Supplementary Information

RNA-Protein Interaction Prediction Using Network-Guided Deep Learning

Haoquan Liu[1], Yiren Jian[2], Chen Zeng[3] and Yunjie Zhao[1]*

[1]Institute of Biophysics and Department of Physics, Central China Normal University, Wuhan, 430079, China

[2]Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA

[3]Department of Physics, The George Washington University, Washington, DC 20052, USA

*Correspondence:

Yunjie Zhao Email: yjzhaowh@ccnu.edu.cn

The Supplementary information includes 5 Supplementary Notes, 30 Supplementary Figures, and 5 Supplementary Tables.

Supplementary Notes

Supplementary Note 1 Comparison of different aggregation functions graph neural network embedding layer.

Various aggregator architectures can aggregate the neighbor representations in the graph neural network embedding layer, including the mean aggregator and the max pooling aggregator. The mean aggregator takes the elementwise mean of the vectors computed. Specifically, the mean aggregator computes:

$$\mathrm{AGGREGATE}_k^{\mathrm{mean}} = \mathrm{MEAN}\left(\left\{\bar{x}_u^{(k-1)}, \forall u \in \mathcal{N}(v)\right\}\right), \quad (1)$$

where $k = 1,2.\ \mathcal{N}(v)$ denotes the neighborhood of the node v. The max pooling approach processes each neighbor's vector separately through a fully connected neural network. After this transformation, an element-wise max pooling operation is performed to aggregate the information from the entire set of neighbors. Specifically, the max pooling aggregator computes:

$$\mathrm{AGGREGATE}_k^{\mathrm{max-pooling}} = \max\left(\left\{\sigma\left(W_{\mathrm{pool}}\bar{x}_u^{(k-1)} + b\right), \forall u \in \mathcal{N}(v)\right\}\right) \quad (2)$$

We compared the two aggregators, and the mean aggregator achieves 0.911 and 0.831 MCC on the benchmark sets NPInter2 and RPI7317, outperforming the max pooling aggregator, which achieves 0.879 and 0.806, respectively (see Supplementary Figure 21a and 21b). In the TheNovel set, the mean aggregator also achieves 0.798 AUROC and 0.820 AUPRC, surpassing the max pooling aggregator's 0.787 AUROC and 0.814 AUPRC (see Supplementary Figure 21c). These results demonstrate the better performance of the mean aggregator.

Supplementary Note 2 Comparison of different feature extracting layers in the VecNN architecture.

In ZHMolGraph VecNN, we use a 1D convolutional layer to extract features of the RNA and protein embeddings. We have also explored alternative architectures by replacing the 1D convolutional layer (VecNN) with a transformer encoder layer (TransformerNN). The Transformer encoder layer has four attention heads in the multi-head attention mechanism and a hidden dimension 2048.

The results indicate that different network structures perform at 0.911 MCC for VecNN and 0.894 MCC for TransformerNN on the NPInter2 set (see Supplementary Figure 24a). And 0.831 MCC for VecNN and 0.823 MCC for TransformerNN on the RPI7317 set (see Supplementary Figure 24b). The performance of the TransformerNN is worse than that of the convolutional layer by 7.3% in AUROC and 6.0% in AUPRC on the TheNovel dataset (see Supplementary Figure 24c). These results demonstrate the accuracy and generalizability of the VecNN architecture.

Supplementary Note 3 The detail of the hyperparameter optimization and loss convergence curve.

We utilized Optuna[1], a widely used hyperparameter optimization framework that assists in finding the optimal combination of hyperparameters during the training process of deep learning models by employing efficient search algorithms such as Tree-structured Parzen Estimator (TPE). We utilize the MCC as the objective value during five-fold validation on NPInter2. Following this, we conduct 100 trials to explore the hyperparameter search space and identify an optimized set of parameters for training (see Supplementary Figure 25a). The hyperparameter searches focus on the learning rate, which is sampled logarithmically between 1e-5 and 1e-1; the batch size, which varies among 16, 32, 64, 128, and 256; the number of epochs, which ranges from 30, 60, 90, 120 to 150; the convolution kernel size, which includes options of 3, 5, 7, and 9; and the number of output layers, which can be 1, 2, 3, and 4.

The calculations of hyperparameter importance indicate that the learning rate is the most significant factor affecting performance (see Supplementary Figure 25b). Supplementary Figure 25c displays the distribution of objective values for various hyperparameters: batch size, epochs, convolution kernel size, learning rate, and the number of output layers. Our model's parameters (learning rate of 1e-3, batch size of 128, 120 epochs, a convolution kernel size of 3, and one output layer) are within a reasonable range. Our goal is to develop a generalized AI-physical model for predicting RNA-protein interactions. Therefore, we chose a kernel size 3 for each residue or nucleotide to capture local and adjacent characteristics, following the principles established in our previous work[2].

We calculated the Matthews Correlation Coefficient (MCC) and monitored the loss convergence during the five-fold validation process for each fold (Supplementary Figure 26) and the average values across all five folds (Supplementary Figure 27). The results indicate that the loss curve converged throughout the five-fold validation process. Additionally, the MCC values stabilized around 0.9, which suggests that the model does not show signs of overfitting. Together, the results indicate a balanced and effective training.

Supplementary Note 4 Analysis of the model's computational complexity and scalability regarding the time and GPU memory requirements.

The training for 120 epochs with a batch size of 128 on a single NVIDIA GeForce RTX 3090 takes approximately 20 minutes. It utilizes around 2 GB of GPU memory in a 5-fold cross-validation on the NPInter2 dataset, which has 10,412 positive interactions.

We further analyze the time and space requirements, which become crucial as the RPI network size increases. For training, we select 10% to 90% of the interactions with a step size of 10% from the NPInter2 dataset. This allows us to examine how the time and memory requirements change as the size of the RPI network increases. We evaluate the size of the RPI network using two metrics: the number of edges and the average degree. As shown in Supplementary Figure 28, an increase in the number of edges leads to linear growth in computational time (see Supplementary Figure 28a) and GPU memory (see Supplementary Figure 28c). In contrast, as the average degree increases, computational time (see Supplementary Figure 28b) and GPU memory (see Supplementary Figure 28d) exhibit quadratic growth trends. Our method, ZHMolGraph, can efficiently predict the RNA-protein interaction.

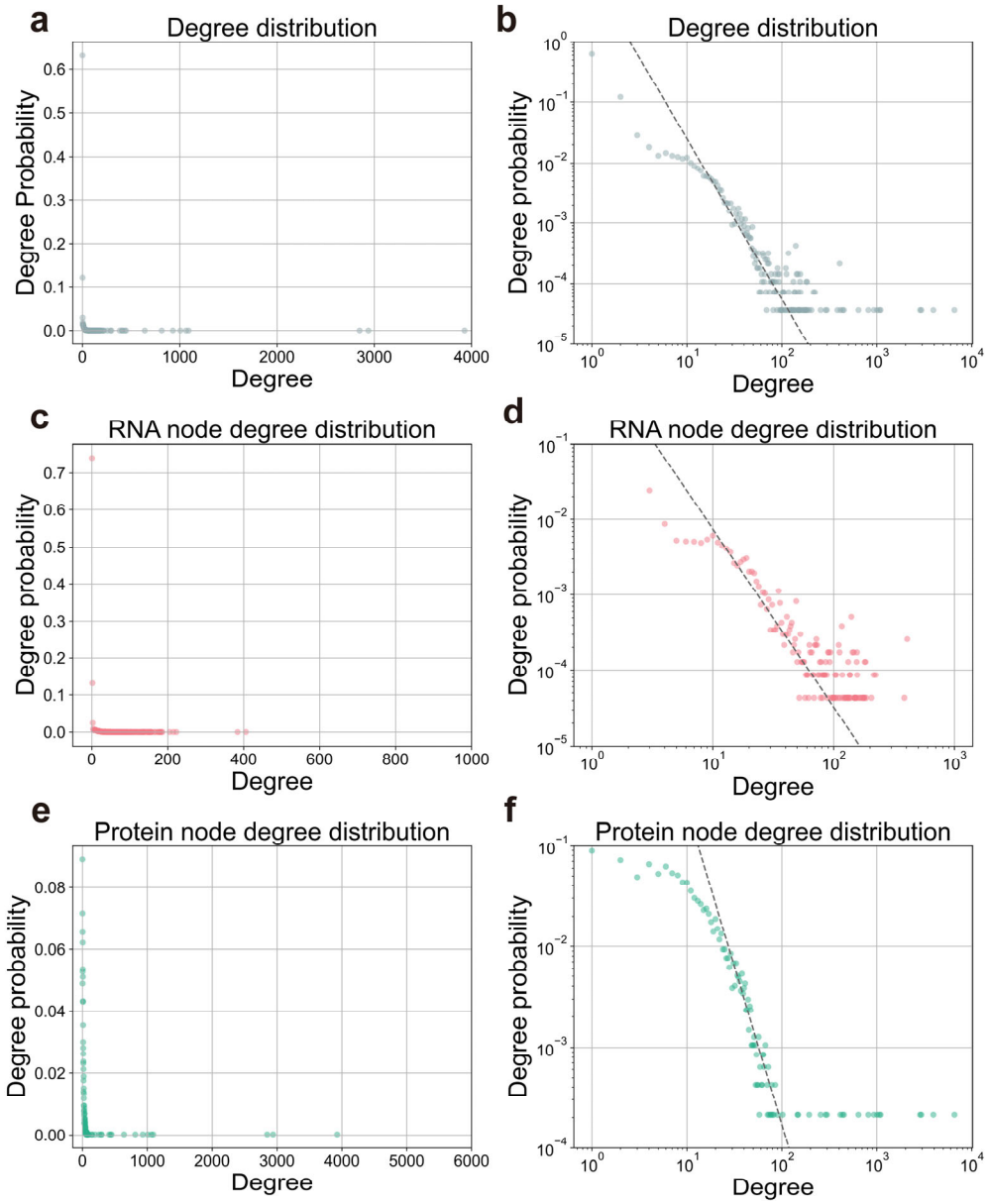
Supplementary Note 5 Comparison of different protein embedding models.

In ZHMolGraph, we utilize a graph neural network model to sample and aggregate features from the local neighborhoods of nodes for RNA-protein interaction prediction. We incorporated large language models into the graph neural network to enhance node characterization, improving prediction accuracy.

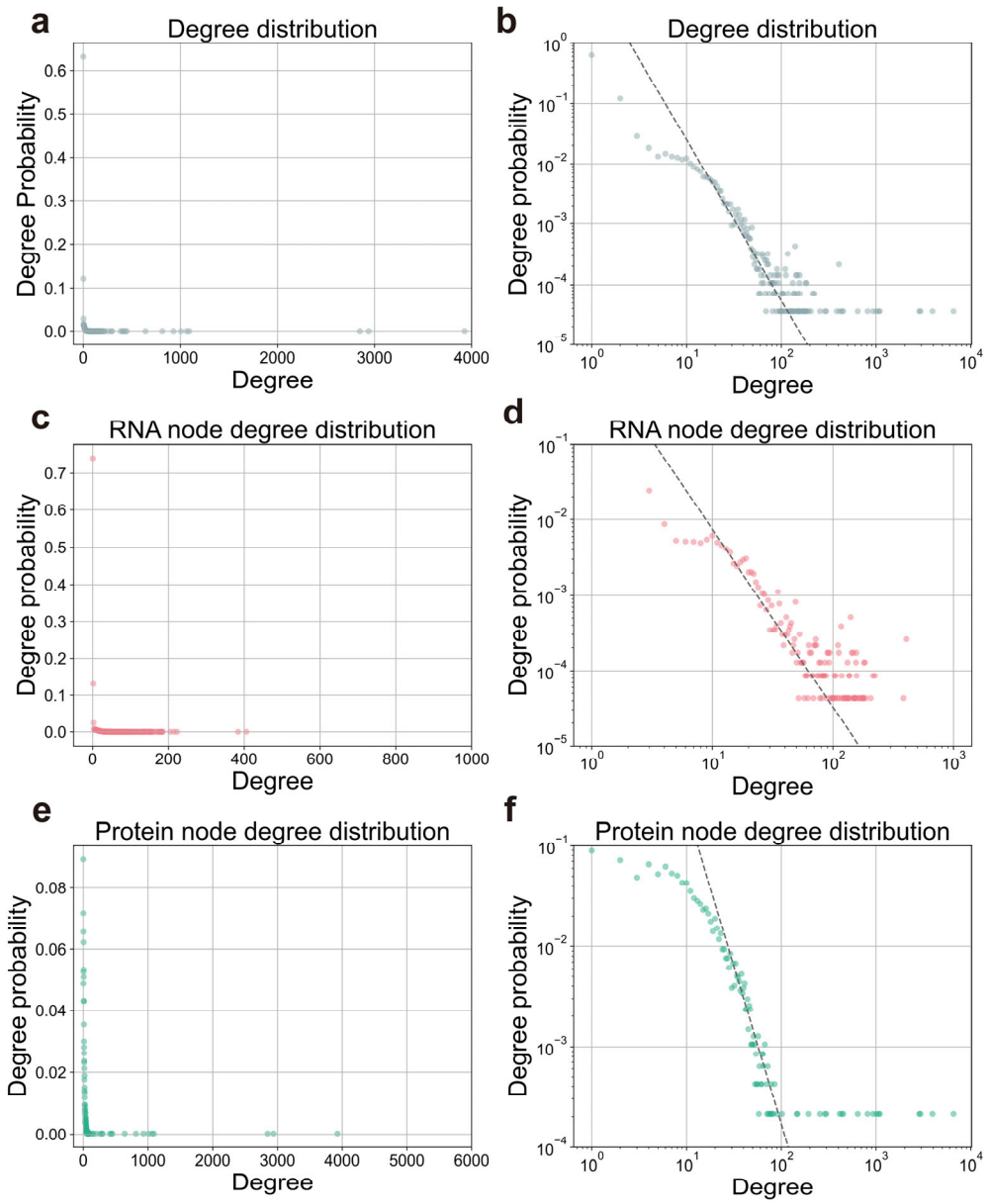
Since there are other LLMs for proteins like ESM-2[3], we conducted a comparative analysis between ProtTrans and ESM-2. For ESM-2, we used esm2_t33_650M_UR50D as the pre-trained model.

