Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PMTIDMA: plant miRNA target interaction prediction framework based on denoising autoencoder and multi-scale spectral graph convolutional attention

Jiangqing Wang
South-Central Minzu University
Xin Zhou
South-Central Minzu University
Jin Xie
jinxie@scuec.edu.cn

South-Central Minzu University
Jun Tie
South-Central Minzu University
Kexin Hu
South-Central Minzu University
Yuwei Wan
South-Central Minzu University
Junrui Liu
South-Central Minzu University
Wenzhen Zhang
South-Central Minzu University

Research Article

Keywords: Heterogeneous graph, DAE, multi-scale spectral graph convolution, scale-level attention fusion, JK module

Posted Date: April 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9127892/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PMTIDMA: plant miRNA target interaction prediction framework based on denoising autoencoder and multi-scale spectral graph convolutional attention

Jiangqing Wang, Xin Zhou, Jin Xie, Jun Tie, Kexin Hu, Yuwei Wan, Junrui Liu, Wenzhen Zhang

School of Computer, South-Central Minzu University, Wuhan, 430000, Hubei, China.
Hubei Provincial Engineering Research Center of Agricultural Blockchain and Intelligent Management, South-Central Minzu University, Wuhan, 430000, Hubei, China.
Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprise, South-Central Minzu University, Wuhan, 430000, Hubei, China.

*Corresponding author(s). E-mail(s): jinxie@scuec.edu.cn;
Contributing authors: wjqing2000@mail.scuec.edu.cn; 2947613887@qq.com; tiejun@mail.scuec.edu.cn; 1548403343@qq.com; 2067246847@qq.com; 1062282469@qq.com; 2504030929@qq.com;

Abstract

Accurate prediction of plant miRNA–target interactions(MTIs) is fundamental to understanding gene regulation in plant growth, development, and stress responses. Experimental validation of MTIs is accurate but costly, time-consuming, and limited in scalability, especially in plant species. In contrast, computation-based methods can efficiently mine potential regulatory relationships, but existing models still exhibit deficiencies in feature representation capability, structural information modeling, and performance under limited-sample conditions. This paper proposes PMTIDMA, a novel plant miRNA–target gene prediction framework based on graph neural networks. It first uses a Denoising Autoencoder(DAE) to reduce dimensionality of miRNA and mRNA sequence features, enhancing robustness and reducing noise. Then, heterogeneous graphs

are built from miRNA–miRNA and mRNA–mRNA sequence similarities and known miRNA–mRNA interactions. On this basis, the sequence features and heterogeneous graphs are jointly input into the multi-scale spectral graph convolutional network for joint training. This network performs weighted fusion on the convolution results of spectral graphs of different scales through the attention mechanism, and further combines the standard Graph convolutional network layer and the Jumping Knowledge module to fully capture the local and global structural information. Ultimately, the multi-layer perceptron was used to predict the interaction probability between miRNA and target genes. Experimental results on multiple plant datasets show that PMTIDMA consistently outperforms existing mainstream methods. Specifically, PMTIDMA achieves an AUPRC of 0.9302, an AUROC of 0.9369, and an F1-score of 0.9351. The ablation experiment further verified the effectiveness of the denoising autoencoder, multi-scale spectral graph convolution and attention fusion mechanism in the model. In conclusion, PMTIDMA provides an efficient and robust computational framework for predicting the interaction between plant mirnas and target genes, which is conducive to the systematic analysis of the plant miRNA regulatory network.

Keywords: Heterogeneous graph, DAE, multi-scale spectral graph convolution, scale-level attention fusion, JK module

1 Introduction

There are various non-coding RNAs, which play important roles in biological evolution and the regulation of life activities. Among them, microRNAs (miRNAs), have gradually begun to attract high attention due to their extensive and delicate regulatory roles[1, 2] . miRNA is a type of endogenous non-coding small RNA with a length of approximately 20 to 25 nucleotides. By complementing and binding to target mRNAs, it negatively regulates their transcription or translation processes, thereby influencing various biological processes in animals and plants, including growth and development, metabolic regulation, and stress response[3, 4]. Although miRNA does not encode proteins, it occupies a core position in the gene regulatory network and is one of the most important components among non-coding RNAs[5]. In plants, miRNA is widely involved in key life processes such as organ morphogenesis, growth and development, and seed formation[6]. At the same time, it plays an important regulatory role in responding to biological and abiotic stresses such as pests and diseases, drought, high temperature, salinity and alkalinity, and nutrient deficiency[7]. Therefore, in-depth research on the regulatory relationship between plant mirnas and their target genes holds significant biological and agricultural application value for enhancing the yield, quality, and environmental adaptability of crops and energy plants.

However, conventional experimental methods such as RT-qPCR and Northern blotting can validate only limited miRNA–target interactions (MTIs), making them labor-intensive and costly[8, 9]. With advances in high-throughput sequencing, computational approaches have become valuable complements to experimental

validation[10, 11]. Deep learning, in particular, has enabled large-scale MTI prediction with improved efficiency and reduced cost[12, 13]. Early studies relied on traditional machine learning techniques (e.g., SVM, RF, k-NN) using manually engineered features such as nucleotide composition, minimum free energy, or structural information. Tools like miRanda and TargetScan exemplify such feature-based prediction methods[14]. Yet biological data commonly exhibit the "few samples, many features" challenge and contain non-linear hierarchical relationships[15, 16]. Traditional feature-dependent methods often struggle with overfitting and noise, especially under high dimensionality and limited samples[17, 18]. In contrast, deep learning can automatically learn abstract hierarchical features, enabling better modeling of complex, non-linear interactions and improving predictive accuracy[19, 20]. Representative works include DeepMirTar, a stacked denoising autoencoder model achieving 93.48% accuracy for site-level MTI prediction[21], and a hybrid CNN–RNN framework by Tongjun et al. (2020), achieving 95.45% accuracy on datasets like DeepMirTar and miRAW[22]. Despite such progress, most studies focus on human or animal data, whereas plant MTI research remains limited. Plant datasets are often smaller, and plant miRNAs typically exhibit near-perfect complementarity with their targets, unlike animals where seed-region pairing dominates[23, 24]. These differences mean that animal-based algorithms cannot be directly applied to plants without tailored adaptation[25]. Zhang et al. proposed kmerPMTF, which constructs k-mer-based similarity networks and applies spectral graph convolution to predict plant miRNA–target associations. Although it demonstrates certain effectiveness, kmerPMTF primarily employs a relatively simple model architecture, which restricts its ability to capture complex feature interactions and fully exploit structural information in miRNA–target interaction networks.

To address these limitations, we propose PMTIDMA, a plant MTI prediction framework that integrates a denoising autoencoder (DAE) with a scale-aware multi-scale spectral graph convolutional network. PMTIDMA first constructs a unified miRNA–target (M–T) heterogeneous graph by integrating miRNA similarity, mRNA similarity, and known miRNA–target interactions. Sequence-derived k-mer features are concatenated with the heterogeneous graph structure matrix and encoded through a DAE to obtain compact and noise-robust latent embeddings, effectively alleviating sparsity while preserving intrinsic attribute–structure correlations.The graph learning module employs multi-branch Chebyshev spectral convolutions with different polynomial orders to capture structural dependencies across multiple topological ranges. A scale-level attention mechanism is introduced to adaptively fuse multi-scale representations instead of simple concatenation, enabling the model to emphasize the most informative structural scale. Subsequently, a GCN layer and a Jumping Knowledge (JK) mechanism further integrate hierarchical features, enhancing node representation expressiveness and improving predictive performance for plant miRNA–target interaction identification.

