Personal Statement

Name	王 鹏 程	Major applied for	系统合成和生物信息学

1 Academic Research Expericence

I received my master's degree in Biology and medicine from Soochow University in 2024 under the mentorship of Professor Sidong Xiong and Professor Hang Ruan. My research primarily focused on the regulation of A-to-I RNA editing and its implications in autoimmune diseases. Through this work, I gained extensive training in bioinformatics and immunology, providing me with a strong foundation for tackling complex biomedical questions. During this period, I developed a strong interest in statistics, machine learning, and gene regulation. However, I still feel there is much more to learn, and I intend to pursue further studies in order to make meaningful contributions in these fields. Here is an introduction to my research achievements and experiences:

1.1 Academic Paper

1 Ruan, H., **Wang, P.-C.**, & Han, L. (2022). Characterization of circular RNAs with advanced sequencing technologies in human complex diseases. WIREs RNA, e1759. https://doi.org/10.1002/wrna.1759. (**Second author, with the supervisor as the first author; Q1, 6.4**)

2 Weng, S.; Yang, X.; Yu, N.; **Wang, P.-C.**; Xiong, S.; Ruan, H. Harnessing ADAR-Mediated Site-Specific RNA Editing in Immune-Related Disease: Prediction and Therapeutic Implications. Int. J. Mol.Sci.2024,25,351. https://www.mdpi.com/1422-0067/25/1/351. (**Co-author; Q2, 4.9**)

3 Zhou, K., Wei, W., Yang, D. et al. Dual electrical stimulation at spinal-muscular interface reconstructs spinal sensorimotor circuits after spinal cord injury. Nat Commun 15, 619 (2024). https://www.nature.com/articles/s41467-024-44898-9. (**Co-author; Q1, 14.9**)

1.2 Research Projects

1.2.1 Master Thesis

Characterization and Functional Analysis of A-to-I RNA Editing in Autoimmune Diseases: This is my independently conducted Master's thesis, which explores the characteristics and functions of A-to-I RNA editing across six common autoimmune diseases. Using RNA sequencing data from multiple cohorts, the study investigates A-to-I RNA editing levels, ADAR gene expression, and dsRNA-related signaling pathways. It identifies and functionally characterizes A-to-I RNA editing sites, shedding light on their roles in gene regulation, miRNA binding, alternative splicing, and peptide immunogenicity. To further clarify the impact of RNA editing in autoimmunity, Mendelian randomization (MR) analysis is used to distinguish editing events with protective effects from those with pathogenic consequences. In addition, patients are stratified into clusters based on their interferon levels to explore how differences in interferon signaling correlate with RNA editing patterns (Figure 1). This study ultimately aims to construct a comprehensive atlas of A-to-I RNA editing in autoimmune diseases, offering novel perspectives that may contribute to the identification of potential therapeutic targets.

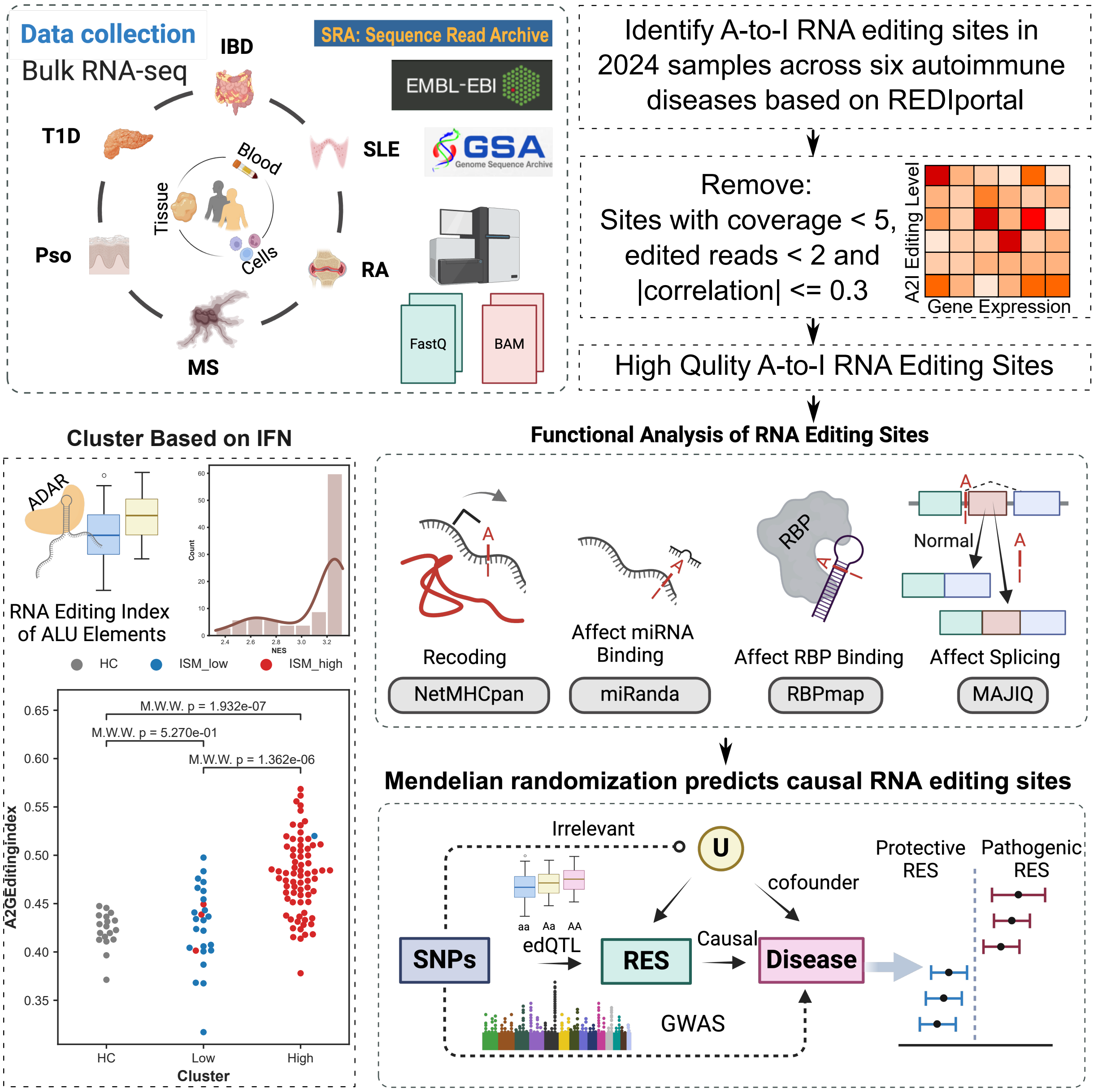


Figure 1. Schematic overview of the study design. Bulk RNA-seq data from 2,024 samples covering six autoimmune diseases—Type 1 diabetes (T1D), inflammatory bowel disease (IBD), systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), psoriasis (Pso), and multiple sclerosis (MS)—were retrieved from public repositories. After filtering for high-quality A-to-I RNA editing sites, samples were stratified by interferon levels, and functional analyses were performed to assess recoding, miRNA binding, RBP binding, and alternative splicing. Finally, Mendelian randomization identified causal editing sites, distinguishing those with protective from those with pathogenic effects.

