SUPPLEMENTARY MATERIAL

A. Experimental Settings and Computational Analysis

1) Implementation Details: Experiments are conducted using an Intel (R) Core (TM) i5-10400F CPU (2.90 GHz) and an NVIDIA GeForce RTX 3080 GPU with 10 GB memory. The implementation is based on Python 3.7.1 and PyTorch 1.7.1. We set the hidden layer size of both the encoder and the projection head to 512. For the two views, the edge addition, edge dropout, and feature masking rates are set to 0.2, 0.3, and 0.1 for the first view and 0.3, 0.4, and 0.0 for the second view, respectively. The temperature parameter τ is set to 0.7. Model training is performed over 1000 epochs using the RReLU activation function and the Adam optimizer, with a learning rate of 0.0001. The range of hyperparameters J and J' is set to [0, 2] and [0, 4], respectively. For the benchmark methods, we adopt the optimal parameter settings reported in their respective publications.

2) Computational Efficiency Analysis: To assess deployment feasibility, we evaluate inference latency (ms/protein) and peak GPU memory consumption for KEGCL, STRPCI [1], and AdaPPI [2] on the BioGRID [3] dataset using a single NVIDIA RTX 3080 GPU (CUDA 11.1).

TABLE S1: Inference latency and peak GPU memory for KEGCL, AdaPPI, and STRPCI

Method	Latency (ms/protein)	Peak Memory (MB)
KEGCL	4.595	353.52
AdaPPI	10.056	909.13
STRPCI	13.561	15126.41

KEGCL achieves the lowest inference latency, being 2.19× faster than AdaPPI and 2.95× faster than STRPCI. Its peak GPU memory consumption (354 MB) is approximately 2.57× lower than AdaPPI and 42.79× lower than STRPCI. These results demonstrate KEGCL's superior computational efficiency and practical deployability in resource-constrained environments, including clinical laboratories and portable analysis platforms, while maintaining competitive predictive performance.

B. Comparative Analysis of Similarity Metrics

To evaluate the impact of different similarity measures in the contrastive learning framework, we compare cosine similarity and Mahalanobis distance within KEGCL.

Cosine Similarity. The original InfoNCE loss in KEGCL is:

$$\mathcal{L}_{\cos} = -\frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \log \frac{\exp(sim_{\cos}(Z_1^i, Z_2^j)/\tau)}{\sum_{o\in\mathcal{N}}\exp(sim_{\cos}(Z_1^i, Z_o)/\tau)} \tag{S1}$$

where $sim_{\cos}(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}^\top\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}$, τ is a temperature parameter. The set $\mathcal{P}$ denotes positive protein pairs (i.e., interacting proteins across augmented PPI views), and $\mathcal{N}$ denotes negative protein pairs, which include non-interacting proteins within the same augmented view as well as across different views. Cosine similarity emphasizes directional consistency and maintains scale invariance. Such characteristics make it particularly suitable for protein interaction networks, where functional similarity is often reflected by the orientation of embedding vectors rather than their magnitude. This methodological rationale supports its adoption as a reliable metric for enhancing the accuracy of functional relationship modeling and the robustness of protein complex identification.

Mahalanobis Distance Variant. To account for correlated embedding dimensions, we employ a Mahalanobis distance-based similarity [4]:

$$sim_{\text{mah}}(\mathbf{u},\mathbf{v}) = -\sqrt{(\mathbf{u}-\mathbf{v})^\top\Sigma^{-1}(\mathbf{u}-\mathbf{v})} \tag{S2}$$

where Σ is the covariance matrix of the batch embeddings. The InfoNCE loss is then obtained by replacing $sim_{\cos}$ in Eq. (S1) with sim_{mah}. This formulation captures anisotropic variance and finer-grained functional relationships, but covariance estimation and inversion can be unstable in small mini-batches and introduce gradient variance. Therefore, while the Mahalanobis distance provides a theoretically grounded alternative, its practical effectiveness is sensitive to the stability of covariance estimation.

Experimental Results. We evaluate four metrics on the BioGRID dataset. Table S2 summarizes the results. Fig. S1 presents the training loss curves for both similarity measures.