The main contributions of this study are summarized as follows:

1. We developed a high-performance plant miRNA–target interaction (MTI) prediction framework PMTIDMA based on RNA sequences. The framework is

characterized by high robustness and strong generalization ability, and it can accurately predict MTIs even under limited-sample conditions.

2. This framework combines denoising autoencoders and attention-weighted multi-scale spectral graph convolution for miRNA-target gene interaction prediction. By reducing noise interference and comprehensively capturing local and global structural information, it significantly improves the accuracy and generalization ability of the model.
3. The framework achieves consistently strong performance across diverse plant species while maintaining a lightweight and computationally efficient architecture, making it particularly well suited for plant miRNA–target interactions (MTIs) prediction in species with limited available data.

2 Materials and methods

2.1 Datasets

In this study, we selected three representative plant species as experimental datasets: Arabidopsis thaliana, Solanum lycopersicum, and Prunus persica. Arabidopsis thaliana is the most widely used model organism in plant genetics and molecular biology, serving as a key reference system for functional gene studies in higher plants [26, 27]. Solanum lycopersicum (tomato) and Prunus persica (peach) represent vegetable and fruit crops, respectively, both of which are of significant importance for research on crop improvement and trait development [28, 29]. All miRNA sequences and their experimentally validated target pairs were obtained from the Plant miRNA Encyclopedia (PmiREN) [30]. This comprehensive database integrates multiple authoritative data sources and experimental validation results,providing high-confidence miRNA sequences, target relationships, and conservation annotations across a wide range of plant species. Compared with other public resources, PmiREN offers broader coverage and higher annotation accuracy, thereby ensuring reliable data support for model training and evaluation.The genomic datasets used in this study contain 26,873–34,658 genes, 221–324 miRNAs, and 2,181–6,251 miRNA–target pairs across the three plant species. As shown in Table 1:

Table 1: Description of the dataset

Species	Gene number	MIR number	Pairs
Arabidopsis thaliana	27,416	221	3,646
Solanum lycopersicum	34,658	324	2,181
Prunus persica	26,873	293	6,251

2.2 Heterogeneous Graph Construction

In the plant MTIs prediction task, the complex regulatory relationships between miRNAs and their target genes naturally form a heterogeneous information network, where nodes represent miRNAs and target genes, and edges consist of three types: miRNA–miRNA similarity edges, gene–gene similarity edges, and experimentally validated miRNA–target interaction edges. We constructed a heterogeneous graph according to Figure 1. To fully capture the multi-level association information between miRNAs and target genes, sequence information is first converted into a computable representation. Specifically, for each miRNA and gene sequence, we extract its k-mer frequency matrix [31]. Given a sequence $S = (s_1, s_2, \ldots, s_L)$ of length L, the sequence can be divided into a set of consecutive k-length nucleotide fragments:

$$\mathcal{K}(S) = S_i = s_i s_{i+1} \ldots s_{i+k-1} \mid i = 1, 2, \ldots, L - k + 1 \tag{1}$$

Let c_j denote the number of occurrences of the j-th k-mer in the sequence S. Then, the k-mer vector of the sequence can be represented as:

$$x_s = [c_1, c_2, \ldots, c_{4^k}] \tag{2}$$

This vector records the occurrence frequencies of all possible k-mers, capturing the distribution patterns of nucleotide fragments within the sequence [32]. Based on the k-mer frequency vectors of all nodes, miRNA–miRNA and target–target similarities are computed. Since miRNA–target interactions are predominantly driven by base-pairing features, especially the partial complementarity in the seed region, the Jaccard coefficient is employed to measure the proportion of shared k-mer patterns, providing a direct reflection of the structural similarity between sequences at the short-fragment level. This aligns with the biological mechanism underlying miRNA target recognition.

Furthermore, the Jaccard coefficient demonstrates robustness to noise and missing data, which is particularly advantageous given the short length of miRNA sequences (typically 20–24 nucleotides) compared to the significantly longer gene sequences. This characteristic helps mitigate the "dilution" of similarity within the high-dimensional space of long sequences, thereby preserving the integrity of the similarity measure [33].Therefore, we adopt the Jaccard coefficient as the similarity metric:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

Here, A and B represent the nonzero k-mer sets of two sequences. If the Jaccard similarity J(A,B) exceeds the threshold of 0.5, a similarity edge is established between the corresponding nodes. In this way, the similarity matrices S_{miRNA} and S_{target} are constructed, respectively. Furthermore, based on high-confidence miRNA-target interaction pairs obtained from the PmiREN database, a third type of edge is introduced. Let $A_{miRNA-target}$denote the binary association matrix between miRNAs and their target genes, where the element $a_{ij} = 1$ indicates a regulatory relationship between $miRNA_i$ and $target_i$, and $a_{ij} = 0$ otherwise. Finally, the heterogeneous graph is

represented by the following adjacency matrix:

$$A_H = \begin{bmatrix} S_{\text{miRNA}} & A_{\text{miRNA-target}} \\ A^T_{\text{miRNA-target}} & S_{\text{target}} \end{bmatrix} \in \mathbb{R}^{(N_{\text{miRNA}}+N_{\text{target}})\times(N_{\text{miRNA}}+N_{\text{target}})} \quad (4)$$

In this formulation, N_{miRNA} and N_{target} denote the numbers of miRNA and target nodes,respectively. The matrix A_H encodes both intra-type similarities and cross-type regulatory relationships through its block structure, ensuring that the graph structure remains symmetric and computationally tractable in the spectral domain.

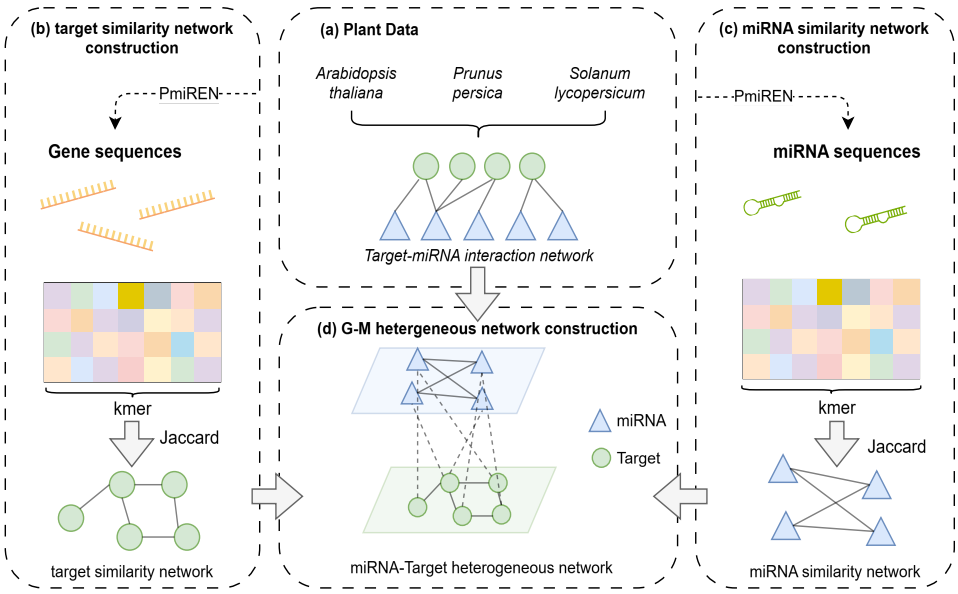


Fig. 1: The construction of the M-T heterogeneous network

2.3 Node Embedding Feature Construction

Previous research has indicated that the recognition region of plant miRNAs generally spans approximately 10–13 nucleotides, a feature critical for determining target specificity and cleavage efficiency within this region [34]. Accordingly, the selection of k-mer fragment lengths is optimized to capture these local pairing patterns while preserving the surrounding upstream and downstream sequence context, thereby ensuring both predictive accuracy and computational efficiency. Fragments that are too short may weaken the discriminative signal of complementarity, whereas overly long fragments may introduce redundant features [35]. Let the graph contain N nodes, with each node represented by a k-mer feature vector of length d The initial feature matrix can then be expressed as:

$$\mathbf{X} = [x_1, x_2, \ldots, x_N]^T \in \mathbb{R}^{N \times d} \tag{5}$$

Given that the plant miRNA–target interaction network constitutes a heterogeneous graph with multiple relation types, relying solely on sequence-derived node features (e.g., k-mer frequency vectors) is insufficient to fully characterize functional associations between nodes. In plant systems, k-mer–based representations are typically high-dimensional and highly sparse, which may limit their discriminative capacity when used independently. Meanwhile, heterogeneous relations—such as miRNA–gene and gene–gene interactions—encode essential topological information reflecting potential regulatory mechanisms and functional dependencies. To enhance node representation, we concatenate the attribute feature matrix X with the heterogeneous adjacency matrix A_H, thereby integrating explicit sequence information with first-order structural connectivity patterns, As shown in panel (a) of Figure 2. This fusion strategy enables each node representation to jointly capture local sequence characteristics and its structural position within the interaction network:

$$\mathbf{Z} = [\mathbf{X}|\mathbf{A}_H] \tag{6}$$

Although concatenation enriches node representations, the resulting feature space remains high-dimensional and potentially noisy, which may introduce redundancy and increase the risk of overfitting. To address this issue, we employ a Denoising Autoencoder (DAE) to learn a compact latent embedding from the augmented feature matrix. By injecting controlled noise into the input and optimizing the reconstruction objective, the DAE is able to filter irrelevant variations while preserving salient attribute–structure correlations[36]. Compared to conventional linear dimensionality reduction techniques, the DAE captures nonlinear dependencies within the fused feature space, making it particularly suitable for sparse and heterogeneous biological data. As a result, the high-dimensional concatenated representation is mapped into a lower-dimensional latent space that is more robust, compact, and information-dense:

$$\mathcal{L}_{rec} = \|Z - Z'\|_2^2 \tag{7}$$

The compressed embedding obtained from the DAE serves as the input to the heterogeneous graph neural network for subsequent message passing. This two-stage design—fusion, compression, and propagation—offers several advantages. First, dimensionality reduction alleviates sparsity and redundancy, improving computational efficiency and reducing overfitting risk during training. Second, by explicitly encoding structural information prior to graph propagation, the model mitigates excessive smoothing effects that may otherwise obscure discriminative sequence patterns. Third, because first-order structural connectivity has already been incorporated into the latent representation, the subsequent heterogeneous message passing process can focus on modeling higher-order interactions and relation-specific dependencies. Collectively, this strategy enhances representation expressiveness and improves the predictive performance and stability of plant miRNA–target interaction modeling.

2.4 Multi-Scale Graph Convolution Network

After obtaining the heterogeneous adjacency matrix A_H and the node embedding matrix Z', both are fed into a designed multi-scale spectral graph convolutional network for association prediction. The proposed architecture consists of three layers and aims to integrate neighborhood information from multiple receptive fields while capturing both local and global structural dependencies. As shown in panel (c) of Figure 2. The first layer is a multi-branch Chebyshev graph convolution, employing three different polynomial orders $K = \{2, 4, 6\}$ to extract node features from neighborhoods of varying depths. Let the normalized Laplacian matrix be defined as:

$$\tilde{\mathbf{L}} = \mathbf{I} - \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2} \tag{8}$$

A = A_H + I denotes the adjacency matrix with self-loops added, and D is thecorresponding degree matrix.

To avoid the high computational cost of spectral decomposition, we adopt the Chebyshev polynomial approximation form of spectral convolution:

$$H_b^{(1)} = \sum_{k=0}^{K_b} \Theta_k^{(b)} T_k(\tilde{L}) Z', \quad b \in \{1, 2, 3\} \tag{9}$$

Where:

- b denotes the branch index;
- $K_b \in \{2, 4, 6\}$ represents the polynomial order for each branch;
- $\Theta_k^{(b)} \in \mathbb{R}^{F_{in} \times F_{out}}$ denotes the learnable weight matrix of the k-th order in branch b;
- $T_k(\cdot)$ is the Chebyshev polynomial of order k, recursively defined as:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \tag{10}$$

Each branch represents a distinct neighborhood receptive field, with the low-order branch (K=2) emphasizing local connectivity patterns, and the high-order branch (K=6) capturing global dependencies across clusters. Unlike conventional multi-branch architectures that directly concatenate branch outputs, we introduce a scale-level attention mechanism to adaptively fuse multi-scale representations. Specifically, attention coefficients are learned for each branch to quantify its relative importance, and the fused representation is computed as:

$$H^{(1)} = \sum_{b=1}^{3} \alpha_b H_b^{(1)} \tag{11}$$

where α_b denotes the learnable attention weight for branch b, normalized via a softmax function. This adaptive weighting strategy enables the model to dynamically emphasize the most informative structural scale for different nodes, thereby improving representation flexibility and robustness.

The second layer serves as a feature integration module, employing a standard GCN [37]to further propagate and refine the attention-weighted multi-scale representation:

$$\mathbf{H}^{(2)} = \sigma\left(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(1)}\mathbf{W}^{(1)}\right) \tag{12}$$

where:

- $\mathbf{W}^{(1)} \in \mathbb{R}^{F_1 \times F_2}$ denotes the learnable weight matrix for linear transformation;
- $\sigma(\cdot)$ represents the activation function;
- $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{D}}$ are defined as above.

To mitigate representation degradation and over-smoothing in deeper architectures, the third layer incorporates a Jumping Knowledge (JK) aggregation mechanism[38]. This module adaptively integrates representations from multiple depths:

$$\mathbf{H}_{JK} = \text{Concat}\left(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}\right) \tag{13}$$

By preserving both shallow multi-scale structural features and deeper propagated representations, the JK mechanism ensures that discriminative local patterns and high-order dependencies are simultaneously retained. This hierarchical integration significantly enhances the model's ability to capture multi-level structural information and improves generalization performance across heterogeneous input patterns.