1.2.2 CoWorks

Dual electrical stimulation at spinal-muscular interface reconstructs spinal sensorimotor circuits after spinal cord injury[1] (Research Article 3): The neural signals produced by varying electrical stimulation parameters lead to characteristic neural circuit responses. However, the characteristics of neural circuits reconstructed by electrical signals remain poorly understood, which greatly limits the application of such electrical neuromodulation techniques for the treatment of spinal cord injury. Here, we develop a dual electrical stimulation system that combines epidural electrical and muscle stimulation to mimic feedforward and feedback electrical signals in spinal sensorimotor circuits. We demonstrate that a stimulus frequency of 10−20Hz under dual stimulation conditions is required for structural and functional reconstruction of spinal sensorimotor circuits, which not only activates genes associated with axonal regeneration of motoneurons, but also improves the excitability of spinal neurons. Overall, the results provide insights into neural signal decoding during spinal sensorimotor circuit reconstruction, suggesting that the combination of epidural electrical and muscle stimulation is a promising method for the treatment of spinal cord injury.

Characterization of circular RNAs with advanced sequencing technologies in human complex diseases[2] (Review Article 1): Circular RNAs (circRNAs) are one category of non-coding RNAs that do not possess 5′ caps and 3′ free ends. Instead, they are derived in closed circle forms from pre-mRNAs by a non-canonical splicing mechanism named “back-splicing”. CircRNAs were discovered four decades ago, initially called “scrambled exons”. Compared to linear RNAs, the expression levels of circRNAs are considerably lower, and it is challenging to identify circRNAs specifically. Thus, the biological relevance of circRNAs has been underappreciated until the advancement of next generation sequencing (NGS) technology. The biological insights of circRNAs, such as their tissue-specific expression patterns, biogenesis factors, and functional effects in complex diseases, namely human cancers, have been extensively explored in the last decade. With the invention of the third generation sequencing (TGS) with longer sequencing reads and newly designed strategies to characterize full-length circRNAs, the panorama of circRNAs in human complex diseases could be further unveiled. In this review, we first introduce the history of circular RNA detection. Next, we describe widely adopted NGS-based methods and the recently established TGS-based approaches capable of characterizing circRNAs in full-length. We then summarize data resources and representative circRNA functional studies related to human complex diseases. In the last section, we reviewed computational tools and discuss the potential advantages of utilizing advanced sequencing approaches to a functional interpretation of full-length circRNAs in complex diseases.

[1] In this article, I provided suggestions for experiments from a bioinformatics perspective and analyzed the sequencing data, with the main results presented in Figure 9.

[2] In this paper, my primary contributions included literature collection and figure creation.

Harnessing ADAR-Mediated Site-Specific RNA Editing in Immune-Related Disease: Prediction and Therapeutic Implications[3] (Review Article 2): ADAR (Adenosine Deaminases Acting on RNA) proteins are a group of enzymes that play a vital role in RNA editing by converting adenosine to inosine in RNAs. This process is a frequent post-transcriptional event observed in metazoan transcripts. Recent studies indicate widespread dysregulation of ADAR-mediated RNA editing across many immune-related diseases, such as human cancer. We comprehensively review ADARs'function as pattern recognizers and their capability to contribute to mediating immune-related pathways. We also highlight the potential role of site-specific RNA editing in maintaining homeostasis and its relationship to various diseases, such as human cancers. More importantly, we summarize the latest cutting-edge computational approaches and data resources for predicting and analyzing RNA editing sites. Lastly, we cover the recent advancement in site-directed ADAR editing tool development. This review presents an up-to-date overview of ADAR-mediated RNA editing, how site-specific RNA editing could potentially impact disease pathology, and how they could be harnessed for therapeutic applications.

[3] In this paper, I primarily presented the cutting-edge research status of this field.

2 Research Proposal

2.1 Introduction

Recent advancements in high-throughput sequencing technologies, particularly single-cell RNA sequencing (scRNA-seq), have revolutionized our ability to explore gene expression heterogeneity at cellular resolution[1,2]. These techniques enable unprecedented exploration of dynamic biological processes, ranging from cellular differentiation to disease progression[3,4]. A central challenge in this domain, however, lies in deciphering causal molecular interactions from observational omics data[5]. This is particularly critical for reconstructing gene regulatory networks (GRNs), in which directed graphs encoding transcription factors' (TFs) regulatory influences on target genes[6].

Existing computational methods for GRN inference, including correlation-based approaches like SCENIC[8,9] that rely on co-expression patterns, and regression models like GRNBoost2[10] that model target gene expression based on transcription factors, fundamentally depend on association metrics (Figure 1). These metrics often conflate causal and correlative relationships[3,7,11], leading to the inclusion of spurious edges from indirect regulation[12]. This limitation is exacerbated by technical artifacts in scRNA-seq data[13], where introduce high dropout rates and significant noise. While model-based methods[4,11], such as differential equations and probability models (e.g., Bayesian networks) (Figure 1), attempt to explicitly model system dynamics or probabilistic dependencies, they often impose restrictive structural or parametric assumptions ill-suited to the complex, nonlinear dynamics of gene regulation and can be computationally expensive for large-scale networks[11,12]. Similarly, deep learning-based approaches (Figure 1), excel at capturing complex, non-linear patterns directly from multi-omics data, they often suffer from poor interpretability, limiting mechanistic biological insights, and typically require large datasets and substantial computational resources for effective training[2,7,14]. Furthermore, these traditional approaches may overlook other crucial regulatory mechanisms beyond transcript levels, such as chromatin accessibility.

Traditional methods for gene regulatory network (GRN) inference often lack inherent mechanisms for encoding causality[5,15]. In contrast, Graph Neural Networks (GNNs) can capture the underlying topological patterns of gene regulation through embedding learning, where gene nodes learn feature representations by aggregating information from their network neighbors[3,16,17]. Notably, attention mechanisms, such as the Graph Attention Network (GAT), exhibit a natural connection to causal strength estimation by assigning differential weights to neighboring nodes, allowing the model to prioritize influential regulators[13,16,18]. Furthermore, graph autoencoders (GAEs), including causal variants, demonstrate the capability to denoise sparse single-cell data by learning latent representations and reconstructing the input, potentially mitigating the impact of technical noise like dropouts[3,13,18]. Additionally, research suggests that GNNs can function as a type of neural Structural Causal Model (SCM)[18], underscoring their potential for advancing causal inference in GRN analysis.

While correlation-based methods can indicate potential regulatory relationships by identifying co-expressed genes[3,7], causal inference offers a more robust approach by aiming to establish direct cause-and-effect relationships, thus revealing which genes actively regulate others[5,13,19]. This move beyond mere association is crucial for overcoming the inherent challenges in deciphering the complexities of GRNs[20]. Causal inference seeks to disentangle the intricate network of interactions and provide a more accurate and mechanistic understanding of gene regulation[5,11,13,16]. Given that genes within GRNs primarily act as regulators (causes) influencing the expression of other genes (effects), the study of these networks is particularly well-suited for the application and development of new causal inference methodologies. Indeed, there is a growing and significant trend in utilizing causal inference techniques to analyze biological datasets[2,5,11,21,22] (Table 1), notably to reconstruct gene regulatory networks from gene expression data and other omics modalities, promising deeper insights into cellular processes and disease mechanisms.

