Supplementary Materials for MTPrior: A multi-task hierarchical graph embedding framework for prioritizing hepatocellular carcinoma-associated genes and long non-coding RNAs

Fatemeh Keikha[1] and Zhi-Ping Liu[1, *]

[1]Department of Biomedical Engineering, School of Control Science and Engineering,
Shandong University, Jinan, Shandong 250061, China
[*]Shandong National Center for Applied Mathematics, Shandong University, Jinan,
Shandong 250100, China. E-mail:zpliu@sdu.edu.cn

S1. Expanded Background

As outlined in the main manuscript, computational methodologies for identifying and prioritizing disease–gene and lncRNA associations can broadly be categorized into three distinct groups: machine learning-driven methods, network-based approaches, and graph representation learning techniques.

Initial research endeavors harnessed machine learning techniques to discern and prioritize gene-disease associations by extracting salient features from intricate networks [1]. Tran et al. [2] integrated multiple biological data sources into a cohesive network and applied a regularized SVM to predict disease-related genes, while Xu et al. [3], on the other hand, employed a random forest classifier specifically for Alzheimer's disease. The exploration of lncRNA-disease associations through computational models initiated with Chen et al.'s semi-supervised framework (LRLSLDA) [4], which was subsequently succeeded by models such as IPCARF, ILDMSF, and MFLDA. The latter models emphasized the integration of similarities and the reconstruction of association matrices [5, 6, 7]. To transcend MFLDA's constraints, Wang et al. introduced WMFLDA [8] and SelMFDF [9], while other existing models like SIMCLDA [10], and Deng et al.'s Gradient Boosting Tree, and semi-supervised frameworks [11] including LRLSLDA [4] and DNILMF-LDA [12], emerged to identifying disease-associated lncRNAs by integrating multifaceted data sources.

Numerous network-centric models have been deployed to predict disease-gene and lncRNA-disease associations, predominantly relying on random walk algorithms executed on networks grounded in the premise that functionally similar proteins are more densely interconnected within protein-protein interaction (PPI) networks [13]. These models integrate an array of data sources, encompassing gene expression profiles and protein interactions, to prioritize disease-relevant genes and anticipate lncRNA-disease linkages [14, 15]. Strategies like Vavien [13], Arete [14] and SLN-SRW [16] argument predictive capabilities by incorporating random walk with restart, and Laplacian normalization techniques. In the realm of lncRNA-disease prediction, models such as BPLLDA [17] RWRlncD [18], and GrWLDA [19] leverage random walk-based methods, whereas others like MHRWR [20], IRWRLDA [21], BRWLDA [22], and multi-layered network models including LRWRHLDA [23] integrate lncRNA and disease similarity networks with association networks. Despite their notable achievements, integrating heterogeneous similarity data to further refine prediction accuracy remains a formidable challenge.

While network-based methods demonstrate commendable performance, they confront stemming from network topology biases and intricacies in fusing multiple data sources. Matrix completion approaches, presupposing a linear gene-disease relationship, frequently struggle to encapsulate the nonlinear intricacies of reality. Recent strides in graph representation learning have bolstered gene-disease relationship by leveraging network structures [24]. Zhu et al. harnessed graph learning coupled with clustering loss enhance disease-gene predictions [25], whereas Li et al. employed GCNs on distinct networks to refine prioritization [24]. Han et al. integrated GCN with matrix factorization to grasp nonlinear relationships [26], and the PGCN model fused genetic and disease data to amplify accuracy [24]. In the realm of lncRNA-disease prediction, models such as GAIRD [27], HGC-GAN [28], GANLDA [29], CNNDLP [30], and CapsNetLDA [31] utilize diverse techniques, including graph attention, and convolutional networks, to decipher complex patterns within lncRNA-disease networks.

Advancements in graph representation learning have ushered in fresh avenues for dissecting genomes within the biological networks [32]. These methodologies, encompassing node embedding methods techniques DeepWalk [33], and Node2Vec [34], which map nodes into vectorial representations, as well as Graph Convolutional Networks [26] and Graph Attention Networks [35] that use graph neural network architectures for node representation, adeptly encapsulate both qualitative and quantitative attributes of network nodes within a mathematically rigorous framework. Network biology harnesses these embedding models to transform intricate biological data into manageable forms, thereby facilitating similarity searches, clustering, and visualization [36]. Node embedding methods meticulously preserve the relative properties among nodes, ensuring that similar nodes are mirrored similarly in their representations, a cornerstone for effective analysis [32]. Beyond computational efficiency, there embedding techniques also capture functional properties and bolster robustness by mitigating noise [36]. As a result, graph embedding methods streamline, visualize, and enhance the analysis of complex biological networks, rendering them indispensable tools for comprehending and interpreting biological data [32, 36]. Recently, heterogeneous graph embedding has garnered significant attention for its prowess in capturing diverse structural and semantics nuances by embedding complex networks into lower-dimensional spaces. By using meta-paths, this approach delves into the various types of relationships, fostering a deeper appreciation of the diverse local structures inherent in heterogeneous graphs [37].

