BMC Bioinformatics

RESEARCH

Open Access

Cross-attention graph neural networks for inferring gene regulatory networks with skewed degree distribution

Jiaqi Xiong[1†], Nan Yin[2*†], Shiyang Liang[3†], Haoyang Li[1], Yingxu Wang[4], Duo Ai[5] and Jingjie Wang[6*]

†Jiaqi Xiong, Nan Yin and Shiyang Liang have contributed equally to this work.

*Correspondence:
Nan Yin
yinnan8911@gmail.com
Jingjie Wang
jingjiefmmu@163.com

Full list of author information is available at the end of the article

Abstract

Background Inferring Gene Regulatory Networks (GRNs) from gene expression data is a pivotal challenge in systems biology. Most existing methods fail to consider the skewed degree distribution of genes, complicating the application of directed graph embedding methods.

Results The Cross-Attention Complex Dual Graph Embedding Model (XATGRN) was proposed to address this issue. It employs a cross-attention mechanism and a dual complex graph embedding approach to manage the skewed degree distribution, ensuring precise prediction of regulatory relationships and their directionality. The model consistently outperforms existing state-of-the-art methods across various datasets.

Conclusions XATGRN provides an effective solution for inferring GRNs with skewed degree distribution, enhancing the understanding of complex gene regulatory mechanisms. The codes and detailed requirements have been released on Github: (https://github.com/kikixiong/XATGRN).

Keywords Gene regulatory network, Direct graph embedding, Cross attention network

Introduction

Cells execute diverse functions by expressing various genes and through the interplay of regulatory relationships among these genes. Bulk sequencing [1] enables the profiling of gene expression within specific tissues. From such gene expression matrices, researchers can extract a Gene Regulator Network (GRN). GRNs are crucial in different biological processes; they can aid in studying developmental biology, unraveling the mechanisms behind various diseases, and identifying new therapeutic targets [2–5].

Although numerous databases have accumulated extensive regulatory relationships, each tissue's GRN reveals its unique regulatory characteristics [6–8]. Hence, it is impractical to validate the specific GRN regulatory network in each tissue solely through wet experiments. Many researchers have proposed computational methods

for bulk sequencing GRN inference. Prior to the pprevalence of deep learning models, researchers proposed various GRN inference methods based on conventional machine learning and statistical approaches, including correlation-based methods [9], Bayesian network-based methods [10, 11], and hybrid methods [12, 13]. In recent years, with the accumulation of sequencing data [14, 15] and the development of deep learning technology, numerous deep learning-based bulk GRN inference methods have emerged. For instance, the CNNGRN [16] model utilizes a convolutional neural network to reconstruct gene regulatory networks from large-scale gene expression data. This CNNGRN model leverages known gene regulatory networks as prior knowledge to capture gene neighborhood information, and incorporates it as network structural features to enhance the predictive power over gene-gene relations. This improvement is particularly effective in inferring gene regulatory networks in real species.

Although effective, CNN-based approaches are not naturally designed for tackling data in non-euclidean space. Hence, more recent models aim to incorporate graph-based methods. GRGNN [17] was the first to introduce Graph Neural Networks (GNNs) [18–24] into GRN research. It transforms GRN inference into a graph classification task. However, the GNN used by GRGNN does not consider the directionality of gene regulation and stipulates that a gene can only be a transcription factor (TF) or a target gene. In reality, many gene nodes in actual GRNs act as both TFs and target genes. Thus, this method can only infer the existence of regulatory relationships between genes, not the precise regulatory direction. DGCGRN [25] is a GRN inference method based on Directed Graph Convolutional Networks [26], specifically designed to handle directed graph-structured data. Compared to GRGNN, DGCGRN can better capture the directionality of gene regulation. DGCGRN employs a local enhancement strategy and dynamic updating strategy, generating enhanced features through Conditional Variational Autoencoders [27] to address the issue of low-degree nodes, and updates edge weights in each iteration to improve predictive performance. Additionally, DGCGRN incorporates sequence features, providing a more comprehensive inference of gene regulatory relationships for real biological data. However, neither GRGNN nor DGCGRN incorporates prior known gene regulatory networks as prior knowledge. DeepFGRN [28] effectively reconstructs large-scale and sparse gene regulatory networks by combining correlation analysis with directed graph embedding techniques. This work incorporates known gene regulatory networks to assist in network construction. It is capable of identifying gene regulatory relationships and discerning the types and directions of gene regulation.

When inferring GRNs using directed graph neural networks, genes are treated as nodes and regulatory relationships between TFs and target genes are treated as edges. Although previous methods such as DeepFGRN have incorporated the directed GNN to represent the directionality of GRNs, they neglect the fact that number of in/out edges can have a significant gap for each node. In particular, some genes may regulate the expression of multiple other genes, thus having a higher out-degree. Conversely, if a gene is regulated by multiple factors, it will have a higher in-degree.Such a phenomenon is known as graph with skewed degree distribution [29, 30]. This challenge is more common in directed graphs, where the separation of in and out neighbours often results in a higher proportion of nodes with a skewed degree distribution compared to undirected graphs. However, existing graph-based bulk GRN inference efforts have not

considered this issue, which will affect their prediction performance. In addition, existing approaches use shallow embedding methods such as CNN to capture the correlation between regulator genes and target genes from the bulk gene expression profiles. However, we argue that a more advanced attention mechanism can better capture the complex relations between two genes compared with shallow embedding methods.

To overcome the aforementioned limitations, we introduce a novel approach named Cross-Attention Complex Dual Graph Attention Network Embedding Model (XATGRN). Our XATGRN model is designed to provide a comprehensive understanding of GRNs by predicting the existence of regulatory relationships and determining their directionality and types. In particular, XATGRN utilizes a cross-attention mechanism to capture the complex interactions reflected in the bulk gene expression profiles of regulator and target genes, thereby enhancing the model's ability to represent these interactions accurately.

Furthermore, our model employs a sophisticated directed graph representation learning method i.e., DUPLEX [29] to encode the gene-gene relations. Such a DUPLEX method consists of a dual Graph Attention encoder for directional neighbour modelling using the generated amplitude and phase embeddings. By leveraging the cross-attention module and the DUPLEX method, our XATGRN can effectively capture the connectivity and directionality of regulatory interactions within the network and alleviate the issue due to skewed degree distribution in GRNs.

The main contributions of this paper are as follows:

- We introduce the XATGRN model, which is capable of predicting the existence, directionality, and type of regulatory relationships in Gene Regulatory Networks (GRNs). This model offers a comprehensive understanding of the intricate mechanisms of gene regulation.
- The model utilizes a cross-attention mechanism to focus on the most informative features within bulk gene expression profiles of regulator and target genes, enhancing the model's representational power.
- By employing the dual complex graph embedding method, our model generates amplitude and phase embeddings that capture both the connectivity and directionality of regulatory interactions, effectively alleviating the issue due to skewed degree distribution in GRNs.
- We conduct extensive experiments on multiple benchmark datasets, demonstrating XATGRN's proficiency in uncovering unseen regulatory mechanisms and potential therapeutic targets for complex diseases.

Methods

Problem definition

In Gene Regulatory Networks (GRNs), regulator genes R interact with target genes T to control cellular processes. These interactions can be activating, when the regulator enhances the target's expression, or repressing, when it decreases the expression. The regulatory relationships are directional, flowing from the regulator gene R to the target gene T. We represent these relationships as directed edges e_{RT}. The goal is to predict the type of regulation r_{RT} between regulator gene R and target gene T, which can be activation, repression, or non-regulated.

Overview of the XATGRN

Our Cross-Attention Complex Dual Graph Embedding Model (XATGRN) is designed to infer the regulation types for Gene Regulatory Networks (GRNs). In particular, our XATGRN can distinguish between the activation type and repression type. Our model operates by treating the GRN inference problem as the link prediction task between regulator genes R and target genes T. As shown in Fig. 1, our model extracts key features from both bulk gene expression data and existing databases that detail prior regulatory associations with regulation types, refining these features through a softmax classifier to predict the regulatory relationships as either activation, repression, or non-regulated interactions.

Initially, the gene expression profiles of regulator-target gene pairs (R, T) are used by the fusion module (Fig. 1a), yielding the fusion embedding vector. This vector encapsulates the gene expression features and the correlation information between the regulator and target genes. Subsequently, our Relation Graph Embedding Module The complex embeddings capture both the connectivity and directionality within the network.

Ultimately, the fusion embedding, along with the complex embeddings of regulator gene R and target gene T, are concatenated to form a comprehensive feature set in the prediction module (Fig. 1c). This aggregated information is then fed into a softmax classifier for predicting the GRN relation.

Fusion module

Our Fusion Module extracts gene expression features from both the regulator gene R and the target gene T. This module captures the interactions between the gene pair (R, T), which are essential for predicting regulatory mechanisms within gene regulatory networks (GRNs), as shown in Fig. 1a.

