Science Advances
AAAS

Supplementary Materials for

Deep representation learning of protein-protein interaction networks for enhanced pattern discovery

Rui Yan *et al.*

Corresponding author: Md Tauhidul Islam, tauhid@stanford.edu; Lei Xing, lei@stanford.edu

Sci. Adv. **10**, eadq4324 (2024)
DOI: 10.1126/sciadv.adq4324

This PDF file includes:

Sections S1 to S8
Figs. S1 to S11
Tables S1 and S2

1 Model performance of link prediction across diverse PPI datasets

Figs. S1-S4 below illustrate the model's performance on all four PPI datasets in terms of ROC-AUC, PR-AUC, balanced accuracy, and F1-score, respectively. Figs. S5 and S6 illustrate the corresponding ROC and PR curves for each dataset, respectively. These results demonstrate that DNE surpasses all eleven other network embedding methods across all PPI datasets.

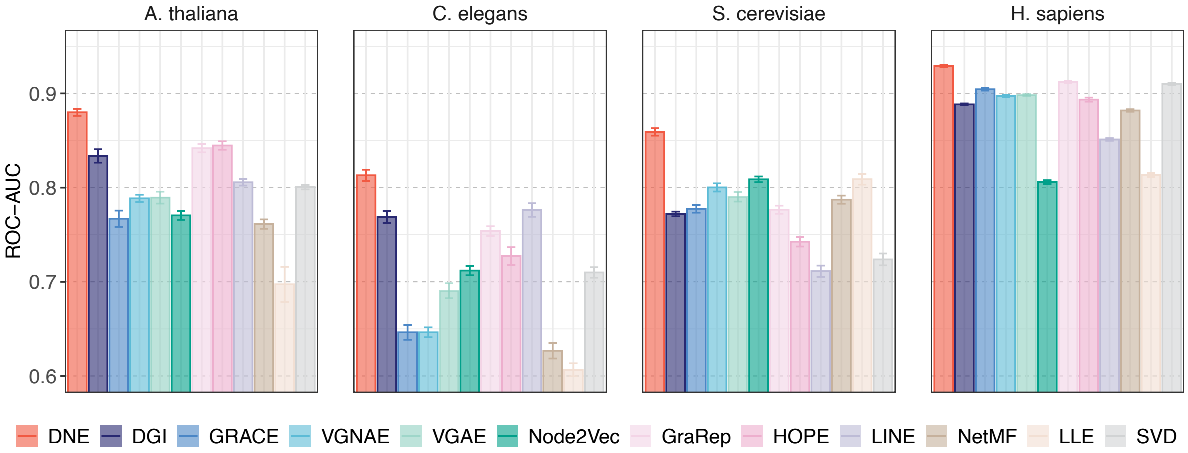

Fig. S1. ROC-AUC scores computed from 10 independent runs across four PPI datasets. Mean values are reported, with error bars representing the standard deviations of the scores.