S2. Algorithms

Algorithm S1 The Overall Process of the Proposed Model

```
Input:
    • The heterogeneous network $G = \{V, E, A, R\}$
    • Node features $h = \{h_{(v_j)}, \forall v_j \in V\}$
    • Node labels $l = \{l_{(v_j)}, \forall v_j \in V\}$
    • The meta-path set $\Phi_{\text{het}} = \{\Phi_1, \ldots, \Phi_p\}$
    • Maximum Meta-path Hops Depth $D$
    • Number of attention heads $M$
Output:
    • Final node embeddings $Z$
    • Final node ranking $R$
$\Phi_{\text{het}}^{\text{New}}, d \leftarrow$ Optimal Multi-Hop Finder Module$(G, h, \Phi_{\text{het}}, D, M)$
$Z \leftarrow$ Multi-hop Embedding Module$(G, h, \Phi_{\text{het}}^{\text{New}}, d, M)$
MLP$(Z, \Phi_{\text{het}}^{\text{New}}, l)$, Predict $f^* = \{f^*_{(v_j)}, \forall v_j \in V\}$
$R \leftarrow$ Assign ranks in descending order of $f^*$
    return $Z, R$
```

Algorithm S2 Optimal Multi-hop Finder Module

```
Input:
    • The heterogeneous network $G$
    • Node features $h$
    • Meta-path set $\Phi_{\text{het}}$
    • Maximum Meta-path Hops Depth $D$
    • Number of attention heads $M$
    • Node labels $l$
Output:
    • New multi-hop meta-path space $\Phi_{\text{het}}^{\text{New}} = \{\Phi^{(\text{hop}_1)}, \ldots, \Phi^{(\text{hop}_d)}\}$
    • Depth $d$
$\Phi^{(\text{hop}_1)} \leftarrow \Phi_{\text{het}}$
$\Phi_{\text{set}} \leftarrow \Phi^{(\text{hop}_1)}$
$\Phi_{\text{het}}^{\text{New}} \leftarrow \Phi_{\text{set}}$
$BestValue \leftarrow \text{Evaluate}(G, h, \Phi^{(\text{hop}_1)}, 1, M, l)$
for $y = 2$ to $D$ do
    for $\Phi_i \in \Phi_{\text{set}}, i \in \{1, \ldots, P\}$ do
        $\Phi_{\text{temp}} \leftarrow \Phi_i^{(\text{hop}_{(y-1)})} + \Phi_i^{(\text{hop}_1)}$
        $TempSubset \leftarrow \Phi_{\text{het}}^{\text{New}} \cup \Phi_{\text{temp}}$
        $Temp \leftarrow \text{Evaluate}(G, h, TempSubset, y, M, l)$
        if $Temp > BestValue$ then
            $\Phi^{(\text{hop}_y)} \leftarrow \Phi^{(\text{hop}_y)} + \Phi_{\text{temp}}$
            $BestValue \leftarrow Temp$
            $d \leftarrow y$
        end if
    end for
end for
    return $\Phi_{\text{het}}^{\text{New}}, d$
```

