ACS OMEGA

Open Access



http://pubs.acs.org/journal/acsodf

Article

# GCN-Based Heterogeneous Complex Feature Learning to Enhance Predictability for LncRNA−Disease Associations

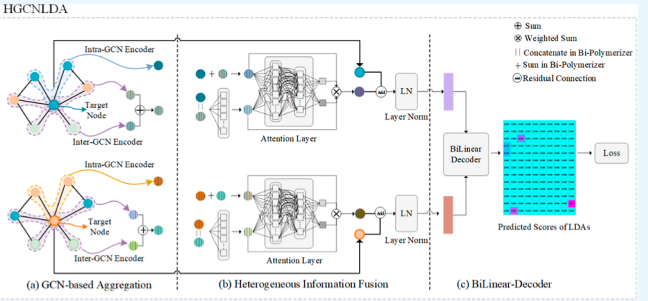
Yi Zhang,* Gangsheng Cai, Xin Li, and Min Chen

Cite This: *ACS Omega* 2024, 9, 1472−1484

Read Online

ACCESS | Metrics & More | Article Recommendations

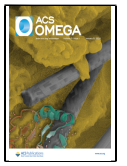
**ABSTRACT:** Using computational models to predict potential lncRNA-disease associations (LDAs) has emerged as an effective supplement to bioexperiments for exploring the pathogenesis of diseases. However, current computational models still face limitations in their ability to learn the complex features of bionetworks. In this study, HGCNLDA, a model which combines graph convolutional network (GCN)-based aggregation, heterogeneous information fusion, and a bilinear-decoder to infer LDAs was proposed. Recognizing the need to extract essential features during data processing, our HGCNLDA explored four key steps for uncovering interaction patterns within the bionetwork: (1) a novel type of tripartite heterogeneous network, known as the lncRNA-disease-miRNA network (LDMN), was constructed using computed similarities and known associations. (2) Homogeneous and heterogeneous features of nodes were extracted from domains within the LDMN by a GCN-based encoder. (3) Feature fusions, including bipolymerization operations and attention mechanism, were employed to capture a more accurate and comprehensive representation of nodes. (4) Bilinear-decoder was used to rebuild the edge type (or rating type) for a specific node pair, resulting in the predicted association score. Through a 5-fold cross-validation on two data sets, namely, data set1 and data set2, our HGCNLDA consistently demonstrated superior performance compared to five related models. It almost achieved the highest AUROC and AUPR values on both data sets, especially on data set2 where the results obtained were more challenging and objective. Case studies involving three real cancer scenarios further validated the practicality of HGCNLDA in identifying potential LDAs in real-world contexts. The source code and data for this study are available at https://github.com/zywait/HGCNLDA.




## 1. INTRODUCTION

Long noncoding RNAs (lncRNAs), which have been arbitrarily defined as transcripts containing more than 200 nucleotides (200 nt), play crucial roles as regulators of gene expression and are involved in diverse biological processes. LncRNAs exhibit several characteristics with protein-coding genes, including promoters, multiple exons, alternative splicing, characteristic chromatin signatures, regulation by morphogens and conventional transcription factors, and altered expression in various diseases, including cancer.[1] The potential of lncRNAs as diagnostic and prognostic biomarkers, as well as therapeutic targets, has garnered significant attention. LncRNAs have great potential to be diagnostic and prognostic biomarkers and therapeutic targets.[2,3] Identifying lncRNA-disease associations (LDAs) is significant for disease prevention, diagnosis, treatment, and prognosis, especially for cancer. While in vivo or in vitro experiments can provide insights into specific LDAs and the pathogenic mechanisms of lncRNAs, conducting traditional low-throughput biological experiments can be a time-consuming, expensive, and inefficient process, especially when dealing with tens of thousands of lncRNAs with unknown functions. In recent years, high-throughput technologies such as microarrays and next-generation sequencing have emerged, allowing for the identification of a large number of dysregulated lncRNAs associated with diseases. However, the results from high-throughput technologies often contain significant noise, and most of the dysregulated lncRNAs identified may not be directly related to the causal lncRNAs responsible for the associated diseases.[4] With large-scale available biological databases being set up, such as LncRNADisease,[5] Lnc2Cancer,[6] HMDD,[7] computer-aided inference of disease-associated lncRNAs as the system-level inference, has become a valuable complementary complement to wet-lab experiments. Graph-based deep learning methods have been applied to various aspects of computational biology.[8−14] These computational approaches help address the

Received: October 10, 2023
Revised: November 20, 2023
Accepted: November 28, 2023
Published: December 22, 2023



challenges posed by the vast amount of data and provide insights into disease-associated lncRNAs.

Computer-aided inference models proposed in recent years can be classified into three categories: (1) network propagation-based methods, leverage-known biological information to construct heterogeneous networks on which applying random walk or some propagation algorithms to infer LDA. In 2019, Wang et al.[15] proposed a multiple biodata set-based model LncDisAP, which utilized random walk with restart (RWR) on related networks to infer LDAs. In the same year, Li et al.[16] proposed an improved model called LRWHLDA based on local random walking, which overcome the limitation of RWR-based models by known LDAs to walk. In 2019, Zhang et al.[17] proposed a new propagation method LncRDNetFlow, which used priority-based ranking to integrate and propagate information in heterogeneous networks for inferring LDAs. In 2018, Ding et al.[18] proposed the model TPGLDA to infer LDAs, which employed resource allocation to integrated heterogeneous features on the lncRNA-disease-gene heterogeneous network. (2) Matrix completion-based methods use matrix factorization to optimize an object function, completing the missing elements in a matrix composed of biodata. In 2018, Fu et al.[19] proposed MFLDA, a matrix factorization-based model that decomposed the matrix of heterogeneous biodata into a low-rank matrix. MFLDA then optimized the low-rank matrix through iteration to reconstruct the matrix of LDAs. In 2020, Zeng et al.[20] proposed SDLDA, a framework that combined SVD with deep learning to extract linear and nonlinear features of lncRNA and diseases. In 2018 and 2019, Lu et al. proposed SIMCLDA[21] based on inductive matrix completion and GMCLCA[22] based on geometric matrix completion to infer LDAs, making effective use of the inner structure embedded in the matrix of LDAs. In 2021, Zeng et al.[23] proposed DMFLDA, a deep matrix factorization-based model to predict LDAs, capturing complex nonlinear relationships between lncRNAs and diseases with a cascade of nonlinear hidden layers. (3) Deep learning-based methods use neural networks to extract deep and complex features from bioinformation networks, leading to improved performance.[24] In 2019, Xuan et al.[25] proposed GCNLDA, which utilized graph convolutional network (GCN) and convolutional neural network (CNN) to learn the local representational structure of lncRNA-disease-miRNA heterogeneous network for inferring LDAs. In 2021, Shi et al.[26] proposed VGAELDA, an end-to-end model that combined VGAE for graph representation learning and alternate training via variational inference, enhancing the capability to capture efficient low-dimensional representations from high-dimensional features for predicting unknown LDAs. In 2022, Wang and Zhong[27] proposed gGATLDA, which extracted closed subgraphs from the LDA matrix and integrated similarities to construct feature vectors for training graph neural networks (GNNs) to infer LDAs. In the same year, Xuan et al.[28] proposed MGLDA, which learned the local and global topology and pairwise attributes to encode and integrate the semantics of multiple meta-paths in a heterogeneous graph, aiding in LDA inference. In 2022, Fan et al.[29] proposed GCRFLDA, a novel prediction method that constructed an encoder with a conditional random field and attention mechanism to learn efficient embeddings of nodes, alongside a decoder layer to score LDAs. Also, in 2022, Zhou et al.[30] proposed LDAformer, a novel LDA prediction model based on topological feature extraction and a transformer encoder. LDAformer designed a topological feature extraction process to capture multihop topological pathway features latent in the heterogeneous network and used a transformer encoder based on global self-attention to infer LDAs by capturing interdependencies between heterogeneous pathways.

However, the models in the above three categories still exhibit the following limitations in learning complex features from heterogeneous bionetwork:

- Underutilization of rich interaction information: the complex mechanisms and functions within bionetworks are not fully leveraged because the aforementioned models treat information from nodes of different types equally, without taking into account the heterogeneity of the network.
- Focus on linear interaction information: the linear interaction information derived from direct or indirect neighbors has been focused on, rather than the semantic information embedded in the heterogeneous network.
- Inherent sparsity challenges: the intrinsic sparsity of heterogeneous bionetworks can introduce bias and instability into the model outcomes.

To address these limitations, we made two key steps. First, we constructed a heterogeneous network known as LDMN (lncRNA-disease-miRNA). Second, we proposed a novel computational model, HGCNLDA, which efficiently integrates heterogeneous features using GCN for the identification of LDAs. In summary, our model offers the following contributions:

- Constructed a heterogeneous LDMN network by introducing miRNA nodes into the known lncRNA-disease bipartite network, thereby enriching the amount of information embedding in the bionetwork.
- Developed two types of encoders within GCN (intra-GCN and inter-GCN) to extract features, while considering sematic relationships and interactions between heterogeneous nodes.
- Integrated information coming from diverse domains composed of homogeneous or heterogeneous nodes, using a bipolymerizer and attention mechanism.
- Enhanced the model's generalization capability and stability by conducting the residual connection and layer normalization (Layer Norm).
- Strengthened the model's robustness by optimizations that avoided the adverse impact of extremely unbalanced positive and negative samples in sparse data sets.

## 2. RESULTS

**2.1. Experiment Data Set.** The performance of our HGCNLDA was evaluated on two benchmark data sets with collection and preprocession described in the literature[19] and literature,[30] respectively:

- Data set1, there are 2697 known LDAs coming from LncRNADisease[5] and Lnc2Cancer,[6] 13562 miRNA-disease associations sourced from HMDD v2.0,[7] as well as 1002 lncRNA-miRNA interactive relationships from starBase v2.0.[31] Data set1 covers 240 lncRNAs, 412 diseases, and 495 miRNAs.
- Data set2, there are 3833 known LDAs coming from Lnc2Cancer v3.0[32] and LncRNADisease v2.0,[33] 8540 miRNA-disease associations sourced from HMDD v3.0,[34] as well as 2108 lncRNA-miRNA interactive relationships


https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

from starBase v2.0. Data set2 covers 665 lncRNAs, 316 diseases, and 295 miRNAs.