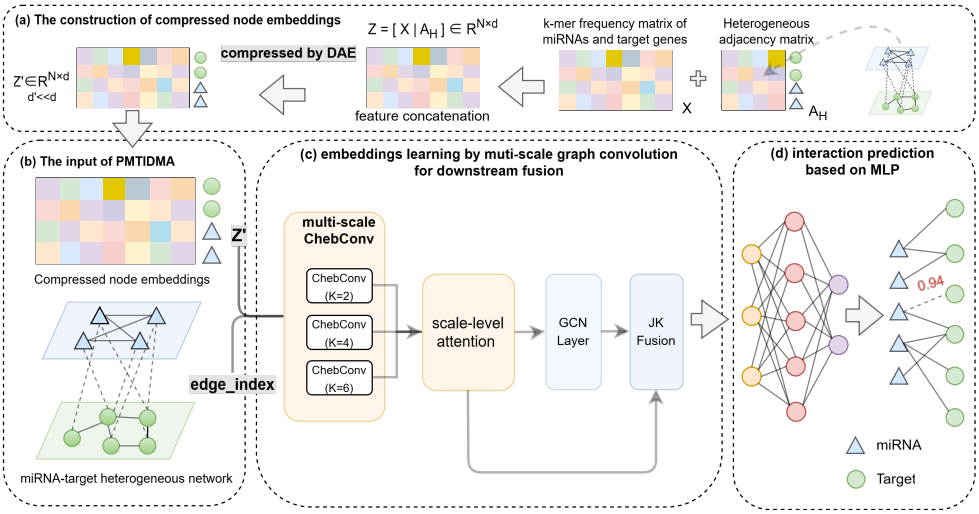


Fig. 2: Overview of the proposed PMTIDMA framework network

2.5 Edge Representation Construction and Probability Prediction

After multi-scale spectral graph convolution, the representation vectors of miRNA nodes and target gene nodes are obtained, denoted as H_{miRNA} and H_{target} respectively.For each miRNA–target pair, their node representations are concatenated to construct an edge feature vector:

$$H_{pair} = [H_{miRNA} \parallel H_{target}] \tag{14}$$

Subsequently, a MLP [39] is employed as a supervised learning model to predict the association probability. As shown in panel (d) of Figure 2. The MLP consists of three fully connected layers, taking as input the miRNA-target pair feature vectors extracted by the multi-scale graph convolution. Since association prediction is treated as a binary classification problem, a Sigmoid activation function is applied at the output layer to generate the predicted probability for each miRNA-target pair.

$$\begin{aligned} z &= \mathrm{MLP}(H_{pair}) \\ \hat{y} &= \sigma(z) \\ L &= -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \end{aligned} \tag{15}$$

2.6 Loss Function and Optimizer

During model training, BCE) is used as the loss function to measure the discrepancy between the model's predictions and the true labels. BCE is widely applied in binary classification tasks, effectively capturing the deviation between predicted probability distributions and the true distribution, thereby guiding the model to learn the probabilistic mapping between input features and target variables [40].For a given dataset $\{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{0, 1\}$ represents the true label and $\hat{y}_i \in [0, 1]$ denotes the model-predicted probability of the positive class, the BCE loss is defined as:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + (1-y_i)\log\left(1-\hat{y}_i\right)\right] \tag{16}$$

For the optimization algorithm, this study adopts the Adam optimizer. Adam combines the advantages of AdaGrad and RMSProp by adaptively estimating the first and second moments of the gradients, thereby enabling dynamic adjustment of the

learning rate [41]. Its update rule is defined as follows:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \\ \theta_{t+1} &= \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \end{aligned} \tag{17}$$

Here, g_t denotes the gradient at time step t, α is the learning rate, β_1 and β_2 are the decay rates for the first and second moment estimates, respectively, and ε is a smoothing term to prevent division by zero. Experimental results show that the Adam optimizer demonstrates good convergence speed and stability for the task in this study.

3 Results and discussion

3.1 Experimental Setup and Evaluation Metrics

The proposed PMTIDMA model was implemented using the PyTorch framework (version 2.0.0) [42], recognized for its high computational efficiency and modular design. Data preprocessing and matrix operations were performed using NumPy (version 1.23.5) and Pandas (version 1.5.3). Furthermore, PyTorch Geometric (version 2.3.0) [43], a geometric deep learning extension library for PyTorch, was employed to efficiently construct and train graph neural network models. Standard machine learning utilities were accessed via the Scikit-Learn library (version 1.2.2) [44]. Model training and evaluation were conducted on an NVIDIA GeForce RTX 3090 GPU under a CUDA 12.4 environment, where GPU acceleration substantially enhanced computational efficiency during both parameter optimization and prediction phases.

To ensure robustness and generalizability, model performance was evaluated using a 5-CV strategy. Experimentally validated miRNA–target pairs constituted the positive sample set, which was evenly divided into five subsets, while unverified potential pairs formed the negative sample set. In each iteration, 80% of the positive samples along with an equal number of randomly selected negative samples were employed for training, and the remaining 20% of positive samples together with the rest of the negative samples were used for testing [45].

Comprehensive performance assessment was conducted using both threshold-independent metrics—such as the area under the precision–recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC)—and threshold-dependent metrics, including accuracy (ACC), recall (REC), precision (PRE), and

F1-score. The threshold-dependent metrics were computed as follows:

$$\begin{cases} ACC = \dfrac{TP + TN}{TP + TN + FP + FN} \\ REC = \dfrac{TP}{TP + FN} \\ PRE = \dfrac{TP}{TP + FP} \\ F1 = \dfrac{2 \times REC \times PRE}{REC + PRE} \end{cases} \tag{18}$$

3.2 Hyperparameter Analysis

This section systematically evaluates the key hyperparameters of the PMTIDMA model to verify its stability and robustness in plant miRNA–target prediction. Specifically, we investigate the effects of the k-mer window size, sequence similarity calculation method, embedding dimension, as well as the learning rate and training epochs in the node representation learning stage. For the k-mer setting, multiple repeated experiments were conducted with $k \in [5, 6, 7, 8, 9, 10]$, as shown in Fig. 3, the results show that k = 7 achieves the highest median AUPRC with a more concentrated distribution, indicating superior performance and stability. To assess the impact of similarity construction strategies, Jaccard and Cosine similarities were compared, as illustrated in Fig. 4, although their overall performances are highly comparable, Jaccard demonstrates a slight advantage, suggesting that the model is not overly sensitive to the choice of similarity metric. Regarding the embedding dimension, as it increases from 16 to 512, the AUPRC consistently improves but exhibits a clear diminishing return pattern, As shown in Fig. 5,while 512D achieves the best performance, 256D offers a more favorable balance between representational capacity and memory consumption. For the training parameters, experiments under different learning rates [0.001, 0.0015, 0.0005, 0.0003, 0.0001] reveal that the model converges stably after approximately 3000 steps without signs of overfitting ,as presented in Fig. 6, and the optimal performance is obtained at a learning rate of 0.0001. Overall, a single-factor control strategy was employed to optimize each hyperparameter sequentially, and the final configuration was determined based on these findings.

3.3 Comparison with Other Models

To assess the effectiveness and performance of the proposed PMTIDMA model, systematic comparisons were conducted against several state-of-the-art deep learning frameworks as well as classical machine learning algorithms. Within the deep learning category, kmerPMTF, MRMTI and the GAT were selected as baseline models. The kmerPMTF framework utilizes a single spectral graph convolution network to extract latent semantic embeddings from sequence data [46], whereas MRMTI adopts a multi-relational integration strategy to fuse heterogeneous interaction information[47]. GAT employs graph embeddings in combination with attention mechanisms to achieve weighted information aggregation among nodes [48]. Both models were implemented

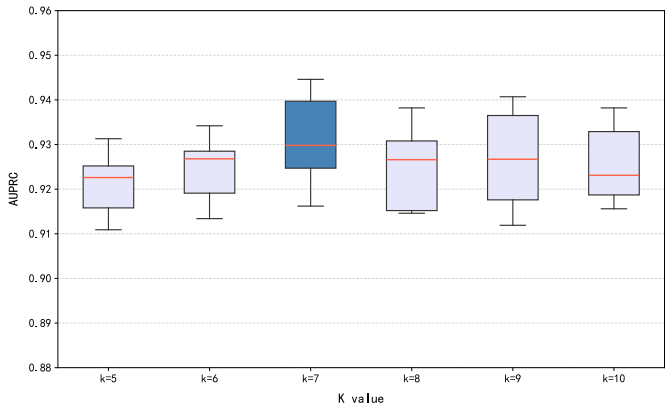


Fig. 3: The boxplot of AUPRC values under different k-mer window sizes (k = 5–10). Each box represents the distribution of results over multiple runs, with the median indicated by the central line.

