PERGAT: Pretrained Embeddings of Graph Neural Networks for miRNA-Cancer Association Prediction

Sa Li, Jonah Shader, Tianle Ma
Oakland University
{sa,jonahshader,tianlema}@oakland.edu

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822135

***Abstract*— MicroRNAs (miRNAs) play a crucial role in the regulation of gene expression and have been implicated in the pathogenesis of various cancers. Predicting miRNA-cancer associations is essential for understanding cancer mechanisms and developing therapeutic strategies. In this work, we introduce a computational method named PERGAT (Pretrained Embeddings based on the Residual Graph Attention Network) for predicting miRNA-cancer associations. Extensive experimental results demonstrate the superior performance of PERGAT compared to other state-of-the-art methods. We achieved the AUC as high as 0.9641 in terms of the area under the Receiver Operating Characteristic (ROC) curves, and the area under the precision-recall curves as 0.9377. Our approach leverages the structure of miRNA-disease networks to capture complex relationships efficiently and improves prediction accuracy. Additionally, PERGAT exhibits exceptional performance in three cancer case studies, underscoring its reliability for studying miRNA-disease associations. The data and source codes are available at https://github.com/ericsaliway/PERGAT.**

***Index Terms*—miRNA-disease association, link prediction, graph neural network, node embeddings.**

I. Introduction

MICRORNAS (miRNAs) are small non-coding RNAs that regulate gene expression by targeting messenger RNAs (mRNAs) and triggering their degradation or translational repression [2], [3]. Research has shown that abnormal miRNA expression is closely linked to the onset and progression of various diseases, including cancer, cardiovascular diseases, neurological disorders, and more [4]–[6]. miRNAs can regulate the expression of key genes, thereby affecting cellular processes such as proliferation, differentiation, and apoptosis, playing a crucial role in the mechanisms of disease development. Consequently, designing an effective method to predict potential associations between miRNAs and diseases is essential [7]–[9].

Graph-based learning approaches, such as Graph Neural Networks (GNNs) [10], exploit this structural information to enhance the learning process. These models are designed to capture the interactions between nodes, considering both local and global graph properties. By incorporating the rich information present in the graph, we can learn more accurate and robust hypotheses for tasks like node classification, link prediction, and graph classification. This graph-based approach often leads to improved performance compared to traditional methods that do not consider the underlying graph structure [11]–[17]. As a graph-based semi-supervised learning method, GNNs do not require labels for all nodes. This feature is particularly powerful for inferring miRNA-associated diseases, as many miRNAs have not been thoroughly investigated in relation to diseases [18]–[21]. Furthermore, a single miRNA can be associated with multiple diseases, allowing the prediction of disease-miRNA associations to be formulated as a multi-label classification problem [22].

In this paper, we introduce deep learning (unsupervised feature learning) [23] techniques into the training of GNN model on multi-omics data. We present PERGAT, a Pretrained Embeddings based on the Residual Attention Graph Neural Network for prediction of miRNA-cancer associations. An overview of PERGAT is shown in Figure 1.

This study makes four major contributions to the understanding of disease-miRNA associations:

- We incorporate link prediction methods to infer potential relationships in miRNA-disease network with pretrained

embeddings.

- Our approach includes embedding clustering techniques to group structurally similar miRNAs and diseases, enhancing the understanding of miRNA disease associations, and facilitating better understanding and visualization of graph structures.
- We propose a novel GNN model, PERGAT, which leverages residual connections and multi-head attention mechanisms to enhance graph node representation learning.
- We evaluate the model's performance through comprehensive experiments, highlighting its predictive accuracy and effectiveness in embedding clustering and link prediction.

II. Related Work

In [24], the authors introduce a Multi-view Multichannel Attention Graph Convolutional Network (MMGCN) to predict potential miRNA–disease associations. Unlike traditional multisource information integration methods, MMGCN uses a GCN encoder to independently capture features of miRNAs and diseases from various similarity views.

The framework presented in [25] explores three graph construction methods and investigates seven GCN models with four distinct graph pooling techniques. Another deep learning approach, EOESG [26], predicts potential miRNA-disease associations by leveraging embeddings within embeddings and a simplified convolutional network. This model integrates disease similarity, miRNA similarity, and the miRNA-disease association network to form a coupled heterogeneous graph. Li et al. [27] introduce PGCN, which leverages graph convolutional neural networks (GCNs) to prioritize genes associated with specific diseases. The method focuses on embedding both diseases and genes into a shared latent space where proximity in this space signifies potential disease-gene associations.

Other GCN-based models have been employed to learn gene feature representations by aggregating features from neighboring miRNA nodes in the network [28]–[30]. Furthermore, MAMFGAT [31] uses GAT as its core for feature aggregation and integrates a multi-modal adaptive fusion module to extract features, and incorporates multi-modal residual feature fusion to address the issue of excessive feature smoothing in GATs.

Previous efforts to predict miRNA-disease associations have utilized a range of machine learning and deep learning techniques, such as matrix factorization, random walks, and convolutional neural networks. However, to the best of our knowledge, the use of Graph Neural Networks (GNNs) for predicting miRNA-cancer associations through pretraining embeddings remains largely underexplored.

III. Methods

In this study, we introduce PERGAT, a Pretrained Embeddings approach utilizing a Residual Graph Attention Network to predict miRNA-cancer associations. PERGAT is developed with the objective of learning representation vectors for both miRNAs and diseases to enhance the prediction of disease-related miRNAs. The model is designed to preserve the original features of miRNAs and diseases, focusing on known miRNA-disease associations with differential expression analysis of miRNAs [32].

A. miRNA Enrichment Analysis

We use Fisher's exact test [33] to compute p-value for contingency table. The test evaluates the null hypothesis that the proportions of diseases associated with each miRNA are independent. Similar to miRNA enrichment tool miEAA [34], we perform the enrichment analysis directly at the level of miRNAs.

