Briefings in Bioinformatics, 2025, **26(2)**, bbaf118
https://doi.org/10.1093/bib/bbaf118
Problem Solving Protocol

AttentionGRN: a functional and directed graph transformer for gene regulatory network reconstruction from scRNA-seq data

Zhen Gao [1], Yansen Su[2], Jin Tang[1], Huaiwan Jin[2], Yun Ding[2], Rui-Fen Cao[1], Pi-Jing Wei[3,*], Chun-Hou Zheng[2,*]

[1]The Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, 111 Jiulong Road, Hefei 230601, Anhui, China
[2]The Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Artificial Intelligence, Anhui University, 111 Jiulong Road, Hefei 230601, Anhui, China
[3]Information Materials and Intelligent Sensing Laboratory of Anhui Province, The Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Institute of Physical Science and Information Technology, Anhui University, 111 Jiulong Road, Hefei 230601, Anhui, China

*Corresponding authors. Pi-Jing Wei, Information Materials and Intelligent Sensing Laboratory of Anhui Province, The Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Institute of Physical Science and Information Technology, Anhui University, 11 1 Jiulong Road, Hefei 230601, Anhui, China. E-mail: weipj@ahu.edu.cn; Chun-Hou Zheng, the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Artificial Intelligence, Anhui University, 111 Jiulong Road, Hefei 230601, Anhui, China. E-mail: zhengch99@126.com

Abstract

Single-cell RNA sequencing (scRNA-seq) enables the reconstruction of cell type-specific gene regulatory networks (GRNs), offering detailed insights into gene regulation at high resolution. While graph neural networks have become widely used for GRN inference, their message-passing mechanisms are often limited by issues such as over-smoothing and over-squashing, which hinder the preservation of essential network structure. To address these challenges, we propose a novel graph transformer-based model, AttentionGRN, which leverages soft encoding to enhance model expressiveness and improve the accuracy of GRN inference from scRNA-seq data. Furthermore, the GRN-oriented message aggregation strategies are designed to capture both the directed network structure information and functional information inherent in GRNs. Specifically, we design directed structure encoding to facilitate the learning of directed network topologies and employ functional gene sampling to capture key functional modules and global network structure. Our extensive experiments, conducted on 88 datasets across two distinct tasks, demonstrate that AttentionGRN consistently outperforms existing methods. Furthermore, AttentionGRN has been successfully applied to reconstruct cell type-specific GRNs for human mature hepatocytes, revealing novel hub genes and previously unidentified transcription factor-target gene regulatory associations.

Keywords: gene regulatory network; graph transformer; directed structural encoding; functional module

Introduction

Gene regulatory networks (GRNs) are complex, directed networks composed of transcription factors (TFs), target genes, and their regulatory relationships. These networks control essential biological processes, including cell differentiation, apoptosis, and organismal development [1]. Reconstructing GRNs from gene expression profiles provides critical insights into normal and pathological changes during tissue or organism development, serving as a foundation for advancing disease treatments and biotechnological applications [2, 3].

Numerous computational methods have been developed to infer GRNs from bulk RNA sequencing (RNA-seq) data [4]. Bulk RNA-seq generates averaged transcriptional profiles from mixed RNA samples, such as tissues or organs, which masks the cellular heterogeneity present in complex biological systems. While bulk RNA-seq allows for the reconstruction of global GRNs in diseased organs, it lacks the resolution to distinguish between different cell types. In contrast, single-cell RNA sequencing (scRNA-seq) [5] enables high-throughput sequencing of individual cells, allowing for the reconstruction of cell type-specific GRNs with greater precision. This technology offers new opportunities to study regulatory mechanisms at the single-cell level, facilitating a deeper understanding of both normal and disease states in specific cell populations, such as healthy cells, diseased cells, or immune cells [6].

Methods for GRN inference from scRNA-seq data can be broadly classified into unsupervised and supervised approaches. Supervised methods [7, 8], compared to unsupervised approaches [9, 10], offer improved predictive performance and can be applied to GRNs of varying scales. Based on how they process scRNA-seq data, current supervised techniques can be grouped into three main types. The first type transforms scRNA-seq data of TF-target gene pairs into histograms, which are then fed into convolutional neural networks (CNNs) for feature extraction [11–14]. The second type inputs scRNA-seq data directly to deep learning models for feature extraction [15–19]. The third type models scRNA-seq data as node attributes in undirected or directed graphs, where TF–target gene interactions are represented as edges [20–24].

Received: November 20, 2024. **Revised:** February 12, 2025. **Accepted:** February 27, 2025

While the first category of methods [11–14] converts gene expression matrices into image-like histograms, which are then processed by CNNs, this transformation introduces noise and can obscure important features of the original data, making it time-consuming and less efficient. The second category [15–19] addresses this issue by directly learning from gene expression data, significantly enhancing computational efficiency and prediction accuracy. However, both of these approaches overlook the network structure inherent in GRNs, which is crucial for understanding the functional relationships between genes.

The network structure of a GRN plays a key role in inferring regulatory relationships, as genes or proteins with similar functions are often topologically related. Therefore, methods that incorporate network structure information are essential for accurately predicting gene regulatory relationships. Recent approaches based on graph neural networks (GNNs), such as scSGL [20], GENELink [21], GNNLink [22], DGCGRN [23], and GATCL [25], have attempted to leverage this information. These methods represent gene expression data as node features and the regulatory relationships between TFs and target genes as edges within a graph, capturing network structure through message-passing mechanisms. However, a growing body of research recognizes the limitations of GNNs, particularly issues such as over-smoothing and over-squashing [26]. Over-smoothing occurs when repeated message passing causes node representations to converge to similar values, while over-squashing refers to the ineffective propagation of information across distant nodes due to excessive compression in deep models. The limitations of GNNs stem from their hard-encoded message-passing paradigm, which constrains the flexibility of information flow. In contrast, graph transformers (GTs) employ soft encoding, which incorporates the structural and positional information of nodes into their features. To this end, this study introduces GTs into GRN inference, leveraging the self-attention mechanism to capture global network features and guide the model to learn directed local structural information within GRNs.

In addition to methods that focus on network structure, some approaches integrate both gene expression and network structure features [27, 28]. These hybrid methods have been shown to significantly improve prediction accuracy. However, they typically only identify the neighbors of a gene without effectively propagating information from those neighbors. To address this limitation, our study aims to simultaneously learn gene expression patterns and directed network topologies, facilitating the propagation of features from both local and functionally related neighbors through a message-passing mechanism.

Functional modules, or sets of genes with similar biological functions, are key components of GRNs. These modules regulate similar biological processes or participate in specific pathways. However, many existing methods fail to consider the role of functionally related genes in GRN inference. To address this, we incorporate functional neighbors into the learning process, allowing the model to capture functional relationships alongside structural information. By using the GT framework, we aggregate features from both r-hop neighbors and functionally related neighbors, overcoming the sparsity of high-order neighbors in some GRNs.

In summary, this study presents AttentionGRN, a novel model for inferring GRNs from scRNA-seq data. By overcoming the limitations of traditional GNNs and incorporating GTs, AttentionGRN captures both global and local network features through a self-attention mechanism. Moreover, our model introduces task-oriented message aggregation strategies, which enhance the self-attention mechanism for GRN inference, allowing it to learn both directed structural and functional information. Extensive experiments on 88 benchmark datasets demonstrate the effectiveness of AttentionGRN in GRN inference. Furthermore, its application to human mature hepatocytes (hHEP) reveals novel hub genes and previously unknown regulatory relationships. The key contributions of this work are as follows:

- We propose AttentionGRN, a novel GT-based model for GRN reconstruction from scRNA-seq data, overcoming the limitations of traditional GNNs.
- AttentionGRN designs directed structure encoding strategy to help capture directed and local network structure information.
- AttentionGRN integrates both functionally related genes and r-hop neighbors to learn functional information and global network features.
- We curate 88 benchmark datasets for GRN reconstruction and successfully apply AttentionGRN to reconstruct cell type-specific GRNs in hHEP, leading to the discovery of new hub genes and regulatory associations. These findings highlight the robustness and versatility of AttentionGRN in decoding the complex regulatory mechanisms underlying biological systems.

Materials and methods

Overview of AttentionGRN

AttentionGRN is a supervised learning model designed to infer GRNs from scRNA-seq data. The AttentionGRN process involves four key stages: input preparation, information pre-extraction, dual-stream feature extraction, and GRN inference, as illustrated in Fig. 1.

The model can incorporate prior GRNs as input. We utilize 88 benchmark datasets, derived from curated resources and strategies in BEELINE [29], as detailed in Supplementary Table S1. The prior GRN is first processed by the information pre-extraction module, which generates gene expression sub-vectors, functionally related neighbor genes (*FN*), and directed structure identity (*DI*). Gene expression sub-vectors are directly input into the Transformer framework, while functionally related genes (*FN*) assist the GT in learning information from distant nodes. Directed structure identity (*DI*) provides guidance for learning directed network structure information.

The dual-stream feature extraction module is designed to capture both gene expression features and directed network structure features related to TF–target gene interactions, thus offering a comprehensive understanding of gene regulatory mechanisms. To obtain robust functional and directed network structure features, we leverage the GT to overcome the over-smoothing problem typically encountered in traditional GNNs. Additionally, this module integrates directed structure encoding to help the model learn local and asymmetric semantic information inherent in GRNs. During message aggregation in the GT, AttentionGRN does not propagate messages to all nodes, but instead focuses on r-hop neighbors and functionally related nodes, enabling the model to learn both functional information and global network structure of the GRN. For gene expression feature extraction, AttentionGRN inputs the gene expression sub-vectors of TF–gene pairs into the Transformer module [19], which includes a positional encoding component and an encoding module. The positional encoding module helps the model identify regulatory patterns within gene expression profiles, while the encoding module calculates the correlations between TFs and target genes.