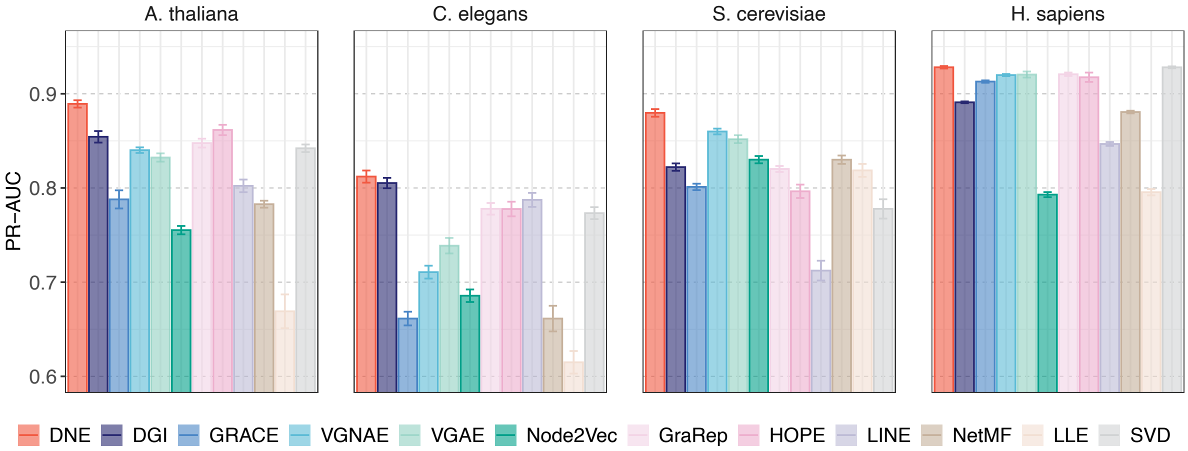

Fig. S2. PR-AUC scores computed from 10 independent runs across four PPI datasets. Mean values are reported, with error bars representing the standard deviations of the scores.

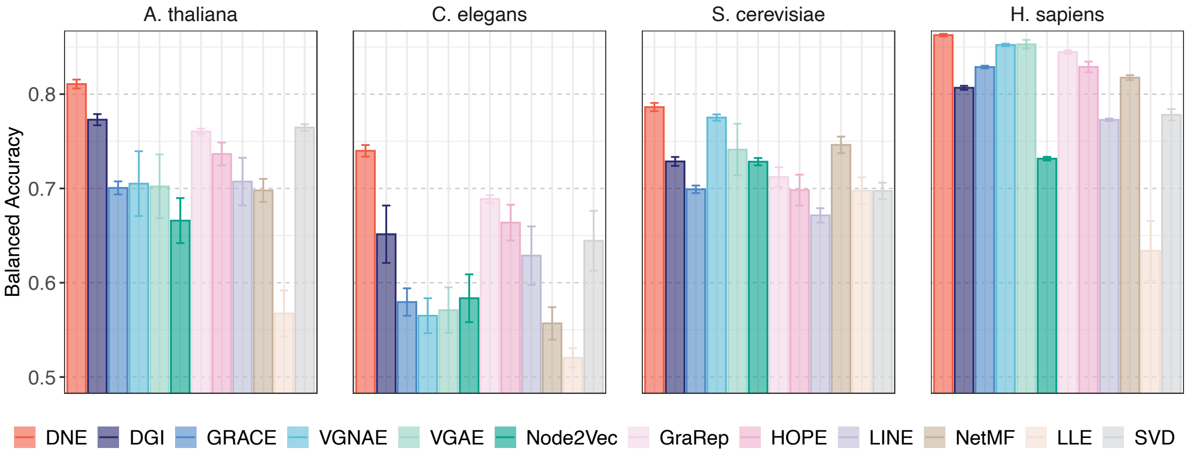

Fig. S3. Balanced accuracy computed from 10 independent runs across four PPI datasets.

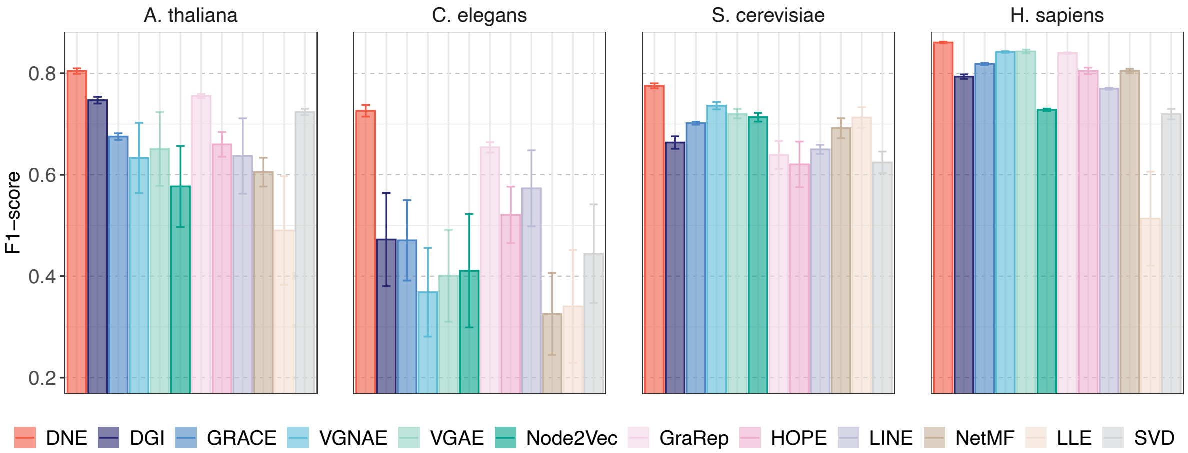

Fig. S4. F1 scores computed from 10 independent runs across four PPI datasets.