In the benchmark sets NPInter2 and RPI7317, ProtTrans outperformed ESM-2, achieving MCC values of 0.911/0.901 and 0.831/0.783, respectively (see Supplementary Figure 29a, b and Supplementary Table 4). Notably, on the TheNovel test set, ProtTrans also exceeded ESM-2's performance, with AUROC and AUPRC values of 0.798 and 0.820 for ProtTrans compared to 0.780 and 0.807 for ESM-2 (see Supplementary Figure 29c and Supplementary Table 5). This highlights the better performance of ProtTrans over ESM-2, emphasizing the generalizability of Large Language Models.

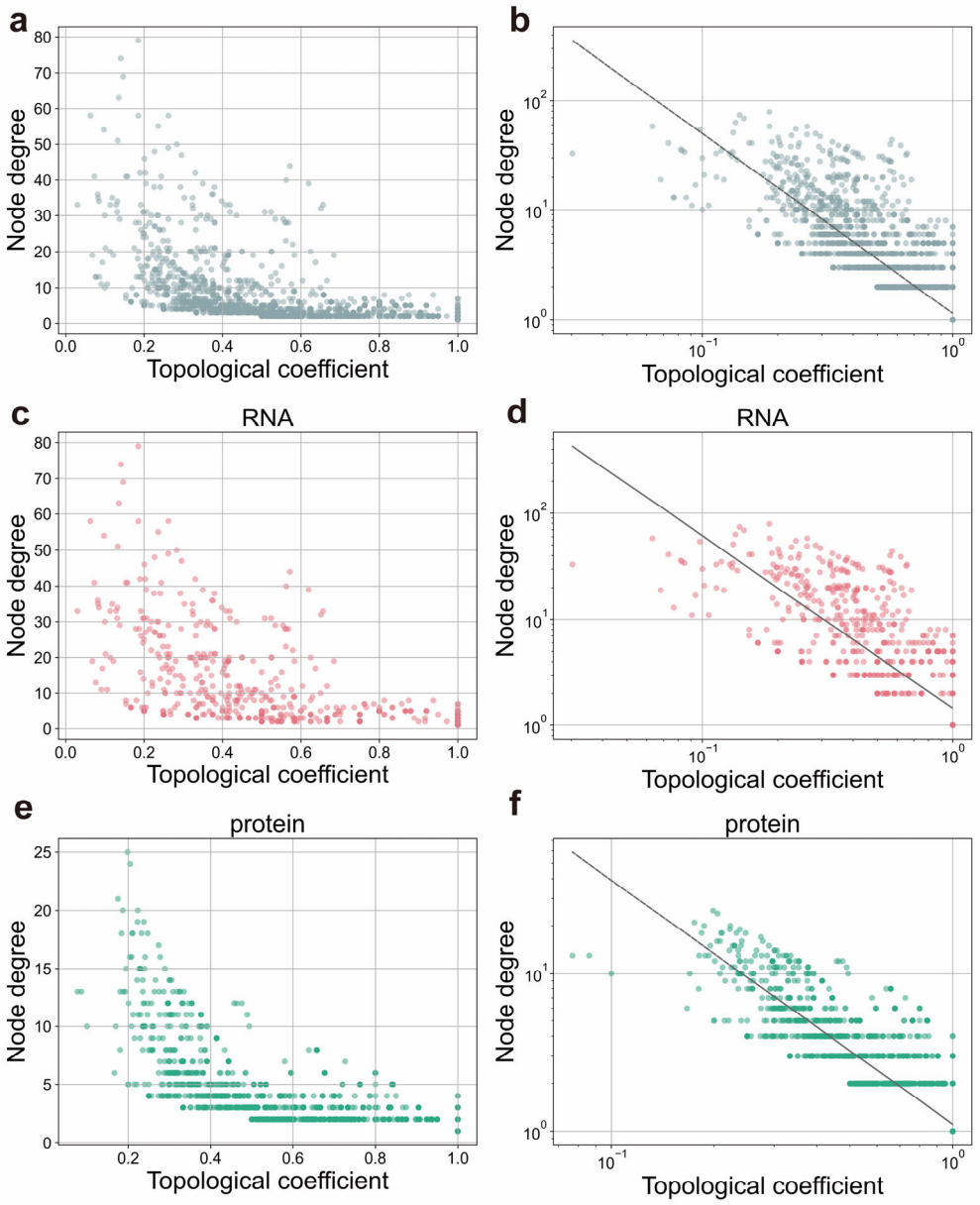
Supplementary Figures



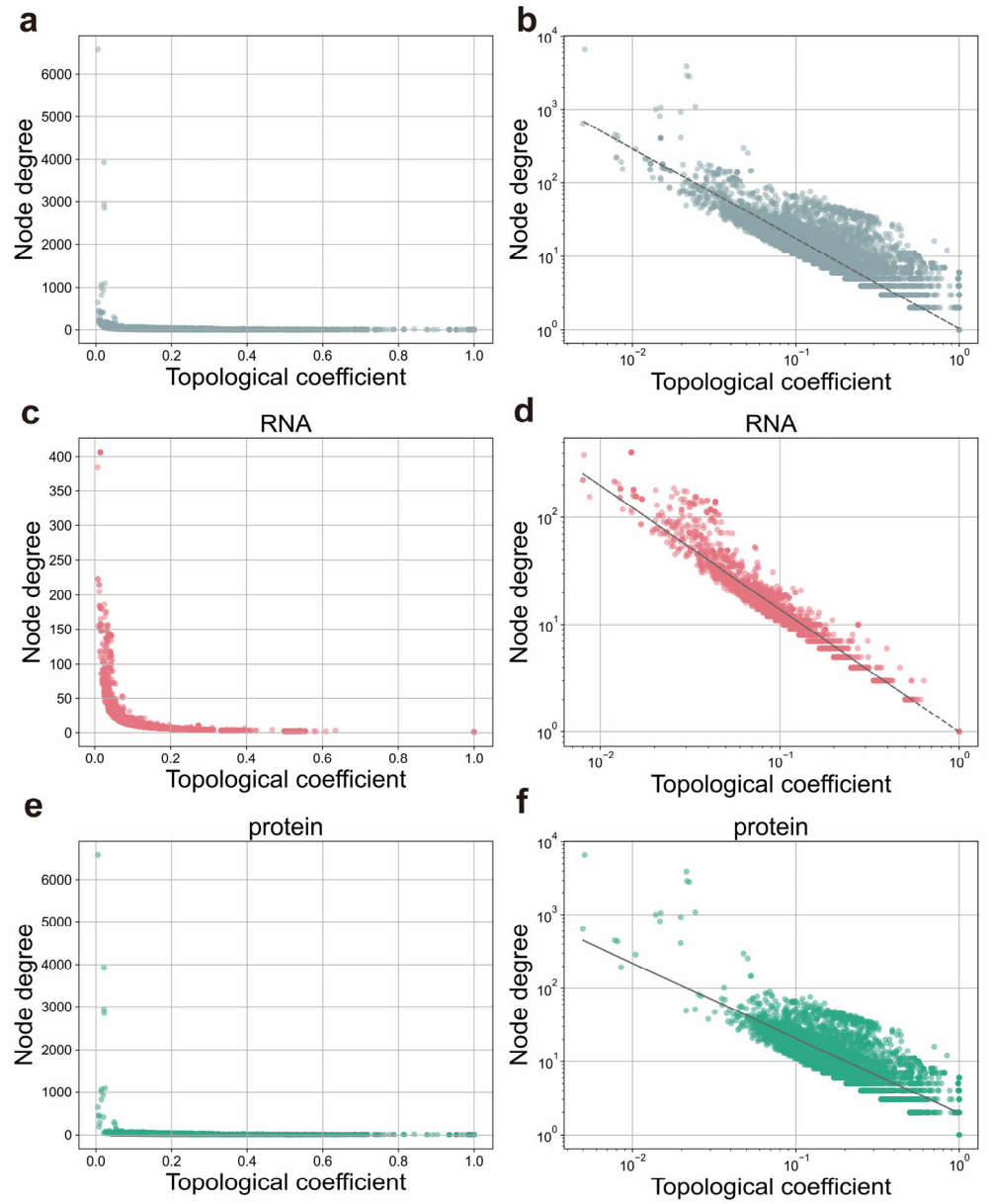
Supplementary Figure 1. Degree Distribution of RNA-protein high-throughput interaction network. a, c, e The degree distributions of (**a**) all nodes, (**c**) RNA nodes, and (**e**) protein nodes in RNA-protein high-throughput interaction network. **b, d, f** The degree distributions of (**b**) all nodes, (**d**) RNA nodes, and (**f**) protein nodes in the RNA-protein high-throughput interaction network are shown in double logarithmic axes (log-log plot) indicate that the degree probability of all nodes, RNA nodes, and protein nodes are well approximated by a power law with degree exponents 2.665, 2.361, and 3.139 respectively. Source data are provided in Supplementary Data 6.



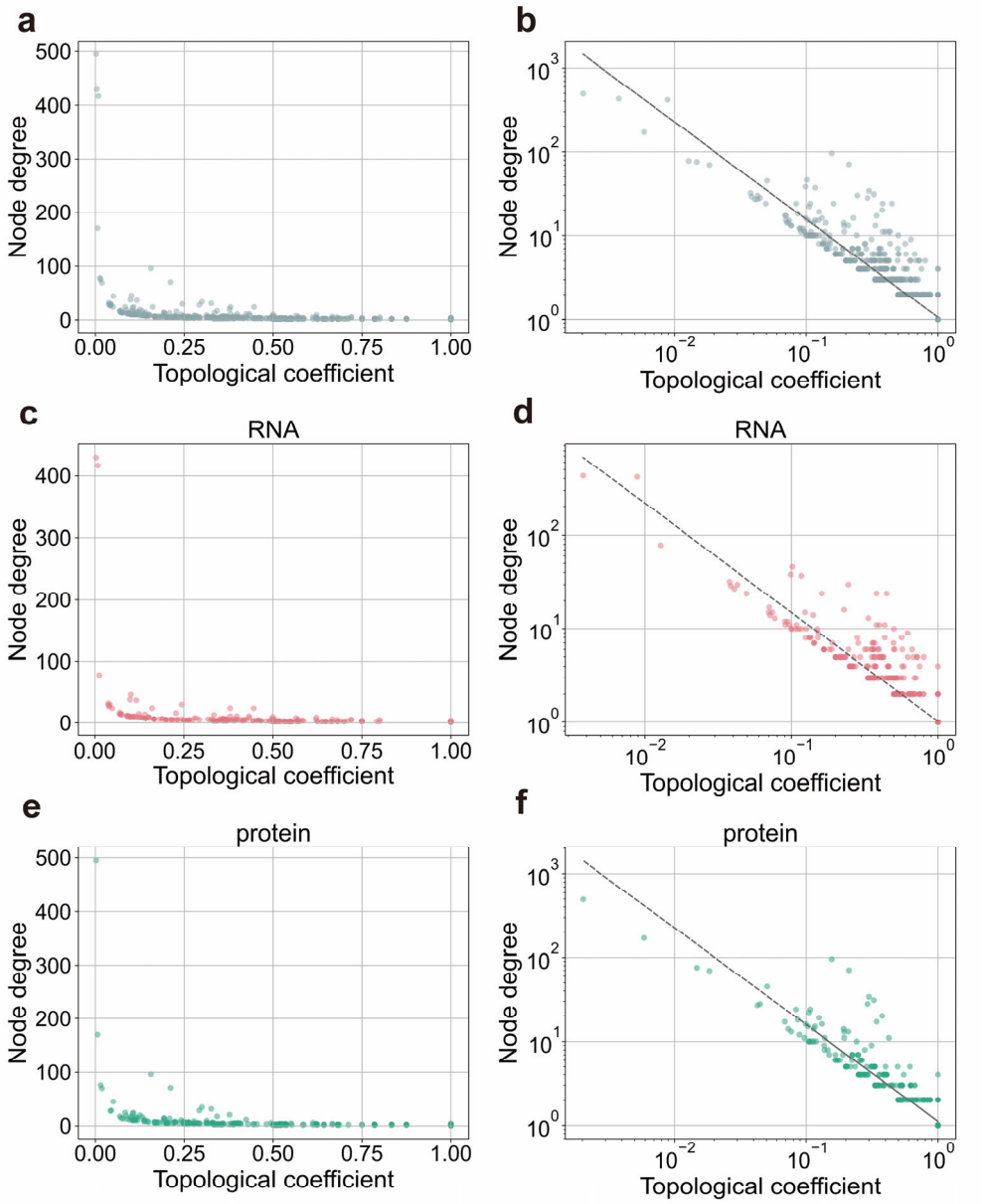
Supplementary Figure 2. Degree Distribution of RNA-protein literature-mining validated interaction network. a, c, e The degree distributions of (**a**) all nodes, (**c**) RNA nodes, and (**e**) protein nodes in RNA-protein literature-mining validated interaction network. **b, d, f** The degree distributions of (**b**) all nodes, (**d**) RNA nodes, and (**f**) protein nodes in RNA-protein experimentally validated interaction network are shown in double logarithmic axes (log-log plot), indicate that the degree probability of all nodes, RNA nodes, and protein nodes are well approximated by a power law with degree exponents 2.537, 2.629, and 2.543 respectively. Source data are provided in Supplementary Data 6.