To address the shortcomings of conventional one-dimensional CNNs used by DeepFGRN [28], we introduce the Fusion Module. This module is based on the Cross-Attention

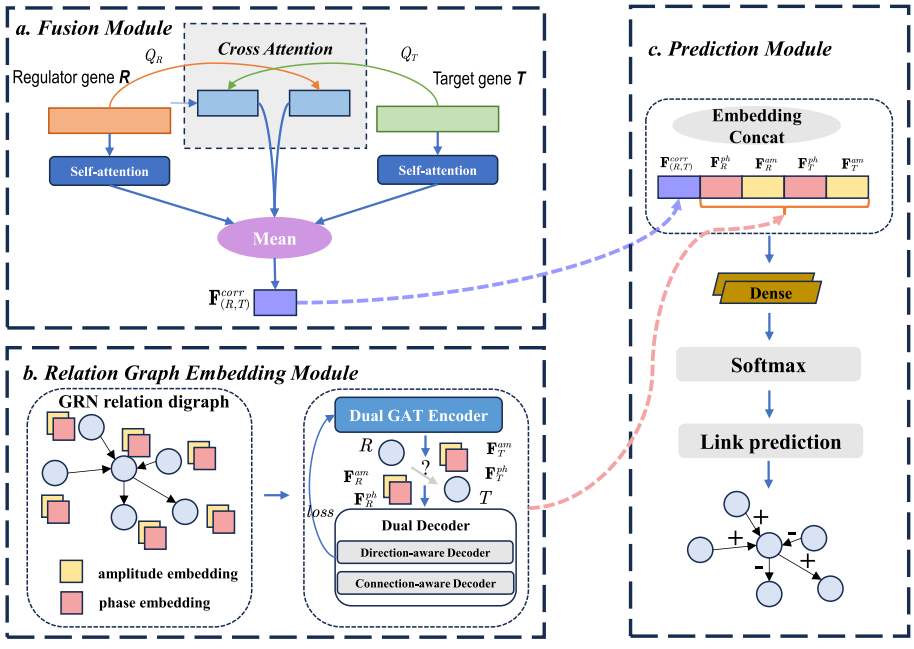

Fig. 1 Overview of the XATGRN method

Network (CAN), inspired by FusionDTI [31]. In particular, CAN enables our model to focus on the most relevant aspects of the gene expressions, which significantly improves its capacity to extract meaningful representations.

The gene expression data for the regulator gene R and the target gene T are processed to generate queries, keys, and values for the cross-attention mechanism, denoted as Y_R and Y_T, respectively. The queries, keys, and values for gene R are represented as Q_R, K_R, V_R and for gene T as Q_T, K_T, V_T. The projection matrices W_q^R, W_k^R, W_v^R and W_q^T, W_k^T, W_v^T map the gene data into the corresponding representations. Specifically, those matrices are defined as follows:

$$Q_R = Y_R W_q^R, \quad K_R = Y_R W_k^R, \quad V_R = Y_R W_v^R, \tag{1}$$

$$Q_T = Y_T W_q^T, \quad K_T = Y_T W_k^T, \quad V_T = Y_T W_v^T. \tag{2}$$

Multi-head self-attention and cross-attention mechanisms are subsequently applied. Notably, each gene retains half of its original self-attention embedding and half of its cross-attention embedding, which allows the model to better handle the intrinsic features of each gene while capturing the complex interactions between them. These embeddings encapsulate the intricate regulatory interactions, thereby enhancing our model's capacity to discern the relationship between the genes.

The embeddings from both the regulator and target genes are then concatenated to form a combined embedding. Such a combined embedding is processed to produce the correlation embedding, which represents the regulatory relationship between the gene pair (R, T).

The fusion module and the subsequent steps are defined by the following equations:

$$R^* = \frac{1}{2}\left[MHA\left(Q_R, K_R, V_R\right) + MHA\left(Q_T, K_R, V_R\right)\right], \tag{3}$$

$$T^* = \frac{1}{2}\left[MHA\left(Q_T, K_T, V_T\right) + MHA\left(Q_R, K_T, V_T\right)\right], \tag{4}$$

$$\mathrm{F}_{(R,T)}^{fusion} = \mathrm{Concat}\left(\left(\mathrm{MeanPool}(R^*), \mathrm{MeanPool}(T^*)\right), 1\right), \tag{5}$$

where the embeddings R^* and T^* represent the enhanced representations of the regulator and target genes, respectively. Specifically, R^* integrates information from both the regulator's self-attention and the cross-attention with the target, while T^* integrates information from both the target's self-attention and the cross-attention with the regulator. The mean pooling operation is denoted by $\mathrm{MeanPool}$, and the concatenation operation is denoted by Concat. The final fusion embedding $\mathrm{F}_{(R,T)}^{fusion}$ represents the regulatory relationship between the regulator and target genes.

Relation graph embedding module

The Relation Graph Embedding Module addresses the complexity of representing nodes within Gene Regulatory Networks (GRNs). It handles the challenges posed by high-dimensional, sparse, and directed interactions. Specifically, it leverages the skewed degree of genes, which is crucial for differentiating between regulator and target nodes in GRNs.

To effectively embed the nodes in a GRN, we adopt the Complex Dual Graph Embedding approach from the DUPLEX framework [29]. As shown in Fig. 1b, this approach generates amplitude and phase embeddings for both regulator and target genes, which encode both the connectivity and directionality of the regulatory interactions.

We model a directed graph (digraph) $G = (V, E)$, where V represents the nodes and E represents the directed edges. Each edge $(R, T) \in E$ symbolizes a regulatory link from the regulator gene R to the target gene T. Here, *R* and *T* are specific instances of genes *u* in the graph, where *R* acts as a regulator and *T* as a target in the regulatory relationship. Our objective is to map each gene u to a d-dimensional vector $\mathrm{x}_u \in \mathbb{C}^{d\times 1}$.

To represent the directionality and connectivity of edges in GRN, our XATGRN leverage the Hermitian Adjacency Matrix (HAM). This approach is particularly effective in addressing the challenge of asymmetric digraphs in GRNs. We use H to denote the HAM, which is defined in polar form as:

$$H = A_s \odot \exp\left(i\frac{\pi}{2}\Theta\right), \tag{6}$$

where i is the imaginary unit and $\odot$ represents the Hadamard product. The symmetric binary matrix A_s is defined as:

$$A_s(u,v) = \begin{cases} 1 & \text{if } (u,v) \in E \text{ or } (v,u) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The antisymmetric matrix Θ contains elements from the set $\{-1, 0, 1\}$ defined as:

$$\Theta(u,v) = \begin{cases} 1 & \text{if } (u,v) \in E, \\ -1 & \text{if } (v,u) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Hence, the HAM's entries $H(u, v)$, taking values in the set $\{i, -i, 1, 0\}$, effectively capture the relationships in GRNs, representing four different types of status between *u* and *v* including forward, reverse, bidirectional interactions and non-existing. This representation is particularly suited to GRNs, where regulatory interactions can be both directional and varied in nature. In contrast, the traditional asymmetric adjacency matrix A necessitates separate entries for $A(u, v)$ and $A(v, u)$, each restricted to $\{0, 1\}$, to encode the same diversity of relationships. The HAM's ability to integrate directionality and connectivity in a single, symmetric matrix offers a more comprehensive and efficient representation, aligning perfectly with the intricate patterns observed in GRNs.

Furthermore, the matrix decomposition $H = X^\top \overline{X}$, where X is the node embedding matrix, represents the inner product between X and its complex conjugate $\overline{X}$. This decomposition facilitates the expression of the node embedding x_u in polar form. Specifically, the embedding is written as follows:

$$\mathrm{x}_u = a_u \odot \exp\left(i\frac{\pi}{2}\theta_u\right), \tag{9}$$

$$\overline{\mathrm{x}}_u = a_u \odot \exp\left(-i\frac{\pi}{2}\theta_u\right). \tag{10}$$

where a_u represents the amplitude and θ_u the phase of the embedding x_u. These complex conjugate embeddings x_u and $\overline{\mathrm{x}}_u$ are interpreted as the representations of the

regulator and target roles of gene node u, respectively. This joint embedding strategy, in contrast to using separate embeddings for the regulator and target, enables co-optimized learning from both the incoming and outgoing edges of the node u. Such an approach effectively addresses the challenge of the imbalance between in-degrees and out-degrees, which can significantly affect the embedding quality of nodes within Gene Regulatory Networks (GRNs).

Dual GAT Encoder

Based on HAM, we will introduce a dual encoder architecture that comprises an amplitude encoder, a phase encoder, and a fusion layer. The encoders refine node embeddings by aggregating information from both incoming and outgoing edges of node$u$, effectively addressing the issue of skewed degree distribution.

The amplitude encoder employs GAT to aggregate information from both incoming and outgoing edges for each node. This process captures the node's overall connectivity, ensuring that even nodes with varying in-degrees and out-degrees are embedded with high quality in the context of the network's topology:

$$\mathrm{a}'_u = \mathrm{ReLU}\left(\sum_{v\in\mathcal{N}(u)} \alpha_{uv}^{(am)} \cdot \mathrm{a}_v\right), \tag{11}$$

where $\mathcal{N}(u)$ denotes the set of neighboring nodes connected to node *u* via either incoming or outgoing edges, $\alpha_{uv}^{(am)}$ is the attention coefficient for amplitude embedding, and a'_u represents the updated amplitude embedding for node u.