TABLE S2: Performance Comparison of Similarity Metrics in KEGCL on the BioGRID Dataset

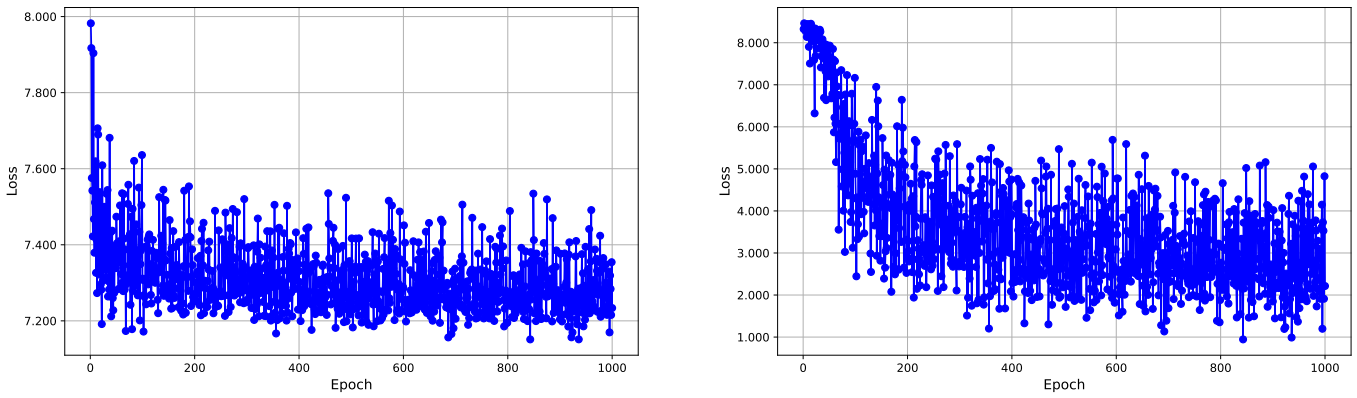
Similarity Metric	Precision	Recall	F1	ACC
Cosine Similarity	**0.527**	0.739	**0.615**	**0.319**
Mahalanobis Distance	0.516	**0.747**	0.610	0.292

The experimental results indicate that cosine similarity provides slightly higher predictive accuracy and more stable convergence. In contrast, the Mahalanobis distance accounts for the covariance structure of embedding dimensions but introduces greater training variance.

C. In-depth Functional Enrichment Analysis of CLPC1

To further elucidate the functional roles, structural organization, and biological context of the identified complex CLPC1, we performed a detailed GO enrichment analysis of this complex, with results visualized in Fig. S2. Specifically, the top 20 GO terms were selected based on the smallest adjusted p-values, highlighting the most statistically significant biological functions. The analysis, presented through a circular plot and a bar plot, reveals the significant enrichment of this complex in RNA splicing and metabolism.

The circular plot provides a multi-layered analysis of GO enrichment. Specifically, the outermost circle marks the enriched categories and a gene count scale, representing the broad background gene distribution. Meanwhile, the second circle, shaded in deep red (p-value $< 1 \times 10^{-10}$), highlights highly enriched processes within the differential gene set. In addition, the third circle, dominated by downregulated genes (light purple, e.g., a significant proportion for GO:0000398), suggests reduced splicing activity under specific cellular conditions. These conditions, such as stress or disease states, may lead



(a) Cosine similarity (b) Mahalanobis distance

Fig. S1: Training loss curves of KEGCL using (a) cosine similarity and (b) Mahalanobis distance, illustrating the optimization convergence behavior under different similarity metrics.