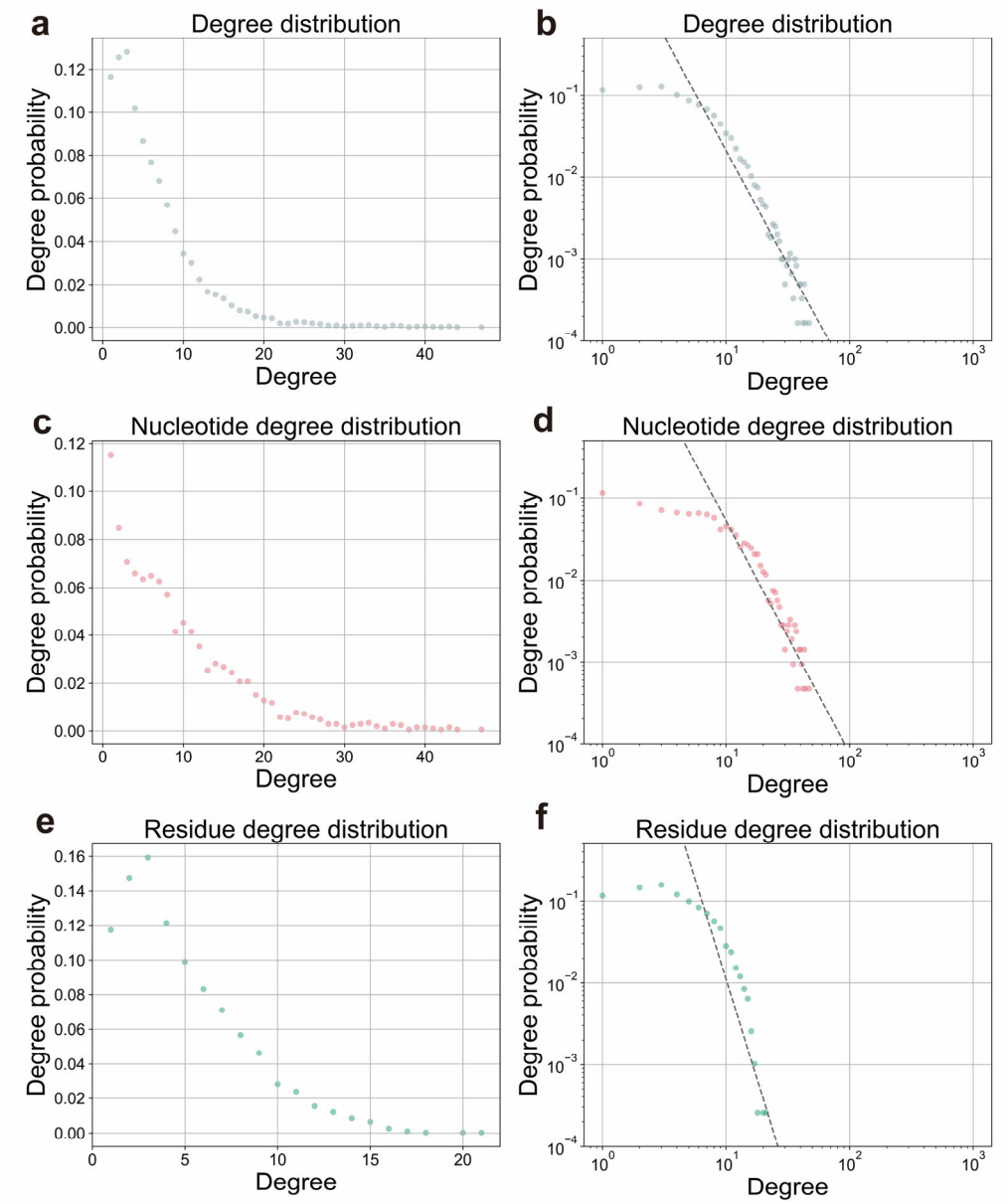
Supplementary Figure 3. Topological coefficient of RNA-protein structural interaction interaction network. a, c, e The correlation between the topological coefficient and node degree of (**a**) all nodes, (**c**) RNA nodes, and (**e**) protein nodes in the RNA-protein structural interaction network. **b, d, f** The correlation between the topological coefficient and node degree of (**b**) all nodes, (**d**) RNA nodes, and (**f**) protein nodes in the RNA-protein structural interaction network are shown in double logarithmic axes (log-log plot) indicate that the topological coefficient and node degree of all nodes, RNA nodes, and protein nodes are anti-correlated with Spearman correlation coefficients -0.927, -0.856, and -0.944, respectively. Source data are provided in Supplementary Data 6.



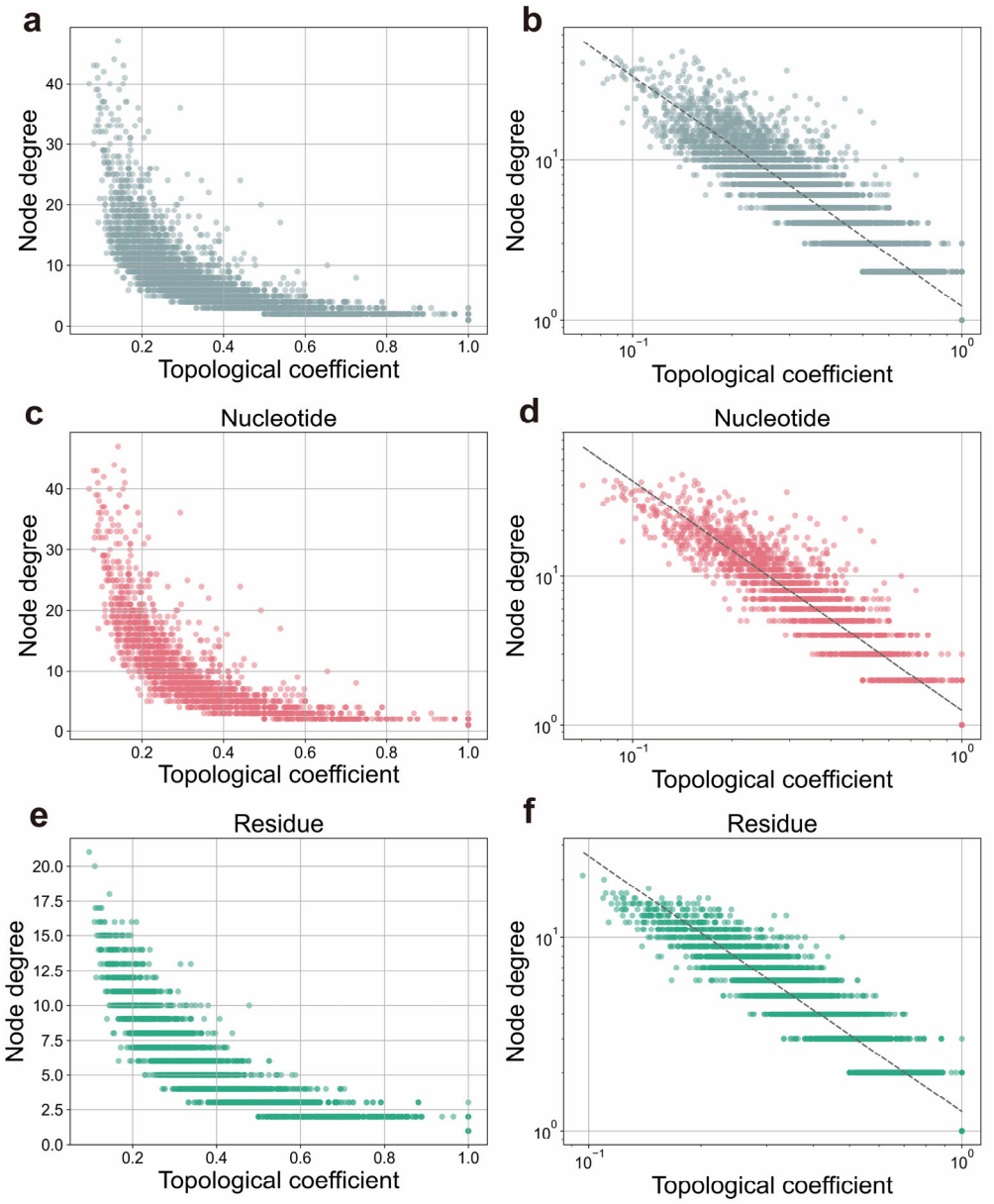
Supplementary Figure 4. Topological coefficient of RNA-protein high-throughput interaction interaction network. a, c, e The correlation between the topological coefficient and node degree of (**a**) all nodes, (**c**) RNA nodes, and (**e**) protein nodes in RNA-protein high-throughput interaction network. **b, d, f** The correlation between the topological coefficient and node degree of (**b**) all nodes, (**d**) RNA nodes, and (**f**) protein nodes in the RNA-protein high-throughput interaction network is shown in double logarithmic axes (log-log plot) indicates that the topological coefficient and node degree of all nodes, RNA nodes, and protein nodes are anti-correlated with Spearman correlation coefficients -0.977, -0.998, and -0.841, respectively. Source data are provided in Supplementary Data 6.



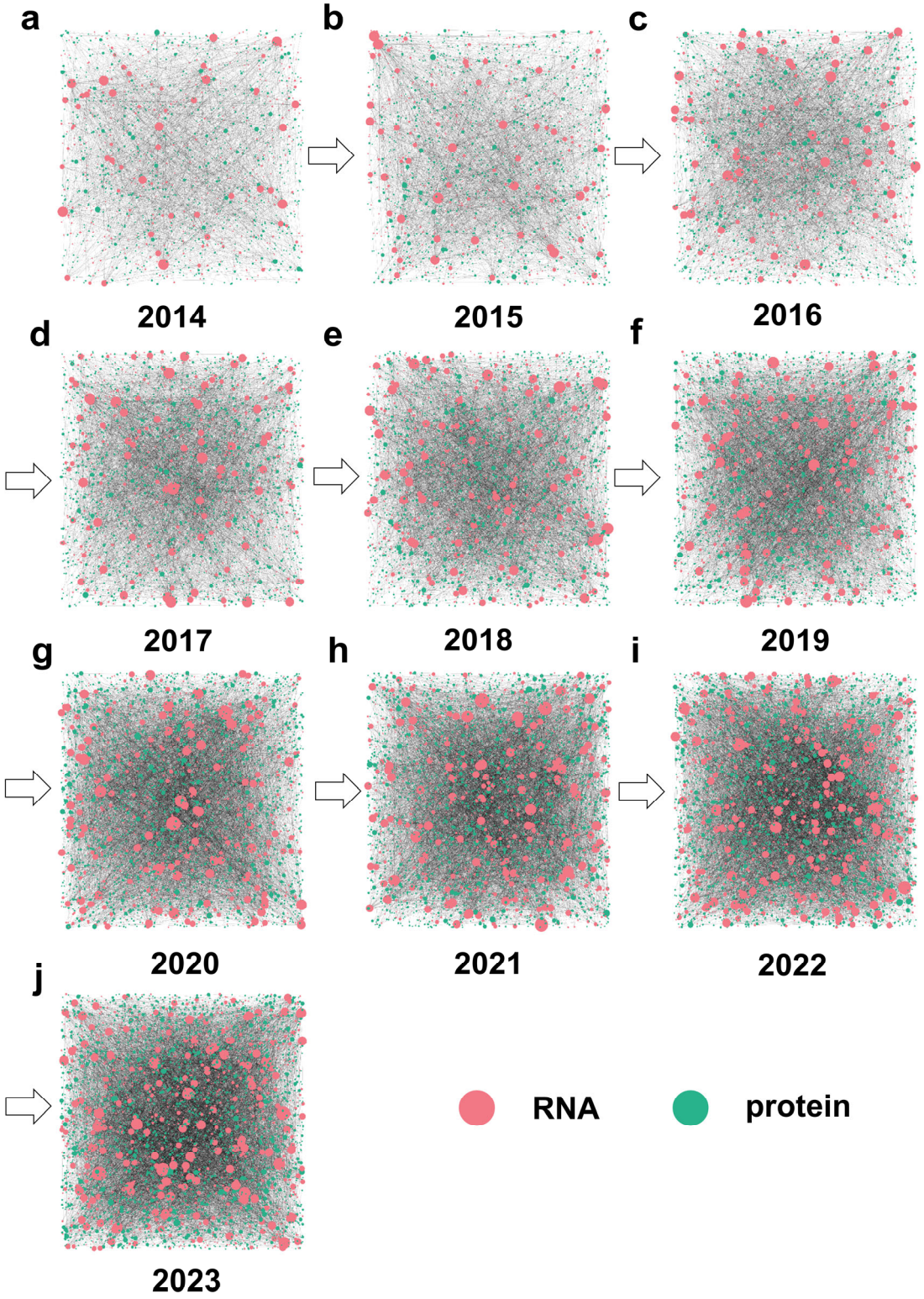
Supplementary Figure 5. Topological coefficient of RNA-protein literature-mining validated interaction network. a, c, e The correlation between the topological coefficient and node degree of (**a**) all nodes, (**c**) RNA nodes, and (**e**) protein nodes in RNA-protein literature-mining validated interaction network. **b, d, f** The correlation between the topological coefficient and node degree of (**b**) all nodes, (**d**) RNA nodes, and (**f**) protein nodes in the RNA-protein experimentally validated interaction network is shown in double logarithmic axes (log-log plot) indicates that the topological coefficient and node degree of all nodes, RNA nodes, and protein nodes are anti-correlated with Spearman correlation coefficients -0.982, -0.985, and -0.975, respectively. Source data are provided in Supplementary Data 6.