The hypergeometric probability that gives the probability of obtaining a specific arrangement of the contingency table is:

$$P(T) = \frac{\binom{K}{n}\binom{N-K}{M-n}}{\binom{N}{M}}$$

where:

- N: Total number of unique diseases.
- K: Number of diseases associated with miRNA$_i$.
- M: Number of diseases associated with miRNA$_j$.
- T: Number of shared diseases between miRNA$_i$ and miRNA$_j$.
- $\binom{K}{n}$: The binomial coefficient, defined as:

$$\binom{K}{n} = \frac{K!}{n!(K-n)!}$$

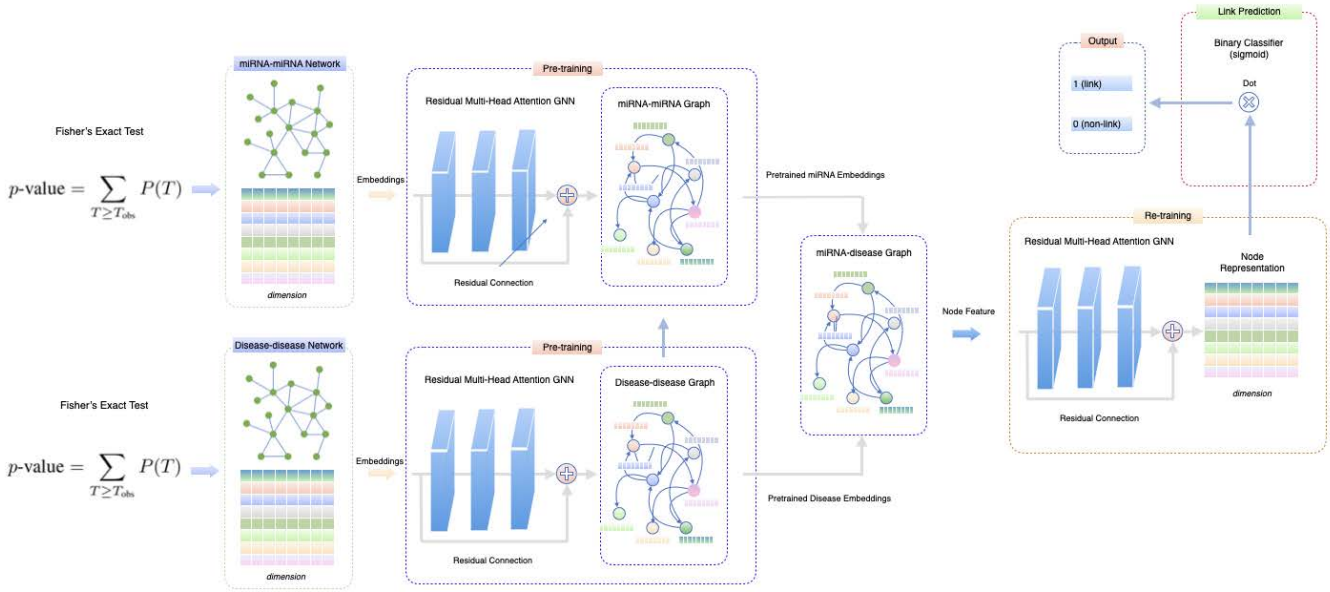


FIG. 1 – Overview of PERGAT framework. The framework follows four key steps. It begins by applying Fisher's Exact Test to miRNA-miRNA and disease-disease networks to identify significant associations. Next, it performs pre-training using a multi-head self-attention layer with residual connections. Finally, a link prediction task is performed using the multi-head self-attention layer, and an output prediction is generated with a fully connected layer to identify potential miRNA-disease links.

For the observed table T_{obs}, the p-value is the sum of the probabilities of all possible tables as following:

$$p\text{-value} = \sum_{T \geq T_{\text{obs}}} P(T)$$

B. Weighted Enrichment Graph Construction

The construction of the miRNA-disease association graph is based on the representation of miRNAs and diseases as nodes, and their associations as directed edges. For each miRNA-disease (cancer) association a_i, nodes and edges are added to the graph $G = (V, E)$ as follows:

1) Nodes:

$$V = \{\text{miRNA}_i, \text{disease}_i \mid i = 1, \ldots, n\}$$

2) Edges: Directed edges $E \subseteq V \times V$ are added as:

$$E = \{(\text{miRNA}_i, \text{disease}_i) \mid i = 1, \ldots, n\}$$

Each node $v \in V$ has the following attributes:

- p-value p_v: The adjusted p-value indicating the strength of the association.
- Significance s_v: A binary value indicating whether the p-value is statistically significant (`significant`) or not (`non-significant`).
- Type t_v: The type of the node, either `miRNA` or `disease`.

Each directed edge $e \in E$ has the following attributes:

- Weight w_e: The adjusted p-value p_e associated with the miRNA-disease pair, indicating the strength of the association.
- Significance s_e: A binary value indicating whether the association is statistically significant.

C. Residual Multi-Head Attention Graph Neural Network

In this research, a new GNN model, named Residual Multi-Head Attention Graph Neural Network (R-MHAGNN), was used for knowledge representation learning. The attention mechanism was implemented with a single linear transformation to aggregate node features from source and destination nodes, utilizing learnable parameters for attention and performing feature transformations via linear layers. Specifically, residual connections are integrated with enhanced attention mechanisms for graph data, and dropout regularization is used

to improve both the performance and robustness of the network when applied to the downstream tasks.

