BMC Bioinformatics

RESEARCH

Open Access

SPGA: graph representation learning and attention fusion for enhanced disease-associated snoRNA prediction

Hailin Chen[1*] and Zirui Song[1]

*Correspondence:
Hailin Chen
chenhailin@ecjtu.edu.cn
[1]School of Information and Software Engineering, East China Jiaotong University, Nanchang 330013, China

Abstract

Background Small nucleolar RNAs (snoRNAs) are increasingly recognized for their involvement in human diseases. Accurate and robust prediction of disease-associated snoRNAs is crucial for accelerating drug discovery and disease treatment. However, the limited availability of biomedical data in this domain poses a significant challenge to the generalization ability of machine learning models. While existing computational methods have made progress, their performance is often constrained by data scarcity.

Results To overcome these limitations, we introduce SPGA, a novel graph representation learning framework designed to enhance snoRNA-disease association prediction. SPGA leverages intrinsic structural features of snoRNAs and diseases to construct interaction-aware graph representations. It then employs residual graph convolutional networks augmented with a hierarchical attention mechanism to learn robust node embeddings for association predictions. Comprehensive experiments demonstrate that SPGA effectively alleviates the data scarcity problem for model generalization and significantly improves state-of-the-art methods in terms of prediction accuracy, achieving an AUC of 0.9812 and an AUPR of 0.9749 under five-fold cross-validation setting. Case studies further validate its efficacy in identifying novel snoRNA-disease associations.

Conclusions Our study provides a potent computational tool for prioritizing candidate disease-related snoRNAs, thereby facilitating the discovery of potential diagnostic biomarkers and therapeutic targets.

Keywords Disease-snoRNA associations, Graph representation learning, Attention fusion

Background

snoRNAs constitute a diverse class of short non-coding RNAs (ncRNAs) primarily renowned for their crucial roles in guiding the post-transcriptional modification of ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs) through methylation and pseudouridylation [1, 2]. Traditionally, snoRNAs were perceived to be confined to the nucleolus and exclusively involved in these housekeeping functions essential for

ribosome biogenesis and spliceosome assembly. However, accumulating evidence now paints a more complex picture, revealing that snoRNAs are far from mere cellular laborers. They exhibit remarkable functional versatility, participating in diverse biological processes beyond guiding modifications, including alternative splicing regulation [3], telomere maintenance [4], mRNA stability [5], and even acting as precursors for smaller regulatory RNAs like microRNAs (miRNAs) [6, 7].

This expanded understanding of snoRNA functionality has illuminated their significant involvement in human disease pathogenesis. Dysregulation of snoRNA expression or function has been identified in a wide spectrum of disorders, ranging from neurodegenerative conditions like Prader-Willi Syndrome [8] and Alzheimer's disease [9] to various types of cancers [10, 11] and metabolic diseases [12]. In oncology, for instance, specific snoRNAs are emerging as oncogenes or tumor suppressors, serving as potential biomarkers for diagnosis, prognosis, and even therapeutic targets [13]. The intricate interplay between altered snoRNA profiles and disease progression underscores their immense potential as novel diagnostic tools and therapeutic avenues.

Despite the growing recognition of their disease relevance, the comprehensive identification of disease-associated snoRNAs remains a significant challenge. Experimental confirmation and functional validation of snoRNAs, especially in the context of specific diseases, are costly, time-consuming, and often limited by sample accessibility and throughput [14]. Furthermore, the indirect nature of many snoRNA-mediated effects and their potential for context-dependent roles make direct experimental elucidation of their disease implications particularly complex. This highlights a critical need for robust and efficient methods to systematically identify and prioritize disease-related snoRNAs using existing biomedical information.

Computational approaches offer a powerful and indispensable means to bridge this knowledge gap. By leveraging advancements in genomics, transcriptomics, and computational biology, it is possible to infer potential disease-associated snoRNAs [15–19]. These snoRNA–disease association prediction methods have largely followed network- or representation-learning paradigms. For example, Zhang et al. [19] developed a computational model that combined graph convolution networks and self-supervised learning for snoRNA–disease association prediction. It should be noted that current computational frameworks for inferring disease-associated snoRNAs are often limited in their scope or prediction ability. There is a pressing need to develop more advanced models to provide more accurate predictions of snoRNA-disease associations.

Moreover, a significant obstacle in developing computational methodologies for predicting disease-related snoRNAs is the limited availability of relevant data. Compared to more extensively studied ncRNAs, such as miRNAs, the research data pertaining to snoRNA-disease associations is considerably less extensive. This data scarcity poses a challenge to the generalizability and prediction accuracy of machine learning models employed in this domain.

To address these challenges, we propose a novel computational framework entitled SPGA for inferring disease-associated snoRNAs in this study. Our approach integrates intrinsic structural features of snoRNAs and diseases, constructs interaction-aware graph representations, and employs residual graph convolutional networks augmented with a hierarchical attention mechanism to learn robust node embeddings for predictions. Through comprehensive rigorous tests and case studies, we demonstrate the

efficacy of our framework in identifying previously unrecognized snoRNA-disease associations and enhancing our understanding of their pathogenic roles. Our work represents a significant step towards accelerating the discovery of novel therapeutic targets and diagnostic markers, paving the way for more precise and personalized medicine.

Results

Experimental setting

In this section, we adopt five-fold cross-validation to assess the generalization ability of our method. The full dataset is randomly partitioned into five mutually exclusive, size-balanced folds. In each round, one fold serves as the test set while the remaining four folds are merged to form the training set; this procedure is repeated five times so that every fold is used exactly once for testing. Final performance values are reported as the arithmetic mean over the five runs, thereby reducing the influence of any single train-test split. Performance is evaluated with three complementary metrics: area under the ROC curve (AUC), area under the precision-recall curve (AUPR), and accuracy (ACC).

Performance comparison with other models

To benchmark the performance of the proposed approach on snoRNA-disease association prediction, we compare it against seven recent baselines:

NIMCGCN [20]: A model constructs a heterogeneous graph with graph convolutional networks (GCNs) and infers latent miRNA-disease links via neural inductive matrix completion.

AMHMDA [21]: A model combines multi-view similarity networks and hypergraph learning, employing attention mechanisms and virtual hyper-nodes to predict miRNA-disease associations.

NSAMDA [22]: A model fuses multi-source similarities into a heterogeneous graph, performs dynamic neighborhood aggregation with a graph attention network and decodes miRNA-disease associations via an inner product.

iPiDA-GCN [23]: A model integrates heterogeneous-network features within a dual-channel GCN architecture to identify piRNA-disease associations.

VGAMF [24]: A model couples a variational graph auto-encoder with non-negative matrix factorization to predict miRNA-disease associations.

IGCNSDA [25]: A model employs subgraph generation and neighbor-information aggregation to provide interpretable and efficient snoRNA-disease association prediction.

DK-SDA [26]: A model builds on a variational graph auto-encoder, applying decoupled representation learning and Kolmogorov-Arnold networks; the final graph is reconstructed via inner-product similarity and learned edge weights for snoRNA-disease prediction.

Tables 1 and 2 summarize the results based on the RNADisease dataset [27] and independent ncRPheno dataset [28], respectively. For RNADisease, our proposed model SPGA achieves an average AUC of 0.9812, AUPR of 0.9749, and ACC of 0.9452, outperforming all competing methods. Compared to the second-best baseline (DK-SDA), our model gains improvement of 2.28% in AUC, 0.26% in AUPR, and 2.51% in ACC. For the independent tests in the ncRPheno test set, SPGA also attains the highest AUC and AUPR compared with DK-SDA [26]. These results confirm that our model SPGA not

Table 1 Comparison results of all models based on the RNA disease dataset

Model	AUC	AUPR	ACC
AMHMDA	0.6572	0.7244	0.6506
NIMCGCN	0.6803	0.6655	0.6141
NSAMDA	0.7023	0.7588	0.6826
VGAMF	0.7304	0.7665	0.6809
iPiDA_GCN	0.8101	0.8173	0.7054
IGCNSDA	0.8438	0.8744	0.7831
DK-SDA	0.9584	0.9723	0.9201
SPGA (ours)	**0.9812**	**0.9749**	**0.9452**

The best results present in the bold

Table 2 Independent test results of all models based on the ncRPheno dataset

Model	AUC	AUPR	ACC
AMHMDA	0.5010	0.4945	0.4897
NIMCGCN	0.7084	0.7781	0.5205
NSAMDA	0.5072	0.6726	0.6911
VGAMF	0.7069	0.7049	0.6457
iPiDA_GCN	0.6182	0.6185	0.6651
IGCNSDA	0.7142	0.8746	0.7290
DK-SDA	0.8033	0.8192	**0.7517**
SPGA (ours)	**0.8529**	**0.8448**	0.7323

The best results present in the bold

only excels on the in-domain benchmark but also retains strong predictive power when evaluated on unseen data drawn from a different repository.

Hyperparameter analysis

Model performance is, to a certain extent, contingent on several key hyper-parameters. Five factors are examined here:

(i) The target dimensionality d' retained after Principal Component Analysis (PCA)
(ii) The dimensionality of the subsequent feature embeddings
(iii) The dropout rate
(iv) The number of GCN layers
(v) The number of attention heads. Using five-fold cross-validation, we systematically evaluate how different combinations of these parameters affect AUROC, AUPRC and ACC.