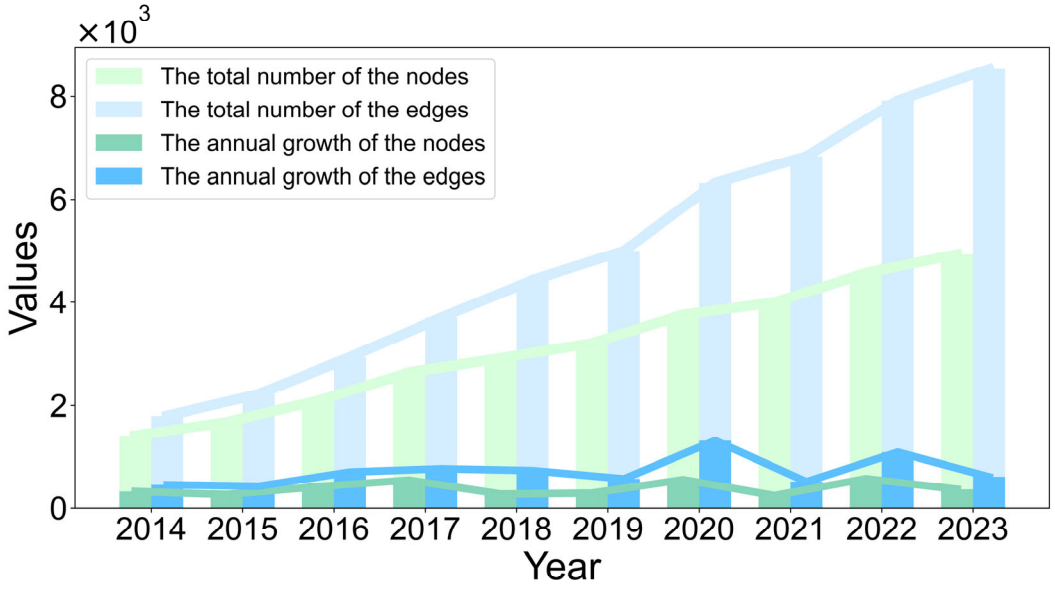
Supplementary Figure 6. Degree Distribution of RNA-protein residue interaction network. a, c, e The degree distributions of (**a**) all nodes, (**c**) RNA nodes, and (**e**) protein nodes in the RNA-protein residue interaction network. **b, d, f** The degree distributions of (**b**) all nodes, (**d**) RNA nodes, and (**f**) protein nodes in the RNA-protein residue interaction network are shown in double logarithmic axes (log-log plot), indicating that the degree probability of all nodes, RNA nodes, and protein nodes are well approximated by a power law with degree exponents 2.790, 2.838, and 4.876 respectively. Source data are provided in Supplementary Data 6.



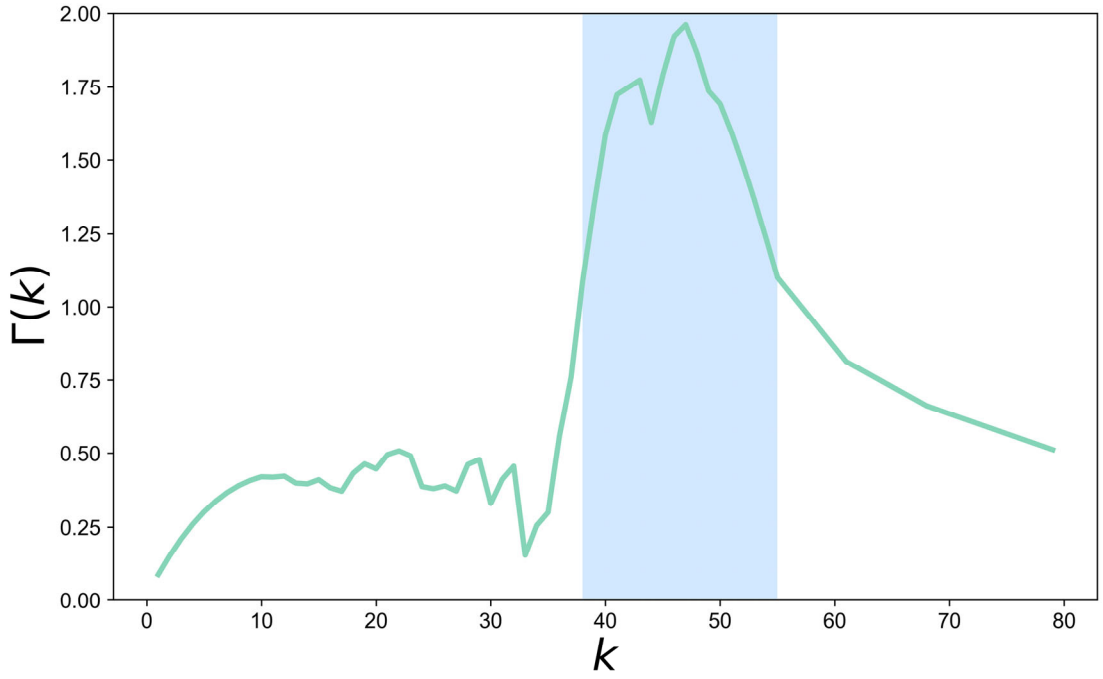
Supplementary Figure 7. Topological coefficient of RNA-protein residue interaction network. a, c, e The correlation between the topological coefficient and node degree of (**a**) all nodes, (**c**) RNA nodes, and (**e**) protein nodes in the RNA-protein residue interaction network. **b, d, f** The correlation between the topological coefficient and node degree of (**b**) all nodes, (**d**) RNA nodes, and (**f**) protein nodes in the RNA-protein residue interaction network are shown in double logarithmic axes (log-log plot) indicate that the topological coefficient and node degree of all nodes, RNA nodes, and protein nodes are anti-correlated with Spearman correlation coefficients -0.938, -0.929, and -0.952, respectively. Source data are provided in Supplementary Data 6.



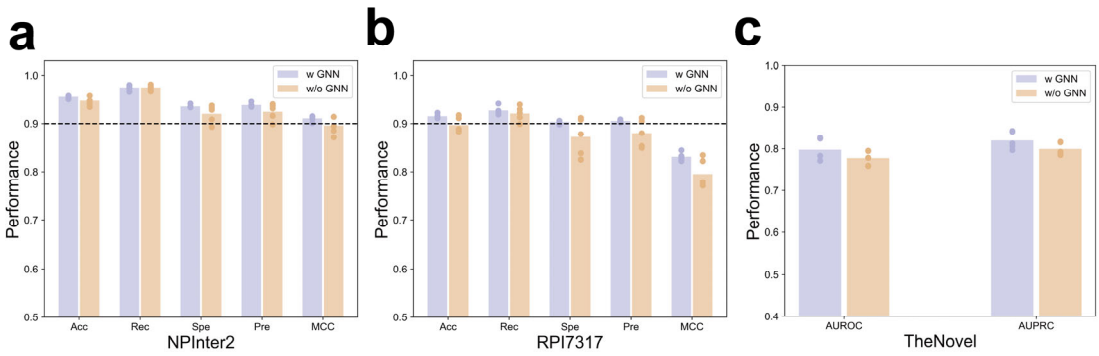
Supplementary Figure 8. The dynamic process of RNA-protein structural interaction network. The network of RNA-protein structural interactions from 2014 to 2023 is visualized as nodes in pink (RNA) and green (protein). Source data are provided in Supplementary Data 6.



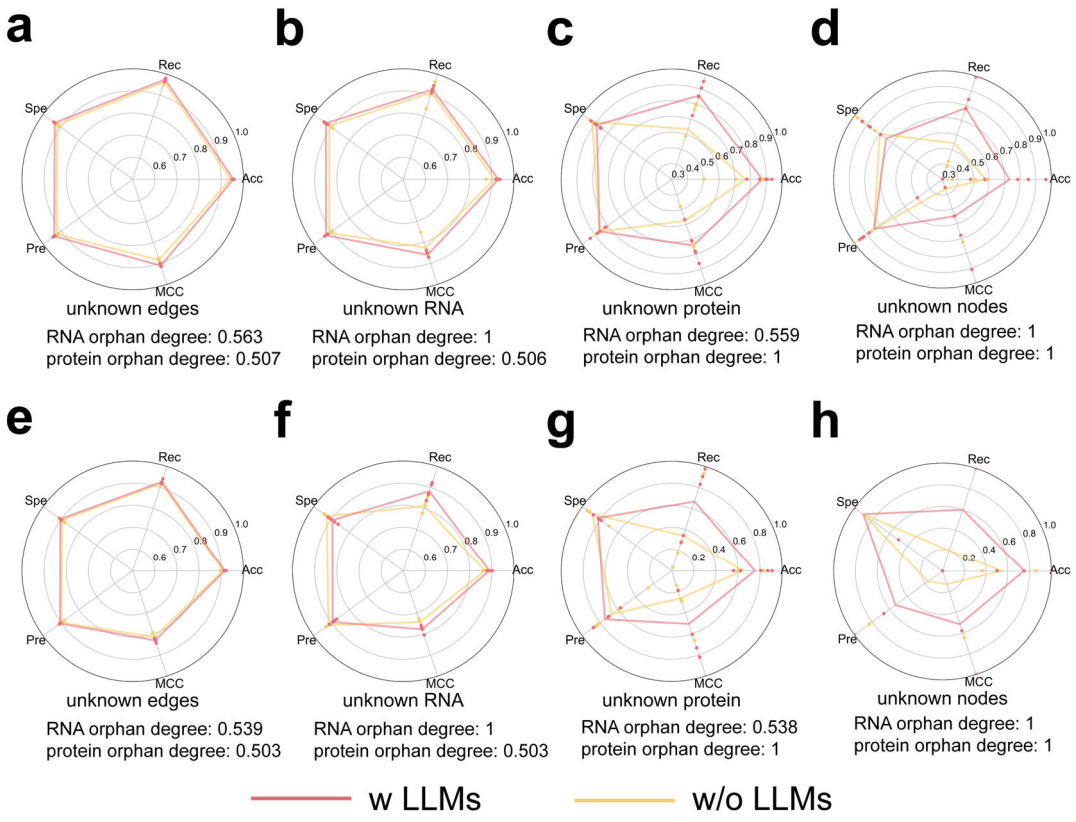
Supplementary Figure 9. The growth of the RNA-protein structural interaction network. The total number and the annual growth of the nodes and edges in the RNA-protein structural interaction network from 2014 to 2023. Source data are provided in Supplementary Data 6.



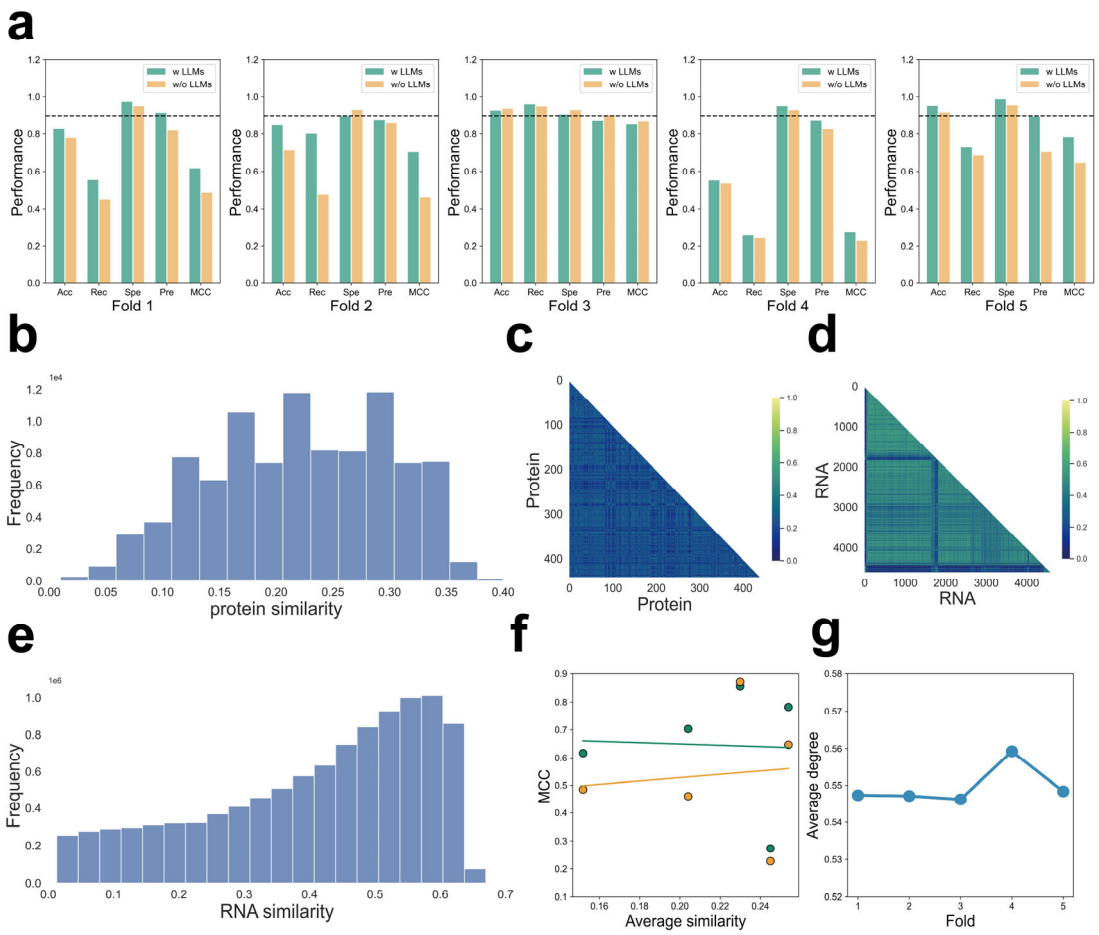
Supplementary Figure 10. The ability of nodes to acquire new links in the RNA-protein structural interaction network. The average capacity of nodes with distinct degrees to acquire new connections spans from 2014 to 2023. Source data are provided in Supplementary Data 6.