The final step is GRN inference. In this stage, the gene expression features, along with functional and directed network

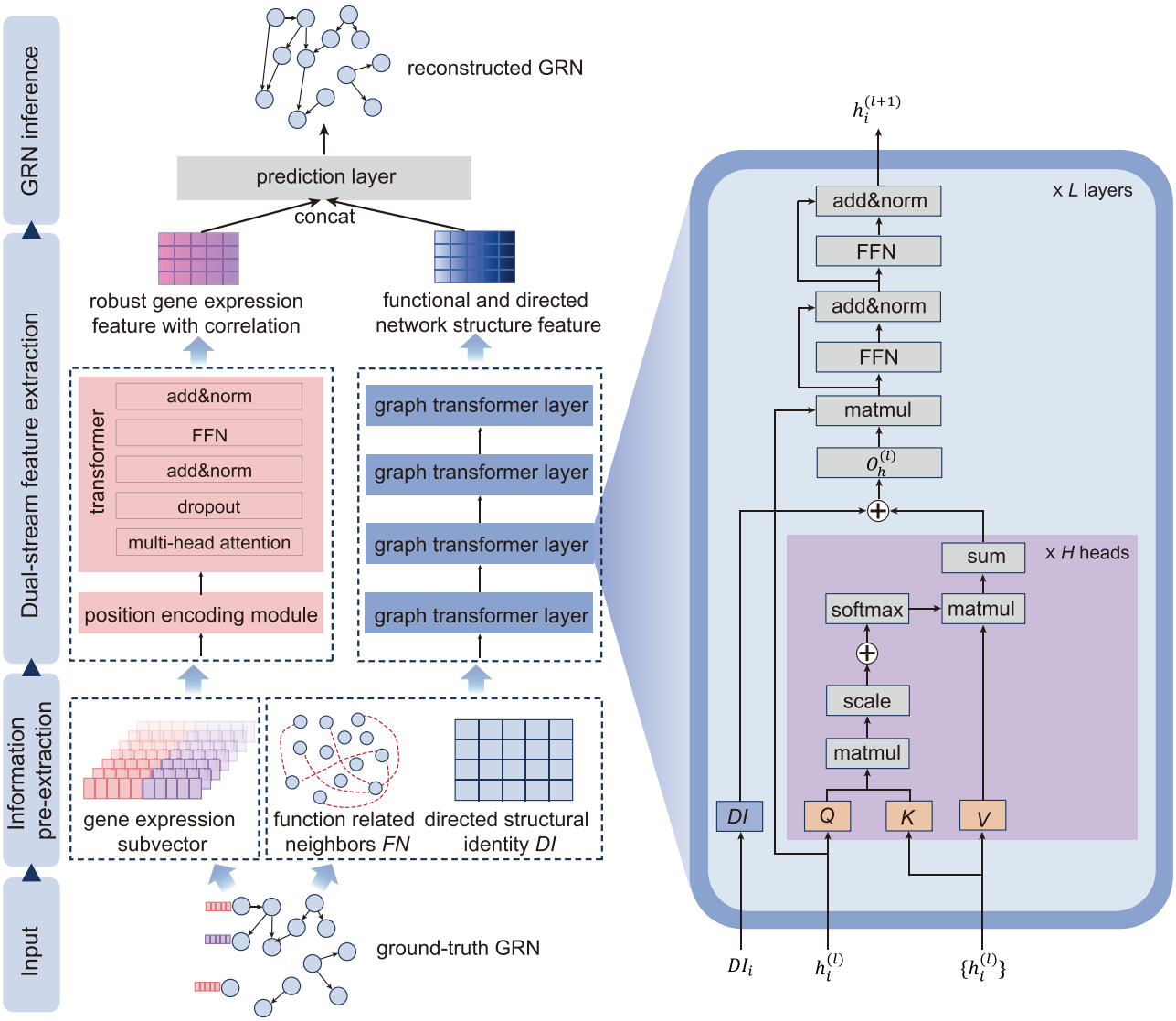

Figure 1. Overview of the AttentionGRN framework. Step1: Preparation of prior GRNs. Step2: Information pre-extraction, the aims are to obtain the gene expression sub-vector, functionally related genes FN and directed structure identity DI. Step3: Dual-stream feature extraction, the aims are to learn gene expression feature by Transformer and functional and directed network structure feature based on GT. Step4: GRN inference. Integrates the above two features and predicts candidate GRN.

structure information, are combined to form the final feature representation for each TF–gene pair. Specifically, the previous step extracted 64-dimensional gene expression features and 64-dimensional network structure features for each TF or gene. Here, by concatenating the two types of features, we obtain the features for the TF or gene (128 dimensions). Then, we can obtain the final features for the TF–target gene pair (256 dimensions). These final features are then passed to the prediction layer, which determines whether a TF regulates its target gene. The prediction layer contains two fully connected layers. Once the GRN is predicted, downstream analyses, such as hub gene identification and the discovery of novel regulatory associations, can be performed.

Model input

The BEELINE dataset [29], derived from real-world data, serves as the original dataset. It includes scRNA-seq data from seven distinct cell types and four categories of prior GRNs. These cell types are: human embryonic stem cells (hESC) [30], hHEP [31], mouse dendritic cells (mDC) [32], mouse embryonic stem cells (mESC) [33], mouse hematopoietic stem cells of the erythroid-lineage (mHSC-E) [34], granulocyte-monocyte-lineage (mHSC-GM) [34], and lymphoid-lineage (mHSC-L) [34]. The four types of prior GRNs include: cell type-specific GRNs, non-specific GRNs, functional interaction GRNs (STRING), and loss/gain of function (LOF/-GOF) GRNs, with the latter being exclusively applicable to mESC.

Two preprocessing methods [35] were applied to the datasets, referred to as DATA1 and DATA2. Both methods filtered out genes expressed in fewer than 10% of cells and selected genes with a p-value less than 0.01. The key difference between them lies in the selection of TFs. In DATA1, the top N ($N = 500, 1000$) most variable genes are identified first, and TFs among them are recorded. In DATA2, all highly variable TFs are identified initially, and then the top N ($N = 500, 1000$) most variable genes are selected from this subset. As a result, DATA2 allows the model to consider TFs that exhibit minimal changes in gene expression but can still regulate their target genes. This distinction led to the creation of four subsets: DATA1-500, DATA1-1000, DATA2-500, and

DATA2-1000, each containing 22 networks. In total, 88 prior GRNs were used in the study, as detailed in Supplementary Table S1.

The GRN can be represented as $G = \{V, E\}$, in which $V = \{v_1, v_2, ..., v_n\}$ and $E = \{e_{11}, e_{12}, ..., e_{nn}\}$ denote gene set and directed regulatory associations set, respectively. $e_{ij} = 1$ indicates that gene v_i regulates v_j, $e_{ij} = 0$ indicates that gene v_i does not regulate v_j. In addition, the node feature of scRNA-seq data is represented as X.

Functionally related genes

In the AttentionGRN model, the message passing of the GT is restricted to a predefined set of neighbors, rather than all nodes within the network. Specifically, we define the set of relevant neighbors to include both known neighboring nodes and functionally related nodes, as shown in Equation 1:

$$N_i = N_i^{(r)} \cup N_i^{(fn)} \tag{1}$$

where $N_i^{(r)}$ and $N_i^{(fn)}$ are r-hop neighbors and functionally related neighbors of gene v_i, respectively.

To identify functionally related genes, we first compute the cosine similarity score based on the gene expression matrix. A higher cosine similarity score indicates a stronger correlation between two genes. Two genes are considered functionally related if their cosine similarity score exceeds a predefined threshold. The threshold is determined as $mean+std*a$, where $mean$ is the average correlation score between the gene in question and all other genes, std is the standard deviation of these correlation scores, and a is a hyperparameter (see section Parameter sensitivity analysis).

Directed structure encoding

While the GT model addresses the limitations of traditional GNNs by employing soft encoding, effectively capturing positional or structural information within graphs remains a challenge. As GRNs are directed graphs, this study emphasizes the importance of directionality and local structure when designing structural encodings. Specifically, the directed structural identity (DI) in this study is composed of the in-degree, out-degree, degree centrality, r-hop source neighbors, and r-hop target neighbors of a node, as outlined in Equation 2:

$$DI_i = \{d_i^{in}, d_i^{out}, I_i^c, T_{ir}^s, T_{ir}^t\} \tag{2}$$

where the d_i^{in} and the d_i^{out} are in-degree and out-degree of gene v_i, respectively. The degree centrality information of gene v_i, denoted as I_i^c, is provided by Equation 3. The structure information of r-hop source neighbors and target neighbors of gene v_i are represented by T_{ir}^s and T_{ir}^t, respectively, as shown in Equation 4.

$$I_i^c = \{idc_i, odc_i, cc_i, bc_i, katzec_i\} \tag{3}$$

where idc_i denotes the in-degree centrality which reflects the capacity of gene v_i for receiving information; odc_i denotes the out-degree centrality which reflects the ability of gene v_i to propagate information; cc_i denotes the closeness centrality which measures the average distance of gene v_i to all other nodes within the GRN, with higher values implying a more rapid establishment of connections; bc_i denotes the betweenness centrality which quantifies the frequency of appearance of gene v_i on the shortest paths connecting other node pairs; $katzec_i$ denotes the Katz centrality which evaluates the importance of gene v_i within the network by considering both direct and indirect paths between nodes.

$$T_{ir}^{(s/t)} = \{I_{ir}^{in}, I_{ir}^{out}, I_i^{uc}\} \tag{4}$$

where I_{ir}^{in}, I_{ir}^{out}, and I_{ir}^{uc} are in-degree information, out-degree information, and degree centrality information of r-hop source neighbors or target neighbors of gene v_i, respectively, as shown in Equations 5–7.

$$I_{ir}^{in} = \{min_{ir}^{in}, max_{ir}^{in}, \mu_{ir}^{in}, \delta_{ir}^{in}\} \tag{5}$$

$$I_{ir}^{out} = \{min_{ir}^{out}, max_{ir}^{out}, \mu_{ir}^{out}, \delta_{ir}^{out}\} \tag{6}$$

$$I_{ir}^{uc} = \{\mu_{ir}^{idc}, \mu_{ir}^{odc}, \mu_{ir}^{cc}, \mu^{bc}, \mu_{ir}^{katzec}\} \tag{7}$$

where min, max, μ, and δ are the minimum, maximum, average, and standard deviation of the r-hop neighbor.