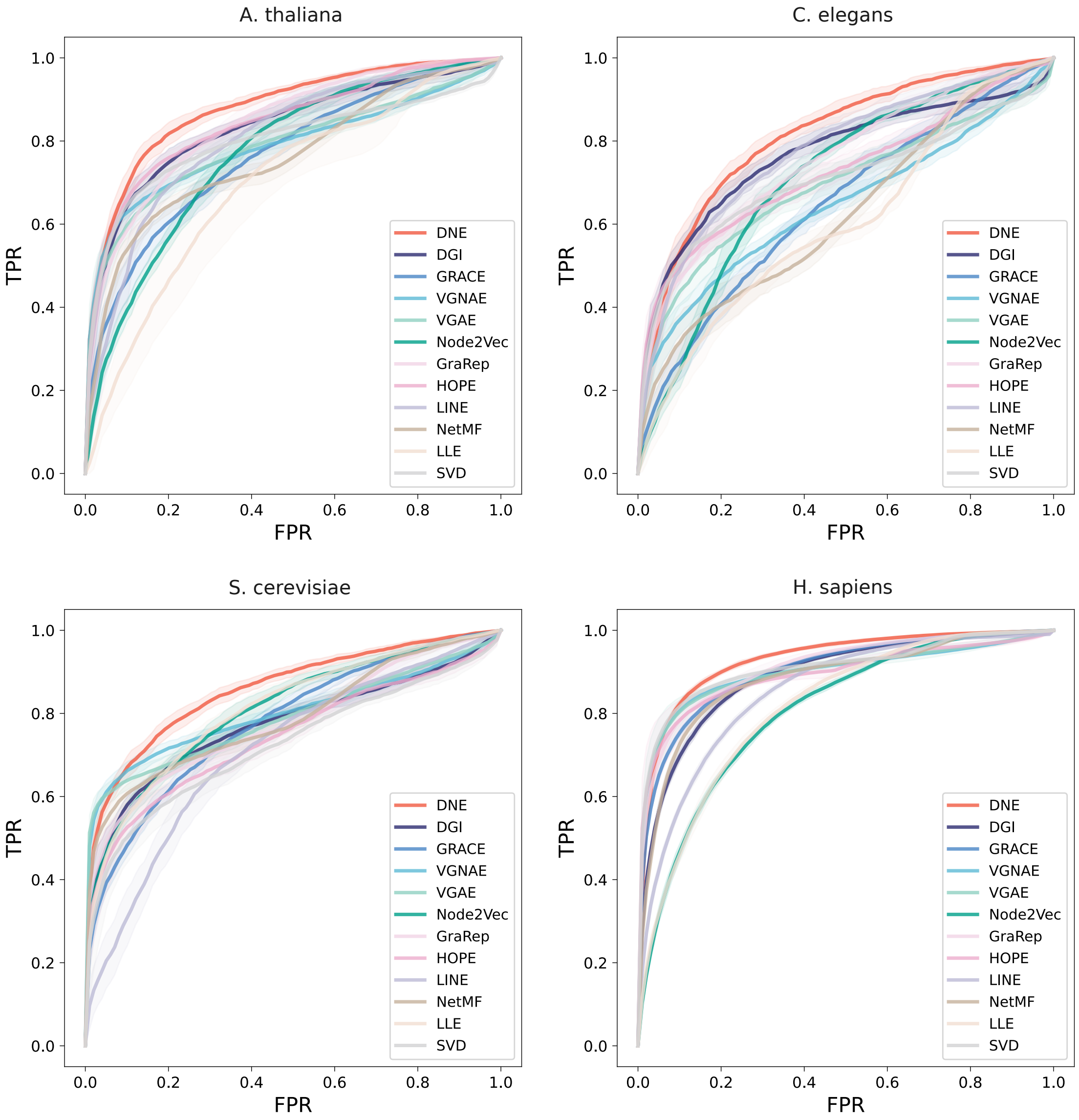

Fig. S5. ROC curves of DNE compared with eleven other network embedding methods for PPI prediction on four PPI datasets.