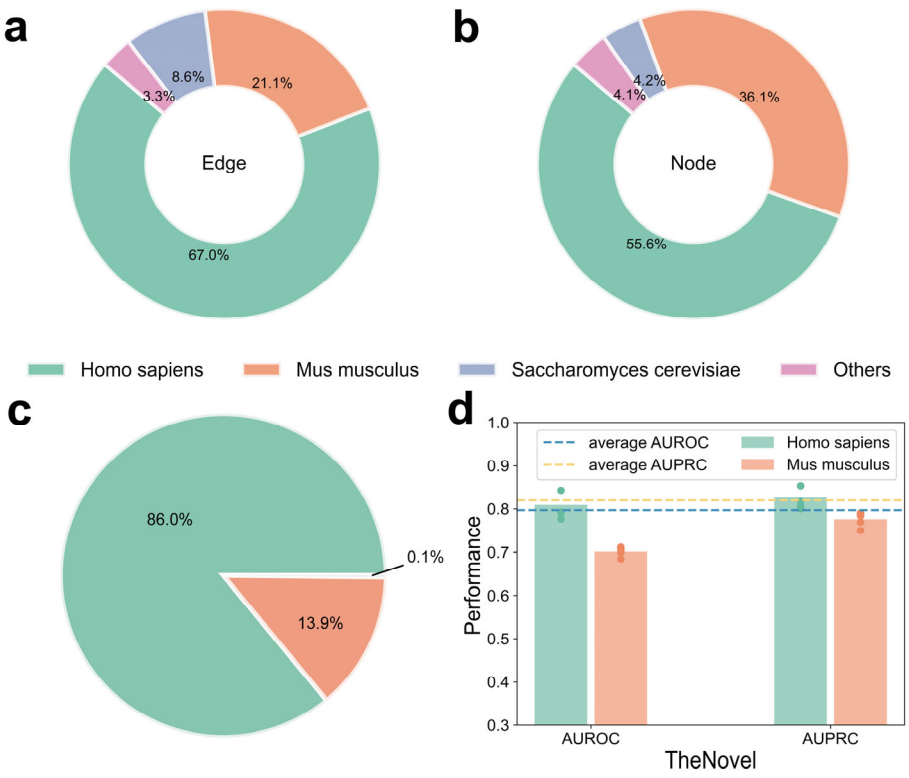
Supplementary Figure 11. Performance comparison with and without GNN Embeddings. a Performance in benchmark NPInter2 in a n=5-fold cross-validation. **b** Performance in benchmark RPI7317 in a n=5-fold cross-validation. **c** Performance in the unknown dataset TheNovel using the n=5 models trained on the NPInter2 dataset. The dashed lines in the figures mark the position where the y-axis value is 0.9. In (a) and (b), dots represent the performance of each fold, and bar height corresponds to the mean. In (c), dots represent the performance of each model, bar height corresponds to the mean. Source data are provided in Supplementary Data 6.



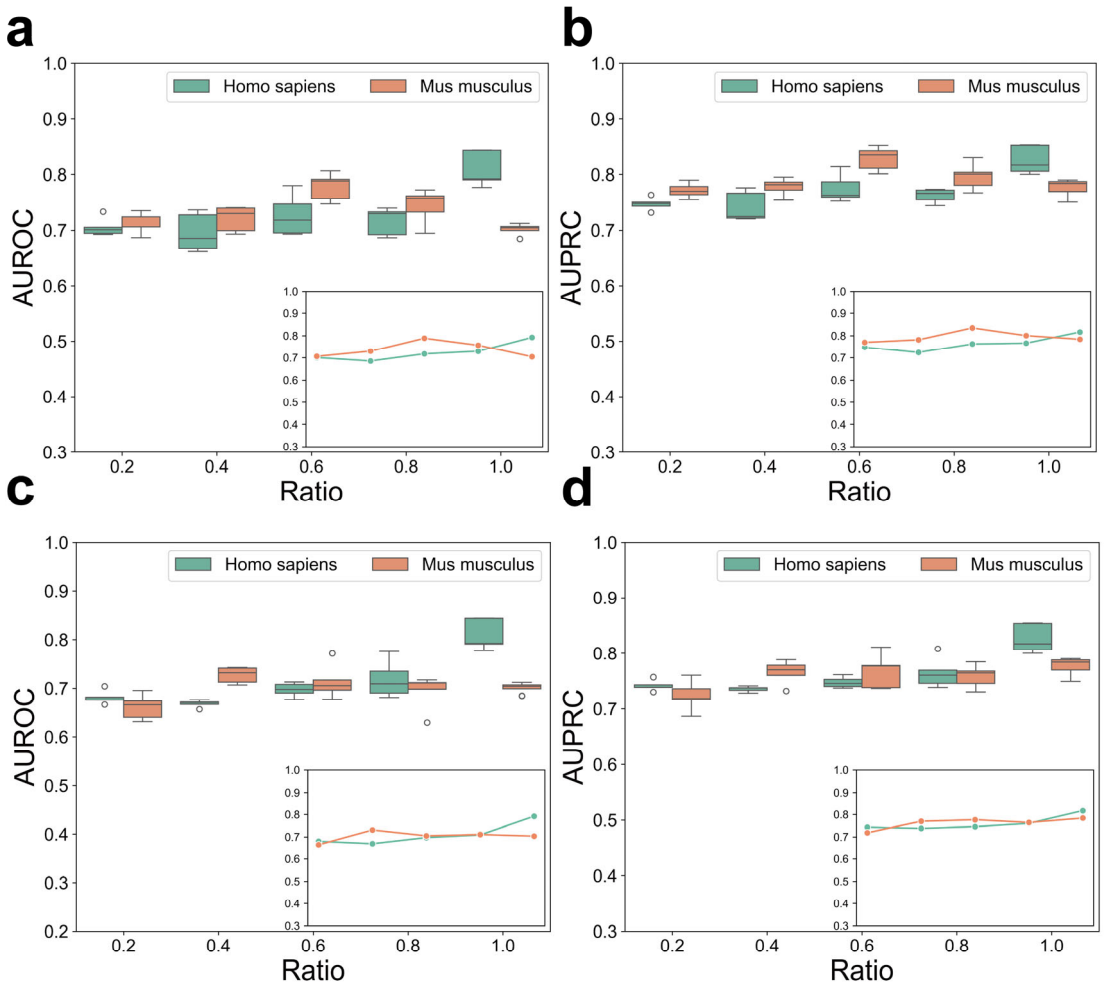
Supplementary Figure 12. Performance evaluation of "orphan" nodes effect using five-fold cross-validation on benchmark datasets NPInter2 (a-d) and RPI7317 (e-h) with and without large language models (LLMs). Performance in the case of **a, e** unknown edges, **b, f** unknown RNA, **c, g** unknown protein, **d, h** unknown nodes. RNA/protein degree refers to the average degree of the testing set RNA/protein nodes in the training set network. Dots in all figures represent the performance of each fold in a n=5-fold cross-validation, the data lines on the radar charts correspond to the mean. Source data are provided in Supplementary Data 6.



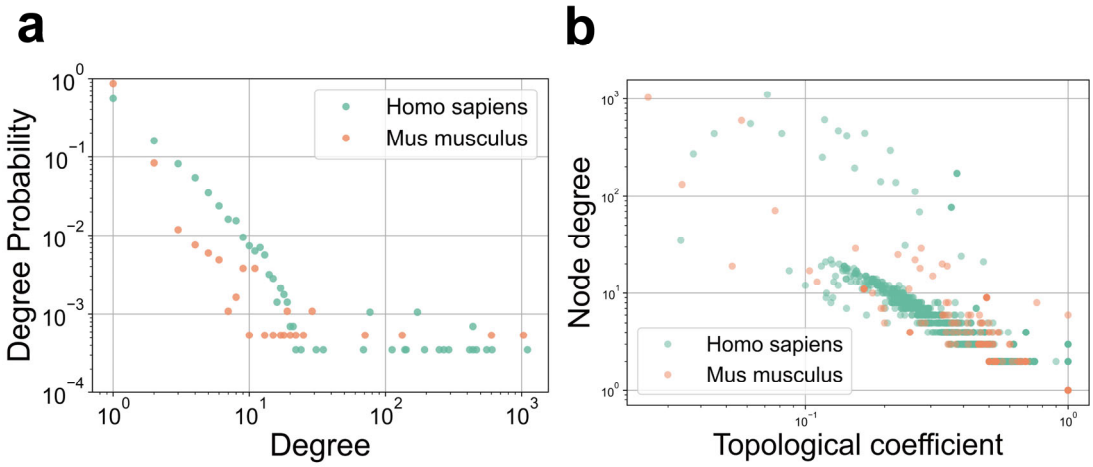
Supplementary Figure 13. Performance of the sequence similarity. a Performance of five-fold cross-validation based on the test set partitioned by protein sequence similarity. The dashed line in the figures marks the position where the y-axis value is 0.9. **b** The histogram of the distribution of pairwise protein sequence similarity. **c** The pairwise protein sequence similarity of the NPInter2 dataset. **d** The pairwise sequence RNA similarity of the NPInter2 dataset. **e** The histogram of the distribution of pairwise protein sequence similarity. **f** The relation between average similarity and the MCC for each fold. **g** The orphan degree of RNA nodes from the testing set for each fold. Source data are provided in Supplementary Data 5 (d, e) and Supplementary Data 6 (a, b, c, f, g).



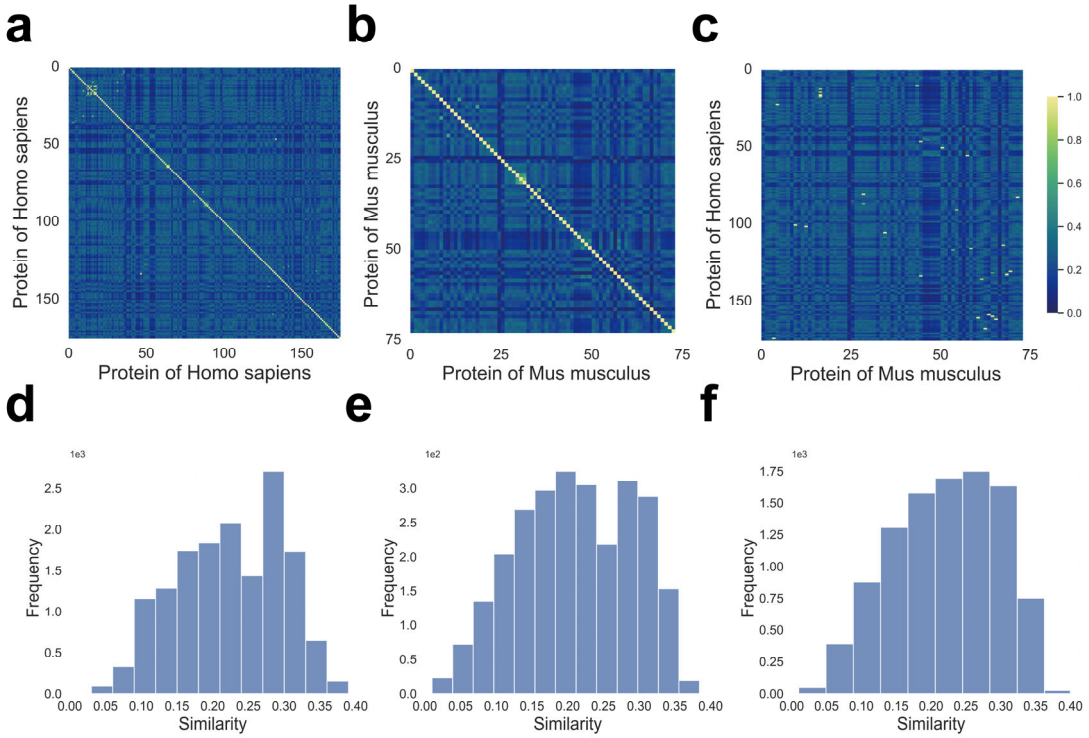
Supplementary Figure 14. The distribution and performance across various species. a The distribution of interaction edges across different species in the NPInter2 set. **b** The distribution of interaction nodes across different species in the NPInter2 set. **c** The distribution of interactions across different species in the NPInter2 set. **d** The AUROC and AUPRC performance for *Homo sapiens* and *Mus musculus* interactions in the TheNovel set using n=5 models trained on the NPInter2 dataset (dots in the figure represent the performance of each model, bar height corresponds to the mean, the dashed lines representing the average performance in the whole TheNovel set). Source data are provided in Supplementary Data 6.