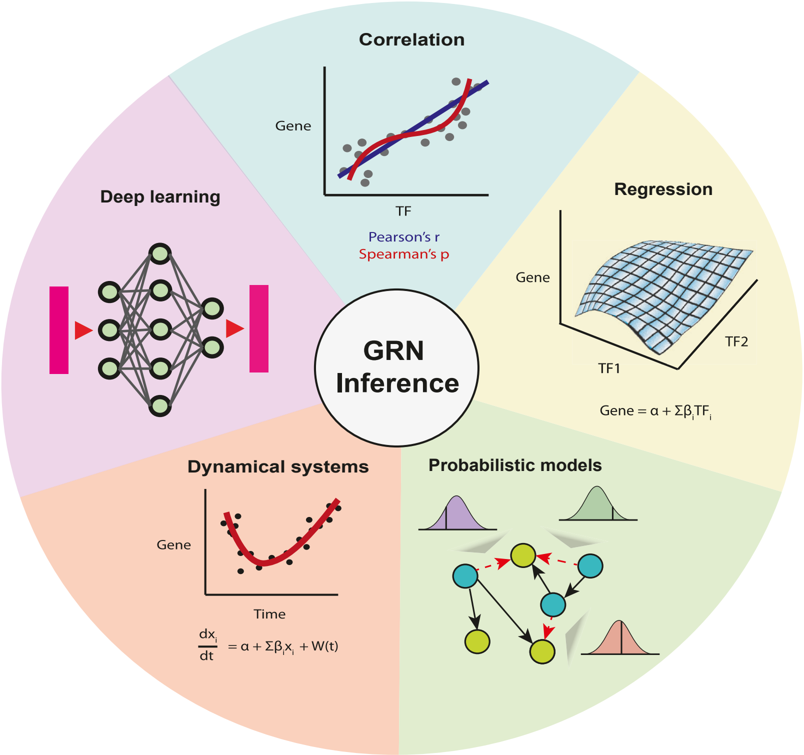


Figure 1. The major classes of methods for paired single-cell multi-omics GRN inference methods[7]. Correlation-based methods, which identify co-variation patterns between molecular features (e.g., TF expression, gene expression, CRE accessibility); Regression approaches, modeling target gene expression as a function of potential predictors like TF expression and/or CRE accessibility; Probabilistic models, aiming to infer the most likely network structure or regulatory interactions based on probability theory; Dynamical systems, using mathematical equations (e.g., differential equations) to model the temporal dynamics of gene expression influenced by regulators and other factors; and Deep learning approaches, employing neural networks to learn complex, potentially non-linear relationships among TFs, CREs, genes, and cellular contexts.

Moving beyond the limitations of correlation-based methods that identify co-expression patterns, this research aims to provide a more mechanistic understanding of gene regulation. To achieve this, this study will couple graph attention mechanisms with causal inference, making edge weights reflect estimated causal strengths rather than mere associations, thereby improving both GRN reconstruction accracy and interpretability.

Correlation Does Not Imply Causation.

2.2 Literature Review

Causal inference combined with Graph Neural Networks (GNNs) for reconstructing gene regulatory networks (GRNs) from single-cell omics data is now a rapidly advancing field[18,23]. Traditional correlation-based methods, such as those employing Pearson correlation or mutual information, and tools like SCENIC[8,9], while effective at identifying gene co-expression patterns, often fail to distinguish between direct and indirect regulatory relationships, leading to the inclusion of spurious edges[12,14,24]. Furthermore, these methods typically lack explicit mechanisms to model the inherent causality within GRNs, making it challenging to predict the impact of perturbations or discern cause-and-effect relationships[14]. These limitations are compounded by the technical noise, sparsity, and dropout events prevalent in single-cell RNA sequencing (scRNA-seq) data[3,4,25]. While model-based approaches like Boolean and Bayesian networks, and regression-based methods such as GENIE3[26] and GRNBoost2[10], offer alternative frameworks, they often suffer from parameter dependency, computational complexity, assumptions about data distributions, or an inability to fully capture the complexity of regulatory mechanisms[4,12] (Table 1).

The integration of causal inference[27] with the representational power of GNNs[28] offers a promising approach to overcome these limitations. Causal inference aims to uncover the underlying mechanisms of gene regulation, moving beyond mere statistical associations. GNNs, with their ability to model complex graph-structured data, are well-suited for capturing the intricate dependencies within GRNs and have shown potential in handling the noise and sparsity inherent in single-cell data by leveraging network structure for information propagation and noise smoothing[29–31]. Moreover, GNNs can naturally model non-linear relationships often missed by linear approaches[11,16,19]. Some methods initialize GRN structures based on prior knowledge and then use GNN-based encoders like Graph Convolutional Networks (GCNs) to refine gene features and identify interdependencies[3,32–35].

For instance, GENELink[23] employs a supervised learning approach using the Graph Attention Network (GAT) to learn low-dimensional embeddings of genes from scRNA-seq data and prior interaction knowledge (Figure 2A, Table 1). These embeddings are then used for downstream tasks such as similarity measurement and causal inference of gene regulation. The attention mechanism in GAT allows the model to adaptively weight the importance of neighboring genes, potentially mitigating the effects of data sparsity. In contrast, CgNN[22] presents a GNN-driven instrumental variable (IV) approach (Figure 2B). It leverages the network structure itself as an instrument to address the challenge of hidden confounders, enabling the estimation of direct and indirect causal effects of treatments within the network. This approach directly tackles the confounding noise issue that challenges traditional correlation-based methods. Both GENELink and CgNN demonstrate the advantage of using GNNs to integrate gene expression features with network topology for improved GRN inference and causal effect estimation.

However, despite these advancements, significant challenges remain in effectively and systematically integrating causal inference principles with GNN architectures, particularly when faced with the extreme heterogeneity of single-cell data. Most existing methods, like those mentioned above, primarily focus on static snapshots of cellular states, often neglecting the crucial temporal dynamics of gene regulation that unfold during processes like development or disease progression[1,36]. This limitation can mask the identification of key causal drivers and regulatory mechanisms that are inherently dynamic. Furthermore, while incorporating prior knowledge through GNNs is beneficial, the inherent noise and heterogeneity in single-cell omics data can still lead to overfitting or the misidentification of spurious regulatory edges[37]. Current causal inference strategies often rely on simplifying assumptions, such as linearity or specific partial correlation structures[12,38,39], which may fail to capture the complex, non-linear, and context-dependent nature of biological regulatory systems[40–42].

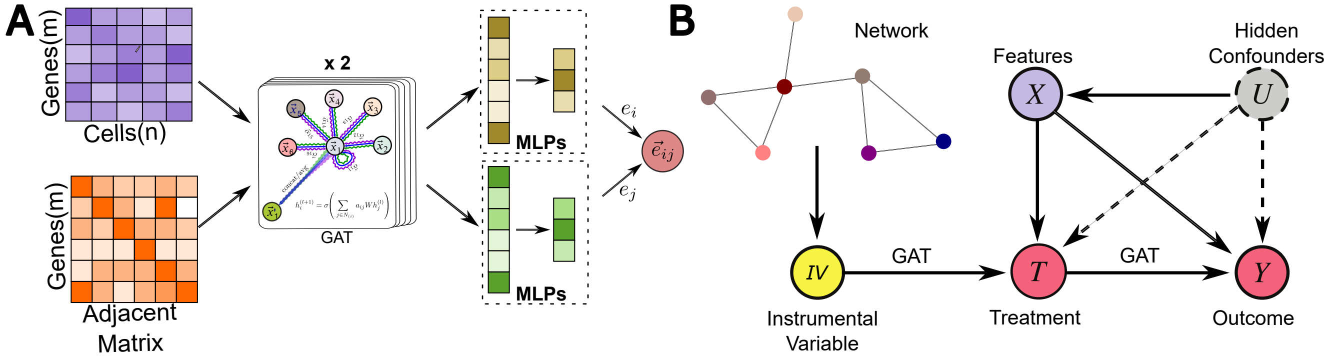


Figure 2. Illustrative applications of GNNs in causal inference. A. CgNN framework use a graph as input and combines the GAT with instrumental variable (IV) to remove hidden confounders and estimate causal effects. B. GENELink processes gene expression and TF-gene adjacency matrices through stacked GAT layers with multi-head attention, generating low-dimensional embeddings for causal GRN reconstruction.