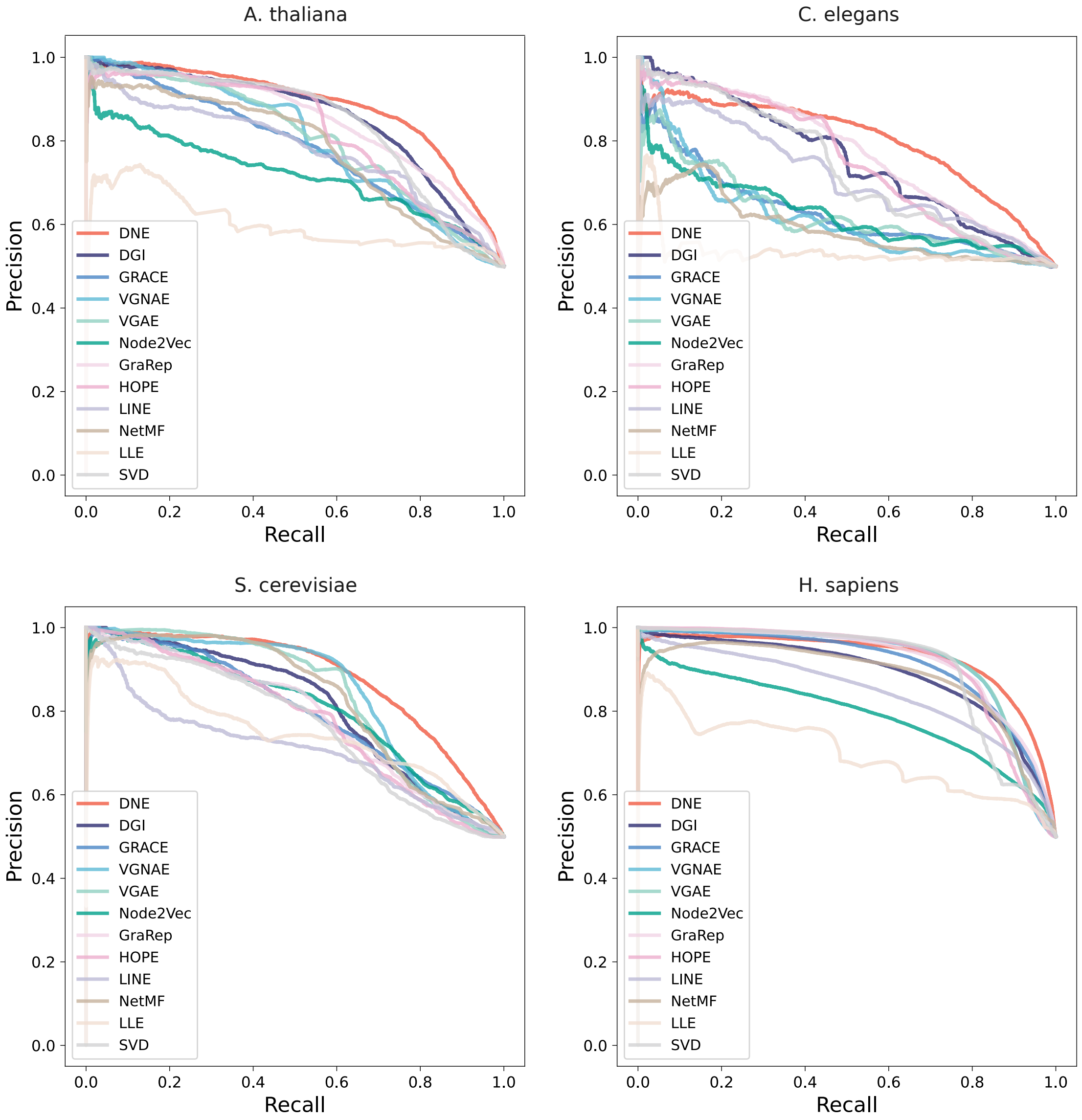

Fig. S6. PR curves of DNE compared with eleven other network embedding methods for PPI prediction on four PPI datasets.

