Towards Network-Guided Large-Scale Foundation Models on Single-Cell Transcriptomics

Sindhura Kommu

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Xuan Wang, Chair
Yue (Joseph) Wang
Dawei Zhou

May 1st, 2025
Blacksburg, Virginia

Keywords: Gene Regulatory Networks, Single-Cell Foundation Models, Graph Neural Networks

Towards Network-Guided Large-Scale Foundation Models on Single-Cell Transcriptomics

Sindhura Kommu

(ABSTRACT)

Large-scale pretrained models known as foundation models, have made breakthrough progress in the fields like NLP and computer vision. Recently, transformer-based foundation models tailored for single-cell RNA sequencing (scRNA-seq) data have shown significant potential in interpreting the 'languages' of cells through self-supervised learning on huge amounts of unlabeled scRNA-seq datasets [10, 21, 43, 50]. These models could significantly enhance our understanding of cellular functions and disease mechanisms [43]. However, unlike text data, scRNA-seq data is high-dimensional, inherently noisy and sparse, posing unique challenges [28]. We hypothesize that a major limitation of current single-cell foundation models (scFMs) lies in their inability to effectively leverage prior biological knowledge that could provide valuable complementary insights on relationships between various genes. One of the most critical applications of scRNA-seq is the inference of gene regulatory networks (GRNs), which represent the intricate interactions between transcription factors (TFs) and their target genes [3]. In the first part of this thesis, we propose **scRegNet** [30], an innovative framework that combines scFMs with graph-based learning by incorporating experimentally validated transcription factor-DNA binding data in the form of networks with known regulatory interactions for the GRN inference task. **scRegNet** [30] achieved state-of-the-art results in the gene regulatory link prediction task when compared to nine baseline methods across seven scRNA-seq benchmark datasets and demonstrated greater robustness. In the second part of the thesis, we systematically explored incorporating prior GRNs into the pretraining

of scFMs. This exploration provided valuable insights into the benefits and limitations of network guidance, revealing varied effects on predictive accuracy across different downstream tasks related to chromatin and network dynamics.

Towards Network-Guided Large-Scale Foundation Models on Single-Cell Transcriptomics

Sindhura Kommu

(GENERAL AUDIENCE ABSTRACT)

Every cell in our body contains thousands of genes working together in complex networks to control how cells grow, respond to stress, or become specialized. Understanding these gene regulatory networks is crucial for studying diseases, development, and treatment responses. With recent advances in single-cell RNA sequencing, scientists can examine gene activity in individual cells, but making sense of this data requires powerful tools. This thesis explores how large-scale "foundation models" trained on millions of single cells can help uncover hidden gene interactions. In the first part of the study, we introduce a new method called scRegNet, which combines these foundation models with graph-based learning to accurately predict missing or unknown gene connections. This method showed superior performance across a variety of cell types, especially when dealing with noisy data. In the second part, we investigate whether incorporating prior biological knowledge, such as known gene regulatory networks, into the training of foundation models can improve their performance on related tasks. By guiding the learning process with real-world biological graphs, we show that these models become better at identifying important gene regulators. Together, these contributions provide new ways to blend data-driven learning with expert knowledge, helping advance biomedical research and precision medicine.

Dedication

I dedicate this thesis to my big brother, Prateek - my first teacher, my lifelong cheerleader, and the quiet architect of every dream I've dared to chase.

Acknowledgments

I extend my immeasurable gratitude to my advisor, Dr. Xuan Wang. She has been a remarkable mentor and role model, profoundly influencing my growth into a confident researcher. To simply call her "patient" does not capture the unwavering warmth and wisdom with which she met my countless drafts, half-formed ideas, and persistent questions. She always led by example: working tirelessly, treating every student's curiosity with deep respect, and fiercely believing in the power of collaboration. For her profound contribution to my work, my curiosity, and fostering a great sense of what I am capable of, I am eternally grateful. I am also deeply grateful to Dr. Yue (Joseph) Wang for turning collaboration into camaraderie. His valuable insights and consistent encouragement to think critically significantly enriched my research experience. Our conversations were a constant reminder that science is as much about asking the right questions as it is about finding the answers. I will always cherish our intellectually stimulating and personally meaningful collaborations. My heartfelt appreciation also goes to Dr. Dawei Zhou for his constructive feedback and insightful recommendations as a committee member. His course, "Learning on Graphs," provided a strong and essential foundation that greatly benefited my work. I am incredibly thankful for my lab family—Meng, Zhenyu, Gaurav, Jun, Manar, Daniel, and Priya. The whiteboard debates, shared laughter, and warm friendship made this journey both enjoyable and memorable. Lastly, I would like to express my appreciation to my loving husband and family members for their unconditional love and motivation. Their support has been a constant source of strength, especially during the most challenging times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity by enabling precise transcriptomic profiling at unprecedented resolution [25]. This technological advancement has transformed our ability to characterize cellular identity and function across tissues, developmental stages, and disease states, revealing complex regulatory mechanisms at fine-grained resolution. As the volume of scRNA-seq data has grown exponentially, with repositories now containing millions of single-cell transcriptomes across diverse biological contexts, computational methods to extract meaningful biological insights have become increasingly critical.

At the core of cellular identity and function lies the intricate web of gene regulatory networks (GRNs), which orchestrate the precise expression patterns governing cellular differentiation, response to stimuli, and disease progression [3]. Accurate inference of these networks represents one of the most fundamental challenges in computational biology, with profound implications for understanding development, disease mechanisms, and potential therapeutic interventions. Despite the transformative potential of scRNA-seq for GRN inference, several substantial challenges remain. First, scRNA-seq data is inherently sparse, with dropout rates frequently exceeding 80%, creating significant noise that complicates regulatory relationship identification. Second, the high dimensionality of the data often encompassing thousands of genes across thousands to millions of cells creates computational challenges for traditional modeling approaches [23, 34]. Finally, the complex, non-linear relationships

between regulators and their targets demand sophisticated computational strategies capable of capturing multilayered dependencies beyond simple co-expression patterns. While substantial progress has been made through both unsupervised and supervised computational approaches [6, 26, 33, 52], significant gaps remain in our ability to accurately and efficiently reconstruct GRNs from single-cell data. Traditional methods often struggle to integrate prior biological knowledge effectively with the rich information contained in large-scale scRNA-seq datasets.

This thesis addresses two fundamental questions at the intersection of machine learning and network biology: (1) How can we leverage both the contextual information captured by large-scale single-cell foundation models (scFMs) and the structural patterns encoded in known regulatory interactions to improve GRN inference? (2) Can prior biological knowledge be effectively incorporated into the pretraining process of scFMs to enhance their ability to capture meaningful gene-gene relationships?

To address these questions, we develop a dual approach. First, we present **scRegNet** [30], which combines pretrained single-cell foundation models with graph neural networks (GNNs) to create a powerful framework for gene regulatory link prediction. Second, we systematically explore strategies for incorporating prior biological knowledge in the form of GRNs directly into the pretraining process of foundation models, creating network-guided variants that better capture biologically relevant gene relationships.

The primary contributions of this thesis are:

1. Development of **scRegNet** [30], a novel framework that integrates single-cell foundation models with graph neural networks for state-of-the-art gene regulatory link prediction.

2. Comprehensive evaluation demonstrating consistent improvements over existing methods across diverse cell types and experimental conditions.

3. Systematic exploration of strategies for incorporating biological prior knowledge into foundation model pretraining.

4. Empirical demonstration that network-guided foundation models exhibit enhanced performance on downstream tasks related to chromatin and network dynamics.

The remainder of this thesis is organized as follows: Chapter 2 provides a comprehensive review of related work in GRN inference and foundation models for single-cell data. Chapter 3 details the **scRegNet** methodology, experimental design and results. Chapter 4 explores approaches for network-guided pretraining of foundation models and presents results and analyses. Finally, Chapter 5 concludes with a discussion of summary, implications, limitations, and directions for future research.

Chapter 2

Review of the Literature

2.1 Gene Regulatory Network Inference

The challenge of inferring gene regulatory networks has been approached through increasingly sophisticated computational methods, evolving from simple statistical associations to complex deep learning frameworks.

2.1.1 Classical Unsupervised Methods

Early approaches to GRN inference from expression data relied primarily on unsupervised methods, using statistical associations between gene expression patterns to infer potential regulatory relationships. Methods such as GENIE3 [23] and GRNBoost2 [34] employ tree-based regression techniques to identify gene sets co-expressed with transcription factors (TFs), using the importance of a potential target gene in predicting a TF's expression as a proxy for regulatory strength. Similarly, correlation-based approaches such as Pearson correlation coefficient (PCC) quantify linear relationships between gene pairs. These methods, while computationally efficient and applicable without labeled training data, face significant limitations. As noted by Freytag et al. [13], the high dimensionality of gene expression data relative to sample size creates numerous spurious correlations, with many co-expression signals arising purely from chance or systematic noise rather than genuine regulatory relation-

ships. Furthermore, these approaches often struggle to capture non-linear and combinatorial regulatory mechanisms that frequently characterize gene regulation.

2.1.2 Supervised Learning for Regulatory Link Prediction

The growing availability of experimentally validated TF-DNA binding data from resources such as ENCODE [9], ChIP-Atlas [35], and ESCAPE [47] has enabled the development of supervised learning approaches that significantly outperform unsupervised methods. These approaches frame GRN inference as a supervised link prediction task, using known TF-gene interactions as training data to predict novel regulatory relationships. Early supervised methods such as CNNC [52] transformed the problem into image classification by converting gene pair co-expression profiles into histogram-like representations processed by convolutional neural networks (CNNs). This approach marked an important shift toward deep learning for GRN inference but was limited by its reliance on pairwise representations that failed to capture broader network context. More recent approaches have incorporated network structure more explicitly. Kc et al. [26] proposed gene network embedding (GNE) that uses multilayer perceptrons (MLPs) to jointly encode gene expression profiles and network topology, while DeepDRIM [8] extended CNN-based approaches to consider potential neighboring genes. The GRN-transformer [39] leveraged the powerful attention mechanisms of transformer architectures in a weakly supervised framework, demonstrating improved capture of long-range dependencies between genes.

2.1.3 Graph-Based Learning for Regulatory Networks

Graph neural networks (GNNs) have emerged as particularly promising for GRN inference due to their intrinsic ability to model complex interaction patterns within networks. Re-

cent approaches such as GENELink [6] and GNNLink [33] employ graph attention networks (GATs) and graph convolutional networks (GCNs), respectively, to encode both local and global topological features of regulatory networks. GENELink [6] utilizes a GAT-based approach that allows the model to assign different weights to different neighboring nodes, capturing the varying importance of different regulatory relationships. GNNLink [33] employs GCN-based graph encoders to aggregate information from a node's neighborhood, effectively modeling the influence of network structure on gene regulation. Both approaches have demonstrated substantial improvements over previous methods, highlighting the value of graph-based representations for capturing the complex inter-dependencies in gene regulatory systems. Despite these advances, current graph-based approaches face important limitations. They primarily rely on topological features derived from limited training data, without fully leveraging the rich biological context encoded in large-scale transcriptomic datasets. Additionally, these approaches often struggle with the high sparsity and noise characteristic of regulatory networks, particularly when available training data is limited.

2.2 Foundation Models for Single-Cell Transcriptomics

The remarkable success of large-scale pre-trained foundation models have significantly impacted fields such as natural language understanding and computer vision by utilizing deep learning models pre-trained on large-scale datasets, which can then be used for various downstream tasks with limited task-specific data [20, 22, 37]. Similarly, for scRNA-seq data, large-scale pre-trained foundation models have become essential tools for interpreting the 'languages' of cells [10, 21, 43, 50].

2.2.1 Self-Supervised Pretraining Strategies

Foundation models for scRNA-seq data typically employ self-supervised learning objectives on massive unlabeled datasets. The most common approach, analogous to masked language modeling (MLM) in NLP [11], involves randomly masking a subset of gene expression values and training the model to predict these values based on the surrounding genomic context. This pretraining strategy enables models to learn comprehensive representations of gene-gene relationships across diverse cellular contexts without requiring labeled data [50]. These models, known as single-cell foundation models (scFMs), trained on large-scale scRNA-seq data spanning millions of samples, provide rich informative representations for advancing network biology. Several key scFMs have emerged, each with distinct architectural choices and pretraining strategies. scBERT [50] adapts the BERT architecture to scRNA-seq data, using a combination of gene identity embeddings and expression level embeddings processed through transformer encoder layers. Geneformer [43] employs a rank value encoding strategy to represent gene expression, prioritizing genes based on their relative expression within each cell and using a deeper transformer architecture with upto 20 layers. scFoundation [21] introduces an asymmetric encoder-decoder architecture specifically designed to handle the sparsity of scRNA-seq data, with the encoder processing only non-zero expression values and the decoder integrating zero-expressed genes. Most recently, scGPT [10] has adapted the GPT architecture to the single-cell domain, enabling both representation learning and generative capabilities.