The field has seen a clear progression from correlation-based methods to model-based and regression-based approaches, and more recently, to the integration of machine learning, particularly GNNs, with causal inference frameworks for GRN reconstruction. Early GNN-based methods often focused on link prediction, aiming to identify potential regulatory interactions by leveraging network structure and gene features. More recent works, such as CgNN[22], are explicitly incorporating causal inference techniques like instrumental variables to address confounding and infer causal effects. Another emerging trend involves using GNNs for tasks beyond simple link prediction, such as identifying causal driver genes and performing computational intervention analysis[43,44]. Methods like DeepGRNCS[4] are exploring cell subpopulation-specific regulation by training models to predict gene expression based on transcription factor activity.

Despite this progress, several key limitations highlight the need for further research. A major open question is how to effectively model the dynamic nature of GRNs using GNNs and causal inference, moving beyond static datasets to capture temporal dependencies. Addressing the inherent noise and heterogeneity of single-cell data while avoiding overfitting remains a significant challenge. Developing causal inference strategies that can capture non-linear and context-dependent regulatory mechanisms is also cru-

Correlation Does Not Imply Causation.

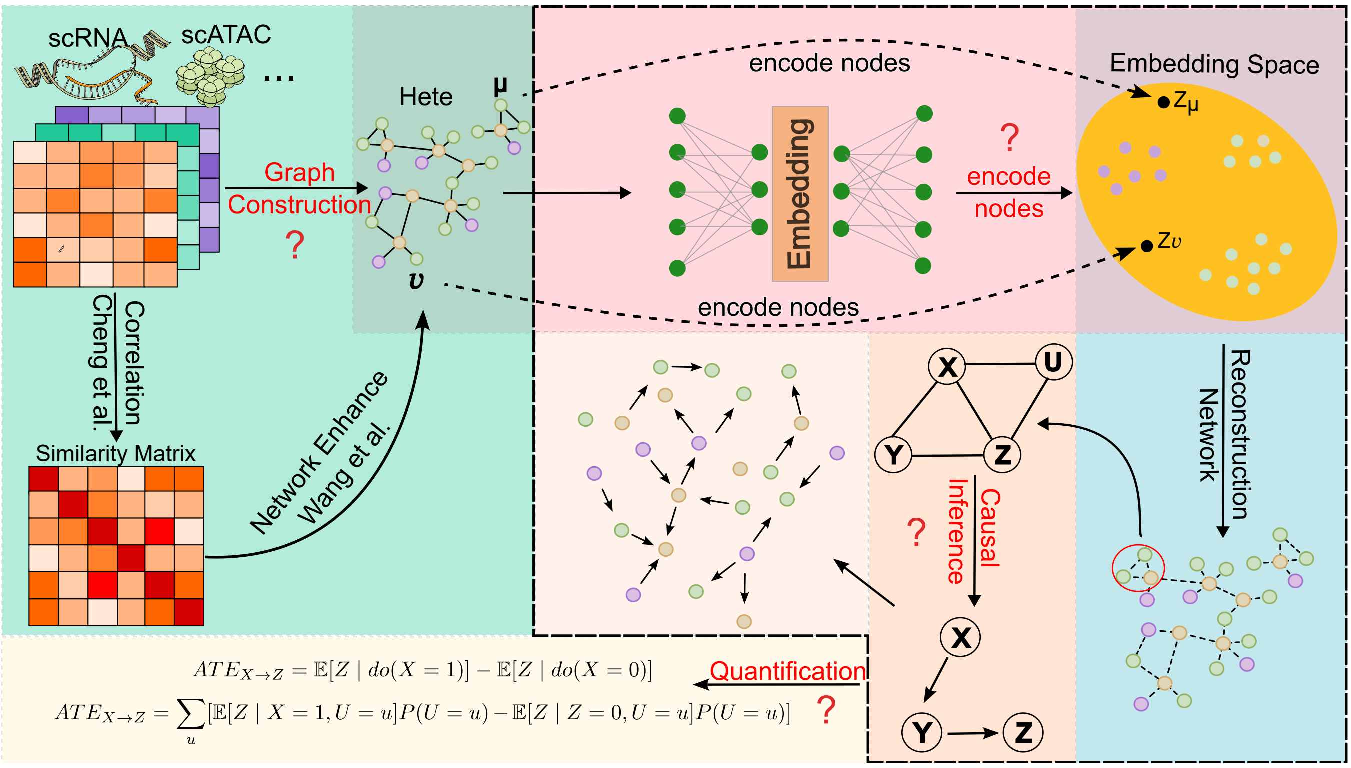
cial. Furthermore, the interpretability of complex GNN models and the validation of inferred causal relationships in biological systems are ongoing areas of research. Future work should focus on developing novel GNN architectures and causal inference techniques that are specifically tailored to the unique challenges of single-cell omics data, enabling a more robust and accurate reconstruction of gene regulatory networks and a deeper understanding of their causal underpinnings.

Table 1. Comparative analysis of GRN inference methods

Method	Core Approach	Causality	Limitations	Year
WGCNA[45]	Weighted correlation network	-	Confounding Noise	2008
GENIE3[26]	Tree-based ensemble	-	Sensitive to parameters	2010
SCENIC[8]	Co-expression & TF motifs	-	Motif Requirement	2017
GRNBoost2[10]	Gradient Boosting Machine	-	Sensitive to hyperparameters	2019
CNNC[46]	Convolutional Neural Network	+	Supervised Nature & Labeled Data Dependency	2019
Scribe[47]	Information theory-based	-	Requires Temporal Coupling	2020
DeepMAPS[48]	Heterogeneity Graph Transformer	-	Low Computational Efficiency	2021
FigR[49]	Correlation & TF motifs	-	Confounding Noise & Prior Knowledge Bias	2022
scREMOTE[38]	Regression	-	Linear-only Modeling	2022
GENELink[23]	Graph Attention Network	+	No Reliable Negatives	2022
GLUE[35]	Graph-Linked Embedding	-	Sensitivity to Noise	2022
Pando[50]	Regression & TF-gene prior	-	Linear-only Modeling & Prior knowledge biasi	2023
NME/cNME[24]	Cross-Mapping Entropy	+	High Computational Cost	2023
Dictys[21]	Stochastic Differential Equation	-	Linear-only Modeling	2023
Swift-DynGFN[51]	Generative Flow Network	+	Limited Generalization	2023
GRINCD[11]	Graph Neural Network & Additive Noise Model	+	Initial Network Dependency	2023
LINGER[52]	Neural Network	-	Reliance on known TF motifs	2024
scMultiomeGRN[53]	Graph Convolutional Network	-	Transcription Factors Only	2025
DigNet[20]	Diffusion-based Model	-	Random Sampling Variability	2025
CIMLA[5]	Machine Learning & Attribution Models	+	Assumption Dependence	2025

2.3 Methodology

This research aims to develop a novel method for reconstructing GRNs by combining Graph Neural Networks (GNNs) with causal inference. Leveraging GNNs can uncover complex relationships in single-cell omics data and causal inference pinpoints true cause-and-effect links among genes. Figure 1 illustrates the overall research framework. The following sections detail our approach to addressing the framework's key challenge.



$$ATE_{X\to Z} = \mathbb{E}[Z \mid do(X=1)] - \mathbb{E}[Z \mid do(X=0)]$$

$$ATE_{X\to Z} = \sum_u [\mathbb{E}[Z \mid X=1, U=u]P(U=u) - \mathbb{E}[Z \mid Z=0, U=u]P(U=u)]$$

Figure 3. Overview of research framework. The proposed methodology integrates Graph Neural Networks (GNNs) with causal inference techniques to reconstruct gene regulatory networks (GRNs) from single-cell omics data. By combining the strengths of GNNs and causal inference, the framework aims to capture the causal, directionally-aware relationships among genes, transcription factors, other regulatory elements and finally quantify the causal effects (such as do-calculus).

