BMC Bioinformatics

RESEARCH Open Access

GRLGRN: graph representation-based learning to infer gene regulatory networks from single-cell RNA-seq data

Kai Wang[1], Yulong Li[1], Fei Liu[1], Xiaoli Luan[1], Xinglong Wang[2,3,4,5] and Jingwen Zhou[2,3,4,5*]

*Correspondence:
zhoujw1982@jiangnan.edu.cn

[1] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, 1800 Lihu Road, Wuxi 214122, Jiangsu, China
[2] Science Center for Future Foods, Jiangnan University, 1800 Lihu Road, Wuxi 214122, Jiangsu, China
[3] Key Laboratory of Industrial Biotechnology, Ministry of Education and School of Biotechnology, Jiangnan University, 1800 Lihu Road, Wuxi 214122, Jiangsu, China
[4] Engineering Research Center of Ministry of Education on Food Synthetic Biotechnology, Jiangnan University, 1800 Lihu Road, Wuxi 214122, Jiangsu, China
[5] Jiangsu Province Engineering Research Center of Food Synthetic Biotechnology, Jiangnan University, 1800 Lihu Road, Wuxi 214122, Jiangsu, China

Abstract

Background: A gene regulatory network (GRN) is a graph-level representation that describes the regulatory relationships between transcription factors and target genes in cells. The reconstruction of GRNs can help investigate cellular dynamics, drug design, and metabolic systems, and the rapid development of single-cell RNA sequencing (scRNA-seq) technology provides important opportunities while posing significant challenges for reconstructing GRNs. A number of methods for inferring GRNs have been proposed in recent years based on traditional machine learning and deep learning algorithms. However, inferring the GRN from scRNA-seq data remains challenging owing to cellular heterogeneity, measurement noise, and data dropout.

Results: In this study, we propose a deep learning model called graph representational learning GRN (GRLGRN) to infer the latent regulatory dependencies between genes based on a prior GRN and data on the profiles of single-cell gene expressions. GRLGRN uses a graph transformer network to extract implicit links from the prior GRN, and encodes the features of genes by using both an adjacency matrix of implicit links and a matrix of the profile of gene expression. Moreover, it uses attention mechanisms to improve feature extraction, and feeds the refined gene embeddings into an output module to infer gene regulatory relationships. To evaluate the performance of GRLGRN, we compared it with prevalent models and performed ablation experiments on seven cell-line datasets with three ground-truth networks. The results showed that GRLGRN achieved the best predictions in AUROC and AUPRC on 78.6% and 80.9% of the datasets, and achieved an average improvement of 7.3% in AUROC and 30.7% in AUPRC. The interpretation discussion and the network visualization were conducted.

Conclusions: The experimental results and case studies illustrate the considerable performance of GRLGRN in predicting gene interactions and provide interpretability for the prediction tasks, such as identifying hub genes in the network and uncovering implicit links.

Keywords: Gene regulatory network, Gene expression data, Graph representation learning, Implicit links

Background

The regulation of gene expression determines cell identity and function during the development of an organism. The expression of specific genes in distinct cells leads to the formation of different types of cells, where transcription factors (TFs) generated by the regulatory genes bind to specific regions of the target genes to regulate the levels of gene expression [1, 2]. A gene regulatory network (GRN) is a graph-level representation that describes the TFs and target genes in cells, where each node represents a gene and each edge represents the regulatory relationship between genes [3]. Previous research has shown that GRNs can provide important insights into cellular dynamics [4–6], the development of target drugs [7–9], and the formulation of models of metabolic systems and their optimization [10–12]. Therefore, it is theoretically and practically valuable to investigate the reconstruction of GRNs, as this can provide accurate insights into cellular phenotypes from the genomic perspective.

Experiments on molecular interactions involving microarrays [13] and chromatin immunoprecipitation sequencing (ChIP-seq) [14] have been used to investigate gene regulatory relationships in the past few years, and can contribute to reconstructing GRNs. However, the relevant methods are typically time consuming, labor intensive, and highly dependent on the experimental conditions. With the development of single-cell high-throughput sequencing technology, researchers can now obtain a large amount of data on various types of omics at the cellular level. A number of methods for inferring GRNs based on traditional machine learning technologies have been developed to reconstruct GRNs, including GENIE3 [15] and GRNBoost2 [16]. Moreover, many deep learning-based approaches have been applied to enhance the quality of the resulting inferences. CNNC [17] and DeepDRIM [18] can be used to convert the data on gene expression into images and leverage convolutional neural networks (CNNs) to infer GRNs. STGRNS [19] makes use of a BERT-based model to fully extract the features of data on gene expression to infer GRNs through transfer learning. Further improving the quality of such inferences requires leveraging their known regulatory relationships and using the profiles of gene expression from scRNA-seq datasets. Recent research [20–22] has proposed several methods to infer GRNs that consider not only the data on gene expressions, but also prior topological information on the GRNs. For instance, PMF-GRN [23] employs the probabilistic matrix factorization and variational inference as core methods to capture the inferred transcription factor activity and latent regulatory relationships in GRNs. Compared with other benchmark models, PMF-GRN demonstrates superior inference performance and reliable uncertainty estimation capabilities. VMPLN [24] is based on the mixture Poisson-lognormal (MPLN) model to infer GRNs from count data of mixed populations. In addition to achieving strong predictive performance, VMPLN demonstrates the reliability of inferring results from highly mixed multi-cell type data. These methods can provide novel solutions for reconstructing GRNs using graph-based information. The CNNGRN [25] incorporates prior link-related information into the CNN to extract the features of genes, GCNG [26] and GNNLINK [27] use graph convolutional networks (GCNs) [28] to obtain low-dimensional embeddings of genes, while GENELINK [29] uses graph attention neural networks [30] to obtain gene representations. Due to the sparsity and heterogeneity of the graphs of GRNs, directly using the explicit link-related information contained in them to obtain the gene embeddings fails

to fully exploit the topological features of the prior GRNs [31]. Advanced approaches to learning graph representations need to be applied to fully learn the implicit links between genes.

In this study, we propose a deep learning model for reconstructing GRNs based on graph representation learning, called GRLGRN. It is designed to infer the latent regulatory dependencies between genes according to a prior GRN and data on the profiles of single-cell gene expressions. GRLGRN leverages a graph transformer network [32] in its gene embedding module to extract implicit links from the graph of the prior GRN, and to further encode the features of the genes from an adjacency matrix of implicit links and the corresponding matrix of the profile of gene expression. Furthermore, a convolutional block attention module (CBAM) [33] is used to enhance feature extraction, and the refined gene embeddings are fed into an output module to infer the gene regulatory relationships. Moreover, a regularization term of graph contrastive learning [34, 35] is introduced when optimizing the loss function during the training of the model, with the aim of preventing model over-fitting owing to the excessive smoothing of the features of genes. We compared the inferences obtained by the GRLGRN with those of six models on benchmark datasets across seven cell lines, and corresponding to three types of ground-truth networks. The results showed that GRLGRN outperformed all other models on most datasets. We also conducted ablation experiments to verify the effectiveness of its modules and compared the computational costs of GRLGRN and its variants. Furthermore, we verified that GRLGRN improved model inference accuracy by combining graph contrastive learning and the automatic weighted loss training techniques [42]. In addition, GRLGRN can visualize the results to provide insights into the GRNs and to reveal the contributions of all the explicit and implicit links to the representation.

The key contributions of our research are as follows: (i) We first leverage a graph transformer network to extract implicit links from a prior GCN, and then generate a GCN to obtain the gene embeddings. (ii) We use a CBAM to refine the features of the genes. (iii) We introduce graph contrastive learning to reduce the excessive smoothing of the features of genes by GRLGRN.