2.2.2 Transfer Learning and Downstream Applications

These foundation models, pretrained on millions of single-cell transcriptomes spanning diverse tissues and conditions, have demonstrated remarkable transfer learning capabilities

across a range of downstream tasks. As shown by Yang et al. [50] and Theodoris et al. [43], pretrained representations from these models capture meaningful biological patterns that generalize across cell types and experimental conditions. Applications have included cell type annotation, perturbation response prediction, and trajectory inference, with pretrained models consistently outperforming specialized models trained from scratch. This success demonstrates the value of knowledge transfer from large-scale pretraining, especially when labeled data for specific tasks is limited. Despite these advances, current foundation models for single-cell data face important limitations. Most critically, they are trained exclusively on gene expression data without incorporating prior biological knowledge encoded in gene regulatory networks, protein-protein interaction networks, or other structured biological data. This represents a missed opportunity to guide these models toward biologically meaningful representations that align with established knowledge of cellular processes.

Chapter 3

scRegNet: Prediction of Gene Regulatory Connections with Joint Single-Cell Foundation Models and Graph-Based Learning

3.1 Method

In this section, we discuss each element of **scRegNet** (Fig 3.1) in detail.

3.1.1 Gene Representations from Foundation Models

Recent studies [10, 21, 43, 50], have demonstrated that large-scale pre-trained foundation models possess a strong capacity to model gene-gene interactions across cells, achieving state-of-the-art performance in various single-cell analysis tasks. In this study, we explore three single-cell foundation models (scFMs), scBERT [50], Geneformer [43], and scFoundation [21], to capture the context-aware gene-gene relationships of the scRNA-seq data.

A summary of these three scFMs, including their architectures and key features, is provided in Table 3.1. All of these three scFMs rely on attention-based Transformer architectures [44]

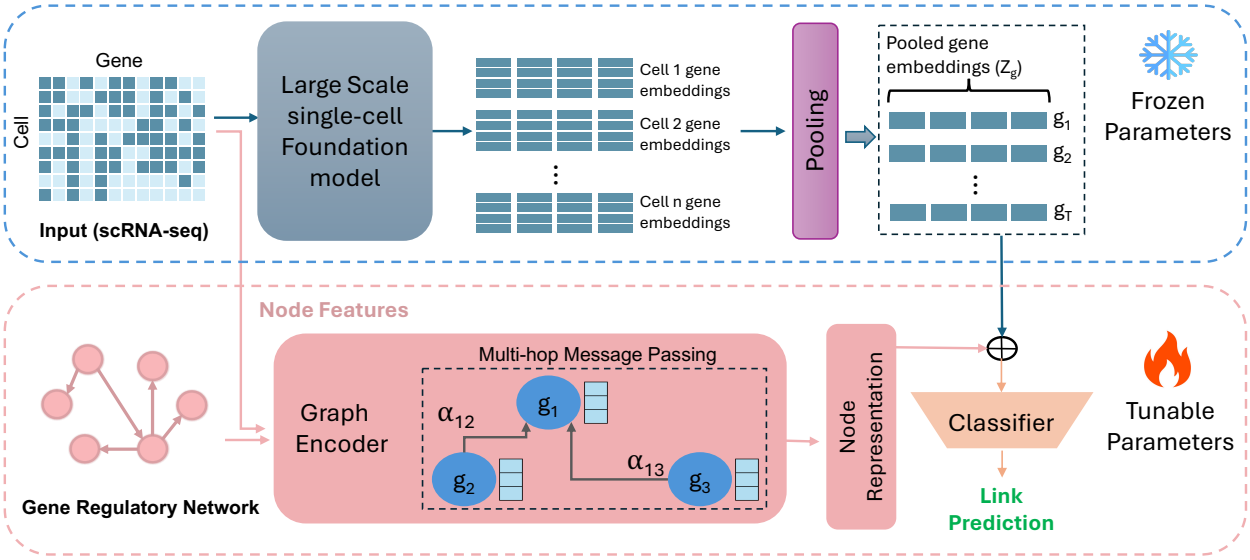


Figure 3.1: **Overview of the scRegNet framework for GRN inference**. SCREGNET utilizes a pre-trained single-cell foundation model (top; Section 3.1.1) to generate gene embeddings from scRNA-seq input which are integrated with outputs from a Graph Encoder (bottom left; Section 3.1.2). The combined representations are fed into a classifier (bottom right; Section 3.1.4) for link prediction, enabling the identification of missing regulatory interactions among genes. The architecture incorporates both frozen parameters for leveraging pre-trained knowledge and tunable parameters for domain-specific learning, facilitating the seamless integration of biological context with learned embeddings.

for processing gene-level vector representations of the scRNA-seq data and employ masked language modeling (MLM) as a self-supervised pre-training strategy to learn multifaceted internal patterns of cells from millions of single-cell transcriptomes. The MLM strategy is the same as that used in pre-training the large language models (LLMs), such as ChatGPT, allowing the LLMs to learn human knowledge from huge archives of natural language texts. However, these three scFMs differ in how they represent the input scRNA-seq data, their model architectures, and their training procedures. Specifically, the input design and pre-processing steps vary for each model as detailed below.

First, we formally define the input scRNA-seq data as a cell-by-gene matrix, $\mathbf{X} \in \mathbb{R}^{N \times T}$, where each element represents the RNA abundance for gene t in cell n. This matrix, referred to as the raw count matrix, is normalized using a log transformation and feature scaling to

Table 3.1: Comparision of the large-scale single-cell foundation models (scFMs)

	Geneformer	**scFoundation**	**scBERT**
Training Size: #single-cells	95M	50M	1M
Model architecture	Encoder Only	Asymmetric Encoder-Decoder	Encoder Only
Design (Layer-Head-Dim)	Transformer: 20-14-896	Encoder Transformer: 12-12-768 Decoder Performer: 6-8-512	Performer: 6-10-200
Input value	Ranked normalized expression values	continuous normalized expression values	binned normalized expression values
Number of input genes	4096 genes with different ranks	19,264 protein-coding or mitochondrial genes	16,906 genes
Masking	Non-Zero genes only	Zero and Non-Zero	Non-Zero genes only
Pre-training Date	Apr-24	Jun-24	Dec-21

ensure compatibility with attention-based architectures. To create a sequence suitable for input for these models, we define a sequence of gene tokens as $\{g_1, ..., g_T\}$, where T is the total number of selected genes in the dataset. Then we go into details of each scFM in how they handle this input data.

scBERT

scBERT [50] utilizes a combination of two features for each gene: (1) a gene ID feature with gene2vec [12] that represents individual genes in a pre-defined vector space, and (2) a gene expression level feature. For each gene token g_t, the initial input representation is constructed as $h_t^0 = emb_{gene2vec}(g_t) + emb_{expr}(g_t)$, where $\text{emb}_{\text{gene2vec}}(\cdot)$ denotes the gene identity embedding and $\text{emb}_{\text{expr}}(\cdot)$ represents the expression level embedding. These input

representations are processed through $L = 6$ successive transformer encoder layers:

$$h_t^{(l)} = Transformer(h_t^{(l-1)}),\ \ l = 1, 2, ..., L. \tag{3.1}$$

The final hidden states $\{h_t^L\}_{t=1}^T$ serve as the 200-dimensional gene-level embeddings, suitable for downstream tasks.

scBERT [50] employs a matrix decomposition variant of the Transformer, known as Performer, to handle longer sequence lengths efficiently. The model is pre-trained via imputation on 5 million cells belonging to a variety of cell types from different sources. To generate emeddings for scBERT, we first requested the checkpoint and data from the corresponding authors. The environment was set up using the scBERT GitHub repository. Log-normalization was performed and cells with less than 200 expressed genes were filtered out.

scFoundation

scFoundation [21] utilizes an asymmetric encoder-decoder architecture [17] that employs attention mechanisms to optimize gene dependency extraction in sparse single-cell data. It also includes an embedding module that converts continuous gene expression scalars into high-dimensional vectors, allowing the model to fully retain the information from raw expression values, rather than discretizing them like other methods. The encoder is designed to only process non-zero and non-masked gene expression embeddings. These encoded embeddings are then recombined with the zero-expressed gene embeddings at the decoder stage to produce final 512-dimensional gene-level representations. These vector representations capture detailed gene dependencies, making them suitable for downstream network biology-based tasks.

scFoundation is pre-trained over 50 million single-cells sourced from a wide range of organs and tissues originating from both healthy and donors with a variety of diseases and cancer types. It employs xTrimogene [18] as a backbone model, a scalable transformer-based architecture that includes an embedding module and an asymmetric encoder-decoder. The encoder is designed to only process nonzero and non-masked gene expression embeddings from the input matrix $\mathbf{X}$, reducing computational load and thus enabling the application of "vanilla transformer blocks to capture gene dependency without any kernel of low-rank approximation". The encoder input is formed as:

$$I = Autobin(X \odot M_{nonzero}) + Lookup(genes) \tag{3.2}$$

where $\odot$ denotes the element-wise product, M_{nonzero} is a mask identifying non-zero elements, and Autobin converts expression values into discretized tokens. The encoder generates gene-level representations using multi-head self-attention:

$$I_{encoder} = Transformer(f_Q(I), f_K(I), f_V(I)) \tag{3.3}$$

where f_Q, f_K, and f_V are linear projections for query, key, and value. These encoded embeddings are then recombined with the zero-expressed gene embeddings at the decoder stage to reconstruct transcriptome-wide embedded representations.

$$I_{full} = W_p(I_{encoder} \oplus I_{zero} \oplus I_{masked}) + b_p \tag{3.4}$$

where $\oplus$ represents concatenation, and W_p, b_p are learned parameters that project the decoder's embedding size. The final gene-level representations are then produced by the de-

coder through:

$$I_{decoder} = Transformer(f_Q(I_{full}), f_K(I_{full}), f_V(I_{full})) \tag{3.5}$$

scFoundation is pre-trained using read-depth-aware (RDA) modeling, an extension of masked language modeling developed to take the high variance in read depth of the data into account. The raw gene expression values are pre-processed using hierarchical Bayesian downsampling in order to generate the input vectors, which can either be the unchanged gene expression profile or where downsampling has resulted in a variant of the data with lower total gene expression counts. After gene expression has been normalized, raw and input gene expression count indicators are represented as tokens which are concatenated with the model input, allowing the model to learn relationships between cells with different read depths.

To generate scFoundation embeddings, we initialized the scFoundation class shared at the official scFoundation GitHub repository. The *01B-resolution* pre-trained model checkpoint was loaded and the embeddings were generated while setting the $input\_type = singlecell$ and $tgthighres = f2$.

Geneformer

Geneformer [43] employs a rank value encoding strategy to represent input scRNA-seq data, prioritizing genes based on their expression value within a cell. To prepare the input data for Geneformer, we utilize the token dictionary (TokenDict) and the gene median file (GeneMedian) provided in the model's repository. These resources ensure that the input is accurately tokenized based on the rank value encoding strategy, maintaining consistency with the pre-trained model. Each gene's expression is normalized relative to a median reference and then

converted into ranked tokens. The input matrix $\mathbf{R}$ is constructed as follows:

$$R = RankValue(X; TokenDict; GeneMedian). \tag{3.6}$$

The RankValue function normalizes each gene's expression using the GeneMedian values and maps them to discrete tokens via the TokenDict. For each single-cell transcriptome, Geneformer embeds each gene into an 896-dimensional space that captures the gene's contextual characteristics within the cell. These contextual embeddings are generated via multi-layer attention mechanisms similar to the equation 3.1, where $h_t^{(0)}$ is the initial embedding for token R_t, and $L = 20$ layers are used in the encoder. The final hidden state, $h_t^{(L)} \in \mathbb{R}^{896}$, represents the context-aware embedding for gene t. To obtain robust and generalizable gene representations, embeddings are extracted from the penultimate layer of the model, as it captures a more abstract and general feature space compared to the final layer.

The model is pre-trained on Genecorpus-95M, which comprises approximately 95 million human single-cell transcriptomes from a broad range of tissues obtained from publicly available data. This method provides a non-parametric representation of each single-cell transcriptome by ranking genes based on their expression levels in each cell and normalizing these ranks within the entire dataset. Consequently, housekeeping genes, which are ubiquitously highly expressed, are normalized to lower ranks, reducing their influence. Only genes detected within each cell are stored, thus reducing the sparsity of the data. The model is pre-trained using a masked learning objective, masking a portion of the genes and predicting the masked genes, which is intended to allow the model to learn gene network dynamics.