The sparsity is defined as the ratio of the number of known LDAs to the number of all possible associations. Data set1, with a sparsity ratio of 1:37, has been widely utilized for performance evaluation since its construction in 2018. On the other hand, Data set2, a newly constructed data set in 2022, exhibits a sparsity ratio of 1:55. While Data set1 is a well-established data set with extensive usage, its construction process relies on certain logical presuppositions as outlined in the original literature.[30] In contrast, Data set2 was constructed respecting the original literature's evidence records in Lnc2Cancer and LncRNADisease, without introducing any logical presuppositions in the process. Moreover, Data set2 is even sparser than Data set1, despite both having a significant imbalance between positive and negative samples. As a result, the performance evaluation on Data set2 is considerably more challenging and objective compared to Data set1.

**2.2. Evaluation Metric and Method.** When the association score of an lncRNA-disease node pair surpasses a given threshold, it is classified as a positive sample. Otherwise, it is designated as a negative sample. The corresponding true positive rate (TPR) and false positive rate (FPR) at a specific threshold were computed. For various threshold values, multiple sets of TPR and FPR were obtained, and subsequently, a receiver operating characteristic (ROC) curve was generated according to these TPRs and FPRs. Two common metrics, namely, the area under the ROC curve (AUROC) and the area under the precision-recall (PR) curve (AUPR), were employed to assess the predictive performance of the models included in the comparison. To mitigate the impact of randomness in experimental results, a 5-fold cross-validation approach was repeated 10 times for evaluation. The average values derived from these repetitions were then calculated to serve as the final evaluation results.

**2.3. Experimental Environment and Parameters.** The pyTorch framework was selected as the experimental environment. Drawing from previous experience, hyperparameters were set at fixed values to attain optimal model performance. To prevent the model from overfitting, Dropout[35] was applied during GCN training to randomly discard network edges with a fixed probability before performing the convolution operation. The precise values of each hyperparameter are detailed in Tables 1 and 2.

**Table 1. Some Detailed Hyperparameter Setup in Experiment**

| hyperparameter | value |
|---|---|
| nearest $K$ neighbors | 3 |
| GCN layers $n$ | 2 |
| learning rate | 0.001 |
| weight attenuation coefficient | 0.00001 |
| dropout probability | 0.4 |

**Table 2. Different Values of Some Hyperparameters in Different Data Sets**

| hyperparameter | value in Data set1 | value in Data set2 |
|---|---|---|
| epoch | 300 | 150 |
| GCN hidden layers $h$ | 256 | 64 |

**2.4. Parameter Selection.** In the tables above, the optimum values of the nearest neighbors ($K$) and GCN hidden layers ($h$) were determined by using a grid search method within the specified ranges of {5, 10, 15, 20} and {32, 64, 128, 256}, respectively. The selection process for these optimal values was visually represented by the heatmaps in Figure 1. From the results presented in the heatmaps, it was observed that AUROC and AUPR achieved higher values with smaller values of $K$, indicating that the inclusion of more neighbors in constructing the adjacent matrix introduced more noise. Regarding the GCN hidden layers ($h$), the model achieved higher AUROC and AUPR values on Data set1 with the addition of $h$. However, in Data set2, this trend was opposite. This contrasting performance on Data set1 and Data set2, with varying values of $h$, indicated that the model had difficulty in learning the low-dimensional representation of nodes on Data set1 with a smaller $h$, while it tended to overfit on Data set2 with a larger $h$. Therefore, based on the analysis above, the optimal values of $K$ and $h$ were set to be 3 and 256 on Data set1, 3, and 64 on Data set2, respectively.

**2.5. Evaluation Result and Analyzation.** The related state-of-the-art models, including SIMCLDA[21] in 2018, DMFLDA[23] in 2020, SDLDA[20] in 2020, GCRFLDA[29] in 2022, and LDAformer[30] in 2022, were compared with our HGCNLDA in the same experimental environment and data sets. The obtained AUROC and AUPR values are detailed in Table 3, Figures 2, and 3.

From the results in Table 3, Figures 2, and 3, our HGCNLDA outperforms the other models on both evaluation metrics on Data set1, especially in terms of AUPR, except for a slightly 0.55% lower AUROC value compared to LDAformer. On Data set2, which is notably more challenging and objective compared to Data set1, our HGCNLDA achieved the highest AUROC and AUPR values among all models. Specifically, our HGCNLDA's AUROC value is 0.73% higher than that of LDAformer ranking second in this metric, and its AUPR value is 25.7% higher than that of SDLDA ranking second in this metric. This significant increase in the AUPR value on Data set2 demonstrated the effectiveness of our HGCNLDA in predicting LDAs on highly imbalanced data sets.

**2.6. Ablation Experiment.** *2.6.1. Crucial-Component Combination.* Our HGCNLDA is composed of five crucial components: ① intra-GCN encoder for aggregating features from the homogeneous domain; ② inter-GCN encoder for aggregating features from the heterogeneous domain; ③ summation polymerizer for fusing extracted features; ④ concatenation polymerizer for fusing extracted features; and ⑤ global attention layer for obtaining low-dimensional representations. The ablation experiments, which were designed to assess the impact of working in various combinations, including or excluding different crucial components, are detailed in Table 4. The corresponding experimental results are presented in Table 5.

In Table 5, the highest AUROC and AUPR values achieved by HGCNLDA explicitly demonstrate the impact of incorporating the crucial components into the model. On Data set1, our HGCNLDA demonstrates that the improvements in AUROC values by 1.7, 1, 1.7, 1.6, and 2.3%, and AUPR values exhibit enhancements of 20.8, 12.9, 18.5, 18.3, and 22.7% compared to model variants A-only, R-only, AR-S, AR-C, and AR-SC, respectively. Although our HGCNLDA's performance has not significantly improved in terms of the AUROC metric, there has been a substantial increase in the AUPR metric. These experimental results illustrate that, on one hand, only



https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

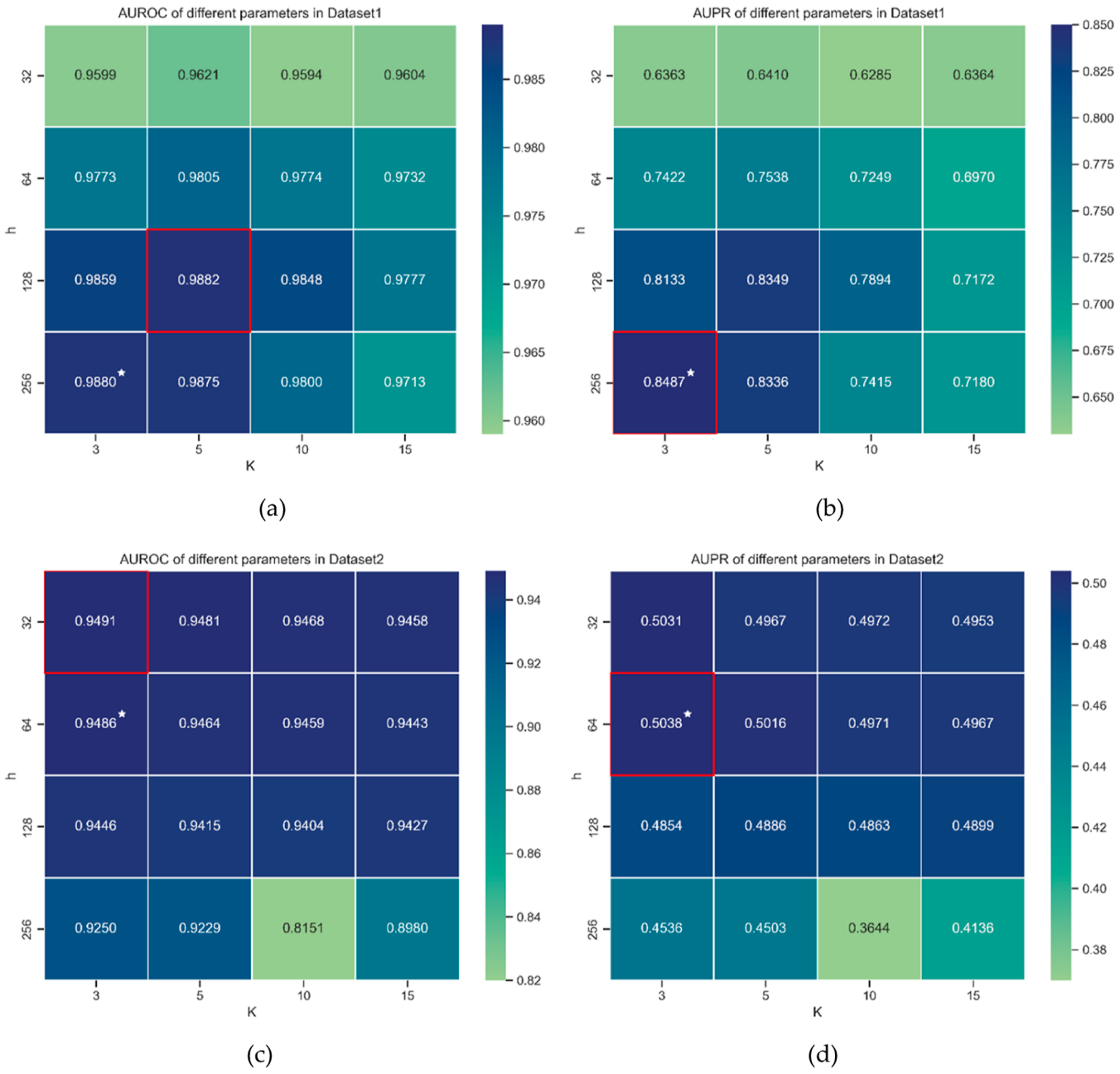



**Figure 1.** AUROC and AUPR with different $K$ and $h$ values on Data set1 and Data set2. (a) AUROC on Data set1; (b) AUPR on Data set1; (c) AUROC on Data set2; and (d) AUPR on Data set2. In each panel, the best performance indicator is highlighted with a red box, and the ultimately selected parameter combination is marked with a white five-pointed star.