PCA's target feature dimension controls the projection dimension of the original features into the new feature space; we first compute the variance contribution of all principal components, and then select the dimension corresponding to the elbow point of the cumulative variance curve. Specifically, we obtain the diagonalized eigenvalue matrix $\mathbf{\Lambda}$ through Eq. 1:

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_d), \quad \lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_d \tag{1}$$

then we compute the variance contribution of each principal component and accumulate the contributions of the first d' principal components to obtain the cumulative contribution rate:

$$\mathrm{CR}_{\mathrm{cumulative}}(d') = \frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \tag{2}$$

As illustrated in Fig. 1, the cumulative-variance curve rises steeply at the beginning, implying that each additional principal component markedly increases the amount of explained variance. Beyond the elbow point, however, the curve flattens out, indicating that the marginal benefit of adding further components diminishes.

We empirically set the cumulative-variance threshold in the 0.95–0.99 range. Under this setting, d'varies between 43 and 152. We test four representative dimensionalities (43, 64, 128, and 152), as shown Fig. 2, and discover that our model obtains its best performance when $d' = 128$.

The embedding dimension partly determines both representational capacity and computational load. To assess its impact on prediction performance, we change the hidden dimension of the GCN across 16, 32, 64, and 128. As shown in Fig. 3, our model achieves its highest accuracy when dim = 32.

Dropout is commonly employed to mitigate overfitting. We test four dropout rates (0.10, 0.25, 0.50, and 0.80) and report the mean performance across the five folds. As illustrated in the Fig. 4, our model achieves its optimal results when the dropout rate is set to 0.25.

The number of GCN layers affects the extent of information propagation between paired entities. To examine its influence, we vary the GCN depth from 1 to 4. As shown in Fig. 5, the model achieves the best performance when two GCN layers are adopted.

The number of attention heads controls the diversity of feature aggregation in the attention mechanism. We evaluate three configurations with 2, 4, and 8 heads. As illustrated in Fig. 6, the model attains optimal performance when the number of attention heads is set to 4.

To sum up, we set d' in the model to be 128, the hidden dimension dim to 32, dropout to 0.25, the number of GCN layers to 2, and the number of attention heads to 4.

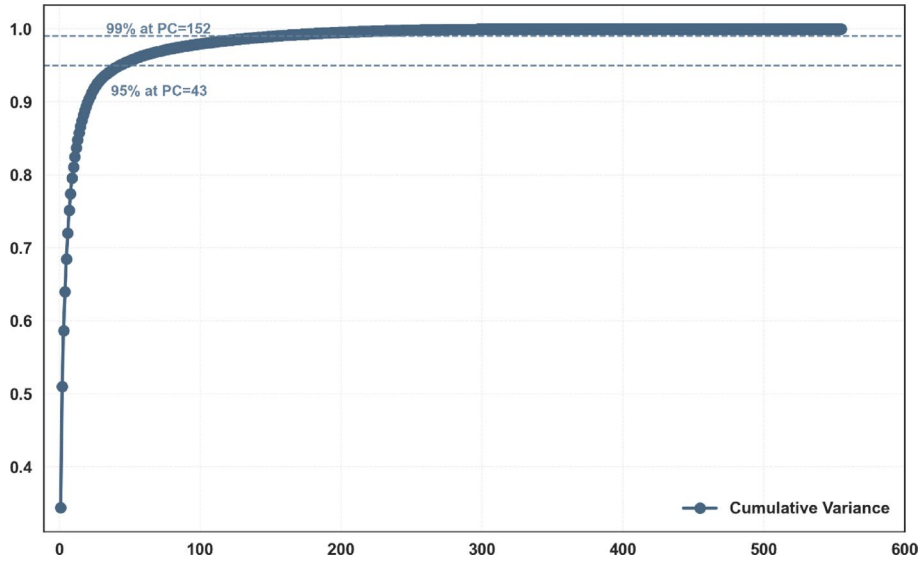

Fig. 1 PCA cumulative variance

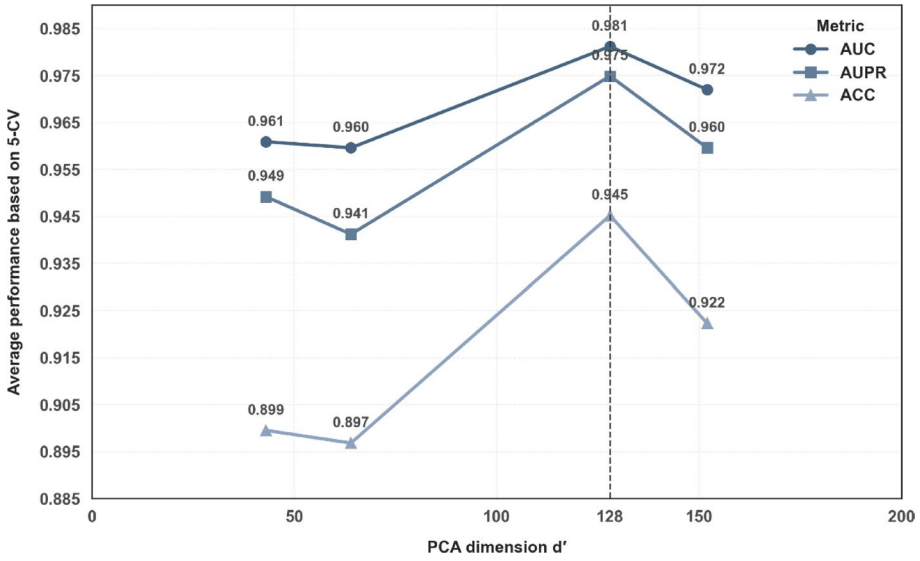

Fig. 2 Performance analysis of PCA dimension d'

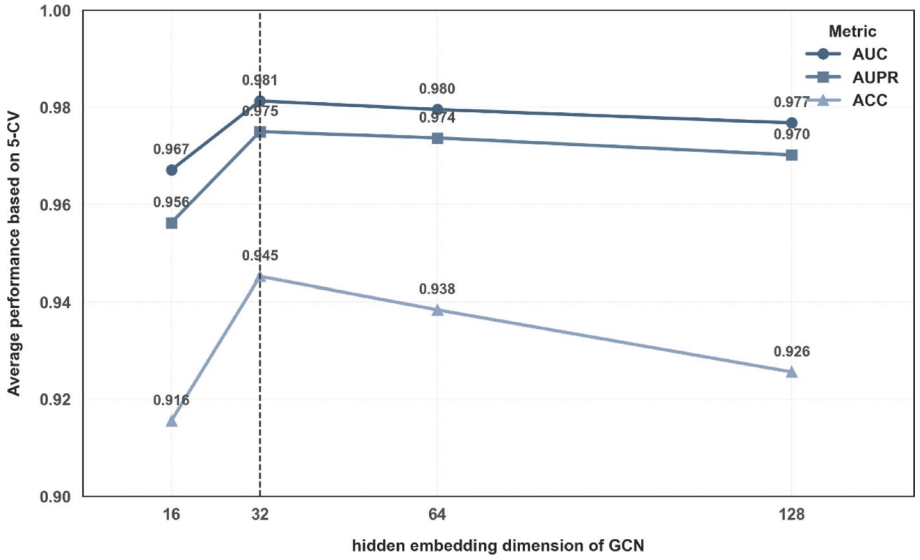

Fig. 3 Performance analysis on the hidden embedding dimension of GCN

Ablation study

To evaluate the performance of the model's core components, we design three ablation configurations and systematically analyze the contribution of each module through progressive removal or functional substitution. All configurations are trained under the same protocol (identical datasets, hyper-parameters, and evaluation metrics) to ensure fair comparison. The following tests are conducted:

(i) PCA variant: This version removes the PCA module while keeping the rest of the model unchanged. As a critical component for dimensionality reduction, the PCA module projects high-dimensional input features into a lower-dimensional space while preserving their primary variance. Removing it disables effective compression