The update rule for the feature vector of node i at layer $l+1$ with multi-head attention is defined as follows:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}(i)}\text{Dropout}\left(\alpha_{ij,k}^{(l)}\right)W_k^{(l)}\text{Dropout}\left(h_j^{(l)}\right)+\mathbf{b}\right) + W_{res}h_i^{(l)} \tag{1}$$

where:

- $\mathbf{h}_i^{(l+1)}$: The hidden state of node i in the $(l+1)$-th layer.
- $\sigma(\cdot)$: The activation function, specifically $LeakyReLU$ in this case.
- $\sum_{j\in\mathcal{N}(i)}$: The sum over the neighbors j of node i.
- Dropout $\left(\alpha_{ij,k}^{(l)}\right)$: The dropout applied to the edge weight $\alpha_{ij,k}^{(l)}$ for the k-th head.
- Dropout $\left(h_j^{(l)}\right)$: The dropout applied to the hidden state $h_j^{(l)}$ of neighbor j.
- $W_k^{(l)}$: The weight matrix of the k-th attention head.
- W_{res}: The residual weight matrix applied to the hidden state of node i.
- $h_i^{(l)}$ and $h_j^{(l)}$: The feature vectors of node i and its neighbor j at layer l.
- K: The number of attention heads.
- $\mathbf{b}$: The bias vector.

The attention score calculation for the k-th head is defined as:

$$\alpha_{ij,k}^{(l)} = \text{softmax}_i\left(e_{ij,k}^{(l)}\right) \tag{2}$$

where:

$$e_{ij,k}^{(l)} = \text{LeakyReLU}\left(\vec{a}_k^T\left[W_k^{(l)}h_i \parallel W_k^{(l)}h_j\right]\right) \tag{3}$$

Here:

- $\alpha_{ij,k}^{(l)}$ is the normalized attention score between nodes i and j for the k-th head.
- $e_{ij,k}^{(l)}$ is the unnormalized attention score between nodes i and j for the k-th head.
- softmax_i denotes the softmax function applied across all neighbors j of node i.
- $\text{LeakyReLU}(\cdot)$ is the Leaky ReLU activation function applied to the attention score.
- $\vec{a}_k$ is the learnable weight vector for the k-th attention head.
- $\parallel$ denotes the concatenation operation between the transformed feature vectors $W_k^{(l)}h_i$ and $W_k^{(l)}h_j$.

D. Reconstructing the miRNA-Disease Association Network

We constructed a miRNA-disease association network by the learned embedding for the downstream tasks. We first extracted miRNA and disease embeddings to map each entity to its corresponding low-dimensional vector representation. Relationships between miRNAs and diseases were then derived from the dbDEMC database [35], which identified specific pairs based on biological data. For each valid miRNA-disease pair, we created a JSON structure that encapsulated the miRNA and disease properties, including their embeddings, and defined their interaction as a "ASSOCIATED" relationship.

E. Negative Sampling

Negative sampling [38]–[40], provides an efficient method for approximating the partition function of an unnormalized distribution. This approach significantly accelerates the training process of models by selecting a set of negative examples that do not follow the target distribution, making it computationally more tractable [41]. In PERGAT, negative samples, which represent non-existent or unknown associations between miRNAs and diseases, are generated and splited into training, validation, and test sets, mirroring the distribution of positive samples. The model learns to distinguish between existing and non-existing associations, thereby improving its predictive capability.

F. Training and Evaluation

We utilize a k-fold cross-validation method to achieve robust evaluation. For each fold, distinct DGL graphs [42] are constructed for the training, validation, and test phases. These graphs are built using the positive and negative samples. Importantly, the edges corresponding to the test and validation sets are removed from the training graph to ensure that the model does not have access to these edges during training. This removal process guarantees that the model's predictions on the test set are based solely on the information learned from

the training data. The model's performance is evaluated on the test set using a range of metrics, including AUC, F1-score, precision, recall, focal loss, accuracy, and mAP (Table II). The test set contains edges that were not seen during training. A Multi-Layer Perceptron (MLP) is used to predict the existence of edges based on the node embeddings. This decouples the embedding generation process from the prediction process, allowing for more flexibility in how predictions are made. Focal Loss is used as the loss function, which is designed to handle class imbalance by focusing more on difficult-to-classify examples [43]. This is particularly useful in link prediction tasks, where the majority of edges are negative examples.

IV. Results

A. Data selection and processing

We validated PERGAT's predicted miRNA-disease associations using experimental data from the dbDEMC database, which is a specialized resource focusing on the differential expression of microRNAs (miRNAs) in human cancers. The latest version includes 3,268 differentially expressed miRNAs across 40 cancer types, with 2,584 of these specifically related to humans, encompassing 46,388 interactions. The dbDEMC 3.0 includes p-value of miRNA returned by the enrichment analysis of the differentially expressed miRNA (DEM) targets on GO terms and KEGG pathways [36], which is available for download at https://www.biosino.org/dbDEMC.

To further confirm our results, miRNA-disease association records were also retrieved from the most recent release of HMDD (v4.0) [44] and miR2Disease [45].

B. Learning Similarities Between miRNA Entities

The similarity between two feature vectors reflects their inherent similarity within the original network. As a result, the vector similarities in PERGAT models correspond to the relationships between entities in the miRNA-disease network. For example, Figure 3 shows the similarity matrices, which depict the embedding similarities between miRNA entities and disease entities, respectively, with scores generated for the most frequent entities found in the dbDEMC database. We can see that the similarity of the miRNA pair hsa-miR-200b and hsa-miR-374a has the highest score, 0.9997, indicating that they have almost identical learned embeddings. These findings can be supported by the fact that the two miRNAs share common attributes in terms of cancer type and their targets. Positive values indicate upregulation, while negative values indicate downregulation of the miRNA in the disease condition compared to the control, as shown in Table I.

