https://doi.org/10.54589/aol.39/1/43

Comparative Analysis of Graph Attention and Graph Convolutional Networks for Predicting Candidate SNP–miRNA–Drug Interactions in Integrated Regulatory Datasets

Varun Batra[1], Pradeep K Yadalam[1], Carlos M Ardila[1,2]

1. Saveetha University, Saveetha Dental College & Hospital, Department of Periodontics, Chennai, Tamil Nadu, India
2. Universidad de Antioquia, Facultad de Odontología, Department of Ciencias Básicas, Medellin, Colombia

ABSTRACT

Complex genetic and epigenetic factors, such as single-nucleotide polymorphisms (SNPs) that influence the function of microRNAs (miRNAs), are involved in periodontal inflammation and tissue regeneration. However, computational approaches for predicting potential SNP–miRNA interactions that may influence inflammatory signaling and drug responses remain limited in precision periodontology. ***Aim:*** *To implement and compare two graph neural network–based machine learning models—Graph Attention Network (GAT) and Graph Convolutional Network (GCN)— for predicting potential SNP–miRNA–drug interactions that may be relevant to inflammatory and regenerative pathways associated with periodontal disease.* ***Materials and Method:*** *SNPs found in miRNA precursor and target regions associated with immune and drug-response genes were included in three selected miRNA SNP v3.0 database datasets. After preprocessing, composite vectors were created by early fusion of heterogeneous features. Principal component analysis (PCA) was then used to normalize and reduce the composite vectors. Using matched hyperparameters, the GAT and GCN models were built in PyTorch Geometric and trained with cross-entropy loss. Regression metrics (RMSE, MAE, R2) and classification metrics (accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC) were used to evaluate performance.* ***Results:*** *With a perfect recall (1.0), F1-score of 0.666, and PR-AUC of 0.750, GAT performed better in classification than GCN, demonstrating higher sensitivity for identifying candidate interactions within the integrated dataset. Although R^2 values were close to zero, indicating limited variance explanation, regression performance was comparable for both models (RMSE ≈ 0.524). Visual diagnostics showed that GAT maintained more prediction variability, while GCN predictions collapsed around mean values.* ***Conclusions:*** *Because of its attention-based architecture, GAT showed improved performance in classification metrics compared with GCN, suggesting potential advantages for prioritizing candidate SNP–miRNA–drug interactions within integrated regulatory datasets. However, the limited predictive performance observed indicates that these findings should be interpreted primarily as a comparative computational evaluation rather than as direct biological evidence. Future studies incorporating periodontal-specific molecular datasets and experimentally validated interactions will be necessary to improve biological interpretability and translational relevance.*

Key words: *periodontitis - polymorphism - single nucleotide - microRNAs - machine learning*

Análisis comparativo de redes de atención de grafos y redes convolucionales de grafos para predecir interacciones candidatas entre SNP, miRNA y fármacos en conjuntos de datos regulatorios integrados

RESUMEN

Factores genéticos y epigenéticos complejos, como los polimorfismos de nucleótido único (SNPs) que influyen en la función de los microARNs (miRNAs), participan en la inflamación periodontal y la regeneración tisular. Sin embargo, en el marco de la periodoncia de precisión siguen siendo limitados los enfoques computacionales para predecir interacciones potenciales SNP–miRNA que podrían influir en la señalización inflamatoria y en las respuestas a fármacos. ***Objetivo:*** *Implementar y comparar dos modelos de aprendizaje automático basados en redes neuronales de grafos—Graph Attention Network (GAT) y Graph Convolutional Network (GCN)— para predecir posibles interacciones SNP–miRNA–fármaco que podrían ser relevantes para las vías inflamatorias y regenerativas asociadas con la enfermedad periodontal.* ***Materiales y Método****: Se incluyeron SNPs ubicados en regiones precursoras de miRNA y en regiones diana, asociados con genes vinculados a respuesta inmune y respuesta a fármacos, a partir de tres conjuntos de datos seleccionados de la base miRNA SNP v3.0. Tras el preprocesamiento, se construyeron vectores compuestos mediante fusión temprana de características heterogéneas. Posteriormente, se aplicó análisis de componentes principales (PCA) para normalizar y reducir la dimensionalidad. Los modelos GAT y GCN se implementaron en PyTorch Geometric, se entrenaron con pérdida de entropía cruzada y se evaluaron usando métricas de regresión (RMSE, MAE, R^2) y de clasificación (exactitud, precisión, sensibilidad, F1-score, ROC-AUC, PR-AUC).* ***Resultados****: Con una sensibilidad perfecta (1,0), una puntuación F1 de 0,666 y un área bajo la curva precisión–recobrado (PR-AUC) de 0,750, el modelo GAT superó al modelo GCN en el desempeño de clasificación, mostrando una mayor sensibilidad para la identificación de interacciones candidatas dentro del conjunto de datos regulatorio integrado. Aunque los valores de R^2 fueron cercanos a cero—lo que sugiere una explicación limitada de la varianza—el rendimiento en regresión fue comparable entre ambos modelos (RMSE ≈ 0,524). Los diagnósticos visuales evidenciaron que GAT conservó mayor variabilidad en las predicciones, mientras que las predicciones de GCN tendieron a concentrarse alrededor de valores medios.* ***Conclusiones:*** *Debido a su arquitectura basada en mecanismos de atención, el modelo GAT mostró un mejor desempeño en métricas de clasificación en comparación con GCN, lo que sugiere posibles ventajas para priorizar interacciones candidatas SNP–miRNA–fármacos dentro de conjuntos de datos regulatorios integrados. No obstante, los valores de R^2 cercanos a cero indican una capacidad limitada para modelar la fuerza continua de interacción, por lo que los resultados deben interpretarse principalmente como una evaluación computacional comparativa. Estos hallazgos destacan el potencial de los enfoques de aprendizaje automático basados en grafos para explorar interacciones regulatorias candidatas y sugieren la necesidad de futuros estudios que incorporen conjuntos de datos moleculares específicos de enfermedad, modelos más interpretables y arquitecturas híbridas para mejorar la relevancia biológica y la aplicabilidad translacional.*

Palabras claves: *periodontitis - polimorfismo de nucleótido único - microARNs - aprendizaje automático*

To cite: Batra V, Yadalam PK, Ardila CM. Comparative Analysis of Graph Attention and Graph Convolutional Networks for Predicting Candidate SNP–miRNA–Drug Interactions in Integrated Regulatory Datasets. Acta Odontol Latinoam. 2026 Jun 26;39(1):43-53. https://doi.org/10.54589/aol.39/1/43