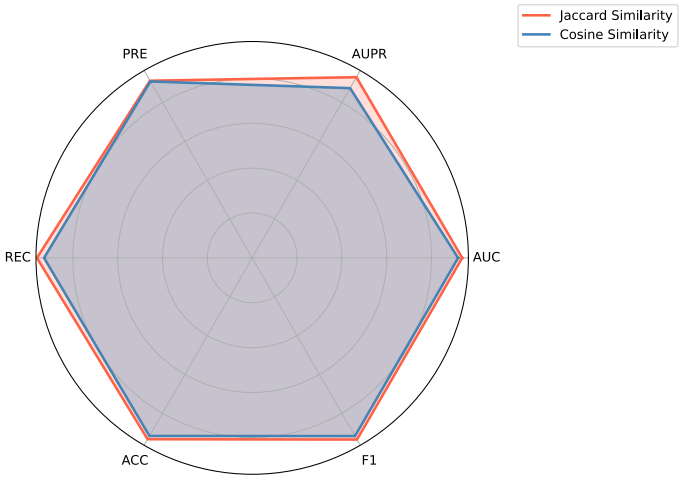


Fig. 4: Radar plot comparing the performance of PMTIDMA using different sequence similarity measures across six evaluation metrics.

and trained strictly according to the parameter settings recommended in their original publications, ensuring fairness and comparability of results.

Additionally, two representative traditional machine learning algorithms SVM-SVC and RF—were incorporated for contrastive evaluation [49, 50]. These multi-level comparative experiments enabled a comprehensive assessment of PMTIDMA's relative advantages, highlighting its superior ability to integrate sequence-specific and graph-based relational features. The evaluation further demonstrated the model's robustness and generalization capability across diverse methodological paradigms.

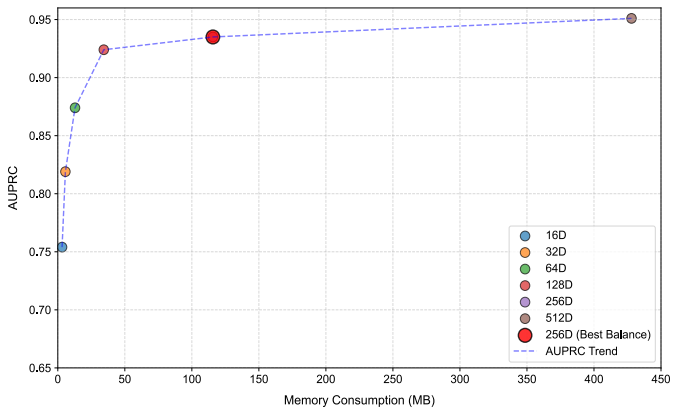
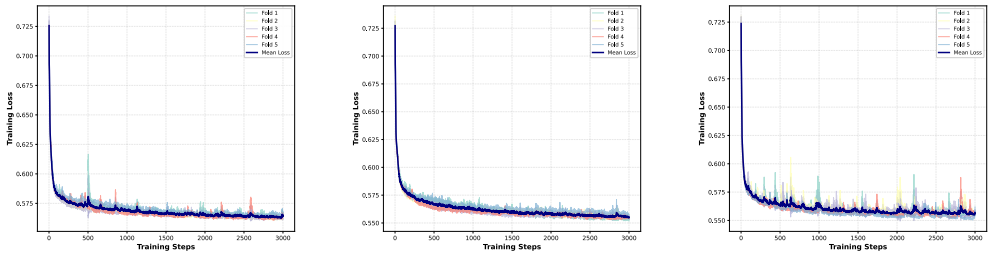


Fig. 5: AUPRC versus memory consumption under different embedding dimensions. While performance improves with higher dimensions, 256D achieves the best balance between accuracy and computational cost.



(a) A 5Fold Loss Curve (b) P 5Fold Loss Curve (c) S 5Fold Loss Curve

Fig. 6: Training loss curves of the model on three species under five-fold cross-validation. The loss decreases rapidly during the early training stage and gradually stabilizes, converging to a relatively steady state after approximately 3000 training steps.

As illustrated in Fig. 7, PMTIDMA consistently achieved superior performance across all datasets and evaluation metrics. On the Arabidopsis dataset, PMTIDMA attained an AUPRC and AUROC of approximately 0.93 and 0.94, respectively, markedly outperforming kmerPMTF (approximately 0.82 and 0.85), MRMTI (approximately 0.70 and 0.73), and GAT (approximately 0.70 and 0.75). Similarly, on the peach and tomato datasets, PMTIDMA maintained a leading performance across all metrics, with an average F1-score improvement of 10%–20% over competing models, demonstrating a clear advantage in distinguishing positive and negative samples as well as achieving overall prediction stability. Furthermore, as shown in Fig. 8, the PR and ROC curves on the Arabidopsis dataset further confirm the strong discriminative ability and robustness of PMTIDMA under five-fold cross-validation.

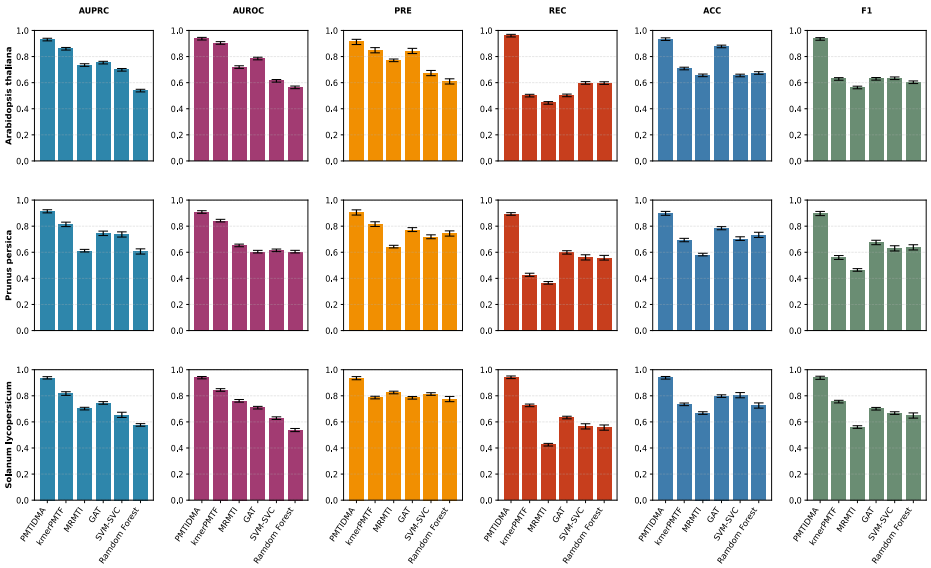


Fig. 7: The results of comparison experiments

When comparing model performance, kmerPMTF generally performed worse than PMTIDMA, highlighting that relying solely on sequence features and a single spectral graph convolution representation is insufficient to capture the complex topological relationships between miRNAs and genes. Specifically, the kmerPMTF model, with its single-feature learning approach, tends to struggle near the decision boundary between positive and negative samples, leading to lower recall (REC). In other words, this model tends to make conservative predictions, resulting in some true positive samples being misclassified as negative. MRMTI, although integrating multiple relational signals, was originally designed and optimized for animal miRNA-target interaction datasets. The biological regulatory mechanisms and interaction patterns in plant systems differ substantially from those in animals, particularly in terms of complementarity patterns and target recognition rules. GAT, on the other hand, primarily aggregates information within local neighborhoods, which makes it challenging to capture global dependencies between distant nodes. Additionally, the attention weights in GAT are computed based on feature similarity, but in a miRNA–gene heterogeneous graph, the feature distributions between different types of nodes vary significantly. This discrepancy limits the attention mechanism's ability to accurately reflect the true biological interaction strength, thereby hindering its feature expression power and predictive performance. Meanwhile, SVM-SVC and Random Forest showed the weakest performance across all metrics, particularly in AUPRC, REC, and F1-score, highlighting the challenges traditional machine learning methods face in modeling complex, high-dimensional biological sequences and network structural features.