Methods

Benchmark datasets

The BEELINE database [36] comprises scRNA-seq data from seven types of cell lines: (i) human embryonic stem cells (hESCs), (ii) human mature hepatocytes (hHEPs), (iii) mouse dendritic cells (mDCs), (iv) mouse embryonic stem cells (mESCs), (v) mouse hematopoietic stem cells with erythroid lineage (mHSC-E), (vi) mouse hematopoietic stem cells with granulocyte-monocyte lineage (mHSC-GM), and (vii) mouse hematopoietic stem cells with lymphoid lineage (mHSC-L). Each cell line corresponded to three ground-truth networks with varying densities that have been documented in STRING [37], cell type-specific ChIP-seq [38], and non-specific ChIP-seq [39]. In addition, the 500 and 1000 gene sets that varied by the largest magnitude for each ground-truth network of a cell line, including TFs with a corrected P-value of less than 0.01, were identified to generate GRNs at two scales. We considered a total of 42 benchmark scRNA-seq datasets, each of which comprised a ground-truth GRN and a matrix of the profile of

gene expressions describing their intensity. Detailed information on the benchmark datasets is listed in Supplementary Table S1, Additional file 1.

Proposed framework

The architecture of the proposed model of GRN inference, called GRLGRN, is shown in Fig. 1. It consists of a gene embedding module, a feature enhancement module, and an output module. The main task of GRLGRN is to infer the potential regulatory dependencies between genes from a prior GRN graph and data on the profile of gene expressions.

Gene embedding module

The gene embedding module is shown in Fig. 1A. It uses a graph transformer network [32] to extract implicit links in a graph transformer layer and obtain gene representations in a subsequent GCN layer.

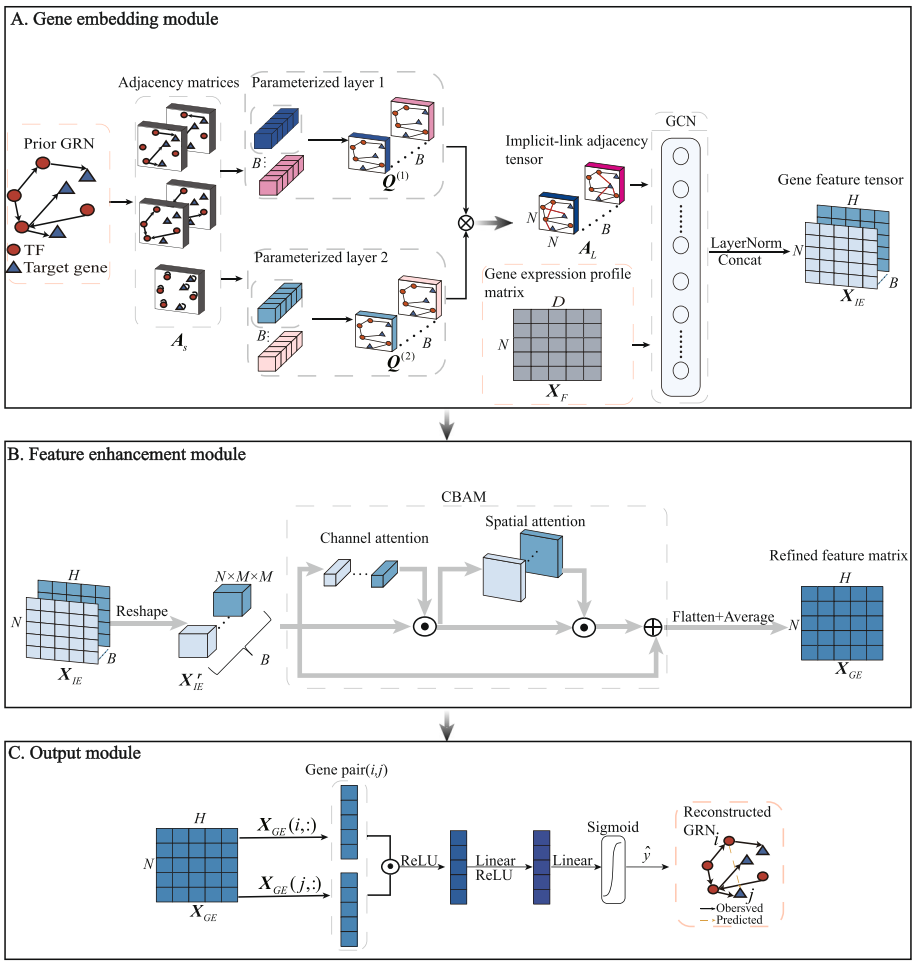

Fig. 1 Overall architecture of GRLGRN. **A** Gene embedding module. We extracted implicit links by using a graph transformer network and obtained gene embeddings by using a GCN. **B** Feature enhancement module. We leveraged channel and spatial attention mechanisms to obtain a refined feature matrix $\boldsymbol{X}_{GE}$. **C** Output module. We computed the Hadamard product of a given pair of genes and fed it into a two-layer perceptron to obtain the scores of inferences of gene regulatory interactions

We can use the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which describes any prior GRN, to formulate five graphs: the directed subgraph $\mathcal{G}_1$, the edges of which represent the regulatory relationships from the TFs to the target genes in $\mathcal{G}$, the directed graph $\mathcal{G}_2$, the edges of which are in the opposite directions to those of $\mathcal{G}_1$, the directed subgraph $\mathcal{G}_3$, the edges of which represent TF–TF regulatory relationships in $\mathcal{G}$, the directed graph $\mathcal{G}_4$, the edges of which are in the opposite directions to those of $\mathcal{G}_3$, and the self-connected gene graph $\mathcal{G}_5$. We then concatenate the adjacency matrices of these five graphs $\boldsymbol{A}_s \in \{0,1\}^{5\times N\times N}$, where N is the number of genes, i.e., $N = |\mathcal{V}|$. Furthermore, individually passing through two parallel, parameterized layers yields the two tensors $\boldsymbol{Q}^{(1)}$ and $\boldsymbol{Q}^{(2)} \in \mathbb{R}^{B\times N\times N}$. This process satisfies

$$\boldsymbol{Q}^{(i)}(j,:,:) = \sum_{k=1}^{5} \varepsilon_{k,j}^{(i)} \boldsymbol{A}_s(k,:,:), \tag{1}$$

where $i = 1, 2$ and $j = 1, 2, ..., B$, $\boldsymbol{Q}^{(i)}(j,:,:)$ denotes the j-th channel matrix of $\boldsymbol{Q}^{(i)}$, $\boldsymbol{A}_s(k,:,:)$ denotes the adjacency matrix of $\mathcal{G}_k$, B denotes the number of output channels of each parameterized layer, and $\varepsilon_{k,j}^{(i)}$ denotes the training parameter $\gamma_{k,j}^{(i)}$ after normalization in the i-th layer:

$$\varepsilon_{k,j}^{(i)} = \frac{\exp(\gamma_{k,j}^{(i)})}{\sum_{k=1}^{5} \exp(\gamma_{k,j}^{(i)})}. \tag{2}$$

Then, the inner product of $\boldsymbol{Q}^{(1)}$ and $\boldsymbol{Q}^{(2)}$ in each channel yields a tensor $\boldsymbol{A}_L \in \mathbb{R}^{B\times N\times N}$. The above process can be expressed as

$$\boldsymbol{A}_L(j,:,:) = \boldsymbol{Q}^{(1)}(j,:,:) \cdot \boldsymbol{Q}^{(2)}(j,:,:), \tag{3}$$

where $j = 1, 2, ..., B$. $\boldsymbol{A}_L(j,:,:) \in \mathbb{R}^{N\times N}$ is the adjacency matrix of the j-th generated two-length meta-path graph, and is called the adjacency matrix of the implicit link. As discussed in [32], the concentrated tensor $\boldsymbol{A}_L$ contains information on the adjacency of B meta-path graphs with at most two lengths, comprising explicit links (one length) and implicit links (two lengths).

