinate:

$$\theta_i(\mathbf{p}) = \omega_i \cdot p^{\,c(i)}, \tag{16}$$

$$\mathrm{PaCoRoPE}(x, \mathbf{p})_{\{2i,2i+1\}} = \begin{pmatrix} \cos\theta_i(\mathbf{p}) & -\sin\theta_i(\mathbf{p}) \\ \sin\theta_i(\mathbf{p}) & \cos\theta_i(\mathbf{p}) \end{pmatrix} x_{\{2i,2i+1\}}. \tag{17}$$

D.2 Interpretation

Consider protein A as an example. Let protein A consist of L'_A effective residue-level tokens after tokenization, indexed by

$$r \in \{0, \ldots, L'_A - 1\}. \tag{18}$$

Protein-A tokens are mapped to positional coordinates via

$$p^{\theta}(r) = \left\lfloor \frac{r}{s} \right\rfloor, \tag{19}$$

where s is the token stride.

The pair-map is constructed by partitioning the same residue axis into H bins using average pooling. Define the binning function

$$\pi_H(r) = \left\lfloor \frac{r}{L'_A/H} \right\rfloor. \tag{20}$$

Each pair-map row index $m \in \{0, \ldots, H-1\}$ corresponds to the residue interval

$$r \in \left[m \cdot \frac{L'_A}{H}, (m+1) \cdot \frac{L'_A}{H} \right), \tag{21}$$

and is assigned the centered coordinate

$$p^{\theta}_{\text{pair}}(m) = \left(m + \tfrac{1}{2}\right) \frac{L'_A}{H}. \tag{22}$$

For any residue r such that $m = \pi_H(r)$, we have

$$\left| r - p^{\theta}_{\text{pair}}(m) \right| \le \frac{L'_A}{2H}. \tag{23}$$

Thus, protein-A tokens and pair-map tokens share a common discretization of the same underlying residue axis, up to bounded quantization error $\mathcal{O}(L'_A/H)$.

An identical construction applies to protein B, yielding the same alignment property along the second spatial axis:

$$\left| r - p^{\varphi}_{\text{pair}}(n) \right| \leq \frac{L'_B}{2W}. \tag{24}$$

Therefore, protein-B tokens and pair-map tokens share the same partitioning of the residue axis up to bounded quantization error $\mathcal{O}(L'_B/W)$.

Thus, PaCo-RoPE does not impose exact equality between token and pair-map coordinates. Instead, both representations are derived from a shared discretization of the same underlying residue axis, ensuring consistent spatial correspondence under bounded resolution error induced by pooling and tokenization.

E Hyperparameters and Computational Resources

We use `esm3-sm-open-v1` as the single protein encoder with 1.4B parameters and 1536 as the hidden dimension. Due to the limitation of the encoder, PPIs with proteins no longer than 2048 amino-acids are kept. Both stride-2 Conv1D layers use kernel of size 4, resulting in a four times length reduction with the pipeline hidden dimension of 1024. The compression factor of 4 corresponds to the average size of $alpha$-helical turns, and the receptive field of 10 covers the natural length scale of local secondary-structure elements. The adaptive mean pooling targets a 32×32 grid, resulting in 1024 pair map tokens to the decoder, as a trade-off between information granularity and computation efficiency. We choose `Qwen3-4B-Instruct` as the frozen base decoder with excellent generic knowledge in biology. LoRA adapter of rank 32 and 0.1 dropout rate is applied to provide enough flexibility and room for the new knowledge to be injected through supervised fine-tuning (SFT). The end-to-end SFT takes 200 GPU hours on NVIDIA H100 80GB for 5 epochs with effective batch size of 64. The peak learning rate for LoRA components is 1e-4, and 5e-5 for modules initialized from scratch, both managed by warm-up and cosine scheduling.