The phase encoder captures the directionality of regulatory relationships by distinguishing between nodes acting as regulators and targets. The phase embedding is updated similarly using a direction-sensitive attention mechanism:

$$\theta'_u = \mathrm{ReLU}\left(\sum_{v\in\mathcal{N}_{\mathrm{in}}(u)} \alpha_{uv}^{(ph)} \cdot \theta_v - \sum_{v\in\mathcal{N}_{\mathrm{out}}(u)} \alpha_{uv}^{(ph)} \cdot \theta_v\right), \tag{12}$$

where $\mathcal{N}_{\mathrm{in}}(u)$ and $\mathcal{N}_{\mathrm{out}}(u)$ denote the sets of in-neighbors and out-neighbors of node *u*, respectively, and $\alpha_{uv}^{(ph)}$ is the attention coefficient for phase embeddings.

The refined embeddings θ'_u and a'_u from each layer become the input features for the next graph attention layer, allowing hierarchical abstraction of regulatory patterns.

The fusion layer is a critical component that combines the amplitude and phase embeddings, which carry distinct yet complementary information. This layer ensures that the embeddings from both encoders are effectively integrated to capture the comprehensive regulatory interactions within the GRN. The fusion process is designed to balance the contributions from both the amplitude and phase embeddings, thereby enhancing the model's ability to represent complex gene interactions accurately.

The fusion layer operates by combining the information from the amplitude and phase embeddings at each layer of the encoder. This is achieved through differentiated aggregation mechanisms where amplitude fusion preserves undirected connectivity while phase fusion maintains directional relationships: For amplitude embeddings:

$$\mathrm{F}_u^{am} = \mathrm{ReLU}\left(\sum_{v\in\mathcal{N}(u)} \alpha_{uv}^{(am)} \cdot \mathrm{a}_v + \sum_{v\in\mathcal{N}(u)} \alpha_{uv}^{(ph)} \cdot \theta_v\right), \tag{13}$$

where F_u^{am} represents the updated amplitude embedding for node u, $\alpha_{uv}^{(am)}$ and $\alpha_{uv}^{(ph)}$ are the attention coefficients for amplitude and phase embeddings, respectively.

For the phase embeddings, it applies a directional aggregation operator for fusion of the phase embedding:

$$\mathrm{F}_u^{ph} = \mathrm{ReLU}\left(\oplus\{\theta_v\} + \oplus\{\mathrm{a}_v\}\right) \tag{14}$$

where F_u^{ph} represents the updated phase embedding for node u and the directional aggregation operator $\oplus$ is defined as:

$$\oplus\{x_v\} \triangleq \sum_{v\in\mathcal{N}_{\mathrm{in}}(u)} \alpha_{uv}^{(ph)} \cdot x_v - \sum_{v\in\mathcal{N}_{\mathrm{out}}(u)} \alpha_{uv}^{(ph)} \cdot x_v \tag{15}$$

These updated embeddings, a'_u and θ'_u, are then propagated to the subsequent layers in their respective encoders. The fusion layer ensures that the information from both the amplitude and phase embeddings is effectively utilized in the following steps.

Dual Decoders and Loss Functions

After obtaining the embedding features for each gene node, we introduce 2 parameter-free decoders to reconstruct the Hermitian Adjacency Matrix (HAM). These decoders are designed to ensure that the embeddings capture both the connectivity and directionality of the regulatory interactions within the GRN.

The HAM reconstruction process explicitly preserves four edge types in $\mathcal{R} = \{i, -i, 1, 0\}$, corresponding to forward, reverse, bidirectional, and non-existent regulatory relationships. Our dual decoding strategy separates directional semantics from topological connectivity through specialized modules:

The direction-aware decoder aims to reconstruct the directionality of regulatory interactions. This task is formulated as a classification problem in GRN, where each edge (u, v) is assigned probabilities that correspond to its edge type (forward, reverse, bidirectional, or no edge). The predicted edge type is determined by the minimum distance between the estimated matrix element $\hat{H}(u, v)$ and the possible edge types *r*:

$$\text{pred. type} = \underset{r}{\operatorname{argmin}} \left(\mathrm{Dist}(\hat{H}(u, v), r)\right), \quad \forall r \in \mathcal{R}, \tag{16}$$

where $\mathrm{Dist}(\hat{H}(u, v), r)$ is the distance between $\hat{H}(u, v)$ and *r*.

The probabilities $P(\hat{H}(u, v) = r)$ are calculated as follows:

$$P(\hat{H}(u, v) = r) = \frac{\exp(-|\mathrm{x}_u^\top \overline{\mathrm{x}}_v - r|)}{\sum_{r'\in\mathcal{R}} \exp(-|\mathrm{x}_u^\top \overline{\mathrm{x}}_v - r'|)}, \quad \forall r \in \mathcal{R}. \tag{17}$$

The self-supervised direction-aware loss is then defined as:

$$\mathcal{L}_d = -\sum_{r\in\mathcal{R}} \sum_{H(u,v)=r} \log P(\hat{H}(u, v) = r). \tag{18}$$

The connection-aware decoder is designed to reconstruct the binary presence of connections between genes, which reconstruct A_s—the amplitude component of the HAM defined in Eq. 6. It models the connection probability for an edge (u, v) as:

$$P(\hat{A}_s(u,v)=1)=\sigma(a_u^\top a_v), \tag{19}$$

where σ is the sigmoid function and $\hat{A}_s$ is the estimated amplitude component of the HAM representing the probability that a connection exists between genes *u* and *v*. The connection-aware loss function $\mathcal{L}_c$ adheres to the same negative log-likelihood minimization principle as the direction-aware loss:

$$\mathcal{L}_c=-\sum_{u,v}\log P(\hat{A}_s(u,v)=Atwosta_s(u,v)). \tag{20}$$

Total Loss

The total loss function combines the direction-aware and connection-aware losses to ensure that the model learns both the connectivity and directionality of the regulatory interactions. The total loss is given by:

$$\mathcal{L}_{total}=\mathcal{L}_d+\lambda_k\mathcal{L}_c, \tag{21}$$

where the connection-aware weight $\lambda_k=\lambda_0 q^k$ dynamically evolves with training epoch *k*, starting from initial value λ_0 and decaying via decay rate $q<1$.

By leveraging the fusion module and the Relation Graph Embedding Module, our XATGRN can effectively capture the connectivity and directionality of regulatory interactions within the network and alleviate the issue due to skewed degree distribution in GRNs.

Prediction module

The prediction module in our model is designed to leverage the embeddings generated by the Fusion Module and the Relation Graph Embedding Module to make accurate predictions about the regulatory relationships within the Gene Regulatory Network (GRN). As shown in Fig. 1c, this module integrates the complex embeddings of genes and employs a series of neural network layers to classify the interactions between gene pairs (*R*, *T*).

For each sample, the module processes the feature and label data for a gene pair (*R*, *T*), where *R* is the regulator gene and *T* is the target gene. The features from the Fusion Module are denoted as $\mathrm{F}_{(R,T)}^{fusion}$, while the features from the Relation Graph Embedding Module includes the amplitude and phase embeddings for both genes: F_R^{am}, F_T^{am}, F_R^{ph}, and F_T^{ph}.

These embeddings are concatenated to form a comprehensive feature vector:

$$\mathrm{F}=\mathrm{concat}(\mathrm{F}_{(R,T)}^{fusion},\mathrm{F}_R^{am},\mathrm{F}_T^{am},\mathrm{F}_R^{ph},\mathrm{F}_T^{ph}). \tag{22}$$

The concatenated feature vector F is first processed through a 1-dimensional convolutional layer with batch normalization (BN) and max pooling, followed by ReLU activation to extract and refine spatial features:

$$x_1=\mathrm{ReLU}(\mathrm{MaxPool}(\mathrm{BN}(\mathrm{Conv1D}(\mathrm{F})))), \tag{23}$$

where the output is x_1, and the vector x_1 is pass to the global average pooling layer:

$$x_2 = \text{GlobalAvgPool}(x_1), \tag{24}$$

where the resulting vector is denoted as x_2.This step condenses the feature map into a fixed-size vector that captures the essential information for classification.

Subsequently, the flattened vector x_2 is passed through two fully connected layers(FC_1 and FC_2) with dropout for regularization, where a dropout rate of $p = 0.3$ is applied:

$$x_3 = \text{FC}_2(\text{Dropout}(\text{ReLU}(\text{FC}_1(x_2)), p = 0.3)), \tag{25}$$

where x_3 represents the output of the second fully connected layer, and x_3 is then passed through a softmax function to produce the final classification probabilities over *C* classes:

$$\text{output} = \text{softmax}(x_3) \tag{26}$$

To address class imbalance, we employ a weighted cross-entropy loss function $\mathcal{L}$. The weights w_c for each class c are inversely proportional to their frequency in the dataset:

$$w_c = \frac{N_{\text{total}}}{N_c}, \tag{27}$$

where N_{total} is the total number of samples in the dataset, and N_c is the number of samples in class c.

The loss function $\mathcal{L}$ is then defined as:

$$\mathcal{L} = -\sum_{c=1}^{C} w_c \cdot y_c \cdot \log(\hat{y}_c), \tag{28}$$

where y_c is the true label for class c, and $\hat{y}_c$ is the predicted probability for class c.