Corresponding Author:
Carlos M Ardila
martin.ardila@udea.edu.co

Received: September 2025
Accepted: June 2026

INTRODUCTION

Chronic inflammatory diseases known as periodontal diseases majorly impact oral and systemic health because they cause tooth-supporting tissues to deteriorate gradually. Although regenerative therapies that aim to repair periodontal tissues, including bone, cementum, and connective tissue, have gained clinical traction, patient response variability remains a significant obstacle[1–3]. Small noncoding RNA molecules known as microRNAs (miRNAs) are essential for post-transcriptional regulation. They can alter gene regulation by binding to the 3′ untranslated region of mRNAs. Gene regulation may be modulated by genetic variations in miRNAs or target binding sites that impact miRNA biogenesis or target binding. Research has indicated that human diseases such as autoimmune disorders and breast cancer are linked to an increase in miRNA-related SNPs[4,5]. Understanding miRNA regulation and choosing functional variants depend on identifying and analyzing these miRNA-related SNPs. Although many resources have been created to find and examine miRNA-related SNPs, most mainly concentrate on humans. SNPs have become much more numerous, especially for animals. It is now crucial to have a thorough database that includes miRNA-related SNPs and their functional profiles across several species[3-5].

According to recent developments in precision periodontology, interindividual genetic variations, specifically single-nucleotide polymorphisms (SNPs) that impact regulatory noncoding RNAs like miRNAs, may be partially responsible for this variability. Important post-transcriptional modulators of gene expression are miRNAs. SNPs in miRNA sequences or their target locations can alter miRNA biogenesis, binding affinity, and ultimately, downstream gene expression linked to tissue remodeling and inflammation. The importance of these SNP–miRNA interactions in modifying drug response and periodontal tissue regeneration is becoming more widely acknowledged. The intricate interactions among genomic variations, miRNA regulation, and pharmacological response are still poorly understood in periodontal disease[6,7].

Graph Neural Networks (GNNs)[8,9] since the advent of graph-based deep learning, which encodes relational structures between entities like genes, miRNAs, SNPs, and medications, has demonstrated promise in modeling such complex biological systems. Graph Attention Networks (GATs) and Graph Convolutional Networks (GCNs), in particular, provide strong frameworks for learning informative node representations and capturing topological dependencies in biological networks. While GCNs aggregate neighborhood information based on graph Laplacians, GATs employ attention mechanisms to learn adaptive weights for neighboring nodes—an advantage in heterogeneous networks where regulatory influence varies across connections[10,11].

A systematic comparison of GCN and GAT architectures for predicting SNP–miRNA–drug interactions in periodontal inflammation and regeneration is still largely unexplored despite the growing use of GNNs in systems biology. Furthermore, not many integrated models capture the complexity of regulatory SNPs influencing miRNA-mediated gene control within pharmacogenomic landscapes pertinent to periodontal health. Instead, most studies concentrate either on genomic features or on drug-target prediction separately. GCNs and GATs[9,12] have been effectively used in earlier research to model intricate biological networks, such as multi-view attention-based regulatory networks, drug–target interactions, and miRNA–disease associations. However, their use in SNP-modulated post-transcriptional regulation in regenerative periodontology has not yet been investigated. This study addresses this gap by constructing a fused multi-modal dataset integrating SNPs in pre-miRNA and miRNA target regions with immune- and drug-response-related genes. We evaluate the performance of GAT and GCN models in both regression and classification settings to predict functionally relevant SNP–miRNA–drug interactions. Our goal is to identify critical regulatory variants that may influence individual responses to regenerative therapies and inflammatory modulation, offering a step toward more personalized and predictive periodontal care.

Therefore, this study aims to implement and compare two graph-based deep learning models—GAT and GCN—for predicting SNP–miRNA–drug interactions relevant to periodontal inflammation and tissue regeneration. By constructing a biologically curated fused dataset that integrates SNPs in pre-miRNA and target site regions with drug-responsive and inflammation-related genes,

we leverage the power of graph representation learning to model complex regulatory relationships. Specifically, we evaluate the models in regression and classification settings to assess their ability to prioritize SNPs that may modulate miRNA function and therapeutic response. Through this comparative framework, we seek to uncover novel regulatory insights and provide a foundation for precision-guided therapeutic strategies in periodontology.

MATERIALS AND METHOD

Data Source and Retrieval

The datasets used in this study came from the publicly available miRNA SNP v3.0 database[4], which provides comprehensive annotations of SNPs that affect miRNA function, such as interactions between miRNA target sites and miRNA biogenesis. The "utr_immune" dataset contains predicted SNPs within miRNA target sites located in the 3′ untranslated regions (3′UTRs) of immune-related genes; the "utr_drug_response" dataset, which includes SNPs within miRNA target sites of genes associated with pharmacogenomic responses; and the "DRV_premiRNA" dataset, which contains information on SNPs located within precursor miRNA (pre-miRNA) sequences that may have an impact on miRNA maturation, stability, or expression. It is important to note that the filtering strategy was designed to enrich the dataset with genes associated with inflammatory and regenerative processes, which are relevant to periodontal biology but not exclusive to periodontal disease. Therefore, the dataset should be interpreted as a general regulatory interaction resource rather than a periodontal-specific dataset. All three datasets were specifically retrieved in tab-separated values (TSV) format.

An illustration of the comprehensive workflow of this study, from data acquisition to final evaluation, is shown in Figure 1. The analysis began with the extraction of SNPs located in miRNA precursor and target regions from the miRNA SNP v3.0 database, specifically focusing on genes associated with immune response and drug metabolism. Following data collection, composite feature vectors were created through early fusion of heterogeneous features, which were then normalized using principal component analysis (PCA) and divided into training and validation datasets. For model development, we implemented both GAT and GCN architectures in PyTorch Geometric, ensuring consistent hyperparameters across both models. The evaluation phase employed rigorous metrics, including regression analysis (RMSE, R^2) and classification performance indicators (F1-score, ROC-AUC), ultimately revealing GAT's superior sensitivity in detecting biologically significant SNP-miRNA-drug interactions.