F Dataset Splitting Protocol and Experiments

Temporal holdout: To evaluate performance of our model in a realistic prospective setting, we introduce a temporal split at May 1, 2025, so that all PPIs first annotated after this date form the holdout set, simulating prediction on newly discovered interactions. To prevent information leakage from homologous proteins, we perform sequence-based decontamination using MMSeq2, defining two proteins A and B as similar ($A \sim B$) if they share more than 50% sequence identity over at least 80% coverage. For the holdout setting, the decontamination is imposed at the pair level: a training PPI (A, B) is removed if there exists a test PPI (C, D) that:

$$\Big((A \sim C) \cap (B \sim D)\Big) \cup \Big((A \sim D) \cap (B \sim C)\Big) \tag{25}$$

C3-hard: Following the idea of Bernett (Bernett et al., 2024), we construct a more stringent C3-hard test set to evaluate out-of-distribution generalization. The leakage criterion is defined at the single-protein level, such that a training interaction (A, B) is excluded if there exists a test interaction (C, D) satisfying:

$$(A \sim C) \cup (A \sim D) \cup (B \sim C) \cup (B \sim D) \tag{26}$$

This ensures that no homologous proteins are shared between the training and test sets, effectively removing any opportunity for the model to exploit sequence similarity or evolutionary relatedness. As a result, C3-hard represents a highly controlled and intentionally stringent evaluation setting that stresses compositional generalization beyond realistic application scenarios.

While this setting is valuable as a worst-case robustness test, it is more extreme than typical pharmaceutical and wet-lab discovery conditions, where newly studied proteins often retain varying degrees of sequence or functional similarity to previously characterized ones and such biological priors are actively leveraged in practice. Consequently, C3-hard should be interpreted as an upper-bound stress

test rather than a direct proxy for standard real-world use cases. In practice, this setting is expected to significantly increase task difficulty for all methods, while more closely reflecting worst-case extrapolation behavior compared to more realistic in-distribution or temporally shifted evaluation settings.

In addition to Figure 4, Table 5 provides a more detailed comparison of the baseline model and ablation variants on the proposed C3-hard split. The results follow the same overall trend observed in Table 1. Specifically, removing the encoder leads to the weakest performance across all settings, with a particular degradation in the raw evidence-based metrics (see Appendix G for further details on these metrics). In contrast, the best performance is consistently achieved by configurations that incorporate all three embedding components, with the full PPI2Text model obtaining the highest scores across the majority of evaluation metrics.

Table 5: Baseline and ablation results on C3-hard split. Lexical metrics includes BLEU-2/4 F1 scores (B-2/4), ROUGE-1/2/L F1 scores (R-1/2/L), with semantic BERTScore using RoBERTa (RBT) and BioBERT (BBT) embeddings. LLM-as-a-judge scores against the raw evidence cover Entities (Ent), Interaction (Int), Mechanism (Mec), and the Average (Avg).

Model	(against Synthesized Text)							(against Raw Evidence)			
	B-2	B-4	R-1	R-2	R-L	RBT	BBT	Ent	Int	Mec	Avg
Seq+Qwen3	34.16	17.82	51.16	20.11	26.60	87.11	80.26	1.45	4.71	0.38	2.18
MINT+Qwen3	34.87	18.47	51.71	22.23	27.53	87.32	81.10	2.53	5.68	1.50	3.24
SingProt-Only PPI2Text	34.87	18.74	52.81	23.11	28.58	87.72	81.52	3.01	5.99	1.24	3.41
PairMap-Only PPI2Text	35.35	19.26	53.26	23.98	28.81	87.93	81.73	4.13	6.49	2.25	4.29
1D-RoPE PPI2Text	34.99	19.00	52.81	23.61	28.69	87.72	81.10	2.42	6.18	2.63	3.74
No-Cross PPI2Text	35.35	18.87	52.48	23.23	28.23	87.52	80.89	4.02	7.27	4.58	5.29
PPI2Text	**37.26**	**20.57**	**55.02**	**25.48**	**29.74**	**88.23**	**82.57**	**5.73**	**7.88**	**5.33**	**6.31**

G Evaluation Against Raw Evidence: LLM-as-a-Judge

Linguistic metrics evaluated on generated text against the synthesized reference text are well established and widely accepted evaluation protocol, but the augmented nature of the reference text adds a layer of uncertainty. Despite careful examination by human expert biologists on the quality of the synthesized reference text, we cannot rule out edge cases where minor details are misinterpreted or missing. In this case, we propose to evaluate the prediction against the aggregated raw evidence for each PPI, since the annotation from human-curated sources are well accepted as gold-standard ground-truth.