C. Prediction of miRNA-Disease Association

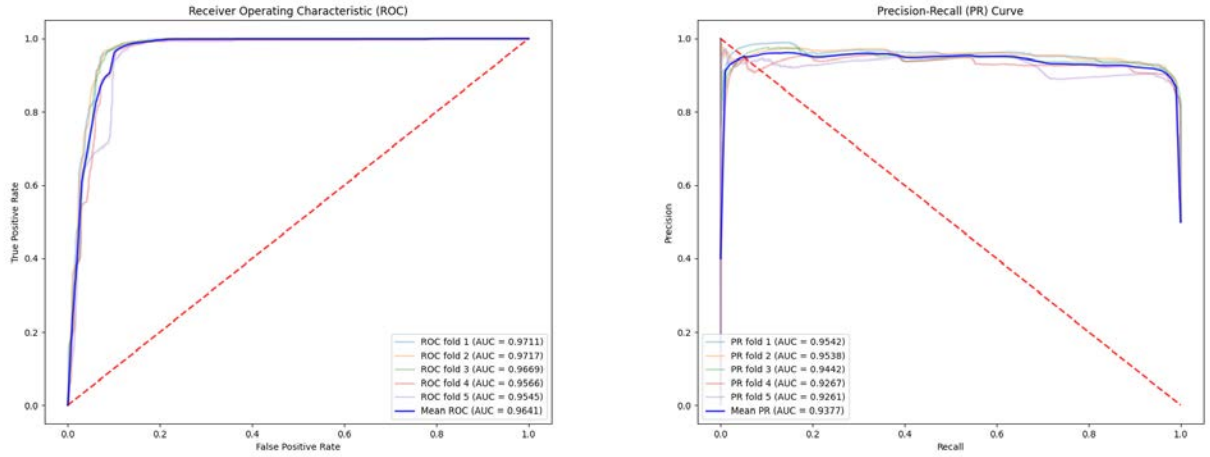
We assessed PERGAT's performance by calculating the area under the receiver operating characteristic curve (AUC). To achieve robust evaluation for PERGAT, we utilize a k-fold cross-validation method (Figure 2). The average metrics for AUC, F1-score, accuracy, precision, recall, and mAP are 96.41%, 91.21%, 90.45%, 84.50%, 95.10%, and 92.30%, respectively, as shown in Table II.

D. Optimizing the Dimensionality of Node Representation

The node representation implies the complex information in latent feature space, and its dimensionality affects the predictive performance of the model. A low number of dimensions may lead to the loss of information, while a high number of dimensions will lead to the introduction of noise and time consuming for calculation. Thus, discovery of the optimized dimension of the node representation is attempted based on the 5-CV through changing dimension in the range of $\{16, 32, 64, 128, 256, 512\}$. The experiment is repeated 10 times for each dimension. Here, Acc, Roc and Aupr are utilized to evaluate the effect of dimension on model performance and statistical average results are shown in Table 4. We can conclude that higher dimensionality tends to be better performance. However, a high feature vector can lead to a huge computational burden and long model training time. Therefore, the optimal feature dimension for node representation is set to 256.

E. Model Stability

Dropout improves the model's generalization and reduces overfitting by randomly deactivating certain neurons in the graph neural network during training. To assess the stability of PERGAT, we used the same setup described in [46]–[48]. The dropout probability p was varied from 0.1 to 0.9 in increments of 0.1, and the resulting AUC values were examined to evaluate the model's stability. As shown in Fig. 6,



(a) (b)

FIG. 2 – Our approach for predicting miRNA-disease associations using 5-fold cross-validation. (a) The ROC curves. (b) The PR curves. As a result, PERGAT achieved a AUC of 0.9641, demonstrating the reliable predictive ability of PERGAT.

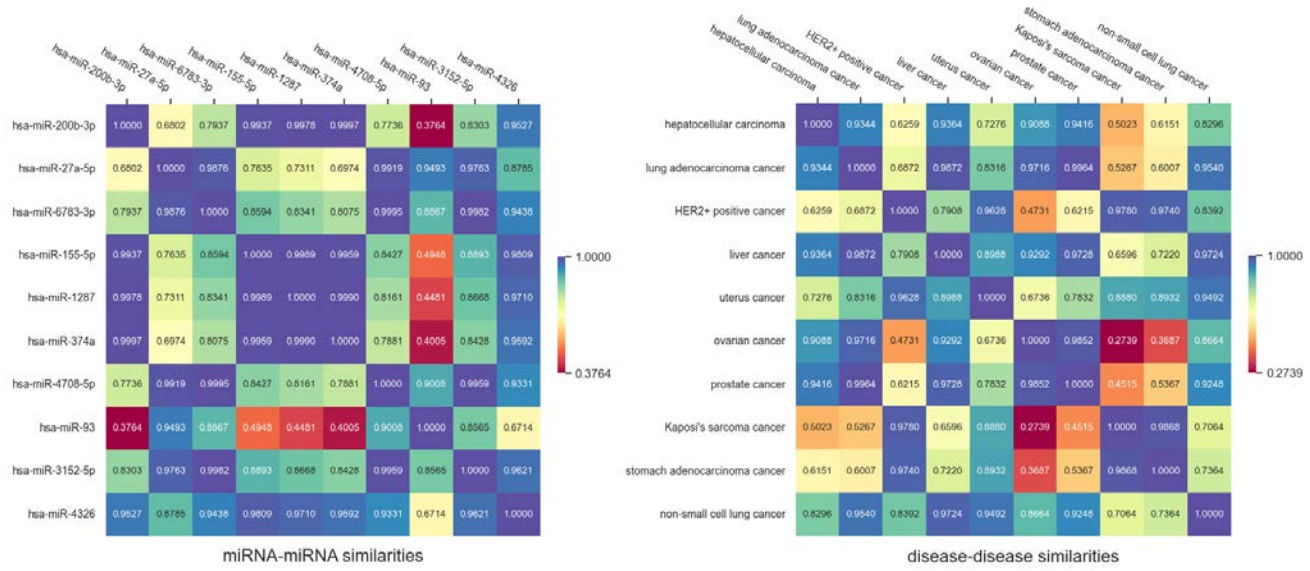


FIG. 3 – Similarity matrices represent miRNA-miRNA similarities and disease-disease similarities. These similarity values are obtained by calculating the cosine similarity between the embeddings of the respective pairs of entities being compared.