**Table 3. Evaluation Results Are for Comparison**

| model | Data set1 | | Data set2 | |
|---|---|---|---|---|
| | AUROC | AUPR | AUROC | AUPR |
| SIMCLDA | 0.8385 ± 0.0329 | 0.1755 ± 0.0925 | 0.6446 ± 0.1588 | 0.0525 ± 0.0472 |
| DMFLDA | 0.8485 ± 0.1699 | 0.3422 ± 0.1720 | 0.8575 ± 0.1668 | 0.2205 ± 0.1093 |
| SDLDA | 0.8518 ± 0.1728 | 0.5113 ± 0.2441 | 0.8447 ± 0.1741 | 0.3759 ± 0.1813 |
| GCRFLDA | 0.9596 ± 0.0026 | 0.4130 ± 0.0292 | 0.9476 ± 0.0225 | 0.2308 ± 0.1056 |
| LDAformer | **0.9935** ± 0.0019 | 0.7325 ± 0.0186 | 0.9423 ± 0.0038 | 0.2354 ± 0.0130 |
| HGCNLDA (ours) | 0.9880 ± 0.0008 | **0.8501** ± 0.0193 | **0.9492** ± 0.0044 | **0.5056** ± 0.0170 |

considering homogeneous information interactions without distinguishing heterogeneous information interactions could lead to a significant decrease in the model's ability to classify positive samples. On the other hand, the fusion methods used for extracted features play a crucial role in enhancing and stabilizing the model's predictive performance. On Data set2, the AUROC and AUPR values obtained by our HGCNLDA did not exhibit significant improvements compared to those of the



https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

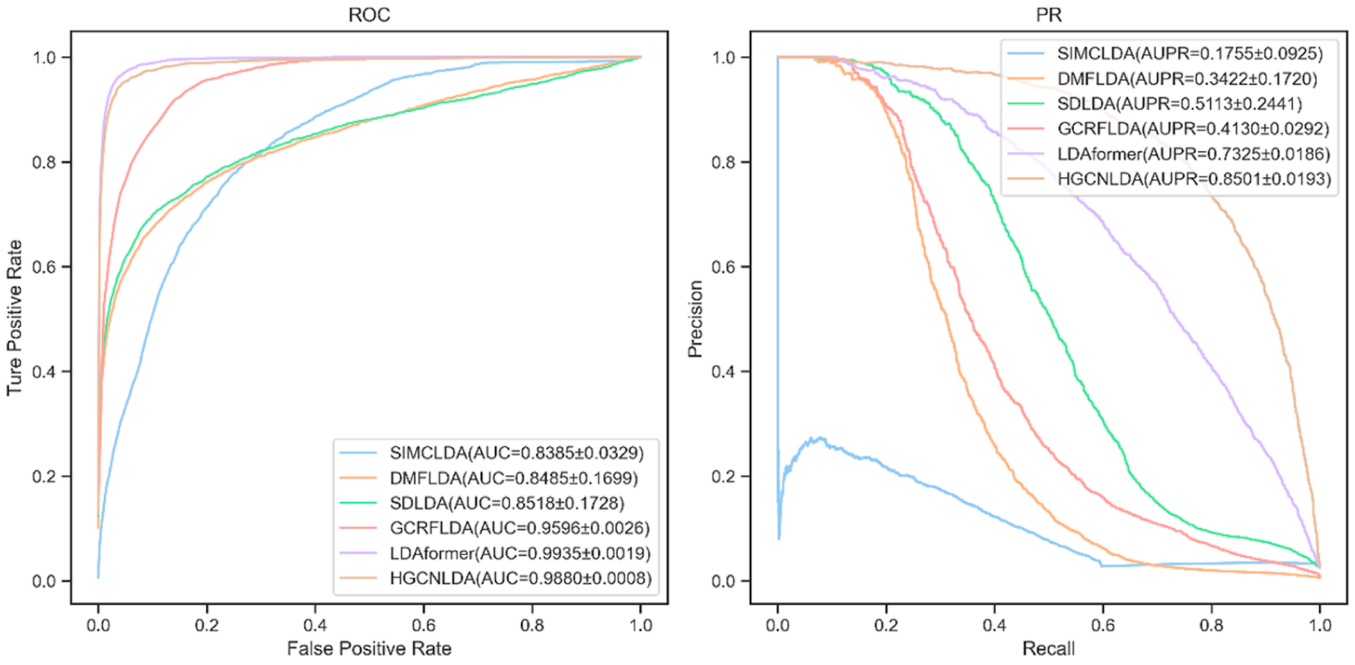



**Figure 2.** Performance of models engaged in comparison on Data set1.

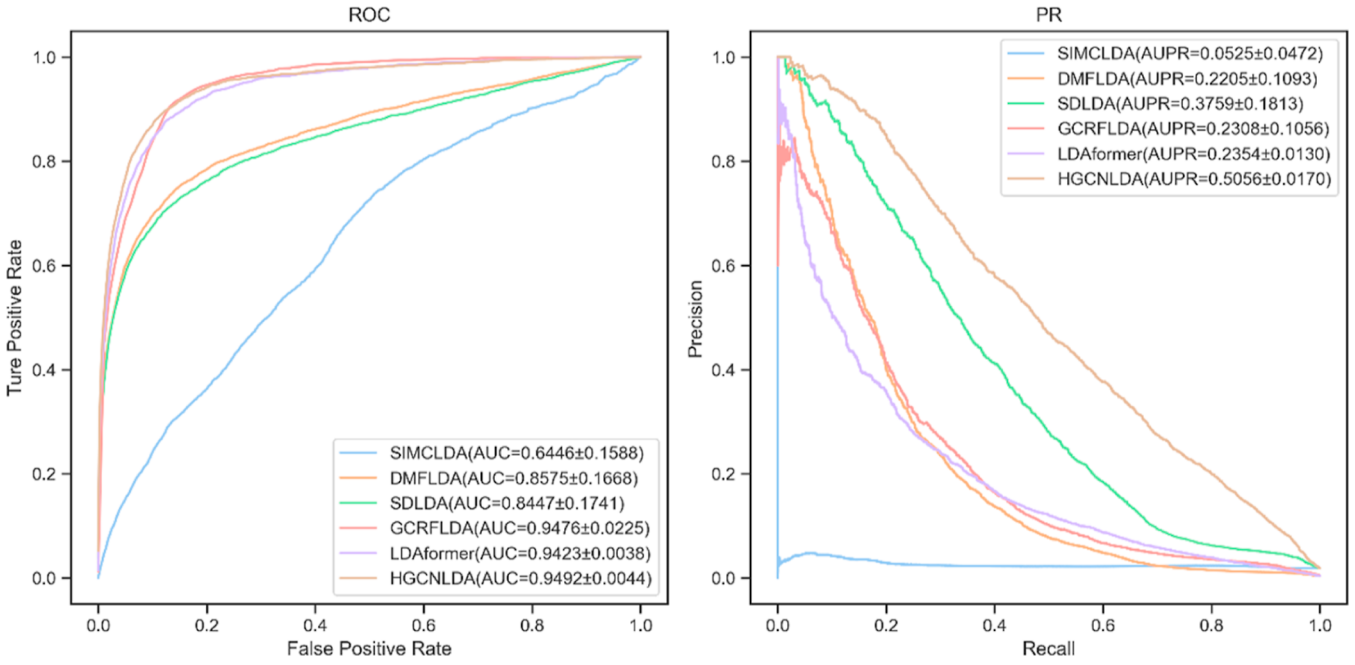



**Figure 3.** Performance of models engaged in comparison on Data set2.

**Table 4. Various Combinations of Crucial Parts Involved**

| model variant | variant implication | GCN-based aggregator | | bipolymerizer | | attention layer |
|---|---|---|---|---|---|---|
| | | intra-GCN encoder | inter-GCN encoder | summation | concatenation | |
| A-only | only crucial part ① included | Inc. | Excl. | Excl. | Excl. | Excl. |
| R-only | only crucial part ② included | Excl. | Inc. | Excl. | Excl. | Excl. |
| AR-S | crucial parts ①②③ included | Inc. | Inc. | Inc. | Excl. | Excl. |
| AR-C | crucial parts ①②④ included | Inc. | Inc. | Excl. | Inc. | Excl. |
| AR-SC | crucial parts ①②③④ included | Inc. | Inc. | Inc. | Inc. | Excl. |
| HGCNLDA | all five crucial parts are included | Inc. | Inc. | Inc. | Inc. | Inc. |



https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

**Table 5. Performances of Each Crucial−Part Combination on Two Data Sets**

| model variant | Data set1 | | Data set2 | |
|---|---|---|---|---|
| | AUROC | AUPR | AUROC | AUPR |
| A-only | 0.9711 ± 0.0063 | 0.6730 ± 0.0435 | 0.9356 ± 0.0051 | 0.4109 ± 0.0258 |
| R-only | 0.9782 ± 0.0067 | 0.7408 ± 0.0601 | 0.9484 ± 0.0037 | 0.5044 ± 0.0183 |
| AR-S | 0.9710 ± 0.0092 | 0.6931 ± 0.0453 | 0.9476 ± 0.0043 | 0.5009 ± 0.0162 |
| AR-C | 0.9726 ± 0.0075 | 0.6943 ± 0.0428 | 0.9475 ± 0.0047 | 0.5038 ± 0.0219 |
| AR-SC | 0.9652 ± 0.0101 | 0.6573 ± 0.0475 | 0.9478 ± 0.0036 | 0.5020 ± 0.0156 |
| HGCNLDA | **0.9880** ± 0.0008 | **0.8501** ± 0.0193 | **0.9492** ± 0.0044 | **0.5056** ± 0.0170 |