By applying a parameter-shared GCN to each channel of an adjacency tensor of the implicit link $\boldsymbol{A}_L$ and the matrix of the profile of gene expression $\boldsymbol{X}_F \in \mathbb{R}^{N\times D}$ for N genes, we can implement gene embeddings to obtain a tensor of the embeddings of the gene $\boldsymbol{X}_{IE} \in \mathbb{R}^{B\times N\times H}$ that satisfies

$$\boldsymbol{X}_{IE}(j,:,:) = \mathrm{LayerNorm}(\mathrm{ReLU}(\tilde{\boldsymbol{D}}_{\mathrm{j}}^{-1}\tilde{\boldsymbol{A}}_{\mathrm{L,j}}\boldsymbol{X}_{\mathrm{F}}\boldsymbol{W})), \tag{4}$$

where $j = 1, 2, ..., B$, H denotes the number of dimensions of the gene embedding, $\boldsymbol{X}_{IE}(j,:,:)$ denotes the matrix of the embedding of the j-th gene in $\boldsymbol{X}_{IE}$, $\tilde{\boldsymbol{A}}_{L,j} := \boldsymbol{A}_L(j,:,:) + \boldsymbol{I}$, $\boldsymbol{I}$ denotes an identity matrix, $\tilde{\boldsymbol{D}}_j$ denotes the degree matrix of $\tilde{\boldsymbol{A}}_{L,j}$, and $\boldsymbol{W}$ denotes the training parameters of the GCN.

Feature enhancement module

The feature enhancement module is shown in Fig. 1B. It is used to further refine the gene representations by leveraging channel and spatial attention

mechanisms. Assuming that the square root of H is M, we can convert the tensor of gene embeddings $\boldsymbol{X}_{IE}$ into $\boldsymbol{X}_{IE}^{r} \in \mathbb{R}^{B\times N\times M\times M}$ by reshaping $\boldsymbol{X}_{IE}(i,j,:) \in \mathbb{R}^{H}$ into $\boldsymbol{X}_{IE}^{r}(i,j,:,:) \in \mathbb{R}^{M\times M}, i = 1,2,...,B, j = 1,2,...,N$. Furthermore, we pass B reshaped tensors $\boldsymbol{X}_{IE}^{r}(i,:,:,:) \in \mathbb{R}^{N\times M\times M}$, $i = 1...B$ through a parameter-sharing CBAM that consists of a channel attention module and a spatial attention module. Finally, we flatten each resulting tensor into a matrix and take the average of B flattened matrices to obtain the refined feature matrix $\boldsymbol{X}_{GE} \in \mathbb{R}^{N\times H}$. It satisfies the following:

$$\boldsymbol{X}_{GE} = \frac{1}{B}\sum_{i=1}^{B} \mathrm{Flatten}(\mathrm{CBAM}(\boldsymbol{X}_{\mathrm{IE}}^{\mathrm{r}}(\mathrm{i},:,:,:))), \tag{5}$$

where the details of the CBAM are provided in Supplementary Note S1, Additional file 1.

Output module

The output module is shown in Fig. 1C. It provides the results of inference of regulatory associations between genes according to the refined feature matrix $\boldsymbol{X}_{GE}$. The refined feature vectors of any gene pair (i, j) are represented by $\boldsymbol{u}_i := \boldsymbol{X}_{GE}(i,:)$ and $\boldsymbol{u}_j := \boldsymbol{X}_{GE}(j,:)$, where $1 \le i < j \le N$. We take the Hadamard product of these two vectors, apply an ReLU activation function to them, and pass them through a two-layer perceptron with activation functions ReLU and Sigmoid. This yields the output score $\hat{y} \in [0,1]$ of inference of the regulatory association between the genes i and j, and satisfies

$$\hat{y} = \mathrm{Sigmoid}(\mathrm{Linear}(\mathrm{ReLU}(\mathrm{Linear}(\mathrm{ReLU}(\boldsymbol{u}_{\mathrm{i}} \odot \boldsymbol{u}_{\mathrm{j}}))))). \tag{6}$$

For causal inference of potential regulatory relationships in GRNs, the output module is slightly modified as shown in Fig. 2 $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are concatenated fed into the subsequent layers to yields the output score $\hat{p} \in [0,1]$ of inferring the causal regulatory from gene i to gene j. the process can be expressed as

$$\boldsymbol{p} = \mathrm{Softmax}(\mathrm{Linear}(\mathrm{ReLU}(\mathrm{Linear}(\mathrm{ReLU}(\mathrm{Concat}(\boldsymbol{u}_{\mathrm{i}}, \boldsymbol{u}_{\mathrm{j}})))))). \tag{7}$$

When prior information about the transcription factor and target gene in the input gene pair (i, j) is available, the number of output neurons is two, and the task is to determine whether there is an interaction between the gene pair. When this prior information is

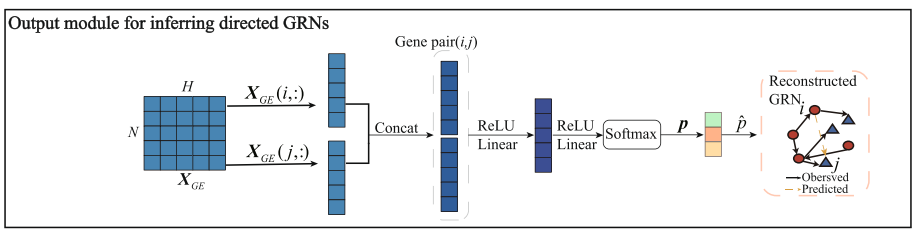

Fig. 2 Output module that can provide inference results of directed regulatory relationships, where the representations of genes i and j are combined by the concatenation instead of the Hadamard product in Fig. 1

unknown, the number of output neurons is three, and the task involves determining: i) whether there is an interaction between the gene pair, ii) whether gene *i* regulates gene *j*, and iii) whether gene *j* regulates gene *i*.

Model training

As shown in Fig. 3, we regarded any gene pair whose regulatory dependency appeared in the ground-truth GRN of a dataset as a positive sample, and labeled it as 1. Conversely, we regarded any gene pair whose regulatory dependency was not in the ground-truth GRN of a dataset as a negative sample, and labeled it as 0 [40]. We subsequently randomly chose positive and negative samples to form training, validation, and test sets according to a certain ratio. Specifically, to ensure the proportion of positive and negative samples for each TF in the partitioned datasets to be consistent with the benchmark dataset, we randomly assigned 66% of the positive and negative samples of each TF to the training set. Larger size of training set can ensure that the model could effectively learn and extract features related to gene pair interactions. Furthermore, we used a smaller proportion, 3.4%, of positive and negative samples for the validation set to evaluate hyperparameter selections. The remaining 30.6% of gene pair samples were randomly assigned to the test set, and model performances were evaluated on this relatively larger test set to ensure that the performance metrics were robust. To prevent information leakage, we made sure that the prior GRN input to the model during training consisted only of gene pairs that interacted in the training set.