miRNA ID	Source ID	Cancer Type	Design	logFC	Expression Status	Experiment ID
hsa-miR-200b	GSE47841	ovarian cancer	cancer vs normal	4.55	UP	EXP00259
hsa-miR-200b	E_MTAB_408	prostate cancer	cancer vs normal	1.50	UP	EXP00638
hsa-miR-200b	GSE40525	breast cancer	cancer vs normal	3.01	UP	EXP00192
hsa-miR-200b	GSE33743	gastric cancer	cancer vs normal	0.92	UP	EXP00175
hsa-miR-200b-3p	TCGA_READ	colorectal cancer	cancer vs normal	2.23	UP	EXP00391
hsa-miR-374a	GSE47841	ovarian cancer	cancer vs normal	-0.28	DOWN	EXP00259
hsa-miR-374a	E_MTAB_408	prostate cancer	cancer vs normal	1.17	UP	EXP00638
hsa-miR-374a	GSE40525	breast cancer	cancer vs normal	1.97	UP	EXP00192
hsa-miR-374a*	GSE33743	gastric cancer	cancer vs normal	0.56	UP	EXP00175
hsa-miR-374a-3p	TCGA_READ	colorectal cancer	cancer vs normal	6.62	UP	EXP00391

TABLE I – Differentially Expressed miRNAs (DEM) for hsa-miR-200b and hsa-miR-374a in dbDEMC database.

Fold	AUC (%)	F1_Val. (%)	Acc. (%)	Prec. (%)	Recall (%)	mAP (%)
1	97.11±0.39	91.47±0.50	90.88±0.59	85.89±0.78	93.83±0.42	93.85±0.86
2	97.17±0.44	90.76±0.49	89.82±0.59	83.15±0.83	95.89±0.10	92.72±1.06
3	96.69±0.45	91.71±0.50	90.39±0.57	85.20±0.79	95.29±0.24	93.34±0.73
4	95.66±0.50	91.68±0.48	90.02±0.58	85.40±0.82	94.96±0.28	91.96±1.02
5	95.45±0.43	90.45±0.47	89.49±0.62	82.88±0.76	95.53±0.19	93.11±0.95
Mean	**96.41±0.44**	**91.21±0.49**	**90.45±0.59**	**84.50±0.80**	**95.10±0.25**	**92.30±0.92**

TABLE II – The effectiveness of our approach assessed using 5-fold cross-validation. The best score in each column is shown in bold.

Models	AUC (%)	F1_Val.(%)	Acc. (%)	Prec. (%)	Recall(%)	mAP (%)
GIN	93.80±0.32	89.58±0.26	89.34±0.26	82.89±0.78	87.70±0.24	90.36±0.59
GAT	94.04±0.26	90.72±0.25	90.31±0.26	84.03±0.37	87.93±0.20	91.72±0.06
GCN	92.50±0.27	89.03±0.26	84.17±0.24	84.35±0.27	91.83±0.40	88.63±0.16
GraphSage	94.02±0.07	90.69±0.36	89.67±0.29	**85.98±0.22**	81.02±0.56	88.92±0.10
PERGAT	**96.41±0.44**	**91.21±0.49**	**90.45±0.59**	84.50±0.80	**95.10±0.25**	**92.30±0.92**

TABLE III – The comparison of PERGAT with state-of-the-art models across 5-fold cross-validation on the dbDEMC dataset. The best score in each column is shown in bold.

the mean AUC across different datasets fluctuated slightly within a range of 0.05, suggesting that varying dropout rates have minimal effect on PERGAT's performance, confirming the model's robustness and stability.

F. Comparison with the Baseline Models

To illustrate the effectiveness of our method, we compare PERGAT model with several baseline models. Both the comparison models and PERGAT are evaluated on the same dataset in this study. We employ 10 repetitions of 5-fold cross-validation to assess the performance of the PERGAT model in comparison with other baseline methods on the dbDEMC dataset. As depicted in Figures 4 and Table III and detailed in Tables 1 and 2, our proposed PERGAT method demonstrates competitive performance.

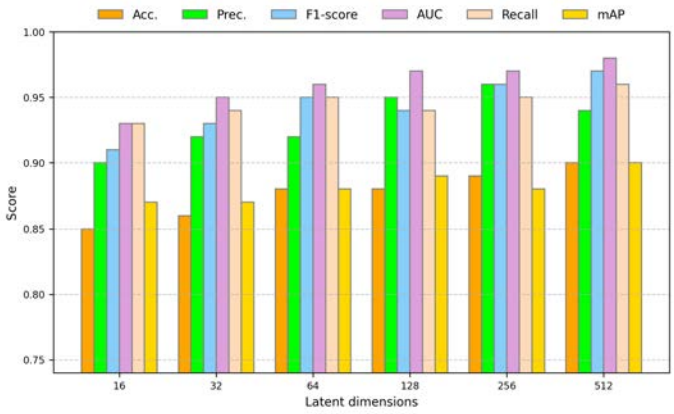


FIG. 4 – The average accuracy, precision, recall, F1-score and AUC values of PERGAT under different latent dimensions upon 5-fold cross-validation.

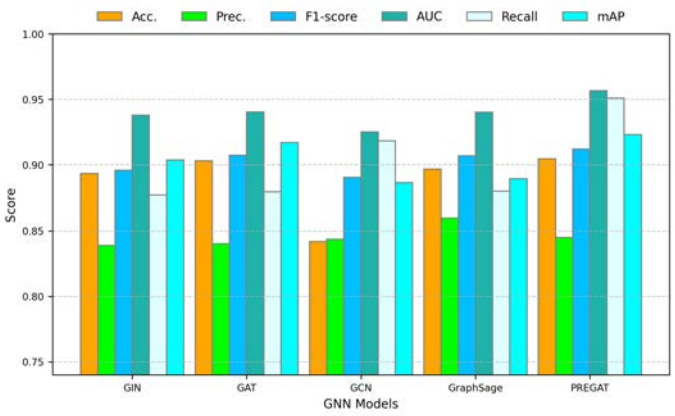
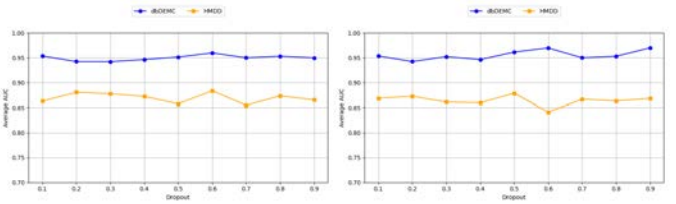


FIG. 5 – The average values for accuracy, precision, recall, F1-score, and AUC of PERGAT across different GNN models using 5-fold cross-validation.