Our model optimizes this loss function using the Adam optimizer, with an exponential learning rate scheduler to ensure stable and efficient convergence.

Model inference

The XATGRN inference workflow for new datasets comprises three key phases: Given *N* genes $\{g_1, ..., g_N\}$, we enumerate all possible directed regulatory pairs (g_i, g_j) with $i \neq j$. This produces $N(N-1)$ candidate relationships that form the prediction space.

Each candidate pair is processed through the trained XATGRN model to estimate regulatory probability P_{ij}. The dual-module architecture ensures simultaneous evaluation of both connection existence ($\mathcal{L}_c$) and interaction type ($\mathcal{L}_d$). A probability threshold τ is applied to generate the final adjacency matrix:

$$A_{ij} = \mathbb{I}(P_{ij} \geq \tau) \tag{29}$$

where $\mathbb{I}$ is the indicator function. The default $\tau = 0.8$ was established through validation on holdout biological datasets.

Table 1 Summary of benchmark datasets statistics

Species	Datasets	numG	dimG	numA	numR
DREAM5	Network1	1643	805	2236	1776
E.coli	Cold stress	2205	24	2070	2034
E.coli	Heat stress	2205	24	2070	2034
E.coli	Lactose stress	2205	12	2070	2034
E.coli	Oxidative stress	2205	33	2070	2034
Human	COVID-19	2478	42	6452	1888
Human	Breast cancer	2478	24	6452	1888
Human	Liver cancer	2478	10	6452	1888
Human	Lung cancer	2478	130	6452	1888

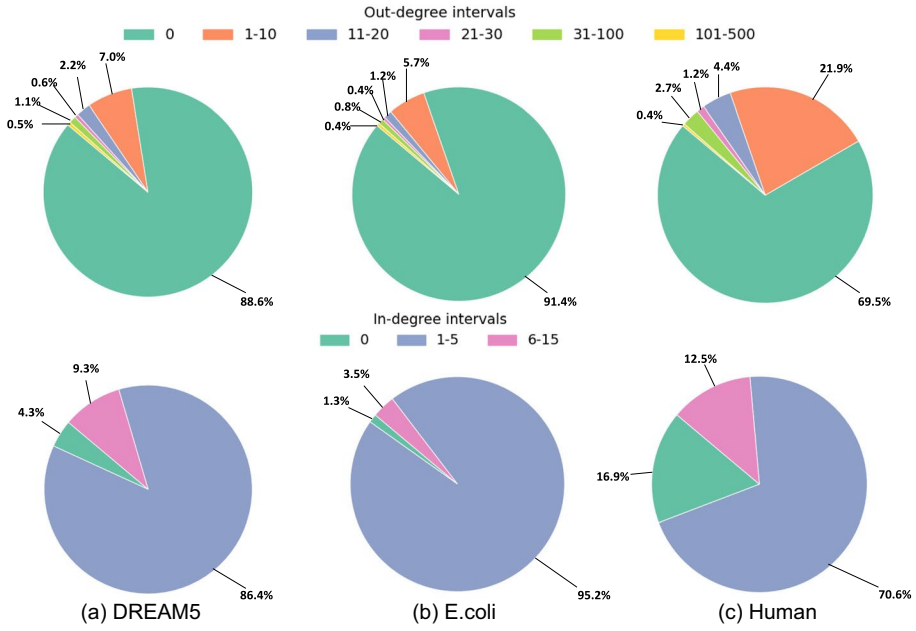

Fig. 2 Visualization of the in-degree and out-degree distribution in DREAM5, Human and E.coli datasets

Dataset and experiment

Datasets

To examine the performance of our XATGRN model, we use the FGRN benchmark, which is introduced in the DeepFGRN [28]. The benchmark collects bulk gene expression data and prior regulatory gene pairs with regulation types across 9 distinct datasets. These datasets include the DREAM5 challenge network1, E.coli under 4 different stress conditions (cold, heat, lactose, and oxidative stress), and 4 types of human diseases (Covid-19, breast cancer, lung cancer, and liver cancer). These datasets are particularly relevant for studying disease mechanisms from a gene regulatory perspective due to their significant implications in understanding disease pathology.

The detailed statistics of these datasets are summarized in Table 1. The table includes the number of genes (numG), the dimension of gene expression data (dimG), and the number of regulatory associations for known activation types (numA) and repression types (numR).

To illustrate the skewed degree distribution in these datasets, we provide a visualization of the in-degree and out-degree distributions for the DREAM5, Human, and E.coli datasets (Fig. 2). This visualization highlights the significant variation in the number of

incoming and outgoing edges for different genes, a characteristic that poses a challenge for traditional graph embedding methods.

Experiment setting

To better evaluate the performance of the model, we adapt 10 times 5-fold cross-validation. For each fold, we calculate the mean of Area Under the Receiver Operating Characteristic curve (AUC), precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to accurately predict both the presence and type of regulatory interactions within gene networks. For the Fusion module,we use PyTorch for implementation. We employed the Adam optimizer with a learning rate of 0.001 to update the model parameters during training. For the Relation Graph Embedding Module, we use the DGL library and PyTorch for implementation, with a learning rate of 1e-3. We choose the optimal hidden dimension in {128,256} and the dropout rate to 0.5. We select the optimal hyperparameter for initial loss weight λ_0 in 0.1,0.3 and the decay rate q in 0,1e-4,1e-2. To demonstrate the effectiveness of our model, we compare XATGRN with state-of-the-art GRN inference models, including CNNGRN [32], DGCGRN [33], and DeepFGRN [3].We also employ the advanced Graph Neural Network Model, including Graph Convolutional Networks(GCN)[34], Isomorphism Network (GIN)[35]and Graphormer[36] to compare with our Relation Graph Embedding Module.

Experiment result

As shown in Table 2 and Fig. 3, our XATGRN model consistently outperforms state-of-the-art models and GNN baselines across all datasets. In particular, CNNGRN focuses on extracting and reconstructing features from bulk gene expression data but disregards the original structure of the gene regulatory network (GRN). While DGCGRN and DeepFGRN construct directed graph embeddings for GRN inference, they fail to address the challenge posed by skewed degree distributions, which can lead to suboptimal performance, particularly for genes with significant disparities between in-degree and out-degree variances.

Our XATGRN achieves the highest AUC, recall, F1-score, and precision across the DREAM5 network1 and all four E.coli datasets. These results demonstrate that XATGRN effectively captures the complex regulatory interactions within gene networks and accurately predicts both the presence and types of regulatory relationships. Notably, XATGRN's robust performance highlights its ability to handle the skewed degree challenge more effectively compared to DeepFGRN and DGCGRN.

Notably, when comparing with advanced GNN architectures, XATGRN demonstrates significant advantages over both GCN and GIN variants. While GCN+CAN achieves competitive AUC scores (e.g., 0.9397 in DREAM5, outperforming previous SOTA models like DeepFGRN), Graphormer+CAN reaches 94.42% AUC on DREAM5 which is closest to our method, our model shows improvements across all datasets. The GIN+CAN variant underperforms consistently due to its isomorphism-focused design being unsuitable for directional GRN structures. This validates that our Relation Graph Embedding Module effectively captures regulatory patterns.

However, it is worth noting that in certain cases, such as the human COVID-19, breast cancer, and lung cancer datasets, the recall of XATGRN is slightly lower than that of

Table 2 Average and standard deviation results over 10 times of 5-fold cross-validation for model performance comparisons across 9 datasets (DREAM5 network1, E.coli cold stress, E.coli heat stress, E.coli lactose stress, E.coli oxidative stress, Human COVID-19, Human Breast cancer, Human Liver cancer, Human Lung cancer.)