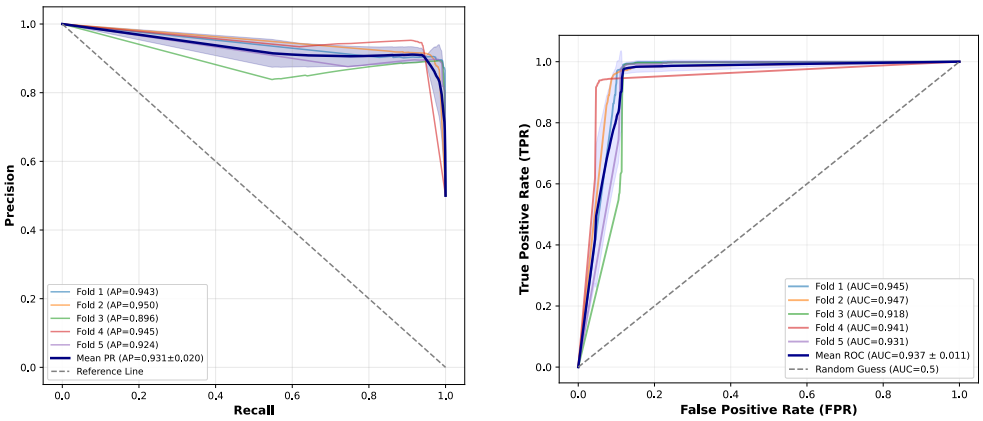


Fig. 8: PR curve and ROC curve on the Arabidopsis thaliana dataset

3.4 Ablation Study

To systematically evaluate the independent contributions of the denoising autoencoder (DAE), multi-scale spectral graph convolutional layer, standard GCN layer and Jumping Knowledge (JK) module in the framework of this paper, we designed five ablation variants (No1 - No5) based on the complete model structure, represented by the Arabidopsis thaliana dataset. A comparative analysis was conducted with the complete model (No6) in six evaluation indicators. Specifically, Model No1 retains multi-scale spectral graph convolution, standard GCN layers, and JK modules in the node representation learning stage, but replaces the denoising autoencoder with a common autoencoder in the feature compression stage to evaluate the denoising modeling ability of DAE. Model No2 employs a denoising autoencoder for feature compression, but only uses a single-scale spectral graph convolution in the graph convolution stage to verify the effectiveness of the multi-scale structure. Model No3 removes the attention weighting mechanism among the integral branches of multi-scale spectral graph volumes on the basis of the complete model and replaces it with simple feature concatenation to investigate the effect of the attention fusion strategy. Models No4 and No5 respectively remove the standard GCN layer and JK module to analyze the significance of deep structure modeling and cross-layer feature aggregation.

As shown in Table 2, When the denoising autoencoder (DAE) was replaced with a conventional autoencoder (No1), multiple metrics declined, indicating that DAE enhances feature robustness and suppresses noise propagation during graph convolution, thereby improving discriminative ability. In No2, using single-scale convolution achieved the highest precision but reduced recall and F1, suggesting that fixed receptive fields favor conservative predictions while lacking the capability to capture long-range dependencies. The multi-scale structure, in contrast, improves global structural modeling and overall balance.

Removing the attention mechanism (No3) led to performance instability despite higher recall, implying that simple concatenation introduces redundant information and weakens feature discrimination. The attention mechanism adaptively weights

multi-scale features, enhancing fusion quality. In No4, eliminating the standard GCN layer resulted in the most significant performance drop, demonstrating its essential role in local neighborhood aggregation and stable representation learning. In No5, removing the JK module slightly improved AUPRC but reduced overall robustness, indicating that cross-layer feature aggregation alleviates over-smoothing and enhances generalization.

Table 2: Ablation Study Results

No.	DAE	Multi-scale chebconv	Attn	GCN	JK	AUPRC	AUROC	PRE	REC	ACC	F1
1	×	✓	✓	✓	✓	0.86699	0.92127	0.89968	0.91031	0.90414	0.90452
2	✓	×	/	✓	✓	0.91080	0.91859	**0.92248**	0.91689	0.91977	0.91945
3	✓	✓	×	✓	×	0.92857	0.93673	0.91023	0.94980	0.92732	0.92874
4	✓	✓	✓	×	/	0.92441	0.92162	0.90968	0.86833	0.89001	0.88411
5	✓	✓	✓	✓	×	**0.93391**	0.92454	0.91026	0.91935	0.91305	0.91185
6	✓	✓	✓	✓	✓	0.93021	**0.93690**	0.91207	**0.96023**	**0.93335**	**0.93505**

Overall, the complete model (No6) achieves the most stable and balanced performance, as each module contributes complementary functionality and forms a synergistic effect, leading to improved global discrimination and positive sample recognition.

3.5 Case Study and Validation

To further assess the predictive ability and biological reliability of PMTIDMA in real-world applications, we conducted a case study using two soybean (Glycine max) miRNAs involved in stress response and hormone signaling—gma-miR172m and gma-miR396c. The verified target of gma-miR172m is the AP2 transcription factor family, while gma-miR396c primarily targets GRFs[51, 52].We input soybean transcriptome data, mature miRNA sequences, and interaction information into PMTIDMA to predict candidate target genes and their corresponding scores. For each miRNA, the top five genes with the highest prediction scores were selected as high-confidence candidates. The predictions were then cross-validated using the PMRD and SoyBase databases[53, 54]. Validation criteria included:

(1) Whether the target gene has been experimentally validated or computationally predicted as an miRNA target in the database;

(2) Whether the annotation information of the target gene in the database is consistent with the known miRNA target gene family.

Upon investigation, all high-confidence predicted target genes were supported by entries in the PMRD database. For example, Glyma11g15650 is annotated in SoyBase as an "ethylene-responsive transcription factor 3," belonging to a major subclass of the AP2/ERF superfamily. Similarly, Glyma19g36200 carries the same functional

annotation, and its gene family and domain, IPR016177 (DNA-binding domain), correspond to the DNA-binding domain characteristic of the AP2/ERF family.On the other hand, the target gene Glyma06g13960 is described as a "growth-regulating factor 7-like isoform X1," which aligns with the canonical known targets of gma-miR396c. The second-ranked gene, Glyma01g34650, is annotated as "growth-regulating factor 9," confirming its membership in the GRF family.