(a) AUC for validation (b) AUC for test

FIG. 6 – The mean AUC of R-MHAGNN evaluated across datasets dbDEMC and HMDD v.3 with varying dropout probabilities.

G. Ablation Experiments

The PERGAT model integrates miRNA features learned from the miRNA-disease network. To assess the contribution of different components to predicting miRNA-cancer associations, we conducted several ablation experiments with various model variants:

1) RGAT: This variant uses the original features of both cancer and miRNA nodes without pre-training the miRNA-disease network, while leaving the rest of the model unchanged.

2) PRGAT: Residual connections in the Graph Attention Network (GAT) layer are disabled to investigate whether removing these connections enhances the embedding representation learning of PERGAT and improves the model's prediction performance.

3) PERGAT-atte: Setting the attention dropout rate to zero aims to explore if this compels the model to focus on a broader set of neighbors, thus reducing over-reliance on a few connections and potentially improving prediction performance.

Rank	Disease	miRNA	Score	Evidence
1	lung adenocarcinoma	hsa-miR-221	0.8732	miR2Disease
2	lung adenocarcinoma	hsa-miR-153	0.8729	miR2Disease
3	lung adenocarcinoma	hsa-miR-182	0.8716	miR2Disease
4	lung adenocarcinoma	hsa-miR-424	0.8715	miR2Disease
5	lung adenocarcinoma	hsa-miR-184	0.8708	miR2Disease
6	lung adenocarcinoma	hsa-miR-129	0.8708	miR2Disease
7	lung adenocarcinoma	hsa-miR-522-5p	0.8636	HMDD
8	lung adenocarcinoma	hsa-miR-208	0.8247	miR2Disease
9	lung adenocarcinoma	hsa-miR-33	0.8244	miR2Disease
10	lung adenocarcinoma	hsa-miR-190	0.8242	miR2Disease
11	breast cancer	hsa-let-7c	0.8881	HMDD
12	breast cancer	hsa-miR-99a	0.8879	miR2Disease
13	breast cancer	hsa-miR-522-5p	0.8875	HMDD
14	breast cancer	hsa-miR-424	0.8867	miR2Disease
15	breast cancer	hsa-miR-7515	0.8852	**Unconfirmed**
16	breast cancer	hsa-miR-199a-3p	0.8783	HMDD
17	breast cancer	hsa-miR-21	0.8761	miR2Disease
18	breast cancer	hsa-miR-125b	0.8710	miR2Disease
19	breast cancer	hsa-miR-142-3p	0.8582	miR2Disease
20	breast cancer	hsa-miR-30b	0.8243	miR2Disease
21	pancreatic cancer	hsa-miR-99a	0.8720	miR2Disease
22	pancreatic cancer	hsa-miR-182	0.8704	miR2Disease
23	pancreatic cancer	hsa-miR-9*	0.8700	miR2Disease
24	pancreatic cancer	hsa-miR-21	0.8691	miR2Disease
25	pancreatic cancer	hsa-miR-29b	0.8687	miR2Disease
26	pancreatic cancer	hsa-miR-522-5p	0.8683	**Unconfirmed**
27	pancreatic cancer	hsa-miR-30b	0.8675	miR2Disease
28	pancreatic cancer	hsa-miR-3180@	0.8660	**Unconfirmed**
29	pancreatic cancer	hsa-miR-188	0.8654	miR2Disease
30	pancreatic cancer	hsa-let-7c	0.8614	HMDD

TABLE IV – The prediction outcomes for the top 10 potential miRNAs associated with lung adenocarcinoma, breast cancer, and pancreatic cancer, respectively, as identified by PERGAT based on known associations in HMDD v4 and miR2Disease, are as follows: 10 out of 10 for lung adenocarcinoma, 9 out of 10 for breast cancer, and 8 out of 10 for pancreatic cancer were confirmed.

4) PERGAT-feat: By setting the feature dropout rate to zero, we examine whether the absence of feature dropout affects the model's generalization ability, potentially making it more robust to missing information.

We perform 5-fold cross-validation on the deDEMC dataset to evaluate and compare the performance of these models. As illustrated in Fig. 7, the results indicate that all variant models show lower performance compared to the original model.

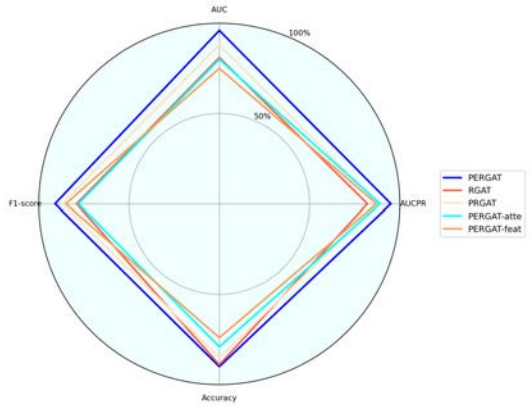


FIG. 7 – Comparative analysis of various models through ablative experiments.