2.3.1 Initial Graph Construction

Graph-based representations of single-cell omics data offer a powerful framework for capturing and analyzing complex cellular and molecular interactions. However, constructing such graphs directly from raw data can introduce noise, redundant edges, or suboptimal connectivity[5,20,37], thereby hindering downstream analyses, such as gene regulatory network inference and cell-cell communication modeling. Fortunately, many studies have provided valuable solutions to these challenges[24,54–56] (Figure 4). As a result, refining the graph structure becomes a critical step for extracting high-resolution biological insights.

In this study, we will construct the initial graph using a k-nearest neighbors (k-NN) approach based

Correlation Does Not Imply Causation.

on gene expression similarity. To determine the optimal value of "k", we will employ a parameter sweep, evaluating the performance of downstream GNN tasks (e.g., graph reconstruction, node clustering) for different "k" values using metrics such as the Area Under the Receiver Operating Characteristic curve (AUROC) and Normalized Mutual Information (NMI). We will select the "k" that yields the best performance. To mitigate noise in the initial graph, we will explore techniques such as edge filtering based on the similarity score or using more robust graph construction methods like mutual information networks (MINs) as a baseline comparison. Additionally, we will investigate the integration of prior knowledge from existing gene regulatory databases (e.g., KEGG, STRING) to inform the initial graph structure by weighting or prioritizing edges supported by these databases. By systematically reducing noise and optimizing connectivity, these refinements not only improve the biological relevance of the resulting graphs but also enable more robust and interpretable downstream applications, including graph neural network modeling and causal inference.

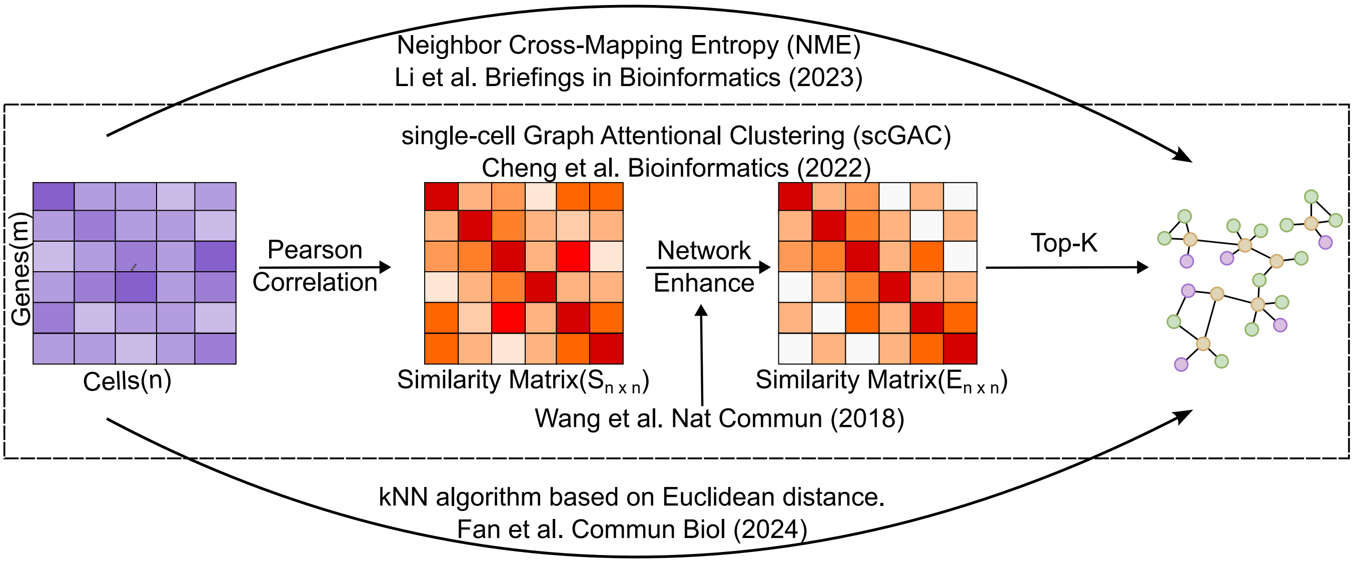

Neighbor Cross-Mapping Entropy (NME)
Li et al. Briefings in Bioinformatics (2023)
single-cell Graph Attentional Clustering (scGAC)
Cheng et al. Bioinformatics (2022)
Genes(m)
Cells(n)
Pearson Correlation
Similarity Matrix($S_{n \times n}$)
Network Enhance
Similarity Matrix($E_{n \times n}$)
Wang et al. Nat Commun (2018)
Top-K
kNN algorithm based on Euclidean distance.
Fan et al. Commun Biol (2024)

Figure 4. Methods for graph construction from single-cell data. This diagram illustrates three distinct approaches for converting a gene expression matrix into a graph representation. (1)Cell similarity is calculated using Pearson Correlation, optionally refined via Network Enhancement[55], and a graph is constructed using Top-K edge selection[54]. (2) The Neighbor Cross-Mapping Entropy (NME) method offers an alternative route for graph construction directly from the expression data[24]. (3) Another approach employs a k-Nearest Neighbors (KNN) algorithm based on Euclidean distance to identify connections and build the graph[56].

2.3.2 Graph Neural Network Architecture

Graph Autoencoder (GAE)[57,58] will be employed to learn low-dimensional representations of genes from single-cell data. The encoder of the GAE will be selected from a set of powerful GNN architectures, including Graph Convolutional Network (GCN), GraphSAGE, Graph Isomorphism Network (GIN), and Graph Attention Network (GAT)[17,57,59,60] (Figure 5). Each of these architectures offers unique strengths in capturing different aspects of the gene regulatory network. For instance, GAT excels at assigning different weights to neighboring nodes, effectively capturing the varying strengths of regulatory relationships. The GAE model will utilize a multi-head attention mechanism, where each head focuses on distinct facets of gene interactions. The outputs from these attention heads will then be aggregated to produce a comprehensive embedding of the network. The choice of the encoder and the specific hyperparameters (number of layers, attention heads, etc.) will be optimized based on the characteristics of the single-cell data and the research question, potentially using techniques like cross-validation on downstream tasks. Non-linear activation functions (e.g., ReLU) will be incorporated to effectively model the complex, non-linear relationships present in the data.