Algorithm S3 Multi-hop Embedding Module

```
Input:
    • The heterogeneous network $G$, Node features $h$
    • New meta-path set $\Phi_{\text{het}}^{\text{New}}$, Maximum Meta-path Hops Depth $D$
    • Number of attention heads $M$
Output:
    • The final node embedding $Z$
for $\Phi_i^{(\text{hop}_1)} \in \{\Phi_{\text{het}}^{\text{New}}\}, i \in \{1, \dots, q\}$ do
    for $\Phi_i^{(\text{hop}_y)} \in \{\Phi_i^{(\text{hop}_1)}, \dots, \Phi_i^{(\text{hop}_d)}\}, y \leq d$ do
        for $m = 1 \dots M$ do
            for $v_j \in V$ do
                Find meta-path-based neighbors $\Phi_i^{(\text{hop}_y)}$
                for $v_k \in N_{(v_j)}^{(\Phi_i^{(\text{hop}_y)})}$ do
                    Calculate the weight coefficient $\alpha_{(v_j v_k)}^{(\Phi_i^{(\text{hop}_y)})}$
                end for
                Learn node-level embedding:
                    $z_{(v_j)}^{(\Phi_i^{(\text{hop}_y)})} \leftarrow \sigma\left(\sum_{v_k \in N_{(v_j)}^{(\Phi_i^{(\text{hop}_y)})}} \alpha_{(v_j v_k)}^{(\Phi_i^{(\text{hop}_y)})} \cdot h_{(v_k)}\right)$
            end for
        end for
        Concatenate learned embeddings from all attention heads:
            $z_{(v_j)}^{(\Phi_i^{(\text{hop}_y)})} \leftarrow \|_{m=1}^{M} \sigma\left(\sum_{v_k \in N_{(v_j)}^{(\Phi_i^{(\text{hop}_y)})}} \alpha_{(v_j v_k)}^{(\Phi_i^{(\text{hop}_y)})} \cdot h_{(v_k)}\right)$
        Calculate weight $\beta^{(\Phi_i^{(\text{hop}_y)})}$ of information-flow $\Phi_i^{(\text{hop}_y)}$ in $\Phi_i$
    end for
    Learn information-flow level node embedding:
        $Z^{(\Phi_i)} \leftarrow \sum_{y=1}^{d} \beta^{(\Phi_i^{(\text{hop}_y)})} \cdot z_{(v_j)}^{(\Phi_i^{(\text{hop}_y)})}$
    Calculate the meta-path attention weight vector $\gamma^{(\Phi_i)}$
end for
Fuse the meta-path embeddings: Z = $\sum_{i=1}^{q} \gamma^{(\Phi_i)} \cdot Z^{(\Phi_i)}$
Calculate Cross-Entropy loss function: $L = -\sum_{j=1}^{N} Y_j \cdot \log(C \cdot Z_j)$
Backpropagation and update model parameters
    return $Z$
```

S3. Details of highly related methods

- Vavien [13] is a network-based method that leverages the topology of protein-protein interaction networks and employs random walk with restart to prioritize candidate disease genes based on their topological similarity to known disease genes.

- Arete [14] is another network-based method that combines biological network structure with various types of evidence to prioritize candidate genes. It offers flexible and interoperable solutions for enhancing gene prioritization beyond traditional network-based methods, also utilizing random walk with restart.

- netSVM [38] develops an integrated approach using network-constrained support vector machines to identify cancer biomarkers by integrating gene expression data with protein-protein interaction networks.

- Hetio [39] constructs a heterogeneous network and extracts features describing the network topology between genes and diseases. It trains a model on GWAS associations to predict associations between protein-coding genes and complex diseases.

- MALANI [40] integrates gene expression data across multiple cancer types employs SVM models trained on gene-wise and gene-pair interactions. It assesses thousands of genes for their potential roles in cancer networks and uncovers novel candidates associated with cancer outcomes.

- PGCN [24] involves the use of a graph convolutional network to embed a heterogeneous network of genes and diseases, along with their features, into a latent space for further analysis.

- RWRHLD [41] is a rank-based method that utilizes a random walk with restart to prioritize candidate lncRNA-disease associations. The study integrates an lncRNA-lncRNA crosstalk network based on shared miRNA response elements, disease-disease similarity, and known lncRNA-disease association networks into a heterogeneous network.

- RisklncRNAs [42] presents a method for prioritizing risk lncRNAs in cancer by constructing a Gene-LncRNA Co-expression Network, integrating expression and protein interaction data, and applying a random walk algorithm combined with disease phenotype similarity scores.

- TPGLDA [43] is a computational method that predicts lncRNA-disease associations by integrating lncRNA-disease and gene-disease associations into a tripartite graph. It effectively captures the heterogeneity of coding and noncoding gene-disease relationships.

- LRWRHLDA [23] introduces a computational framework for predicting lncRNA-disease associations by constructing isomorphic networks for lncRNA, disease, gene, and miRNA similarities, integrating them into heterogeneous networks, and applying a Laplace normalized random walk with restart algorithm to predict associations.

- GANLDA [29] integrates heterogeneous data of lncRNAs and diseases, applies PCA for noise reduction, utilizes graph attention mechanisms to extract relevant features, and employs a multilayer perceptron for association inference.

- GAIRD [27] employs a hybrid random walk strategy to integrate diverse network information. It includes a module to separate and refine attributes and structures, utilizing techniques like group convolution and deep separable convolution to enhance feature learning.

- LDAGM [44] uses a Graph Convolutional Autoencoder and Multilayer Perceptron to predict lncRNA-disease associations. It fuses similarities from six homogeneous networks into a multi-view heterogeneous network, extracts nonlinear and deep topological features, and employs an enhanced MLP with an aggregation layer for improved prediction accuracy.