To generate embeddings for Geneformer, we downloaded the repository, including pre-trained model checkpoints, from Hugging Face. We pre-processed the raw expression files to ensure the correct naming of columns and then fed them into the Geneformer tokenizer (Transcrip-

tomeTokenizer). Once the dataset had been tokenized, we extracted embeddings using the pre-trained checkpoint (20-layer model) with the *EmbExtractor* method.

Mean Pooling

Upon extracting the gene representations as gene embeddings from the three scFMs as described above, the embeddings corresponding to each gene are further aggregated across all cells within a specific benchmark dataset to establish a cohesive representation for each gene within a given cell type. This is accomplished through mean pooling. For each cell-type specific benchmark dataset, the mean-pooled embedding for gene t is computed as follows:

$$Z_{\text{scFM}}[t] = \frac{1}{N} \sum_{n=1}^{N} h_{t(n)}^{(L)}, \tag{3.7}$$

where $h_{t(n)}^{(L)}$ represents the extracted embedding of gene t within cell n and N indicates the total number of cells within the cell-type-specific benchmark dataset (as detailed in sections 3.1.1-3.1.1). This pooling methodology facilitates a balanced representation that encapsulates the average gene activity throughout the entire dataset.

3.1.2 Graph-based Learning with GNNs

In addition to the context-aware gene-level representations extracted from the scFMs mentioned above, we also extract gene representations that encode the regulatory network topology using GNNs.

The regulatory network topology comes from the gene interactions in the training data using experimentally validated TF-DNA binding data from resources such as ENCODE [9], ChIP-Atlas [35], and ESCAPE [47]. We formulate these gene interactions in the training data as

a known graph between TFs and target genes, with nodes denoting TFs or genes and links symbolizing their regulatory associations. The gene regulatory link prediction task aims to discover any missing interactions between gene pairs that are not included in the training data. Specifically, given the gene interactions in the training data, the graph encoders (GNNs) learn a mapping function that can generate low-dimensional gene embeddings that capture the underlying structure of the gene interactions.

Let the gene interactions in the training data be represented as $G = \{V, E\}$, where V is the set of nodes (genes) and E is the set of edges (regulatory interactions). The goal is to learn effective node representations through message passing, which embeds into each node—information about its multi-hop neighbors. Specifically, each node receives and aggregates messages (i.e., features or embeddings) from its neighboring nodes recursively in multiple layers. Formally, the updated representation ${v_t}^l$ of each node, in l-th layer is given by:

$$ {v_t}^l = \mathcal{M}(\{{v_s}^{l-1} : s \in \eta_t\}, {v_t}^{l-1}; \theta^l) \tag{3.8} $$

where η_t represents the set of neighboring nodes for an arbitrary node t, and $\mathcal{M}(\cdot)$, parameterized by θ_l in the l-th layer, is the message passing function for neighborhood aggregation. The neighborhood aggregation varies depending on the type of GNN.

To derive the initial features of the genes, we apply pre-processing operations to the raw single-cell expression data. Here in scRegNet, a simple Graph Convolutional Network (GCN) [29] is employed. We noticed that this simple architecture is adequate to reach similar performance compared to computation-demanding GNN frameworks such as Graph Attention Networks (GAT) [45] and GraphSAGE (SAmple and aggreGatE) [19]. A detailed GNN framework comparison and analysis can be found in the Results Section 3.3.

3.1.3 Unified Gene Representations

After extracting gene representations from both the scFMs and the GNNs for a given cell-type-specific benchmark dataset, we integrate them into a unified gene representation for each gene as illustrated in Fig 3.1. This integration involves concatenating the representations from scFM (capturing contextual gene interactions) and GNN (capturing network topology) as follows:

$$Z_{\text{scFM}}[t] \oplus Z_{\text{GNN}}[t] = Z_{\text{joint}}[t], \tag{3.9}$$

where $Z_{\text{scFM}}[t]$ represents the foundation model-derived representation for gene t based on the gene expression profiles in the benchmark dataset (section 3.1.1) and $Z_{\text{GNN}}[t]$ represents the node representation of gene t derived from the GNN encoder trained on the corresponding cell-type-specific network (section 3.1.2). The concatenated representation $Z_{\text{joint}}[t]$ serves as the unified representation for gene t, effectively capturing both gene expression context and network structural information specific to the given dataset.

3.1.4 Link Prediction Layer

The link prediction module constitutes the final component of scRegNet, specifically designed to evaluate the likelihood of unseen regulatory interactions between the gene pairs. We employ MLP networks integrated with ReLU activation functions and Dropout regularization for this task.

For each gene pair (i, j), unified feature representations $Z_{\text{joint}}[i]$ and $Z_{\text{joint}}[j]$ as derived above are processed through MLP. The two outputs from the MLP are concatenated to form a combined representation that captures the joint features of the gene pair. This concatenated representation is passed to a fully connected classification layer. This classification layer

predicts the likelihood of a regulatory interaction by outputting a score for each possible class (presence or absence of an interaction). The predicted scores are normalized using a softmax function, yielding probabilities for each class, as shown in the following equation:

$$\hat{P} = \text{Softmax}(FCN(\text{MLP}(Z_{\text{joint}}[i]) \oplus \text{MLP}(Z_{\text{joint}}[j]))), \tag{3.10}$$

where $\oplus$ represents the concatenation operation, and FCN denotes the final fully connected network that maps the combined representation to the output probabilities. The predicted probabilities correspond to the likelihood of the presence ($\hat{Y} = 1$) or absence ($\hat{Y} = 0$) of a regulatory interaction.

3.1.5 Model Training

To train the scRegNet model, we employ the Binary Cross-Entropy (BCE) loss function, which measures the difference between the predicted regulatory interaction probabilities and the ground-truth labels in the training dataset:

$$\text{BCE} = -\sum_{i=1}^{K} \left[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i) \right], \tag{3.11}$$

where K is the total number of gene pairs in training data, y_i is the ground-truth label for the i-th gene pair ($y_i = 1$ for interaction, $y_i = 0$ for no interaction), p_i is the predicted probability of a regulatory interaction for the i-th pair. The BCE loss is backpropagated through the scRegNet framework, enabling end-to-end optimization of the model parameters. The parameters of GNN layers are updated while the parameters of the scFM remain frozen during training as shown in Fig 3.1.

3.2 Experimental Setup

3.2.1 Datasets and Pre-processing

We evaluate **scRegNet** on seven scRNA-seq benchmark datasets provided in BEELINE [36], more specifically 1) human embryonic stem cells (hESC), 2) human mature hepatocytes (hHEP), 3) mouse dendritic cells (mDC), 4) mouse embryonic stem cells (mESC), 5) mouse hematopoietic stem cells of the erythroid lineage (mHSC-E), 6) mouse hematopoietic stem cells with a granulocyte-monocyte lineage (mHSC-GM), and 7) mouse hematopoietic stem cells with a lymphoid-like lineage (mHSC-L). Following GENELink [6] and GNNLink [33], we adopt the cell-type-specific ChIP-seq networks from the aforementioned datasets as ground truth to evaluate the performance of **scRegNet** and baseline methods.

Following the original paper of BEELINE [36] that provided the seven benchmark datasets, we pre-process each scRNA-seq dataset by only inferring the interactions outgoing from TFs. Following BEELINE [36], we respectively select 500 and 1000 significantly most-varying genes with all TFs whose corrected P-value (Bonferroni method) of variance is lower than 0.01 as the ground truth network for gene regulatory link prediction. The seven scRNA-seq datasets can be downloaded from Gene Expression Omnibus with the accession numbers GSE81252 (hHEP), GSE75748 (hESC), GSE98664 (mESC), GSE48968 (mDC) and GSE81682 (mHSC).

For a fair comparison with existing state-of-the-art baseline models (Section **??**), we follow the same evaluation strategy as GENELink [6] to split the ground truth networks into training/validation/test sets in all benchmark datasets. In these ground truth networks, the number of TFs is limited, and most of them are with high degrees. To validate that the supervised model can distinguish the much more subtle differences between target and non-target genes for each TF, we divide the positive and negative target genes of each TF

in proportion to the training and test datasets.

Specifically, for each transcription factor (TF), the interactions (edges) with target genes are categorized into positive and negative samples. Positive samples represent true regulatory relationships supported by experimental evidence, such as ChIP-seq data from ground-truth networks. Negative samples, on the other hand, consist of gene pairs with no known regulatory interactions. To ensure a robust evaluation framework, the positive and negative samples for each TF are divided proportionally into training and test sets, maintaining a fixed ratio of 67% for training and 33% for testing. This partitioning ensures a consistent evaluation process across all TFs. Additionally, a small subset of the training data (10%) is reserved as a validation set for hyperparameter tuning and early stopping during model training. The data splitting is performed per TF, ensuring that all TFs contribute examples to both the training and test sets. Crucially, this partitioning strategy prevents data leakage by ensuring that the same gene does not appear in both the training and test sets for the same TF. This approach guarantees that the model's performance is evaluated on entirely independent data, maintaining the integrity of the evaluation process. The sizes of each ground-truth network training set are listed in Table 3.2.

Table 3.2: The statistics of prior networks with TFs and 500 (1000) most-varying genes

Cell Type	**Species**	**#Cells**	**#TFs**	**#Genes**	**Density**	**Training Size**	**Test Size**
hESC	human	759	34 (34)	815 (1260)	0.164 (0.165)	20677 (32065)	7142 (11047)
hHEP	human	426	30 (31)	874 (1331)	0.379 (0.377)	19002 (30026)	6563 (10348)
mDC	mouse	384	20 (21)	443 (684)	0.085 (0.082)	10969 (18556)	3792 (6395)
mESC	mouse	422	88 (89)	977 (1385)	0.345 (0.347)	65895 (96460)	22736 (33229)
mHSC-E	mouse	1072	29 (33)	691 (1177)	0.578 (0.566)	13632 (26565)	4718 (9164)
mHSC-GM	mouse	890	22 (23)	618 (1089)	0.543 (0.565)	9280 (17406)	3216 (6003)
mHSC-L	mouse	848	16 (16)	525 (640)	0.525 (0.507)	5976 (7392)	2076 (2560)

3.2.2 Parameter Settings

We leveraged Optuna [2], a powerful hyperparameter optimization framework, to systematically explore the search space. Detailed information on the hyperparameter settings can be found in Table 3.3. For each combination of dataset and model, we conducted 50 optimization trials, ensuring a thorough investigation of potential configurations. Optuna's Tree-structured Parzen Estimator (TPE) was employed to intelligently navigate the search space, prioritizing regions with higher potential based on previous trials. We also applied early stopping criteria to save computational resources by halting underperforming trials early.

Table 3.3: * Applicable only for Graph Attention Network models

hyperparameter	search space	type
learning rate	[1e-5, 1e-2]	continual
weight decay	[1e-5, 1e-4]	continual
dropout	[0.1, 0.8]	continual
#GNN layers	[1, 6]	discrete
#MLP layers	[1, 6]	discrete
batch size	[32, 56]	discrete
hidden dimension	[4, 256]	discrete
#attention heads*	[1, 8]	discrete
reduction*	[concatenate, mean]	discrete
alpha*	[0.01, 0.5]	continual
optimizer	[Adam, RMSprop, SGD]	discrete

3.2.3 Baseline models and evaluation metrics

We compare scRegNet against nine baseline methods for gene regulatory link prediction from single-cell RNA-seq data, which have been proven to achieve good performance. These methods include traditional statistical techniques, machine learning algorithms, and deep learning models, applied to single-cell RNA-seq. We use the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) as the evaluation metrics.

- GNNLink [33] is a graph neural network model that uses a GCN-based interaction graph encoder to capture gene expression patterns.
- GENELink [6] proposes a graph attention network (GAT) approach to infer potential GRNs by leveraging the graph structure of gene regulatory interactions.
- GNE (gene network embedding) [26] proposes a multilayer perceptron (MLP) approach to encode both gene expression profiles and network topology for predicting gene dependencies.
- CNNC [52] proposes inferring GRNs using deep convolutional neural networks (CNNs).
- DeepDRIM [8] is a supervised deep neural network that utilizes images representing the expression distribution of joint gene pairs as input for binary classification of regulatory relationships, considering both target TF-gene pairs and potential neighbor genes.
- GRN-transformer [39] is a weakly supervised learning method that utilizes axial transformers to infer cell type-specific GRNs from single-cell RNA-seq data and generic GRNs.
- Pearson correlation coefficient (PCC) [38] is a traditional statistical method for measuring the linear correlation between two variables, often used as a baseline for GRN inference.

- GRNBoost2 [34] is a gradient boosting-based method for GRN inference.
- GENIE3 [23] is a random forest-based machine learning method that constructs GRNs based on regression weight coefficients, and won the DREAM5 In Silico Network Challenge in 2010.