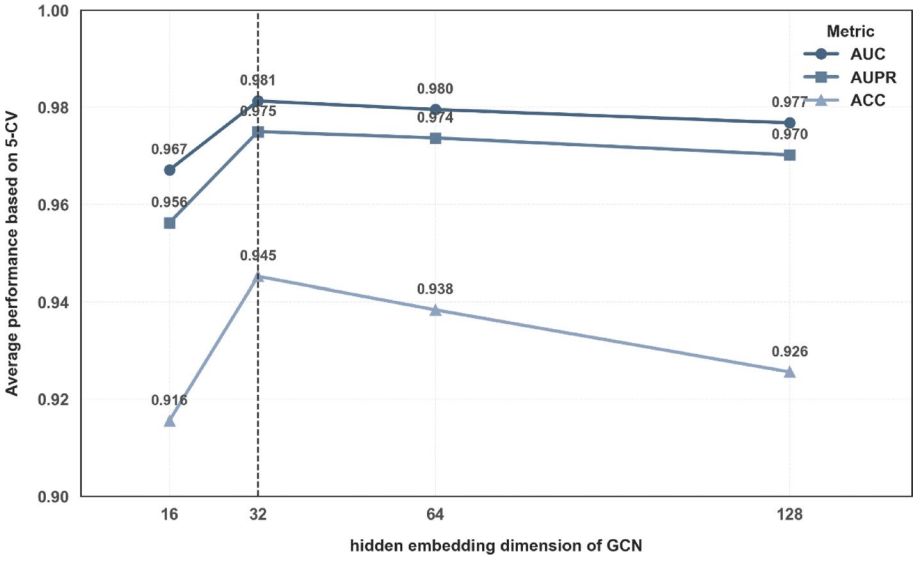

Fig. 4 Performance analysis on dropout value

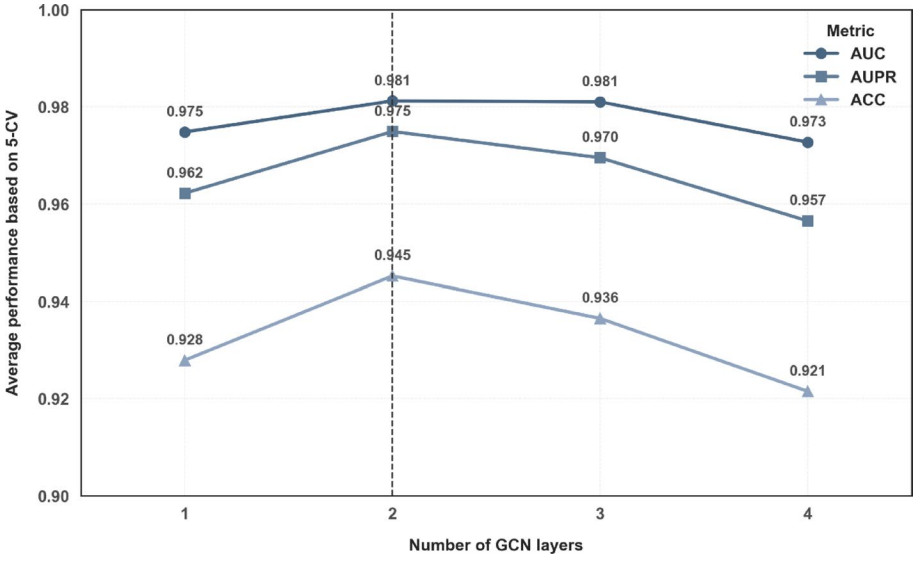

Fig. 5 Performance analysis on the number of GCN layers

based on principal components, which may negatively impact downstream graph learning and classification performance.

(ii) pL variant: In this setting, the subgraph construction module is removed and replaced with a standard fully connected layer. Eliminating this module means forfeiting the structural advantages introduced by graph representations, relying solely on dense node embeddings. This variant is designed to evaluate the importance of graph structure modelling versus conventional dense representations in the context of snoRNA-disease association prediction.

(iii) ATT variant: This version deletes the layer-wise attention mechanism prior to classification. Layer attention is used to assign varying weights to different graph convolution layers, enabling the model to focus on structurally informative layers.

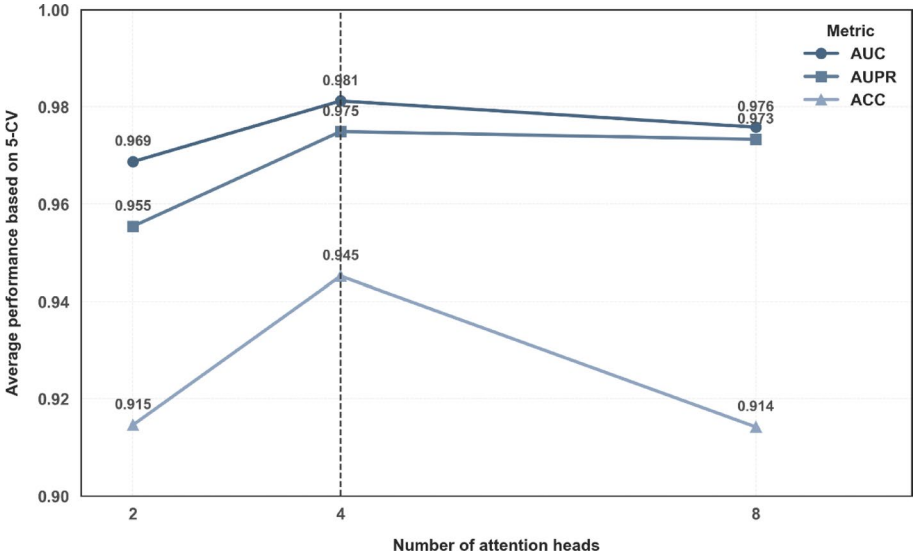

Fig. 6 Performance analysis on the number of attention heads

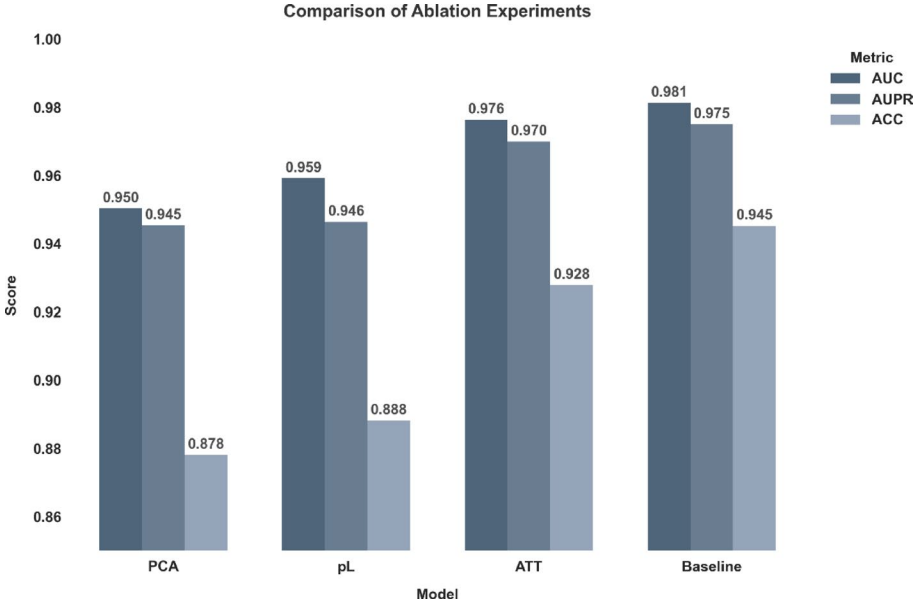

Fig. 7 Results of ablation study

Without it, our model treats all layer outputs equally, which may result in suboptimal feature fusion and reduced predictive accuracy.

Figure 7 shows the performance comparison of the three variants. The PCA variant yields an AUC of 0.950, AUPR of 0.945, and ACC of 0.878, significantly lower than the complete model, which achieves 0.981, 0.975, and 0.945 on the same metrics. This drop confirms the essential role of PCA-based dimensionality reduction in preserving key variance directions for effective feature compression.

The pL variant, which replaces the subgraph construction mechanism with a fully connected layer, shows moderate improvement over the PCA variant (AUC = 0.959, AUPR = 0.946, ACC = 0.888), but still underperforms relative to the full model. This

suggests that while dense representations can retain some predictive signal, graph-based relational encoding provides a more discriminative structure for capturing snoRNA-disease relationships.

The ATT variant, which omits the layer-attention mechanism, receives relatively strong performance (AUC = 0.9763, AUPR = 0.9699, ACC = 0.9279), though still consistently below the full model. This indicates that layer-level attention contributes to the final prediction by adaptively emphasizing structurally informative layers during representation fusion.

Taken together, these results demonstrate that all the three components contribute meaningfully to our model's overall performance.

Impact of different fusion strategies

We investigate three different graph construction strategies (pLmax, pLave and pLmin) to determine the final adjacency matrix. Specifically, pLmax adopts an element-wise maximum fusion to retain the strongest connections emphasized by either metric, which is effective in highlighting highly confident interactions but may suppress weaker yet informative links. pLave integrates the two similarity views by averaging them, providing a balanced graph structure while potentially smoothing out more complex relational patterns. In contrast, pLmin applies an element-wise minimum fusion, enforcing consistency across different similarity measures and yielding a more conservative graph construction. These strategies regulate the graph structure by combining the cosine similarity graph and the Manhattan distance graph, ultimately influencing the model's performance.

Figure 8 presents the performance of the three graph construction strategies based on their impact on the final graph structure and downstream prediction accuracy. Among them, pLmin obtains the best overall results, with an AUC of 0.981, AUPR of 0.975, and ACC of 0.945. The results show the effectiveness of a conservative, element-wise minimum strategy that enforces consistency across different similarity metrics.