We applied binary cross-entropy as the loss function when training the model to calculate the loss between the actual values of the labels and their predicted values:

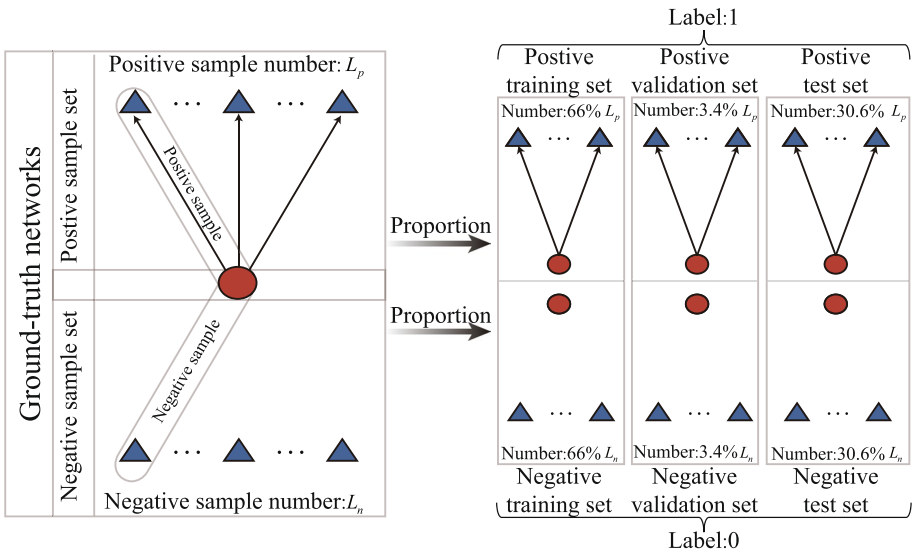

Fig. 3 Partition of the dataset. Gene pairs with observed interactions in the ground-truth network are regarded as positive samples, labeled as 1; conversely, any gene pair whose regulatory dependency is not present in the ground-truth GRN of a dataset is considered a negative sample, labeled as 0. The positive and negative samples for each TF are randomly selected into training, validation, and test subsets according to a specific ratio

$$\text{Loss}_{bc}(y,\hat{y}) = -\frac{1}{n}\sum_{s=1}^{n}\left(y_s \log\left(\hat{y}_s\right) + \left(1-y_s\right)\log\left(1-\hat{y}_s\right)\right), \tag{8}$$

where y_s denotes the value of the label of any given gene pair, $\hat{y}_s$ denotes its value predicted by GRLGRN (Eqs. (1)–(6)), and n denotes the number of gene pairs in a training batch.

Moreover, as shown in Fig. 4, we applied graph contrastive learning to GRNs provided by non-specific CHIP-seq [41] to prevent model over-fitting owing to the excessive smoothing of the features of the genes. The gene representations $\boldsymbol{X}_{EE} \in \mathbb{R}^{N\times H}$ corresponding to information on explicit links were contained in a GCN that shared the same parameters with the gene embedding module. By contrasting the gene embeddings of $\boldsymbol{X}_{EE}$ and $\boldsymbol{X}_{GE}$, the regular term N_{gc} of graph contrastive learning could then be calculated as

$$N_{gc} = -\frac{1}{2N}\sum_{m=1}^{N}\left(P(\boldsymbol{u}_m,\boldsymbol{v}_m) + P(\boldsymbol{v}_m,\boldsymbol{u}_m)\right), \tag{9}$$

where $\boldsymbol{u}_m := \boldsymbol{X}_{GE}(m,:)$, $\boldsymbol{v}_m := \boldsymbol{X}_{EE}(m,:)$, N is the number of genes, and the value of the gene pair $P(\boldsymbol{u}_m,\boldsymbol{v}_m)$ is defined in Supplementary Note S2, Additional file 1. Then, the loss function containing the regularization term for graph contrast learning can be defined as

$$\text{Loss} = \alpha\text{Loss}_{bc} + \beta N_{gc} + b, \tag{10}$$

where $\alpha = \frac{1}{2\sigma_1^2}$ and $\beta = \frac{1}{2\sigma_2^2}$ are weight coefficients, σ_1 and σ_2 are trainable parameters, and $b = \text{Log}\sigma_1\sigma_2$ is used as a loss function bias term. During the training process, the automatic weighted loss approach proposed in [42] was used to dynamically balance two different loss function terms in Eq. 10 and to enhance the stability of model parameter update. This addressed the limitation of manually setting weight coefficients based on empirical values when dealing with datasets of varying scales, thereby improving the model applicability across different scenarios.

The GRLGRN inference model can be trained and applied on PyTorch version 2.1. We set the hyperparameters corresponding to different ground-truth networks in Supplementary Table S2, Additional File 1. As shown in Fig. 5, we compared the model

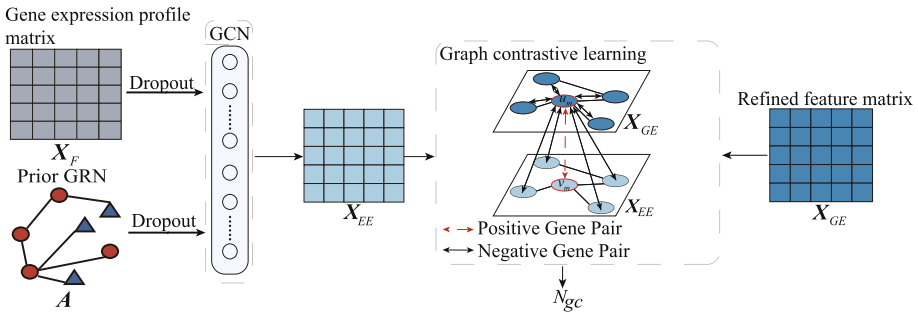

Fig. 4 Framework of graph contrastive learning. During training, the regularization term N_{gc} is calculated by contrasting two matrices of gene features $\boldsymbol{X}_{GE}$ and $\boldsymbol{X}_{EE}$, and the loss function containing N_{gc} is optimized for the GRNs.

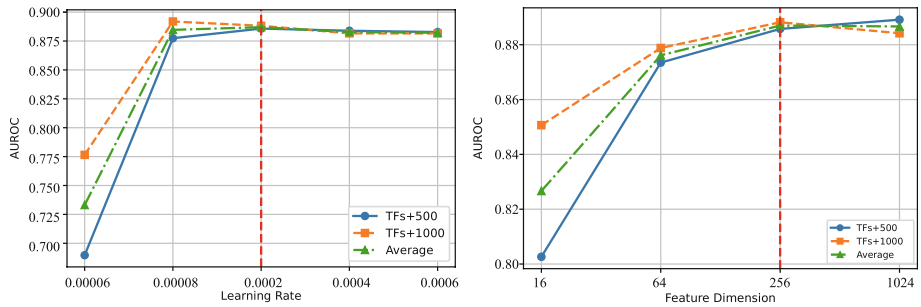

Fig. 5 AUROC metric values of GRLGRN under different learning rates and feature dimensions

performance of AUROC scores under different learning rates and feature dimensions (AUPRC scores can be found in Figure S1 of Additional file1). Therefore, during the training process, we set the batch size to 1024, with the learning rate and feature dimension set to 0.0002 and 256, respectively. We used the Adam optimizer to update GRLGRN parameters iteratively according to the gradient descent strategy. All experiments were conducted on a computer equipped with an Intel(R) Xeon(R) Gold 6226R CPU, 256GB of RAM, and three NVIDIA GTX 4090 GPUs.

Results

Performance on benchmark datasets

To evaluate the performance of the proposed GRLGRN, we compared it with six models (GNNLINK, GENELINK, STGRNS, GNE, GRNBoost2, and GENIE3) across seven cell lines under three ground-truth networks provided by STRING, cell type-specific ChIP-seq, and non-specific CHIP-seq. GNNLINK and GENELINK use GNNs to directly encode the features of genes from their explicit links (original edges), STGRNS uses a BERT-based encoder to learn the features of genes only from the data on gene expression, GNE [43] uses the topological structure of the GRN and data on gene expression to learn the low-dimensional features of genes by using a multi-layer perceptron (MLP), while GRNBoost2 and GENIE3 are built based on the random forest algorithm.