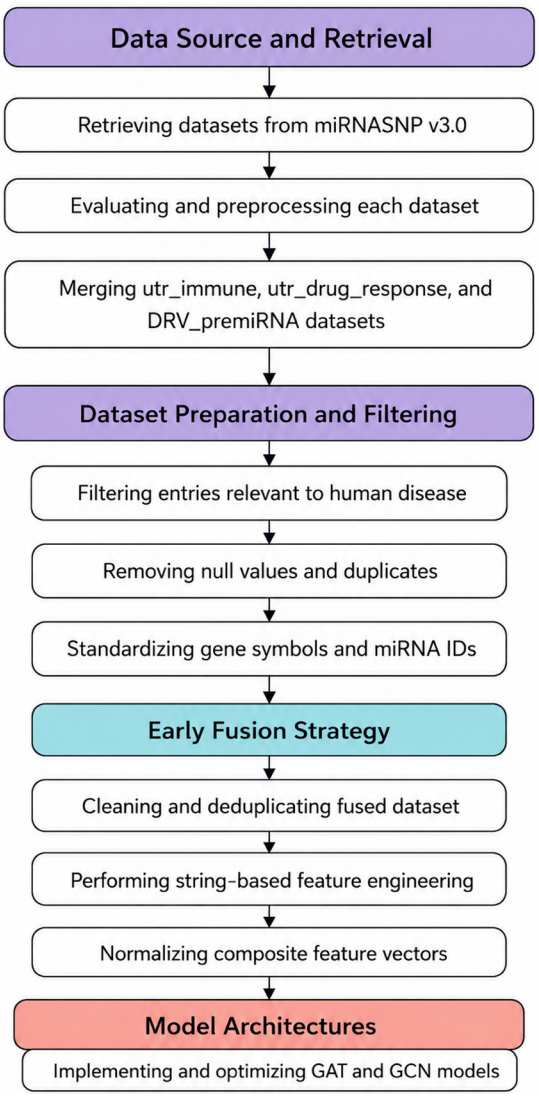

Fig. 1: Workflow of graph-based machine learning analysis of SNP–miRNA–drug interaction candidates using an integrated miRNA-SNP dataset.

Data Preprocessing and Quality Control

Before data preprocessing using Python (Pandas Library) standardized protocol, each dataset was independently assessed for formatting consistency, null values, and duplication. Column headers, including miRNA_ID, Gene_Symbol, SNP_ID,

and Chromosome_Position, were aligned across datasets to enable precise merging and downstream analysis. The analysis did not include records that lacked critical characteristics such as chromosomal coordinates, gene symbols, miRNA names, or SNP identifiers. To guarantee consistency across datasets, chromosome identifiers were standardized by eliminating the "chr" prefixes, and SNP positions were converted to integer formats.

Dataset Merging and Fusion Strategy

We first used key-based merging to combine the utr_immune and utr_drug_response datasets through an outer join using shared keys (miRNA_ID, Gene_Symbol, SNP_ID) to preserve unique features. This allowed us to integrate the SNP information about immune response, drug response, and miRNA biogenesis. The alignment of SNPs that affect both miRNA processing and target site binding was then made possible by integrating the DRV_premiRNA dataset using miRNA_ID as the linking field. Then, using distinct tuples of SNP_ID, miRNA_ID, and Gene_Symbol, we concentrated on deduplication and redundancy removal, locating and removing duplicate entries produced from shared SNP-miRNA mappings. To improve biological interpretation, we finally consolidated functional annotation by combining metadata from various datasets, such as predicted ΔΔG binding energy changes, seed-region mismatches, and drug/immune pathway associations, into enriched annotation fields.

Dataset Preparation and Filtering

The initial combined dataset consisted of 52,129 entries, which were then filtered using biologically significant terms like inflammation, cancer, regeneration, tumor, oncology, and healing to determine their relevance to human disease and tissue remodeling. These keywords were compared to miRNA associations, functional descriptors, and gene annotations. Several quality control procedures were used to guarantee data integrity: duplicate rows, particularly from overlapping annotations or redundant gene-SNP relationships, were eliminated to preserve uniqueness; entries with null values in essential fields were systematically removed; and final feature standardization was completed by harmonizing gene symbols to HGNC-approved nomenclature, normalizing miRNA IDs, and unifying column data types. The final curated and fused dataset is a thorough, highly reliable resource designed for AI-based computational modeling. It facilitates multi-domain analyses of how SNP variations affect the maturation of miRNAs and their downstream regulatory functions in pharmacogenomics, immune modulation, and disease-related regenerative pathways.

Early Fusion Strategy

Data Cleaning, Feature Engineering, and Dimensionality Reduction

The fused dataset underwent systematic preprocessing using the Pandas and Scikit-learn Python libraries to guarantee logical consistency and high-quality input for further analysis. After importing the original fused dataset from a fused_early_dataset.csv CSV file, several procedures were carried out to guarantee consistency and remove duplication. To start, a clean and legitimate subset was created by deleting entries in the early_fusion column that contained null or empty values. To enable consistent string comparisons, the leading and trailing whitespace in each early_fusion entry was removed using the.str.strip function. Additionally, redundant entries in the early_fusion column were found and eliminated using drop_duplicates to ensure a non-redundant data representation. To confirm the integrity of the transformation, a preview was printed after the cleaned dataset was saved as cleaned_fused_dataset.csv.

Feature Engineering via Early Fusion

String-based features from several columns were concatenated into a single composite column called early_fusion to incorporate heterogeneous features into a single vector representation. This method guarantees a linearized representation appropriate for numerical transformation while maintaining semantic dependencies among various modalities. A delimiter-based string merge was used to apply concatenation across the chosen feature columns.

Feature Normalization and Dimensionality Reduction

A high-dimensional feature matrix was then produced by vectorizing the early_fusion strings using tokenization and numerical encoding methods (such as TF-IDF or embedding-based models). The resultant feature vectors were normalized to unit length using L2 normalization before dimensionality

reduction to ensure scale invariance across features and prevent dominance by higher magnitude components. The normalized vectors were then subjected to Principal Component Analysis (PCA) to minimize dimensionality while maintaining maximum variance. Using cumulative variance thresholds (>90%), the principal components were chosen to preserve important signal properties while increasing computational efficiency and reducing noise. The input for ensuing machine learning tasks was this converted low-dimensional feature space.

The fused dataset was randomly partitioned into training and validation subsets using a standard data-splitting procedure implemented in the Scikit-learn Python library. The split preserved the distribution of positive and negative interaction labels to avoid class imbalance between training and validation sets.

Model Architectures

Graph Neural Network Model Architectures

We implemented and optimized two different GNN models, GAT and the GCN, to investigate graph-based feature representation learning on the fused dataset. Both models had similar architectural depths and hyperparameters to guarantee a controlled experimental evaluation. Model performance was evaluated using multiple classification metrics including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC.