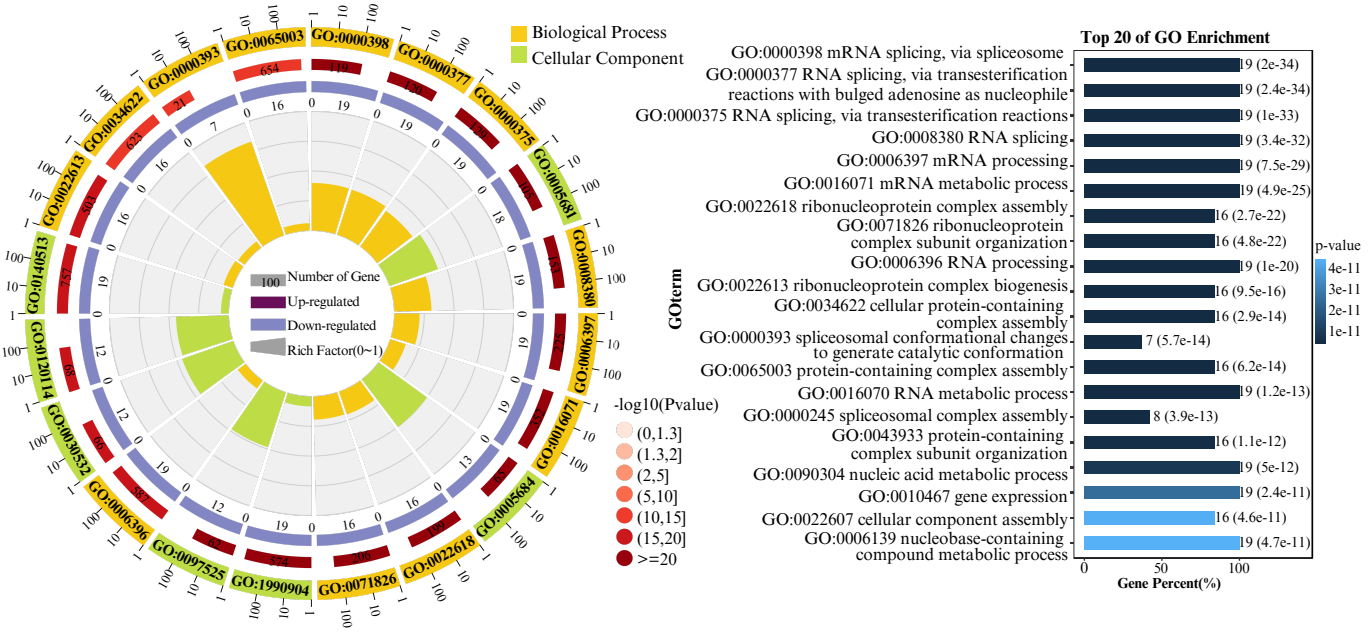


Fig. S2: Top 20 GO terms identified in the enrichment analysis of the CLPC1 complex, selected based on the smallest p-value.

TABLE S3: KEGG Pathway Enrichment Analysis for the Unmatched Protein Complex CLPC1

Pathway	Candidate Genes with Pathway Annotation (14)	All Genes with Pathway Annotation (2,124)	p-value	Pathway ID
Spliceosome	14 (100%)	75 (3.53%)	1.341e-21	ko03040
mRNA surveillance pathway	1 (7.14%)	50 (2.35%)	2.843e-1	ko03015
Nucleocytoplasmic transport	1 (7.14%)	65 (3.06%)	3.537e-1	ko03013
RNA degradation	1 (7.14%)	65 (3.06%)	3.537e-1	ko03018

to the suppression or reprogramming of RNA processing functions. Finally, the innermost circle displays enrichment factors (e.g., *splicesomal conformational changes to generate catalytic conformation* > 0.3, GO:0000393). This quantifies the ratio of foreground to background genes and further confirms the prioritization of RNA splicing-related processes.

The bar plot further supports this conclusion, showing that among the top 20 enriched GO terms, RNA splicing-related processes (e.g., the top 6 terms) involve up to 19 genes, with extremely low p-value ($< 1 \times 10^{-24}$), indicating their high enrichment in this complex. *Ribonucleoprotein complex assembly* (GO:0022618) and *cellular protein-containing complex assembly* (GO:0034622), although involving slightly fewer genes (16), remain statistically significant (p-value $< 1 \times 10^{-14}$), supporting the critical roles of these complexes in RNA processing and structural assembly.

To further complement the GO enrichment analysis and provide a pathway-level understanding of the biological function of CLPC1, we conducted a KEGG pathway enrichment analysis, as presented in Table S3. The results reveal that all 14

proteins in the CLPC1 complex (100%) are associated with the *spliceosome* pathway (ko03040), which represents a substantial enrichment compared to the background distribution—only 3.53% of all annotated proteins in the PPI network are linked to this pathway. This indicates that CLPC1 is structurally and functionally integrated into the spliceosomal machinery with high specificity. In contrast, other pathways showed no significant enrichment (p-values > 0.28), suggesting that CLPC1 is tightly related to RNA splicing rather than broadly associated with general RNA processing.

Our findings validate the core functional roles of the CLPC1 complex in spliceosome-mediated RNA splicing, highlighting its potential involvement in RNA processing dysregulation under pathological or stress conditions. The observed downregulation pattern of CLPC1 may reflect suppression or disruption of splicing activity, particularly in diseases associated with splicing defects, such as cancer and neurodegenerative disorders. Moreover, the statistically significant GO enrichment patterns and the low p-value from LAGO confirm the biological relevance of the identified complexes, underscoring their potential implications. These conclusions are further substantiated by the KEGG enrichment result, which offers additional functional evidence linking CLPC1 specifically to the spliceosome.

References

[1] Zeqian Li, Shilong Wang, Hai Cui, Xiaoxia Liu, and Yijia Zhang. Spatiotemporal constrained rna–protein heterogeneous network for protein complex identification. *Briefings in Bioinformatics*, 25(4), 2024.

[2] Hongwei Chen, Yunpeng Cai, Chaojie Ji, Gurudeeban Selvaraj, Dongqing Wei, and Hongyan Wu. Adappi: identification of novel protein functional modules via adaptive graph convolution networks in a protein–protein interaction network. *Briefings in Bioinformatics*, 24(1):bbac523, 2023.

[3] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, et al. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541, 2019.

[4] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80:S1–S7, 2018.