Dual-stream feature extraction

This module consists of two key components: gene expression feature extraction and functional and directed network structure feature extraction. A Transformer-based approach is employed to extract gene expression features, while a GT-based approach is used to capture functional and directed network structure features.

Gene expression features are extracted by passing the gene expression sub-vectors of TF-gene pairs through a positional encoding module, which helps identify regulatory patterns. The output is then passed through an encoding module to determine correlations between the sub-vectors [19].

To extract functional and directed network structure feature, L GT layers are established. Each GT layer contains multi-head self-attention mechanism, directed structural encoding, feedforward layer, as depicted in Equations 8–13.

The attention output of a single head is calculated (Equation 8), followed by integrating the attention output of multiple heads (Equation 9), adding it to the directed structure identity (10), performing layer normalization (Equation 11), and passing it through two feedforward layers (Equations 12–13).

$$\alpha_{ij}^{k,l} = softmax\left(\frac{Q^{k,l}h_i^l \cdot K^{k,l}h_j^l}{\sqrt{d_k}}\right) \tag{8}$$

$$\widehat{h}^{l+1} = O_h^l \prod_{k=1}^{H}\left(\sum_{j\in N_i}\alpha_{ij}^{k,l}V^{k,l}h_j^l\right) \tag{9}$$

$$\widehat{\widehat{h}}_i^{l+1} = Dropout(\widehat{h}^{l+1}) + f(DI_i^l) \tag{10}$$

$$\widehat{\widehat{h}}_i^{l+1} = LayerNorm\left(\widehat{\widehat{h}}_i^{l+1} + h_i^l\right) \tag{11}$$

$$\widehat{\widehat{h}}_i^{l+1} = W_B^l ReLU\left(W_A^l \widehat{\widehat{h}}_i^{l+1}\right) \tag{12}$$

$$h_i^{l+1} = LayerNorm\left(\widehat{\widehat{h}}_i^{l+1} + \widehat{\widehat{h}}_i^{l+1}\right) \tag{13}$$

where $Q^{k,l} \in R^{d_k \times d}$, $K^{k,l} \in \mathbb{R}^{d_k \times d}$, $V^{k,l} \in \mathbb{R}^{d_k \times d}$, $O_h^l \in \mathbb{R}^{d \times d}$. H is the number of attention head. | denotes concatenation. h_i and h_j are

features of genes v_i and v_j. l denotes the lth GT layer. $W_A^l \in \mathbb{R}^{2d \times d}$, $W_B^l \in \mathbb{R}^{d \times 2d}$. d_k is the dimension of key vector.

Evaluation strategies

There are two main types of tasks in the GRN inference field: gene–gene regulatory network (GRN) and TF–gene regulatory network (TRN) [11, 19]. The distinction between these tasks lies in the fact that TFs can be regulated by other TFs in GRNs, but not in TRNs [11, 19]. Most current studies use two different strategies for GRN inference and TRN inference: independent testing for GRN inference and TF-aware three-fold cross-validation (TTCV) for TRN inference [19].

For the independent testing strategy, positive samples consist of ground-truth TF–gene pairs, while negative samples are some unknown TF–target gene pairs. Specifically, the TFs in negative samples are the same as those in positive samples, and the target genes in negative samples are all genes that are not regulated by the aforementioned TFs. It can be observed that the dataset obtained in this way is imbalanced, with a significantly higher number of negative samples compared to positive samples. Subsequently, we combine the positive and negative samples and shuffle them. Then, divide the dataset into training, validation, and test sets in a 3:1:1 ratio, ensuring that the ratio of positive to negative samples is the same in each subset. The detailed process is shown in Supplementary Table S7.

TTCV considers ground-truth TF–gene pairs as positive samples, too. Then, for each positive sample, one hard negative sample is identified. For a positive sample A→B, the hard sample is A→C, where C is random selected from gene list without TFs. It can be observed that each TF in ground-truth TF–gene pairs has a positive sample and a negative sample. TTCV randomly divides all TFs into three equal parts, identifying corresponding positive and negative samples to form three subsets. Subsequently, TTCV conducts training and validation on two subsets, and tests on the remaining subset. By successively using each subset as the test set, three sets of results are obtained. Averaging these results provides the final TTCV results. The detailed process is shown in Supplementary Table S7.

For fairness, baseline models (CNNC [11], STGRNS [19], GENELink [21], GNNLink [22], and DeepFGRN [27]) were evaluated on DATA1 and DATA2 using the strategies described above. Specifically, 44 networks in DATA1-500 and DATA1-1000 were used for inferring GRNs, while the remaining 44 networks in DATA2-500 and DATA2-1000 were used for inferring TRNs. Evaluation metrics include the average AUROC and AUPR values across multiple tests.

Results and discussion

Parameter sensitivity analysis

Considering the diversity and complexity of GRNs, it is crucial to set appropriate parameters for different GRNs. Training the AttentionGRN model involves many parameters, some of which significantly impact predictive performance of AttentionGRN. Therefore, we conducted parameter tuning experiments for these parameters. Additionally, there are other parameters that have a smaller impact on the predictive performance, thus we set their values based on experience, as detailed in Supplementary Table S4.

The significant parameters include threshold a used in functionally related gene sampling, r-hop neighbors, the number L of GTs layers and training epochs. We conducted parameter analysis for each of the above parameters sequentially and found out appropriate values for each datasets, the experimental results are shown in Supplementary Figs S1–S7 and Tables S2–S4. This section only exhibits part of experimental results, namely the results of cell type-specific GRNs reconstruction on DATA1-500 and cell type-specific TRNs reconstruction on DATA2-500, as shown in Fig. 2.

First, we analyzed threshold a used in functionally related gene sampling. Observing that very few functionally related genes were identified when $a > 3.0$, and an excessive number of genes were selected when $a < 2.5$. Based on these observations, the threshold a was set within the range of $\{2.5, 2.6, 2.7, 2.8, 2.9, 3.0\}$, and the results are displayed in Fig. 2a–b and Supplementary Fig. S1. Then, the optimal a for each dataset was summarized in Supplementary Table S2.

Second, the analysis about r-hop neighbors was implemented. Many genes in GRNs have more low-order neighbors and fewer high-order neighbors. Setting high-order neighbors may introduce noise, thus we only tested one-hop, two-hop, and three-hop neighbors, as shown in Fig. 2c–d and Supplementary Fig. S2. The experimental results indicate that there are not many differences in predictive performance between the neighbors with one hop, two hops, and three hops. The predictive performance of one-hop neighbors is relatively better, and the computational efficiency is also relatively higher. Therefore, we ultimately set r to 1.

Third, the impact of the number of GT layers was evaluated, with results presented in Fig. 2e–f and Supplementary Fig. S3. On most datasets, increasing the number of layers resulted in higher mean prediction accuracy and lower standard deviation. However, the improvement in accuracy was minimal, and the standard deviation did not further decrease when $L \geq 10$. Given the trade-off between performance and computational cost, the number of GT layers was set to 10 for the final model.

Finally, loss curves for both training and validation sets were plotted for selecting optimal training epochs, as shown in Fig. 2g–h and Supplementary Figs S4–S7. Figure 2g shows the results of cell type-specific GRNs reconstruction for hESC. We initially set the epoch to 200 and found that the fitting of the model was optimal and had already converged after 100 epochs (Fig. 2g hESC-default parameters). Therefore, we trained this dataset for 100 epochs, resulting in a better fitting situation (Fig. 2g hESC-optimal parameters). Similarly, the appropriate epoch was selected for cell type-specific TRNs for hESC, as shown in Fig. 2h. The final epoch was set to 70. The optimal training epochs for all datasets are shown in Supplementary Table S3.

Performance on GRN and TRN inference

To evaluate the prediction performance and computational efficiency of AttentionGRN for GRN inference from scRNA-seq data, we performed a comparative analysis with state-of-the-art models on both GRN and TRN inference tasks, following the evaluation strategies outlined in section Evaluation strategies.

The baseline models used for comparison are:

- CNNC [11]: A CNN-based model that converts TF–gene pair gene expression data into histograms for feature extraction, which are then used for GRN reconstruction.
- STGRNS [19]: This model directly extracts features from TF–gene pair gene expression data using a Transformer architecture.
- GENELink [21] and GNNLink [22]: These methods treat GRN inference as a link prediction task. GENELink utilizes a Graph Attention Network (GAT) to extract topological structural features, while GNNLink employs Graph Autoencoders (GAE).