- MMHGAN [45] is a deep learning model that predicts lncRNA-disease associations by leveraging hierarchical graphical attention networks. It constructs heterogeneous and homogeneous graphs, uses multihead attention for feature aggregation, and evaluates metapath importance to extract features. Predictions are made by reconstructing these features through a fully connected layer.

The ACC metric indicates that our method accurately classifies genes and lncRNAs related to the disease. The AUROC metric demonstrates its effectiveness in distinguishing between positive and negative samples, while the AUPRC metric confirms that our method predicts a high percentage of true disease-related genes and lncRNAs. The results show that our proposed method achieves the highest scores in ACC, AUROC, and AUPRC metrics for both gene and lncRNA prediction. Specifically, for gene prediction, our method achieves scores of 0.9238, 0.95686, and 0.94675, respectively. For lncRNA prediction, our method achieves scores of 0.89526, 0.93494, and 0.90184, respectively. Additionally, MALANI ranks second in ACC, AUROC, and AUPRC for gene prioritization, with scores of 0.82828, 0.86471, and 0.88534, respectively. For lncRNA prioritization models, GANLDA ranks second in ACC with a score of 0.88666, while RWRHLD achieves the second-highest scores in AUROC and AUPRC, with values of 0.88547 and 0.88356, respectively.

The Matthews Correlation Coefficient (MCC) metric demonstrates that our method effectively aligns predictions with true disease-related genes and lncRNAs. The F1 score further confirms its high precision and recall in predicting these targets. Specifically, our method ranks first in MCC and F1 for gene prioritization, with scores of 0.84761 and 0.92380, respectively. Hetio follows in second place, with scores of 0.7912 and 0.80061, respectively, in these metrics. For lncRNA prediction, our proposed method ranks third in MCC, with a score of 0.76969, while GANLDA and LRWRHLDA secure the first and second ranks with scores of 0.83333 and 0.81818, respectively. However, in the same comparison category, our proposed model achieves the highest score of 0.86484 in F1, ranking first, GAIRD ranks second with a score of 0.84523.

Sensitivity reflects the effectiveness of our method in detecting true disease-related genes and lncRNAs, while specificity demonstrates its proficiency in correctly identifying non-disease-related cases. Our method excels in both gene and lncRNA prioritization models, achieving the highest scores in sensitivity and specificity metrics. Specifically, for gene prediction, our method scores 0.928844 and 0.940369 in sensitivity and specificity, respectively, outperforming Hetio, which ranks second with scores of 0.869047 and 0.870012. In the realm of lncRNA prediction, our model similarity shines, scoring 0.903771 and 0.903816 in sensitivity and specificity, respectively, earning the top spot among models predicting lncRNA-disease associations. LRWRHLDA follows closely in the second place with scores of 0.90123 and 0.901452. Sensitivity underscores the effectiveness of our method in detecting true disease-related genes and lncRNAs, while specificity underscores its proficiency in accurately identifying non-disease-related cases. The results clearly demonstrate that our proposed method holds an advantage over other comparison approaches in predicting disease-related genes and lncRNAs.

The proposed method attains an average AUROC of 0.9459, making a substantial 24.72% improvement over network-based methods that average an AUROC of 0.75841. These network-based methods employ the random walk with restart algorithm on biological network topologies to prioritize disease-related genes or lncRNAs. When compared to ML-based models trained on integrated gene expression data and protein-protein interaction networks, our method also demonstrates a notable performance enhancement of 23.99% in average AUROC. Furthermore, our method surpasses GNN-based models by 10.98%, underscoring the significant impact of integrating heterogeneous data, attributes, and structures with embedding models on enhancing prediction performance.

We attribute the enhanced performance of our method to several pivotal factors. We refine the backbone of the heterogeneous network to facilitate the identification of more pertinent meta-paths associated with the disease, acknowledging that the backbone nodes encompass those not directly targeted in our prediction. Furthermore, beyond merely considering direct meta-path neighbors, we delineate neighbors and their interconnections through multiple iterations of each meta-path, regarding them as multi-hop meta-path neighbors. This strategy delves deeper into the network's structure. Additionally, we deploy a two-layered hierarchical attention model that captures the attention weights of various nodes, different multi-hop meta-paths, thereby augmenting the method's precision in prioritizing disease-related nodes.

S4. Experimental results with varying sample ratios

Experiments were conducted utilizing diverse ratios of positive to negative samples to investigate their influence on the proposed method for prioritizing genes and lncRNAs. Notably, variations in the ratios between positive and negative samples had a substantial impact on the method's performance.