To enable large-scale evaluation of generated predictions, relying solely on biochemical experts would be time consuming, as validating thousands of outputs requires substantial manual effort. To address this limitation, we propose the use of a large language model (Claude-Opus-4.7) as an automated judge, enabling a faster and more scalable evaluation of our pipeline. The prompt instructs the LLM to act as an expert biochemist and systematically compare the predicted interaction description against the evidence card along three orthogonal evaluation axes.

First, entity grounding measures whether the prediction correctly identifies proteins, families, structural features, organism, and cellular context without introducing unsupported or fabricated details. This dimension is critical because errors at the level of entity identity or biological context fundamentally undermine the validity of any downstream interpretation.

Second, interaction topology evaluates whether the predicted relationship type (e.g., direct binding, complex co-membership, or co-occurrence) matches the level of evidence supported by curated databases such as UniProt and STRING (Appendix B.1), penalizing both over- and under-interpretation. Accurately capturing interaction topology ensures that the model's predictions respect the evidence strength and biological meaning encoded in source databases.

Third, mechanism fidelity assesses whether mechanistic details, such as directionality, regulatory effects, or domain-level interfaces, are accurately preserved when present in the underlying evidence. Finally, the framework accounts for cases where mechanistic information is absent, ensuring that

models are not rewarded for fabricating unsupported mechanistic claims. Incorrect mechanistic predictions underlies its utility for explaining biological processes and guiding experimental design. Distortion of such details reduces rich biochemical knowledge to vague associations, while incorrect mechanisms can lead to misleading hypotheses about regulation or function.

Each axis is scored independently on a 0–10 scale, enabling a fine-grained decomposition of model performance across factual correctness, interaction classification, and mechanistic reasoning. In Figure 14 we provide the curated evaluation prompt.

> You are an expert biochemist and protein-protein interaction reviewer. You are given (1) a compact RAW EVIDENCE CARD describing what is known about a protein-protein interaction from primary databases (UniProt, IntAct, STRING, Reactome, 3DID, SIGNOR, CORUM, ComplexPortal, PubMed abstracts) and (2) a model PREDICTION — a free-text description of that same interaction.
>
> Your job is to score the PREDICTION against the RAW EVIDENCE on THREE ORTHOGONAL AXES, each on an integer scale 0-10:
>
> **1. ENTITY GROUNDING —** Are the proteins, families, Pfam/structural folds, organism, subcellular localization, and individual protein functions described in the prediction consistent with the evidence card? Penalize fabricated families/folds, wrong organism, confidently wrong substrate names, invented cell-line context not in the abstracts.
>
> **2. INTERACTION TOPOLOGY —** Does the prediction correctly classify the *type* of association? Categories: self/homo-oligomer; direct binary binding; complex co-membership (named complex); database-level co-occurrence (single-report proteomic survey, no defined coupling). Award high scores for matching the truth's category and naming the right complex (when one exists). Penalize topology inflation (calling a DB-level co-occurrence a direct binding) or topology deflation (calling a known complex member just co-localized).
>
> **3. MECHANISM FIDELITY —** If the evidence carries mechanistic content (interface_summary, causal_effect, named complex membership requirement, mech_detail = present, 3DID Pfam pair), does the prediction preserve it: directional action (A acts on B), regulatory consequence, interface region/domain pair? Award 0-3 for flattening rich mechanism into "uncharacterized"; 4-6 for partial preservation; 7-10 for faithful capture. Use "NA" if the truth has no mechanism (mech_detail = absent and no interface_summary and no causal_effect) AND the prediction also avoids inventing one. If the truth has no mechanism but the prediction fabricates one, score LOW (don't use NA).
>
> Output in a JSON with three scores plus a one-sentence note (about 25 words).

Figure 14: Structured prompt that instructs the LLM to act as an expert biochemist and compare free-text interaction predictions against aggregated raw evidence.