3.3 Results and Discussion

3.3.1 Performance on Benchmark Datasets

As demonstrated in Tables 3.4 and 3.5, all variants of **scRegNet** (w/ scBERT, Geneformer, and scFoundation) consistently surpass existing baseline models in both AUROC and AUPRC metrics across all seven cell-type-specific datasets (hESC, hHEP, mDC, mESC, mHSC-E, mHSC-GM, mHSC-L). The Geneformer- and scFoundation-based configurations achieved the highest performance, with **scRegNet**(w/ Geneformer) delivering an average improvement of 7.4% and 6.9% in AUROC and 18.6% and 4.1% (Table 3.4) in AUPRC over GNNLink and GENELink respectively, on datasets with 500 most-variable genes (TFs+500). Similarly, for TFs+1000 datasets, **scRegNet**(w/ Geneformer) outperformed GENELink and GNNLink by 6.2% and 7.5% in AUROC and 16.5% and 3.9% in AUPRC (Table 3.5) respectively. Among the three scFM backbone configurations, Geneformer and scFoundation yielded slightly better results compared to scBERT.

Traditional methods, such as GRNBOOST2, GENIE3, and PCC, demonstrated limited predictive accuracy, particularly in AUPRC, due to their reliance on simplistic pairwise correlation metrics. In contrast, graph-based deep learning frameworks like GNNLink and GENELink improved performance by leveraging gene-gene interactions, but their effectiveness remained limited. **scRegNet** consistently outperformed these approaches across di-

Table 3.4: Link prediction performance on seven scRNA-seq datasets with **500 most-variable genes**. Each dataset includes a cell-type-specific ground-truth network. The values reported are averages from 50 independent evaluations per cell type. **scRegNet** utilizing the three backbone models—scBERT, Geneformer, and scFoundation—consistently outperforms the baselines.

Method		Cell Type						
		hESC	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GRNBOOST2	AUROC	0.49	0.52	0.52	0.53	0.53	0.50	0.52
Moerman et al. [34]	AUPRC	0.15	0.38	0.06	0.32	0.57	0.52	0.5
GENIE3	AUROC	0.50	0.54	0.50	0.50	0.52	0.53	0.52
Huynh-Thu et al. [23]	AUPRC	0.15	0.39	0.05	0.31	0.56	0.53	0.50
PCC	AUROC	0.47	0.49	0.54	0.51	0.49	0.54	0.55
Salleh et al. [38]	AUPRC	0.14	0.35	0.06	0.31	0.56	0.53	0.52
GRN-Transformer	AUROC	0.51	0.49	0.50	0.53	0.64	0.50	0.64
Shu et al. [39]	AUPRC	0.15	0.35	0.06	0.49	0.71	0.66	0.64
DeepDRIM	AUROC	0.63	0.52	0.50	0.51	0.56	0.64	0.58
Chen et al. [8]	AUPRC	0.13	0.39	0.06	0.46	0.76	0.64	0.59
CNNC	AUROC	0.68	0.64	0.54	0.73	0.67	0.69	0.67
Yuan and Bar-Joseph [52]	AUPRC	0.25	0.46	0.06	0.48	0.74	0.68	0.64
GNE	AUROC	0.67	0.80	0.52	0.81	0.82	0.83	0.77
Kc et al. [26]	AUPRC	0.34	0.65	0.06	0.64	0.80	0.78	0.70
GENELink	AUROC	0.82	0.84	0.71	0.88	0.87	0.89	0.83
Chen and Liu [6]	AUPRC	0.50	0.70	0.11	0.75	0.89	0.89	0.83
GNNLink	AUROC	0.85	0.82	0.70	0.84	0.83	0.89	0.84
Mao et al. [33]	AUPRC	0.52	0.75	**0.25**	0.76	0.88	0.89	0.85
scRegNet	AUROC	0.88±0.00	**0.90**±0.00	0.75±0.01	0.92±0.00	**0.92**±0.00	0.92±0.00	0.85±0.01
(w/ scBERT)	AUPRC	0.61±0.00	0.83±0.00	0.12±0.01	0.84±0.00	**0.94**±0.00	0.93±0.00	0.85±0.01
scRegNet	AUROC	**0.89**±0.00	**0.90**±0.00	**0.81**±0.00	**0.93**±0.00	**0.92**±0.00	**0.93**±0.00	**0.88**±0.00
(w/ scFoundation)	AUPRC	**0.62**±0.00	0.83±0.00	0.15±0.01	**0.86**±0.00	**0.94**±0.00	**0.94**±0.00	**0.88**±0.00
scRegNet	AUROC	**0.89**±0.00	**0.90**±0.00	**0.81**±0.00	**0.93**±0.00	**0.92**±0.00	**0.93**±0.00	**0.88**±0.00
(w/ Geneformer)	AUPRC	**0.62**±0.00	**0.84**±0.00	0.17±0.00	**0.86**±0.00	**0.94**±0.00	**0.94**±0.00	**0.88**±0.00

verse cell types and datasets. For example, as shown in Table 3.4, **scRegNet** achieved an AUPRC of 0.62 in hESC, representing a +24% improvement over GENELink (AUPRC = 0.50) and a +21.6% improvement over GNNLink (AUPRC = 0.51) under comparable conditions. **scRegNet** demonstrated superior performance in challenging scenarios with sparse regulatory signals, emphasizing the value of integrating foundation model embeddings to capture context-aware gene relationships and overcome the limitations of correlation-based and graph-only methods.

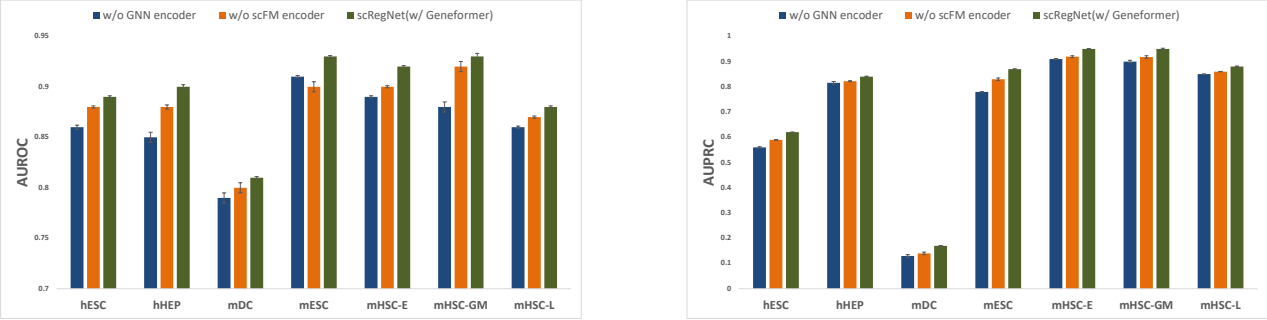
Table 3.5: Link prediction performance on seven scRNA-seq datasets with **1000 most-variable genes**. Each dataset includes a cell-type-specific ground-truth network. The values reported are averages from 50 independent evaluations per cell type. **SCREGNET** utilizing the three backbone models—scBERT, Geneformer, and scFoundation—consistently outperforms the baselines.

Method		Cell Type						
		hESC	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GRNBOOST2	AUROC	0.49	0.52	0.53	0.53	0.51	0.49	0.53
Moerman et al. [34]	AUPRC	0.14	0.37	0.05	0.32	0.54	0.52	0.48
GENIE3	AUROC	0.50	0.54	0.52	0.50	0.50	0.51	0.52
Huynh-Thu et al. [23]	AUPRC	0.15	0.38	0.05	0.31	0.54	0.53	0.48
PCC	AUROC	0.47	0.50	0.54	0.51	0.49	0.54	0.55
Salleh et al. [38]	AUPRC	0.14	0.34	0.05	0.31	0.53	0.54	0.51
GRN-Transformer	AUROC	0.67	0.58	0.57	0.50	0.59	0.53	0.58
Shu et al. [39]	AUPRC	0.16	0.53	0.05	0.51	0.69	0.61	0.52
DeepDRIM	AUROC	0.56	0.63	0.50	0.62	0.50	0.66	0.57
Chen et al. [8]	AUPRC	0.19	0.46	0.06	0.46	0.73	0.64	0.48
CNNC	AUROC	0.72	0.66	0.56	0.73	0.72	0.69	0.62
Yuan and Bar-Joseph [52]	AUPRC	0.27	0.49	0.05	0.50	0.77	0.73	0.56
GNE	AUROC	0.68	0.81	0.52	0.82	0.84	0.84	0.77
Kc et al. [26]	AUPRC	0.34	0.66	0.05	0.65	0.81	0.81	0.68
GENELink	AUROC	0.83	0.85	0.74	0.89	0.90	0.90	0.84
Chen and Liu [6]	AUPRC	0.50	0.71	0.12	0.76	0.90	0.91	0.81
GNNLink	AUROC	0.80	0.84	0.78	0.84	0.87	0.92	0.86
Mao et al. [33]	AUPRC	0.51	0.78	**0.21**	0.78	0.93	0.93	0.86
SCREGNET	AUROC	0.86±0.00	0.90±0.00	0.75±0.01	**0.93**±0.01	0.92±0.00	0.92±0.00	0.81±0.01
(w/ scBERT)	AUPRC	0.56±0.01	0.83±0.00	0.14±0.01	0.86±0.01	0.94±0.00	0.94±0.00	0.79±0.01
SCREGNET	AUROC	**0.88**±0.00	**0.91**±0.00	0.82±0.00	**0.93**±0.00	**0.94**±0.00	**0.94**±0.00	0.87±0.00
(w/ scFoundation)	AUPRC	**0.62**±0.00	**0.85**±0.00	0.15±0.00	0.86±0.00	**0.95**±0.00	**0.95**±0.00	0.86±0.00
SCREGNET	AUROC	**0.88**±0.00	0.90±0.00	**0.84**±0.00	**0.93**±0.00	**0.94**±0.00	**0.94**±0.00	**0.88**±0.00
(w/ Geneformer)	AUPRC	**0.62**±0.00	0.84±0.00	0.17±0.01	**0.87**±0.00	**0.95**±0.00	**0.95**±0.00	**0.87**±0.00

3.3.2 Ablation Study

To understand the contribution of each component within **SCREGNET**, we performed a series of ablation studies, with the results displayed in Figure 3.2. The first experiment involved removing the GNN encoder, which led to a substantial decline in performance, highlighting the critical role of graph-based representation learning in refining gene embeddings. In the second ablation, we excluded the pre-trained foundation model embeddings. This omission impaired performance, demonstrating the importance of capturing diverse cellular contexts through pre-trained embeddings for accurate GRN inference. In these experiments, we uti-

lized Geneformer as the scFM backbone and GCN as the GNN backbone for the model. For each dataset, the AUROC score is calculated as the average of the AUROC values from the TF+500 and TF+1000 datasets. Similarly, the AUPRC score is computed as the average of the AUPRC values from these two networks.



(a) AUROC (b) AUPRC

Figure 3.2: Ablation study validating the contributions of the GNN encoder and scFM (w/ Geneformer) encoder in scRegNet, evaluated using cell-type-specific GRNs. The analysis considers networks with TFs + 500 and TFs + 1000 genes, and the reported scores represent the average AUROC (left) and AUPRC (right) across both configurations, highlighting the impact of each component on model performance.

3.3.3 Impact of GNN architecture

To understand the effective choice of GNN architecture, we employ three distinct GNN architectures: GCN [29], GraphSAGE [19], and GAT [45]. *GCN* is designed to capture the local structure of the graph by aggregating features from a node's immediate neighbors. *GraphSAGE* builds upon GCN by enabling the aggregation of information from sampled neighborhoods, rather than requiring all neighbors to be included. *GAT* introduces attention mechanisms that weigh the importance of different neighbors during the message-passing process.

We evaluated scRegNet(w/ Geneformer) with all three of the above-mentioned GNN architectures. The evaluation results are detailed in Table 3.6. This analysis reveals an interesting

phenomenon: after extensive hyperparameter tuning, all GNN variants show similar performance, with only slight differences observed among them. This similarity in performance can be attributed to the sparsity of the networks [4]. This phenomenon underscores the importance of considering graph sparsity when designing and applying GNNs. It suggests that in some cases, simpler GNN architectures may be sufficient for sparse biological networks, and that efforts to improve performance might be better directed towards graph construction and feature engineering rather than increasing model complexity.

The experiments were conducted using two different sets of most variable genes combined with transcription factors (TFs): (a) 500 most variable genes and (b) 1000 most variable genes.

3.3.4 Robustness Study

While our methodology leverages experimentally validated gene regulations, we acknowledge that in practice, these interactions may contain noise and false positives. Therefore, it is essential to evaluate the robustness of our model when subjected to noisy priors. To address this, we assessed the performance of scRegNet(w/ Geneformer), under various levels of noise-corrupted training data. We introduced controlled perturbations to the priors by flipping the labels of positive instances to negative and vice versa, simulating noise levels of 1%, 2%, 3%, and 4% in the training data. For each noise level, we generated 10 distinct noise-corrupted priors to ensure diverse variations. The performance of scRegNet(w/ Geneformer) was evaluated against each corrupted prior, with results visualized through box plots to illustrate performance variations.