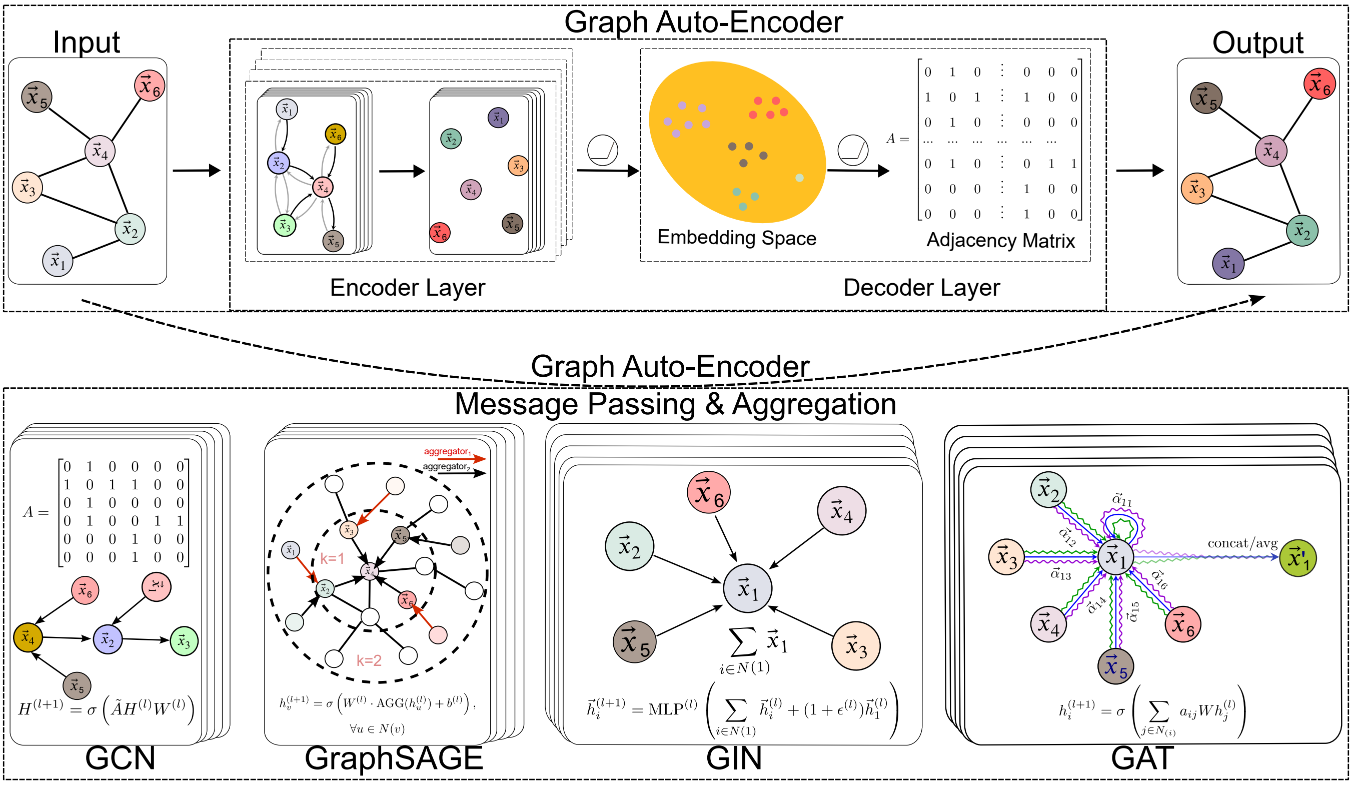

Graph Auto-Encoder
Input
Output
Encoder Layer
Embedding Space
Adjacency Matrix
Decoder Layer
Graph Auto-Encoder
Message Passing & Aggregation
$H^{(l+1)} = \sigma\left(\tilde{A}H^{(l)}W^{(l)}\right)$
aggregator$_1$
aggregator$_2$
k=1
k=2
$h_v^{(l+1)} = \sigma\left(W^{(l)} \cdot \mathrm{AGG}(h_u^{(l)}) + b^{(l)}\right), \forall u \in N(v)$
$\sum_{i \in N(1)} \vec{x}_1$
$\vec{h}_i^{(l+1)} = \mathrm{MLP}^{(l)}\left(\sum_{i \in N(1)} \vec{h}_i^{(l)} + (1+\epsilon^{(l)})\vec{h}_1^{(l)}\right)$
concat/avg
$h_i^{(l+1)} = \sigma\left(\sum_{j \in N_{(i)}} a_{ij} W h_j^{(l)}\right)$
GCN
GraphSAGE
GIN
GAT

Figure 5. Graph autoencoder for learning gene embeddings. (Top) Overview of a Graph Auto-Encoder (GAE), where an input graph is processed by an encoder to generate node embeddings, which are then used by a decoder to reconstruct the graph structure (adjacency matrix). (Bottom) Illustration of common message passing and aggregation strategies used in GNN layers within GAEs, including GCN, GraphSAGE, GIN, and GAT.

2.3.3 Causal Inference Methods for Static and Dynamic GRNs

To identify true causal relationships, the GNN will be integrated with causal inference techniques. Specifically, an additive noise model[11] will test causal direction between gene pairs by comparing model fit in both directions using the low-dimensional gene embeddings learned by the GAE as input features. For static GRN inference, we will apply the additive noise model to all gene pairs in the inferred graph.

For dynamic GRN modeling, if time-series single-cell data is available, we will explore using Granger causality based on the learned embeddings over time. This involves fitting autoregressive models for each gene's expression using the past expression of other genes to determine if one gene's past activity can predict another's future activity.

$$P(X,Y,Z,W) = P(X)P(Y)P(Z \mid X,Y)P(W \mid Z) \qquad BIC = -2log(L) + plog(n)$$