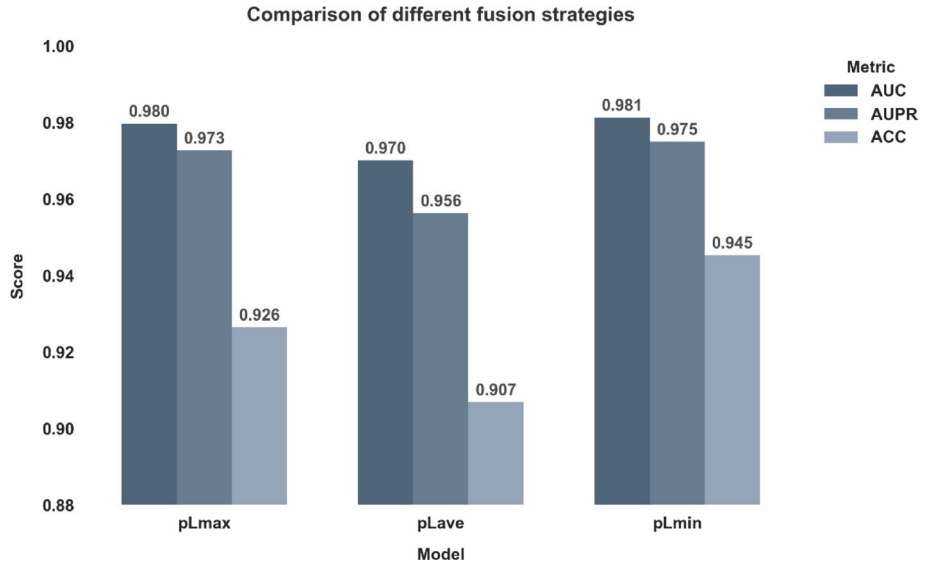

Fig. 8 Results of different fusion strategies

Statistical significance analysis

To ensure that the observed performance differences among models are not merely due to random fluctuations, we conduct a statistical significance analysis based on five-fold cross-validation results. As reference [29] emphasized, relying solely on average performance scores (e.g., AUC or AUPR) can be misleading, and conclusions about model superiority should be grounded in robust statistical tests. The resulting p-value heatmaps are illustrated in Figs. 9 and 10.

Our model SPGA consistently outperforms all competing baseline methods, including IGCNSDA, iPiDA_GCN, VGAMF, NSAMDA, NIMCGCN, and AMHMDA, with p-values well below the 0.001 threshold in both AUC and AUPR metrics, indicating highly significant performance differences. These results confirm that SPGA not only exhibits superior and stable predictions but also provides statistically validated improvements over these models.

Robustness analysis under negative sampling imbalance

In our prediction setting, negative examples are far more abundant than known positives, and improper handling of this imbalance can degrade model reliability. To assess the stability of SPGA under such conditions, we conduct experiments by progressively increasing the ratio of negative to positive samples in the training set. The test set remains fixed with a balanced ratio to ensure consistent evaluation.

We select DK-SDA as the comparative baseline. As illustrated in Fig. 11, both SPGA and DK-SDA suffer performance degradation as the level of imbalance increases. However, SPGA consistently maintains superior performance across all sampling

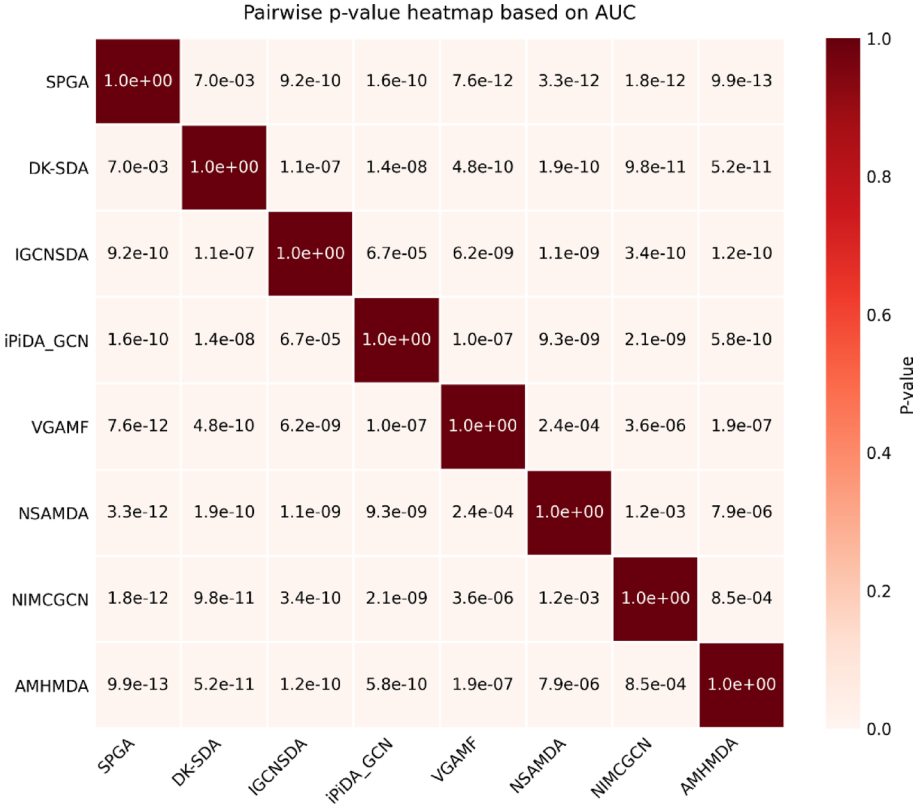

Fig. 9 p-value heatmap of pairwise statistical significance tests based on AUC

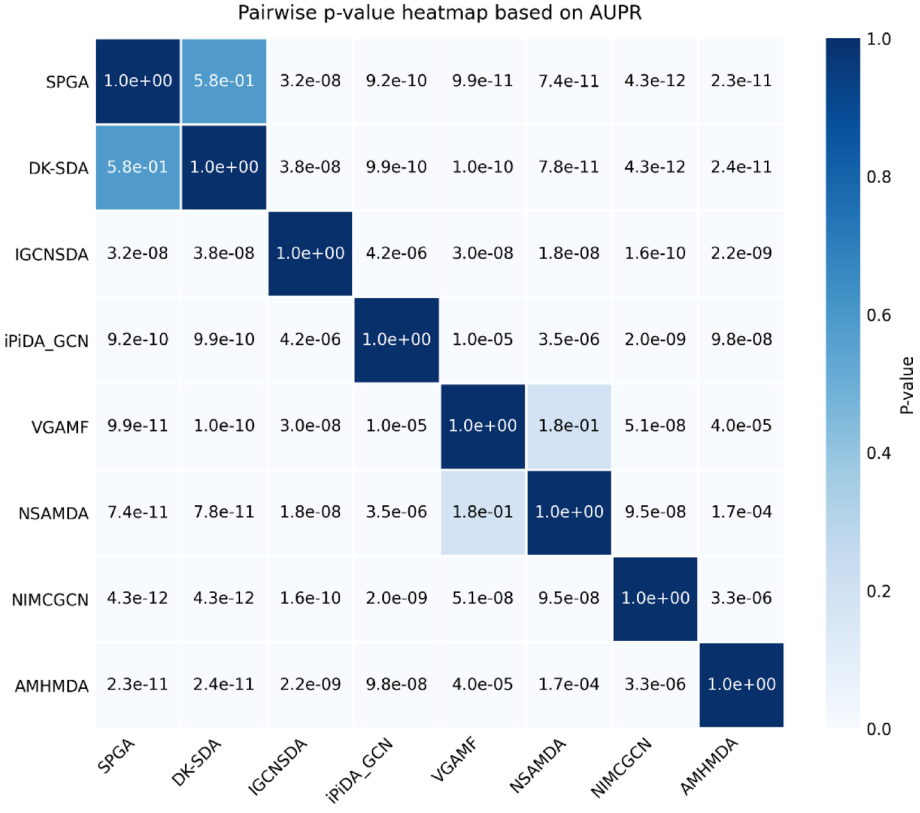

Fig. 10 *p*-value heatmap of pairwise statistical significance tests based on AUPR

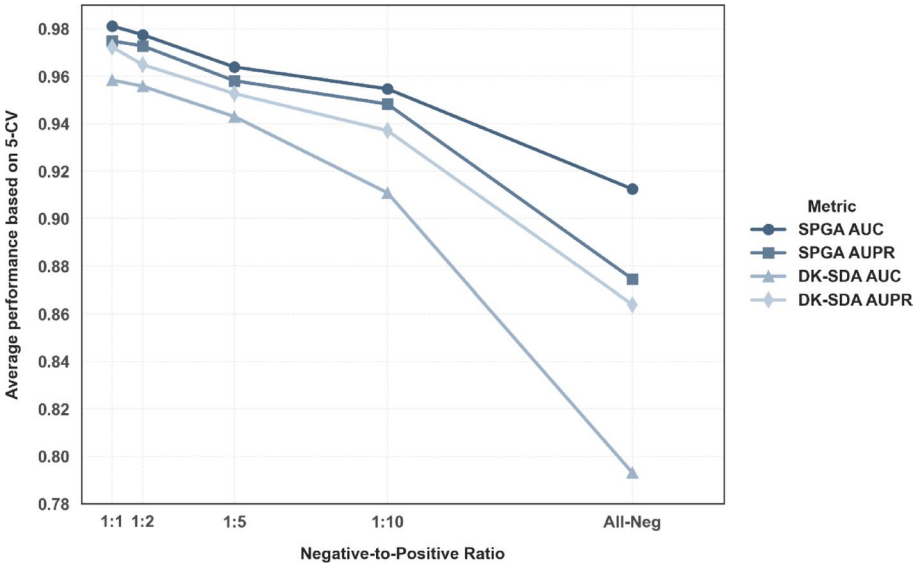

Fig. 11 Performance comparison between SPGA and DK-SDA under different negative-to-positive sample ratios

configurations. This trend becomes particularly evident under extreme imbalance, where the baseline model showed a more severe decline. These results suggest that SPGA is less susceptible to noise introduced by excessive negative instances and better preserves informative patterns when facing heavily skewed training distributions. Such resilience may stem from its structure-aware representation learning and adaptive attention

mechanisms, which together enhance the model's ability to distinguish true associations from spurious ones.