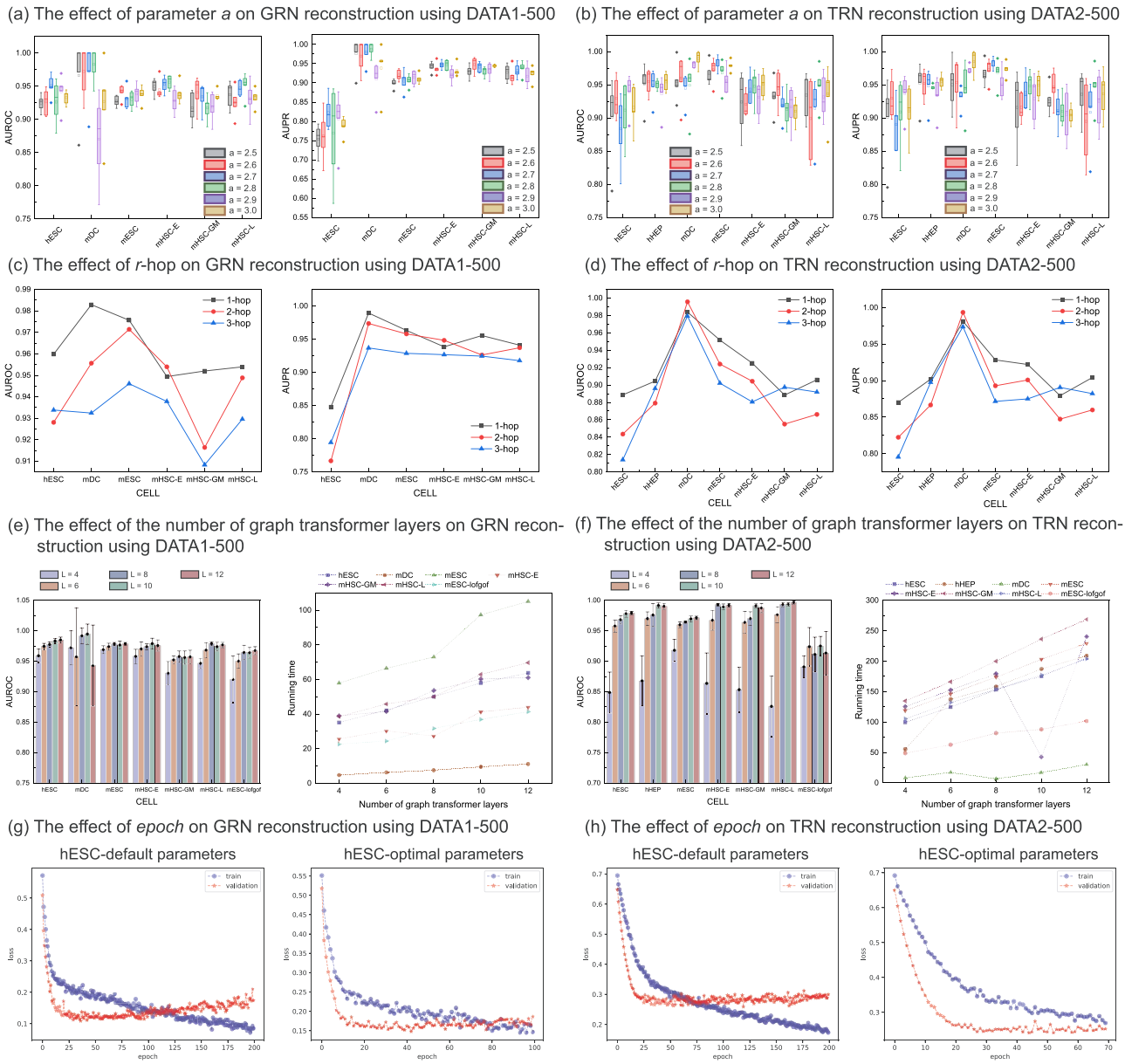

Figure 2. The effect of four parameters on the AttentionGRN model in predicting cell type-specific GRNs or TRNs. (a) and (b) represent the influence of parameter a on the model's prediction performance, (c) and (d) represent the loss curves of the model across different epochs, (e) and (f) represent the influence of r-hop neighbors on the model's prediction performance, and (g) and (h) represent the prediction performance and running times of different GT layers.

- DeepFGRN [27]: This model combines both gene expression features and directed network structural features for GRN reconstruction.

Performance on GRN inference

The experimental results of the aforementioned models are shown in Fig. 3 and supplementary comparison results. AttentionGRN outperforms existing methods on most datasets, as highlighted in Fig. 3. Specifically, the following key observations can be made: (i) AttentionGRN achieves superior performance across various types of GRNs. This demonstrates its ability to predict diverse GRNs, including those where other methods perform poorly, such as cell type-specific GRN inference for mDC in the DATA1-1000 dataset (Fig. 3b). (ii) Larger datasets improve performance. The models perform better on the DATA1-1000 dataset than on DATA1-500, suggesting that deep learning methods, including AttentionGRN, benefit from the increased data size, which aids in capturing more complex patterns. (iii) AttentionGRN achieves the best overall performance. STGRNS outperforms CNNC, likely due to the loss of important features when CNNC converts gene expression data into histograms, as well as the more effective feature extraction by the Transformer in STGRNS compared to the CNN in CNNC. Comparing CNNC and STGRNS with GENELink and GNNLink, the former focus more on gene expression features, while the latter concentrate on network structural features. The varying complexities of GRNs and the dimensionality of gene expression data make it difficult to establish clear superiority between these methods. DeepFGRN outperforms the other four methods on cell type-specific GRNs by integrating both gene expression and directed network structural features, but it still struggles with certain

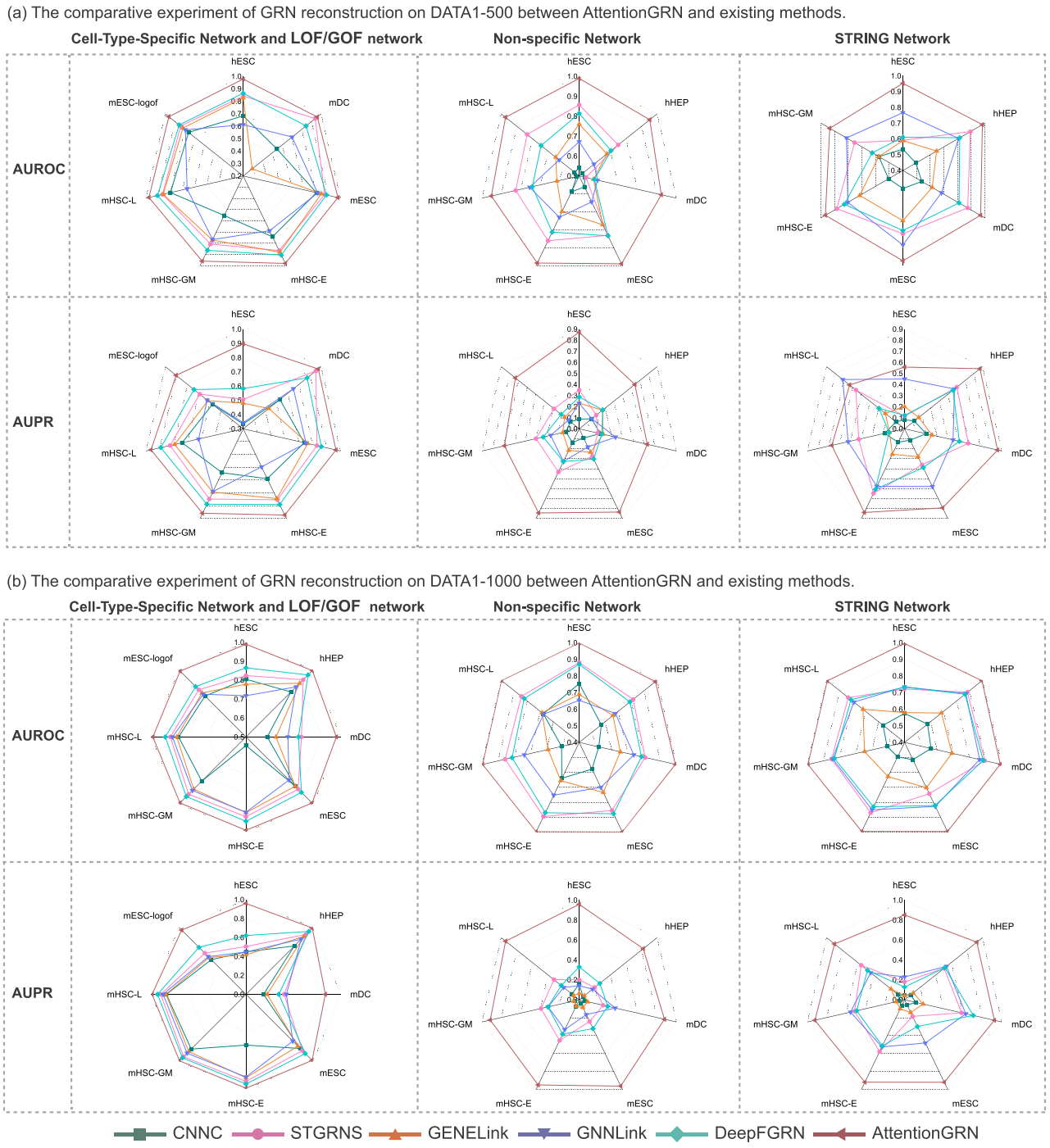

Figure 3. The results of AttentionGRN and other state-of-the-art methods for GRN inference. (a) and (b) show the results of models for GRN inference on DATA1-500 and DATA1-1000. The rows are AUROC and AUPR, respectively. The columns are results for cell type-specific networks and LOF/GOF networks, non-specific networks, STRING networks, respectively. The radar charts use different line colors to represent comparative methods, with each axis corresponding to a cell type.

types of GRNs. In contrast, AttentionGRN's use of a Transformer for gene expression feature extraction and a GT for network structure features enables it to achieve competitive performance compared to the other models.