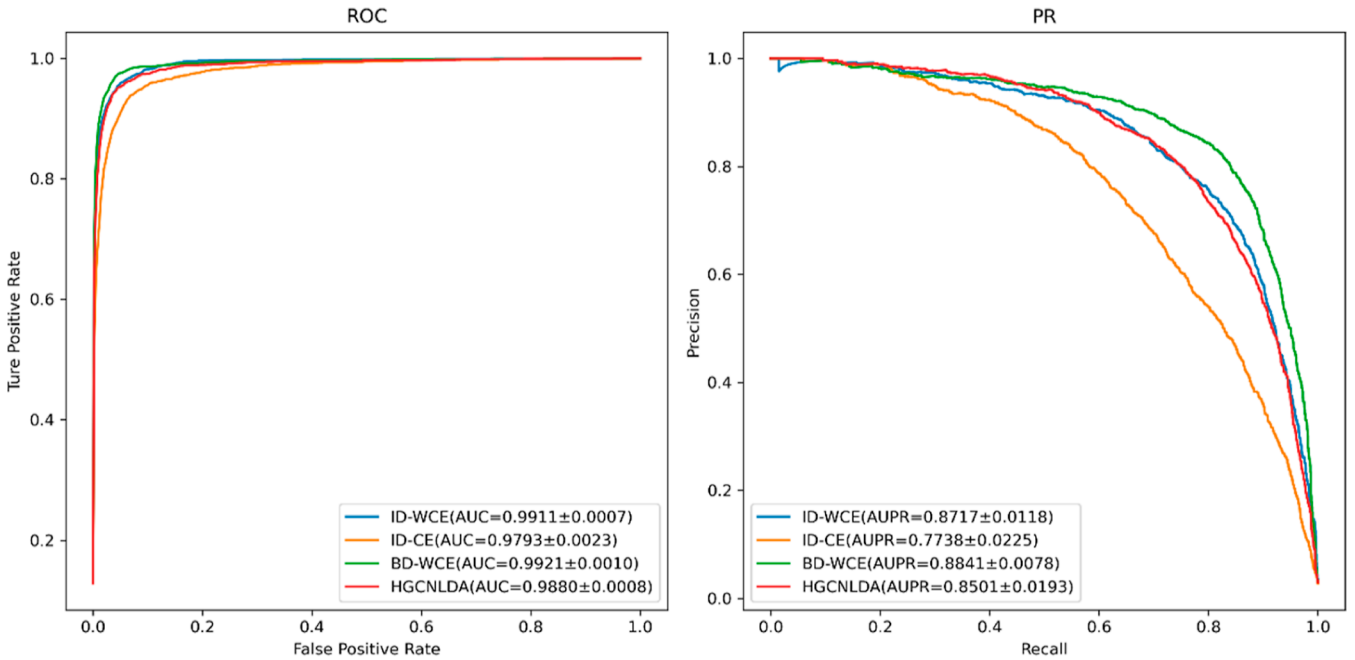



**Figure 4.** Performance of the model variants on Data set1.

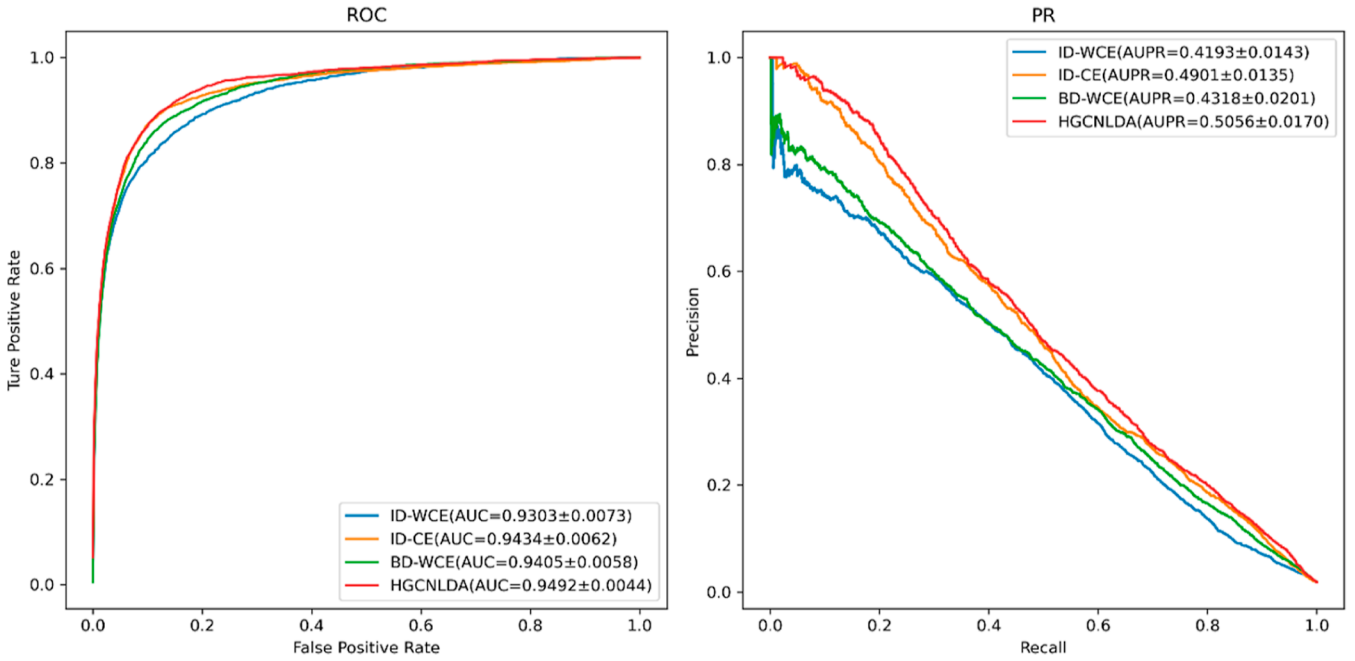



**Figure 5.** Performance of the model variants on Data set2.



https://doi.org/10.1021/acsomega.3c07923
*ACS Omega* 2024, 9, 1472−1484

**Table 6. Top 10 Breast Cancer-Related LncRNAs in Potential**

| case | ranking | LncRNA | PMID | case | ranking | LncRNA | PMID |
|---|---|---|---|---|---|---|---|
| breast cancer | 1 | BDNF-AS | 32521278 | breast cancer | 6 | TDRG1 | 33822672 |
| | 2 | FOXD2-AS1 | 34043149 | | 7 | TUSC7 | 34305410 |
| | 3 | DLEU2 | unconfirmed | | 8 | HCP5 | 36980766 |
| | 4 | MIR100HG | 33088216 | | 9 | DGCR5 | 32521856 |
| | 5 | C5orf66-AS1 | 35499320 | | 10 | HIF1A-AS1 | 26339353 |

**Table 7. Top 10 Lung Cancer-Related LncRNAs in Potential**

| case | ranking | LncRNA | PMID | case | ranking | LncRNA | PMID |
|---|---|---|---|---|---|---|---|
| lung cancer | 1 | XIST | 31553952 | lung cancer | 6 | SNHG16 | 33015794 |
| | 2 | TUG1 | 35249784 | 7 | | TP73-AS1 | 36118078 |
| | 3 | CRNDE | 35611803 | 8 | | BDNF-AS | 31421833 |
| | 4 | DLX6-AS1 | 36017915 | 9 | | HULC | 30575912 |
| | 5 | ZFAS1 | 36569479 | 10 | | SNHG6 | 32590190 |

**Table 8. Top 10 Colorectal Cancer-Related LncRNAs in Potential**

| case | ranking | LncRNA | PMID | case | ranking | LncRNA | PMID |
|---|---|---|---|---|---|---|---|
| colorectal cancer | 1 | CDKN2B-AS1 | 34436551 | colorectal cancer | 6 | SNHG6 | 31322251 |
| | 2 | MIAT | 35607443 | | 7 | PCAT1 | 33277833 |
| | 3 | TP73-AS1 | 35896939 | | 8 | SNHG3 | 34661273 |
| | 4 | CRNDE | 33891491 | | 9 | SNHG14 | 31273190 |
| | 5 | DLX6-AS1 | 32785606 | | 10 | SNHG7 | 35747807 |

model variants, including A-only, R-only, AR-S, AR-C, and AR-SC. This indicates that performance improvement on Data set2 presents a greater challenge, as discriminating between positive and negative samples on Data set2 is notably more complex than on Data set1.

*2.6.2. Optimization Combination.* To determine the optimization scheme, ablation experiments were designed to assess the impact of various combination of decoders and loss functions: ① the inner-product decoder and weighted cross entropy were included to create the model variant ID-WCE; ② the inner-product decoder and cross entropy were included to create the model variant ID-CE; ③ the bilinear-decoder and weighted cross entropy were included to create the model variant BD-WCE; and ④ the bilinear-decoder and cross entropy were included in our HGCNLDA. The corresponding experimental results were presented in Figures 4 and 5.

From the experimental results depicted in the figures above, the model variants using the bilinear-decoder achieved superior performance with higher AUROC and AUPR values compared with those using the inner-product decoder. As for the choice of loss functions, on Data set1, the model variants employing weighted cross entropy (ID-WCE and BD-WCE) exhibited better performance with higher AUROC and AUPR values than those using cross entropy (ID-CE and HGCNLDA). However, this trend was reversed for weighted cross entropy on Data set2. It indicated that cross entropy, when applied in our HGCNLDA, performed better on data sets that are sparser and more seriously imbalanced between positive and negative samples, such as Data set2.

**2.7. Case Study.** Global cancer statistics[36] reported that breast cancer is the most prevalent type of cancer in women worldwide and ranks second in terms of death tolls. Lung cancer is the second most common cancer in both males and females when combined. Colorectal cancer (CRC) (colon + rectum) is the third leading cause of cancer-related mortality worldwide. These three representative cancer types were selected as the real cases to further investigate HGCNLDA's ability in predicting lncRNAs related to these significant specific diseases—breast cancer, lung cancer, and colorectal cancer. To assess this, known associations related to breast cancer were masked in Data set2, and the remaining known associations served as the training samples to train HGCNLDA. A similar process was repeated for lung cancer and colorectal cancer. The associations between lncRNAs and the three aforementioned cancers, predicted by HGCNLDA, were then sorted by scores in descending order. The top 10 associations for each of these cancers were selected based on their scores. Detailed verification results were presented in Tables 6−8, along with the corresponding evidence found in the PubMed database (https://pubmed.ncbi.nlm.nih.gov/).

In Table 6, only one out of ten lncRNAs predicted has not been found to have any evidence described in the literature of PubMed database. Although there is no direct description of the association between "DLEU2" and breast cancer in the literature so far, the literature[37−39] have demonstrated that DLEU2 binds to miR-30a-5p through the same binding site, facilitating the expression of ETS2. ETS2 is overexpressed in breast carcinoma. Moreover, the miR-30a-5p axis regulates breast cancer cell proliferation and migration. Therefore, this indirectly supports the existence of an association between DLEU2 and breast cancer. In Tables 7 and 8, all lncRNAs predicted by our HGCNLDA have been found to have the evidence of associations with lung cancer and colorectal cancer.