Robustness analysis under positive sample reduction

We further conduct a complementary experiment by reducing the number of positive samples in the training set while leaving the negative samples and the test set unchanged. This setting simulates real-world scenarios where known disease associations are scarce and difficult to obtain.

As depicted in Fig. 12, both SPGA and DK-SDA show declining performance as fewer positive samples are kept for training. Nevertheless, SPGA demonstrates more stable trends across most conditions and preserves its prediction advantage. These results indicate that SPGA can effectively leverage limited labeled data and maintain discriminative power under sparse supervision, reinforcing its practical utility in data-scarce biomedical environments.

Case study

To assess the reliability of SPGA's predictions in realistic scenarios, we conduct case studies involving two diseases (gastric cancer and colorectal cancer) and two snoRNAs (SNORD89 and SNORD104). For the two diseases, we first remove all known associations related to each of the target diseases, then rank all candidate snoRNAs in descending order by prediction scores and select the top 20 candidates for confirmation. A similar procedure is applied to the two snoRNAs, where the top 10 predicted diseases are selected for validation. To confirm our predictions, we first consult the RNADisease v4.0 [27] database as the primary reference source, and use the PubMed literature repository as a supplementary resource.

Gastric cancer is a malignant tumor originating from the gastric mucosal epithelium and is highly prevalent in East Asia, particularly in China. Known risk factors include Helicobacter pylori infection, high-salt diets, and tobacco use. As shown in Table 3, among the top 20 predicted snoRNA candidates associated with gastric cancer, all are

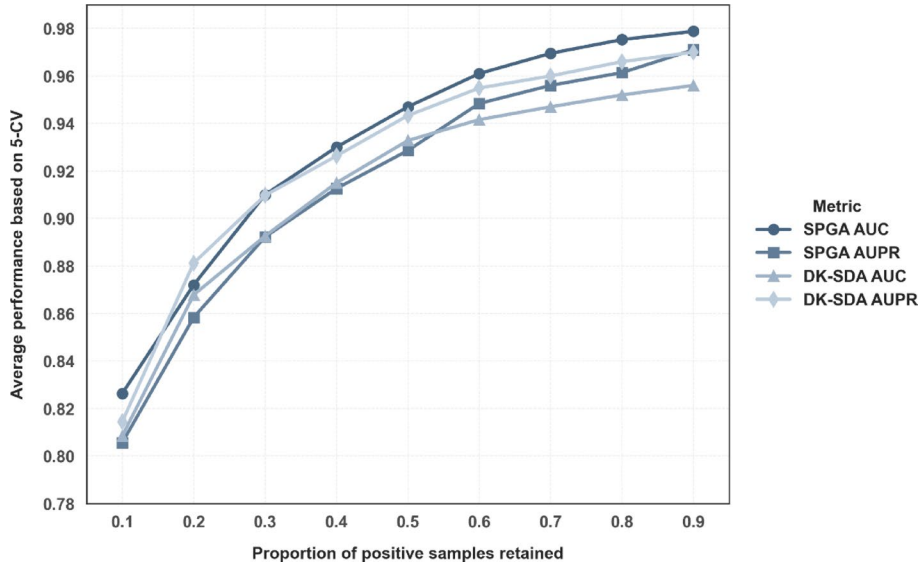

Fig. 12 Performance comparison between SPGA and DK-SDA under different positive training ratios

Table 3 The top 20 predicted snoRNAs associated with gastric cancer

Rank	snoRNA	Evidence	Rank	snoRNA	Evidence
1	SNORD3D	RNADisease	11	SNORD113-8	RNADisease
2	SNORD99	RNADisease	12	SNORD66	RNADisease
3	SNORD27	RNADisease	13	SNORD73A	RNADisease
4	SNORD25	RNADisease	14	SNORD59B	RNADisease
5	SNORD38B	RNADisease	15	SNORD72	RNADisease
6	SNORD19B	RNADisease	16	SNORD77	RNADisease
7	SNORD82	RNADisease	17	SNORD75	RNADisease
8	SNORD52	RNADisease	18	SNORD104	RNADisease
9	SNORD60	RNADisease	19	SNORD78	RNADisease
10	SNORD37	RNADisease	20	SNORA37	PMID:39815331

Table 4 The top 20 predicted snoRNAs associated with colorectal cancer

Rank	snoRNA	Evidence	Rank	snoRNA	Evidence
1	SNORA24	RNADisease	11	SNORD44	PMID: 28530127
2	SNORA9	RNADisease	12	SNORA15	Unconfirmed
3	SNORA17B	RNADisease	13	SNORA48	RNADisease
4	SNORA68	RNADisease	14	SNORD1C	Unconfirmed
5	SNORD126	RNADisease	15	SNORD74	Unconfirmed
6	SNORA50C	RNADisease	16	SNORD123	RNADisease
7	SNORD12C	RNADisease	17	SNORA70	RNADisease
8	SNORA70C	RNADisease	18	SNORA21	RNADisease
9	SNORD76	RNADisease	19	SNORA29	RNADisease
10	SNORD78	RNADisease	20	SNORD42	RNADisease

successfully confirmed. For example, as found in the reference [30], SNORA37 promotes the growth, invasion, and metastasis of cancer cells both in vitro and in vivo. This gene has been identified as an upregulated snoRNA, which is crucial for the tumorigenesis and invasiveness of cancer.

Colorectal cancer is a malignant tumour that occurs in the colon or rectum and has a high global incidence rate. Contributing risk factors include high-fat, low-fibre diets, obesity, and genetic predisposition. As shown in the Table 4, among the top 20 predicted snoRNA candidates, 17 are successfully identified, with 16 validated by the RNADisease v4.0 [27] database and 1 supported by evidence from PubMed. Although SNORA15 and SNORD1C have not yet been directly confirmed as colorectal cancer-associated snoRNAs in existing databases, existing studies offer some supporting clues. For example, reference [31] identified differentially expressed snoRNAs in cancer (CRC) through microarray and qPCR analysis, indicating that SNORA15 may play a role in the progression of rectal cancer. Similarly, reference [32] used qRT PCR to detect the expression of SNORD1C in the serum of CRC patients and identified it as a potential non-invasive biomarker for the diagnosis of colorectal cancer. These findings further support the predictive potential of our model in uncovering novel and biologically meaningful associations.

SNORD89 is a C/D box small nucleolar RNA that forms a snoRNP complex through its conserved C box (RUGAUGA) and D box (CUGA) motifs, guiding site-specific modification of rRNA [33]. Increased expression of SNORD89 has been associated with poor prognosis in ovarian cancer patients [34]. Similarly, SNORD104 has been shown to promote the growth of endometrial cancer both in vivo and in vitro [35].

Table 5 The top 10 predicted diseases associated with SNORD89

Rank	Disease	Evidence	Rank	Disease	Evidence
1	Gastric cancer	RNADisease	6	Rheumatoid arthritis	RNADisease
2	Dysregulated immunity	RNADisease	7	Breast cancer	RNADisease
3	Prostate cancer	RNADisease	8	Fetal growth restriction	Unconfirmed
4	Lung cancer	RNADisease	9	Intracranial aneurysm	Unconfirmed
5	Multiple sclerosis	RNADisease	10	Cerebral cavernous malformation	Unconfirmed

Table 6 The top 10 predicted diseases associated with SNORD104

Rank	Disease	Evidence	Rank	Disease	Evidence
1	Multiple sclerosis	RNADisease	6	Rheumatoid arthritis	RNADisease
2	Dysregulated Immunity	RNADisease	7	Tuberculosis	RNADisease
3	Gastric cancer	RNADisease	8	Oral squamous cell carcinoma	Unconfirmed
4	Hyperlipidemia	RNADisease	9	Sarcoma	Unconfirmed
5	Lung cancer	RNADisease	10	Alzheimer disease	RNADisease

Table 7 Information of the two benchmark datasets

Data sets	Number of snoRNAs	Number of diseases	Number of snoRNA-disease associations
RNADisease	471	84	1095
ncRPheno	13	82	439

Table 5 shows that, among the top 10 predicted disease candidates for SNORD89, 7 are confirmed in the RNADisease v4.0 [27] database. The remaining three Fetal Growth Restriction, Intracranial Aneurysm, and Cerebral Cavernous Malformation have not yet been validated but may warrant further investigation. As shown in Table 6, in the fourth case study focusing on SNORD104, 8 out of the top 10 predictions are validated by RNADisease v4.0 [27], with only Oral Squamous Cell Carcinoma and Sarcoma lacking direct evidence. These results further demonstrate our model's robustness in identifying high-confidence, biologically possible snoRNA–disease associations.