Performance on TRN inference

The experimental results for TRN inference are shown in Fig. 4 and supplementary comparison results. AttentionGRN achieves the optimal predictive performance across all datasets. Key findings include: (i) network structure features are more important than gene expression features for TRN inference. GENELink outperforms STGRNS and CNNC on most datasets, emphasizing the crucial role of network structure in TRN inference. (ii) Directional semantic information is critical. The poor performance of GNNLink can be attributed to its failure to account for the directionality of TRNs when embedding them into the latent space using GCN. (iii) Joint learning of gene expression features

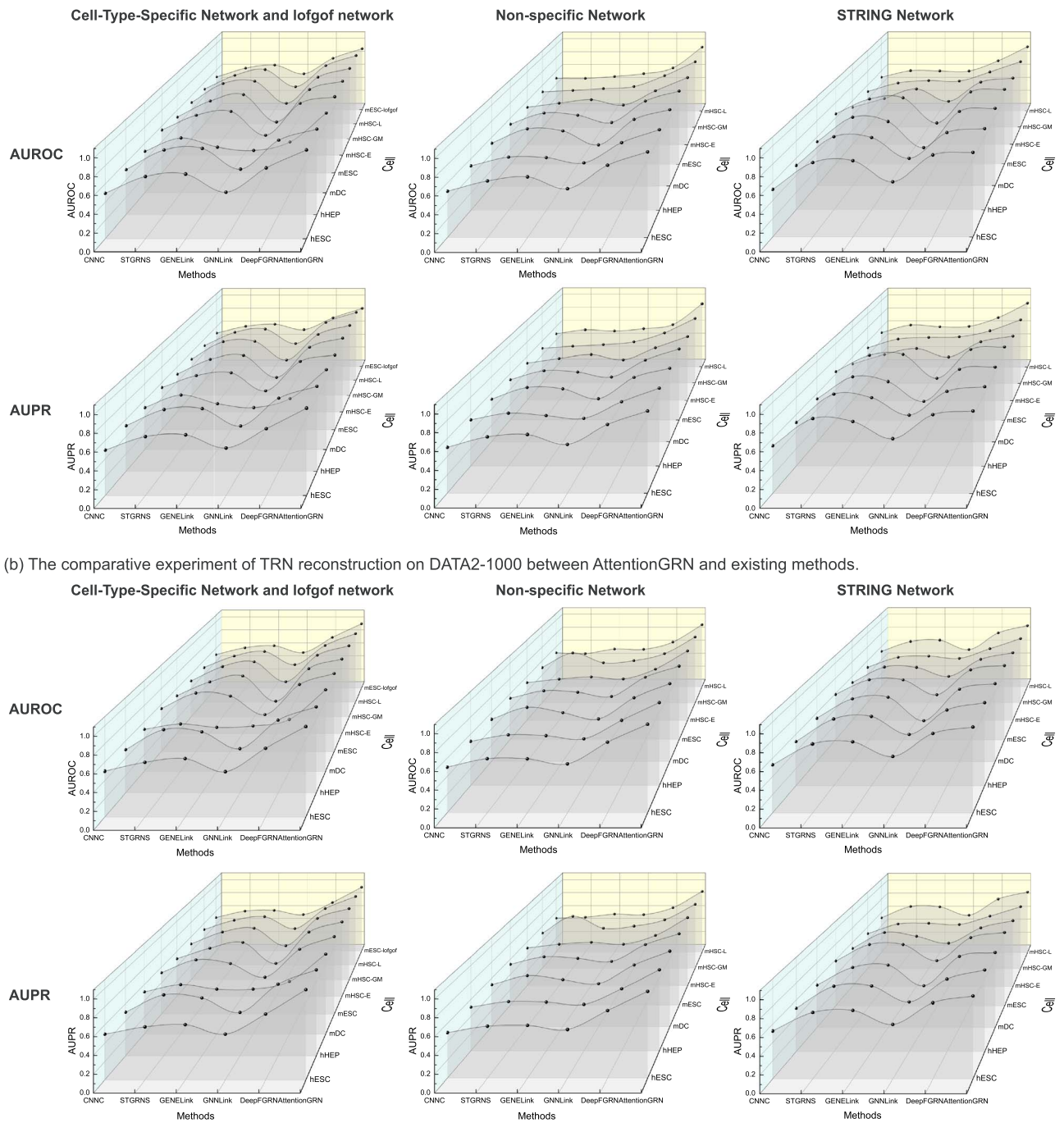

Figure 4. The results of AttentionGRN and other state-of-the-art methods for TRN inference. (a) and (b) show results for TRN inference on DATA2-500 and DATA2-1000. The rows represent AUROC and AUPR, and the columns correspond to cell type-specific networks and LOF/GOF networks, non-specific networks, and STRING networks. The waterfall charts display different methods along the x-axis, AUROC or AUPR on the y-axis, and cell types on the z-axis.

and directed network structure features is effective. DeepFGRN outperforms STGRNS and GENELink on many datasets by considering both gene expression and network structure features. (iv) AttentionGRN achieves superior performance because it integrates both gene expression features and functional, directed network structure features. This comprehensive approach enables the model to capture long-range dependencies and adapt to the diverse characteristics of TRNs, resulting in enhanced predictive accuracy across different network complexities and scales.

Time complexity analysis

We compared the runtime of AttentionGRN with other methods, and the experimental results are shown in Supplementary Tables S8–S9. GENELink has the fastest runtime while AttentionGRN has the longest runtime on GRN inference task. Although the runtime of AttentionGRN exceeds other methods, the prediction performance of AttentionGRN outperforms other methods, and the runtime is reasonable. The average runtime of AttentionGRN on all datasets in DATA1 for GRN inference is 536.0184 s, which

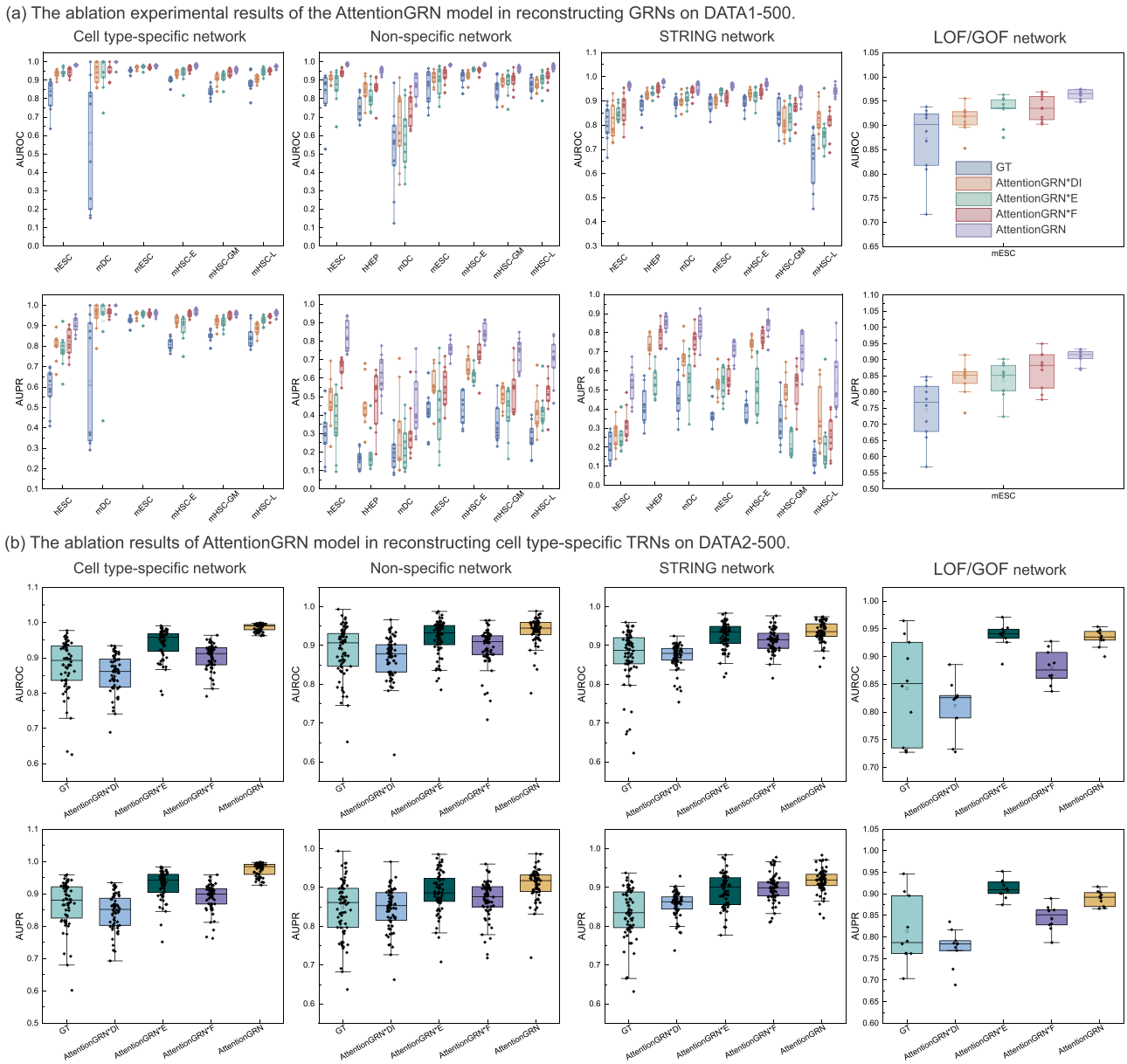

Figure 5. Ablation results of AttentionGRN for GRN and TRN inference. (a) presents GRN inference results for each cell in DATA1-500. (b) presents TRN inference results for all cells in DATA2-500. The rows represent AUROC and AUPR, and the columns correspond to different network types.

is less than 9 min. The TRN inference task exhibited the opposite situation compared to the GRN inference task. AttentionGRN has the fastest runtime while GENELink has the longest runtime. Their average runtimes on all datasets in DATA2 for TRN inference are 107.9796 and 902.8129 s, respectively. In summary, AttentionGRN demonstrated superior performance compared to state-of-the-art methods while maintaining a reasonable runtime, which is a critical factor for practical applications in the field. This achievement underscores the effectiveness of our approach in GRN inference tasks.