The results of the performances of all models in terms of inferring the GRNs on 42 benchmark datasets are shown in Fig. 6 (AUROC metric) and Supplementary Figure S2, Additional file 1 (AUPRC metric). The scRNA-seq datasets contained data on seven types of cell lines and three types of ground-truth networks, with TFs+ 500 and TFs+ 1000. As shown in Table 1, the proposed GRLGRN achieved the best predictive performance, in terms of the average AUROC and AUPRC, on 78.6% (33 of 42) and 80.9% (34 of 42) of the benchmark datasets, with average improvements of 7.3% and 30.7%, respectively. Figure 6A shows that compared with the second-best model (CNNLINK or GENELINK) on the AUROC metric across datasets, the proposed GRLGRN yielded average improvements of approximately 6.7% (12 of 14), 3.6% (8 of 14), and 6.35% (13 of 14) across the three ground-truth networks. It yielded average improvements of at least approximately 8.8% over the other models across the three ground-truth networks in terms of AUPRC scores. The box plots in Fig. 6B and Supplementary Figure S2B, Additional file 1 respectively show the distributions of the AUROC and AUPRC scores of GRLGRN and the other models on the benchmark datasets of ground-truth networks

Fig. 6 **A** The performance of the proposed GRLGRN and six prevalent models in terms of inferring GRNs on all benchmark datasets of ground-truth networks with TFs and 500 most significantly varying genes (top), and ground-truth networks with TFs and 1000 most significantly varying genes (bottom) on the AUROC metric. Each row corresponds to a different type of cell, while each column represents a model of inference. The depths of colors of the heatmap correspond to the magnitudes of the values. **B** The corresponding box plots of the AUROC values on 21 benchmark datasets of ground-truth networks with the TFs and 500 most significantly varying genes (top), and on 21 benchmark datasets of ground-truth networks with the TFs and 1,000 most significantly varying genes (bottom). The bars inside the boxes represent the median values, their top and bottom edges represent the upper and lower quartiles, respectively, the bars outside the boxes represent the maximum and minimum values, and the small circles represent outliers

Table 1 Average values of the AUROC and AUPRC obtained by GRLGRN and other prevalent six on 42 benchmark datasets. The bold values represent the superior evaluation metric values

Model	AUROC	AUPRC
GRLGRN	**0.88 (33/42)**	**0.51 (34/42)**
GNNLINK	0.82	0.39
GENELINK	0.81	0.29
STGRNS	0.67	0.24
GNE	0.64	0.19
GRNBoost2	0.57	0.16
GENIE3	0.60	0.15

with TFs+ 500 (above) and TFs+ 1000 (below). The results graphically depict the significant improvements in performance brought about by the proposed GRLGRN in terms of the AUROC and AUPRC metrics. They thus demonstrate that our graph representation learning-based approach to GRN inference yields superior performance to other methods. Furthermore, Fig. 7 and Supplementary Figure S3, Additional file 1 presents the values of performance metrics (AUROC and AUPRC) by applying GRLGRN to the non-causal inference (output module in Fig. 1) and causal inference (output module in Fig. 2) of potential regulatory relationships in cell-type-specific GRNs, the results show that the non-causal inference performances are better than the causal inference performances on 71.4% (10/14) and 78.6% (11/14) of the datasets, with improvements of approximately 0.97% and 1.3%, respectively (More details can be found in Table S3 and S4 of the Additional file 1). Moreover, the results also demonstrate that GRLGRN can provide relatively significant causal inference performances for directed GRNs.

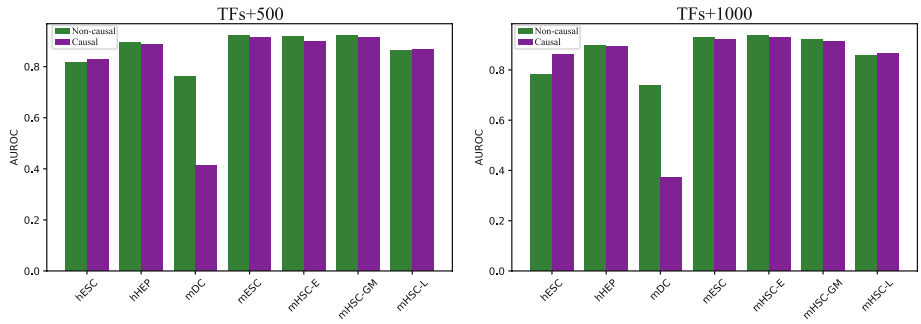

Fig. 7 AUROC scores of GRLGRN for non-causal inference and causal inference of potential regulatory relationships in cell-type-specific GRNs

In addition, the above results show that the models that incorporated graphs of prior GRNs—GRLGRN, GNNLINK, and GENELINK—provided superior inference-related performance on most benchmark datasets, which verifies the advantages of graph representation learning. The two models based on traditional machine learning (GRNBoost2 and GENIE3) exhibited an inferior inferential capacity to the other models, which demonstrates the advantages of deep learning methods.

Ablation study

To verify the effectiveness of each module of the proposed model, we conducted ablation experiments in which the inferential performance of GRLGRN and its variants was compared on the benchmark datasets. By simultaneously removing the graph transformer layer from the gene embedding module and the CBAM from the feature enhancement module, we obtained a variant model of GRLGRN called GRLGRN-v1. Moreover, only removing the CBAM from GRLGRN yielded another variant model called GRLGRN-v2. The architectures of these two variants are shown in Supplementary Figure S4, Additional file 1.

Figure 8 presents the AUROC scores of GRLGRN, GRLGRN-v1, and GRLGRN-v2 on the benchmark datasets of ground-truth networks with TFs+ 500 and TFs+ 1000, while their detailed AUPRC scores are shown in Supplementary Figure S5, Additional file 1. The proposed GRLGRN, which contained both the graph transformer layer and the CBAM, outperformed its two variant models overall (GRLGRN-v1 and GRLGRN-v2). GRLGRN was superior to GRLGRN-v2 in terms of AUROC and AUPRC scores on 80.9% (34 of 42) and 83.3% (35 of 42) of the scRNA-seq datasets, respectively. It recorded average improvements over GRLGRN-v2 in terms of AUROC of approximately 5.7% (14 of 14), 0.8% (7 of 14), and 1.02% (13 of 14) on the three ground-truth networks. Its average values of improvement in terms AUPRC scores over GRLGRN-v2 were approximately 25.9% (13 of 14), 2.41% (10 of 14), and 14.5% (12 of 14) on the three different ground-truth networks. GRLGRN-v2 was superior to GRLGRN-v1 and yielded average improvements of approximately 3.8% and 11.9% in AUROC and AUPRC scores respectively across the three ground-truth networks.

Moreover, we compared the computational scales of GRLGRN and its variants in terms of the average training time (see Table 2) and GPU memory usage (see Table 3). The results show that GRLGRN that incorporates the graph transformer layer and CBAM

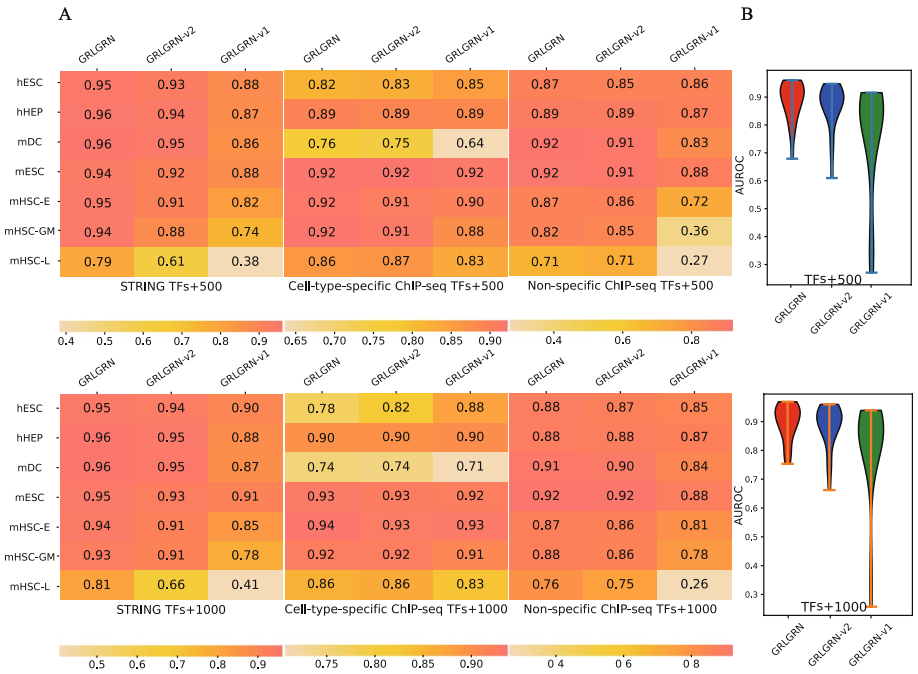

Fig. 8 Performances of GRLGRN and two variants (GRLGRN-v1 and GRLGRN-v2) in terms of inferring GRNs on the AUROC metric values. **A** The AUROC metric values on ground-truth networks with TFs and 500 most significantly varying genes (top), and ground-truth networks with TFs and 1000 most significantly varying genes (bottom). **B** Distributions of AUROC metric values on the TFs+500 (top) and TFs+1000 (bottom) benchmark datasets