To contextualize the robustness of scRegNet(w/ Geneformer), we benchmarked its performance against the baseline method, GENELink. The results demonstrated the stability and

Table 3.6: Comparative Analysis of Link Prediction Performance in GRNs Using Popular GNN Variants as Backbone Graph-Based Encoders for scRegNet, with Geneformer Serving as the Foundation Model Backbone.

Cell Type		**GCN**	**SAGE**	**GAT**
hESC	AUROC	0.886 ± 0.002	0.884 ± 0.004	0.886 ± 0.004
	AUPRC	0.622 ± 0.004	0.623 ± 0.003	0.614 ± 0.002
hHEP	AUROC	0.902 ± 0.002	0.901 ± 0.003	0.901 ± 0.003
	AUPRC	0.841 ± 0.003	0.839 ± 0.004	0.838 ± 0.003
mDC	AUROC	0.808 ± 0.003	0.788 ± 0.001	0.808 ± 0.002
	AUPRC	0.167 ± 0.002	0.158 ± 0.002	0.167 ± 0.001
mESC	AUROC	0.926 ± 0.001	0.926 ± 0.004	0.926 ± 0.002
	AUPRC	0.860 ± 0.001	0.859 ± 0.001	0.859 ± 0.003
mHSC-E	AUROC	0.923 ± 0.001	0.923 ± 0.001	0.921 ± 0.004
	AUPRC	0.941 ± 0.002	0.941 ± 0.003	0.939 ± 0.002
mHSC-GM	AUROC	0.931 ± 0.002	0.930 ± 0.002	0.929 ± 0.003
	AUPRC	0.939 ± 0.001	0.939 ± 0.001	0.938 ± 0.003
mHSC-L	AUROC	0.879 ± 0.002	0.877 ± 0.002	0.878 ± 0.002
	AUPRC	0.879 ± 0.002	0.874 ± 0.001	0.876 ± 0.003

(a) TFs + 500 Most Variable Genes

Cell Type		**GCN**	**SAGE**	**GAT**
hESC	AUROC	0.883 ± 0.002	0.882 ± 0.002	0.885 ± 0.002
	AUPRC	0.624 ± 0.003	0.620 ± 0.004	0.625 ± 0.001
hHEP	AUROC	0.905 ± 0.004	0.905 ± 0.003	0.905 ± 0.002
	AUPRC	0.844 ± 0.001	0.844 ± 0.002	0.843 ± 0.002
mDC	AUROC	0.838 ± 0.002	0.835 ± 0.003	0.838 ± 0.004
	AUPRC	0.162 ± 0.002	0.146 ± 0.002	0.162 ± 0.003
mESC	AUROC	0.931 ± 0.002	0.931 ± 0.001	0.930 ± 0.003
	AUPRC	0.866 ± 0.002	0.866 ± 0.002	0.865 ± 0.004
mHSC-E	AUROC	0.941 ± 0.004	0.941 ± 0.002	0.940 ± 0.002
	AUPRC	0.956 ± 0.003	0.954 ± 0.001	0.953 ± 0.002
mHSC-GM	AUROC	0.936 ± 0.003	0.937 ± 0.004	0.935 ± 0.003
	AUPRC	0.949 ± 0.002	0.950 ± 0.001	0.948 ± 0.003
mHSC-L	AUROC	0.876 ± 0.001	0.876 ± 0.003	0.876 ± 0.002
	AUPRC	0.865 ± 0.001	0.865 ± 0.001	0.864 ± 0.004

(b) TFs + 1000 Most Variable Genes

resilience of **scRegNet** under noisy training data, with consistently superior performance compared to GENELink, even as noise levels increased. These findings, depicted in Figure 3.3, underscore the reliability of **scRegNet**(w/ Geneformer) in leveraging experimentally validated gene regulations, even in the presence of noise. This robustness positions the model as a reliable choice for real-world applications with noisy training data.

To further identify the threshold at which the performance of our model significantly deteriorates, we extended the noise perturbation analysis to higher noise levels (Table 3.4).

3.3.5 scRegNet Infers Biologically Meaningful Interactions

Human embryonic stem cells (hESCs) rely on complex GRNs to maintain pluripotency and orchestrate early development. First, we observe that **scRegNet** predicts several important regulations in the test set that were otherwise missed by the baseline methods such as GENELink. We further analyze novel regulatory interactions that were not part of the original curated dataset and were missed by baseline methods. The analysis concentrated on predicted positive interactions with a probability greater than 85%, involving key TFs such as classic pluripotency factors NANOG and SOX2, early developmental regulator OTX2, a forkhead factor FOXP1, and stress/differentiation-associated factors like JUND (AP-1 family), each known for roles in stem cell fate decisions. **scRegNet** identified 109 such novel TF→target interactions in the hESC context. The predicted target genes range from signaling receptors and transcriptional regulators to cell-cycle and epigenetic factors. We highlight several specific TF-target pairs from this novel set and examine supporting experimental evidence for their functional roles in stem cell biology, pluripotency maintenance, or early differentiation.

Our analysis demonstrates that novel interactions identified by **scRegNet** in the TFs+500

network for hESCs reveal biologically significant regulatory relationships supported by experimental literature. For instance, the prediction of NANOG→BMPR1A aligns with evidence that NANOG antagonizes BMP signaling pathways, helping to maintain pluripotency by limiting differentiation cues [42]. Additionally, the predicted JUND→SNAI2 interaction is corroborated by findings that AP-1 family transcription factors, including JUND, directly activate SNAI2, promoting epithelial-to-mesenchymal transitions essential during early development [5]. Furthermore, the novel interaction OTX2→CITED2 fits well with studies demonstrating OTX2's role in driving differentiation by counteracting pluripotency maintenance programs, potentially through downregulating CITED2-mediated stabilization of core pluripotency factors such as NANOG and OCT4 [1, 31]. Collectively, these experimentally supported predictions underscore **scRegNet** the ability to uncover critical, previously unrecognized regulatory links in stem cell biology, providing valuable targets for further experimental validation and functional studies.

Table 3.7: Biologically relevant TF–target interactions predicted by **scRegNet** (Geneformer backbone) on the hESC TFs+500 network. Prior Network indicates whether the interaction was documented in training/test sets.

TF	Target Gene	Biological Role/Pathway	GENELink	Prior Network	Literature Support
FOXP1	NANOG	Pluripotency maintenance	Predicted	Absent	[15]
FOXP1	GDF3	Signaling (Nodal/TGF- pathway)	Predicted	Absent	[15]
NANOG	OTX2	Pluripotency maintenance	Not Predicted	Absent	[41]
NANOG	BMPR1A	Pluripotency maintenance	Not Predicted	Absent	[42]
OTX2	NANOG	Differentiation	Not Predicted	Absent	[16]
OTX2	CITED2	Differentiation	Not Predicted	Absent	[1, 31]
NANOG	GATA6	Differentiation (Endoderm lineage)	Not Predicted	Absent	[40]
JUND	CDH2	EMT (Mesenchymal marker)	Not Predicted	Absent	[51]
JUND	SNAI2	EMT/Differentiation	Not Predicted	Absent	[5]
SOX2	CDK6	Cell cycle regulation	Not Predicted	Absent	[32]
SMAD2/3	NANOG	Signaling (Activin/TGF- pluripotency)	Predicted	Present	[48]

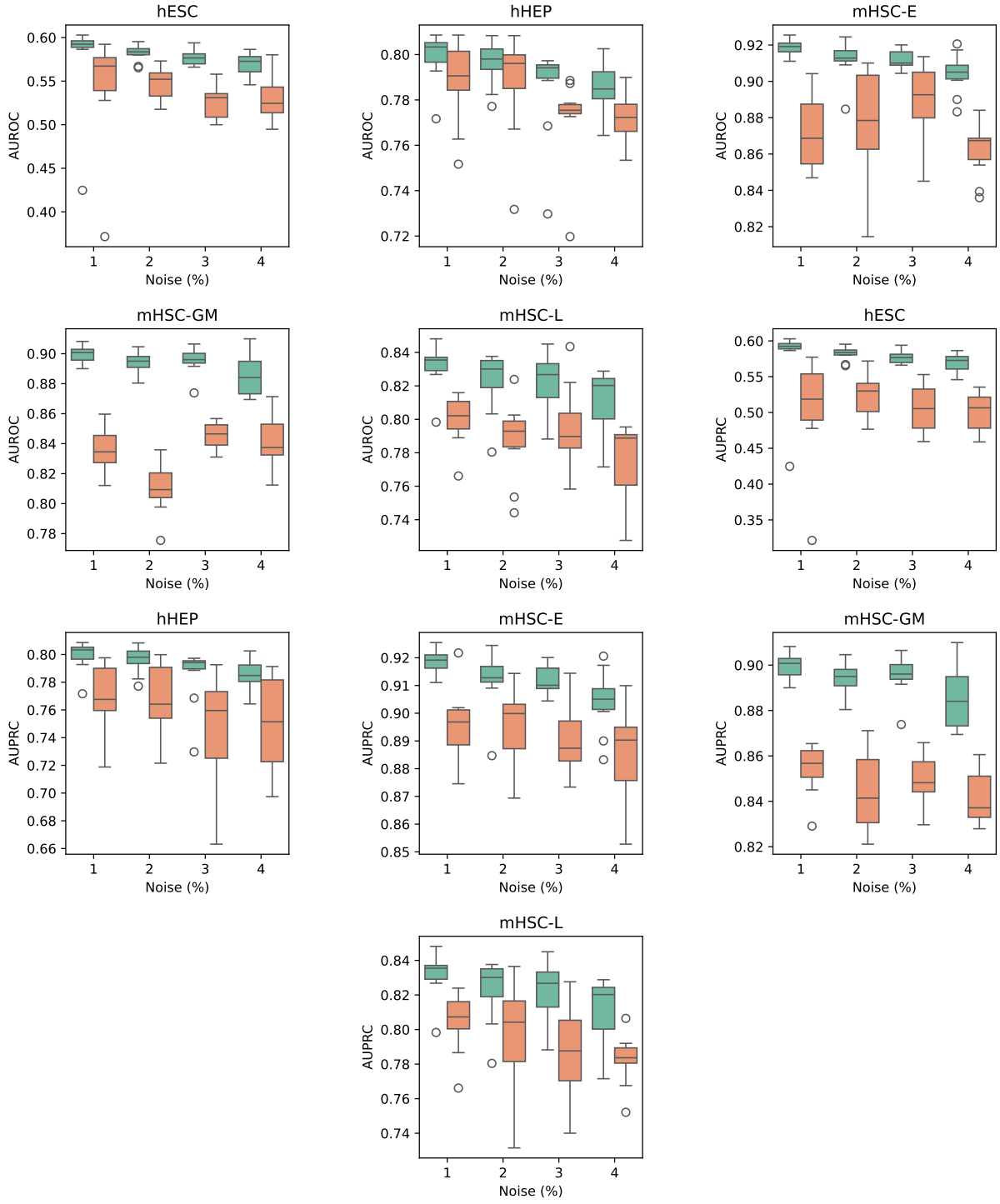


Figure 3.3: Performance comparison of **scRegNet**-Geneformer(green) vs. GENELink(orange) under increasing noise levels in cell-type-specific GRNs. The evaluation was conducted on networks containing TFs + 500 genes, with noise in the training dataset incrementally increased from 1% to 5%. Box plots illustrate the robustness of **scRegNet** in comparison to GENELink as noise levels rise, highlighting the model's stability across varying perturbations.

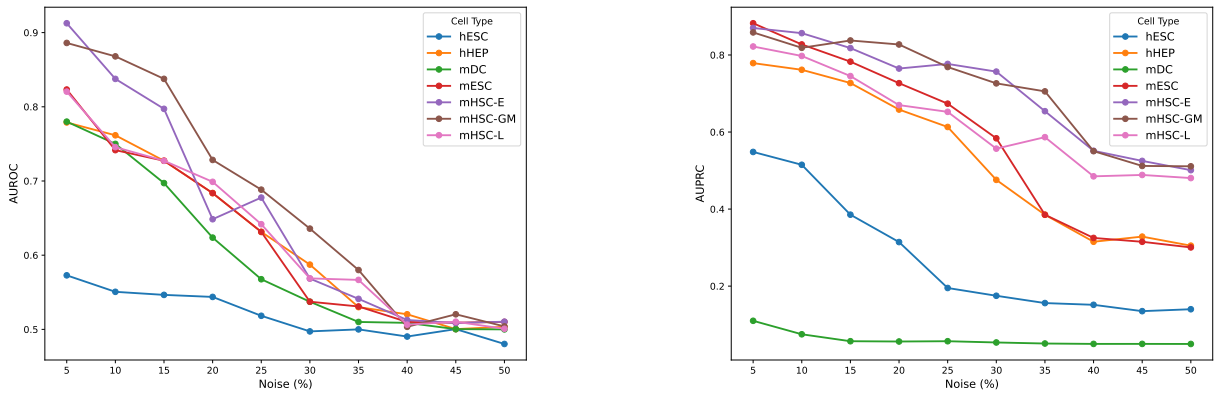


Figure 3.4: Robustness analysis of model performance under varying levels of label noise (from 5% to 50%) measured by AUROC (left) and AUPRC (right) across all the cell types. Performance remains stable at low noise levels (10%), but deteriorates significantly beyond a threshold of approximately 30% label noise, highlighting the practical limits of the model's noise tolerance.