Methods

Datasets

In this study, two benchmark datasets provided in reference [25] is used for performance evaluation and comparison. As shown in Table 7, the first dataset was selected from the RNADisease v4.0 [27] repository. After filtering experimentally verified and predicted association data, and removing redundant records and missing values, a standardized dataset was obtained containing 471 snoRNAs, 84 diseases, and 1,095 associations. This dataset is randomly split into a training subset (80%) and a testing subset (20%) for model test. To validate the generalization ability of our model, another dataset was extracted from the ncRPheno [28] database, which contained 82 diseases, 13 snoRNAs, and 439 associations. It also serves as an external validation set for independent tests.

Method architecture

In this part, we illustrate the workflow of SPGA in predicting snoRNA–disease associations. As shown in Fig. 13, our model consists of four primary components: the feature construction module, graph structure modeling module, hierarchical attention fusion module, and association prediction module.

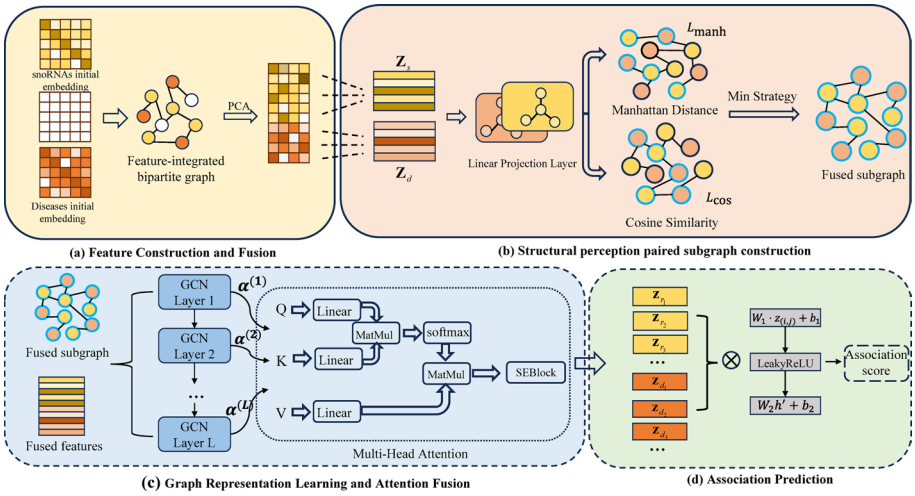

Fig. 13 The overview of SPGA in snoRNA-disease association prediction

Feature construction and fusion

In this part, we assign initial structural features to each snoRNA and disease node, according to previous research [26]. Specifically, we utilize the initial feature matrices for snoRNAs and diseases, where each matrix encodes the structural, attribute, or semantic information of snoRNAs and diseases, rather than arbitrary pairwise similarities or interactions. To integrate the feature sets of these two types of nodes into a unified representation space, we construct a block diagonal matrix $\mathbf{F} \in \mathbb{R}^{(m+n)\times(m+n)}$, as shown below:

$$\mathbf{F} = \begin{bmatrix} \mathbf{X}_{\text{disease}} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{\text{snoRNA}} \end{bmatrix} \tag{3}$$

where m and n represent the number of disease and snoRNA nodes, respectively, $\mathbf{X}_{\text{disease}} \in \mathbb{R}^{m\times m}$ and $\mathbf{X}_{\text{snoRNA}} \in \mathbb{R}^{n\times n}$ denote the disease–disease and snoRNA–snoRNA feature matrices, and $\mathbf{F} \in \mathbb{R}^{(m+n)\times(m+n)}$ is the resulting block-diagonal fused feature matrix in which the off-diagonal regions are filled with zeros. This zero-padding fusion strategy differs from reference [36] of cross-concatenating association relationships. We intentionally avoid introducing any prior interactions between snoRNAs and diseases during the feature fusion phase. This is mainly because zero-padding clearly delineates the feature spaces of snoRNAs and diseases, ensuring that cross-type information is not injected at the input stage, thus making subsequent graph-based structure learning more controllable. Additionally, separating intra-modal feature modeling from inter-modal structural reasoning helps the subsequent graph neural network focus on structural awareness and relationship learning, rather than relying on static cross-domain input matrices that explicitly encode snoRNA–disease interactions.

To compress the dimensions, improve computational efficiency, and enhance the stability of downstream learning, we apply Principal Component Analysis (PCA) to the feature matrix $\mathbf{F}$, projecting it from its original dimensionality to a lower-dimensional subspace. Let the dimensionality be d'. We first centralize the input feature matrix $\mathbf{F}$ by subtracting the column-wise mean vector $\boldsymbol{\mu}$, then compute the covariance matrix and perform eigen decomposition in one unified step, as follows:

$$\mathbf{C} = \frac{1}{m+n}(\mathbf{F} - \mathbf{1} \cdot \boldsymbol{\mu}^\top)^\top (\mathbf{F} - \mathbf{1} \cdot \boldsymbol{\mu}^\top) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top \tag{4}$$

Here, $\mathbf{C}$ is the covariance matrix, and $\mathbf{U}$, $\boldsymbol{\Lambda}$ represent the eigenvectors and eigenvalues of $\mathbf{C}$, respectively. We select the eigenvectors corresponding to the top d' principal components to form the projection matrix $\mathbf{P} \in \mathbb{R}^{(m+n)\times d'}$, and then perform linear dimensionality reduction on the original features:

$$\mathbf{Z} = (\mathbf{F} - \mathbf{1} \cdot \boldsymbol{\mu}^\top) \cdot \mathbf{P} \tag{5}$$

The resulting matrix $\mathbf{Z} \in \mathbb{R}^{(m+n)\times d'}$ represents the unified low-dimensional representation of the fused snoRNA and disease nodes, serving as the input for the subsequent graph embedding and relationship modeling modules.

Structure-aware embedding generation

After obtaining the low-dimensional joint feature matrix $\mathbf{Z}$, we further perform entity separation and embedding mapping. Since the first *m* rows correspond to disease nodes and the last *n* rows correspond to snoRNA nodes, we split $\mathbf{Z}$ into two parts:

$$\mathbf{Z}_{\text{disease}} = \mathbf{Z}(1:m,:), \quad \mathbf{Z}_{\text{snoRNA}} = \mathbf{Z}(m+1:m+n,:) \tag{6}$$

To enhance the representation ability and adapt to subsequent graph modeling tasks, we introduce a linear projection layer based on this split:

$$\mathbf{H}_{\text{disease}} = \sigma(\mathbf{Z}_{\text{disease}}\mathbf{W}) \in \mathbb{R}^{m\times d}, \quad \mathbf{H}_{\text{snoRNA}} = \sigma(\mathbf{Z}_{\text{snoRNA}}\mathbf{W}) \in \mathbb{R}^{n\times d} \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{d'\times d}$ is the weight matrix for the linear projection, $\sigma(\cdot)$ is the activation function, and *d* represents the final embedding dimension used for graph modeling. Notably, $\mathbf{H}_{\text{snoRNA}}$and $\mathbf{H}_{\text{disease}}$are node-level embeddings shared across all candidate pairs, i.e., the same snoRNA or disease keeps an identical base representation regardless of which counterpart it is paired with.

Construction of paired subgraphs

Inspired by graph neural modeling methods such as AMHMDA [21], we propose a local subgraph construction strategy based on paired entities to capture the structurally-aware interactive relationships between snoRNAs and diseases. Unlike the hypergraph modeling mechanism used in AMHMDA, we avoid using virtual hypernodes for structural completion. Instead, for each snoRNA-disease entity pair (r_i, d_j), we construct a local graph consisting solely of the real nodes in that pair, thereby improving the accuracy and controllability of structural representation. Specifically, given the snoRNA representation $\mathbf{h}_{r_i} \in \mathbb{R}^d$ and the disease representation $\mathbf{h}_{d_j} \in \mathbb{R}^d$ obtained from the previous stage, we concatenate them to form a node feature matrix:

$$\mathbf{X}_{(i,j)} = \begin{bmatrix} \mathbf{h}_{r_i} \\ \mathbf{h}_{d_j} \end{bmatrix} \tag{8}$$

Where $X(i,j) \in \mathbb{R}^{2\times d}$ represents the node feature matrix of the local paired subgraph constructed specifically for the (i,j)-th snoRNA-disease candidate, and $\mathbf{X}_{(i,j)}$ is constructed by selecting the corresponding two node embeddings from the global

embedding tables, rather than learning independent features for each pair. Subsequently, we construct the adjacency matrix for these nodes from two structural perspectives.

(1) Cosine similarity graph.

We define the cosine adjacency matrix as:

$$\mathbf{A}_{\cos}[p,q] = \frac{(\mathbf{x}_p, \mathbf{x}_q)}{\| \mathbf{x}_p \| \cdot \| \mathbf{x}_q \| + \varepsilon} \tag{9}$$

Where $\mathbf{x}_p$ and $\mathbf{x}_q$ denote the p-th and q-th row vectors of $X(i,j)$ in Eq. (8), and ε is a small constant to avoid numerical instability. Then, symmetric normalization is applied to obtain the propagation matrix of the cosine graph:

$$\mathbf{L}_{\cos} = \mathbf{D}_{\cos}^{-1/2} \mathbf{A}_{\cos} \mathbf{D}_{\cos}^{-1/2} \tag{10}$$

Where$\mathbf{D}_{\cos} = \operatorname{diag}\left(\sum_q \mathbf{A}_{\cos}[p,q]\right)$.