Table 2 The average training time and GPU memory usage of GRLGRN, GRLGRN-v1, and GRLGRN-v2 across 21 (TFs+ 500) datasets

Model	Average time (s)	GPU memory (MB)
GRLGRN	709.19	1020.0
GRLGRN-v2	616.34	876.0
GRLGRN-v1	530.58	556.0

Table 3 The average training time and GPU memory usage of GRLGRN, GRLGRN-v1, and GRLGRN-v2 across 21 (TFs+ 1000) datasets

Model	Average time (s)	GPU memory (MB)
GRLGRN	1330.07	1434
GRLGRN-v2	1192.5	1354.0
GRLGRN-v1	915.79	993.0

needs the longest average training time and the largest GPU memory usage, GRLGRN-v2 with no CBAM needs shorter average training time and smaller GPU memory usage, and GRLGRN-v1, which has the lowest model complexity, needs the shortest average

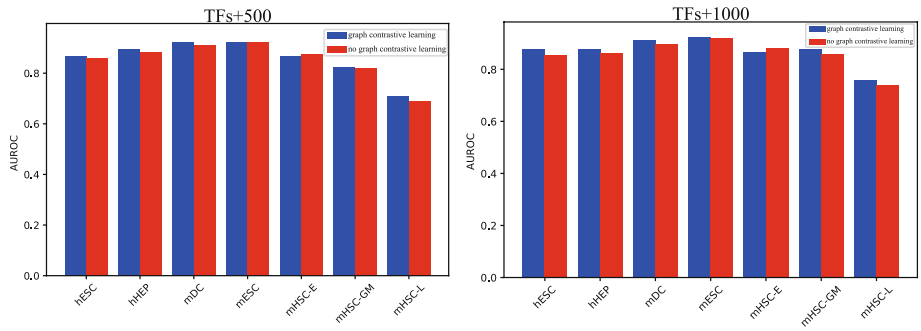

Fig. 9 AUROC scores of the GRLGRN models trained with and without graph contrastive learning on TFs+ 500 (left) and TFs+ 1000 (right) non-specific ChIP-seq datasets

Table 4 Details of AUROC scores of the GRLGRN models trained with (outside brackets) and without (in brackets) graph contrastive learning on TFs+ 500 and TFs+ 1000 non-specific ChIP-seq datasets. The bold values represent the superior evaluation metric values

Dataset	TFs+ 500	TFs+ 1000
hESC	**0.8664** (0.8594)	**0.8770** (0.8540)
hHEP	**0.8948** (0.8838)	**0.8798** (0.8612)
mDC	**0.9206** (0.9104)	**0.9116** (0.8974)
mESC	**0.9226** (0.9208)	**0.9223** (0.9198)
mHSC-E	0.8667 (**0.8739**)	0.8676 (**0.8816**)
mHSC-GM	**0.8240** (0.8212)	**0.8795** (0.8585)
mHSC-L	**0.7103** (0.6890)	**0.7600** (0.7396)

training time and the smallest GPU memory usage. More detailed information can be found in Table S8-S10 in Additional File 1.

The results verify the importance of the graph transformer layer and CBAM to the proposed GRLGRN. The above results show that GRLGRN-v1, which was based on information on explicit links, yielded the worst inferential performance. This shows that using the graph transformer layer in the gene embedding module to extract the features of genes can considerably enhance the inference of GRNs. Although GRLGRN-v2 used the graph transformer layer in the graph feature extraction module, its inferential capacity was still worse than that of GRLGRN. This demonstrates that applying the CBAM can further improve inferential performance, which highlights the advantage of the attention mechanism.

Furthermore, to validate the effectiveness of applying graph contrastive learning, we compared the inference performances, in terms of AUROC and AUPRC metrics, provided by the GRLGRN models trained with and without graph contrastive learning on non-specific ChIP-seq datasets. As shown in Fig. 9 and Table 4, in terms of the AUROC, graph contrastive learning improved inference performances on 85.7% (12/14) of the datasets, achieving an average improvement of approximately 1.8%. Moreover, Figure S6 and Table S5 of Additional file 1 shows that, in terms of AUPRC, graph contrastive learning improved inference performances on 57.1% (8/14) of the datasets, achieving an average improvement of approximately 2.5%. In addition, to further investigate the impact of the automatic weighted loss training techniques on model test performance, as shown

in Table S6 of Additional File 1, in terms of the AUROC and AUPRC, the automatic weighted loss training techniques improved inference performances on 64.2% (9/14) of the datasets, achieving an average improvement of approximately 1.24%. The AUPRC metric shows improvement across all datasets (more details can be found in Table S7 of Additional File 1). The above results demonstrated that combining GRLGRN with graph contrastive learning and automatic weighted loss training techniques can improve model inference accuracy.

Interpretation of graph transformer layer

To interpret the importance of extraction of implicit links in the graph transformer layer for the downstream task of GRN inference, we now discuss the contributions of all the generated meta-path graphs that were learned on the hESC scRNA-seq datasets of cell type-specific ground-truth networks with TFs+ 500 and TFs+ 1000. The contribution of any meta-path (implicit or explicit link) to the j-th output channel of the graph transformer layer, with $j = 1, 2, ..., B$, is $\varepsilon^{(1)}_{k_1,j} \cdot \varepsilon^{(2)}_{k_2,j}, k_1, k_2 = 1, 2, ..., 5$, where $\varepsilon^{(i)}_{k,j}$ is defined in Eq. (2), and is the attention score for the edge in $\mathcal{G}_k$.

Figure 10 interprets the importance of implicit links for ground-truth GRNs at two scales (TFs+ 500 and TFs+ 1000) in the same cell line (hESC), which are learned by the graph transformer layer to extract the features of genes via graph representation learning. Supplementary Tables S11 and S12, Additional file 1 show the contributions of all possible explicit and implicit links that have been learned. For instance, the implicit links between the target genes at two scales are the most important when extracting the features. This illustrates the importance of the association of two target genes regulated by the same TF during feature extraction. Moreover, the skip-level implicit links between the TFs and the target genes play a significant role. This shows that joint expression and skip-level implicit links can be regarded as playing a feed-forward role in a three-node network motif [44]. In addition, the explicit links between the TFs and target genes act as identity matrices.

The above results show that the proposed GRLGRN can automatically learn useful implicit links and optimize the process of feature extraction based on downstream tasks.