2 Model performance of link prediction on other network datasets

Beyond the four PPI datasets analyzed, we expanded our evaluation of DNE's effectiveness to include three distinct networks: (1) The Cora dataset [44], which is a citation network consisting of 2,708 scientific publications and 5,278 citation links; (2) The Power dataset [45], representing the topology of the Western States Power Grid of the United States with 4,941 nodes and 6,594 edges; and (3) The Router dataset [46], showcasing the router-level architecture of the Internet with 5,022 nodes and 6,258 edges. For details on dataset properties and statistics, refer to Supplementary Section 5. Across these diverse network datasets, the DNE method consistently outperformed competing approaches.

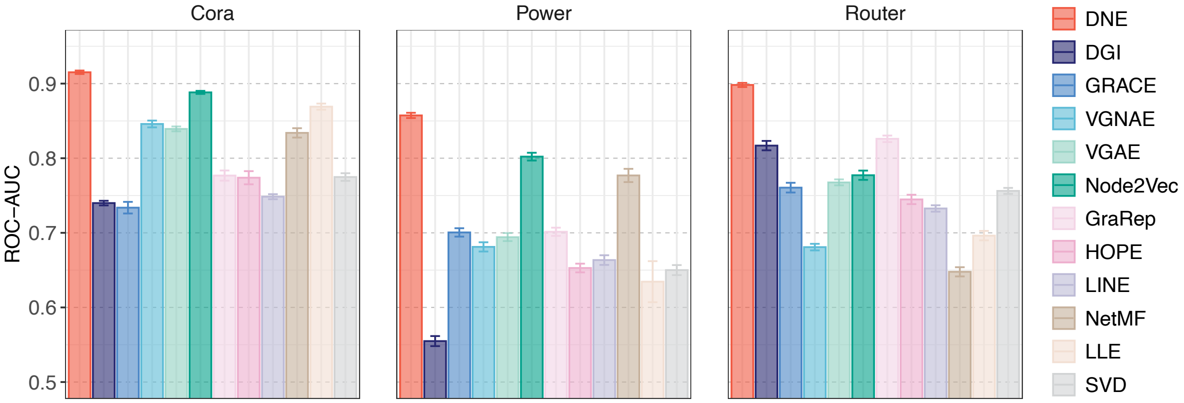

Fig. S7. ROC-AUC scores computed from 10 independent runs across three different network datasets.

3 Model robustness against link perturbations for link prediction

The robustness of the model was evaluated on the A. thaliana dataset by randomly removing varying fractions of edges from the network. DNE exhibits robustness against network perturbations and consistently outperforms other methods across different perturbation ratios (Fig. S8).