Model	DREAM5 network1				E.coli cold stress				E.coli heat stress			
	AUC	Recall	F1	Precision	AUC	Recall	F1	Precision	AUC	Recall	F1	Precision
CNNGRN	76.82±0.45	63.04±0.82	59.32±0.76	55.49±0.68	84.43±0.38	59.41±1.12	55.38±1.05	51.83±0.98	84.94±0.42	58.11±1.08	54.88±1.02	52.36±0.95
DGCGRN	87.51±0.35	77.73±0.78	77.86±0.82	78.23±0.75	81.21±0.48	71.98±1.25	72.09±1.18	72.44±1.32	81.40±0.52	72.09±1.28	72.19±1.35	72.50±1.42
DeepFGRN	92.55±0.14	79.93±0.15	80.00±0.17	80.39±0.25	91.96±0.26	79.35±0.56	79.43±0.57	79.75±0.52	91.80±0.22	78.96±0.37	79.07±0.38	79.38±0.40
GCN+CAN	93.97±0.11	82.72±0.49	82.82±0.51	83.14±0.50	86.01±0.39	72.00±0.55	71.92±0.62	72.28±0.62	87.59±0.32	75.35±0.23	75.31±0.26	75.55±0.30
GIN+CAN	91.93±0.29	78.36±0.56	78.32±0.54	78.58±0.42	79.23±0.32	61.17±0.47	60.88±0.57	61.01±0.53	80.46±0.25	63.23±0.44	62.92±0.43	63.22±0.45
Graphormer+CAN	94.42±0.14	83.67±0.33	83.85±0.33	84.23±0.32	89.10±0.30	77.29±0.54	77.40±0.54	78.18±0.57	89.89±0.26	78.81±0.38	78.96±0.40	79.50±0.38
XATGRN	**94.47**±0.34	**83.92**±0.39	**84.10**±0.38	**84.48**±0.39	**92.17**±0.17	**81.67**±0.32	**81.78**±0.32	**82.19**±0.35	**92.16**±0.19	**81.63**±0.42	**81.74**±0.43	**82.08**±0.41
	E.coli lactose stress				E.coli oxidative stress				Human COVID-19			
	AUC	Recall	F1	Precision	AUC	Recall	F1	Precision	AUC	Recall	F1	Precision
CNNGRN	83.86±0.52	61.29±0.95	59.62±0.88	56.03±0.82	84.02±0.48	59.83±1.15	56.68±1.08	53.81±0.95	86.29±0.45	74.18±1.22	71.34±1.18	68.57±1.08
DGCGRN	82.40±0.62	72.99±1.35	73.08±1.42	73.46±1.38	82.03±0.58	72.69±1.48	72.80±1.52	73.13±1.45	81.46±0.65	70.83±1.55	71.75±1.62	73.30±1.58
DeepFGRN	**92.10**±0.23	79.62±0.49	79.69±0.49	79.99±0.48	91.75±0.34	79.13±0.59	79.23±0.59	79.52±0.59	90.02±0.18	**79.43**±0.25	78.49±0.26	78.33±0.26
GCN+CAN	87.82±0.27	74.82±0.30	74.75±0.27	75.11±0.36	89.12±0.29	76.75±0.38	76.70±0.40	76.88±0.33	83.29±0.39	67.77±0.50	69.73±0.51	73.17±0.68
GIN+CAN	80.87±0.31	63.42±0.45	62.93±0.43	63.23±0.44	82.56±0.29	66.31±0.47	65.98±0.56	66.17±0.49	78.06±0.45	64.77±0.48	64.76±0.58	65.42±0.79
Graphormer+CAN	89.13±0.26	77.33±0.33	77.48±0.33	78.00±0.32	89.87±0.28	78.90±0.38	79.03±0.37	79.68±0.38	88.22±0.24	75.65±0.32	77.25±0.24	79.78±0.30
XATGRN	92.08±0.28	**81.61**±0.34	**81.71**±0.36	**82.09**±0.30	**92.28**±0.26	**81.80**±0.39	**81.91**±0.40	**82.33**±0.40	**91.05**±0.16	79.16±0.19	**80.18**±0.18	**81.92**±0.26
	Human breast cancer				Human liver cancer				Human lung cancer			
	AUC	Recall	F1	Precision	AUC	Recall	F1	Precision	AUC	Recall	F1	Precision
CNNGRN	86.19±0.55	74.86±1.25	68.11±1.08	65.32±0.95	85.94±0.62	78.22±1.35	74.68±1.22	71.83±1.15	84.38±0.58	72.03±1.42	66.48±1.28	62.49±1.18
DGCGRN	81.37±0.65	70.87±1.55	71.68±1.62	72.97±1.58	81.98±0.72	71.60±1.68	72.49±1.72	73.69±1.65	81.99±0.68	71.83±1.75	72.62±1.82	73.90±1.78
DeepFGRN	90.13±0.23	**79.69**±0.21	78.52±0.25	78.39±0.27	89.66±0.16	79.84±0.16	78.75±0.18	78.53±0.20	90.49±0.00	**80.47**±0.00	79.59±0.00	79.38±0.00
GCN+CAN	83.49±0.30	68.85±0.47	70.43±0.38	73.09±0.27	81.08±0.37	65.88±0.56	67.72±0.54	70.80±0.53	84.53±0.26	69.21±0.44	71.19±0.32	74.64±0.25
GIN+CAN	78.29±0.56	64.29±0.61	64.49±0.64	65.47±0.86	76.58±0.71	64.97±0.43	63.72±0.53	63.41±0.09	81.48±0.64	68.30±0.40	68.74±0.73	69.84±0.69
Graphormer+CAN	88.77±0.34	75.72±0.29	77.17±0.30	79.46±0.48	88.94±0.22	76.06±0.32	77.89±0.18	80.81±0.28	89.67±0.36	76.73±0.18	78.37±0.19	80.98±0.36
XATGRN	**90.86**±0.17	79.23±0.03	**80.11**±0.25	**81.53**±0.18	**91.50**±0.14	**80.02**±0.23	**80.90**±0.18	**82.36**±0.21	**91.59**±0.24	80.28±0.31	**81.14**±0.28	**82.57**±0.30

Bold text indicates the highest performance metrics (e.g., AUC, F1-score) achieved by different models across the datasets

DeepFGRN. This observation indicates that it may be necessary for such complex datasets to strike a better balance between addressing the skewed degree problem and optimizing source-target embeddings. In conclusion, our XATGRN model constantly and consistently outperforms competitive baselines for inferring GRNs from different types of gene expression data.

Ablation study

To evaluate the contribution of each module in the XATGRN model, we conducted an ablation study by systematically varying the inclusion of key components. The study was performed on three datasets: DREAM5 Network1, E.coli cold stress, and the Human COVID-19 dataset. We compared three different setups to assess the impact of the Fusion Module and the Relation Graph Embedding Module on the model's performance. Specifically, these setups included a full XATGRN model with both modules, a model with only the Relation Graph Embedding Module, and a model with only the Fusion Module. Detailed results are visualized in Fig. 4.

The full XATGRN model, which includes both the Fusion Module and the Relation Graph Embedding Module, achieved the highest performance across all metrics. This configuration effectively captures the complex interactions within the Gene Regulatory Network (GRN), handling the skewed degree distribution of genes while preserving accuracy and robustness.

When only the Relation Graph Embedding Module was included, the model's performance dropped slightly. These reductions highlight the importance of the Fusion Module in enhancing the model's discriminative power. By focusing on the most relevant gene expression features and the correlations between regulator and target genes, the Fusion Module improves the model's ability to accurately distinguish activation, repression, and non-regulated interactions.

Conversely, when only the Fusion Module was included, the model's performance also declined. This underscores the critical role of the Relation Graph Embedding Module in preserving the model's accuracy and robustness. The Relation Graph Embedding Module captures the connectivity and directionality of interactions within the GRN, allowing the model to effectively handle the skewed degree distribution of genes.

These findings demonstrate that the combination of the Relation Graph Embedding module and the Fusion module greatly enhances the model's ability to accurately predict regulatory interactions. Both components are indispensable for the model's success.

Sensitivity analysis

As shown in Fig. 5, the connection-aware weight λ demonstrated dataset-dependent optimal ranges. DREAM5 networks maintained stable performance (AUC fluctuation <0.3%) across all λ_0 values, whereas Ecoli cold and Human COVID datasets peaked at $\lambda = 0.3$. The decay rate q significantly impacted biological data consistency, with $q = 1e-4$ achieved maximum AUC in all tested set.The multi-head analysis revealed distinct architectural preferences - DREAM5 achieved peak performance with h=4 (AUC 0.9459), while biological networks required expanded capacity with COVID dataset peaking at h=16 (AUC 0.9191), suggesting complex disease relationships demand richer feature representations. Optimal configurations emerged at $\lambda = 0.3$ with $q \leq 0.001$ and h = 4/16, balancing connection regularization and directional refinement.

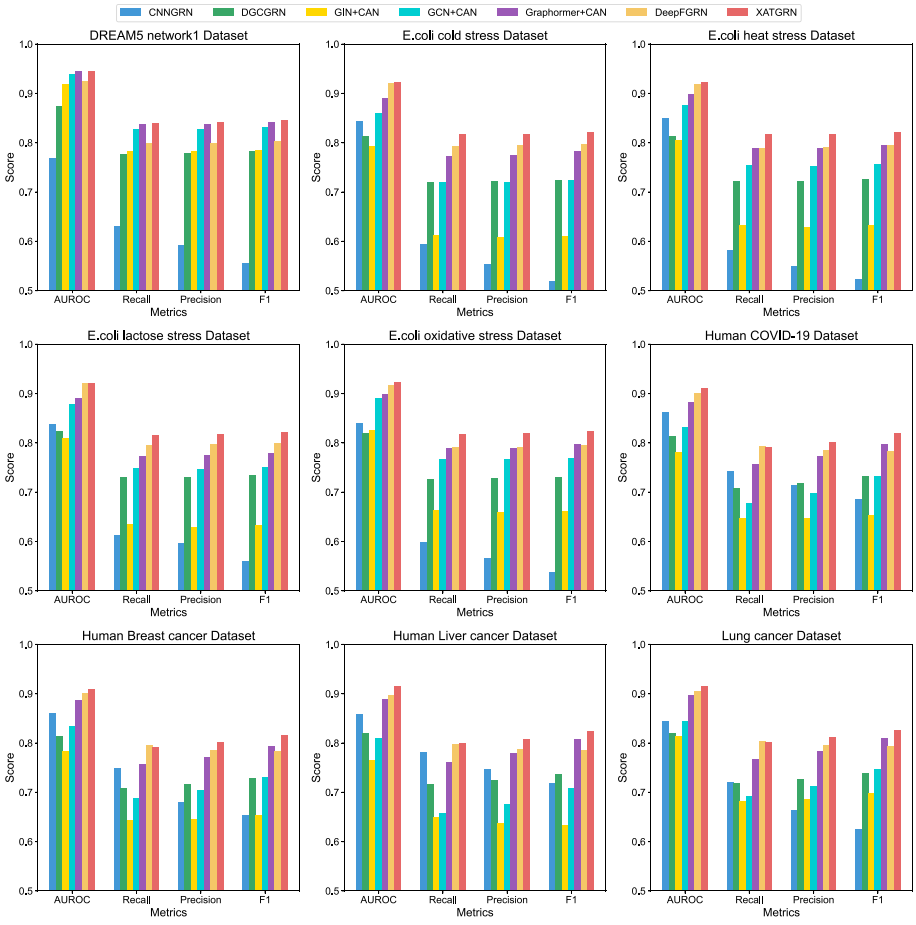

Fig. 3 Average results over 10 times of 5-fold cross-validation for model performance comparisons across 9 datasets