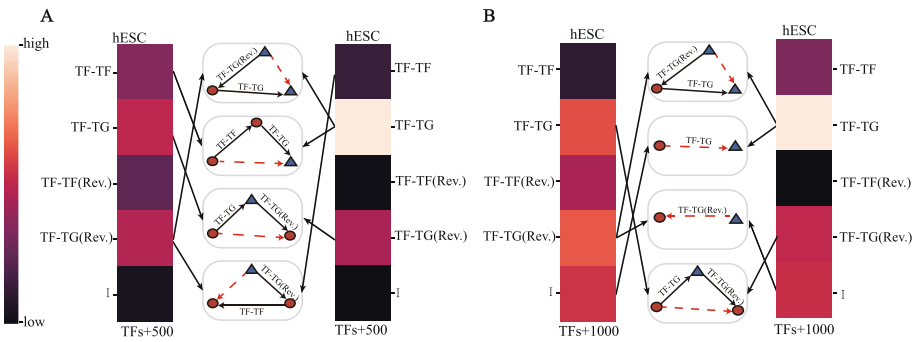

Fig. 10 Visualization of the contributions of different types of implicit (and explicit) links to TFs+ 500 (left) and TFs+ 1000 (right) hESC scRNA-seq datasets. The ground-truth network of hESC was provided by cell type-specific ChIP-seq, while the contribution of each type of link was obtained by computing the product of its attention scores. **A** The attention scores on the hESC with TFs and 500 most significantly varying genes. **B** The attention scores on the hESC with TFs and 1000 most significantly varying genes

Network visualization of prediction results

Finally, Fig. 11A visualizes a reconstructed GRN that was inferred by GRLGRN from the cell-type-specific ground-truth network with TFs and 500 most significantly varying genes for hESC cell line, where the 11 largest nodes are genes with degrees higher than 50. We note that GRLGRN is able to identify hub genes with high degrees, such as TFAP2 A, TEAD4, and JUND. For the sake of simplicity, we do not plot the genes whose degrees are lower than five and the corresponding regulatory associations. Specifically, TF AP- 2α, which is encoded by TFAP2 A and is hub gene with the highest degree, is crucial for various biological processes, including embryonic development, cell differentiation, and disease regulation [45]. In the development of neural crests, AP- 2α regulates the expression of various target genes by binding to specific DNA sequences, contributing to the development of facial structures, ocular tissues, the nervous system, and kidneys [46]. Mutations in TFAP2 A are associated with congenital diseases such as Axenfeld-Rieger syndrome and cleft lip/palate. Additionally, AP- 2α indirectly influences key signaling pathways by regulating the expression of genes such as EGFR, c-MYC, and Cyclin D1 [47], potentially functioning as an oncogene or tumor suppressor at different stages or in distinct tumor microenvironments of cancers, including breast cancer, melanoma, and head and neck squamous cell carcinoma. Therefore, modulating the expression of the hub gene TFAP2 A through gene regulatory networks (GRNs) could offer new insights into the treatment of related diseases. Furthermore, we used the trained GRLGRN to predict TFAP2 A-gene pairs. Figure 11B shows the top twenty gene pair interactions predicted for TFAP2 A, where the green lines represent interactions included in the benchmark training set (prior knowledge), and the red lines represent interactions outside the benchmark training set that have been confirmed (benchmark test set). The records on the bioGRID official website explain how TFAP2 A and DNM3 ta regulate gene expression through methylation [48]. These results illustrate the effectiveness of GRLGRN in predicting novel potential gene pair interactions. Moreover, the TFs+ 1000 scRNA-seq dataset is visualized in Supplementary Figure S7, Additional File

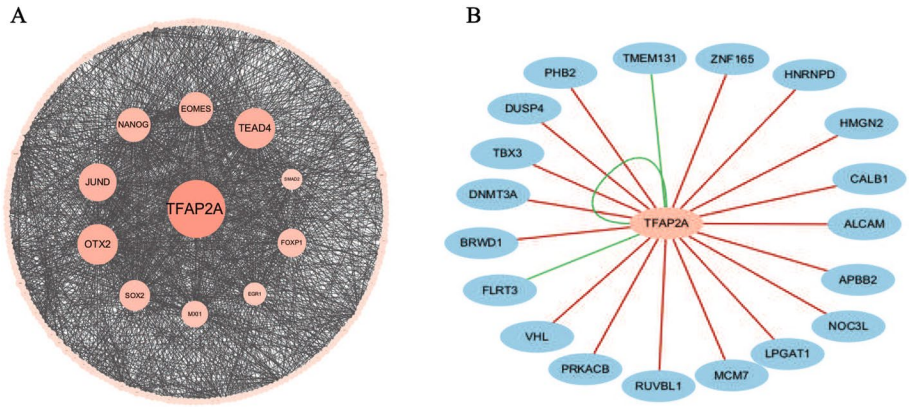

Fig. 11 **A** Visualization of a reconstructed cell type-specific ground-truth network that was inferred by GRLGRN, with TFs and 500 most significantly varying genes on the hESC scRNA-seq dataset. Eleven genes with the highest degrees, including TFAP2 A, TEAD4, and JUND, are represented by the large red nodes. **B** Visualization of 20 strongest newly inferred gene regulatory associations between TFAP2 A (hub gene) and other genes

1, which highlights the effectiveness of GRLGRN in inferring interactions related to hub genes.

Discussion

With the rapid advancement of high-throughput scRNA-seq technology, researchers have obtained various scRNA-seq data at a cellular resolution. Compared with traditional bulk data on omics, scRNA-seq data can avoid diluting information on individual cells to enable a better understanding of the intrinsic activities occurring within cells [36]. A number of methods for inferring GRNs based on traditional machine learning and deep learning have been developed in recent years, and provide new and effective tools based on scRNA-seq data. However, owing to biological heterogeneity, measurement errors, and data dropout, inferring GRNs based on scRNA-seq data still faces significant challenges [19]. Although GRLGRN makes significant progress in predicting potential interactions between genes, it still has certain limitations. GRLGRN is a supervised-learning-based model, which means that its performance depends on the reliability of sample labels. However, obtaining precise labels in biological networks remains a challenging issue, making it one of the factors that limits GRLGRN inference capabilities.

In the future works, unsupervised learning methods such as clustering, dimensionality reduction, and contrastive learning can be explored to uncover latent patterns and structures in gene expression profiles, thereby reducing the dependencies on data labels and providing more information for subsequent gene representations. Moreover, we can further try to apply transfer learning and meta-learning techniques can leverage existing knowledge in the field of bioinformatics, and try to incorporate multi-model data such as DNA methylation, histone modifications, etc., and protein interaction data to provide a more comprehensive view of gene regulations.

Conclusion

In this study, we proposed GRLGRN—a deep learning model for extracting information on implicit links and gene embeddings from structures based on prior GRNs and profiles of gene expression by using a graph transformer network. We used attention mechanisms to optimize the capability of GRLGRN to extract gene representations. We took the Hadamard product of the gene representations of any given gene pair and then passed it through a two-layer perceptron to obtain its regulatory relationship scores. In addition, we used graph contrastive learning during model optimization to reduce the degradation of its inferences due to over-smoothing. A comparison of GRLGRN with six prevalent models on scRNA-seq datasets from seven types of cell lines and three types of ground-truth networks (at scales of TFs+ 500 and TFs+ 1000) showed that our method outperformed the other models in terms of the AUROC and AUPRC metrics on most scRNA-seq datasets. The experimental results indicated that the model was also able to achieve good performance for causal regulation. Furthermore, we conducted ablation experiments that not only verified the effectiveness of the graph transformer layer in the gene embedding module and CBAM in the feature enhancement module but also demonstrated the computational costs of GRLGRN and its variants. Moreover, we also validated the effectiveness of

graph contrastive learning and automatic weighted loss training techniques. Our discussion of the interpretation of its results and network visualization provided insights that can help us better understand GRNs as well as the contributions of all explicit and implicit links to them.

The proposed GRLGRN delivers superior inferential performance to that of GENIE3 [15] and GRNBoost2 [16] by introducing deep learning technology. Moreover, it can use information on prior links through graph representation learning, unlike GNE [43] and STGRNS [19]. Compared with GENELINK [29] and GNNLINK [27], which leverage link-related information, an innovation of GRLGRN lies in its ability to fully exploit the link-related information of prior GRNs through the graph transformer layer, and to identify implicit links of various forms. It then uses a GCN to obtain the corresponding gene embedding. The strength of GRLGRN lies in its ability to learn useful explicit and implicit links adaptively in accordance with downstream tasks. Furthermore, its use of a regularization term for graph contrastive learning during training enables it to maximize the consistency in features between embeddings to mitigate the risk of overfitting.