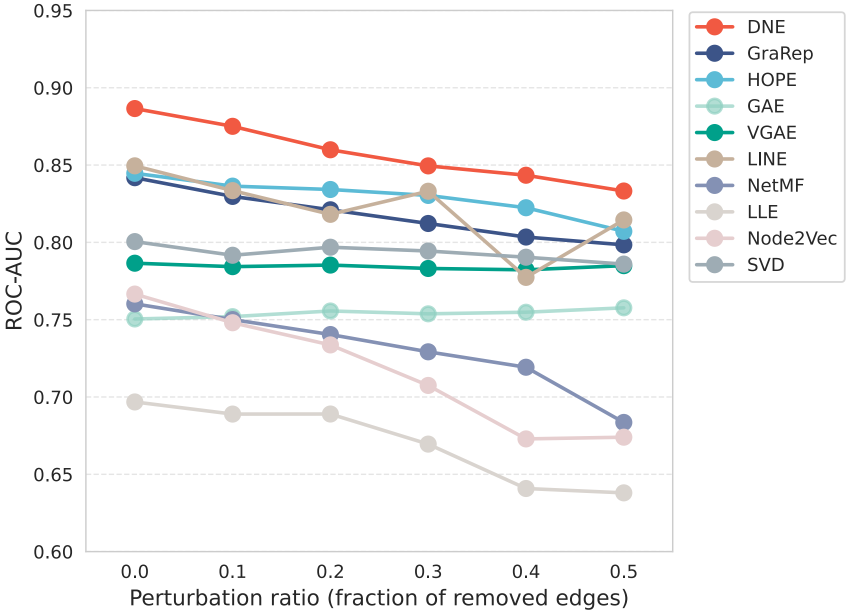

Fig. S8. Robustness evaluation showing the ROC-AUC scores of DNE compared with nine other network embedding methods against link perturbations, where links are randomly removed with different ratios.

4 Model robustness against edge sampling parameters for link prediction

Here we performed an ablation study to examine the effects of edge sampling parameters (walk length l and walk number γ) on link prediction using the S. cerevisiae dataset. In Fig. S9 (left), with the number of walks set at 10, we varied the walk length. The link prediction performance, as measured by ROC-AUC, peaks at $l = 50$ and then declines slightly. This suggests increasing the walk length to a certain threshold captures more neighborhood nodes as positive pairs, but further increases can introduce noise by including distant nodes to positive pairs. In Fig. S9 (right), with the walk length set at 10, we varied the walk number. We observed that the ROC-AUC improves by approximately 1.5% when the number of walks increases within $\gamma < 25$. Beyond this point, the additional gains diminish to less than 0.5% as γ approaches 200. This suggests that increasing the number of walks initially enhances coverage of neighboring nodes effectively for better performance, but further increases lead to redundant sampling of the same neighboring nodes, with minimal further improvements in performance. Overall, the outcome of DNE prediction is not highly sensitive to parameters variations.

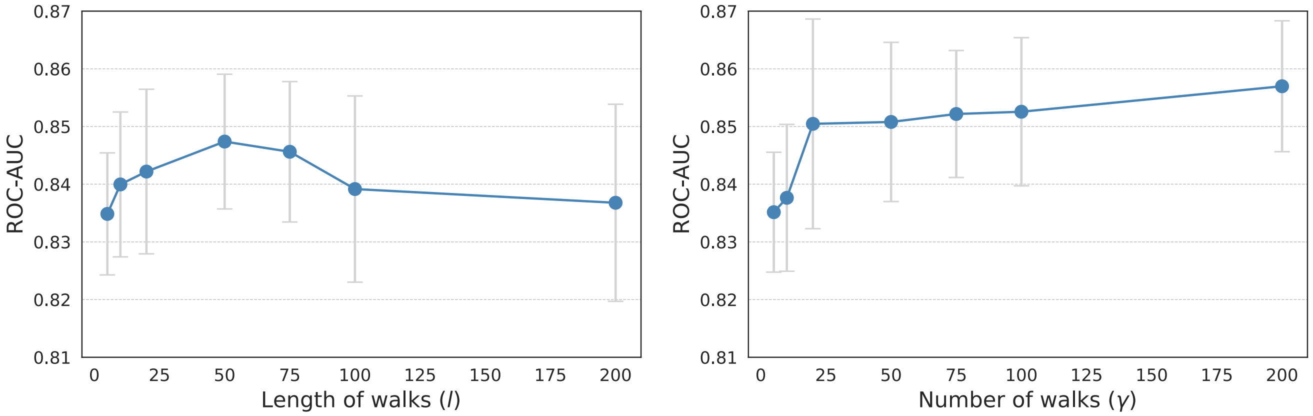

Fig. S9. Sensitivity analysis of walk length l (left) and walk number γ (right) on link prediction performance.

5 Visualization of node embeddings

To validate that DNE (Fig. S10A) effectively captured biologically meaningful signals via its embeddings, we evaluated the correlation between the distances in the embeddings of different proteins and both the functional similarity in terms of GOBP (Fig. S10B) and the proximity within the PPI network (Fig. S10C). Our findings indicate that shorter cosine distances among these embeddings correlate with both greater similarity in GOBP terms and closer connections in the PPI network.