## 3. DISCUSSION

Numerous computational models for predicting LDAs have been developed. However, many of them share some common limitations, including underutilization of rich interaction information, a focus on linear interaction information rather than semantic information, and bias and instability due to the inherent sparsity of biological data. To address these limitations, a novel computational model, HGCNLDA, has been proposed



https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

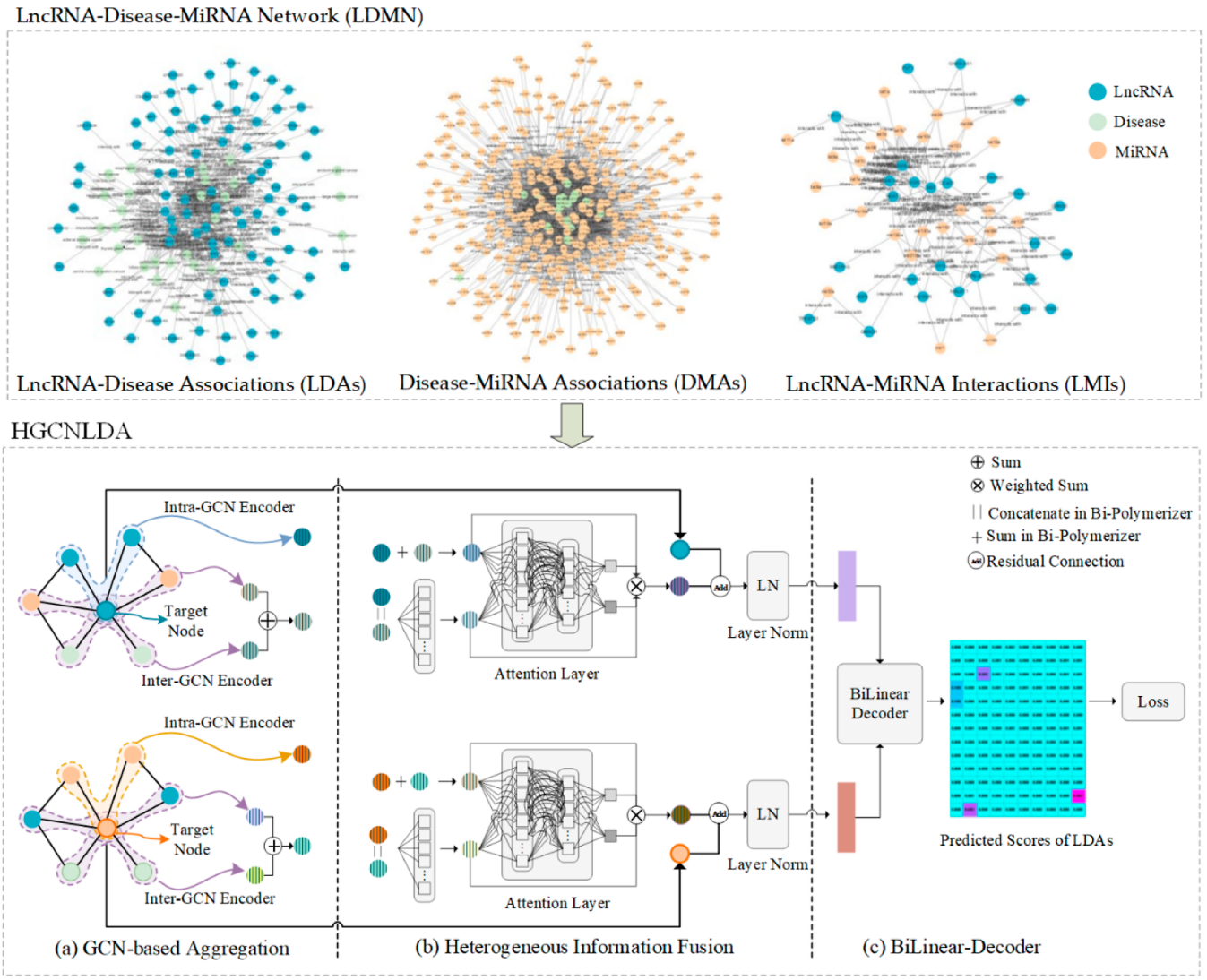



**Figure 6.** Model schematic depiction. First step, the initial features of lncRNAs, diseases, and miRNAs were linearly transformed into the same projection space; second step, the intra-GCN encoder and inter-GCN encoder extracted features from the domains (homogeneous and heterogeneous) in LDMN; third step, bipolymerization operations (summation and concatenation) that fused the features extracted parallelly. Subsequently, a global attention mechanism was designed to further integrate heterogeneous information to obtain low-dimensional representations. To stabilize the gradient and enhance model robustness, a residual connection and layer normalization (Layer Norm) were added between each module's input and output layers; and fourth step, the bilinear-decoder predicted the probability scores of LDAs.

for identifying potential LDAs. HGCNLDA incorporates a GCN-based aggregation module and a heterogeneous information fusion module to acquire semantic relationships from both homogeneous and heterogeneous domains within the heterogeneous network, LDMN. In the information fusion module, a dipolymerizer and global attention mechanism are employed to acquire low-dimensional node representations, considering both local neighborhood attribute information and structural information. Lastly, a bilinear-decoder module rebuilt the LDA matrix to obtain the predicted probability scores of LDAs, through reinforcing common attributes and diminishing differences of vectors. The evaluation results on both Data set1 and Data set2, particularly on Data set2, clearly demonstrate that HGCNLDA outperforms the other five related state-of-the-art models in terms of AUROC and AUPR values. Case studies further confirm that HGCNLDA exhibits excellent predictive capabilities in identifying potential LDAs, which will enable it to contribute to the design of treatment strategies and the development of therapeutics. However, it is worth noting that HGCNLDA still did not fully leverage the abundant semantic information within heterogeneous networks. Therefore, our forthcoming work will center around, on one hand, further enhancing the model's ability in learning complex features on heterogeneous networks by utilizing semantic information and, on the other hand, integrating more diverse and richer biological interaction information into the model's computational process.

## 4. MATERIALS AND METHODS

**4.1. Similarity Network Construction.** *4.1.1. Disease Semantic Similarity.* Wang et al.[40] employed medical subject headings (MeSH), which describes relationships between diseases using a directed acyclic graph (DAG), to calculate the disease semantic similarity. Within the DAG, a disease node $d$ is represented by $DAG_d = (d, T_d, E_d)$, where $T_d$ denotes the set encompassing all ancestors of disease $d$ (inclusive of $d$ itself), and $E_d$ denotes the edges connecting these diseases within the set.


https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

Consequently, the semantic contribution value of any disease $d$ to disease $d_i$ was established using the expression

$$SC_{d_i}(d) = \begin{cases} 1, & \text{if } d = d_i \\ \max\{\gamma \times SC_{d_i}(d') | d' \in \text{child of } d\}, & \text{if } d \neq d_i \end{cases} \quad (1)$$

where $\gamma$ denotes the semantic contribution factor whose value is set to 0.5 by reference to the literature.[40]

The semantic value of disease $d_i$ is denoted by $SV(d_i)$, defined as

$$SV(d_i) = \sum_{d \in T_d} SC_{d_i}(d) \quad (2)$$

Based on this method, we calculated the semantic similarity between any two diseases with which to construct the disease similarity network denoted as $\mathbf{S}_d \in {}^{o \times o}$, where $o$ represents the number of diseases. In $\mathbf{S}_d$, any element representing the semantic similarity between disease $d_i$ and disease $d_j$ is denoted as $\mathbf{S}_d(d_i,d_j)$, and calculated as

$$\mathbf{S}_d(d_i, d_j) = \frac{\sum_{d_k \in T_{d_i} \cap T_{d_j}} (SC_{d_i}(d_k) + SC_{d_j}(d_k))}{SV(d_i) + SV(d_j)} \quad (3)$$

*4.1.2. LncRNA (MiRNA) Functional Similarity.* It is well known that lncRNAs (miRNAs) with similar functions tend to be associated with similar diseases and vice versa.[41,42] Based on this assumption, we employed a method similar to the one described in the literature[43] to calculate the lncRNA functional similarity between any two lncRNAs by calculating disease semantic similarities. Consequently, the lncRNA similarity network and miRNA similarity network were constructed, denoted by $\mathbf{S}_l \in {}^{u \times u}$ and $\mathbf{S}_m \in {}^{v \times v}$, where $u$ and $v$ represent the numbers of lncRNAs and miRNAs, respectively.

Sets $D(i)$ and $D(j)$ represent the sets of disease nodes associated with lncRNA $l_i$ and $l_j$, respectively. Hence, the similarity between disease $d_t \in D(i)$ and $D(j)$ is defined as

$$DS(d_t, D(j)) = \max_{d \in D(j)} (\mathbf{S}_d(d_t, d)) \quad (4)$$

Within $\mathbf{S}_l$, any element representing the functional similarity between lncRNA $l_i$ and $l_j$ is denoted by $\mathbf{S}_l(l_i,l_j)$, and calculated as

$$\mathbf{S}_l(l_i, l_j) = \frac{\sum_{d \in D(i)} DS(d, D(i)) + \sum_{d \in D(j)} DS(d, D(j))}{|D(i)| + |D(j)|} \quad (5)$$

The construction process of miRNA similarity network $\mathbf{S}_m \in {}^{v \times v}$ follows a similar procedure as described above.

*4.1.3. Neighborhood Matrix.* For any disease node in $\mathbf{S}_d$, we selected the nearest $K$ neighbors related to this disease node to construct a disease neighborhood matrix, denoted as $\mathbf{A}_d \in {}^{o \times o}$. Matrix element $\mathbf{A}_d(d_i,d_j) = \mathbf{S}_d(d_i,d_j)$ when disease $d_j$ belongs to one of the $K$ neighbors of disease $d_i$, otherwise $\mathbf{A}_d(d_i,d_j) = 0$. The construction process for lncRNA neighborhood matrix $\mathbf{A}_l \in {}^{u \times u}$ and miRNA neighborhood matrix $\mathbf{A}_m \in {}^{v \times v}$ is similar.