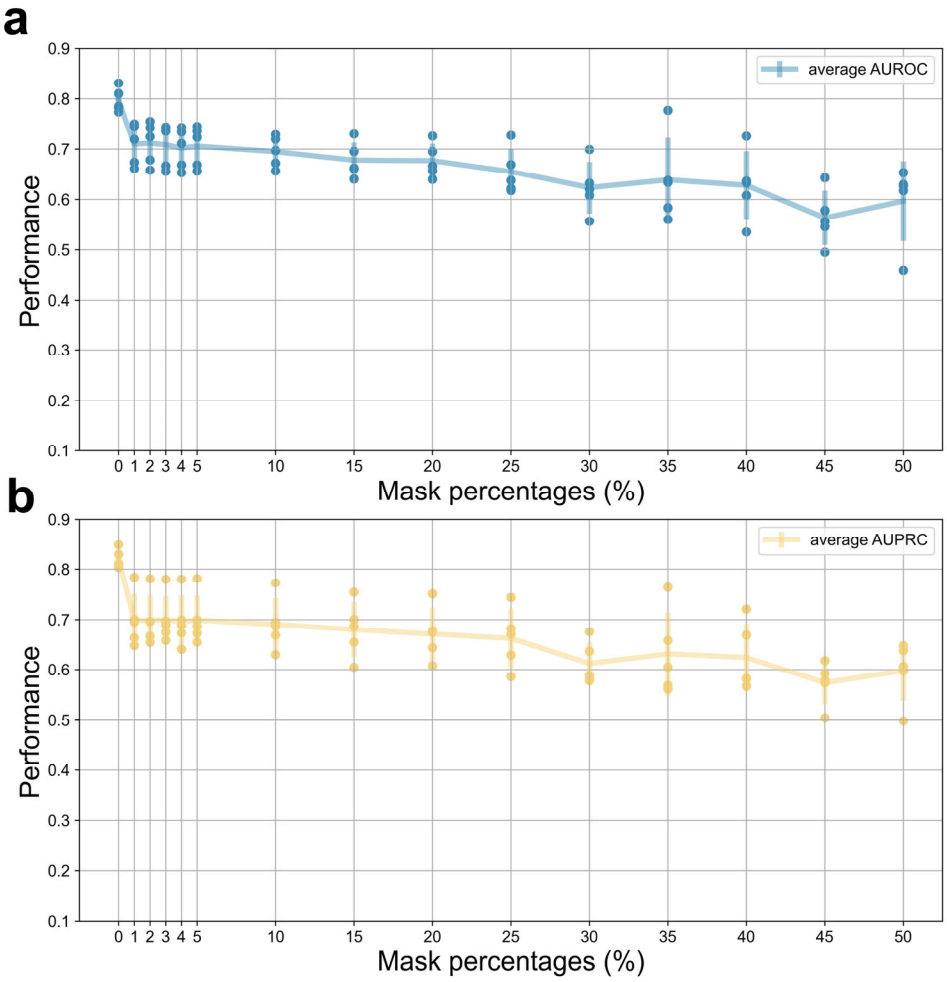
Supplementary Figure 15. The effect of training data size on model performance in TheNovel dataset across different species. a, b Relationship between the ratio of *Homo sapiens* interactions and the performance of AUROC (a) and AUPRC (b) in the n=5 models trained on the NPInter2 dataset. The inset line chart represents the mean. **c, d** Relationship between the ratio of *Mus musculus* interactions and performance in AUROC (c) and AUPRC (d) in the n=5 models trained on the NPInter2 dataset. The inset line chart represents the mean. Source data are provided in Supplementary Data 6.



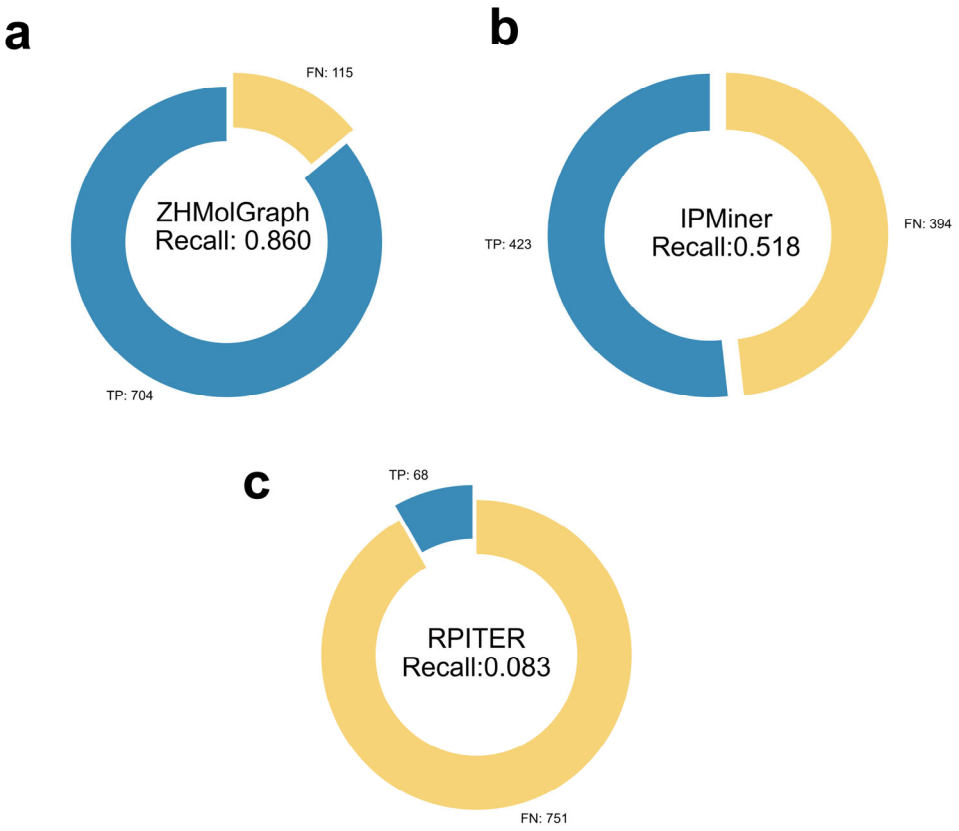
Supplementary Figure 16. RPI network topological properties between species. a The degree distributions of different species are shown in double logarithmic axes (log-log plot). **b** The correlation of different species between the topological coefficient and node degree is shown in double logarithmic axes (log-log plot). Source data are provided in Supplementary Data 6.



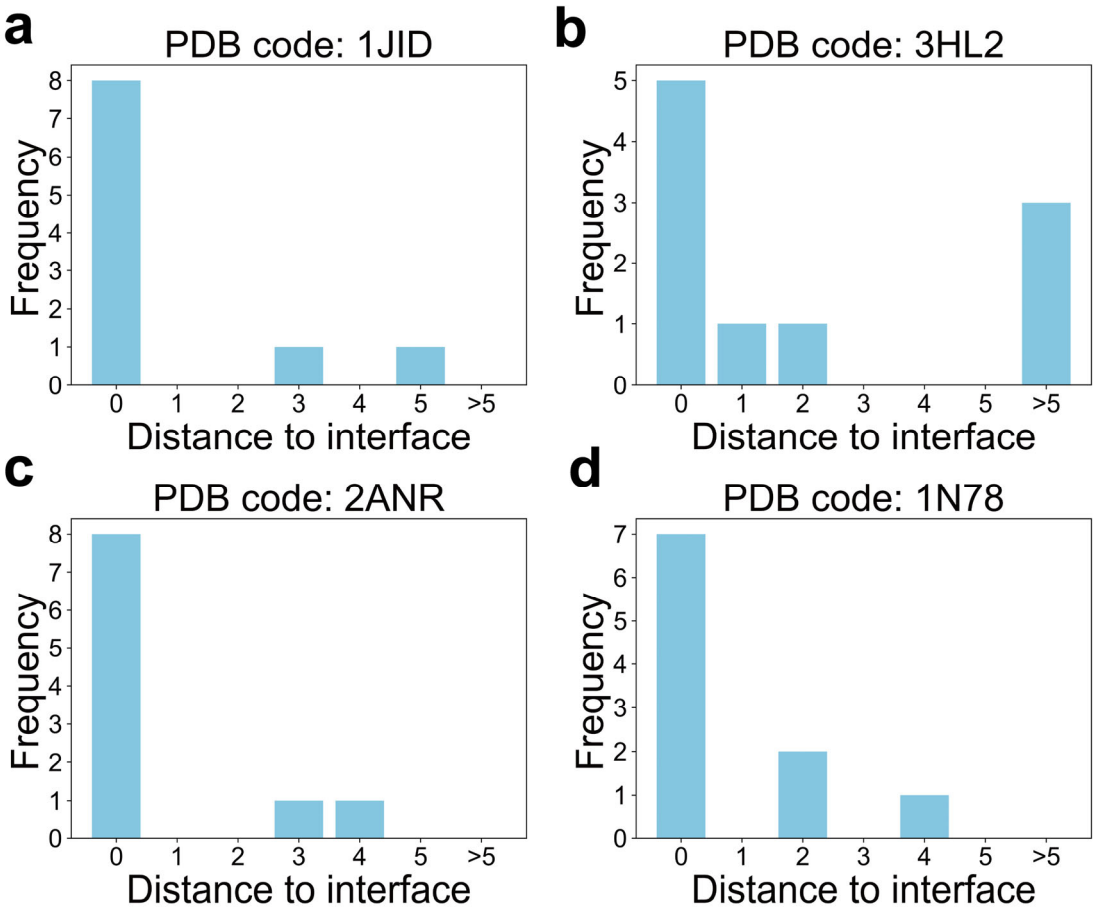
Supplementary Figure 17. RPI network topological properties between species. a The degree distributions of different species are shown in double logarithmic axes (log-log plot). **b** The correlation of different species between the topological coefficient and node degree is shown in double logarithmic axes (log-log plot). Source data are provided in Supplementary Data 6.



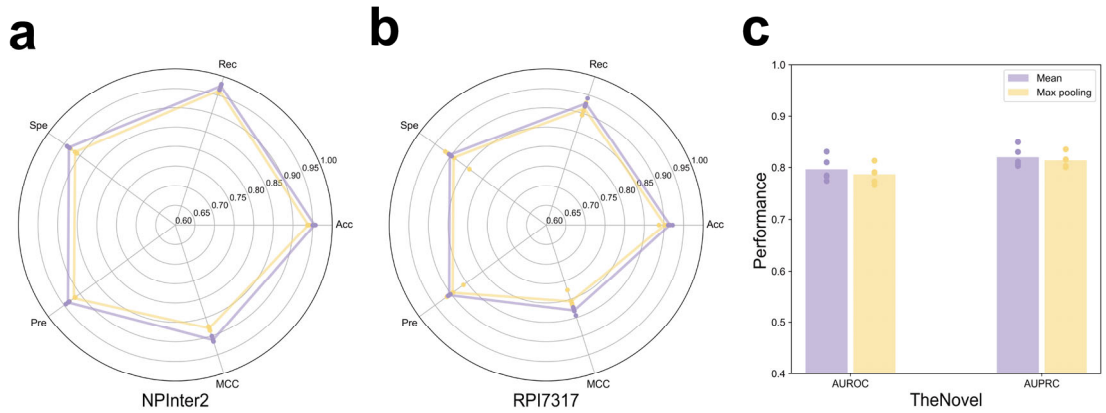
Supplementary Figure 18. Evaluating the Effects of Errors in Input Sequences on the TheNovel set. a Average AUROC and **b** average AUPRC of various sequence mask percentages in the n=5 models trained on the NPInter2 dataset. The error bars represent the standard deviation of performance AUROC and AUPRC. Source data are provided in Supplementary Data 6.



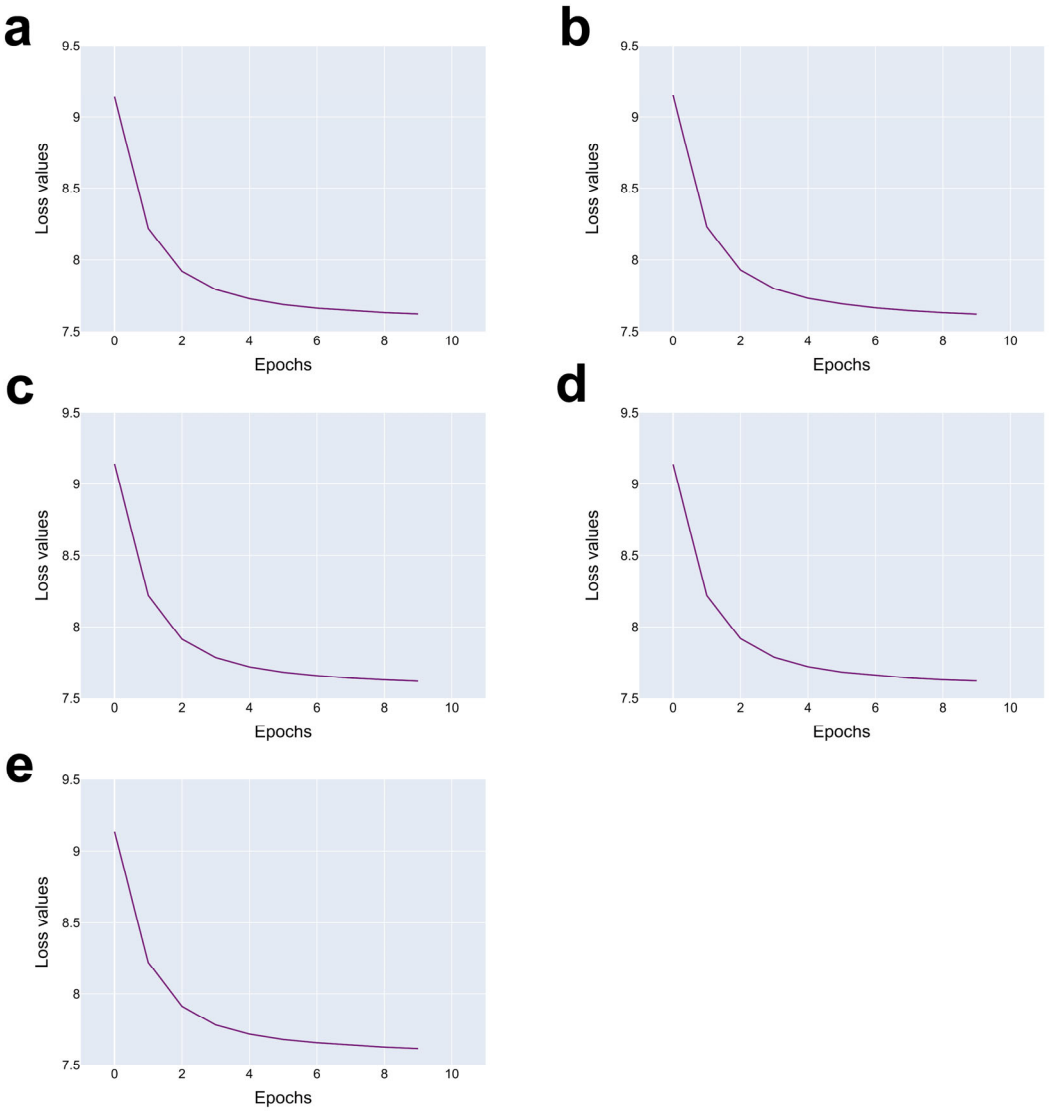
Supplementary Figure 19. Identification performance on RNA-protein interactions related to SARS-CoV-2. a ZHMolGraph. **b** IPMiner. **c** RPITER. TP means true positive, and FN means false negative in prediction. Source data are provided in Supplementary Data 6.