Abbreviations

GRNs	Gene regulatory networks
TFs	Transcription factors
ChIP-seq	Chromatin immunoprecipitation sequencing
CNNs	Convolutional neural networks
GCNs	Graph convolutional networks
CBAM	Convolutional block attention module
hESCs	Human embryonic stem cells
hHEPs	Human mature hepatocytes
mDCs	Mouse dendritic cells
mESCs	Mouse embryonic stem cells
mHSC-E	Mouse hematopoietic stem cells with erythroid lineage
mHSC-GM	Mouse hematopoietic stem cells with granulocyte-monocyte lineage
AUROC	Area Under the Receiver Operating Characteristic Curve
AUPRC	Area Under the Precision-Recall Curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06116-1.

Supplementary file 1.

Acknowledgements
Not applicable.

Author contributions
KW and YL designed this research and wrote the draft. XW validated the results. FL and XL guided this research process. JZ conducted research, guided this research process, and revised the manuscript. All authors read and approved the final draft.

Funding
Natural Science Foundation of Jiangsu Province (BK20202002); National First-class Discipline Program of Light Industry Technology and Engineering (QGJC20230102); National Natural Science Foundation of China (62373166); China Postdoctoral Science Foundation (2022M711362).

Data availability
The original dataset is available in the paper Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data https://doi.org/10.1038/s41592-019-0690-6, and the processed dataset is available for download from https://github.com/zpliulab/GENELink.

Materials availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing of interests
Not applicable.

Received: 22 December 2024 Accepted: 18 March 2025
Published online: 18 April 2025

References

1. Cramer P. Organization and regulation of gene transcription. Nature. 2019;573(7772):45–54.
2. Mao G, Zeng R, Peng J, et al. Reconstructing gene regulatory networks of biological function using differential equations of multilayer perceptrons. BMC Bioinform. 2022;23(1):503.
3. Li X, Ma S, Guo F, et al. Inferring gene regulatory network via fusing gene expression image and RNA-seq data. Bioinformatics. 2022;38(6):1716–23.
4. Ma B, Fang M, Jiao X. Inference of gene regulatory networks based on nonlinear ordinary differential equations. Bioinformatics. 2020;36(19):4885–93.
5. Shu H, Zhou J, Ma J, et al. Modeling gene regulatory networks using neural network architectures. Nat Comput Sci. 2021;1(7):491–501.
6. Van De Sande B, Flerin C, Davie K, et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat Protoc. 2020;15(7):2247–76.
7. Zhang C, Lu Y, Zang T. CNN-DDI: a learning-based method for predicting drug-drug interactions using convolution neural networks. BMC Bioinform. 2022;23(Suppl 1):88.
8. Zhao M, He W, Guo F, et al. A comprehensive overview and critical evaluation of gene regulatory network inference technologies. Briefings Bioinform. 2021;22(5):bbab009.
9. Delgado-Chaves FM, Gómez-Vela F, Divina F, et al. Computational analysis of the global effects of Ly6E in the immune response to coronavirus infection using gene networks. Genes. 2020;11(7):831.
10. Österberg L, Domenzain I, Münch J, et al. A novel yeast hybrid modeling framework integrating Boolean and enzyme-constrained networks enables exploration of the interplay between signaling and metabolism. PLoS Comput Biol. 2021;17(4): e1008891.
11. Düvel K, Yecies JL, Menon S, et al. Activation of a metabolic gene regulatory network downstream of mTOR complex 1. Mol Cell. 2010;39(2):171–83.
12. Zhang Y, Wang H, Tu W, et al. Comparative transcriptome analysis provides insight into spatio-temporal expression characteristics and genetic regulatory network in postnatal developing subcutaneous and visceral fat of bama pig. Front Genet. 2022;13: 844833.
13. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet. 1999;21(1):33–7.
14. Park PJ. ChIP-Seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80.
15. Huynh-Thu VA, Irrthum A, Wehenkel L, et al. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;5(9): e12776.
16. Moerman T, Aibar Santos S, Bravo González-Blas C, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35(12):2159–61.
17. Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. Proc Natl Acad Sci. 2019;116(52):27151–8.
18. Chen J, Cheong C, Lan L, et al. DeepDRIM: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell RNA-seq data. Briefings Bioinform. 2021;22(6):bbab325.
19. Xu J, 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(4):btad165.
20. Wang Q, Guo M, Chen J, et al. A gene regulatory network inference model based on pseudo-siamese network. BMC Bioinform. 2023;24(1):163.
21. Shachaf LI, Roberts E, Cahan P, et al. Gene regulation network inference using k-nearest neighbor-based mutual information estimation: revisiting an old dream. BMC Bioinform. 2023;24(1):84.
22. Hu S, Jing Y, Li T, et al. Inferring circadian gene regulatory relationships from gene expression data with a hybrid framework. BMC Bioinform. 2023;24(1):362.
23. Skok Gibbs C, Mahmood O, Bonneau R, et al. PMF-GRN: a variational inference approach to single-cell gene regulatory network inference using probabilistic matrix factorization. Genome Biol. 2024;25(1):88.
24. Tang J, Wang C, Xiao F, et al. Single-cell gene regulatory network analysis for mixed cell populations. Quantitative Biol. 2024;12(4):375–88.
25. 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 Bioinf. 2023;20(5):2853–61.
26. Yuan Y, Bar-Joseph Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. Genome Biol. 2020;21(1):300.
27. Mao G, Pang Z, Zuo K, et al. Predicting gene regulatory links from single-cell RNA-seq data using graph neural networks. Briefings Bioinform. 2023;24(6):bbad414.

28. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016
29. Chen G, Liu ZP. Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data. Bioinformatics. 2022;38(19):4522–9.
30. Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks. Stat. 2017;1050(20):10–48550.
31. Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. IEEE Trana Neural Netw Learn Syst. 2020;32(1):4–24.
32. Yun S, Jeong M, Kim R, et al. Graph Transformer Networks: learning meta-path graphs to improve GNNs. Neural Netw. 2022;153:104–19.
33. Woo S, Park J, Lee JY, et al. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 3–19.
34. Zhu Y, Xu Y, Yu F, et al. Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131. 2020;.
35. You Y, Chen T, Sui Y, et al. Graph contrastive learning with augmentations. arXiv preprint arXiv:2010.13902. 2021
36. Pratapa A, Jalihal AP, Law JN, et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17(2):147–54.
37. Szklarczyk D, Gable AL, Lyon D, et al. STRING V11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–13.
38. Xu H, Baroukh C, Dannenfelser R, et al. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database. 2013;2013:bat045.
39. Garcia-Alonso L, Holland CH, Ibrahim MM, et al. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019;31(4):745–745.
40. Yang Z, Ding M, Zou X, et al. Region or global a principle for megative sampling in graph-based recommendation. IEEE Trans Knowl Data Eng. 2022;35(6):6264–77.
41. Wang Y, Min Y, Chen X, et al. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In: Proceedings of the Web Conference 2021; 2021. p. 2921–2933.
42. Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7482–7491.
43. Kc K, Li R, Cui F, et al. GNE: a deep learning framework for gene network inference by aggregating biological information. BMC Syst Biol. 2019;13(38):1–14.
44. Liu W, Yang Y, Lu X, et al. NSRGRN: a network structure refinement method for gene regulatory network inference. Briefings Bioinform. 2023;24(3):bbad129.
45. Bumrungthai S, Buddhisa S, Prakobkaew N, et al. Association of HHV-6 reactivation and SLC6A3 (C> T, rs40184), BDNF (C> T, rs6265), and JARID2 (G> A, rs9383046) single nucleotide polymorphisms in depression. Biomed Rep. 2024;21(6):181.
46. Chambers BE, Gerlach GF, Clark EG, et al. Tfap2a is a novel gatekeeper of nephron differentiation during kidney development. Development. 2019;146(13): dev172387.
47. Khachigian LM. The Yin and Yang of YY 1 in tumor growth and suppression. Int J Cancer. 2018;143(3):460–5.
48. Hervouet E, Vallette FM, Cartron PF. Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. Epigenetics. 2009;4(7):487–99.

Publisher's Note