**4.2. LDMN Construction.** According to the literature,[19,44−46] LDAs, disease-miRNA associations (DMAs), and lncRNA-miRNA interactions (LMIs) were acquired to create the corresponding networks represented as matrices, $\mathbf{A}_{ld} \in {}^{u \times o}$, $\mathbf{A}_{dm} \in {}^{o \times v}$, and $\mathbf{A}_{lm} \in {}^{u \times v}$, respectively. Each matrix element $\mathbf{A}_{ld}(l_i,d_j) = 1$ when lncRNA $l_i$ has a known association with disease $d_j$, otherwise $\mathbf{A}_{ld}(l_i,d_j) = 0$. Matrix elements $\mathbf{A}_{dm}(d_i,m_j)$ and $\mathbf{A}_{lm}(l_i,m_j)$ were calculated in a similar manner. Finally, a heterogeneous network named LDMN, represented as an adjacent matrix $\mathbf{A} \in {}^{(u+o+v) \times (u+o+v)}$, was constructed as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_l & \mathbf{A}_{ld} & \mathbf{A}_{lm} \\ \mathbf{A}_{ld}{}^T & \mathbf{A}_d & \mathbf{A}_{dm} \\ \mathbf{A}_{lm}{}^T & \mathbf{A}_{dm}{}^T & \mathbf{A}_m \end{bmatrix} \quad (6)$$

where $\mathbf{A}_{ld}{}^T$, $\mathbf{A}_{lm}^T$, and $\mathbf{A}_{dm}{}^T$ represent the corresponding transpose matrices of $\mathbf{A}_{ld} \in {}^{u \times o}$, $\mathbf{A}_{dm} \in {}^{o \times v}$, and $\mathbf{A}_{lm} \in {}^{u \times v}$, respectively.

In the construction process of the LDMN, only $K$-nearest neighbors of any related node were selected, which may result in the loss of information from some nodes. Therefore, an initial feature matrix, denoted as $\mathbf{X} \in {}^{(u+o+v) \times (u+o+v)}$, preserved the similarity information on all nodes related to LDMN, with definition as

$$\mathbf{X} = \begin{bmatrix} \mathbf{S}_l & 0 & 0 \\ 0 & \mathbf{S}_d & 0 \\ 0 & 0 & \mathbf{S}_m \end{bmatrix} \quad (7)$$

**4.3. Model Structure.** A new computational model to infer LDAs, namely, HGCNLDA, was described in detail in this section. The workflow of HGCNLDA consisted of three modules: GCN-based aggregation, heterogeneous information fusion, and a bilinear-decoder, as briefly depicted in Figure 6.

*4.3.1. GCN-Based Aggregation. 4.3.1.1. Feature Linear Transformation.* The feature vectors of heterogeneous nodes may exist in different dimensions. Even when feature vectors have equal dimensions, they may be located in different feature spaces.[47] To facilitate the exchange of node information between heterogeneous networks and uniformly processed feature vectors, the initial features of heterogeneous nodes were projected into the same potential vector space through a specific linear transformation

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_l \\ \mathbf{H}_d \\ \mathbf{H}_m \end{bmatrix} = XW \quad (8)$$

where $\mathbf{H} \in {}^{(u+o+v) \times h}$ is the matrix obtained after projection with the initial features, with $\mathbf{H}_l \in {}^{u \times h}$, $\mathbf{H}_d \in {}^{o \times h}$, and $\mathbf{H}_m \in {}^{v \times h}$ representing the projection matrices for three types of nodes (lncRNA, disease, and miRNA), respectively. $\mathbf{W} \in {}^{(u+o+v) \times h}$ is a linear transformation matrix, and dimension $h$ is the dimension of the projected vector.

*4.3.1.2. Encoder.* GCN is capable of encoding both graph structure and node features, serving as a powerful feature extractor for nodes in a graph during convolution operations.[48] A specific node, along with its homogeneous neighbors, constituted a homogeneous domain, while a specific node, along with its heterogeneous neighbors, constituted a heterogeneous domain. In LDMN, the process of message delivery among homogeneous nodes differs from that among the heterogeneous nodes. Drawing from the description of inter-



https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

and intradomain feature extraction,[49] we designed two distinct encoders for feature extraction from homogeneous and heterogeneous domains. Consequently, both inter-GCN and intra-GCN encoders aggregated information from the node and its neighbors to extract the local features of the node. This aggregation relied on topological relationships between nodes, enabling it to obtain more accurate feature representations

$$\mathrm{GCN}(\mathbf{A}, \mathbf{H}^{(n)}, \mathbf{W}^{(n)}) = \sigma(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(n)}\mathbf{W}^{(n)}) \tag{9}$$

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I} \tag{10}$$

where $\mathbf{I}$ is an identity matrix with the same dimension as matrix $\mathbf{A}$, $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$,$\mathbf{H}^{(n)} \in {}^{(u+o+v)\times h}$ that represents the matrix of node features input into the $n$th layer of GCN, $\mathbf{W}^{(n)} \in {}^{h\times h}$ represents a learnable weight matrix in the $n$th layer of GCN, and $\sigma$ is the ReLU activation function.[50]

*4.3.1.3. Homogeneous Aggregation.* In LDMN, there are three types of homogeneous domains (lncRNA−lncRNA, disease−disease, and miRNA−miRNA) in which homogeneous features of nodes were extracted by the intra-GCN encoder

$$\bar{\mathbf{H}} = \begin{bmatrix} \bar{\mathbf{H}}_l \\ \bar{\mathbf{H}}_d \\ \bar{\mathbf{H}}_m \end{bmatrix} = \begin{bmatrix} \mathrm{GCN}(\mathbf{A}_l, \mathbf{H}_l, \mathbf{W}_l) \\ \mathrm{GCN}(\mathbf{A}_d, \mathbf{H}_d, \mathbf{W}_d) \\ \mathrm{GCN}(\mathbf{A}_m, \mathbf{H}_m, \mathbf{W}_m) \end{bmatrix} \tag{11}$$

where $\mathbf{W}_l \in {}^{h\times h}$, $\mathbf{W}_d \in {}^{h\times h}$, and $\mathbf{W}_m \in {}^{h\times h}$ are the learnable weight matrices in GCN for three types of homogeneous domains (lncRNA−lncRNA, disease−disease, and miRNA−miRNA) respectively. The dimension of the hidden layer in the GCN is set to be $h$. Consequently, the outputs $\bar{\mathbf{H}}_l \in {}^{u\times h}$, $\bar{\mathbf{H}}_d \in {}^{o\times h}$, and $\bar{\mathbf{H}}_m \in {}^{v\times h}$ represent the homogeneous features extracted with the intra-GCN encoder for three types of homogeneous domains, respectively.

The literature[51] concluded that stacking multiple convolutional layers did not improve performance, and a simple combination of a convolutional layer followed by a dense layer worked best. Consequently, the intra-GCN encoder in this study was implemented with a single convolution layer in GCN.

*4.3.1.4. Heterogeneous Aggregation.* In LDMN, there are three types of interactions between heterogeneous nodes, forming three distinct heterogeneous domains: lncRNA−disease, disease−miRNA, and lncRNA−miRNA. The inter-GCN encoder aggregated interactions between one specific node and the other two types of heterogeneous nodes to extract heterogeneous features. For example, in LDMN, when lncRNA node $l_i$ interacts with two types of heterogeneous nodes (disease and miRNA), the heterogeneous features of node $l_i$ are extracted by the inter-GCN encoder from two types of heterogeneous domains, which are lncRNA-disease and lncRNA-miRNA domains

$$\tilde{\mathbf{H}}_{l_i} = \sigma\left(\frac{\sum_{j=1}^{o}\mathbf{A}_{ld}(l_i, d_j)\mathbf{H}_{d_j}\mathbf{W}_{ld}}{\sum_{j=1}^{o}\mathbf{A}_{ld}(l_i, d_j)} + \frac{\sum_{q=1}^{v}\mathbf{A}_{lm}(l_i, m_q)\mathbf{H}_{m_q}\mathbf{W}_{lm}}{\sum_{q=1}^{v}\mathbf{A}_{lm}(l_i, m_q)}\right) \tag{12}$$

where $\mathbf{H}_{d_j} \in {}^{1\times h}$ and $\mathbf{H}_{m_q} \in {}^{1\times h}$ input into GCN represent the feature vectors of disease node $d_j$ and miRNA node $m_q$ in projection matrices $\mathbf{H}_d \in {}^{o\times h}$ and $\mathbf{H}_m \in {}^{v\times h}$, respectively. $\mathbf{W}_{ld} \in {}^{h\times h}$ and $\mathbf{W}_{lm} \in {}^{h\times h}$ are the learnable weight matrices for the heterogeneous domains (lncRNA−disease and lncRNA−miRNA). Correspondingly, output vector $\tilde{\mathbf{H}}_{l_i} \in {}^{1\times h}$ represents the heterogeneous extracted features of lncRNA node $l_i$.

Similarly, the heterogeneous features of disease node $d_j$ and miRNA node $m_q$ extracted with the inter-GCN encoder were

$$\tilde{\mathbf{H}}_{d_j} = \sigma\left(\frac{\sum_{i=1}^{u}\mathbf{A}_{ld}(l_i, d_j)\mathbf{H}_{l_i}\mathbf{W}_{ld}}{\sum_{i=1}^{u}\mathbf{A}_{ld}(l_i, d_j)} + \frac{\sum_{q=1}^{v}\mathbf{A}_{dm}(d_j, m_q)\mathbf{H}_{m_q}\mathbf{W}_{dm}}{\sum_{q=1}^{v}\mathbf{A}_{dm}(d_j, m_q)}\right) \tag{13}$$

$$\tilde{\mathbf{H}}_{m_q} = \sigma\left(\frac{\sum_{i=1}^{u}\mathbf{A}_{lm}(l_i, \mathrm{m}_q)\mathbf{H}_{l_i}\mathbf{W}_{lm}}{\sum_{i=1}^{u}\mathbf{A}_{lm}(l_i, m_q)} + \frac{\sum_{j=1}^{o}\mathbf{A}_{dm}(d_j, m_q)\mathbf{H}_{d_j}\mathbf{W}_{dm}}{\sum_{j=1}^{o}\mathbf{A}_{dm}(d_j, m_q)}\right) \tag{14}$$

*4.3.2. Feature Fusion. 4.3.2.1. Bipolymerization Operation.* In the heterogeneous network LDMN, features coming from both homogeneous and heterogeneous neighbors of one specific node were aggregated into that node using inter- and intra-GCN encoders. To enhance the accuracy and comprehensiveness of the node's representation, the extracted homogeneous features were fused with the extracted heterogeneous features through a bipolymerization operation (summation and concatenation)

$$\hat{\mathbf{H}}_{l_i} = \bar{\mathbf{H}}_{l_i} + \tilde{\mathbf{H}}_{l_i} \tag{15}$$

$$\breve{\mathbf{H}}_{l_i} = (\bar{\mathbf{H}}_{l_i} \| \tilde{\mathbf{H}}_{l_i})\mathbf{W}_c + \mathbf{B}_c \tag{16}$$

where $\bar{\mathbf{H}}_{l_i} \in {}^{1\times h}$ is the feature vector of lncRNA node $l_i$ in the extracted homogeneous feature matrix $\bar{\mathbf{H}}_l$. Operator + denotes the summation operation, and ‖ denotes the concatenation operation on vectors. In the concatenation operation, $\mathbf{W}_c \in {}^{2h\times h}$ is a linear transformation matrix, and vector $\mathbf{B}_c \in {}^{1\times h}$ is a bias. The resulting output vectors $\hat{\mathbf{H}}_{l_i} \in {}^{1\times h}$ and $\breve{\mathbf{H}}_{l_i} \in {}^{1\times h}$ represent the fused feature vectors after the bipolymerization operation.