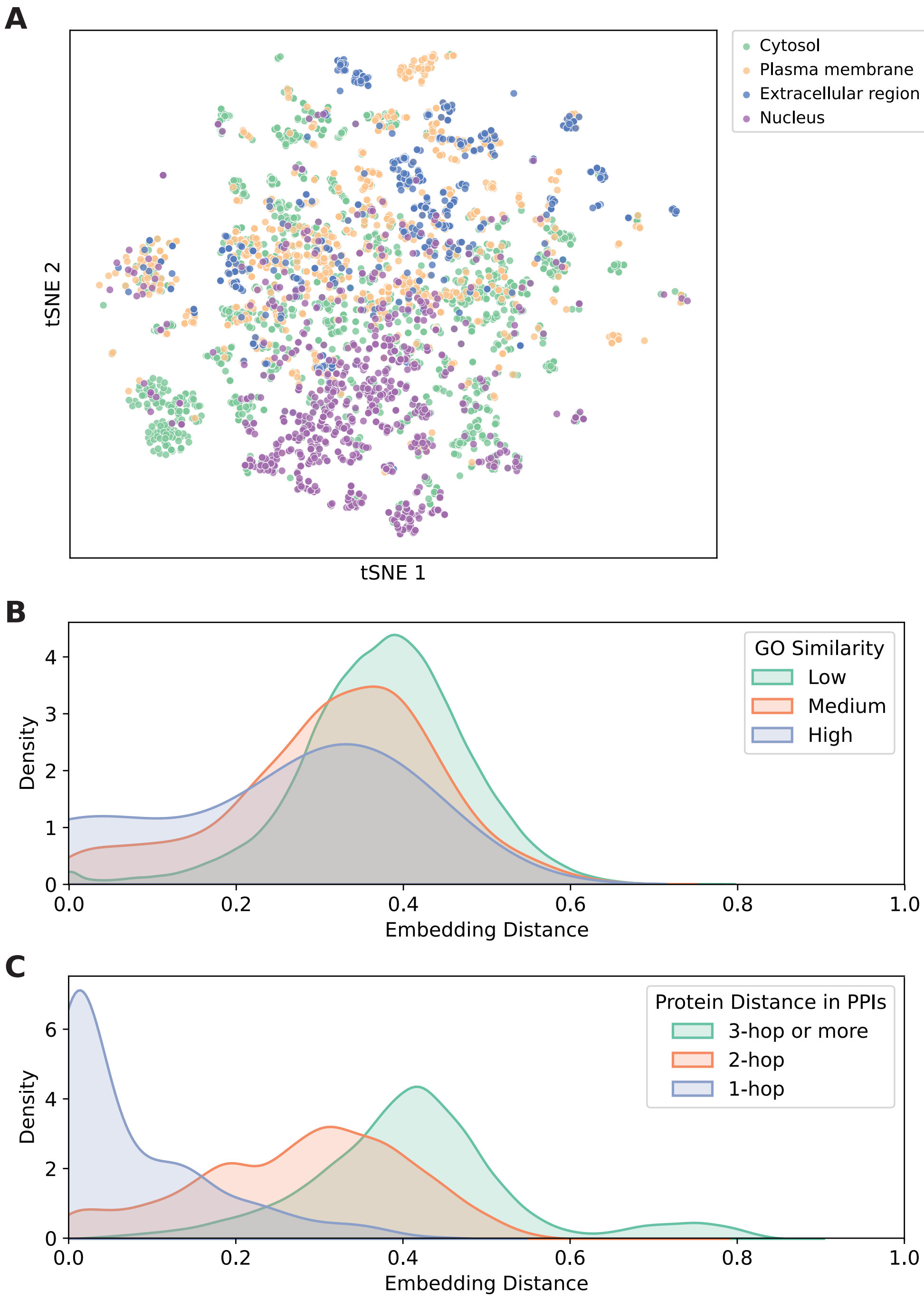

Fig. S10. Visualization of DNE protein embeddings. a) t-SNE projections of protein embeddings obtained from DNE, colored by subcellular locations described by Gene Ontology Cellular Component (GOCC) terms. **c)** Distribution of embedding distances for pairs of proteins sharing low (5th percentile), medium (25th–75th percentile), or high (95th percentile) similarity of GOBP terms, measured by Jaccard similarity. Proteins with higher GOBP similarity have more similar embeddings. **c)** Distribution of embedding distances for pairs of n-hop neighbors in the PPI network. Proteins that are closer together in the network tend to exhibit more similar embeddings.