To assess the model's capacity to classify and prioritize genes and lncRNAs using different sample ratios, multiple experimental datasets were generated. Each dataset consists all positive samples coupled with varying quantities of negative samples: an equivalent number, five times, or ten times the amount of negative samples compared to positives. These curated datasets were employed to evaluate the prediction model's performance in relation to both genes and lncRNAs associated with HCC.

Figure 1 and Table 1 present the results, demonstrating that the proposed method attains average ACC, MCC, and F1 scores of 92.38%, 84.76%, and 92.38%, respectively, for predicting disease genes, and 89.52%, 76.96%, and 86.48%, respectively, for predicting HCC-related lncRNAs, under an equivalent ratio of positive to negative samples. Notably, our method MTPrior achieves the highest Sensitivity and Specificity values with 1:1 ratio of negative to positive samples, particularly excelling in gene prioritization tasks. Additionally, MTPrior demonstrates superior performance in terms of AUROC and AUPRC metrics with a 1:1 ratio for both gene and lncRNA classification, achieving values of 0.9568 and 0.9349 for AUC, and 0.9467 and 0.9018 for AUPRC in gene and lncRNA prediction, respectively. However, for Accuracy, the highest scores are achieved with a 1:5 ratio, yielding 0.9298 for gene prioritization and 0.9025 for lncRNA prioritization. The high AUC and AUPRC scores demonstrates the effectiveness of the proposed method in accurately distinguishing between positive and negative samples. The minimal variation in results across different ratios of positive to negative samples underscores the method's stability and reliability.

Table 1: **Results of the proposed methods under varying ratios (Positive Samples : Negative Samples = 1:1, 1:5, and 1:10).**

Method	Ratio	ACC	AUROC	AUPRC	MCC	F1	Sensitivity	Specificity
Gene prioritizing	1:1	**0.9238**	**0.9569**	**0.9468**	**0.8476**	**0.9238**	**0.9288**	**0.9404**
	1:5	0.9299	0.9251	0.9353	0.7500	0.8742	0.8355	0.8410
	1:10	0.8941	0.8960	0.8940	0.8004	0.8998	0.7485	0.7483
lncRNA prioritizing	1:1	**0.8953**	**0.9349**	**0.9018**	**0.7697**	**0.8648**	**0.9038**	**0.9038**
	1:5	0.9025	0.9021	0.8266	0.6977	0.8423	0.9023	0.9023
	1:10	0.8788	0.8636	0.8829	0.7736	0.8629	0.8674	0.8694

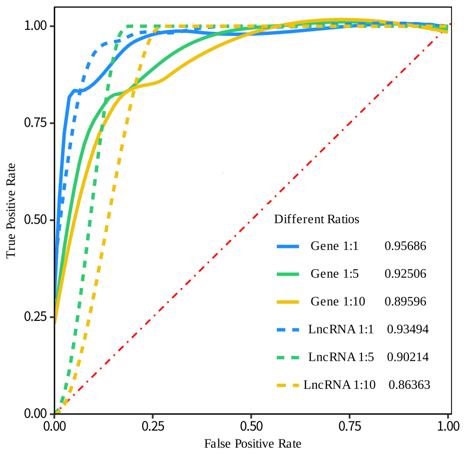

Figure 1: **AUC values for experimental results using different ratios of positive to negative samples.** Solid lines depict the gene identification task, while dashed lines represent the lncRNA identification task. The blue curves indicate a 1:1 sample ratio, green curves indicate a 1:5 sample ratio, and yellow curve indicates a 1:10 sample ratio.

S5. Evaluation metrics across varying embedding sizes

We examine the influence of embedding dimension on our proposed method MTPiror, with the outcomes presented in Fig 2. The embedding sizes considered here encompass 32, 64, 128, 256, 512, and 1024, respectively. Across all evaluation metrics, our method demonstrates optimal performance in gene and lncRNA prioritization when the embedding dimensions are set at 256 and 512. Specifically, for gene prioritization, the performance of our method enhances as the embedding dimension increases from 32 to 256, while for lncRNA prioritization, it improves from 32 to 512. Nevertheless, a decline in performance is observed when the embedding dimension is escalated to 1024. Consequently, when utilizing our proposed method, we select an embedding size of 256 for gene prioritization and 512 for lncRNA prioritization.