Chapter 4

scNetFormer: Network-Guided Pretraining of Single-Cell Foundation Models

While scRegNet [30] demonstrates promising results, it still has several limitations. One notable challenge lies in the reliance on ground-truth TF-DNA interactions in the training data, which may not always be available for new cell types. The next part of this thesis addresses this challenge by focusing on incorporating the graph topology knowledge of prior biological networks into the foundation model pre-training phase, thereby enabling their broader applicability across diverse downstream tasks.

4.1 Collecting and Pre-processing of Prior Knowledge

Gene regulatory networks (GRNs) represent a comprehensive blueprint of cellular identity, delineating the complex interactions between transcription factors (TFs) and their target genes. To leverage this biological knowledge effectively, we integrate prior GRNs into the pretraining process of foundation models, enhancing their capacity to grasp universal regulatory mechanisms. The rapid accumulation of genomics data has enriched our knowledge base with critical regulatory elements and validated interactions among genes, signif-

icantly enhancing our understanding of cellular and biological processes. For our analysis, we acquired genome-scale networks comprising transcription factor and miRNA regulatory interactions from 36 normal human tissues, utilizing data from the GRAND database (https://grand.networkmedicine.org). GRAND constructs these comprehensive GRNs using the GTEx dataset and the PANDA methodology, generating aggregate TF networks consisting of 30,243 genes and 644 regulators. The dataset includes robust GRNs derived from various tissue samples, providing a rich and diverse basis for model training and analysis.

The distribution of tissue samples used from the GRAND database is visualized in Fig 4.1, highlighting the sample count across different tissue types and underscoring the diversity and scale of data employed in our integrative analyses.

4.1.1 Graph Construction

Edge weights in the downloaded GRNs indicate the strength and confidence of regulatory interactions between transcription factors (TFs) and their target genes. These weights are computationally inferred by integrating diverse sources of evidence, including gene co-expression patterns, TF-binding motifs, and protein-protein interactions. Higher edge weights reflect stronger, more confident, and biologically relevant regulatory relationships, whereas lower edge weights indicate weaker or uncertain interactions. To define the edges for our analysis, we employed two thresholds for these edge weights: 0 and 1. This resulted in creating dense and sparse graphs for each tissue type, respectively. On average, dense graphs exhibited a density of approximately 0.3, while sparse graphs had a density of about 0.09. These thresholds facilitated a balanced exploration of different graph densities in regulatory networks.

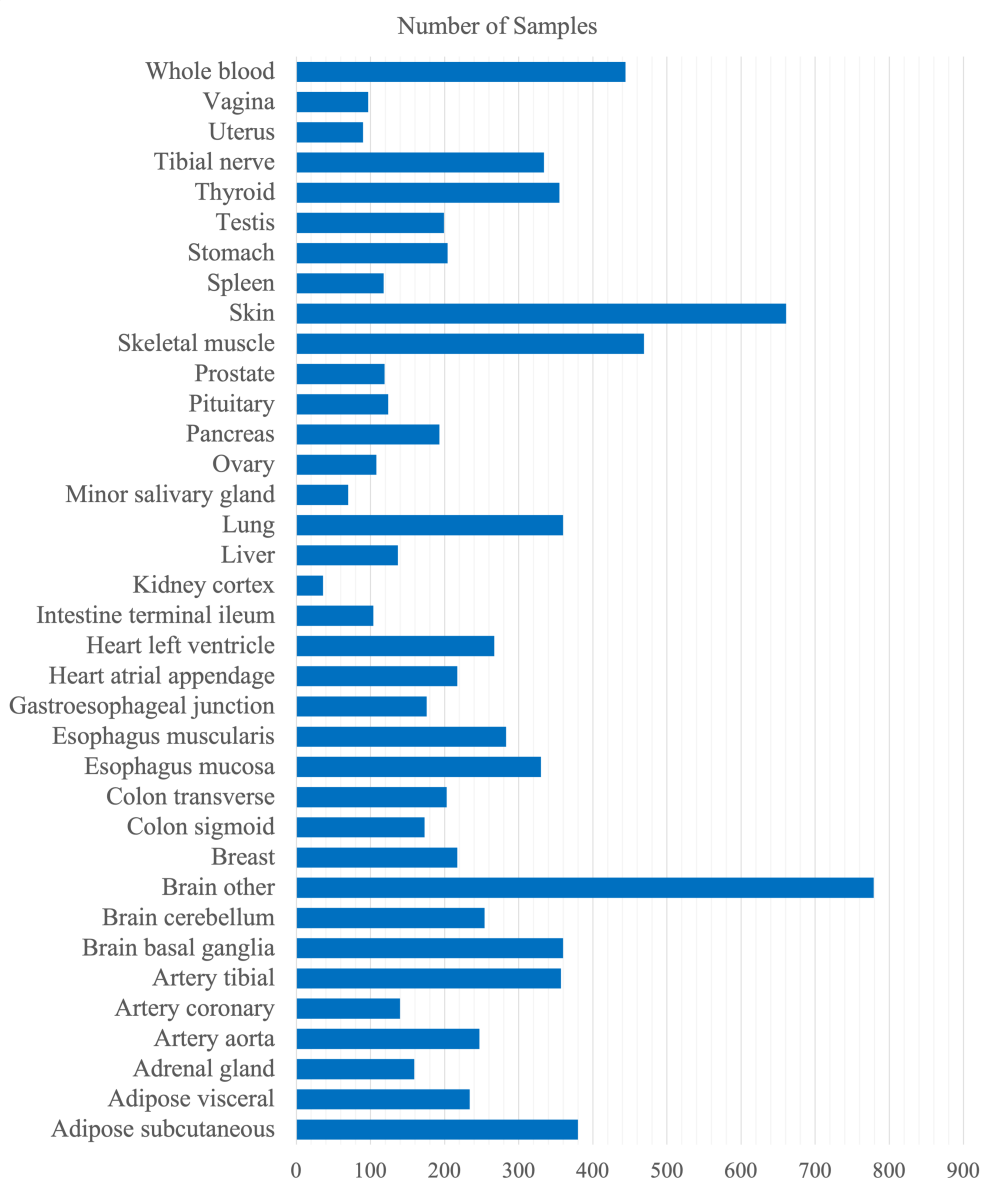


Figure 4.1: Number of samples across different tissue types from the GRAND database used in constructing gene regulatory networks.

4.2 Network-Guided Pre-training Strategies

We explored three distinct fusion strategies to integrate prior biological knowledge into the pretraining process of single-cell foundation model. We use the 6-layer Geneformer as the base model for these experiments.

4.2.1 Early Fusion

Early fusion integrates network-derived gene embeddings at the input level. Using graph neural networks (GNNs), we encoded topological structures of gene interactions from GRNs into numerical embeddings. These embeddings were merged with original gene representations by element-wise addition (Fig 4.2), thus enhancing the input representation's biological relevance.

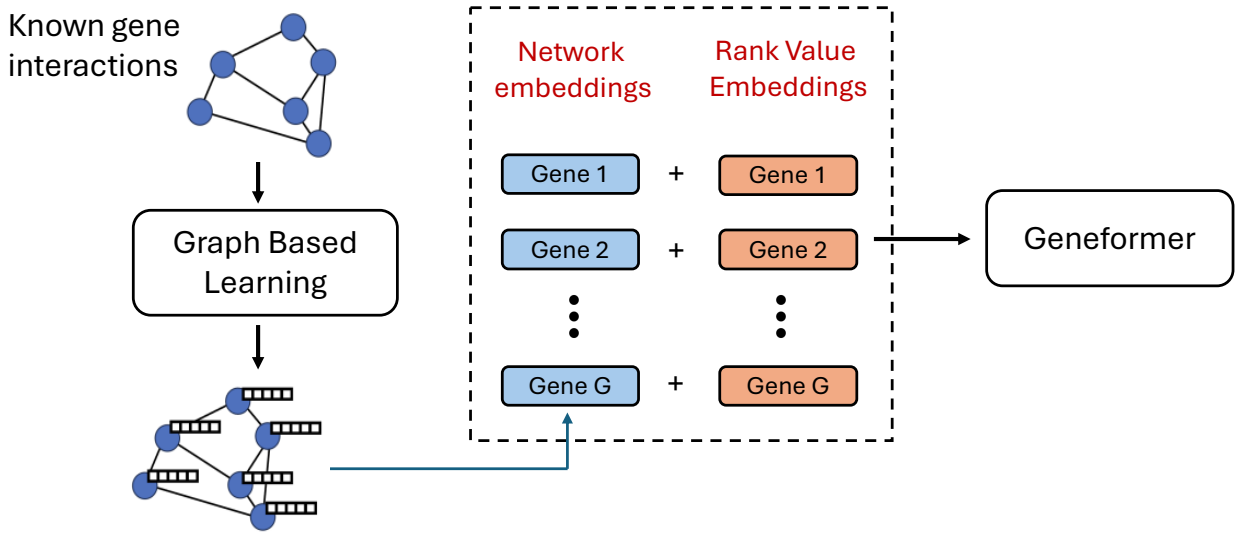


Figure 4.2: Early fusion at input level

4.2.2 Intermediate Fusion

Intermediate fusion incorporates GRN constraints directly into the attention mechanisms of transformer layers during model pretraining. Attention weights between gene pairs were selectively updated based on known regulatory connections, enforcing biologically plausible interaction patterns and improving interpretability.

Masked Attention: In a vanilla transformer, self-attention is computed for all possible pairs of tokens in the input. By contrast, genes typically attend to adjacent genes in GRNs.

Thus, for the network-guided pretraining at intermediate level it can be beneficial to introduce graph priors into the attention mask M, for example by restricting self-attention to local neighborhoods. This can be realized by setting elements of M to 0 for pairs of tokens that should be connected, and to $-\infty$ otherwise.

4.2.3 Late Fusion

Late fusion employs a contrastive learning strategy at the output layer, leveraging known regulatory relationships to regularize gene representations (Fig 4.3). An InfoNCE-based contrastive loss is used, as shown in the following equation, to ensure proximity in the representation space for biologically connected genes, enhancing the foundation model's alignment with regulatory interactions:

$$\ell_{i,j} = -\log \frac{\exp\big(\mathrm{sim}(z_i, z_j)/\tau\big)}{\sum_{\substack{k=1 \\ k\neq i}}^{2N} \exp\big(\mathrm{sim}(z_i, z_k)/\tau\big)}, \tag{4.1}$$

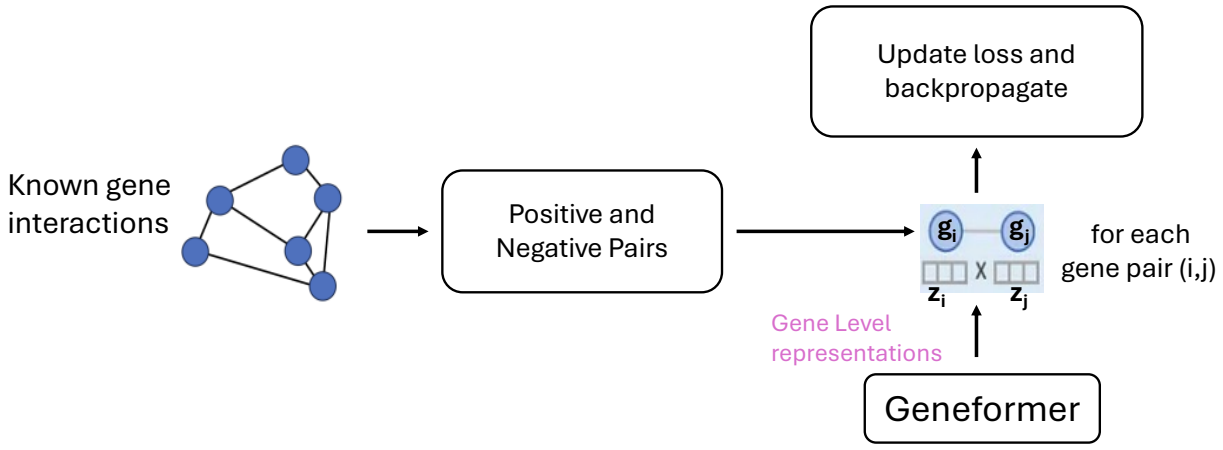


Figure 4.3: Late fusion with contrastive learning

4.3 Pre-training and Optimisation

Our work adopts the **6-layer Geneformer** architecture, where each layer comprises a multi-head self-attention block followed by an MLP. The original model was trained on *Genecorpus-95M* with $\sim$95 million human single-cell transcriptomes spanning a wide range of tissues.