Ablation study

To assess the contribution of each module in GRN and TRN inference, we conducted an ablation study with five different models:

- GT: Graph Transformer without directed identity and functionally related genes.
- AttentionGRN∗DI: removing directed structural identity DI.
- AttentionGRN∗E: removing gene expression features.
- AttentionGRN∗F: removing functionally related genes.
- AttentionGRN: full model.

The results of these models on both GRN and TRN inference tasks are shown in Supplementary Figs S8–S9. Figure 5a–b present a subset of these results, showing the impact of different modules on performance.

Figure 5a illustrates the reconstruction of various GRNs, where GT consistently underperforms across all datasets, especially in cell type-specific GRN inference for mDCs. The incorporation of directed structure encoding (*DI*), gene expression features, and functionally related genes (*FN*) into AttentionGRN significantly boosts predictive accuracy, particularly for non-specific and STRING GRNs. Interestingly, GT already performs well on cell type-specific GRNs for certain cell types (e.g. mDCs, mESCs, mHSC-Es), likely because the GT has already captured robust

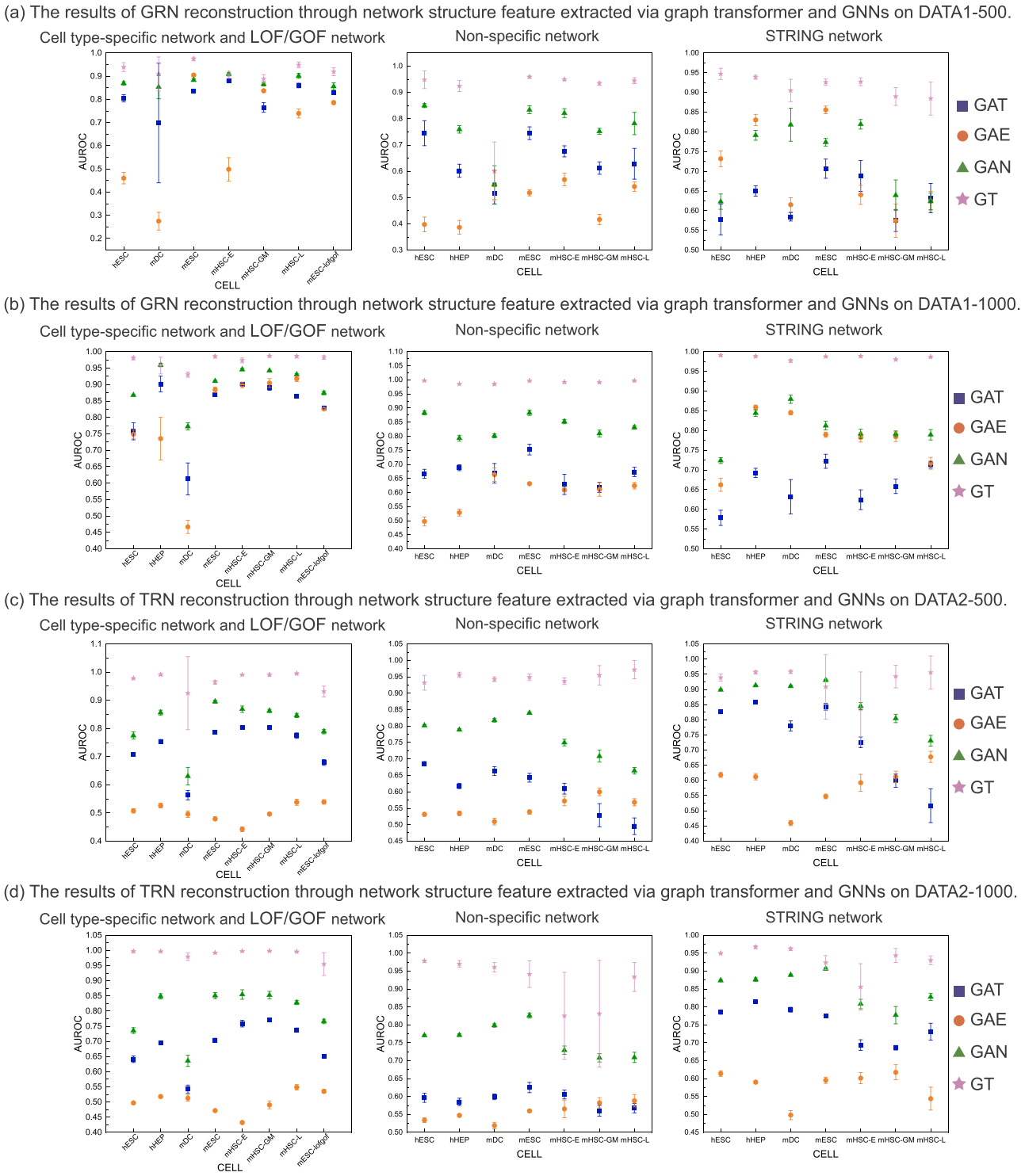

Figure 6. Comparison of graph embedding methods for GRN and TRN inference. (a) and (b) show GRN inference results on DATA1-500 and DATA1-1000. (c) and (d) show TRN inference results on DATA2-500 and DATA2-1000. The columns represent different network types.

network structural features (see next section for detailed analysis).

Figure 5b shows the reconstruction of TRNs for various cell types. While GT underperforms, AttentionGRN emerges as the superior model. Removing the directed structure encoding (*DI*) results in a significant drop in performance, emphasizing the importance of directionality in TRN inference. Although removing gene expression features slightly reduces AUROC and AUPR, the large standard deviation suggests that gene expression features improve model robustness. The performance also decreases when functionally related genes are omitted, highlighting the importance of incorporating functionally related neighbors for more accurate gene embeddings. In the case of LOF/GOF TRN inference, the best performance is observed when gene expression features are excluded, indicating that network structural features dominate in this network type.

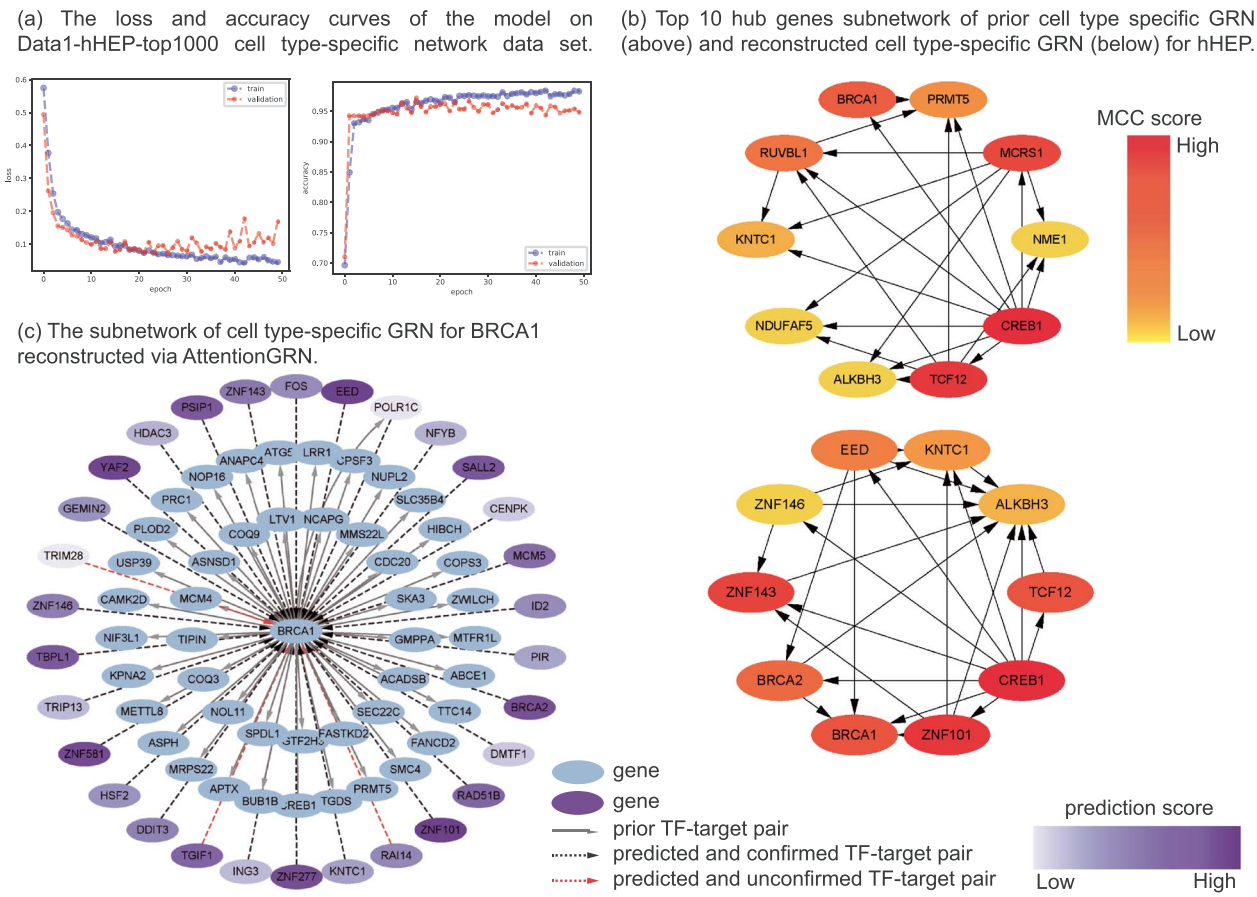

Figure 7. Cell type-specific GRN for hHEP reconstructed by AttentionGRN.