Configurations of the Model

The GAT model improves conventional graph learning by adding attention mechanisms that give neighbor contributions learnable weights. On the other hand, the GCN model uses a normalized graph Laplacian-based transformation to aggregate neighbor features. Both architectures used dropout for regularization and two hidden layers. PyTorch Geometric was used to implement all models, and the Adam optimizer was used for training. Table 1 provides a summary of the hyperparameters used for both models.

This standardized setup enables a robust comparison of attention-based versus convolution-based neighborhood aggregation mechanisms in graph learning tasks applied to SNP–miRNA–gene networks. Both models were trained for 200 epochs with early stopping based on validation loss.

Table 1. Architecture and Hyperparameters of GAT and GCN Models

Hyperparameter	GAT	GCN
Input Dimension	256	256
Hidden Layers	[128, 64]	[128, 64]
Attention Heads	4	—
Activation Function	LeakyReLU (α = 0.2)	ReLU
Dropout Rate	0.2	0.2
Optimizer	Adam	Adam
Learning Rate	0.001	0.001
Loss Function	Cross-Entropy	Cross-Entropy
Framework	PyTorch Geometric	PyTorch Geometric

RESULTS

Regression analysis aimed to estimate a continuous interaction score representing the predicted strength of SNP-modulated miRNA effects on drug-response pathways. When the GAT and GCN models are compared, it can be seen that both have moderate prediction error margins, roughly 36–38%. While GAT performs marginally better than GCN in mean absolute error, it performs worse in terms of MAPE. The root mean square error (RMSE), which is approximately 0.524, is almost the same for both models, indicating a steady distribution of residuals. Nevertheless, the R^2 values of each model are close to zero, indicating a poor fit for continuous outcome prediction, and demonstrating the model's incapacity to capture variance. In particular, GCN's marginally better R^2 (0.0002) suggests that it performs similarly to a mean predictor, whereas GAT's negative R^2 raises concerns about overfitting or inadequate generalization. More sophisticated modeling approaches, such as hybrid graph-transformers or knowledge-guided GNN models, are required, as these high error rates and nearly zero R^2 values highlight the complexity and nonlinear character of SNP-driven effects on miRNA function and drug interactions in periodontal regeneration.

A comparison of the training dynamics and predictive performance of GAT and Graph GCN models in analyzing SNP-miRNA-drug interactions for periodontal applications is shown in Figure 2. The left panel reveals GAT's (blue) superior optimization trajectory, starting with a lower initial loss (~2.3 vs. GCN's ~6.5) and stabilizing faster with minimal oscillations after just three epochs, while the right panel shows both models converging

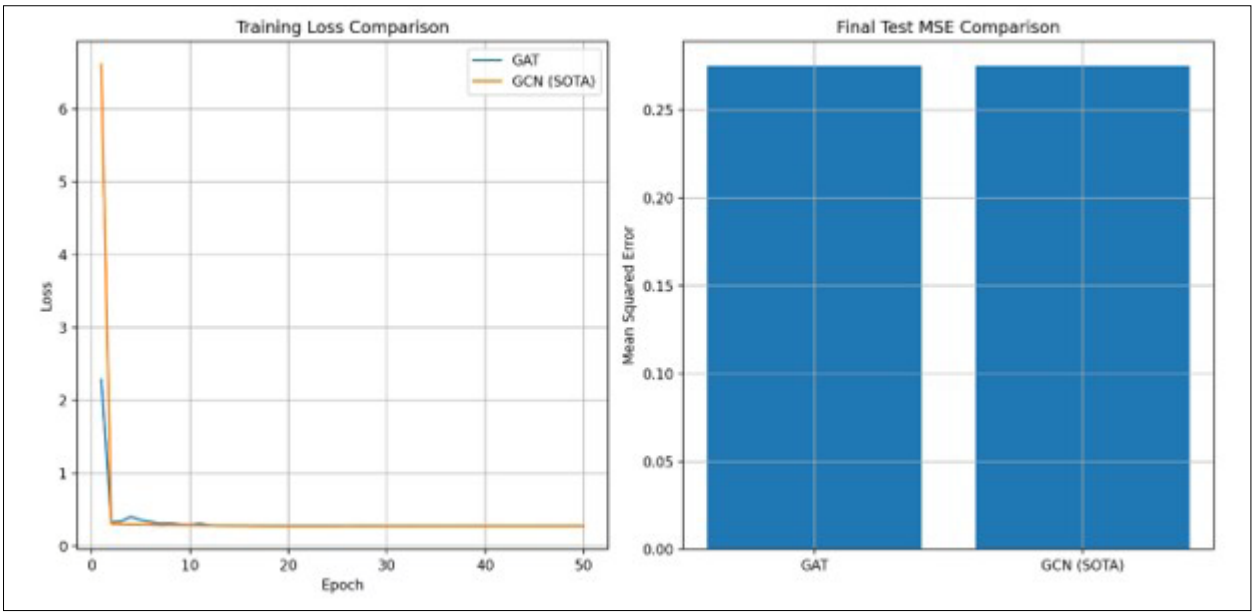

Fig. 2: Training dynamics and regression error comparison between GAT and GCN models on the fused SNP–miRNA dataset. Left panel: training loss trajectories for GAT and GCN across epochs. Right panel: test mean squared error (MSE) comparison after convergence. Although both models reached similar final regression error values, this analysis is intended to describe optimization behavior and comparative model performance on this dataset.

to similar test MSE values (~0.275). However, this apparent parity in regression accuracy masks critical differences: GAT maintains distinct advantages in biological interpretability and prediction variability despite matching GCN's (orange) final error metrics. The ability of the attention mechanism to capture nuanced feature relationships proves particularly valuable in this data-sparse biomedical context, where GCN's constrained expressivity limits its capacity to reveal subtle regulatory patterns. These findings suggest that the GAT architecture may provide advantages for identifying candidate SNP–miRNA–drug interactions in classification tasks. However, the limited regression performance observed in both models indicates that predicting continuous interaction strength remains challenging in this dataset.

The comparative prediction performance and error distributions between GAT and GCN models in analyzing SNP-miRNA-drug interactions relevant to periodontal regeneration are depicted in Figure 3. The top row features a residual plot highlighting GAT's superior performance, with well-dispersed predictions and smaller residuals, while GCN exhibits systematic underfitting, evidenced by large, uniform residuals due to its tendency toward flat predictions. The accompanying kernel density (KDE) plot of errors reveals a subtle right skew in both models, masking GCN's symmetric but biologically implausible error profile.