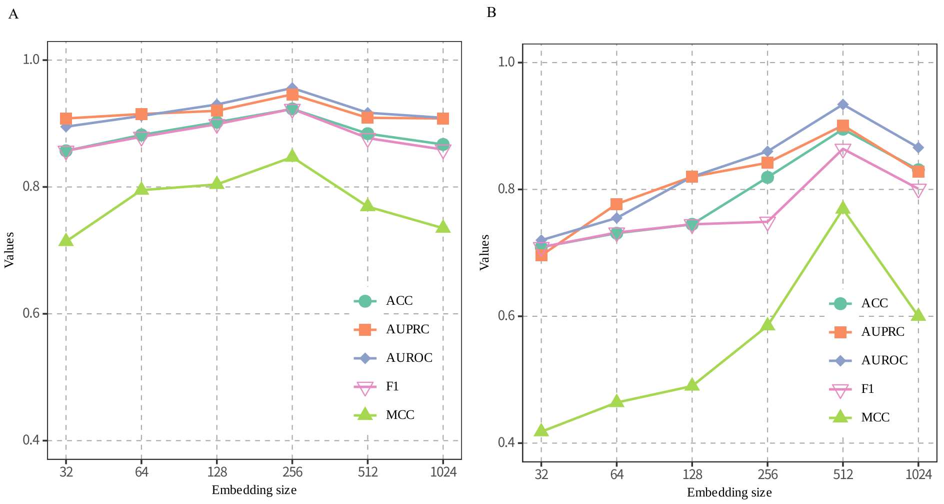

Figure 2: **Evaluation metrics across various embedding dimensions.** This figure contrasts the results of evaluation metrics (ACC, AUROC, AUPRC, F1, MCC) for different embedding sizes in the context of (A) gene identification and (B) lncRNA identification tasks. Each line depicts the performance of a particular evaluation metric as the embedding size varies from 32 to 1024.

S6. Evaluation metrics across different classification methods

In this experiment, we evaluated the performance of various classifiers by comparing a MLP with several other models, namely SVM, Random Forest, logistic regression, Adaboost, Naïve Bayes, and XGBoost. We employed five-fold cross-validation across multiple evaluation criteria to ensure a comprehensive assessment. To maintain fairness and reliability in our comparison, all experimental settings were kept consistent, with the exception of the classifier being tested. Our results, as illustrated in Fig 3, demonstrate that in the context of gene and lncRNA prioritization, the MLP consistently outperforms the other models across nearly all metrics. The only exception is in the AUPRC metric for the lncRNA experiment, where Random Forest achieved a marginally superior result (0.90298) compared to MLP (0.90184).

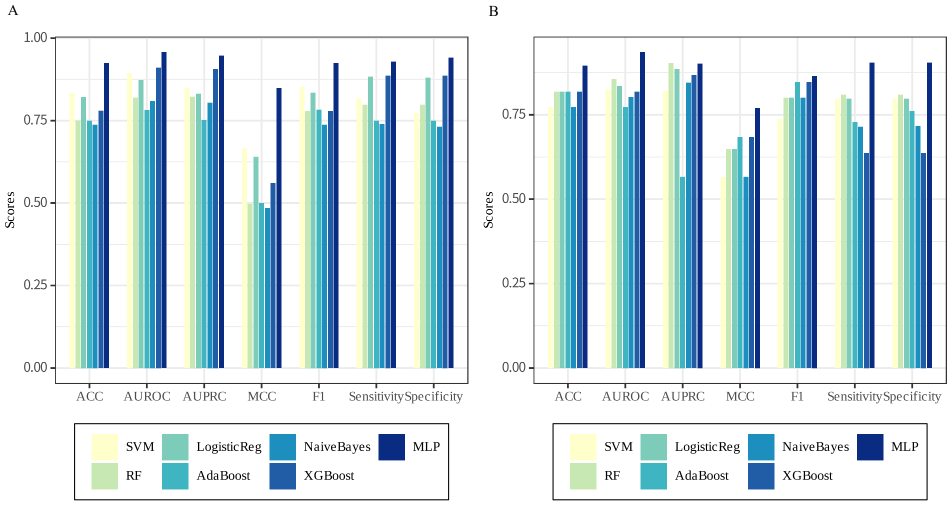

Figure 3: **Evaluation metrics across different classification methods.** This figure compares the different classifier's performance including SVM, Random Forest, logistic regression, Adaboost, Naïve Bayes, XGBoost, and MLP by several evaluation metrics (ACC, AUROC, AUPRC, F1, MCC) for (A) gene identification task, and (B) lncRNA identification task.

References

[1] Li Y, Guo Z, Wang K, Gao X, Wang G. End-to-end interpretable disease-gene association prediction. Briefings in bioinformatics. 2023; 24(3). pmid: 36987781.

[2] Tran VD, Sperduti A, Backofen R, Costa F. Heterogeneous networks integration for disease-gene prioritization with node kernels. Bioinformatics. 2020; 36(9):2649-2656. pmid: 31990289.