*4.3.2.2. Attention Mechanism.* The summation operation in the bipolymerization process aims to incorporate features from all neighbors, including those with different properties (homogeneity and heterogeneity), into the specific node. This process helped to obtain the global-domain information. Meanwhile, the concatenation operation in the bipolymerization operation aims to preserve the diversity of homogeneous and heterogeneous features, contributing to obtain the structure information.

Since the contributions of global-domain and structure information in forming a specific node's representation differ, an attention mechanism was employed to weigh and combine these two types of information, resulting in the node's integrated low-dimensional representation

$$\hat{\phi}_{l_i} = \boldsymbol{\omega}\cdot\tanh(\mathbf{W}_a\hat{\mathbf{H}}_{l_i}^{T} + \mathbf{B}_a) \tag{17}$$



https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

$$\breve{\phi}_{l_i} = \boldsymbol{\omega}\cdot\tanh(\mathbf{W}_a\breve{\mathbf{H}}_{l_i}^{T} + \mathbf{B}_a) \tag{18}$$

$$\hat{\alpha}_{l_i} = \frac{\exp(\hat{\phi}_{l_i})}{\exp(\hat{\phi}_{l_i}) + \exp(\breve{\phi}_{l_i})} \tag{19}$$

$$\breve{\alpha}_{l_i} = \frac{\exp(\breve{\phi}_{l_i})}{\exp(\hat{\phi}_{l_i}) + \exp(\breve{\phi}_{l_i})} \tag{20}$$

$$\ddot{\mathbf{H}}_{l_i} = \hat{\boldsymbol{\alpha}}_{l_i}\hat{\mathbf{H}}_{l_i} + \breve{\boldsymbol{\alpha}}_{l_i}\breve{\mathbf{H}}_{l_i} \tag{21}$$

where $\boldsymbol{\omega} \in {}^{1\times h}$ is a mapping vector, and weight matrix $\mathbf{W}_a \in {}^{h\times h}$ and bias vector $\mathbf{B}_a \in {}^{h\times 1}$ were used for nonlinear transformation of a specific node's features. The vectors $\hat{\mathbf{H}}_{l_i}^{T}$ and $\breve{\mathbf{H}}_{l_i}^{T}$ are the transposed vectors corresponding to $\hat{\mathbf{H}}_{l_i}$ and $\breve{\mathbf{H}}_{l_i}$, respectively. The corresponding results $\hat{\phi}_l^{\,i}$ and $\breve{\phi}_l^{\,i}$ are the attention scores for fused features ($\hat{\mathbf{H}}_{l_i}$ and $\breve{\mathbf{H}}_{l_i}$) obtained with the bipolymerization operation. Weights $\hat{\alpha}_{l_i} \in [0,1]$ and $\breve{\alpha}_{l_i} \in [0,1]$ are obtained with a softmax function that normalized vector $\hat{\mathbf{H}}_{l_i}$ and $\breve{\mathbf{H}}_{l_i}$. Finally, vector $\ddot{\mathbf{H}}_{l_i} \in {}^{1\times h}$ represents the integrated representation of node $l_i$ obtained through further weighted summation.

In order to enhance the model's generalization ability and reduce training time, we introduced a residual connection[52] and applied layer normalization (Layer Norm)[53] to acquire the final representation of node $l_i$, with denotation as $\mathbf{Z}_{l_i} \in {}^{1\times h}$

$$\mathbf{Z}_{l_i} = \text{layer norm}(\ddot{\mathbf{H}}_{l_i} + \mathbf{H}_{l_i}) \tag{22}$$

Similarly, the final representation of disease node $d_j$, denoted as $\mathbf{Z}_{d_j} \in {}^{1\times h}$, was obtained by using the same procedure.

*4.3.3. LDA's Rebuilding. 4.3.3.1. Bilinear-Decoder.* Multiplication between vectors that emphasizes common properties of vectors and diminishes differences could effectively model interactions.[54] Inspired by the description of the bilinear-decoder in the literature,[51] the edge type (or rating type), denoted as $r \in R = \{0, 1\}$ between node $l_i$ and $d_j$, was rebuilt with element $\hat{\mathbf{A}}_{ld}(l_i, d_j)$ in matrix $\hat{\mathbf{A}}_{ld} \in {}^{u\times o}$. When $r = 0$, $\hat{\mathbf{A}}_{ld}(l_i, d_j) = r$ indicates that no association exists between the node pair $(l_i,d_j)$. Otherwise, when $r = 1$, $\hat{\mathbf{A}}_{ld}(l_i, d_j) = r$ indicates that node $l_i$ doses associate with $d_j$. Through a bilinear operation followed by the application of softmax function, the decoder outputs a probability of the possible rating type as a predicted association score for a specific node pair

$$p(\hat{\mathbf{A}}_{ld}(l_i, d_j) = r) = \text{softmax}(\text{bilinear} - \text{decoder}(\mathbf{Z}_{l_i}, \mathbf{Z}_{d_j})) = \frac{e^{(\oplus(\mathbf{Z}_{l_i}\mathbf{Q}_r\odot\mathbf{Z}_{d_j}))}}{\sum_{s=0}^{1} e^{(\oplus(\mathbf{Z}_{l_i}\mathbf{Q}_s\odot\mathbf{Z}_{d_j}))}} \tag{23}$$

where operator $\oplus$ denotes the summation of each element in the vector, operator $\odot$ denotes the dot product between two vectors, and matrix $\mathbf{Q}_r \in {}^{h\times h}$ is a trainable weight matrix.

*4.3.3.2. Model Optimization.* According to the literature,[19] the original LDA matrix $\mathbf{A}_{ld} \in {}^{u\times o}$ is an extremely imbalanced data set, where the known LDAs (positive samples) are significantly fewer in number compared to the unknown or nonexisting LDAs (negative samples). To mitigate the adverse impact of treating positive and negative samples equally during model optimization, the model parameters were learned by minimizing the following negative log likelihood of the predicted probability score[51]

$$\text{loss} = -\sum_{(l_i,d_j)\in U^+}\sum_{r=0}^{R} I[\mathbf{A}_{ld}(l_i, d_j) = r] \log(p(\hat{\mathbf{A}}_{ld}(l_i, d_j) = r)) \tag{24}$$

where $I[\mathbf{A}_{ld}(l_i, d_j) = r] = 1$ when matrix element $\mathbf{A}_{ld}(l_i, d_j) = r$, otherwise $I[\mathbf{A}_{ld}(l_i, d_j) = r] = 0$. Notation $(l_i,d_j) \in U^+$ represents a known LDA in $\mathbf{A}_{ld}$, and $U^+$ represents the collection of all positive samples. Only the positive samples require optimization.

## AUTHOR INFORMATION


### Corresponding Author

**Yi Zhang** − *Guilin University of Technology, Guilin 541004, China; Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China;* orcid.org/0000-0001-6167-7945; Email: zywait@glut.edu.cn

### Authors

**Gangsheng Cai** − *Guilin University of Technology, Guilin 541004, China; Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China*

**Xin Li** − *Guilin University of Technology, Guilin 541004, China; Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China*

**Min Chen** − *School of Computer Science and Technology, Hunan Institute of Technology, Hengyang 421010, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c07923


### Author Contributions

Conceptualization, Y.Z.; Data curation, S.C.; Formal analysis, X.L.; Funding acquisition, Y.Z.; Methodology, Y.Z. and S.C.; Software, S.C.; Validation, X.L.; Writing—original draft, Y.Z.; Writing—review and editing, Y.Z. and M.C.


### Funding

This research was funded by the National Natural Science Foundation of China (grant nos. 62166014 and 62162019) with funder Yi Zhang, and the Natural Science Foundation of Guangxi Zhuang Autonomous Region (grant no. 2020GXNSFAA297255) with funder Yi Zhang.


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS


The authors thank the anonymous reviewers for suggestions that helped improve the paper substantially.


## ABBREVIATIONS

AUROC area under ROC curve
AUPR area under PR curve



https://doi.org/10.1021/acsomega.3c07923
ACS Omega 2024, 9, 1472−1484

| | |
|---|---|
| LDAs | lncRNA-disease associations |
| LDMN | lncRNA-disease-miRNA network |
| GCN | graph convolutional network |

## REFERENCES


(1) Mattick, J. S.; Amaral, P. P.; Carninci, P.; Carpenter, S.; Chang, H. Y.; Chen, L.-L.; Chen, R.; Dean, C.; Dinger, M. E.; Fitzgerald, K. A.; et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **2023**, *24* (6), 430−447.

(2) Fathi Dizaji, B. Strategies to target long non-coding RNAs in cancer treatment: Progress and challenges. *Egypt. J. Med. Hum. Genet.* **2020**, *21* (1), 41.

(3) Geisler, S.; Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* **2013**, *14* (11), 699−712.

(4) Yan, C.; Zhang, Z.; Bao, S.; Hou, P.; Zhou, M.; Xu, C.; Sun, J. Computational methods and applications for identifying disease-associated lncRNAs as potential biomarkers and therapeutic targets. *Mol. Ther.—Nucleic Acids* **2020**, *21*, 156−171.

(5) Chen, G.; Wang, Z.; Wang, D.; Qiu, C.; Liu, M.; Chen, X.; Zhang, Q.; Yan, G.; Cui, Q. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* **2012**, *41* (D1), D983−D986.