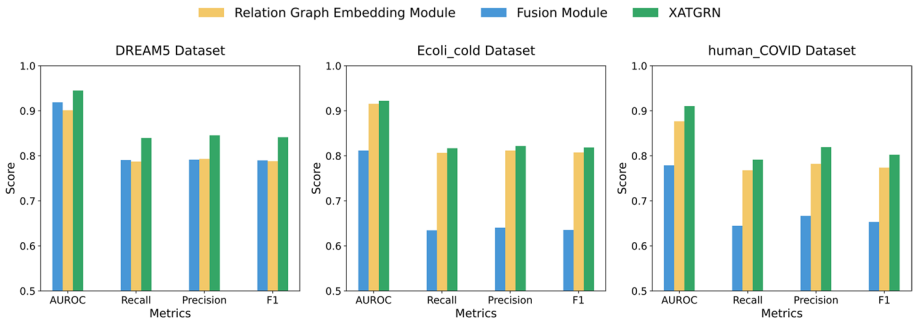

Fig. 4 Ablation study in DREAM5 Ecoli cold and human COVID datasets

Case study

To validate the biological significance of XATGRN, we reconstructed GRN using breast cancer data and employed in-depth analysis, including prediction of biomarkers and enrichment analysis of potential therapeutic drugs. The constructed breast cancer GRN contains 2,478 genes and 8,772 relationships. After reconstructing GRN, we selected ten hub genes with the highest degree (Fig. 6).

The hub genes in Fig. 6 have been fully validated through literature review. Both RELA and NFKB1 are important members of the nuclear factor kappa-B family. In breast

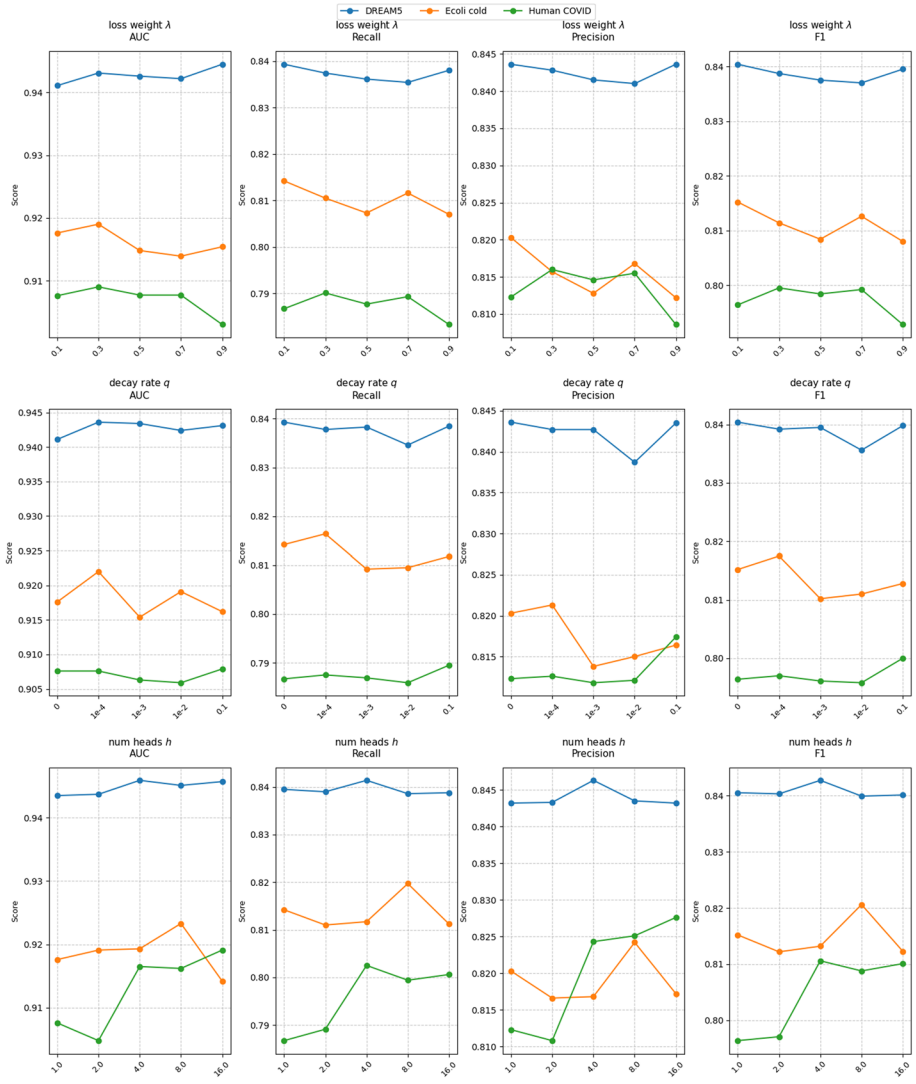

Fig. 5 Sensitivity analysis in DREAM5 Ecoli cold and human COVID datasets

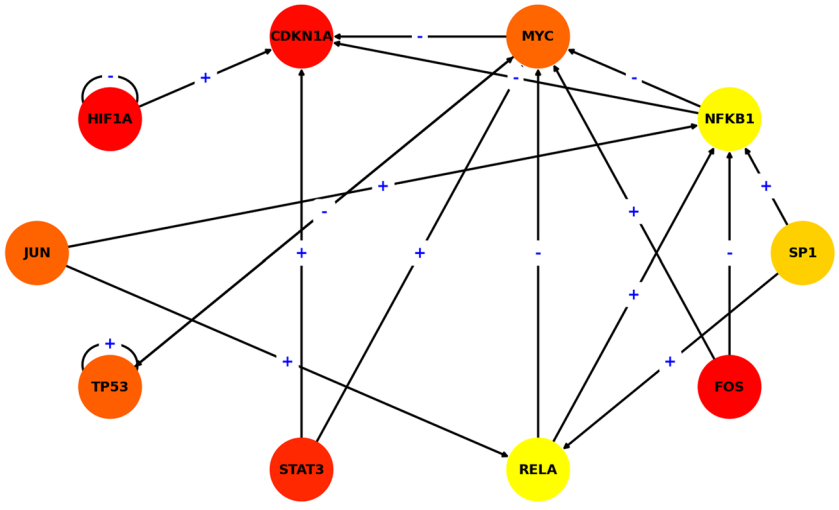

Fig. 6 Hub gene of the breast cancer

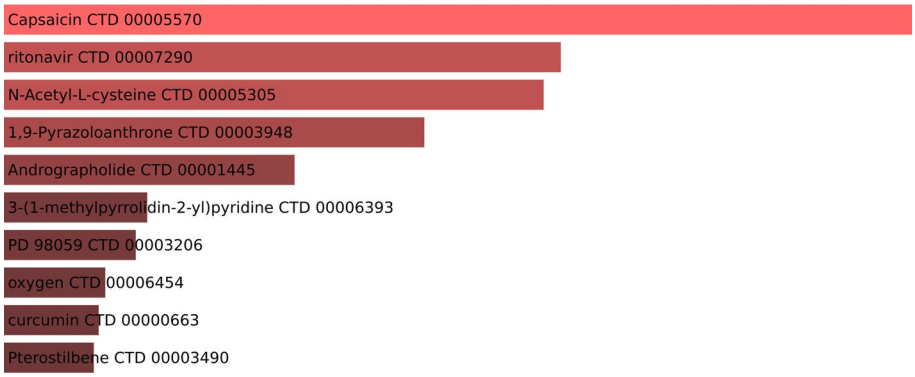

Fig. 7 Drug enrichment analysis of breast cancer

cancer, the abnormal activation of the kappa-B signaling pathway is closely associated with the occurrence, development, invasion, and metastasis of tumors [37, 38]. SP1 is closely related to the staging, invasive potential, and survival rates of breast cancer, and high levels of SP1 often indicate poor prognosis for patients [39]. MYC is a key regulator of cell growth, proliferation, metabolism, differentiation, and apoptosis, and its deregulation contributes to breast cancer development and progression, associated with poor outcomes [40]. The transcription factor Jun is closely associated with metastasis and prognosis in breast cancer, acting as both a suppressor and oncogene [41]. TP53 mutation, frequently occurring in triple-negative breast cancer (TNBC), enhances the correlation between the high-MYC and low-TXNIP gene signature and death from breast cancer [42]. STAT3 plays a crucial role in the regulation of cancer hallmarks in breast cancer, including angiogenesis, metabolism, and invasion, and is involved in tamoxifen resistance [43]. The cytoplasmic localization of CDKN1A/p21 is predominantly associated with cancer, where it serves to promote tumorigenesis and inhibit apoptosis in breast cancer cell lines [44]. Hypoxia inducible factor-1α(HIF-1α) is crucial in the regulation of cancer hallmarks in breast cancer, including angiogenesis, metabolism, and invasion [43]. Lastly, FOS is downregulated in breast cancer tissues and cells, and its overexpression restrains the malignant phenotypes of breast cancer cells [45].