Graph transformer enables to achieve more effective network structure feature

As previously noted, network structure features are essential for both GRN and TRN inference. To analyze which graph embedding methods capture the most effective network structure features, we compared four methods in GRN inference: GAT (from GENELink), GAE (from GNNLink), GAN (from DeepFGRN), and GT-DI (GT with directed identity from AttentionGRN). Results shown in Fig. 6 demonstrate that predictive performance improves sequentially from GAE to GAT, to GAN, and finally to GT-DI. The GCN used in GAE tends to suffer from over-smoothing and over-squashing issues, which GT-DI mitigates effectively. GAT also faces similar challenges. GAN, which incorporates directed network features, achieves better AUROCs compared to GAE and GAT, indicating that directionality is a crucial factor in capturing meaningful network structure. GT-DI outperforms all other methods, illustrating that the GT, with its self-attention mechanism and directed structural encoding, captures the most effective network structural features.

AttentionGRN identifies novel hub genes in human mature hepatocytes

To further assess whether AttentionGRN can predict novel TF–target gene associations and generate functionally meaningful GRNs, we inferred a cell type-specific GRN for hHEP using AttentionGRN. The dataset for hHEP contains 63 TFs, 777 target genes, and 1038 known regulatory associations (see Supplementary Table S1). Following the procedure outlined in section Evaluation strategies, we generated 3880 training samples, consisting of both positive and negative pairs, to train the model. Additionally, 45 008 unknown TF–target pairs were used for prediction. The dataset was split into a 4:1 training-to-validation ratio. AttentionGRN was trained on the training set, and the best model, identified by the lowest loss on the validation set, was selected (Fig. 7). This model was then used to predict the 45 008 unknown TF–target pairs, resulting in 9600 potential TF–target gene pairs.

To assess the functional relevance of the predicted GRN, we performed a hub gene analysis. The top 1000 predicted TF–target pairs were integrated with the 1038 known regulatory associations, yielding a candidate cell type-specific GRN for hHEP. This GRN consisted of 63 TFs, 777 target genes, and 2038 regulatory associations. The 2038 associations were visualized in Cytoscape [36], and Matthews correlation coefficient scores were calculated using Cytohubba [37]. The top 10 hub genes were selected for further analysis (Fig. 7b).

A comparison of the predicted hub genes with those from the prior GRN revealed that several genes—CREB1, TCF12, BRCA1, KNTC1, and ALKBH3—were already present in the known GRN (Fig. 7b). Notably, novel hub genes identified by AttentionGRN included ZNF101, ZNF143, BRCA2, EED, and ZNF146. A subsequent literature review confirmed that these novel hub genes are strongly associated with liver diseases, such as liver fibrosis [38], liver cancer [39–48], chronic acute liver failure [49], non-alcoholic fatty liver disease [50–52], human alcoholic hepatitis, and non-alcoholic steatohepatitis [42], as detailed in Supplementary Table S5. These findings suggest that AttentionGRN can effectively infer biologically relevant GRNs from scRNA-seq data and

identify novel hub genes, providing a valuable resource for further biological research.

Furthermore, AttentionGRN can uncover novel TF–target regulatory associations. To illustrate this, we focused on the top 30 predicted associations for the master gene BRCA1 (Fig. 7c and Supplementary Table S6). Using the Harmonizome database [53], we found supporting evidence for 27 of the predicted associations. This confirms that AttentionGRN is capable of inferring novel TF–target regulatory interactions from scRNA-seq data, offering a powerful framework for experimental design and potentially saving biologists valuable time and resources.

Conclusion

The development of effective computational methods for GRN inference from scRNA-seq data is crucial for advancing our understanding of gene regulatory mechanisms and for supporting the development of disease treatments and biotechnologies. Building on current state-of-the-art approaches, this study introduces AttentionGRN, a novel model for both GRN and TRN inference, and provides preprocessing of 88 benchmark datasets for the broader research community. AttentionGRN integrates gene expression features with directed network structure features, leveraging a GT architecture to mitigate the over-smoothing issues commonly encountered in traditional GNNs. Additionally, the model incorporates directed structural encoding to guide the learning of asymmetric semantic information. By utilizing r-hop neighbors and functionally related genes, AttentionGRN effectively captures local network structures, long-range dependencies, and functional relationships between genes. The advantages of AttentionGRN are validated across 88 benchmark datasets, demonstrating its robustness and versatility.

In particular, experiments on cell type-specific GRN inference for hHEP highlight AttentionGRN's capability to predict novel TF-target gene pairs and identify promising hub genes. These findings underscore the potential of the model to discover new regulatory relationships and improve our understanding of gene expression control in complex biological systems. In general, AttentionGRN offers high prediction accuracy, strong generalizability, and valuable insights into the intricate regulatory networks that govern gene expression, making it a powerful tool for computational and biological research.

Nevertheless, some limitations remain. For instance, alternative methods such as community detection or clustering techniques could be explored to identify functionally related genes. Furthermore, while AttentionGRN excels in GRN and TRN inference within a single species, cross-species GRN inference remains an open challenge. Addressing this limitation will be a focus of our future work, as we aim to extend the model's applicability across species boundaries.

Key Points

- To the best of our knowledge, AttentionGRN is the first model to leverage GTs for GRN inference from scRNA-seq data, achieving state-of-the-art predictive performance.
- AttentionGRN introduces novel message aggregation strategies, specifically designed to capture directed network structures, functional information, and both short- and long-range dependencies within GRNs.
- Experimental results across 88 benchmark datasets demonstrate that AttentionGRN outperforms state-of-the-art methods in reconstructing cell type-specific GRNs and TRNs, highlighting its robustness and generalizability.
- Case studies, including an in-depth analysis of hHEP, reveal that AttentionGRN identifies novel hub genes and uncovers previously unknown regulatory associations.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work was supported by the grant of National Key Research and Development Program of China (No. 2020YFA0908700), the University Synergy Innovation Program of Anhui Province (No. GXXT-2021-039), the National Natural Science Foundation of China (No. 62202004, No. 62433001 and No. 62322301), and Anhui University outstanding youth research project (No. 2022AH020010).

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest

None declared.

Data availability

Raw datasets were downloaded from BEELINE (https://doi.org/-10.5281/zenodo.3378975), and the preprocessed dataset is presented in Supplementary Table S1 and on GitHub. All datasets and code are available at https://github.com/PhoebeGaoZhen/AttentionGRN/tree/master. A detailed reproduction tutorial is also provided on GitHub. Academic peers can reproduce the data preprocessing, AttentionGRN, and five comparison methods as outlined in the reproduction tutorial.

References

1. Huynh-Thu VA, Sanguinetti G. *Gene Regulatory Network Inference: An Introductory Survey, Volume 1883 of Methods in Molecular Biology*. New York, NY: Humana Press, 2019, 1–23. https://doi.org/10.1007/978-1-4939-8882-2_1.
2. Zhao M, He W, Tang J. *et al.* A comprehensive overview and critical evaluation of gene regulatory network inference technologies. *Brief Bioinform* 2021;**22**:bbab009. https://doi.org/10.1093/bib/bbab009
3. Badia-I-Mompel P, Wessels L, Mueller-Dott S. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 2023;**24**:739–54. https://doi.org/10.1038/s41576-023-00618-5
4. Aalto A, Viitasaari L, Ilmonen P. *et al.* Gene regulatory network inference from sparsely sampled noisy data. *Nat Commun* 2020;**11**:3493. https://doi.org/10.1038/s41467-020-17217-1
5. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022;**40**:1458–66. https://doi.org/10.1038/s41587-022-01284-4