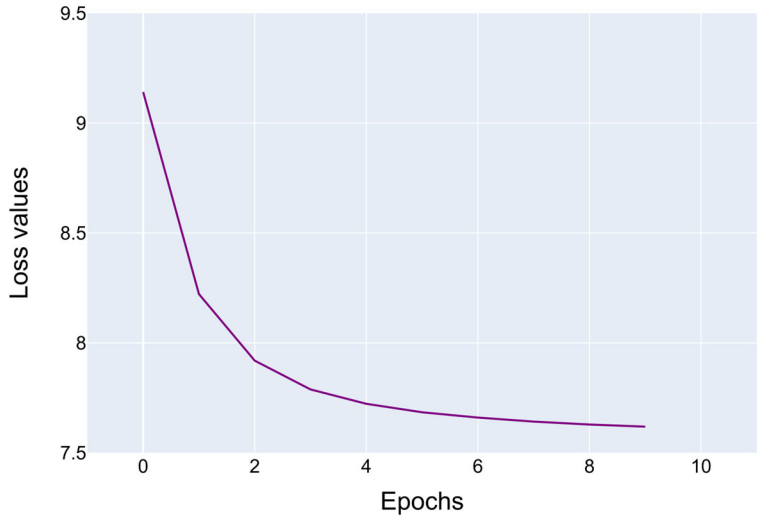
Supplementary Figure 20. The distribution of distances between the predicted RNA binding sites and the RNA-protein complex interface. a Human SRP19 in complex with helix 6 of human SRP RNA (PDB code: 1JID). **b** Human SepSecS-tRNASec complex (PDB code: 3HL2). **c** Nova-1 KH1/KH2 domain tandem with RNA hairpin (PDB code: 2ANR). **d** Glutamyl-tRNA synthetase complexed with tRNA (Glu) and glutamol-AMP (PDB code: 1N78). Source data are provided in Supplementary Data 6.



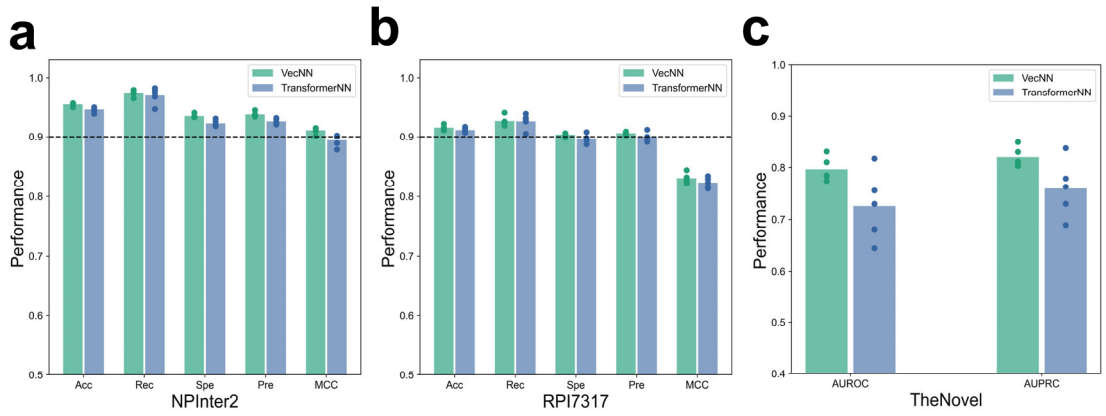
Supplementary Figure 21. Comparison of different aggregation functions in the graph neural network (mean aggregator and max pooling aggregator). a Performance in benchmark NPInter2 in a n=5-fold cross-validation. **b** Performance in benchmark RPI7317 in a n=5-fold cross-validation. **c** Performance in the unknown dataset TheNovel using the n=5 models trained on the NPInter2 dataset. In (a) and (b), the dots in the figures represent the performance of each fold, and the data lines on the radar chart correspond to the mean. In (c), dots in the figure represent the performance of each model, bar height corresponds to the mean. Source data are provided in Supplementary Data 6.



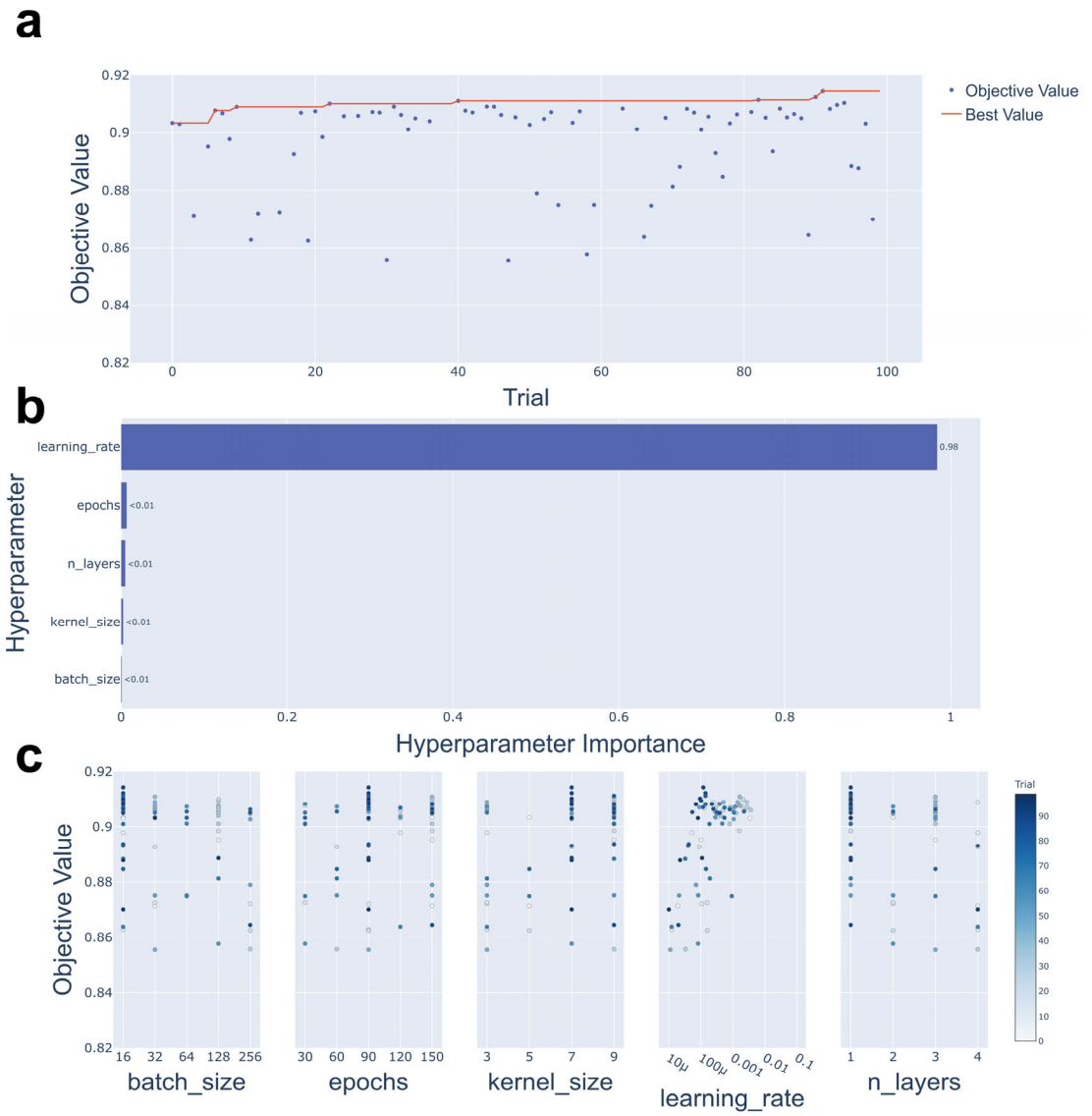
Supplementary Figure 22. Training loss curves for the GNN across five folds. a Fold 1. **b** Fold 2. **c** Fold 3. **d** Fold 4. **e** Fold 5. Source data are provided in Supplementary Data 6.



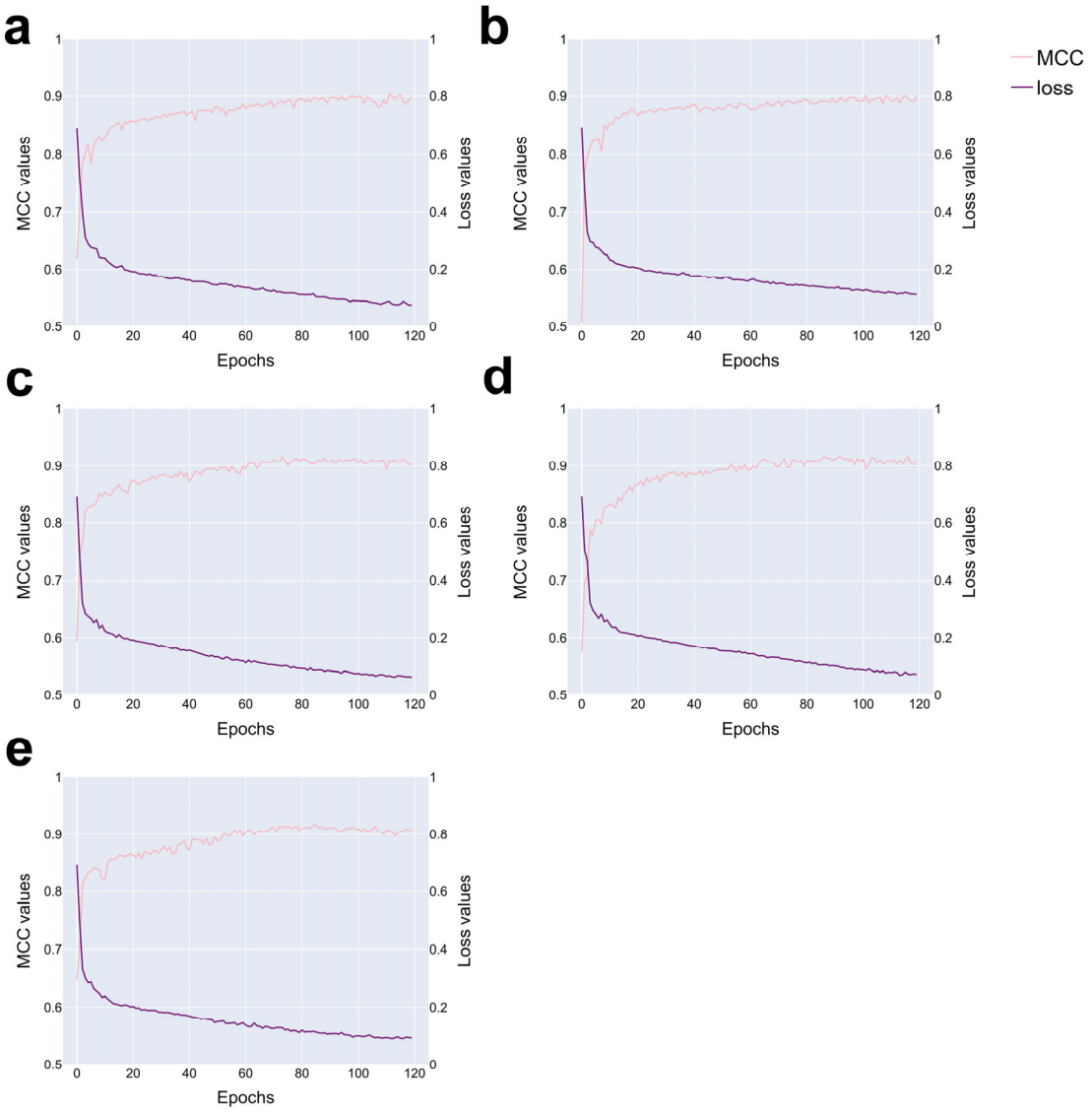
Supplementary Figure 23. Averaged loss curves for the GNN across five folds averaged. The shaded area represents the standard deviation of the loss across n=5 folds. Source data are provided in Supplementary Data 6.



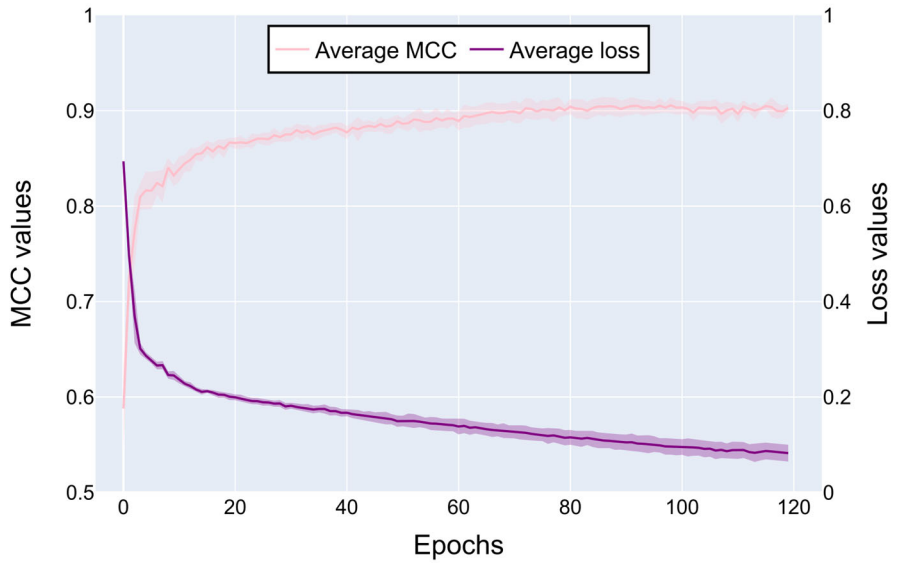
Supplementary Figure 24. Comparison of different feature extraction layers. a Performance in benchmark NPInter2 in a n=5-fold cross-validation. **b** Performance in benchmark RPI7317 in a n=5-fold cross-validation. **c** Performance in the unknown dataset TheNovel using the n=5 models trained on the NPInter2 dataset. In (a) and (b), dots represent the performance of each fold, and bar height corresponds to the mean. In (c), dots represent the performance of each model, and bar height corresponds to the mean. Source data are provided in Supplementary Data 6.