6 Performance comparison of link prediction methods using node features

Our proposed method uses a dual-encoder mechanism to integrate node features into the embedding learning process. In Fig. 5b of the main manuscript, we assessed our method alongside baseline methods (DGI, GRACE, and VGAE) using the S. cerevisiae dataset, which uniquely provides complete protein sequences for each protein in its PPI network, whereas other PPI datasets in our study lack comprehensive protein features and were not suitable for this analysis. Similarly, we conducted an evaluation on the Cora dataset—a citation network where node features are word vectors describing scientific publications (Fig. S11). DNE consistently outperforms other baseline methods (DGI, GRACE, and VGNAE) in scenarios both with and without node features. By incorporating features, DNE achieved an approximate 2% improvement compared to versions not using features.

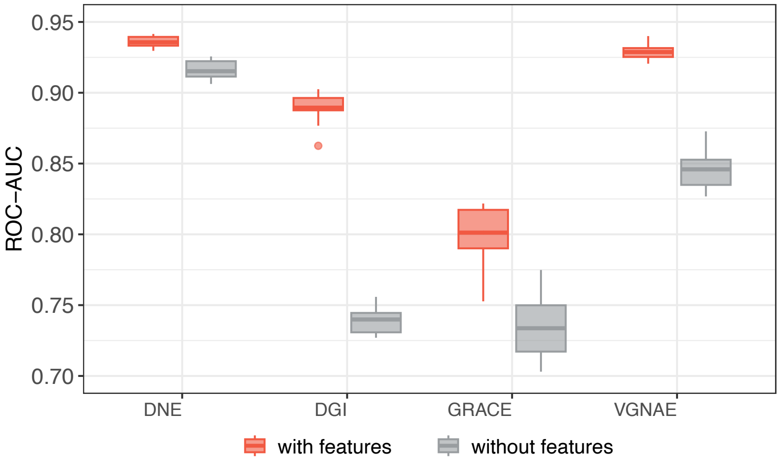

Fig. S11. ROC-AUC scores for DNE and other baseline methods on the Cora dataset, derived from 10 independent runs. Gray boxes indicate cases considering only network structures, while red boxes depict cases incorporating both network structures and node features.

7 Summary of network datasets

Details on the characteristics and statistics of the datasets, such as the number of nodes and edges, average clustering coefficient (ACC), edge density, as well as the average and maximum node degree, are provided in Table S1 below. The ACC calculates network clustering by averaging the individual clustering coefficients of all nodes. It effectively indicates the local connectivity within a social network and the tendency for nodes in a graph to cluster together. Edge density measures the ratio of existing edges to total possible edges in a graph, reflecting network connectivity. The average node degree represents the mean number of edges connected to nodes, providing an overview of network connectivity. Meanwhile, the maximum node degree indicates the highest number of edges connected to a single node.

Table S1. Network dataset characteristics and statistics

Dataset	A. thaliana	C. elegans	S. cerevisiae	H. sapiens	Cora	Power	Router
Num of nodes	2774	2528	2674	8272	2708	4941	5022
Num of edges	6205	3864	7075	52548	5278	6594	6258
ACC	0.049	0.019	0.190	0.059	0.241	0.080	0.012
Density	1.61e-3	1.21e-3	1.98e-3	1.54e-3	1.44e-3	5.40e-4	4.96e-4
Average degree	4.47	3.06	5.29	12.71	3.90	2.67	2.49
Max degree	268	101	140	500	168	19	106

8 Model parameters

The DNE parameters used to reproduce the results are summarized in Table S2 below.

Table S2. DNE model default parameters

Parameter name	Value
Embedding size	128
Walk length	10
Walk number	100
Epochs	10
Learning rate	1e-3
Batch size	1000
Optimizer	Adam
Positional encoding	LE
Dropout rate	0.3
Number of MLP layers	2