In our constructed breast cancer GRN, the regulatory interactions between specific regulator genes and their target genes exhibit both positive and negative regulatory mechanisms. For instance, for positive regulation, STAT3 has been shown to accelerate Myc-induced tumor formation in a mouse model of breast cancer[46]. This finding highlights the role of STAT3 in promoting the expression of Myc, a key oncogene involved in tumor progression. Conversely, for negative regulation, loss of p53 in breast cancer leads to Myc activation, increased cell plasticity, and expression of a mitotic signature with prognostic value [47]. Additionally, p53 represses c-Myc through the induction of the tumor suppressor miR-145 [48]. These studies demonstrate that p53 acts as a negative regulator of Myc, highlighting the critical role of p53 in controlling Myc expression and maintaining cellular homeostasis. The accurate representation of both positive and negative regulatory mechanisms in our GRN highlights its potential to effectively capture the complex gene regulatory dynamics involved in breast cancer.

Furthermore, we performed drug enrichment analysis on top ten hub genes of breast cancer predicted by XATGRN. Figure 7 shows the top 10 enriched potential drugs based on DsigDB obtained through hub genes. It has been confirmed that seven out of the 10

drugs in Fig. 7 may be used for the treatment of breast cancer. Capsaicin, known for its activation of the TRPV1 receptor, has shown potential in inhibiting breast cancer cell growth by inducing apoptosis [49]. Ritonavir, typically used as an HIV protease inhibitor, is being studied for its potential to boost the effectiveness of breast cancer chemotherapy [50]. N-Acetyl-L-cysteine (NAC), with its antioxidant properties, shows promise in mitigating chemotherapy side effects and may play a role in breast cancer management [51]. Andrographolide has been found to suppress breast cancer progression by inhibiting COX-2 expression and angiogenesis, affecting p300 signaling and the VEGF pathway [52]. PD 98059, a MEK inhibitor, reduces the invasive capabilities of breast cancer cells by disrupting the MAPK signaling pathway [53]. Hyperbaric oxygen therapy is being explored as a supportive measure to improve the outcomes of radiotherapy for breast cancer [54]. Curcumin, a component of turmeric, has shown potential in breast cancer treatment due to its anti-tumor, anti-oxidative, and anti-inflammatory properties, coupled with its low toxicity and high safety profile [55].

Conclusion and discussion

In this paper, we have introduced the Cross-Attention Complex Dual Graph Attention Network Embedding Model (XATGRN) for Gene Regulatory Network (GRN) inference. This model addresses several critical challenges in GRN prediction, including the accurate representation of gene regulatory interactions, the handling of skewed degree distributions, and the effective capture of complex gene-gene relationships. By incorporating a cross-attention mechanism, XATGRN enhances the model's ability to predict not only the presence of regulatory relationships but also their directionality and specific types, such as activation or repression.

Our results show that XATGRN consistently outperforms state-of-the-art models across multiple datasets. The cross-attention mechanism allows XATGRN to focus on the most relevant features from bulk gene expression data, while our relation graph embedding module effectively captures both connectivity and directionality within the GRN, even in the presence of imbalanced node degrees. This combination of strategies enables XATGRN to overcome the limitations of existing models, making it more robust and applicable in real-world biological contexts. The strong performance of XATGRN across diverse datasets emphasizes its robustness and generalizability, positioning it as a promising tool for exploring GRNs in a variety of biological contexts.

Extensive experiments on benchmark datasets underscore the model's effectiveness in uncovering previously unknown regulatory mechanisms and its potential to identify novel therapeutic targets for complex diseases. Our XATGRN model provides a comprehensive and powerful framework for advancing our understanding of gene regulatory networks, offering a valuable approach for both basic and applied biological research.

In conclusion, our XATGRN represents a significant step forward in GRN inference, providing a robust and accurate framework for studying gene regulatory mechanisms. By effectively managing skewed degree distributions and leveraging advanced attention mechanisms, XATGRN serves as a powerful tool for uncovering regulatory interactions and identifying potential therapeutic targets.

Author contributions

J.X. coordinated the overall research design and wrote the main manuscript text. N.Y. and S.L. conducted the primary experiments and data collection, with N.Y. also preparing key figures and tables. H.L. provided technical support for the experiments and contributed to the methods section. Y.W. and D.A. contributed to the theoretical framework and data

analysis, respectively. F.P. assisted with the literature review and manuscript structure. J.W. supervised the project and provided critical revisions. All authors reviewed and approved the final manuscript.

Funding
This work was sponsored by Shanxi innovation ability support plan (2024CX-GXPT-44)

Data availability
Benchmark data used in our work is obtained from DeepFGRN https://github.com/PhoebeGaoZhen/DeepFGRN/tree/ master. The codes and detailed requirements have been released on Github: https://github.com/kikixiong/XATGRN.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no Conflict of interest.

Author details
[1]Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Guangzhou 528225, China
[2]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China
[3]Department of Internal Medicine, The No. 944 Hospital of Joint Logistic Support Force of PLA, Xiongguan Road, Jiu Quan 735000, China
[4]Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[5]Department of Dermatology, Xijing Hospital, Fourth Military Medical University, No 127 of West Changle Road, Xi'an 710032, Shaanxi, China
[6]Department of Gastroenterology, Tangdu Hospital, Fourth Military Medical University, Shaanxi 710038, China

Received: 13 February 2025 / Accepted: 10 June 2025
Published online: 16 July 2025

References

1. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, Vieth B, Vick B, Jeremias I, Ziegenhain C. Prime-seq, efficient and powerful bulk rna sequencing. Genome Biol. 2022;23(1):88.
2. Levine M, Davidson EH. Gene regulatory networks for development. Proc Natl Acad Sci. 2005;102(14):4936–42.
3. Dong J, Li J, Wang F. Deep learning in gene regulatory network inference: a survey. IEEE/ACM Trans Comput Biol Bioinform (2024)
4. Kloesch B, Ionasz V, Paliwal S, Hruschka N, Villarreal JM, Öllinger R, Mueller S, Dienes HP, Schindl M, Gruber ES. A gata6-centred gene regulatory network involving hnfs and δnp63 controls plasticity and immune escape in pancreatic cancer. Gut. 2022;71(4):766–77.
5. Meng Z, Liu S, Liang S, Jani B, Meng Z. Heterogeneous biomedical entity representation learning for gene-disease association prediction. Brief Bioinform. 2024;25(5):380.
6. Goldman JA, Poss KD. Gene regulatory programmes of tissue regeneration. Nat Rev Genet. 2020;21(9):511–25.
7. Sonawane AR, Platig J, Fagny M, Chen C-Y, Paulson JN, Lopes-Ramos CM, DeMeo DL, Quackenbush J, Glass K, Kuijjer ML. Understanding tissue-specific gene regulation. Cell Rep. 2017;21(4):1077–88.
8. Yi X, Liu S, Wu Y, McCloskey D, Meng Z. Bpp: a platform for automatic biochemical pathway prediction. Brief Bioinform. 2024;25(5):355.
9. Salleh F, Arif SM, Zainudin S, Firdaus-Raih M. Reconstructing gene regulatory networks from knock-out data using gaussian noise model and pearson correlation coefficient. Comput Biol Chem. 2015;59:3–14.
10. Liu F, Zhang S-W, Guo W-F, Wei Z-G, Chen L. Inference of gene regulatory network based on local bayesian networks. PLoS Comput Biol. 2016;12(8):1005024.
11. Xing L, Guo M, Liu X, Wang C, Wang L, Zhang Y. An improved bayesian network method for reconstructing gene regulatory network based on candidate auto selection. BMC Genomics. 2017;18:17–30.
12. Raza K, Alam M. Recurrent neural network based hybrid model for reconstructing gene regulatory network. Comput Biol Chem. 2016;64:322–34.
13. Ju W, Yi S, Wang Y, Xiao Z, Mao Z, Li H, Gu Y, Qin Y, Yin N, Wang S et al. A survey of graph neural networks in real world: Imbalance, noise, privacy and ood challenges. arXiv preprint arXiv:2403.04468 (2024)
14. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The sequence read archive: a decade more of explosive growth. Nucleic Acids Res. 2022;50(D1):387–90.
15. Shiraishi Y, Okada A, Chiba K, Kawachi A, Omori I, Mateos RN, Iida N, Yamauchi H, Kosaki K, Yoshimi A. Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data. Nat Commun. 2022;13(1):5357.
16. Gao Z, Tang J, Xia J, Zheng C-H, Wei P-J. Cnngrn: A convolutional neural network-based method for gene regulatory network inference from bulk time-series expression data. IEEE/ACM Trans Comput Biol Bioinf. 2023;20(5):2853–61.
17. Wang J, Ma A, Ma Q, Xu D, Joshi T. Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks. Comput Struct Biotechnol J. 2020;18:3335–43.