The middle row's scatter plot demonstrates GAT's ability to capture a broader range of target values with adaptive variability, contrasting sharply with GCN's tightly clustered predictions that confirm its prediction collapse. Supporting boxplots in the bottom row further underscore GCN's instability through higher variance and outlier susceptibility compared to GAT's more consistent predictions. This performance gap stems from GAT's attention mechanism, which dynamically weights node features to capture biologically meaningful signal variance – a capability fundamentally limited in GCN by its rigid aggregation strategy.

Together, these visualizations provide compelling evidence for GAT's superiority in modeling complex regulatory networks, establishing it as the preferred framework for precision applications in periodontal biology.

The 50-epoch training trajectories of GAT and GCN models are illustrated in Figure 4, revealing critical differences in their learning capabilities. The GCN model struggles initially with a significantly higher loss (≈6.5 vs. GAT's ≈2.2), reflecting its inherent challenges in processing heterogeneous fused graph

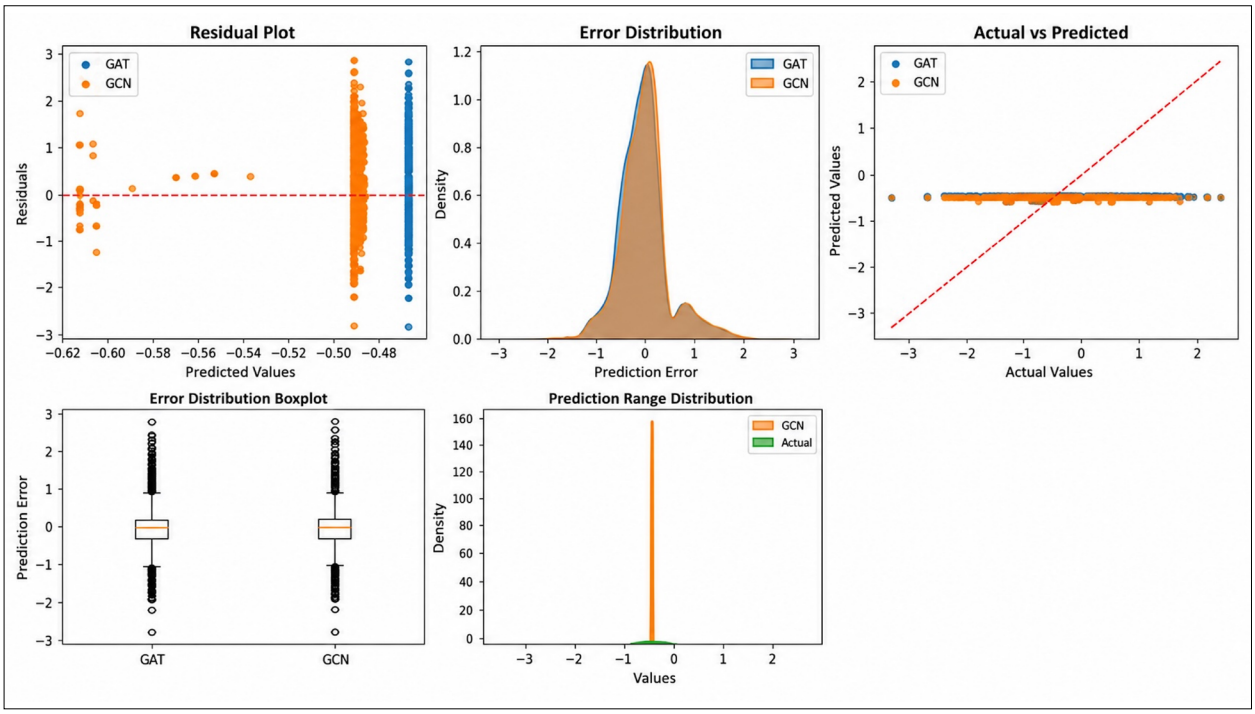

Fig 3: Prediction behavior and error distribution of GAT and GCN in regression analysis.
Panels show residual plots, error density distributions, predicted vs. observed scatter plots, and summary boxplots of errors for both models. The visual diagnostics indicate that GCN predictions tend to collapse toward central values (underfitting behavior), whereas GAT produces a wider spread of predictions. These plots complement the quantitative regression metrics, including near-zero R^2 values, and should be interpreted as model-behavior diagnostics on this dataset rather than as evidence of biological mechanism.

features. While both architectures demonstrate rapid loss reduction during early epochs - emphasizing the importance of initial training phases - their subsequent behaviors diverge markedly. GAT maintains a consistently lower loss trajectory, indicating superior feature adaptation through its attention mechanism, while GCN's flat convergence suggests architectural limitations in capturing complex biological relationships. Notably, neither model shows overfitting beyond epoch 5, but GAT's stable, lower loss plateau demonstrates its enhanced capacity for modeling intricate SNP–miRNA–drug interactions. These dynamics underscore GAT's advantages for periodontal regeneration research: its attention-based weighting excels at parsing high-dimensional biological graphs where GCN's rigid architecture falters, particularly when predicting context-dependent therapeutic targets in inflammatory and regenerative pathways. The figure ultimately positions GAT as the more robust framework for precision periodontology applications requiring nuanced biological inference.

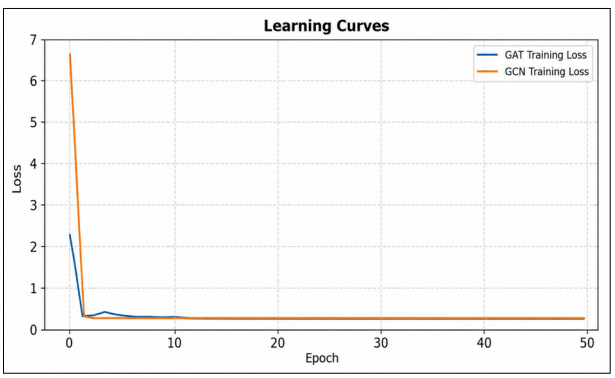

Fig. 4: Comparative training loss trajectories of GAT and GCN across 50 epochs.
The figure illustrates differences in optimization dynamics, with GAT showing a lower loss curve than GCN under matched hyperparameters and training conditions. The plot is presented to summarize training stability and convergence patterns in this comparative evaluation.