H. Case studies

To validate the performance of PERGAT in predicting miRNA-disease associations, we conducted case studies on three major tumor types: lung adenocarcinoma, breast cancer, and pancreatic cancer, using the dbDEMC 3.0 dataset. We constructed training samples by excluding associations with the specific disease under study and including negative miRNA-disease associations alongside experimentally verified positive ones. The specific disease associations were then used to create testing samples. We trained the PERGAT model on these training samples and used it to predict associations between miRNAs and the specific disease. We ranked the predictions and selected the top scores as potential candidates. Subsequently, we verified the top 10 predictions by cross-referencing with the latest data from HMDD v4.0 and miR2Disease for supporting evidence. 10 out of 10 for lung adenocarcinoma, 9 out of 10 for breast cancer, and 8 out of 10 for pancreatic cancer were confirmed as shown in Table IV.

V. CONCLUSION

We present a novel method for predicting miRNA-cancer associations using Pretrained Embedding based on Graph Neural Networks (PERGAT). Our approach leverages the structural information in miRNA-disease networks and demonstrates superior performance compared to existing methods. Our proposed method effectively captures the topological structure

of miRNA-cancer networks and leverages it for accurate association prediction. The ability of GNNs to learn from graph-structured data makes them well-suited for this task. This work highlights the potential of GNNs in biomedical research and opens up new avenues for studying miRNA-related diseases.

However, the model has some limitations, such as the lack of interpretability in the learned embeddings and attention mechanisms, as complex models like GATs may obscure the biological processes behind predictions. In the future, we will integrate multi-omics data such as gene expression, protein interactions, and epigenetic changes, to improve the model's ability to capture deeper biological insights and enhance prediction accuracy across more diseases.

ACKNOWLEDGEMENTS

We sincerely thank the anonymous reviewers for their valuable suggestions and comments, which greatly improved this paper. Their feedback guided us in refining our approach and enhancing the clarity of the research. This work is supported by the National Science Foundation under Grant No. 2245805.

REFERENCES

[1] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. "Pan-cancer analysis of whole genomes." *Nature*, vol. 578, pp. 82–93, 2020. https://doi.org/10.1038/s41586-020-1969-6.

[2] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215-233, Jan. 2009. doi: 10.1016/j.cell.2009.01.002.

[3] C. Liu, C. Yu, G. Song, X. Fan, S. Peng, S. Zhang, X. Zhou, C. Zhang, X. Geng, T. Wang, W. Cheng, and W. Zhu, "Comprehensive analysis of miRNA-mRNA regulatory pairs associated with colorectal cancer and the role in tumor immunity," *BMC Genomics*, vol. 24, no. 1, p. 724, Nov. 2023. doi: 10.1186/s12864-023-09635-4.

[4] G. A. Calin and C. M. Croce, "MicroRNA signatures in human cancers," *Nat Rev Cancer*, vol. 6, no. 11, pp. 857-866, Nov. 2006. doi: 10.1038/nrc1997.

[5] Z. Li, X. Huang, Y. Shi, X. Zou, Z. Li, and Z. Dai, "Identification of MiRNA-Disease Associations Based on Information of Multi-Module and Meta-Path," *Molecules*, vol. 27, no. 14, p. 4443, Jul. 2022. doi: 10.3390/molecules27144443.

[6] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for Pre-training Graph Neural Networks," *arXiv preprint*, 2020. Available: https://arxiv.org/abs/1905.12265.

[7] C. Tang, H. Zhou, X. Zheng, Y. Zhang, and X. Sha, "Dual Laplacian regularized matrix completion for microRNA-disease associations prediction," *RNA Biol*, vol. 16, no. 5, pp. 601-611, May 2019. doi: 10.1080/15476286.2019.1570811.

[8] H. Feng, D. Jin, J. Li, Y. Li, Q. Zou, and T. Liu, "Matrix reconstruction with reliable neighbors for predicting potential MiRNA-disease associations," *Brief Bioinform*, vol. 24, no. 1, p. bbac571, Jan. 2023. doi: 10.1093/bib/bbac571.

[9] Q. Dai, Y. Chu, Z. Li, Y. Zhao, X. Mao, Y. Wang, Y. Xiong, and D. Q. Wei, "MDA-CF: Predicting MiRNA-Disease associations based on a cascade forest model by fusing multi-source information," *Comput Biol Med*, vol. 136, p. 104706, Sep. 2021. doi: 10.1016/j.compbiomed.2021.104706.

[10] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61-80, Jan. 2009. doi: 10.1109/TNN.2008.2005605.

[11] M. Zhang and Y. Chen, "Link Prediction Based on Graph Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 5165-5175.

[12] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How Powerful are Graph Neural Networks?," *ICLR*, 2019.

[13] Li, Qimai, Zhichao Han, and Xiao-Ming Wu. "Deeper insights into graph convolutional networks for semi-supervised learning." Proceedings of the AAAI Conference on Artificial Intelligence 32, no. 1 (2018).

[14] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2017).

[15] Hamilton, William L., Rex Ying, and Jure Leskovec. "Inductive representation learning on large graphs." Advances in neural information processing systems 30 (2017).

[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018. Available: https://openreview.net/forum?id=rJXMpikCZ. [Accepted as poster].

[17] T. Ma and J. Wang, "GraphPath: a graph attention model for molecular stratification with interpretability based on the pathway-pathway interaction network," *Bioinformatics*, vol. 40, no. 4, pp. btae165, Mar. 2024.

[18] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: online learning of social representations," *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2014.

[19] Q. Liao, Y. Ye, Z. Li, H. Chen, and L. Zhuo, "Prediction of miRNA-disease associations in microbes based on graph convolutional networks and autoencoders," *Front Microbiol*, vol. 14, p. 1170559, Apr. 2023. doi: 10.3389/fmicb.2023.1170559.

[20] Z. Zhang, X. Wu, G. Zhu, W. Qin, and N. Liang, "A Graph Attention Network-Based Link Prediction Method Using Link Value Estimation," *IEEE Access*, vol. 12, pp. 34-45, 2024. doi: 10.1109/ACCESS.2023.3346688.