18. Liu S, Meng Z, Macdonald C, Ounis I. Graph neural pre-training for recommendation with side information. ACM Trans Inf Syst. 2023;41(3):1–28.
19. Liu S, Ounis I, Macdonald C. An mlp-based algorithm for efficient contrastive graph recommendations. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022;2431–2436
20. Yin N, Shen L, Wang M, Lan L, Ma Z, Chen C, Hua X-S, Luo X. Coco: A coupled contrastive framework for unsupervised domain adaptive graph classification. In: International Conference on Machine Learning, 2023;40040–40053. PMLR
21. Wang Y, Yin N, Xiao M, Yi X, Liu S, Liang S. Dusego: Dual second-order equivariant graph ordinary differential equation. arXiv preprint arXiv:2411.10000 (2024)
22. Wang Y, Liang V, Yin N, Liu S, Segal E. SGAC: A Graph Neural Network Framework for Imbalanced and Structure-Aware AMP Classification (2024). https://arxiv.org/abs/2412.16276
23. Liang S, Liu S, Song J, Lin Q, Zhao S, Li S, Li J, Liang S, Wang J. Hmcda: a novel method based on the heterogeneous graph neural network and metapath for circrna-disease associations prediction. BMC Bioinform. 2023;24(1):335. https://doi.org/10.1186/s12859-023-05441-7.
24. Wang Z, Liang S, Liu S, Meng Z, Wang J, Liang S. Sequence pre-training-based graph neural network for predicting lncrna–mirna associations. Brief Bioinform. 2023;24(5):317. https://doi.org/10.1093/bib/bbad317.
25. Wei P-J, Guo Z, Gao Z, Ding Z, Cao R-F, Su Y, Zheng C-H. Inference of gene regulatory networks based on directed graph convolutional networks. Brief Bioinform 2024;25(4)
26. Tong Z, Liang Y, Sun C, Rosenblum DS, Lim A. Directed graph convolutional network. arXiv preprint arXiv:2004.13970 (2020)
27. Pol AA, Berger V, Germain C, Cerminara G, Pierini M. Anomaly detection with conditional variational autoencoders. In: 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), 2019;1651–1657. IEEE
28. Gao Z, Su Y, Xia J, Cao R-F, Ding Y, Zheng C-H, Wei P-J. Deepfgrn: inference of gene regulatory network with regulation type based on directed graph embedding. Brief Bioinform. 2024;25(3):143.
29. Ke Z, Yu H, Li J, Zhang H. DUPLEX: Dual GAT for Complex Embedding of Directed Graphs (2024). https://arxiv.org/abs/2406.05391
30. Wang Y, Liu S, Wang M, Liang S, Yin N. Degree distribution based spiking graph networks for domain adaptation. arXiv preprint arXiv:2410.06883 (2024)
31. Meng Z, Meng Z, Yuan K, Ounis I. FusionDTI: Fine-grained Binding Discovery with Token-level Fusion for Drug-Target Interaction (2024). https://arxiv.org/abs/2406.01651
32. Gao Z, Tang J, Xia J, Zheng C-H, Wei P-J. Cnngrn: A convolutional neural network-based method for gene regulatory network inference from bulk time-series expression data. IEEE/ACM Trans Comput Biol Bioinf. 2023;20(5):2853–61. https://doi.org/10.1109/TCBB.2023.3282212.
33. Wei P-J, Guo Z, Gao Z, Ding Z, Cao R-F, Su Y, Zheng C-H. Inference of gene regulatory networks based on directed graph convolutional networks. Brief Bioinform. 2024;25(4):309. https://doi.org/10.1093/bib/bbae309.
34. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. CoRR (2016) arxiv:1609.02907
35. Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? CoRR (2018) arxiv:1810.00826
36. Ying C, Cai T, Luo S, Zheng S, Ke G, He D, Shen Y, Liu T-Y. Do transformers really perform badly for graph representation? In: Thirty-Fifth Conference on Neural Information Processing Systems (2021). https://openreview.net/forum?id=OeWooOxFwDa
37. Karin M, Lin A. Nf-κb at the crossroads of life and death. Nat Immunol. 2002;3(3):221–7.
38. Biswas DK, Dai S-C, Cruz A, Weiser B, Graner E, Pardee AB. The nuclear factor kappa b (nf-κb): a potential therapeutic target for estrogen receptor negative breast cancers. Proc Natl Acad Sci. 2001;98(18):10386–91.
39. Zhao Y, Ma J, Fan Y, Wang Z, Tian R, Ji W, Zhang F, Niu R. Tgf-β transactivates egfr and facilitates breast cancer migration and invasion through canonical smad3 and erk/sp1 signaling pathways. Mol Oncol. 2018;12(3):305–21.
40. Xu J, Chen Y, Olopade OI. Myc and breast cancer. Genes & cancer. 2010;1(6):629–40.
41. Zhu P, Liu G, Wang X, Lu J, Zhou Y, Chen S, Gao Y, Wang C, Yu J, Sun Y. Transcription factor c-jun modulates glut1 in glycolysis and breast cancer metastasis. BMC Cancer. 2022;22(1):1283.
42. Børresen-Dale A-L. Tp53 and breast cancer. Hum Mutat. 2003;21(3):292–300.
43. Mirzaei S, Ranjbar B, Tackallou SH, Aref AR. Hypoxia inducible factor-1α(hif-1α) in breast cancer: The crosstalk with oncogenic and onco-suppressor factors in regulation of cancer hallmarks. Pathol Res Pract 2023;154676
44. Wei C-Y, Tan Q-X, Zhu X, Qin Q-H, Zhu F-B, Mo Q-G, Yang W-P. Expression of cdkn1a/p21 and tgfbr2 in breast cancer and their prognostic significance. Int J Clin Exp Pathol. 2015;8(11):14619.
45. Chang D, Li L, Xu Z, Chen X. Targeting fos attenuates malignant phenotypes of breast cancer: Evidence from in silico and in vitro studies. J Biochem Mol Toxicol. 2023;37(7):23358.
46. Jhan J-R, Andrechek ER. Stat3 accelerates myc induced tumor formation while reducing growth rate in a mouse model of breast cancer. Oncotarget. 2016;7(40):65797.
47. Santoro A, Vlachou T, Luzi L, Melloni G, Mazzarella L, D'Elia E, Aobuli X, Pasi CE, Reavie L, Bonetti P. p53 loss in breast cancer leads to myc activation, increased cell plasticity, and expression of a mitotic signature with prognostic value. Cell Rep. 2019;26(3):624–38.
48. Sachdeva M, Zhu S, Wu F, Wu H, Walia V, Kumar S, Elble R, Watabe K, Mo Y-Y. p53 represses c-myc through induction of the tumor suppressor mir-145. Proc Natl Acad Sci. 2009;106(9):3207–12.
49. Chen M, Xiao C, Jiang W, Yang W, Qin Q, Tan Q, Lian B, Liang Z, Wei C. Capsaicin inhibits proliferation and induces apoptosis in breast cancer by down-regulating fbi-1-mediated nf-κb pathway. Drug Design Dev Therapy 2021;125–140
50. Hendrikx JJ, Lagas JS, Song J-Y, Rosing H, Schellens JH, Beijnen JH, Rottenberg S, Schinkel AH. Ritonavir inhibits intratumoral docetaxel metabolism and enhances docetaxel antitumor activity in an immunocompetent mouse breast cancer model. Int J Cancer. 2016;138(3):758–69.
51. Cheng H, Lee SH, Wu S. Effects of n-acetyl-l-cysteine on adhesive strength between breast cancer cell and extracellular matrix proteins after ionizing radiation. Life Sci. 2013;93(21):798–803.
52. Peng Y, Wang Y, Tang N, Sun D, Lan Y, Yu Z, Zhao X, Feng L, Zhang B, Jin L. Andrographolide inhibits breast cancer through suppressing cox-2 expression and angiogenesis via inactivation of p300 signaling and vegf pathway. J Exp Clin Cancer Res. 2018;37:1–14.

53. Normanno N, Luca AD, Maiello MR, Campiglio M, Napolitano M, Mancino M, Carotenuto A, Viglietto G, Menard S. The mek/mapk pathway is involved in the resistance of breast cancer cells to the egfr tyrosine kinase inhibitor gefitinib. J Cell Physiol. 2006;207(2):420–7.
54. Batenburg MC, Maarse W, Leij F, Baas IO, Boonstra O, Lansdorp N, Doeksen A, Bongard DH, Verkooijen HM. The impact of hyperbaric oxygen therapy on late radiation toxicity and quality of life in breast cancer patients. Breast Cancer Res Treat. 2021;189:425–33.
55. Hu S, Xu Y, Meng L, Huang L, Sun H. Curcumin inhibits proliferation and promotes apoptosis of breast cancer cells. Exp Ther Med. 2018;16(2):1266–72.

Publisher's Note