Table 3: The case study results of PMTIDMA

miRNA	Ranking	Target	Score	Evidence
gma-miR172	1	Glyma11g15650	0.99	PMRD,SoyBase
	2	Glyma12g07800	0.99	PMRD,SoyBase
	3	Glyma15g04930	0.98	PMRD,SoyBase
	4	Glyma19g36200	0.98	PMRD,SoyBase
	5	Glyma01g04770	0.98	PMRD,SoyBase
gma-miR396	1	Glyma06g13960	0.99	PMRD,SoyBase
	2	Glyma01g34650	0.98	PMRD,SoyBase
	3	Glyma12g01730	0.98	PMRD,SoyBase
	4	Glyma13g11630	0.98	PMRD,SoyBase
	5	Glyma19g37740	0.97	PMRD,SoyBase

The experimental results demonstrate that the proposed model effectively captures the key sequence complementarity patterns between miRNAs and their target genes, enabling accurate identification of highly correlated targets. The superior predictive performance can be attributed to the synergistic integration of multiple components. At the node representation level, the model jointly incorporates sequence features and graph topological information, achieving deep fusion of attribute and structural characteristics. The denoising autoencoder enhances feature robustness during compression by introducing perturbation-based learning, extracting informative latent representations while suppressing noise that could otherwise propagate through graph convolution layers. The multi-scale spectral graph convolution module aggregates neighborhood information across different receptive ranges, strengthening the model's ability to capture both local structural patterns and long-range dependencies, thereby improving its representation of complex topological relationships. Furthermore, the attention-based fusion mechanism optimizes the integration of multi-scale features, while the Jumping Knowledge (JK) module aggregates representations from different network depths to mitigate over-smoothing and enhance hierarchical feature integration. The collaborative interaction among these modules ultimately leads to improved prediction accuracy and generalization capability.

4 Conclusion

In this study, we proposed PMTIDMA, a deep learning framework for plant miRNA–target interaction prediction. By integrating sequence-derived k-mer features with heterogeneous graph structural information, the model achieves joint modeling

of attribute and topological representations. Architecturally, PMTIDMA incorporates a denoising autoencoder, multi-scale spectral graph convolution, an attention-based fusion mechanism, and a Jumping Knowledge module, enabling effective capture of local sequence patterns and hierarchical topological dependencies across different receptive ranges and network depths.

Extensive experiments on three plant datasets demonstrate that PMTIDMA consistently outperforms existing methods across multiple evaluation metrics, exhibiting stable and well-balanced predictive performance. Ablation studies further confirm the substantial contribution of each module to overall model effectiveness. A cross-species case study on soybean highlights the framework's generalization capability and its potential to identify novel miRNA–target interactions.

Despite these advantages, the current model primarily relies on sequence-derived representations and does not yet fully incorporate broader biological prior knowledge, such as structural domain information or discontinuous binding patterns. Moreover, similar to many deep graph learning models, PMTIDMA still faces challenges in interpretability. Future work will focus on integrating multimodal biological features and developing interpretable mechanisms to enhance biological transparency and practical reliability.

Declarations

Funding

This work was supported by the Hubei Province Key Research and Development Special Project of Science and Technology Innovation Plan(Grant no. 2023BAB087), the Wuhan Key Research and Development Projects(Grant no. 2023010402010614), the open competition project for selecting the best candidates, Wuhan East Lake High-tech Development Zone(Grant no. 2024KJB328), the Fund for Research Platform of South-Central Minzu University(Grant no. PTZ25001), the Central Government Guides Local Funds for Science and Technology Development(Grant no. ZYYD2024QY08).

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The dataset used in this study can be easily derived from https://doi.org/10.5281/zenodo.19439708.

References

[1] Guo, L., Zhao, Y., Yang, S., Zhang, H., Chen, F.: Integrative analysis of mirna-mrna and mirna-mirna interactions. BioMed Research International **2014**, 1–8 (2014) https://doi.org/10.1155/2014/907420

[2] Deng, L., Ren, S., Zhang, J.: Prediction of lncrna functions using deep neural networks based on multiple networks. BMC Genomics **23** (2023) https://doi.org/10.1186/s12864-023-09578-w

[3] Bartel, D.P.: Micrornas: Target recognition and regulatory functions. Cell **136**(2), 215–233 (2009) https://doi.org/10.1016/j.cell.2009.01.002

[4] Mak, K.W.Y., Mustafa, A.F., Belsham, D.D.: Neuroendocrine micrornas linked to energy homeostasis: future therapeutic potential. Pharmacological Reports **74**(5), 774–789 (2022) https://doi.org/10.1007/s43440-022-00409-5

[5] Rosatti, S., Rojas, A.M.L., Moro, B., Suarez, I., Bologna, N., Chorostecki, U., Palatnik, J.: Principles of mirna/mirna* function in plant mirna processing. Nucleic Acids Research **52**(14), 8356–8369 (2024) https://doi.org/10.1093/nar/gkae458

[6] WEI, Q., LIANG, Y.-H., LI, G.-L.: Evolution of mirna in plants. Hereditas (Beijing) **35**(3), 315–323 (2013) https://doi.org/10.3724/sp.j.1005.2013.00315

[7] Li, J., Song, Q., Zuo, Z.-F., Liu, L.: Microrna398: A master regulator of plant development and stress responses. International Journal of Molecular Sciences **23**(18), 10803 (2022) https://doi.org/10.3390/ijms231810803

[8] Kuhn, D.E., Martin, M.M., Feldman, D.S., Terry, A.V., Nuovo, G.J., Elton, T.S.: Experimental validation of mirna targets. Methods **44**(1), 47–54 (2008) https://doi.org/10.1016/j.ymeth.2007.09.005

[9] Mao, S.: Biochemical prediction of mirna targeting. Science **366**(6472), 1467–11469 (2019) https://doi.org/10.1126/science.366.6472.1467-j

[10] Krol, J., Loedige, I., Filipowicz, W.: The widespread regulation of microrna biogenesis, function and decay. Nature Reviews Genetics **11**(9), 597–610 (2010) https://doi.org/10.1038/nrg2843

[11] Yu, T., Xu, N., Haque, N., Gao, C., Huang, W., Huang, Z.: Popular computational tools used for mirna prediction and their future development prospects. Interdisciplinary Sciences: Computational Life Sciences **12**(4), 395–413 (2020) https://doi.org/10.1007/s12539-020-00387-3

[12] Pan, J.: Deep learning boosts rna prediction. Nature Computational Science **1**(9), 564 (2021) https://doi.org/10.1038/s43588-021-00134-8

[13] Giansanti, V., Castelli, M., Beretta, S., Merelli, I.: Comparing deep and machine learning approaches in bioinformatics: A mirna-target prediction case study. Lecture Notes in Computer Science, 31–44 (2019) https://doi.org/10.1007/978-3-030-22744-9_3

[14] Parveen, A., Mustafa, S.H., Yadav, P., Kumar, A.: Applications of machine learning in mirna discovery and target prediction. Current Genomics **20**(8), 537–544 (2020) https://doi.org/10.2174/1389202921666200106111813

[15] Xu, C., Jackson, S.A.: Machine learning and complex biological data. Genome Biology **20**(1) (2019) https://doi.org/10.1186/s13059-019-1689-0

[16] Allison, D.B., Cui, X., Page, G.P., Sabripour, M.: Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics **7**(1), 55–65 (2006) https://doi.org/10.1038/nrg1749

[17] Lewis, B.P., Burge, C.B., Bartel, D.P.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. Cell **120**(1), 15–20 (2005) https://doi.org/10.1016/j.cell.2004.12.035

[18] Cherkassky, V., Ma, Y.: Another look at statistical learning theory and regularization. Neural Networks **22**(7), 958–969 (2009) https://doi.org/10.1016/j.neunet.2009.04.005

[19] Cao, H., Kapranov, P.: Methods to analyze the non-coding rna interactome—recent advances and challenges. Frontiers in Genetics **13** (2022) https://doi.org/10.3389/fgene.2022.857759