The predictive capabilities of GAT and GCN models in forecasting SNP–miRNA–drug interaction scores relevant to periodontal inflammation and tissue regeneration are compared in Figure 5. The left panel reveals GAT's strength through its unimodal, right-skewed distribution of predicted

values, which effectively captures the biological variability inherent in these interactions. In stark contrast, the middle panel exposes GCN's critical limitation - its predictions collapse around a single point, demonstrating severe underfitting that aligns with previous regression metrics. This fundamental shortcoming becomes even more apparent in the right panel's scatter plot, where GAT maintains a broad, biologically plausible range of predictions while GCN's outputs cluster narrowly, reflecting its inability to learn complex patterns. The comparative analysis highlights how GAT's attention mechanism successfully deciphers subtle regulatory effects on key pathways (e.g., Wnt and NF-κB) through adaptive weighting of node features, whereas GCN struggles with the heterogeneity of molecular interactions. While GAT emerges as clearly superior for this application, the figure also suggests that both models could benefit from further refinement to enhance their regression fidelity and biological interpretability for precision periodontology applications.

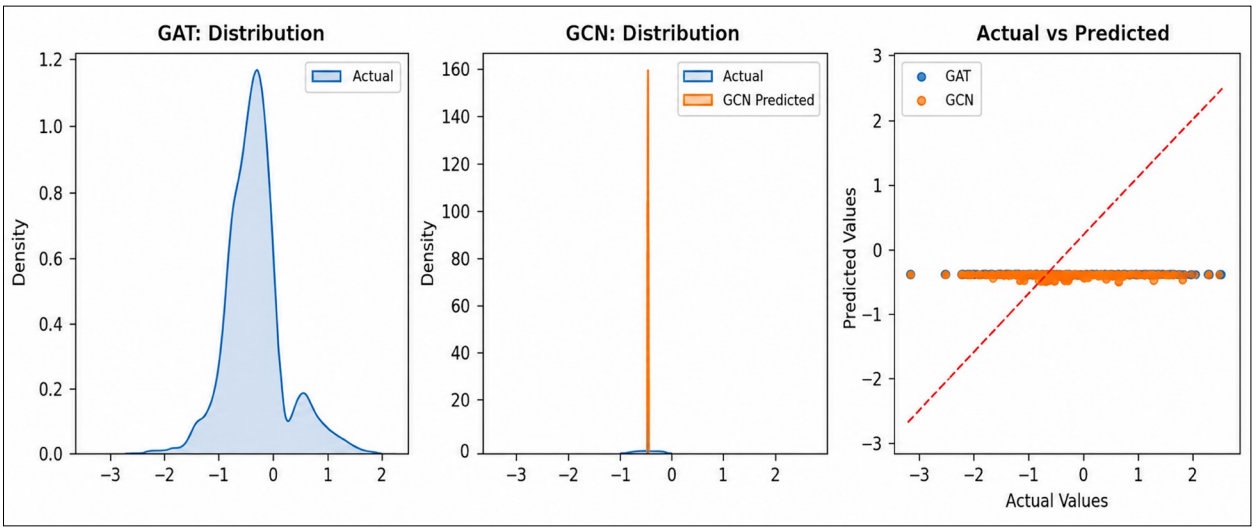

Fig. 5: Distribution of predicted interaction scores and prediction range comparison between GAT and GCN. Left and middle panels show the distributions of predicted scores for GAT and GCN, respectively. Right panel shows predicted vs. observed values, highlighting that GAT outputs span a broader range while GCN outputs are more clustered. These visualizations provide qualitative support for underfitting patterns in GCN and should be interpreted alongside the reported regression metrics (including near-zero R^2) as evidence of limited regression fidelity on this dataset.

Classification Metrics Analysis

Binary classification was used to assess the functional relevance of SNP–miRNA–disease interactions as either neutral or disease-modulating in periodontal disease. While the GAT model had low accuracy and precision (0.499), leading to many false positives, it demonstrated perfect recall (1.0) and an F1-score of 0.666, indicating effective detection of positive cases. Nevertheless, a PR_AUC of 0.750 indicates reasonable performance in ranking relevant interactions. In contrast, the GCN model showed a low PR_AUC (0.461), a very low F1-score (0.118), and poor recall (0.070), underscoring its conservative nature and propensity to overlook important interactions. These results suggest that while GCN's fixed graph propagation rules restrict its ability to capture important biological associations, GAT, with its attention mechanism, better identifies SNP-miRNA–drug relationships relevant to inflammation and regenerative pathways, even in the face of more false positives. The error distributions in both models showed mild right skewness and sporadic high-magnitude underestimations, but their mean error was nearly zero, indicating no significant systematic bias. Accurately predicting miRNA-SNP interactions is crucial for individualized regenerative treatment because SNPs in pre-miRNA and mature miRNA regions can also significantly impact miRNA stability, drug sensitivity, and successful treatment pathways in periodontology. Even though both models showed shortcomings in regression performance, GAT's classification capabilities make it more suitable for giving clinically actionable SNP-miRNA-drug interactions priority in the

complex regulatory environment of periodontal inflammation and regeneration. This suggests that more advanced models and biologically informed priors are required to improve predictive accuracy in precision periodontology.

DISCUSSION

Different strengths and limitations are revealed by comparing the GAT and GCN models. GAT performs better than GCN in the classification domain, showing perfect recall (1.0) and a significantly higher F1-score (0.666 vs. 0.118), albeit at the expense of lower precision (0.499) (Figs 2-5). Because of its attention mechanism, which dynamically weighs node neighbors and captures subtle variations in graph topology—a crucial advantage in sparse biological networks—these findings imply that the GAT model is highly sensitive to detecting SNP–miRNA–drug associations. On the other hand, GCN, which is limited by fixed aggregation rules, exhibits a low F1-score and poor recall (0.070), suggesting an excessively cautious approach that overlooks biologically significant positive interactions. The present results suggest that GAT may offer advantages over GCN in classification-based prioritization of candidate SNP–miRNA–drug interactions. However, the limited regression performance observed in both models indicates that current architectures still face substantial challenges in accurately modeling continuous interaction strength. Furthermore, recent explainable GNN initiatives reflect the need for improved interpretability in SNP–miRNA–-drug modeling. Despite differences in classification performance, both models show similar regression metrics, with mean absolute errors (MAE) and root mean square error (RMSE) values circling 0.36 and 0.524, respectively. However, when predicting continuous interaction strength or influence scores, the R^2 values near zero (or slightly negative) indicate limited explanatory power. This limitation might result from the linearity assumptions of the regression loss function and the high biological heterogeneity inherent in SNP-miRNA-target interactions. These assumptions may fail to capture the intricate nonlinear dependencies involved in pharmacodynamic response and post-transcriptional regulation[9,12].