(2) Manhattan distance graph.

We define the weight plot based on L1 distance as:

$$\mathbf{A}_{\text{manh}}[p,q] = \operatorname{ReLU}\left(1 - \frac{\| \mathbf{x}_p - \mathbf{x}_q \|_1}{\| \mathbf{x}_p \|_1 + \| \mathbf{x}_q \|_1 + \varepsilon}\right) \tag{11}$$

The normalized distance here reflects the proximity between nodes. The ReLU activation ensures non-negativity, and then symmetric normalization is applied:

$$\mathbf{L}_{\text{manh}} = \mathbf{D}_{\text{manh}}^{-1/2} \mathbf{A}_{\text{manh}} \mathbf{D}_{\text{manh}}^{-1/2} \tag{12}$$

where $\mathbf{D}_{\text{manh}} = \operatorname{diag}\left(\sum_q \mathbf{A}_{\text{manh}}[p,q]\right)$.

To integrate the two structural perspectives mentioned above, we use an element-wise minimum strategy to generate the final graph structure:

$$\mathbf{L}_{(i,j)} = \min\left(\mathbf{L}_{\cos}, \mathbf{L}_{\text{manh}}\right) \tag{13}$$

Graph representation learning and attention fusion

After constructing the local subgraph for each snoRNA–disease node pair, we input its structural graph $\mathbf{L}_{(i,j)} \in \mathbb{R}^{2\times 2}$ along with node features $\mathbf{X}_{(i,j)} \in \mathbb{R}^{2\times d}$ into the graph representation learning module to extract structurally-aware interaction representations. This module consists of multiple layers of Graph Convolutional Networks (GCNs), augmented with a residual connection mechanism and multi-head attention fusion module to enhance representation diversity and stability. Notably, this module learns a parameterized representation function with shared weights to produce pair-conditional interaction-aware embeddings from the local subgraph, while the underlying snoRNA or disease node embeddings remain reusable across different candidate pairs.

We use L-layer residual GCN to represent the nodes in the local graph, with the update rule for the graph convolution layer as follows:

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{L}_{(i,j)} \mathbf{H}^{(l)} \mathbf{W}^{(l)}\right) + \mathbf{H}^{(l)} \quad (l = 0, 1, \ldots, L-1) \tag{14}$$

where $\mathbf{H}^{(0)} = \mathbf{X}_{(i,j)}$ represents the initial node features, $\mathbf{W}^{(l)}$ is the learnable parameters of the l-th layer, $\sigma(\cdot)$ is the nonlinear activation function, and the residual connection $+\mathbf{H}^{(l)}$ helps mitigate the vanishing gradient problem and enhances feature flow.

The output of each GCN layer is temporarily stored as intermediate features representing the multi-layer structure awareness.

To fully exploit the hierarchical semantic differences in the multi-layer GCN outputs, we introduce a multi-head attention mechanism to fuse the outputs from each layer. Let the output of the l-th layer GCN be $\mathbf{H}^{(l)} \in \mathbb{R}^{2\times d}$, and we stack it into a 3D tensor:

$$\mathcal{H} = \left[\mathbf{H}^{(0)}, \mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(L)}\right] \in \mathbb{R}^{2\times(L+1)\times d} \tag{15}$$

For each attention head $h \in \{1, \ldots, H\}$, we define the attention weight computation as:

$$\alpha^{(h)} = \mathrm{Softmax}\left(\frac{Q^{(h)}K^{(h)\top}}{\sqrt{d/H}}\right), \quad Q^{(h)} = \mathcal{H}\mathbf{W}_Q^{(h)}, K^{(h)} = \mathcal{H}\mathbf{W}_K^{(h)} \tag{16}$$

The final output is obtained by the weighted fusion of multiple attention heads, represented as the concatenation of their results:

$$Z_{attn} = SEBlock\left(\|_{h=1}^{H}\, \alpha^{(h)} \cdot \left(\mathrm{H}W_V^{(h)}\right) \|\right) \tag{17}$$

where $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{d\times d/H}$ is the projection matrix for each attention head, and $\mathrm{SEBlock}(\cdot)$ represents the channel attention compression mechanism, used to suppress redundant channels and enhance important structural dimensions. Specifically, $\mathrm{SEBlock}(X) = g \odot X$, where $g = \mathrm{Sigmoid}\left(W_2\mathrm{ReLU}\left(W_1\mathrm{GAP}(X)\right)\right)$ is a channel-wise gate produced by global average pooling and a two-layer MLP.

Prediction and optimization

The results after attention fusion are separated by node types (snoRNAs and diseases), and their respective embedding representations $\mathbf{z}_{r_i}$ and $\mathbf{z}_{d_j}$ are taken for the subsequent association prediction module. To capture the relational features between the two nodes in the embedding space, we use element-wise multiplication to fuse the representations of snoRNA and disease:

$$\mathbf{z}_{(i,j)} = \mathbf{z}_{r_i} \odot \mathbf{z}_{d_j} \in \mathbb{R}^d \tag{18}$$

The fused representation is then input into an improved multi-layer perceptron with a two-layer structure to perform the final probability prediction:

$$s_{i,j} = \mathrm{Sigmoid}\left(\mathbf{W}_2 \cdot \mathrm{LeakyReLU}\left(\mathrm{BN}\left(\mathbf{W}_1 \cdot \mathbf{z}_{(i,j)} + \mathbf{b}_1\right)\right) + \mathbf{b}_2\right) \tag{19}$$

where $\mathrm{BN}(\cdot)$ denotes batch normalization, and $s_{i,j} \in (0, 1)$ is the predicted association probability for the (i, j)-th snoRNA–disease pair. We use the Binary Cross-Entropy (BCE) loss function as the model optimization objective, formulated as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{(i,j)}\left[y_{i,j}\log s_{i,j} + (1 - y_{i,j})\log(1 - s_{i,j})\right] \tag{20}$$

where $y_{i,j} \in \{0, 1\}$ indicates whether the snoRNA–disease pair has a true association, and N is the number of training samples. For the optimizer selection, we use the Adam optimizer for gradient updates.

Discussion

As a category of critical biomarkers, accurate identification of disease-related snoRNAs would provide huge benefits to drug discovery and disease treatment. Our proposed computational framework SPGA effectively addresses key challenges in the field, including data scarcity and the need for robust performance. Previous approaches often failed to capture the high-dimensional and complex nature of molecular interactions. Our use of deep learning, particularly with graph representations, addresses this limitation by learning rich and hierarchical representations directly from the data. Comprehensive experimental evaluation demonstrates that SPGA has achieved substantial improvements in prediction performance.

The superior performance of SPGA resides in its graph-based architecture. Initially, it integrates intrinsic structural features of snoRNAs and diseases. PCA is then applied as an unsupervised dimensionality reduction step to improve efficiency and stabilize downstream learning, without using any snoRNA–disease association edges or labels. For each candidate snoRNA-disease pair, a localized subgraph is constructed. This graph's structure is uniquely defined by computing the element-wise minimum of two similarity metrics-cosine similarity and Manhattan distance. This strategy emphasizes only the most consistent and robust structural interdependencies. Subsequently, the graph is processed by a residual GCN to learn node representations. A hierarchical attention mechanism further fuses features derived from different GCN layers, thereby capturing multi-scale structural information.

However, certain limitations need to be considered. While our model is designed to handle data scarcity, the inherent limitation of the available annotated snoRNA-disease associations still imposes constraints on model generalization. The 'black-box' nature of deep learning models, while powerful for prediction, presents a challenge in providing direct, intuitive biological explanations for individual predictions, necessitating experimental validation. Furthermore, the accuracy of SPGA is, to some extent, dependent heavily on graph structure, and the known experimentally supported snoRNA-disease associations are rare. Therefore, integrating other data sources, such as expression profiles, might improve model performance. Future work will focus on the integration of more reliable biomedical data relevant to snoRNAs and diseases to further enhance the robustness and applicability of our model.

Conclusions

We present a novel structure-aware deep learning framework SPGA for predicting associations between snoRNAs and human diseases in this study. Extensive experimental tests have demonstrated that SPGA substantially outperforms seven state-of-the-art baseline methods on both a primary dataset (RNADisease) and an independent external dataset (ncRPheno). Our model's effectiveness and the specific contributions of its principal components are rigorously examined through hyperparameter analysis, ablation studies, and statistical significance tests. Furthermore, case studies focusing on two cancers and two specific snoRNAs demonstrate its practical utility. Many of these top predictions are subsequently confirmed by the RNADisease database and existing literature. We expect that our study would provide a potent and robust computational tool to accelerate the discovery of novel snoRNA-disease relationships, offering significant potential for identifying new diagnostic biomarkers and therapeutic targets.