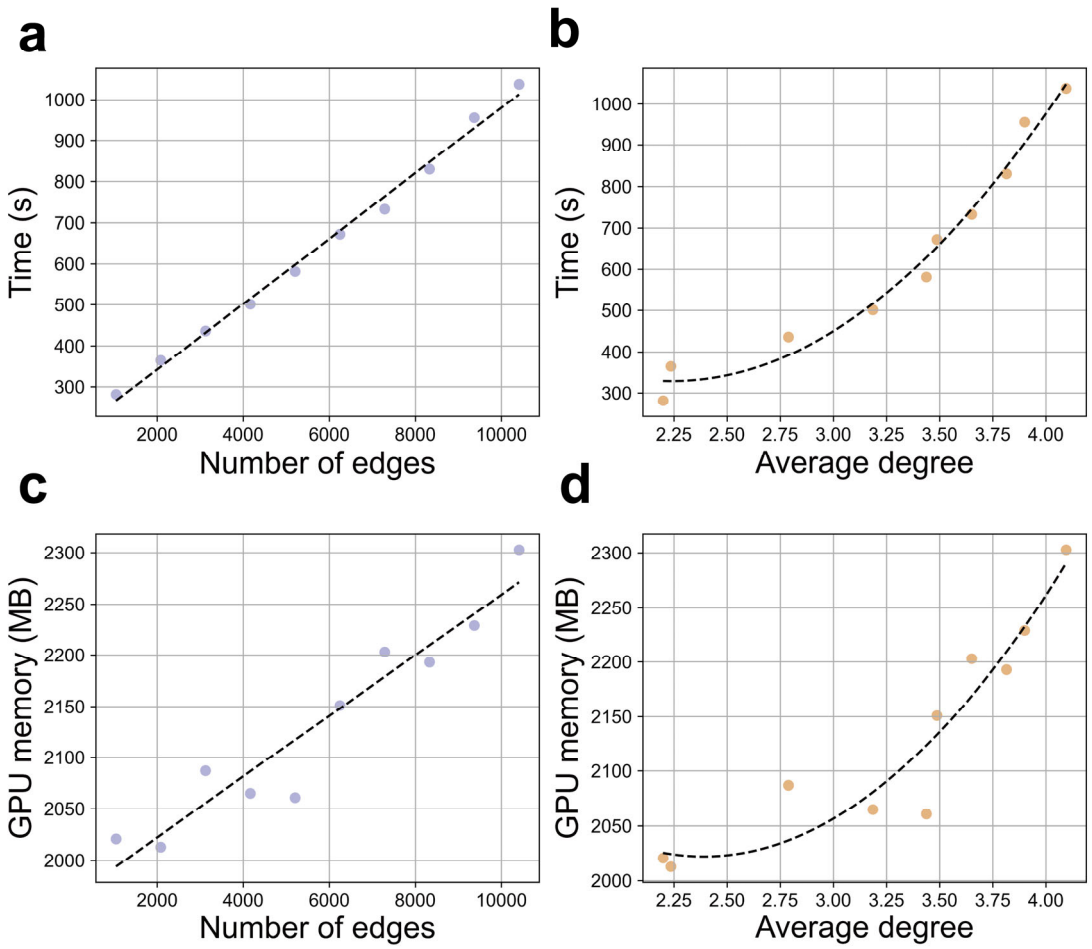
Supplementary Figure 25. Hyperparameter optimization of the neural network architecture. **a** All trials for hyperparameter optimization. **b** Hyperparameter importance estimated by Optuna with different configurations. **c** Plot the parameter relationships as a slice plot to illustrate the performance of each parameter. Source data are provided in Supplementary Data 6.



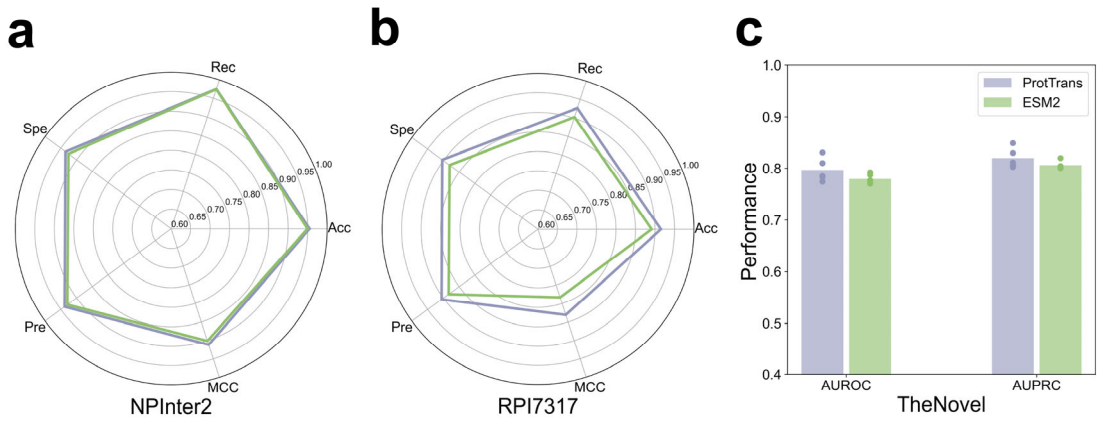
Supplementary Figure 26. Training MCC (pink) and loss (purple) curves for the VecNN architectures across five folds. a Fold 1. **b** Fold 2. **c** Fold 3. **d** Fold 4. **e** Fold 5. Source data are provided in Supplementary Data 6.



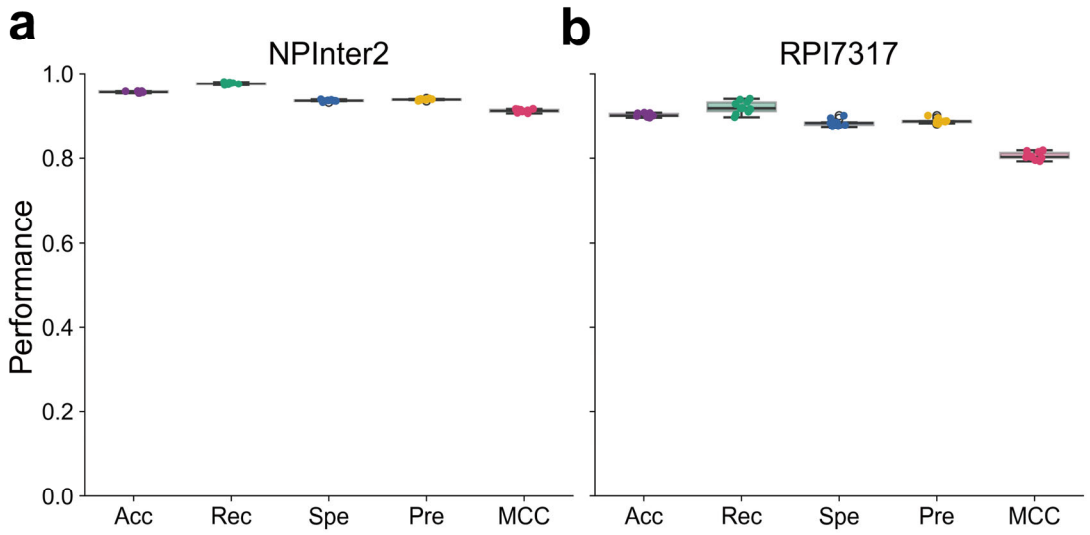
Supplementary Figure 27. Averaged MCC (pink) and loss (purple) curves for the VecNN architectures across five folds averaged. The shaded area represents standard deviation of the MCC and loss across n=5 folds. Source data are provided in Supplementary Data 6.



Supplementary Figure 28. The computational time and memory requirements for training the ZHMolGraph scale with the growth of the RPI network size. The relationship between computational time and **a** the number of edges or **b** the average degree of the RPI network, and the relationship between GPU memory usage and **c** the number of edges or **d** the average degree of the RPI network. Source data are provided in Supplementary Data 6.



Supplementary Figure 29. Comparison of various protein LLMs (ProtTrans and ESM2). **a** Performance in benchmark NPInter2 in a n=5-fold cross-validation. **b** Performance in benchmark RPI7317 in a n=5-fold cross-validation. **c** Performance in the unknown dataset TheNovel using the n=5 models trained on the NPInter2 dataset. In (a) and (b), the dots in the figures represent the performance of each fold, and the data lines on the radar chart correspond to the mean. In (c), dots in the figure represent the performance of each model, bar height corresponds to the mean. Source data are provided in Supplementary Data 6.



Supplementary Figure 30. Performance of n=10 random selections of negative samples on a NPInter2 and **b** RPI7317. Source data are provided in Supplementary Data 6.

Supplementary Tables

Dataset	Methods	Acc	Rec	Spe	Pre	Mcc
NPInter2	ZHMolGraph	**0.955**	**0.975**	**0.935**	**0.938**	**0.911**
	ZHMolGraph_p_LLMs	0.947	0.964	0.931	0.933	0.895
	ZHMolGraph_r_LLMs	0.949	0.969	0.929	0.931	0.898
	ZHMolGraph_n_LLMs	0.942	0.963	0.921	0.924	0.884
	ZHMolGraph_n_GNN	0.943	0.971	0.915	0.919	0.887
RPI7317	ZHMolGraph	**0.915**	**0.927**	**0.904**	**0.906**	**0.831**
	ZHMolGraph_p_LLMs	0.912	0.925	0.900	0.902	0.825
	ZHMolGraph_r_LLMs	0.909	0.923	0.896	0.899	0.819
	ZHMolGraph_n_LLMs	0.907	0.919	0.896	0.898	0.815
	ZHMolGraph_n_GNN	0.897	0.92	0.872	0.88	0.795

Supplementary Table 1. The impact of large language models (LLMs) and graph neural network embeddings on the NPInter2 and RPI7317 benchmark sets. ZHMolGraph_p_LLMs refers to the exclusive use of protein LLMs, ZHMolGraph_r_LLMs indicates the exclusive use of RNA LLMs, ZHMolGraph_n_LLMs denotes the absence of LLMs usage, and ZHMolGraph_n_GNNs refers to the exclusion of graph neural network embeddings. The bold text highlights the best performance among the methods.

Dataset	Methods	AUROC	AUPRC
TheNovel	ZHMolGraph	**0.798**	**0.820**
	ZHMolGraph_p_LLMs	0.636	0.707
	ZHMolGraph_r_LLMs	0.738	0.770
	ZHMolGraph_n_LLMs	0.579	0.669
	ZHMolGraph_n_GNNs	0.777	0.799

Supplementary Table 2. The Impact of Large Language Models (LLMs) and Graph Neural Network Embeddings on the TheNovel set. ZHMolGraph_p_LLMs refers to the exclusive use of protein LLMs, ZHMolGraph_r_LLMs indicates the exclusive use of RNA LLMs, ZHMolGraph_n_LLMs denotes the absence of LLMs usage, and ZHMolGraph_n_GNNs refers to the exclusion of graph neural network embeddings. The bold text highlights the best performance among the methods.

Complex	RNA	Protein	Lowest RMSD (Å)	Lowest TM-score	Average RMSD (Å)
Human SRP19 in complex with helix 6 of human SRP RNA (PDB code: 1JID)	unbound	unbound	**2.64**/4.82	**0.83**/0.79	**9.60**/11.01
Human SepSecS-tRNASec complex (PDB code: 3HL2)	unbound	unbound	**3.71**/5.86	**0.91**/0.88	**13.19**/19.21
Nova-1 KH1/KH2 domain tandem with RNA hairpin (PDB code: 2ANR)	unbound	bound	**0.94**/5.83	**0.98**/0.88	**9.07**/9.46
Glutamyl-tRNA synthetase complexed with tRNA (Glu) and glutamol-AMP (PDB code: 1N78)	bound	unbound	**1.92**/9.73	**0.94**/0.84	**15.32**/18.78
Average performance	N.A.	N.A.	**2.30**/6.56	**0.92**/0.85	**11.80**/14.62

Supplementary Table 3. The performance of unbound RNA-protein complex prediction ranking analysis between 3dRPC and 3dRPC-ZHMolGraph. Bold text highlights the superior result.

Dataset	Protein LLMs	Accuracy	Recall	Specificity	Precision	MCC
NPInter2	ProtTrans	0.955	0.975	0.935	0.938	0.911
	ESM2	0.950	0.974	0.926	0.929	0.901
RPI7317	ProtTrans	0.915	0.927	0.904	0.906	0.831
	ESM2	0.892	0.903	0.88	0.883	0.783

Supplementary Table 4. Comparison of various protein LLMs (ProtTrans and ESM2) in NPInter2 and RPI7317.

Dataset	Protein LLMs	AUROC	AUPRC
TheNovel	ProtTrans	0.798	0.820
	ESM2	0.780	0.807

Supplementary Table 5. Comparison of various protein LLMs (ProtTrans and ESM2) in NPInter2 and RPI7317.

References

[1]T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*). Association for Computing Machinery (2019).

[2]Y. Zhao, Y. Huang, Z. Gong, Y. Wang, J. Man and Y. Xiao. "Automated and fast building of three-dimensional RNA structures," Scientific Reports **2**, 734 (2012).

[3]Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," Science **379**, 1123-1130 (2023).