[3] Xu L, Liang G, Liao C, Chen GD, Chang CC. k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification. Frontiers in genetics. 2019; 10:33. pmid: 30809242.

[4] Chen X, Yan GY. Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics. 2013;29(20):2617-24. pmid: 24002109.

[5] Zhu R, Wang Y, Liu JX, Dai LY. IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier. BMC Bioinformatics. 2021;22(1):175.pmid: 33794766.

[6] Chen Q, Lai D, Lan W, Wu X, Chen B, Liu J, Chen YP, Wang J. ILDMSF: Inferring Associations Between Long Non-Coding RNA and Disease Based on Multi-Similarity Fusion. IEEE/ACM transactions on computational biology and bioinformatics. 2021; 18(3):1106-1112. pmid: 31443046.

[7] Fu G, Wang J, Domeniconi C, Yu G. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. Bioinformatics. 2018; 34(9):1529-1537. pmid: 29228285.

[8] Wang Y, Yu G, Wang J, Fu G, Guo M, Domeniconi C. Weighted matrix factorization on multi-relational data for LncRNA-disease association prediction. Methods. 2020;173:32-43. pmid: 31226302.

[9] Wang Y, Yu G, Domeniconi C, Wang J, Zhang X, Guo M. Selective matrix factorization for multi-relational data fusion. Proceedings of the International Conference on Database Systems for Advanced Applications; 2019. p. 313–329.

[10] Lu C, Yang M, Luo F, Wu FX, Li M, Pan Y, Li Y, Wang J. Prediction of lncRNA-disease associations based on inductive matrix completion. Bioinformatics. 2018; 34(19):3357-3364. pmid: 29718113.

[11] Deng L, Li W, Zhang J. LDAH2V: Exploring Meta-Paths Across Multiple Networks for lncRNA-Disease Association Prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021; 18(4):1572-1581. pmid: 31725386.

[12] Li Y, Li J, Bian N. DNILMF-LDA: Prediction of lncRNA-Disease Associations by Dual-Network Integrated Logistic Matrix Factorization and Bayesian Optimization. Genes. 2019; 10(8):608. pmid: 31409034.

[13] Erten S, Bebek G, Koyutürk M. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. Journal of computational biology. 2011; 18(11):1561-1574. pmid: 22035267.

[14] Lysenko A, Boroevich KA, Tsunoda T. Arete - candidate gene prioritization using biological network topology with additional evidence types. BioData mining. 2017; 10:1-12. pmid: 28694847.

[15] Sheng N, Wang Y, Huang L, Gao L, Cao Y, Xie X, Fu Y. Multi-task prediction-based graph contrastive learning for inferring the relationship among lncRNAs, miRNAs and diseases. Briefings in bioinformatics. 2023; 24(5). pmid: 37529914.

[16] Peng J, Bai K, Shang X, Wang G, Xue H, Jin S, Cheng L, Wang Y, Chen J. Predicting disease-related genes using integrated biomedical networks. BMC Genomics. 2017; 18:1043. pmid: 28198675.

[17] Xiao X, Zhu W, Liao B, Xu J, Gu C, Ji B, Yao Y, Peng L, Yang J. BPLLDA: Predicting lncRNA-Disease Associations Based on Simple Paths With Limited Lengths in a Heterogeneous Network. Frontiers in genetics. 2018;9:411. pmid: 30459803.

[18] Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S, Zhou M. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Molecular BioSystems. 2014; 10(8):2074-81. pmid: 24850297.

[19] Gu C, Liao B, Li X, Cai L, Li Z, Li K, Yang J. Global network random walk for predicting potential human lncRNA-disease associations. Scientific reports. 2017; 7(1):12442. pmid: 28963512.

[20] Zhao X, Yang Y, Yin M. MHRWR: Prediction of lncRNA-Disease Associations Based on Multiple Heterogeneous Networks. IEEE/ACM transactions on computational biology and bioinformatics. 2021; 18(6):2577-2585. pmid: 32086216.

[21] Chen X, You ZH, Yan GY, Gong DW. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget. 2016; 7(36):57919-57931. pmid: 27517318

[22] Yu G, Fu G, Lu C, Ren Y, Wang J. BRWLDA: bi-random walks for predicting lncRNA-disease associations. Oncotarget. 2017;8(36):60429-60446. pmid: 28947982.

[23] Wang L, Shang M, Dai Q, He PA. Prediction of lncRNA-disease association based on a Laplace normalized random walk with restart algorithm on heterogeneous networks. BMC Bioinformatics. 2022;23(1):5. pmid: 34983367.