6. Jovic D, Liang X, Zeng H. *et al.* Single-cell RNA sequencing technologies and applications: a brief overview. *Clin Transl Med* 2022;**12**:e694. https://doi.org/10.1002/ctm2.694
7. Li S, Liu Y, Shen L-C. *et al.* GMFGRN: a matrix factorization and graph neural network approach for gene regulatory network inference. *Brief Bioinform* 2024;**25**:bbad529. https://doi.org/10.1093/bib/bbad529
8. Cui W, Long Q, Xiao M. *et al.* Refining computational inference of gene regulatory networks: integrating knockout data within a multi-task framework. *Brief Bioinform* 2024;**25**:bbae361. https://doi.org/10.1093/bib/bbae361
9. Zhao J, Wong C-W, Ching W-K. *et al.* NG-SEM: an effective non-Gaussian structural equation modeling framework for gene regulatory network inference from single-cell RNA-seq data. *Brief Bioinform* 2023;**24**:bbad369. https://doi.org/10.1093/bib/bbad369
10. Shojaee A, Huang S-SC. Robust discovery of gene regulatory networks from single-cell gene expression data by causal inference using composition of transactions. *Brief Bioinform* 2023;**24**:bbad370. https://doi.org/10.1093/bib/bbad370
11. Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc Natl Acad Sci USA* 2019;**116**:27151–8. https://doi.org/10.1073/pnas.1911536116
12. Yuan Y, Bar-Joseph Z. Deep learning of gene relationships from single cell time-course expression data. *Brief Bioinform* 2021;**22**:bbab142. https://doi.org/10.1093/bib/bbab142
13. Chen J, Cheong CW, Lan L. *et al.* DeepDRIM: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell RNA-seq data. *Brief Bioinform* 2021;**22**:bbab325. https://doi.org/10.1093/bib/bbab325
14. Lin Z, Ou-Yang L. Inferring gene regulatory networks from single-cell gene expression data via deep multi-view contrastive learning. *Brief Bioinform* 2023;**24**:bbac586. https://doi.org/10.1093/bib/bbac586
15. Gan Y, Xin H, Zou G. *et al.* Inferring gene regulatory networks from single-cell transcriptomic data using bidirectional RNN. *Front Oncol* 2022;**12**:899825. https://doi.org/10.3389/fonc.2022.899825
16. Luo Q, Yongzhen Y, Lan X. SIGNET: single-cell RNA-seq-based gene regulatory network prediction using multiple-layer perceptron bagging. *Brief Bioinform* 2022;**23**:bbab547.
17. Zhao M, He W, Tang J. *et al.* A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. *Brief Bioinform* 2022;**23**:bbab568. https://doi.org/10.1093/bib/bbab568
18. Yang B, Bao W, Chen B. *et al.* Single_cell_GRN: gene regulatory network identification based on supervised learning method and single-cell RNA-seq data. *BioData Min* 2022;**15**:13. https://doi.org/10.1186/s13040-022-00297-8
19. Jing X, Zhang A, Liu F. *et al.* STGRNS: an interpretable transformer-based method for inferring gene regulatory networks from single-cell transcriptomic data. *Bioinformatics* 2023;**39**:btad165.
20. Karaaslanli A, Saha S, Aviyente S. *et al.* scSGL: kernelized signed graph learning for single-cell gene regulatory network inference. *Bioinformatics* 2022;**38**:3011–9. https://doi.org/10.1093/bioinformatics/btac288
21. Chen G, Liu Z-P. Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data. *Bioinformatics* 2022;**38**:4522–9. https://doi.org/10.1093/bioinformatics/btac559
22. Mao G, Pang Z, Zuo K. *et al.* Predicting gene regulatory links from single-cell RNA-seq data using graph neural networks. *Brief Bioinform* 2023;**24**:bbad414. https://doi.org/10.1093/bib/bbad414
23. Wei P-J, Guo Z, Gao Z. *et al.* Inference of gene regulatory networks based on directed graph convolutional networks. *Brief Bioinform* 2024;**25**:bbae309. https://doi.org/10.1093/bib/bbae309
24. Ma A, Wang X, Li J. *et al.* Single-cell biological network inference using a heterogeneous graph transformer. *Nat Commun* 2023;**14**:964. https://doi.org/10.1038/s41467-023-36559-0
25. Liu J, Zhou S, Ma J. *et al.* Graph attention network with convolutional layer for predicting gene regulations from single-cell ribonucleic acid sequence data. *Eng Appl Artif Intel* 2024;**136**:108938. https://doi.org/10.1016/j.engappai.2024.108938
26. Kreuzer D, Beaini D, Hamilton WL. *et al.* Rethinking graph transformers with spectral attention. In: Ranzato M, Beygelzimer A, Dauphin Y. *et al.* (eds.), Advances in Neural Information Processing Systems 34 (NEURIPS 2021). *35th Annual Conference on Neural Information Processing Systems (NeurIPS), ELECTR NETWORK, DEC 06-14, 2021*. La Jolla, CA, USA: Neural Information Processing Systems Foundation, Inc.; 2021:21618–29.
27. Gao Z, Yansen S, Xia J. *et al.* DeepFGRN: inference of gene regulatory network with regulation type based on directed graph embedding. *Brief Bioinform* 2024;**25**:bbae143. https://doi.org/10.1093/bib/bbae143
28. Gao Z, Tang J, Xia J. *et al.* CNNGRN: a convolutional neural network-based method for gene regulatory network inference from bulk time-series expression data. *IEEE/ACM Trans Comput Biol Bioinform* 2023;**20**:2853–61. https://doi.org/10.1109/TCBB.2023.3282212
29. Pratapa A, Jalihal AP, Law JN. *et al.* Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 2020;**17**:147–54. https://doi.org/10.1038/s41592-019-0690-6
30. Chu L-F, Leng N, Zhang J. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* 2016;**17**:173. https://doi.org/10.1186/s13059-016-1033-x
31. Gray Camp J, Sekine K, Gerber T. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* 2017;**546**:533–8. https://doi.org/10.1038/nature22796
32. Shalek AK, Satija R, Shuga J. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 2014;**510**:363–9. https://doi.org/10.1038/nature13437
33. Hayashi T, Ozaki H, Sasagawa Y. *et al.* Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat Commun* 2018;**9**:619. https://doi.org/10.1038/s41467-018-02866-0
34. Nestorowa S, Hamey FK, Sala BP. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* 2016;**128**:E20–31. https://doi.org/10.1182/blood-2016-05-716480
35. Shu H, Zhou J, Lian Q. *et al.* Modeling gene regulatory networks using neural network architectures. *Nat Comput Sci* 2021;**1**:491–501. https://doi.org/10.1038/s43588-021-00099-8
36. Shannon P, Markiel A, Ozier O. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504. https://doi.org/10.1101/gr.1239303
37. Chin C-H, Chen S-H, Wu HH *et al.* CytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol* 2014;**8 Suppl 4**:S11.
38. Li G, Jiang Q, Keshu X. CREB family: a significant role in liver fibrosis. *Biochimie* 2019;**163**:94–100. https://doi.org/10.1016/j.biochi.2019.05.014

39. Liu X, Zhang H, Zhou P. *et al.* CREB1 acts via the miR-922/ARID2 axis to enhance malignant behavior of liver cancer cells. *Oncol Rep* 2021;**45**:79. https://doi.org/10.3892/or.2021.8030
40. Ye B, Shen W, Zhang C. *et al.* The role of ZNF143 overexpression in rat liver cell proliferation. *BMC Genomics* 2022;**23**:483. https://doi.org/10.1186/s12864-022-08714-2
41. Wang G-H, Zhao C-M, Huang Y. *et al.* BRCA1 and BRCA2 expression patterns and prognostic significance in digestive system cancers. *Hum Pathol* 2018;**71**:135–44. https://doi.org/10.1016/j.humpath.2017.10.032
42. Liao G, French B, Liu H. *et al.* Upregulation of BRCA1 and BRCA2 in human alcoholic hepatitis and nonalcoholic steatohepatitis. *Am J Clin Pathol* 2015;**144**:A379. https://doi.org/10.1093/ajcp/144.suppl2.379
43. Yang J, Zhang L, Jiang Z. *et al.* TCF12 promotes the tumorigenesis and metastasis of hepatocellular carcinoma via upregulation of CXCR4 expression. *Theranostics* 2019;**9**:5810–27. https://doi.org/10.7150/thno.34973
44. Zhang JYJ, Zhang J, Kiffe M. *et al.* Preclinical pharmacokinetics and metabolism of MAK683, a clinical stage selective oral embryonic ectoderm development (EED) inhibitor for cancer treatment. *Xenobiotica* 2022;**52**:65–78. https://doi.org/10.1080/00498254.2021.2005852
45. Yang F, Yuan C. KNTC1 knockdown inhibits proliferation and metastases of liver cancer. *3 Biotech* 2023;**13**:309. https://doi.org/10.1007/s13205-023-03722-9
46. Tong H, Liu X, Peng C. *et al.* Silencing of KNTC1 inhibits hepatocellular carcinoma cells progression via suppressing PI3K/Akt pathway. *Cell Signal* 2023;**101**:110498. https://doi.org/10.1016/j.cellsig.2022.110498
47. Wang Q, Wang G, Wang Y. *et al.* Association of AlkB homolog 3 expression with tumor recurrence and unfavorable prognosis in hepatocellular carcinoma. *J Gastroenterol Hepatol* 2018;**33**:1617–25. https://doi.org/10.1111/jgh.14117
48. Bao L, Wang M, Fan Q. Hsa:circ_NOTCH3 regulates ZNF146 through sponge adsorption of miR-875-5p to promote tumorigenesis of hepatocellular carcinoma. *J Gastrointest Oncol* 2021;**12**:2388–402. https://doi.org/10.21037/jgo-21-567
49. Chen J, Zhang Q, Wenxiong X. *et al.* Baicalein upregulates macrophage TREM2 expression via TrKB-CREB1 pathway to attenuate acute inflammatory injury in acute-on-chronic liver failure. *Int Immunopharmacol* 2024;**139**:112685. https://doi.org/10.1016/j.intimp.2024.112685
50. Ho B, Thompson A, Jorgensen A. *et al.* Genome wide analysis identifies potential mechanisms of non-alcoholic fatty liver disease. *Gut* 2022;**71**:A14–5. Annual Meeting of the British-Society-of-Gastroenterology (BSG), Birmingham, ENGLAND, JUN 20-23, 2022.
51. Ge J, Bai Y, Tang B. *et al.* The gene signature associated with hepatocellular carcinoma in patients with nonalcoholic fatty liver disease. *J Oncol* 2021;**2021**:1–9. https://doi.org/10.1155/2021/6630535
52. Dong Y, Minjie H, Tan K. *et al.* ZNF143 inhibits hepatocyte mitophagy and promotes non-alcoholic fatty liver disease by targeting increased lncRNA NEAT1 expression to activate ROCK2 pathway. *Epigenetics* 2023;**18**:2239592. https://doi.org/10.1080/15592294.2023.2239592
53. Rouillard AD, Gundersen GW, Fernandez NF. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016;**2016**:baw100.

Briefings in Bioinformatics, 2025, **26(2)**, bbaf118
https://doi.org/10.1093/bib/bbaf118
Problem Solving Protocol