H SAFEGUARDS

PPI2Text demonstrates strong potential for protein–protein interaction description, though we cannot rule out the possibility of occasional inaccuracies, so results should be interpreted with care and verified where possible. While the model is designed for beneficial research applications, we encourage its use in responsible and well-regulated settings to ensure alignment with ethical and safety standards.

I GEMINI SYNTHESIS SYSTEM PROMPT

You are rewriting noisy, multi-source metadata about a protein-protein interaction into a single coherent scientific paragraph. The user prompt is organised in two halves:

(1) DECISION BLOCKS — short lines in the form ‘[TAG] value‘ at the top of the user prompt. They encode pre-computed decisions that have already taken account of the evidence. Treat them as binding constraints. Do not re-derive, override, or question them.

(2) EVIDENCE BLOCKS — labelled multi-source raw evidence about the interaction. These are the factual substrate of the paragraph.

DECISION BLOCK REFERENCE

[Length] e.g. "240-350 words, approximately 9-13 sentences" — the paragraph MUST land inside this range. Stop when the evidence has been covered; do not pad to hit the maximum, and do not truncate to stay under the minimum. If you finish with more to say than the target allows, prioritize interface / mechanism / complex context over single-protein housekeeping.

[Verb] firm | mixed | hedged — governs the verbs you may use.
- firm: "directly binds", "the interface lies between", "directly contacts". No hedges in the main claim.
- mixed: "binds", "contacts", "physically associates". Mild hedging allowed for specific sub-claims only.
- hedged : "reported to co-purify with", "associates with", "appears to co-occur". Never use firm verbs.

[Mechanism] present | absent — when absent, describe each protein individually and label the link as indirect, computational, or database-level. Do NOT invent a specific coupling mechanism in this case (no "B recruits A to . . . ", no "their interaction streamlines X").

[Silence] none | A | B | A+B — if a side is listed, do not assign that protein any functional descriptor. You may state its structural features (domains, localisation) plainly, but not "functions as", "serves as", "acts as", "plays a role in". Its family name, UniProt keywords, and protein-class descriptor are also forbidden as functional glosses for a silenced protein.

OUTPUT

One single paragraph of third-person scientific prose. No bullets, headers, markdown, preamble, trailing commentary, or quotation. The opening sentence must reflect whatever is most informative about THIS specific pair — the interface, the named complex, the asymmetry in function between the two proteins, the mechanism of the coupling, the strength or absence of evidence, or the location pattern.

BINDING RULES

1. Factual accuracy. Every protein-specific claim (partners, diseases, PTMs, activities, domains, tissues, clinical relevance) must come directly from an evidence block. You may add generic biochemical reasoning, but do not invent protein-specific details. If the evidence is silent, omit it.

2. Naming and numbers. Do not name any specific experimental assay, technique, or acronym. This applies to anything listed under "Detection methods" or mentioned by name in an abstract (e.g., affinity-pulldowns, hybrid systems, mass-spectrometric surveys, array platforms, structural methods, complementation assays, proximity-labelling, surface or kinetic assays, kit or platform names). Use only generic terms ("biochemical evidence", "proteomic survey", "binding assay", "structural analysis", "high-throughput screen", "interactome data") or omit methods.

3. Speculative-mechanism ban. Do not introduce regulation, competition, displacement, rescue, recycling, sequestration, or compartment-switching for this pair unless explicitly supported by an evidence block. You may describe each protein’s regulation, but not combine them into a specific mechanism without evidence.

4. PTM specificity. Use PTM terms exactly as given in the evidence ("ubiquitinated", "neddylated", "SUMOylated"). Do not convert the generic "Ubl conjugation" into a specific modifier unless the evidence specifies it.

5. Topology vocabulary. Use the provided qualitative position terms ("N-terminal", "central", "C-terminal", etc.) exactly as written. Do not use residue numbers or invent positions.

6. Naming of the two proteins. Refer to them only as "Protein A" and "Protein B". Do not include real names, gene symbols, accessions, or species synonyms.

7. Stop when the evidence has been covered. Do not add closing remarks about unknowns, future work, or unproven consequences. End with the last supported claim.

Return ONLY the paragraph.

Figure 15: System prompt used to synthesize the free-text descriptions. A decision block reference helps the model to interpret evidence-tiered dynamic controlling labels.