Abbreviations

snoRNAs	Small nucleolar RNAs
ncRNAs	Non-coding RNAs
rRNAs	Ribosomal RNAs
snRNAs	Small nuclear RNAs
AUC	Area under the ROC curve
AUPR	Area under the precision–recall curve
ACC	Accuracy
GCNs	Graph convolutional networks
PCA	Principal component analysis

Author contributions

H.C. conceived and designed this study. Z.S. implemented the experiments. H.C. and Z.S. analyzed the results. H.C. and Z.S. wrote the manuscript. Both authors read and approved the final manuscript.

Funding

This work was supported by National Natural Science Foundation of China (62562031), and Jiangxi Provincial Natural Science Foundation, China (20242BAB25083).

Data availability

The benchmark datasets and source code for this study are available at https://github.com/aisuhdiuasgd/SPGA . The original data is from RNADisease v4.0 (http://www.rnadisease.org/) and ncRPheno (http://lilab2.sysu.edu.cn/ncrpheno).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 25 November 2025 / Accepted: 4 March 2026

Published online: 25 March 2026

References

1. Kiss-László Z, Henry Y, Bachellerie J-P, Caizergues-Ferrer M, Kiss T. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. Cell. 1996;85(7):1077–88.
2. Bachellerie J-P, Cavaillé J, Hüttenhofer A. The expanding snoRNA world. Biochimie. 2002;84(8):775–90.
3. Khanna A, Stamm S. Regulation of alternative splicing by short non-coding nuclear RNAs. RNA Biol. 2010;7(4):480–5.
4. von Stedingk K, Koster J, Piqueras M, Noguera R, Navarro S, Påhlman S, Versteeg R, Øra I, Gisselsson D, Lindgren D. snoRNPs regulate telomerase activity in neuroblastoma and are associated with poor prognosis. Translational Oncol. 2013;6(4):447–IN446.
5. Fafard-Couture É, Jacques P-É, Scott MS. Motif conservation, stability, and host gene expression are the main drivers of snoRNA expression across vertebrates. Genome Res. 2023;33(4):525–40.
6. Holley CL, Topkara VK. An introduction to small non-coding RNAs: miRNA and snoRNA. Cardiovasc Drugs Ther. 2011;25(2):151–9.
7. Ono M, Scott MS, Yamada K, Avolio F, Barton GJ, Lamond AI. Identification of human miRNA precursors that resemble box C/D snoRNAs. Nucleic Acids Res. 2011;39(9):3879–91.
8. Duker AL, Ballif BC, Bawle EV, Person RE, Mahadevan S, Alliman S, Thompson R, Traylor R, Bejjani BA, Shaffer LG. Paternally inherited microdeletion at 15q11. 2 confirms a significant role for the SNORD116 C/D box snoRNA cluster in Prader–Willi syndrome. Eur J Hum Genet. 2010;18(11):1196–201.
9. Potemkin N, Cawood SM, Guévremont D, Mockett B, Treece J, Stanton J-AL, Williams JM. Whole transcriptome RNA-Seq reveals drivers of pathological dysfunction in a transgenic model of Alzheimer's disease. Mol Neurobiol 2025:1–21.
10. Su H, Xu T, Ganapathy S, Shadfan M, Long M, Huang TH, Thompson I, Yuan Z. Elevated snoRNA biogenesis is essential in breast cancer. Oncogene. 2014;33(11):1348–58.
11. Xiao L, Wang J, Ju S, Cui M, Jing R. Disorders and roles of tsRNA, snoRNA, snRNA and piRNA in cancer. J Med Genet. 2022;59(7):623–31.
12. Sletten AC, Davidson JW, Yagabasan B, Moores S, Schwaiger-Haber M, Fujiwara H, Gale S, Jiang X, Sidhu R, Gelman SJ. Loss of SNORA73 reprograms cellular metabolism and protects against steatohepatitis. Nat Commun. 2021;12(1):5214.
13. Liu B, Wu T, Miao BA, Ji F, Liu S, Wang P, Zhao Y, Zhong Y, Sundaram A. Zeng T-B: snoRNA-facilitated protein secretion revealed by transcriptome-wide snoRNA target identification. Cell. 2025;188(2):465–83. e422.
14. Bratkovič T, Božič J, Rogelj B. Functional diversity of small nucleolar RNAs. Nucleic Acids Res. 2020;48(4):1627–51.
15. Liu S, Zhu W, Wang P, Yu S, Wu F. Identifying associations between small nucleolar RNAs and diseases via graph convolutional network and attention mechanism. IEEE J Biomedical Health Inf 2024.
16. Sun Z, Huang Q, Yang Y, Li S, Lv H, Zhang Y, Lin H, Ning L. PSnoD: identifying potential snoRNA-disease associations based on bounded nuclear norm regularization. Brief Bioinform. 2022;23(4):bbac240.

17. Guan Z, Jin X, Zhang X. MFF-nDA: a computational model for ncRNA–disease association prediction based on multimodule fusion. J Chem Inf Model. 2025;65(7):3324–42.
18. Momanyi BM, Zhou Y-W, Grace-Mercure BK, Temesgen SA, Basharat A, Ning L, Tang L, Gao H, Lin H, Tang H. SAGESDA: multi-graphSAGE networks for predicting SnoRNA-disease associations. Curr Res Struct Biology. 2024;7:100122.
19. Zhang L, Chen M, Hu X, Deng L. Graph convolutional network and contrastive learning small nucleolar RNA (snoRNA) disease associations (GCLSDA): predicting snoRNA–disease associations via graph convolutional network and contrastive learning. Int J Mol Sci. 2023;24(19):14429.
20. Li J, Zhang S, Liu T, Ning C, Zhang Z, Zhou W. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics. 2020;36(8):2538–46.
21. Ning Q, Zhao Y, Gao J, Chen C, Li X, Li T, Yin M. AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA–disease associations identification. Brief Bioinform. 2023;24(2):bbad094.
22. Zhao H, Li Z, You Z-H, Nie R, Zhong T. Predicting Mirna-disease associations based on neighbor selection graph attention networks. IEEE/ACM Trans Comput Biol Bioinf. 2022;20(2):1298–307.
23. Hou J, Wei H, Liu B. iPiDA-GCN: Identification of piRNA-disease associations based on graph convolutional network. PLoS Comput Biol. 2022;18(10):e1010671.
24. Ding Y, Lei X, Liao B, Wu F-X. Predicting miRNA-disease associations based on multi-view variational graph auto-encoder with matrix factorization. IEEE J biomedical health Inf. 2021;26(1):446–57.
25. Hu X, Zhang P, Liu D, Zhang J, Zhang Y, Dong Y, Fan Y, Deng L. IGCNSDA: unraveling disease-associated snoRNAs with an interpretable graph convolutional network. Brief Bioinform 2024, 25(3).
26. Wang Z, Chen Y, Ma H, Gao H, Zhu Y, Wang H, Zhang N. Accurate identification of snoRNA targets using variational graph autoencoder to advance the redevelopment of traditional medicines. Front Pharmacol. 2025;15:1529128.
27. Chen J, Lin J, Hu Y, Ye M, Yao L, Wu L, Zhang W, Wang M, Deng T, Guo F. RNADisease v4. 0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. Nucleic Acids Res. 2023;51(D1):D1397–404.
28. Zhang W, Yao G, Wang J, Yang M, Wang J, Zhang H, Li W. ncRPheno: a comprehensive database platform for identification and validation of disease related noncoding RNAs. RNA Biol. 2020;17(7):943–55.
29. Reimers N, Gurevych I. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. arXiv 2018. *arXiv preprint arXiv:180309578*.
30. Bao B, Tian M, Wang X, Yang C, Qu J, Zhou S, Cheng Y, Tong Q, Zheng L. SNORA37/CMTR1/ELAVL1 feedback loop drives gastric cancer progression via facilitating CD44 alternative splicing. J Experimental Clin Cancer Res. 2025;44(1):15.
31. Yang X, Li Y, Li L, Liu J, Wu M, Ye M. SnoRNAs are involved in the progression of ulcerative colitis and colorectal cancer. Dig Liver Disease. 2017;49(5):545–51.
32. Liu Y, Zhao C, Sun J, Wang G, Ju S, Qian C, Wang X. Overexpression of small nucleolar RNA SNORD1C is associated with unfavorable outcome in colorectal cancer. Bioengineered. 2021;12(1):8943–52.
33. Bao H-j, Chen X, Liu X, Wu W, Li Q-h, Xian J-y, Zhao Y, Chen S. Box C/D snoRNA SNORD89 influences the occurrence and development of endometrial cancer through 2′-O-methylation modification of Bim. Cell Death Discovery. 2022;8(1):309.
34. Zhu W, Niu J, He M, Zhang L, Lv X, Liu F, Jiang L, Zhang J, Yu Z, Zhao L. SNORD89 promotes stemness phenotype of ovarian cancer cells by regulating Notch1-c-Myc pathway. J translational Med. 2019;17(1):259.
35. Lu B, Chen X, Liu X, Chen J, Qin H, Chen S, Zhao Y. C/D box small nucleolar RNA SNORD104 promotes endometrial cancer by regulating the 2′-O-methylation of PARP1. J translational Med. 2022;20(1):618.
36. Tian Z, Han C, Xu L, Teng Z, Song W. Mgcnss: mirna–disease association prediction with multi-layer graph convolution and distance-based negative sample selection strategy. Brief Bioinform. 2024;25(3):bbae168.

Publisher's note