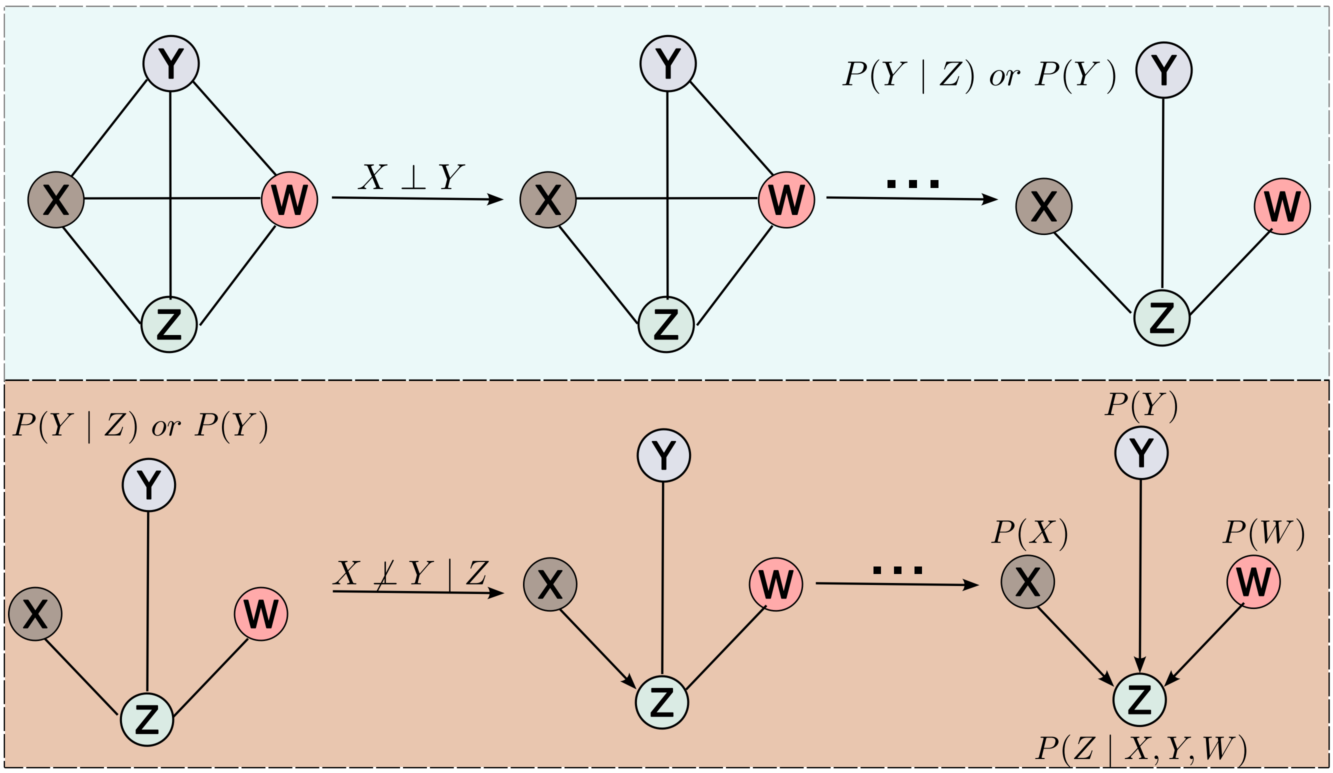


Figure 6. Causal structure learning framework. Illustration of a constraint-based approach to inferring a causal graph. Starting from a fully connected graph (top left), edges are removed based on conditional independence tests (e.g., $X \perp Y$) to identify the graph skeleton. Subsequently (bottom panel), edges are oriented based on conditional independence relations (e.g., identifying v-structures like $X \rightarrow Z \leftarrow Y$ if $X \perp Y|Z$) to obtain a Directed Acyclic Graph (DAG) representing causal relationships.

Do-calculus[18,27] will be incorporated to handle interventions and predict the effects of gene perturbations. Specifically, after inferring the causal structure (represented as a Directed Acyclic Graph - DAG), we will use the do-operator to simulate interventions on specific genes (e.g., setting their expression to a constant value). We will then propagate the effects of these interventions through the inferred causal network to predict changes in the expression of downstream target genes. This will involve applying the rules of do-calculus to estimate the interventional distributions $P(Y|do(X = x))$ for target genes Y given an intervention on gene X. We will leverage the structural information of the inferred DAG to identify appropriate adjustment sets for estimating causal effects and handling potential confounding.

2.3.4 Advantages of the Proposed Methodology

This combined approach offers several advantages over existing methods:

- **Improved Accuracy**: Integrating causal inference with GNNs more accurately identifies regulatory relationships than methods relying solely on correlation.
- **Enhanced Interpretability**: Causal inference clarifies underlying regulatory mechanisms. Additional interpretability comes from embedding visualization (e.g., t-SNE/UMAP) and attention-weight analysis, highlighting the most influential genes and interactions.
- **Robustness to Confounding**: Causal inference methods help control confounding factors, yielding more reliable results.

Moreover, this combined approach aims to capture temporal dependencies and cause-and-effect dynamics within GRNs. By modeling interventions and gene perturbations, we gain deeper insights into how genes influence one another over time, a critical perspective for understanding complex regulatory processes in cellular function and disease.

2.3.5 Validation Strategy

To validate the reconstructed GRNs, we will employ several strategies. First, we will compare the inferred network with known regulatory interactions from established databases. We will evaluate the overlap and accuracy using metrics like precision, recall, and F1-score. Second, we plan to perform in silico perturbation experiments using the inferred causal relationships and do-calculus, where we simulate the effect of knocking out or overexpressing specific genes and compare the predicted downstream effects with existing biological knowledge or experimental data if available. Finally, we will conduct functional enrichment analysis on the identified regulatory modules within the inferred GRN to assess their biological coherence and relevance to known cellular processes. For dynamic GRN validation, if applicable, we will compare our inferred temporal relationships with known temporal regulatory patterns or time-series perturbation experiments from the literature.

2.4 Timeline

(1) **Months 1-3**: Literature review on GNNs, causal inference, and gene regulatory networks. Collection of available single-cell omics datasets for benchmarking and validation [1].

(2) **Months 4-7**: Implement the model architecture, including selection of hyperparameters (layers, attention heads, activation functions).

[1] Literature review will be an ongoing process throughout the project.

(3) **Months 8-11**: Model training and preliminary evaluation on a subset of the data. Explore different causal inference techniques (additive noise model, do-calculus).

(4) **Months 12-14**: Refine the GNN model and causal inference methods based on initial results. Optimize model performance and address challenges related to cyclic relationships in GRNs.

(5) **Months 15-17**: Conduct comprehensive model evaluation using various metrics (accuracy, precision, recall, F1-score, AUROC, AUPRC). Compare the performance of the proposed methodology with existing GRN inference methods.

(6) **Months 18-20**: Focus on model interpretation and visualization. Develop methods to interpret the learned gene embeddings and identify important features contributing to causal relationships.

(7) **Months 21-24**: Validate the inferred GRN using independent datasets or experimental validation techniques.

(8) **Months 25-27**: Write the thesis, incorporating all research findings, methodology, and analysis.

2.5 Conslusion

This research aims to make a significant contribution to the field of gene regulatory inference by developing a novel methodology that combines GNNs and causal inference. By overcoming the limitations of existing methods, this approach has the potential to provide a more accurate, robust, and interpretable understanding of GRNs. The improved accuracy stems from the integration of causal inference, which allows the model to identify true cause-and-effect relationships between genes, going beyond correlation-based analysis. The enhanced interpretability is achieved through the use of visualization techniques and feature importance analysis, providing insights into the underlying mechanisms of gene regulation. The robustness to confounding is addressed by employing causal inference techniques that mitigate the effects of extraneous factors.

This research has significant implications for a wide range of applications, including disease modeling, drug discovery, and personalized medicine. By accurately inferring GRNs, we can gain a deeper understanding of the complex regulatory processes that govern cellular behavior and disease development. This knowledge can be used to identify potential drug targets, predict disease progression, and develop personalized treatment strategies. For example, in the context of cancer research, this approach could be used to identify key genes and pathways that drive tumor growth and metastasis, leading to the development of more effective cancer therapies. Furthermore, this research could contribute to a better understanding of the genetic basis of complex diseases and pave the way for personalized medicine approaches that tailor treatments to individual patients' genetic profiles.

REFERENCES

[1] Nguyen H, Tran D, Tran B, et al. A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. Briefings in bioinformatics, 2021, 22(3):bbaa190.

[2] Dong J, Li J, Wang F. Deep learning in gene regulatory network inference: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2024, 21:2089 - 2101.

[3] Mao G, Pang Z, Zuo K, et al. Predicting gene regulatory links from single-cell rna-seq data using graph neural networks. Briefings in Bioinformatics, 2023, 24(6):bbad414.

[4] Lei Y, Huang X T, Guo X, et al. Deepgrncs: deep learning-based framework for jointly inferring gene regulatory networks across cell subpopulations. Briefings in Bioinformatics, 2024, 25(4):bbae334.

[5] Dibaeinia P, Ojha A, Sinha S. Interpretable ai for inference of causal molecular relationships from omics data. Science Advances, 2025, 11(7):eadk0837.

[6] Hernández-Lemus E, Tovar H. Networks of transcription factors//Genome Plasticity in Health and Disease. Elsevier, 2020:137-155.

[7] Kim D, Tran A, Kim H J, et al. Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. NPJ Systems Biology and Applications, 2023, 9(1):51.

[8] Aibar S, González-Blas C B, Moerman T, et al. Scenic: single-cell regulatory network inference and clustering. Nature methods, 2017, 14(11):1083-1086.

[9] Bravo González-Blas C, De Winter S, Hulselmans G, et al. Scenic+: single-cell multiomic inference of enhancers and gene regulatory networks. Nature methods, 2023, 20(9):1355-1367.

[10] 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):2159-2161.