The error distribution statistics, such as skewness (~0.61) and kurtosis (~2.9), also reveal a slightly right-skewed and leptokurtic pattern, indicating the existence of sporadic large-magnitude prediction errors. Underrepresented SNPs or low-frequency miRNA variants that have disproportionate regulatory effects on inflammation and tissue remodeling, as observed in regenerative periodontology, may be the source of these errors. To enhance predictive performance in overcoming constraints, several future directions are proposed: integrating molecular fingerprints to provide contextual understanding of SNP-miRNA-drug interactions; employing multi-task learning to optimize classification and regression objectives[13–15]. simultaneously, utilizing deeper or hybrid architectures to capture complex dependencies in regulatory landscapes, applying transfer learning to improve model generalizability by fine-tuning on specific datasets, and implementing techniques for model interpretability, including attention visualizations and importance attribution, to derive biological insights about regulatory SNPs and miRNAs in periodontal disease and healing.

An important consideration is that the miRNA SNP v3.0 database is not specific to periodontal disease but represents a broad collection of SNP–miRNA regulatory interactions across multiple biological contexts. Consequently, although our filtering strategy focused on genes associated with inflammation, immune response, and tissue repair, these biological processes are not exclusive to periodontal pathology. The present study should therefore be interpreted as a computational framework for prioritizing candidate interactions that may be relevant to periodontal inflammation and regeneration, rather than as an analysis based on a periodontal-specific molecular dataset.

Notwithstanding the knowledge acquired, this study has several drawbacks[16,17].

First, even after being selected and screened for biological significance, the dataset size is still small enough to train deep learning models, which may restrict generalizability. Second, model performance may be impacted by possible problems with data quality, such as inconsistent annotations, redundant SNP-miRNA mappings, and uneven class distributions. Third, hyperparameter optimization and architecture experimentation were limited due to the models' limited computational resources during training[13,18,19]. Fourth, despite the structural advantages of GNNs, their interpretability is still limited, making it difficult to convert predictions

into molecular biology mechanistic hypotheses. Finally, the continuum of biological interactions, where SNP effects can differ in magnitude, context-dependence, and downstream phenotypic impact—is oversimplified by using binary classification[20–22].

Even though the GAT model shows promise in detecting SNP–miRNA–drug interactions associated with periodontal inflammation and regeneration, substantial improvements in data diversity, model complexity, and interpretability are required to fully utilize graph learning in precision periodontology.

Another limitation is that the dataset was derived from a general miRNA SNP database rather than a periodontal-specific molecular dataset. Although filtering strategies were applied to enrich inflammation- and regeneration-related genes, these biological processes may also be involved in other inflammatory diseases. Future studies should incorporate periodontal-specific molecular datasets, including markers such as MMP-8, MMP-9, and pathways associated with key periodontal pathogens such as *Porphyromonas gingivalis*.

The dataset fusion strategy represents an important methodological contribution of this study, as it integrates heterogeneous SNP–miRNA regulatory information and drug-response annotations into a unified analytical framework. However, the results obtained should be interpreted primarily as a computational proof-of-concept. Because the dataset is derived from a general regulatory database rather than a periodontal-specific molecular resource, the findings cannot be directly extrapolated to all periodontal biological contexts. Future studies incorporating periodontal-specific molecular markers and experimentally validated regulatory interactions will be necessary to strengthen biological interpretability.

Another limitation of this study is the absence of classical machine learning baseline models such as logistic regression or random classifiers. The present work focused primarily on comparing two graph neural network architectures applied to the same fused dataset. Future work should incorporate additional baseline models to better contextualize the predictive performance of graph-based approaches.

CONCLUSIONS

This study compares GAT and GCN to model SNP-modulated miRNA-drug response interactions potentially relevant to inflammatory and regenerative pathways associated with periodontal disease. Our results indicate that the GAT architecture achieved higher sensitivity in classification tasks compared with GCN when identifying candidate SNP–miRNA–drug interactions. However, the low R^2 values observed in regression analyses highlight the difficulty of accurately modeling continuous interaction strength in complex post-transcriptional regulatory systems. Even though both models performed similarly on regression tasks, their low predictive ability (shown by nearly zero R^2 values) highlights the difficulty involved in modeling post-transcriptional regulatory dynamics with existing graph-based architectures. Error statistics and performance metrics analysis provide valuable insights into model behavior provide preliminary evidence that graph neural network architectures can identify patterns within integrated SNP–miRNA–drug datasets despite biological heterogeneity and sparse data. However, the study also points out important drawbacks that limit the translational applicability of existing models, such as dataset scale, interpretability, and architectural depth.

Predictive performance may be greatly improved by incorporating domain-specific features like molecular fingerprints, implementing multi-task learning frameworks, and embracing more expressive and interpretable architectures. These findings highlight the potential of graph-based machine learning approaches for prioritizing candidate regulatory interactions within integrated SNP–miRNA datasets. However, given the general nature of the dataset and the limited predictive performance observed, the results should be interpreted primarily as a computational proof-of-concept. Future studies incorporating periodontal-specific molecular datasets and experimentally validated regulatory interactions will be required to strengthen biological interpretation and translational applicability in periodontology.

CONFLICT OF INTEREST

The authors declare no potential conflicts of interest regarding the research, authorship, and/or publication of this article.