Input representation. Raw UMI counts are first converted to *rank-value encodings*: within every cell, genes are sorted by expression, assigned a rank, and the ranks are then scaled by the global gene-specific median. This non-parametric scheme (i) attenuates ubiquitous housekeeping genes, (ii) preserves relative abundance information, and (iii) yields a compact, sparsity-reducing token sequence because only expressed genes are retained. Each gene token is mapped to an $\mathbf{R}^{896}$ embedding.

Initialisation. We initialise from the publicly released checkpoint `ctheodoris/Geneformer` (6-layer, 95 M cells). After standardising column names, expression matrices are passed through the `TranscriptomeTokenizer`; embeddings are extracted with the `EmbExtractor` utility.

Continued pre-training. For our new datasets we perform further pre-training and minimise a composite loss:

$$\mathcal{L} \;=\; \lambda_{\text{MLM}}\,\mathcal{L}_{\text{MLM}} \;+\; \lambda_{\text{CTR}}\,\mathcal{L}_{\text{CTR}},$$

where $\mathcal{L}_{\text{MLM}}$ is the standard masked-language-model objective and $\mathcal{L}_{\text{CTR}}$ is an InfoNCE contrastive loss that leverages graph-based positive pairs when the strategy requires it. For strategies that do not involve contrastive learning we set $\lambda_{\text{CTR}} = 0$. Hyper-parameters λ_{MLM} and λ_{CTR} are tuned on a held-out validation set.

This continued pre-training distils graph-aware structure into the Transformer while preserving the rich contextual knowledge encoded during the original Genecorpus-95M training.

4.4 Results and Discussion

4.4.1 Preliminary Experiments

To identify the most effective strategy for knowledge-guided pre-training, we began with a 6-layer GENEFORMER as the baseline foundation model and evaluated the three graph-integration strategies described in Section 4.2. Experiments were run on two tissue-specific GRNs: Heart atrial appendage and Heart left ventricle drawn from GRAND.

For each strategy we asked two questions:

1. **Can the network-guided model be fine-tuned to discriminate central from peripheral regulators within an *N1*-dependent sub-network?**

2. **Does incorporating graph information improve discrimination when only single-cell transcriptional profiles (∼30,000 endothelial cells from the Heart Atlas) are available?**

The answer to both is yes. Across five cross-validation folds, the best configuration, contrastive learning on the sparse graph combined with early-fusion 2-layer GCN features from the dense graph achieved an AUC of **0.87**, a substantial gain over the baseline (**0.81**). Results for all tested variants are summarized in Table 4.1.

Table 4.1: Preliminary results of the enhanced foundational models on different knowledge incorporation strategies.

Model	Graph-based Learning	Input GRN	AUC
Geneformer-6L (Baseline)	-	-	0.81
+ Masked Attention	-	Dense	0.74
		Sparse	0.76
+ Early Fusion	Gene2Vec	Dense	0.80
		Sparse	0.81
	GCN (2 layers)	Dense	0.84
		Sparse	0.79
	GCN (3 layers)	Dense	0.82
		Sparse	0.79
+ Contrastive Learning	-	Dense	0.79
		Sparse	0.82
+ Contrastive Learning & Early Fusion	GCN (2 layers)	Dense (Contrastive) + Sparse (Input)	0.76
	GCN (2 layers)	Dense (Input) + Sparse (Contrastive)	**0.87**

4.4.2 Performance on Binary Downstream Tasks

Experimental set-up. The best pre-training recipe from Section 4.4.1, contrastive learning on the *sparse* graph coupled with early-fusion 2-layer GCN features from the *dense* graph was adopted for the final model. Because GPU memory limited the feasible batch size, we selected the 18 GRNs that (i) are most relevant to our downstream evaluations, (ii) contain the largest numbers of single-cell profiles, and (iii) exhibit above-median edge densities. The resulting network-guided foundation model was then fine-tuned independently for four representative gene-level classification tasks and compared against the same model trained *without* graph priors.

Results. Table 4.2 shows that incorporating GRN priors consistently improved discrimination on three of the four tasks, with AUC gains ranging from ΔAUC = +0.04 (chromatin-

state prediction) to ΔAUC = +0.08 (N1 activation status). No loss of accuracy was observed on gene-dosage sensitivity, indicating that the priors do not over-constrain the model when the downstream signal is orthogonal to the pre-training graphs.

Table 4.2: Impact of GRN-based pre-training on four downstream classification tasks. Each value is the average AUC over five cross-validation folds; boldface marks the better score in each task.

	Network dynamics (N1 activated vs. non-target)	**Network topology** (Central vs. peripheral regulators)	**Chromatin dynamics** (Bivalent vs. non-methylated)	**Dosage sensitivity** (Sensitive vs. tolerant)
Logistic Regression (w/ gene counts)	0.62	0.60	0.72	0.61
Random Forest (w/ gene counts)	0.61	0.69	0.51	0.67
SVM (w/ gene counts)	0.53	0.68	0.84	0.67
Logistic Regression (w/ gene ranks)	0.60	0.59	0.51	0.65
Random Forest (w/ gene ranks)	0.58	0.68	0.69	0.72
SVM (w/ gene ranks)	0.60	0.65	0.54	0.75
Geneformer-6L (Baseline)	0.77	0.77	0.84	**0.83**
scNetFormer (ours)	**0.85**	**0.82**	**0.88**	**0.83**

4.4.3 Analysis of Attention Patterns

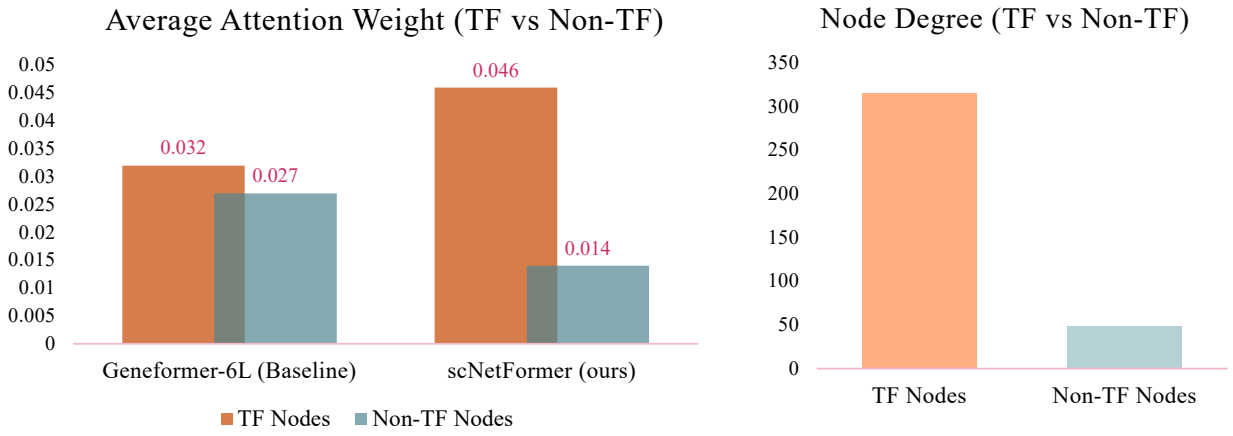


Figure 4.4: Average self-attention importance and mean node degree for transcription-factor (TF) and non-transcription-factor (non-TF) genes. TF genes interact with more partners and consistently attract higher attention weights than non-TF genes.

To probe how *scNetFormer* encodes regulatory information, we inspect its multi-head self-attention maps. For each head h, let $A^{(h)} \in \mathbb{R}^{N\times N}$ and $a_{ij}^{(h)}$ denote the attention weight from query gene i to key gene j, where N is the total number of genes. We summarize the

influence of every gene j with a *gene-wise attention importance score*

$$\phi_j = \frac{1}{H.N} \sum_{h=1}^{H} \sum_{i=1}^{N} a_{ij}^{(h)},$$

where H is the number of attention heads. A larger ϕ_j indicates that gene j is frequently referenced by the representations of other genes, signifying a stronger regulatory signal. Fig 4.4 contrasts these importance scores with node degrees for transcription-factor (TF) and non-TF nodes, underscoring how graph connectivity aligns with attention strength.

Chapter 5

Conclusions and Future Work

5.1 Summary of Contributions

In this thesis, we addressed the limitations of current single-cell foundation models through two innovative approaches: (1) scRegNet, which combines single-cell foundation models with graph neural networks for accurate GRN prediction, and (2) Network-Guided pretraining strategies that incorporate prior biological knowledge into foundation model architectures.

Our comprehensive evaluations across seven scRNA-seq benchmark datasets demonstrated that scRegNet consistently outperforms existing approaches, achieving average improvements of 7.4% in AUROC and 18.6% in AUPRC over GNNLink, and 6.9% in AUROC and 4.1% in AUPRC over GENELink on datasets with 500 most-variable genes. These performance gains were consistent across diverse cell types and experimental conditions, highlighting the robust capabilities of our integrated approach.

The second major contribution of this work explored incorporating prior GRNs directly into foundation model pretraining through three distinct fusion strategies. Our experiments revealed that contrastive learning on sparse graphs combined with early-fusion GCN features from dense graphs yielded optimal performance, improving discriminative capabilities on downstream tasks by 4-8% AUC. This integration demonstrates that biological network priors can effectively guide foundation models to develop more biologically relevant repre-

sentations.

Taken together, these contributions advance the field in two significant ways: (1) by establishing a new state-of-the-art methodology for GRN inference that leverages both large-scale pretrained embeddings and graph-based learning, and (2) by demonstrating the viability and effectiveness of incorporating established biological knowledge directly into foundation model architectures, creating more biologically informed models for single-cell analysis.

5.2 Broader Implications and Conclusion

The methodologies developed in this thesis represent significant steps toward more accurate and biologically informed computational analysis of single-cell transcriptomics. By effectively combining the contextual learning capabilities of foundation models with the structural insights of graph-based approaches, we have demonstrated a pathway toward more comprehensive understanding of gene regulatory networks. The broader implications of this work extend beyond methodological advancement. Improved GRN inference has the potential to accelerate discovery in developmental biology, disease mechanisms, and therapeutic development. Understanding the regulatory programs that define cellular identity and function provides crucial insights into normal development and disease processes, potentially informing precision medicine approaches and cell-based therapies. As single-cell technologies continue to advance and generate increasingly rich datasets, computational methods that effectively integrate prior knowledge with data-driven learning will become essential to extracting biological insights from this wealth of information. The network-guided foundation model paradigm established in this thesis provides a framework for this integration, pointing toward a future where computational models more effectively capture and represent the complex, interconnected nature of biological systems.

5.3 Limitations and Future Work

Despite encouraging performance, our study leaves several open questions and constraints that should guide future research.

Universal *vs.* cell-type–specific foundation models. We evaluated off-the-shelf single-cell foundation models (scFMs) trained on heterogeneous corpora, yet it is unclear whether a *single* universal model can ultimately outperform a collection of cell-type–focused models fine-tuned on niche data sets. Recent head-to-head benchmarks report mixed results, with universal scFMs occasionally lagging behind simpler baselines in zero-shot settings [27]. Conversely, large-scale multi-tissue transformers such as GET demonstrate that sheer data diversity can encode broad regulatory grammars [14]. A systematic comparison, ideally controlling for model capacity and training data size remains an important avenue for exploration.

Static snapshots versus dynamic regulation. Our experiments rely on static scRNA-seq atlases, a design that cannot capture causal or temporal aspects of gene regulation. Time-series or perturbation-based measurements (e.g., CRISPR screens) are starting to bridge this gap by revealing dynamic network rewiring [24]. Incorporating such longitudinal data, either during pre-training or downstream fine-tuning should yield models that better mirror real-world regulatory kinetics and disease progression.

Embedding aggregation and cell-type diversity. Mean pooling provides a simple, memory-efficient way to obtain gene-level embeddings but inevitably blurs cell-state heterogeneity. Alternatives such as graph-aware hierarchical pooling [7] or mixture-of-experts formulations that route information through cell-type–specific experts [49] could preserve

fine-grained context without requiring a prohibitive number of parameters. Implementing and benchmarking such pooling schemes is a priority for future work.