[21] D. Ouyang, Y. Liang, J. Wang, L. Li, N. Ai, J. Feng, S. Lu, S. Liao, X. Liu, and S. Xie, "HGCLAMIR: Hypergraph contrastive learning with attention mechanism and integrated multi-view representation for predicting miRNA-disease associations," *PLoS Comput Biol*, vol. 20, no. 4, p. e1011927, Apr. 2024. doi: 10.1371/journal.pcbi.1011927.

[22] W. Peng, Q. Tang, W. Dai, and T. Chen, "Improving cancer driver gene identification using multi-task learning on graph convolutional network," *Brief Bioinform*, vol. 23, no. 1, p. bbab432, Jan. 2022. doi: 10.1093/bib/bbab432.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *arXiv preprint arXiv:1206.5538*, 2014. Available: https://arxiv.org/abs/1206.5538.

[24] X. Tang, J. Luo, C. Shen, and Z. Lai, "Multi-view Multichannel Attention Graph Convolutional Network for miRNA-disease association prediction," *Brief Bioinform*, vol. 22, no. 6, p. bbab174, Nov. 2021. doi: 10.1093/bib/bbab174.

[25] T. Li, Z. Zhou, S. Li, C. Sun, R. Yan, and X. Chen, "The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study," *Mech. Syst. Signal Process.*, vol. 168, p. 108653, 2022. doi: 10.1016/j.ymssp.2021.108653.

[26] S. Pang, Y. Zhuang, X. Wang, et al., "EOESGC: predicting miRNA-disease associations based on embedding of embedding and simplified graph convolutional network," *BMC Med. Inform. Decis. Mak.*, vol. 21, p. 319, 2021. doi: 10.1186/s12911-021-01671-y.

[27] Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, "PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks," *bioRxiv*, p. 532226, 2019. Cold Spring Harbor Laboratory. 10.1016/j.ymssp.2021.108653.

[28] W. Peng, R. Wu, W. Dai, Y. Ning, X. Fu, L. Liu, and L. Liu, "MiRNA-gene network embedding for predicting cancer driver genes," *Brief Funct Genomics*, vol. 22, no. 4, pp. 341–350, Jul. 2023. doi: 10.1093/bfgp/elac059.

[29] R. Schulte-Sasse, S. Budach, D. Hnisz, and others, "Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms," *Nat Mach Intell*, vol. 3, pp. 513–526, 2021. doi: 10.1038/s42256-021-00325-y.

[30] X. Pan and H. B. Shen, "Inferring Disease-Associated MicroRNAs Using Semi-supervised Multi-Label Graph Convolutional Networks," *iScience*, vol. 20, pp. 265-277, Oct. 2019. doi: 10.1016/j.isci.2019.09.013.

[31] Z. Jin, M. Wang, C. Tang, X. Zheng, W. Zhang, X. Sha, and S. An, "Predicting miRNA-disease association via graph attention learning and multiplex adaptive modality fusion," *Comput. Biol. Med.*, vol. 169, p. 107904, Feb. 2024. doi: 10.1016/j.compbiomed.2023.107904.

[32] J. Wang, J. Chen, and S. Sen, "MicroRNA as Biomarkers and Diagnostics," *J Cell Physiol*, vol. 231, no. 1, pp. 25-30, Jan. 2016. doi: 10.1002/jcp.25056.

[33] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in Statistics*, S. Kotz and N. L. Johnson, Eds., Springer Series in Statistics. New York, NY: Springer, 1992, pp. 66-70. doi: 10.1007/978-1-4612-4380-9.

[34] C. Backes, Q. T. Khaleeq, E. Meese, and A. Keller, "miEAA: microRNA enrichment analysis and annotation," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W110-6, Jul. 2016. doi: 10.1093/nar/gkw345.

[35] F. Xu, Y. Wang, Y. Ling, C. Zhou, H. Wang, A. E. Teschendorff, Y. Zhao, H. Zhao, Y. He, G. Zhang, and Z. Yang, "dbDEMC 3.0: Functional Exploration of Differentially Expressed miRNAs in Cancers of Human and Model Organisms," *Genomics Proteomics Bioinformatics*, vol. 20, no. 3, pp. 446-454, Jun. 2022.

[36] Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., Kanehisa M. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999. https://doi.org/10.1093/nar/27.1.29.

[37] M. Zhang and Y. Chen, "Link Prediction Based on Graph Neural Networks," *arXiv preprint*, 2018. Available: https://arxiv.org/abs/1802.09691.

[38] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *arXiv preprint*, 2013. Available: https://arxiv.org/abs/1310.4546.

[39] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *Journal of Machine Learning Research*, vol. 13, pp. 307-361, Feb. 2012.

[40] A. Mnih and K. Kavukcuoglu, "Learning word embeddings efficiently with noise-contrastive estimation," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 2265-2273, 2013.

[41] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 3111-3119, 2013.

[42] M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li, and Z. Zhang, "Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks," *arXiv preprint*, 2020. Available: https://arxiv.org/abs/1909.01315.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *arXiv preprint*, 2018. Available: arXiv:1708.02002.

[44] C. Cui, B. Zhong, R. Fan, and Q. Cui, "HMDD v4.0: a database for experimentally supported human microRNA-disease associations," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1327–D1332, Jan. 2024. doi: 10.1093/nar/gkad717.

[45] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. suppl_1, pp. D98-D104, Oct. 2008.

[46] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. "Dropout: A simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[47] Zheng Zhang, Lei Cui, Jianping Wu. "Exploring an edge convolution and normalization based approach for link prediction in complex networks." *Journal of Network and Computer Applications*, vol. 189, 103113, 2021. https://doi.org/10.1016/j.jnca.2021.103113.

[48] Zhang Z., Xu H., Zhu G. "Incorporating high-frequency information into edge convolution for link prediction in complex networks." *Scientific Reports*, vol. 14, no. 1, p. 5437, 2024. https://doi.org/10.1038/s41598-024-56144-9.