[24] Li Y, Kuwahara H, Yang P, Song L, Gao X. PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks. biorxiv. 2019; 532226. doi: https://doi.org/10.1101/532226.

[25] Zhu L, Hong Z, Zheng H. Predicting gene-disease associations via graph embedding and graph convolutional networks. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019; pp. 382-389.

[26] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016. Available from: https://doi.org/10.48550/arXiv.1609.02907

[27] Wang S, Hui C, Zhang T, Wu P, Nakaguchi T, Xuan P. Graph Reasoning Method Based on Affinity Identification and Representation Decoupling for Predicting lncRNA-Disease Associations. Journal of Chemical Information and Modeling. 2023; 63(21):6947-6958. pmid: 37906529.

[28] Zhang W, Wei H, Zhang W, Wu H, Liu B. Multiple types of disease-associated RNAs identification for disease prognosis and therapy using heterogeneous graph learning. Science China Information Sciences. 2024; 67(8):1-2.

[29] Lan W, Wu X, Chen Q, Peng W, Wang J, Chen YP. GANLDA: Graph attention network for lncRNA-disease associations prediction. Neurocomputing. 2022; 469:384–393.

[30] Xuan P, Sheng N, Zhang T, Liu Y, Guo Y. CNNDLP. CNNDLP: A method based on convolutional autoencoder and convolutional neural network with adjacent edge attention for predicting lncRNA–disease associations. International Journal of Molecular Sciences, 2019; 20(17):4260. pmid: 31480319.

[31] Zhang Z, Xu J, Wu Y, Liu N, Wang Y, Liang Y. CapsNet-LDA: predicting lncRNA-disease associations using attention mechanism and capsule network based on multi-view data. Briefings in Bioinformatics. 2023; 24(1). pmid: 36511221.

[32] Lagisetty Y, Bourquard T, Al-Ramahi I, Mangleburg CG, Mota S, Soleimani S, Shulman JM, Botas J, Lee K, Lichtarge O. Identification of risk genes for Alzheimer's disease by gene embedding. Cell Genom. 2022; 2(9):100162. pmid: 36268052.

[33] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. InProceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 2014; pp. 701-710.

[34] Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 2016:855-864. pmid: 27853626.

[35] Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. stat. 2017; 1050(20):10-48550.

[36] Nelson W, Zitnik M, Wang B, Leskovec J, Goldenberg A, Sharan R. To Embed or Not: Network Embedding as a Paradigm in Computational Biology. Front Genet. 2019 ; 10:381. pmid: 31118945.

[37] Wang X, Bo D, Shi C, Fan S, Ye Y, Philip SY. A survey on heterogeneous graph embedding: methods, techniques, applications and sources. IEEE Transactions on Big Data. 2022; 9(2):415-36.

[38] Chen L, Xuan J, Riggins RB, Clarke R, Wang Y. Identifying cancer biomarkers by network-constrained support vector machines. BMC systems biology. 2011; 5:1-20. pmid: 21992556.

[39] Himmelstein DS, Baranzini SE. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. PLoS computational biology. 2015; 11(7). pmid: 26158728.

[40] Ghanat Bari M, Ung CY, Zhang C, Zhu S, Li H. Machine learning-assisted network inference approach to identify a new class of genes that coordinate the functionality of cancer networks. Scientific Reports, 2017; 7(1):6993. pmid: 28765560.

[41] Zhou M, Wang X, Li J, Hao D, Wang Z, Shi H, Han L, Zhou H, Sun J. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. Molecular BioSystems. 2015; 11(3):760-9. pmid: 25502053.

[42] Xu C, Qi R, Ping Y, Li J, Zhao H, Wang L, Du MY, Xiao Y, Li X. Systemically identifying and prioritizing risk lncRNAs through integration of pan-cancer phenotype associations. Oncotarget. 2017; 8(7):12041-12051. pmid: 28076842.

[43] Ding L, Wang M, Sun D, Li A. TPGLDA: Novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph. Sci Rep. 2018 Jan 18;8(1):1065. doi: 10.1038/s41598-018-19357-3. PMID: 29348552

[44] Zhang B, Wang H, Ma C, Huang H, Fang Z, Qu J. LDAGM: prediction lncRNA-disease asociations by graph convolutional auto-encoder and multilayer perceptron based on multi-view heterogeneous networks. BMC Bioinformatics. 2024 Oct 15;25(1):332. PMID: 39407120.

[45] Yao D, Deng Y, Zhan X, Zhan X. Predicting lncRNA-disease associations using multiple metapaths in hierarchical graph attention networks. BMC Bioinformatics. 2024 Jan 29;25(1):46. PMID: 38287236.