Iterative refinement of prior networks. An intriguing open question is whether attention weights extracted from scFMs can iteratively refine the initial regulatory priors. Iterative edge-reweighting has proven effective for improving GRN quality in related settings [46]. Future work may close the loop by alternating between graph-aware fine-tuning and prior-network updates.

5.4 Publications Arising from This Thesis

1. Sindhura Kommu, Yizhi Wang, Yue Wang, and Xuan Wang, "Prediction of Gene Regulatory Connections with Joint Single-Cell Foundation Models and Graph-Based Learning," in *Proceedings of the 2025 International Conference on Intelligent Systems for Molecular Biology (ISMB '25)*, Liverpool, United Kingdom, 20–24 July 2025. *Acceptance rate: 17.5%.*

2. Sindhura Kommu, Yizhi Wang, Yue Wang, and Xuan Wang, "Gene Regulatory Network Inference from Pre-Trained Single-Cell Transcriptomics Transformers with Joint Graph Learning," *AI for Science Workshop, 41st International Conference on Machine Learning (ICML '24)*, Vienna, Austria, 26 July 2024.

Bibliography

[1] Domenico Acampora, Silvia Di Giovannantonio, Adriana Garofalo, Maria A. Nigro, Maria Rosaria Omodei, Maria F. De Filippis, Maria F. Meroni, Maria F. De Felice, and Maria F. Simeone. Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition. *Development*, 140 (1):43–55, 2013. doi: 10.1242/dev.084996.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330701. URL https://doi.org/10.1145/3292500.3330701.

[3] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007. doi: 10.1038/nrg2102.

[4] Saurabh Bajaj, Hojae Son, Juelin Liu, Hui Guan, and Marco Serafini. Graph neural network training systems: A performance comparison of full-graph and mini-batch, 2024. URL https://arxiv.org/abs/2406.00552.

[5] Latifa Bakiri, Maria L. Macho-Maschler, Michael Custic, Maria Niemiec, Maria Guío-Carrión, Erwin Hasenfuss, and Erwin F. Wagner. Fra-1/ap-1 induces epithelial-to-mesenchymal transition (emt) in mammary epithelial cells by modulating zeb1/2 and tgf expression. *Cell Death Differentiation*, 22:336–350, 2015. doi: 10.1038/cdd.2014.147.

[6] Guangyi Chen and Zhi-Ping Liu. Graph attention network for link prediction of gene regulations from single-cell rna-sequencing data. *Bioinformatics*, 38(19):4522–4529, 2022.

[7] Huidong Chen, Jayoung Ryu, Michael E. Vinyard, Adam Lerer, and Luca Pinello. Simba: single-cell embedding along with features. *Nature Methods*, 21(6):1003–1013, 2024. doi: 10.1038/s41592-023-01899-8.

[8] Jiaxing Chen, ChinWang Cheong, Liang Lan, Xin Zhou, Jiming Liu, Aiping Lyu, William K Cheung, and Lu Zhang. Deepdrim: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell rna-seq data. *Briefings in bioinformatics*, 22(6):bbab325, 2021.

[9] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012. doi: 10.1038/nature11247.

[10] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, Feb 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[12] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(1):82, Feb 2019. ISSN 1471-2164. doi: 10.1186/s12864-018-5370-x. URL https://doi.org/10.1186/s12864-018-5370-x.

[13] Saskia Freytag, Johann Gagnon-Bartsch, Terence P. Speed, and Melanie Bahlo. Systematic noise degrades gene co-expression signals but can be corrected. *BMC Bioinformatics*, 16(1):309, Sep 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0745-3. URL https://doi.org/10.1186/s12859-015-0745-3.

[14] Xi Fu, Shentong Mo, Alejandro Buendia, Anouchka P. Laurent, Raul Rabadan, et al. A foundation model of transcription across human cell types. *Nature*, 637:965–973, 2025. doi: 10.1038/s41586-024-08391-z.

[15] Mathieu Gabut, Payman Samavarchi-Tehrani, Xinchen Wang, Valentina Slobodeniuc, Dave O'Hanlon, Hoon-Ki Sung, Manuel Alvarez, Shaheynoor Talukder, Qun Pan, Esteban O. Mazzoni, Stephane Nedelec, Hynek Wichterle, Knut Woltjen, Timothy R. Hughes, Peter W. Zandstra, Andras Nagy, Jeffrey L. Wrana, and Benjamin J. Blencowe. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell*, 147(1):132–146, 2011. doi: 10.1016/j.cell.2011.08.023. URL https://pubmed.ncbi.nlm.nih.gov/21924763/.

[16] Silvia Di Giovannantonio, Maria F. Acampora, Maria F. Omodei, Maria F. De Filippis, Maria F. Meroni, Maria F. De Felice, and Maria F. Simeone. The role of Otx2 in pluripotent stem cells and during neural development. *International Journal of Molecular Sciences*, 22(3):1503, 2021. doi: 10.3390/ijms22031503. URL https://www.mdpi.com/1422-0067/22/3/1503.

[17] Jing Gong, Minsheng Hao, Xingyi Cheng, Xin Zeng, Chiming Liu, Jianzhu Ma, Xuegong Zhang, Taifeng Wang, and Le Song. xtrimogene: an efficient and scalable representation learner for single-cell rna-seq data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[18] Jing Gong, Minsheng Hao, Xingyi Cheng, Xin Zeng, Chiming Liu, Jianzhu Ma, Xuegong Zhang, Taifeng Wang, and Le Song. xtrimogene: An efficient and scalable representation learner for single-cell rna-seq data, 2024. URL https://arxiv.org/abs/2311.15156.

[19] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, 2017.

[20] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. Pre-trained models: Past, present and future, 2021. URL https://arxiv.org/abs/2106.07139.

[21] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, Aug 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02305-7. URL https://doi.org/10.1038/s41592-024-02305-7.

[22] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty, 2019. URL https://arxiv.org/abs/1901.09960.

[23] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9): e12776, 2010.

[24] Masato Ishikawa, Seiichi Sugino, Yoshie Masuda, Yusuke Tarumoto, Atsushi Mochizuki, et al. Renge infers gene regulatory networks using time-series single-cell rna-seq data with crispr perturbations. *Communications Biology*, 6(1290), 2023. doi: 10.1038/s42003-023-05594-4.

[25] Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.*, 12(3):e694, March 2022.

[26] Kishan Kc, Rui Li, Feng Cui, Qi Yu, and Anne R Haake. Gne: a deep learning framework for gene network inference by aggregating biological information. *BMC systems biology*, 13:1–14, 2019.

[27] Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, Alex X. Lu, et al. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26 (101), 2025. doi: 10.1186/s13059-025-03574-x.

[28] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[29] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[30] Sindhura Kommu, Yizhi Wang, Yue Wang, and Xuan Wang. Prediction of gene regulatory connections with joint single-cell foundation models and graph-based learning. *bioRxiv*, 2025. doi: 10.1101/2024.12.16.628715. URL https://www.biorxiv.org/content/early/2025/01/29/2024.12.16.628715.

[31] Krzysztof R. Kranc, Maria F. De Filippo, Maria F. De Cegli, Maria F. De Luca, Maria F. De Santis, Maria F. De Simone, Maria F. De Vita, Maria F. De Vivo, Maria F. De Vito, and Maria F. De Vries. Cited2 is an essential regulator of adult hematopoietic stem cells and erythroid differentiation. *Cell Stem Cell*, 16(1):1–14, 2015. doi: 10.1016/j.stem.2014.11.001.

[32] Ying Li, Zhen Luo, Xiao Gao, Yiping Zhu, Yongchao Ren, Amanda Leung, Jianming Liu, Edwin Choy, Heather Harwood, Raphael Guzman, et al. Sox2 co-occupies distal enhancer elements with distinct pou factors in escs and npcs to specify cell state. *PLoS genetics*, 8(2):e1002503, 2012.

[33] Guo Mao, Zhengbin Pang, Ke Zuo, Qinglin Wang, Xiangdong Pei, Xinhai Chen, and Jie Liu. Predicting gene regulatory links from single-cell RNA-seq data using graph neural networks. *Briefings in Bioinformatics*, 24(6):bbad414, 11 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad414. URL https://doi.org/10.1093/bib/bbad414.

[34] Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.

[35] Hiroki Okanishi, Masashi Tanaka, and Yoshihide Suzuki. Chip-atlas: a data-mining suite powered by full integration of public chip-seq data. *EMBO reports*, 22(10):e52171, 2021. doi: 10.15252/embr.202152171.

[36] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, Feb 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0690-6. URL https://doi.org/10.1038/s41592-019-0690-6.

[37] XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, September 2020. ISSN 1869-1900. doi: 10.1007/s11431-020-1647-3. URL http://dx.doi.org/10.1007/s11431-020-1647-3.

[38] Faridah Hani Mohamed Salleh, Shereena Mohd Arif, Suhaila Zainudin, and Mohd Firdaus-Raih. Reconstructing gene regulatory networks from knock-out data using

gaussian noise model and pearson correlation coefficient. *Computational biology and chemistry*, 59:3–14, 2015.

[39] Hantao Shu, Fan Ding, Jingtian Zhou, Yexiang Xue, Dan Zhao, Jianyang Zeng, and Jianzhu Ma. Boosting single-cell gene regulatory network reconstruction via bulk-cell transcriptomic data. *Briefings in Bioinformatics*, 23(5):bbac389, 2022.

[40] Anand M. Singh, Maria F. Hamazaki, Maria F. Hankowski, and Maria F. Terada. A heterogeneous expression pattern for Nanog in embryonic stem cells. *Stem Cells*, 25(10):2534–2542, 2007. doi: 10.1634/stemcells.2007-0126. URL https://stemcellsjournals.onlinelibrary.wiley.com/doi/full/10.1634/stemcells.2007-0126.

[41] Ying Su, Yuchen Wang, Xiaoyan Zhou, Yuxiao Yang, Xiaoyan Lai, Xiaoying Zheng, Yanhong Chen, Jiekai Chen, Lijian Hui, Hongyan Wang, and Guoji Guo. Repression of Otx2 by Nanog mediates the fate commitment from naive pluripotency. *Cell Stem Cell*, 23(3):412–425.e10, 2018. doi: 10.1016/j.stem.2018.08.004. URL https://www.cell.com/cell-stem-cell/fulltext/S1934-5909(18)30353-2.

[42] Atsushi Suzuki, Ángel Raya, Yasuhiko Kawakami, Masanobu Morita, Takaaki Matsui, Kinichi Nakashima, Fred H. Gage, Concepción Rodríguez-Esteban, and Juan Carlos Izpisúa Belmonte. Nanog binds to smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells. *Proceedings of the National Academy of Sciences*, 103(27):10294–10299, 2006. doi: 10.1073/pnas.0506945103. URL https://www.pnas.org/doi/10.1073/pnas.0506945103.

[43] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*,

618(7965):616–624, Jun 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[45] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

[46] Yijie Wang, Dong-Yeon Cho, Hangnoh Lee, Justin Fear, Brian Oliver, and Teresa M. Przytycka. Reprogramming of regulatory network using expression uncovers sex-specific gene regulation in *Drosophila*. *Nature Communications*, 9(4061), 2018. doi: 10.1038/s41467-018-06382-z.

[47] Jianzhu Xu, Zhixing Shao, Xuanlin Li, and et al. Escape: database for integrating high-confidence human tf-target regulatory interactions. *Nucleic Acids Research*, 43 (D1):D929–D934, 2015. doi: 10.1093/nar/gku1161.

[48] Rongwen Xu, Maria F. Sampsell, Maria F. Wang, Maria F. Lee, Maria F. Stiles, Maria F. Powers, and Maria F. Kamps. Activin/tgf- signaling mediates self-renewal and pluripotency of embryonic stem cells. *Cell Stem Cell*, 2(6):523–535, 2008. doi: 10.1016/j.stem.2008.03.011. URL https://www.cell.com/cell-stem-cell/fulltext/S1934-5909(08)00164-5.

[49] Youran Xu, Li Zhou, Qian Liu, Sheng Cao, et al. scMoE: single-cell mixture of experts for learning hierarchical, cell-type-specific and interpretable representations. *bioRxiv*, 2024. doi: 10.1101/2024.10.24.620111.

[50] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4 (10):852–866, Oct 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00534-z. URL https://doi.org/10.1038/s42256-022-00534-z.

[51] Sheng Yang, Maria F. Liu, Maria F. Li, Maria F. Zhang, Maria F. Liu, Maria F. Fang, and Maria F. Zhang. The role of JUN proteins in the regulation of epithelial-to-mesenchymal transition during cancer progression. *Journal of Translational Medicine*, 14(1):1–12, 2016. doi: 10.1186/s12967-016-0815-9. URL https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-016-0815-9.

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