[20] Ai, N., Liang, Y., Yuan, H., Ouyang, D., Xie, S., Liu, X.: Gdcl-ncda: identifying non-coding rna-disease associations via contrastive learning between deep graph learning and deep matrix factorization. BMC Genomics **24**(1) (2023) https://doi.org/10.1186/s12864-023-09501-3

[21] Gu, T., Zhao, X., Barbazuk, W.B., Lee, J.-H.: mitar: a hybrid deep learning-based approach for predicting mirna targets. bioRxiv (2020) https://doi.org/10.1101/2020.04.02.022608

[22] Axtell, M.J., Westholm, J.O., Lai, E.C.: Vive la différence: biogenesis and evolution of micrornas in plants and animals. Genome Biology **12**(4), 221 (2011) https://doi.org/10.1186/gb-2011-12-4-221

[23] Li, Z., Xu, R., Li, N.: Micrornas from plants to animals, do they define a new messenger for communication? Nutrition & Metabolism **15**(1) (2018) https://doi.org/10.1186/s12986-018-0305-8

[24] Jones-Rhoades, M.W., Bartel, D.P., Bartel, B.: Micrornas and their regulatory roles in plants. Annual Review of Plant Biology **57**(1), 19–53 (2006) https://doi.

org/10.1146/annurev.arplant.57.032905.105218

[25] Nie, Z., Gao, M., Jin, X., Rao, Y., Zhang, X.: Mfpinc: prediction of plant ncrnas based on multi-source feature fusion. BMC Genomics **25**(1) (2024) https://doi.org/10.1186/s12864-024-10439-3

[26] Somerville, C., Koornneef, M.: A fortunate choice: the history of arabidopsis as a model plant. Nature Reviews Genetics **3**(11), 883–889 (2002) https://doi.org/10.1038/nrg927

[27] Kurubanjerdjit, N., Huang, C.-H., Lee, Y.-l., Tsai, J.J.P., Ng, K.-L.: Prediction of microrna-regulated protein interaction pathways in arabidopsis using machine learning algorithms. Computers in Biology and Medicine **43**(11), 1645–1652 (2013) https://doi.org/10.1016/j.compbiomed.2013.08.010

[28] Kimura, S., Sinha, N.: Tomato (solanum lycopersicum): A model fruit-bearing crop. Cold Spring Harbor Protocols **2008**(11) (2008) https://doi.org/10.1101/pdb.emo105

[29] Martínez-Gómez, P., Sánchez-Pérez, R., Rubio, M.: Clarifying omics concepts, challenges, and opportunities for prunus breeding in the postgenomic era. OMICS: A Journal of Integrative Biology **16**(5), 268–283 (2012) https://doi.org/10.1089/omi.2011.0133

[30] Guo, Z., Kuang, Z., Zhao, Y., Deng, Y., He, H., Wan, M., Tao, Y., Wang, D., Wei, J., Li, L., Yang, X.: Pmiren2.0: from data annotation to functional exploration of plant micrornas. Nucleic Acids Research **50** (2021) https://doi.org/10.1093/nar/gkab811

[31] Trabelsi, A., Chaabane, M., Ben-Hur, A.: Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. arXiv preprint arXiv:1905.03940 (2019)

[32] Xie, W., Luo, J., Pan, C., Liu, Y.: Sg-lstm-frame: a computational frame using sequence and geometrical information via lstm to predict mirna–gene associations. Briefings in Bioinformatics **22**(2), 2032–2042 (2020) https://doi.org/10.1093/bib/bbaa022

[33] Thadani, R., Tammi, M.T.: Microtar: predicting microrna targets from rna duplexes. BMC Bioinformatics **7** (2006) https://doi.org/10.1186/1471-2105-7-s5-s20

[34] Axtell, M.J., Meyers, B.C.: Revisiting criteria for plant microrna annotation in the era of big data. The Plant Cell **30**(2), 272–284 (2018) https://doi.org/10.1105/tpc.17.00851

[35] Sætrom, P., Heale, B.S.E., Snøve, O., Aagaard, L., Alluin, J., Rossi, J.J.: Distance constraints between microrna target sites dictate efficacy and cooperativity. Nucleic Acids Research **35**(7), 2333–2342 (2007) https://doi.org/10.1093/nar/gkm133

[36] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research (2010)

[37] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2016) https://doi.org/10.48550/ARXIV.1609.02907

[38] Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., Jegelka, S.: Representation learning on graphs with jumping knowledge networks (2018)

[39] Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review **65**(6), 386–408 (1958) https://doi.org/10.1037/h0042519

[40] Klüver, J.: A mathematical theory of communication: Meaning, information, and topology. Complexity **16**(3), 10–26 (2011) https://doi.org/10.1002/cplx.20317

[41] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014) https://doi.org/10.48550/ARXIV.1412.6980

[42] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. Neural Information Processing Systems (2019)

[43] Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric (2019) https://doi.org/10.48550/ARXIV.1903.02428

[44] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in python. arXiv (2012) https://doi.org/10.48550/ARXIV.1201.0490

[45] Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning (2018) https://doi.org/10.48550/ARXIV.1811.12808

[46] Zhang, W., Zhang, P., Sun, W., Xu, J., Liao, L., Cao, Y., Han, Y.: Improving plant mirna-target prediction with self-supervised k-mer embedding and spectral graph convolutional neural network. PeerJ **12** (2024) https://doi.org/10.7717/peerj.17396

[47] Luo, J., Ouyang, W., Shen, C., Cai, J.: Multi-relation graph embedding for predicting mirna-target gene interactions by integrating gene sequence information. IEEE Journal of Biomedical and Health Informatics **26**(8), 4345–4353 (2022) https://doi.org/10.1109/jbhi.2022.3168008

[48] Ben, J., Sun, Q., Liu, K., Yang, X., Zhang, F.: Multi-head multi-order graph attention networks. Applied Intelligence **54**, 8092–8107 (2024) https://doi.org/10.1007/s10489-024-05601-z

[49] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995) https://doi.org/10.1007/bf00994018

[50] Breiman, L.: Random forests. Machine learning (2001)

[51] Wang, Y., Wang, L., Zou, Y., Chen, L., Cai, Z., Zhang, S., Zhao, F., Tian, Y., Jiang, Q., Ferguson, B.J., Gresshoff, P.M., Li, X.: Soybean mir172c targets the repressive ap2 transcription factor nnc1 to activate ¡i¿enod40¡/i¿ expression and regulate nodule initiation. The Plant Cell **26**(12), 4782–4801 (2014) https://doi.org/10.1105/tpc.114.131607

[52] Liu, W., Zhou, Y., Li, X., Wang, X., Dong, Y., Wang, N., Liu, X., Chen, H., Yao, N., Cui, X., Jameel, A., Wang, F., Li, H.: Tissue-specific regulation of gma-mir396 family on coordinating development and low water availability responses. Frontiers in Plant Science **8** (2017) https://doi.org/10.3389/fpls.2017.01112

[53] Zhang, Z., Yu, J., Li, D., Zhang, Z., Liu, F., Zhou, X., Wang, T., Ling, Y., Su, Z.: Pmrd: plant microrna database. Nucleic Acids Research **38**(suppl_1), 806–813 (2009) https://doi.org/10.1093/nar/gkp818

[54] Grant, D., Nelson, R.T.: Soybase: A comprehensive database for soybean genetic and genomic data. Compendium of Plant Genomes, 193–211 (2017) https://doi.org/10.1007/978-3-319-64198-0_13