REFERENCES

1. Ramesh A, Varghese SS, Doraiswamy JN, Malaiappan S. Herbs as an antioxidant arsenal for periodontal diseases. J Intercult Ethnopharmacol. 2016 Jan 27;5(1):92-6. https://doi.org/10.5455/jice.20160122065556. PMID: 27069730; PMCID: PMC4805154.
2. Panda S, Sankari M, Satpathy A, Jayakumar D, Mozzati M, Mortellaro C, Gallesio G, Taschieri S, Del Fabbro M. Adjunctive Effect of Autologus Platelet-Rich Fibrin to Barrier Membrane in the Treatment of Periodontal Intrabony Defects. J Craniofac Surg. 2016 May;27(3):691-6. https://doi.org/10.1097/SCS.0000000000002524. PMID: 27046472.
3. Kaarthikeyan G, Jayakumar ND, Padmalatha O, Sheeja V, Sankari M, Anandan B. Analysis of the association between interleukin-1beta (+3954) gene polymorphism and chronic periodontitis in a sample of the south Indian population. Indian J Dent Res. 2009 Jan-Mar;20(1):37-40. https://doi.org/10.4103/0970-9290.49061. PMID: 19336858.
4. Cao W, He J, Feng J, Wu X, Wu T, Wang D, Min C, Niu X, Gao Z, Guo AY, Gong J. miRNASNP-v4: a comprehensive database for miRNA-related SNPs across 17 species. Nucleic Acids Res. 2025 Jan 6;53(D1):D1066-D1074. https://doi.org/10.1093/nar/gkae888. PMID: 39413204; PMCID: PMC11701550.
5. Barenboim M, Zoltick BJ, Guo Y, Weinberger DR. MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. Hum Mutat. 2010 Nov;31(11):1223-32. https://doi.org/10.1002/humu.21349. Epub 2010 Oct 7. PMID: 20809528; PMCID: PMC3001138.
6. Wen X, Zhou J, Fang H, Li J, Wang R, Zeng D, Xie X, Deng Y, Ren L, Liu E. Allele-specific micro-RNA-mediated regulation of ADAM33 in childhood allergic asthma. Respirology. 2025 Feb;30(2):113-123. https://doi.org/10.1111/resp.14846. Epub 2024 Oct 30. PMID: 39478322.
7. Xu J, Geng J, Zhang Q, Fan Y, Qi Z, Xia T. Association of three micro-RNA gene polymorphisms with the risk of cervical cancer: a meta-analysis and systematic review. World J Surg Oncol. 2021 Dec 16;19(1):346. https://doi.org/10.1186/s12957-021-02463-4. PMID: 34911543; PMCID: PMC8675500.
8. Ning Q, Zhao Y, Gao J, Chen C, Li X, Li T, Yin M. AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA-disease associations identification. Brief Bioinform. 2023 Mar 19;24(2):bbad094. https://doi.org/10.1093/bib/bbad094. PMID: 36907654.
9. Fu Y, Yang R, Zhang L, Fu X. HGECDA: A Heterogeneous Graph Embedding Model for CircRNA-Disease Association Prediction. IEEE J Biomed Health Inform. 2023 Oct;27(10):5177-5186. https://doi.org/10.1109/JBHI.2023.3299042. Epub 2023 Oct 5. PMID: 37494154.
10. Yang Y, Sun Y, Li F, Guan B, Liu JX, Shang J. MGCNRF: Prediction of Disease-Related miRNAs Based on Multiple Graph Convolutional Networks and Random Forest. IEEE Trans Neural Netw Learn Syst. 2024 Nov;35(11):15701-15709. https://doi.org/10.1109/TNNLS.2023.3289182. Epub 2024 Oct 29. PMID: 37459265.
11. Dong B, Sun W, Xu D, Wang G, Zhang T. DAEMDA: A Method with Dual-Channel Attention Encoding for miRNA-Disease Association Prediction. Biomolecules. 2023 Oct 12;13(10):1514. https://doi.org/10.3390/biom13101514. PMID: 37892196; PMCID: PMC10604960.
12. Xie GB, Yu JR, Lin ZY, Gu GS, Chen RB, Xu HJ, Liu ZG. Prediction of miRNA-disease associations based on strengthened hypergraph convolutional autoencoder. Comput Biol Chem. 2024 Feb;108:107992. https://doi.org/10.1016/j.compbiolchem.2023.107992. Epub 2023 Nov 27. PMID: 38056378.
13. Shi H, Zhang X, Tang L, Liu L. Heterogeneous graph neural network for lncRNA-disease association prediction. Sci Rep. 2022 Oct 20;12(1):17519. https://doi.org/10.1038/s41598-022-22447-y. PMID: 36266433; PMCID: PMC9585029.
14. Jin C, Shi Z, Lin K, Zhang H. Predicting miRNA-Disease Association Based on Neural Inductive Matrix Completion with Graph Autoencoders and Self-Attention Mechanism. Biomolecules. 2022 Jan 2;12(1):64. https://doi.org/10.3390/biom12010064. PMID: 35053212; PMCID: PMC8774034.
15. Yu L, Ju B, Ren S. HLGNN-MDA: Heuristic Learning Based on Graph Neural Networks for miRNA-Disease Association Prediction. Int J Mol Sci. 2022 Oct 29;23(21):13155. https://doi.org/10.3390/ijms232113155. PMID: 36361945; PMCID: PMC9657597.
16. Zhang G, Li M, Deng H, Xu X, Liu X, Zhang W. SGNNMD: signed graph neural network for predicting deregulation types of miRNA-disease associations. Brief Bioinform. 2022 Jan 17;23(1):bbab464. https://doi.org/10.1093/bib/bbab464. PMID: 34875683.
17. Liu J, Kuang Z, Deng L. GCNPCA: miRNA-Disease Associations Prediction Algorithm Based on Graph Convolutional Neural Networks. IEEE/ACM Trans Comput Biol Bioinform. 2023 Mar-Apr;20(2):1041-1052. https://doi.org/10.1109/TCBB.2022.3203564. Epub 2023 Apr 3. PMID: 36049014.
18. Wang J, Li J, Yue K, Wang L, Ma Y, Li Q. NMCMDA: neural multicategory MiRNA-disease association prediction. Brief Bioinform. 2021 Sep 2;22(5):bbab074. https://doi.org/10.1093/bib/bbab074. PMID: 33778850.
19. Li X, Yang Y, Xu S, Gui Y, Chen J, Xu J. Screening biomarkers for spinal cord injury using weighted gene co-expression network analysis and machine learning. Neural Regen Res. 2024 Dec 1;19(12):2723-2734. https://doi.org/10.4103/1673-5374.391306. Epub 2023 Dec 21. PMID: 38595290; PMCID: PMC11168503.
20. Ding Y, Tian LP, Lei X, Liao B, Wu FX. Variational graph auto-encoders for miRNA-disease association prediction. Methods. 2021 Aug;192:25-34. https://doi.org/10.1016/j.ymeth.2020.08.004. Epub 2020 Aug 13. PMID: 32798654.
21. Jadhav SP. MicroRNAs in microglia: deciphering their role in neurodegenerative diseases. Front Cell Neurosci. 2024 May 15;18:1391537. https://doi.org/10.3389/fncel.2024.1391537. PMID: 38812793; PMCID: PMC11133688.
22. Yan C, Duan G, Li N, Zhang L, Wu FX, Wang J. PDMDA: predicting deep-level miRNA-disease associations with graph neural networks and sequence features. Bioinformatics. 2022 Apr 12;38(8):2226-2234. https://doi.org/10.1093/bioinformatics/btac077. PMID: 35150255.