(6) Ning, S.; Zhang, J.; Wang, P.; Zhi, H.; Wang, J.; Liu, Y.; Gao, Y.; Guo, M.; Yue, M.; Wang, L.; et al. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* **2016**, *44* (D1), D980−D985.

(7) Li, Y.; Qiu, C.; Tu, J.; Geng, B.; Yang, J.; Jiang, T.; Cui, Q. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* **2014**, *42* (D1), D1070−D1074.

(8) Su, X.; Hu, L.; You, Z.; Hu, P.; Wang, L.; Zhao, B. A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2. *Briefings Bioinf.* **2022**, *23* (1), bbab526.

(9) Zheng, K.; Zhang, X. L.; Wang, L.; You, Z. H.; Ji, B. Y.; Liang, X.; Li, Z. W. SPRDA: a link prediction approach based on the structural perturbation to infer disease-associated Piwi-interacting RNAs. *Briefings Bioinf.* **2023**, *24* (1), bbac498.

(10) Chen, Z.; Zhang, L.; Sun, J.; Meng, R.; Yin, S.; Zhao, Q. DCAMCP: A deep learning model based on capsule network and attention mechanism for molecular carcinogenicity prediction. *J. Cell. Mol. Med.* **2023**, *27* (20), 3117−3126.

(11) Gao, H.; Sun, J.; Wang, Y.; Lu, Y.; Liu, L.; Zhao, Q.; Shuai, J. Predicting metabolite-disease associations based on auto-encoder and non-negative matrix factorization. *Briefings Bioinf.* **2023**, *24* (5), bbad259.

(12) Meng, R.; Yin, S.; Sun, J.; Hu, H.; Zhao, Q. scAAGA: Single cell data analysis framework using asymmetric autoencoder with gene attention. *Comput. Biol. Med.* **2023**, *165*, 107414.

(13) Wang, T.; Sun, J.; Zhao, Q. Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism. *Comput. Biol. Med.* **2023**, *153*, 106464.

(14) Sun, F.; Sun, J.; Zhao, Q. A deep learning method for predicting metabolite-disease associations via graph neural network. *Briefings Bioinf.* **2022**, *23* (4), bbac266.

(15) Wang, Y.; Juan, L.; Peng, J.; Zang, T.; Wang, Y. LncDisAP: a computation model for LncRNA-disease association prediction based on multiple biological datasets. *BMC Bioinf.* **2019**, *20* (S16), 582.

(16) Li, J.; Zhao, H.; Xuan, Z.; Yu, J.; Feng, X.; Liao, B.; Wang, L. A novel approach for potential human LncRNA-disease association prediction based on local random walk. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2021**, *18* (3), 1049−1059.

(17) Zhang, J.; Zhang, Z.; Chen, Z.; Deng, L. Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2019**, *16* (2), 396−406.

(18) 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**, *8* (1), 1065.

(19) 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.

(20) Zeng, M.; Lu, C.; Zhang, F.; Li, Y.; Wu, F.-X.; Li, Y.; Li, M. SDLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning. *Methods* **2020**, *179*, 73−80.

(21) Lu, C.; Yang, M.; Luo, F.; Wu, F.-X.; Li, M.; Pan, Y.; Li, Y.; Wang, J. Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics* **2018**, *34* (19), 3357−3364.

(22) Lu, C.; Yang, M.; Li, M.; Li, Y.; Wu, F.-X.; Wang, J. Predicting human lncRNA-disease associations based on geometric matrix completion. *IEEE J. Biomed. Health Inform.* **2020**, *24* (8), 2420−2429.

(23) Zeng, M.; Lu, C.; Fei, Z.; Wu, F.-X.; Li, Y.; Wang, J.; Li, M. DMFLDA: a deep learning framework for predicting lncRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2021**, *18* (6), 2353−2363.

(24) Lei, X.; Mudiyanselage, T. B.; Zhang, Y.; Bian, C.; Lan, W.; Yu, N.; Pan, Y. A comprehensive survey on computational methods of non-coding RNA and disease association prediction. *Briefings Bioinf.* **2021**, *22* (4), bbaa350.

(25) Xuan, P.; Pan, S.; Zhang, T.; Liu, Y.; Sun, H. Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. *Cells* **2019**, *8* (9), 1012.

(26) Shi, Z.; Zhang, H.; Jin, C.; Quan, X.; Yin, Y. A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations. *BMC Bioinf.* **2021**, *22* (1), 136.

(27) Wang, L.; Zhong, C. gGATLDA: lncRNA-disease association prediction based on graph-level graph attention network. *BMC Bioinf.* **2022**, *23* (1), 11.

(28) Xuan, P.; Zhao, Y.; Cui, H.; Zhan, L.; Jin, Q.; Zhang, T.; Nakaguchi, T. Semantic Meta-Path Enhanced Global and Local Topology Learning for lncRNA-Disease Association Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2023**, *20* (2), 1480−1491.

(29) Fan, Y.; Chen, M.; Pan, X. GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. *Briefings Bioinf.* **2022**, *23* (1), bbab361.

(30) Zhou, Y.; Wang, X.; Yao, L.; Zhu, M. LDAformer: predicting lncRNA-disease associations based on topological feature extraction and Transformer encoder. *Briefings Bioinf.* **2022**, *23* (6), bbac370.

(31) Li, J.-H.; Liu, S.; Zhou, H.; Qu, L.-H.; Yang, J.-H. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **2014**, *42* (D1), D92−D97.

(32) Gao, Y.; Shang, S.; Guo, S.; Li, X.; Zhou, H.; Liu, H.; Sun, Y.; Wang, J.; Wang, P.; Zhi, H.; et al. Lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. *Nucleic Acids Res.* **2021**, *49* (D1), D1251−D1258.

(33) Bao, Z.; Yang, Z.; Huang, Z.; Zhou, Y.; Cui, Q.; Dong, D. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* **2019**, *47* (D1), D1034−D1037.

(34) Huang, Z.; Shi, J.; Gao, Y.; Cui, C.; Zhang, S.; Li, J.; Zhou, Y.; Cui, Q. HMDD v3. 0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* **2019**, *47* (D1), D1013−D1017.

(35) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15* (1), 1929−1958.

(36) Sung, H.; Ferlay, J.; Siegel, R. L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *Ca-Cancer J. Clin.* **2021**, *71* (3), 209−249.

(37) Xu, W.; Wang, B.; Cai, Y.; Guo, C.; Liu, K.; Yuan, C. DLEU2: A Meaningful Long Noncoding RNA in Oncogenesis. *Curr. Pharm. Des.* **2021**, *27* (20), 2337−2343.




https://doi.org/10.1021/acsomega.3c07923
*ACS Omega* 2024, 9, 1472−1484

(38) Zhang, W.; Wang, Y.; Xu, P.; Du, Y.; Guan, W. lncRNA DLEU2 Accelerates Oral Cancer Progression via miR-30a-5p/RAP1B Axis to Regulate p38 MAPK Signaling Pathway. *Dis. Markers* **2022**, *2022*, 9310048.
(39) Li, L.; Kang, L.; Zhao, W.; Feng, Y.; Liu, W.; Wang, T.; Mai, H.; Huang, J.; Chen, S.; Liang, Y.; Han, J.; Xu, X.; Ye, Q. miR-30a-5p suppresses breast tumor growth and metastasis through inhibition of LDHA-mediated Warburg effect. *Cancer Lett.* **2017**, *400*, 89−98.
(40) Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **2010**, *26* (13), 1644−1650.
(41) Chen, X.; Yan, G.-Y. Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* **2013**, *29* (20), 2617−2624.
(42) Lu, M.; Zhang, Q.; Deng, M.; Miao, J.; Guo, Y.; Gao, W.; Cui, Q. An analysis of human microRNA and disease associations. *PLoS One* **2008**, *3* (10), No. e3420.
(43) Chen, X.; Clarence Yan, C.; Luo, C.; Ji, W.; Zhang, Y.; Dai, Q. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* **2015**, *5* (1), 11338.
(44) Wang, W.; Zhang, L.; Sun, J.; Zhao, Q.; Shuai, J. Predicting the potential human lncRNA-miRNA interactions based on graph convolution network with conditional random field. *Briefings Bioinf.* **2022**, *23* (6), bbac463.
(45) Zhang, L.; Yang, P.; Feng, H.; Zhao, Q.; Liu, H. Using Network Distance Analysis to Predict lncRNA-miRNA Interactions. *Interdiscip. Sci.: Comput. Life Sci.* **2021**, *13* (3), 535−545.
(46) Wong, L.; Wang, L.; You, Z. H.; Yuan, C. A.; Huang, Y. A.; Cao, M. Y. GKLOMLI: a link prediction model for inferring miRNA-lncRNA interactions by using Gaussian kernel-based method on network profile and linear optimization algorithm. *BMC Bioinf.* **2023**, *24* (1), 188.
(47) Fu, X.; Zhang, J.; Meng, Z.; King, I. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. *Proceedings of The Web Conference 2020*, 2020; pp 2331−2341.
(48) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:02907. preprint
(49) Cai, L.; Lu, C.; Xu, J.; Meng, Y.; Wang, P.; Fu, X.; Zeng, X.; Su, Y. Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Briefings Bioinf.* **2021**, *22* (6), bbab319.
(50) Nair, V.; Hinton, G. E. Rectified linear units improve restricted boltzmann machines. *the 27th International Conference on Machine Learning (ICML-10)*; Omnipress, 2010.
(51) Berg, R. v. d.; Kipf, T. N.; Welling, M. Graph convolutional matrix completion. *arXiv* **2017**, arXiv:02263. preprint
(52) Li, Q.; Han, Z.; Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. *Thirty-Second AAAI conference on artificial intelligence*; Association for the Advancement of Artificial Intelligence, 2018.
(53) Ba, J. L.; Kiros, J. R.; Hinton, G. E. Layer normalization. *arXiv* **2016**, arXiv:06450. preprint
(54) Zhu, H.; Feng, F.; He, X.; Wang, X.; Li, Y.; Zheng, K.; Zhang, Y. Bilinear graph neural network with neighbor interactions. *arXiv* **2020**, arXiv:2002.03575. preprint


https://doi.org/10.1021/acsomega.3c07923
*ACS Omega* 2024, 9, 1472−1484