[11] Feng K, Jiang H, Yin C, et al. Gene regulatory network inference based on causal discovery integrating with graph neural network. Quantitative Biology, 2023, 11(4):434-450.

[12] Badia-i Mompel P, Wessels L, Müller-Dott S, et al. Gene regulatory network inference in the era of single-cell multi-omics. Nature Reviews Genetics, 2023, 24(11):739-754.

[13] Wang J C, Chen Y J, Zou Q. Grace: Unveiling gene regulatory networks with causal mechanistic graph neural networks in single-cell rna-sequencing data. IEEE Transactions on Neural Networks and Learning Systems, 2024:1-13.

[14] Shu H, Zhou J, Lian Q, et al. Modeling gene regulatory networks using neural network architectures. Nature Computational Science, 2021, 1(7):491-501.

[15] Job S, Tao X, Cai T, et al. Exploring causal learning through graph neural networks: an in-depth review. arXiv preprint arXiv:2311.14994, 2023.

[16] Otal H T, Subasi A, Kurt F, et al. Analysis of gene regulatory networks from gene expression using graph neural networks. arXiv preprint arXiv:2409.13664, 2024.

[17] Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826, 2018.

[18] Zečević M, Dhami D S, Veličković P, et al. Relating graph neural networks to structural causal models. arXiv preprint arXiv:2109.04173, 2021.

[19] Mercatelli D, Scalambra L, Triboli L, et al. Gene regulatory network inference resources: A practical overview. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 2020, 1863(6): 194430.

[20] Wang C, Liu Z P. Diffusion-based generation of gene regulatory networks from scrna-seq data with dignet. Genome Research, 2025, 35(2):340-354.

[21] Wang L, Trasanidis N, Wu T, et al. Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. Nature Methods, 2023, 20(9):1368-1378.

[22] Du X, Yang F, Gao W, et al. Causal gnns: A gnn-driven instrumental variable approach for causal inference in networks. arXiv preprint arXiv:2409.08544, 2024.

[23] Chen G, Liu Z P. Graph attention network for link prediction of gene regulations from single-cell rna-sequencing data. Bioinformatics, 2022, 38(19):4522-4529.

[24] Li L, Xia R, Chen W, et al. Single-cell causal network inferred by cross-mapping entropy. Briefings in Bioinformatics, 2023, 24(5):bbad281.

[25] Dai H, Ng I, Luo G, et al. Gene regulatory network inference in the presence of dropouts: a causal view. arXiv preprint arXiv:2403.15500, 2024.

[26] Huynh-Thu V A, Irrthum A, Wehenkel L, et al. Inferring regulatory networks from expression data using tree-based methods. PloS one, 2010, 5(9):e12776.

[27] Pearl J. Causal inference in statistics: An overview. 2009.

[28] Wein S, Malloni W M, Tomé A M, et al. A graph neural network framework for causal inference in brain networks. Scientific reports, 2021, 11(1):8061.

[29] Zhao M, He W, Tang J, et al. A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. Briefings in bioinformatics, 2022, 23(2):bbab568.

[30] Wang J, Ma A, Chang Y, et al. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. Nature communications, 2021, 12(1):1882.

[31] Gu H, Cheng H, Ma A, et al. scgnn 2.0: a graph neural network tool for imputation and clustering of single-cell rna-seq data. Bioinformatics, 2022, 38(23):5322-5325.

[32] Mao G, Pang Z, Zuo K, et al. Gene regulatory network inference using convolutional neural networks from scrna-seq data. Journal of Computational Biology, 2023, 30(5):619-631.

[33] Zhao M, He W, Tang J, et al. A comprehensive overview and critical evaluation of gene regulatory network inference technologies. Briefings in bioinformatics, 2021, 22(5):bbab009.

[34] Keyl P, Bischoff P, Dernbach G, et al. Single-cell gene regulatory network prediction by explainable ai. Nucleic Acids Research, 2023, 51(4):e20-e20.

[35] Cao Z J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nature Biotechnology, 2022, 40(10):1458-1466.

[36] Keil A P, Zadrozny S, Edwards J K. A review and synthesis of multi-level models for causal inference with individual level exposures. Current Epidemiology Reports, 2024, 11(1):54-62.

[37] Wang Q, Guo M, Chen J, et al. A gene regulatory network inference model based on pseudo-siamese network. BMC bioinformatics, 2023, 24(1):163.

[38] Tran A, Yang P, Yang J Y, et al. scremote: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model. NAR genomics and bioinformatics, 2022, 4(1):lqac023.

[39] Wen Y, Huang J, Guo S, et al. Applying causal discovery to single-cell analyses using causalcell. Elife, 2023, 12:e81464.

[40] Shojaee A, Huang S s C. Robust discovery of gene regulatory networks from single-cell gene expression data by causal inference using composition of transactions. Briefings in Bioinformatics, 2023, 24(6):bbad370.

[41] Jereesh A, Kumar G S, et al. Reconstruction of gene regulatory networks using graph neural networks. Applied Soft Computing, 2024, 163:111899.

[42] Yazdani A. Mendelian randomization and causal networks for systematic analysis of omics. 2020. https://arxiv.org/abs/2004.06958.

[43] Tejada-Lapuerta A, Bertin P, Bauer S, et al. Causal machine learning for single-cell genomics. arXiv preprint arXiv:2310.14935, 2023.

[44] Deshpande A, Chu L F, Stewart R, et al. Network inference with granger causality ensembles on single-cell transcriptomics. Cell reports, 2022, 38(6).

[45] Langfelder P, Horvath S. Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics, 2008, 9:1-13.

[46] Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. Proceedings of the National Academy of Sciences, 2019, 116(52):27151-27158.

[47] Qiu X, Rahimzamani A, Wang L, et al. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. Cell systems, 2020, 10(3):265-274.

[48] Ma A, Wang X, Wang C, et al. Deepmaps: Single-cell biological network inference using heterogeneous graph transformer. Biorxiv, 2021:2021-10.

[49] Kartha V K, Duarte F M, Hu Y, et al. Functional inference of gene regulation using single-cell multi-omics. Cell genomics, 2022, 2(9).

[50] Fleck J S, Jansen S M J, Wollny D, et al. Inferring and perturbing cell fate regulomes in human brain organoids. Nature, 2023, 621(7978):365-372.

[51] Nguyen T, Tong A, Madan K, et al. Causal inference in gene regulatory networks with gflownet: Towards scalability in large systems. arXiv preprint arXiv:2310.03579, 2023.

[52] Yuan Q, Duren Z. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. Nature Biotechnology, 2024:1-11.

[53] Xu J, Lu C, Jin S, et al. Deep learning-based cell-specific gene regulatory networks inferred from single-cell multiome data. Nucleic Acids Research, 2025, 53(5):gkaf138.

[54] Cheng Y, Ma X. scgac: a graph attentional architecture for clustering single-cell rna-seq data. Bioinformatics, 2022, 38(8):2187-2193.

[55] Wang B, Pourshafeie A, Zitnik M, et al. Network enhancement as a general method to denoise weighted biological networks. Nature communications, 2018, 9(1):3108.

[56] Fan X, Liu J, Yang Y, et al. scgraphformer: unveiling cellular heterogeneity and interactions in scrna-seq data using a scalable graph transformer network. Communications Biology, 2024, 7(1): 1463.

[57] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. arXiv preprint arXiv:1710.10903, 2017.

[58] Pan S, Hu R, Long G, et al. Adversarially regularized graph autoencoder for graph embedding. arXiv preprint arXiv:1802.04407, 2018.

[59] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. Advances in neural information processing systems, 2017, 30.

[60] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.