GeneNet Transformer : A Novel Transformer-Based Architecture for Gene Network Inference

A thesis
Submitted towards the partial fullfilment of
BS-MS dual degree programme
by

PRANAV ANAND SADHU

DATE:

under the guidance of

DR TANMAY BASU

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH
BHOPAL

from Jan 2024 to Oct 2024

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH
PUNE

Certificate

This is to certify that this dissertation entitled ”GeneNetTransformer : A Novel Transformer-Based Architecture for Gene Network Inference” submitted towards the partial fulfillment of the BS-MS degree at the Indian Institute of Science Education and Research, Pune represents original research carried out by Pranav Anand Sadhu at Indian Institue of Science Education and Research Bhopal, under the supervision of Dr Tanmay Basu during academic year Jan 2024 to Oct 2024.

Supervisor:
DR TANMAY BASU
ASSISTANT
PROFESSOR IISER
BHOPAL

PRANAV SADHU
ROLL NO.
20191152
BS-MS
IISER PUNE

DATE:
20/10/2024

Declaration

I, hereby declare that the matter embodied in the report titled "GeneNet-Transformer : A Novel Transformer-Based Architecture for Gene Network Inference" is the result of the investigations carried out by me at the "Indian Institue of Science Education and Research Bhopal" from the period 01-01-2024 to 30-10-2024 under the supervision of Dr Tanmay Basu and the same has not been submitted elsewhere for any other degree.

Basu

Supervisor:
DR TANMAY BASU
ASSISTANT
PROFESSOR IISER
BHOPAL

PRANAV ANAND
SADHU
20191152
BS-MS
IISER PUNE

Date:
20/10/2024

Acknowledgements

I would like to extend my heartfelt gratitude to everyone who has supported me throughout this thesis journey.

First and foremost, I would like to express my deepest appreciation to my mother, father, sister, cousins and my whole joint family, whose love and encouragement have been a source of strength. I am forever grateful for their unwavering support.

I would like to thank my supervisor, Dr. Tanmay Basu, for his invaluable guidance, insightful discussions, and continuous encouragement. I am also deeply grateful to Dr.Leelavati Naralikar for her advice throughout the thesis.

A special thanks goes to my friends from School, IISER Pune and IISER Bhopal, for their camaraderie and encouragement and "BDS lab" for making my time here so memorable.

Lastly, I would like to acknowledge my gaming buddies for providing much-needed breaks and fun during challenging times.

Thank you all!

Abstract

Precise inference of gene interaction networks is crucial for understanding complex biological processes and disease mechanisms. Traditional methods often rely on curated databases, which may overlook important but undiscovered interactions. Various deep learning based approaches have been developed over the last two years to address this issue. However, most of these method are either database specific or consider specific kinds of genes. Therefore, a deep learning-based transformer model is introduced in this thesis to predict missing edges in gene interaction networks using the existing databases. The proposed method integrates heterogeneous gene interaction data with microarray expression data, leveraging the attention mechanisms in transformer models to uncover intricate relationships.

In the first stage, the model processes a candidate gene's one-hot encoding and its microarray expression values, constructing a fully connected network to generate individual embeddings, each of size d. These embeddings, concatenated into a vector of size $2d$ are passed through a standard transformer encoder, which reduces them to d-dimensional embeddings to extract significant information from both gene identity and expression. In the second stage, these transformer-generated embeddings for gene pairs are used to train an SVM classifier. The input to the classifier is the element-wise product of a gene pair's embeddings, along with their known interaction labels.

The performance of the proposed model is compared with the state of the arts in terms of AUC-ROC and AUPR using seven standard datasets, each corresponding to cell-type-specific ChIP-seq, Non-Specific ChIP-seq and STRING dataset ground truth networks. The empirical analysis shows that the proposed one outperforms the state of the arts, which indicates its potential for predicting new or undocumented interactions in biological networks. in future, the performance and scalability of the model need to be tested on various other types of reasonably large networks.

Contents

Chapter 1

Introduction

The rapid progress of single-cell RNA sequencing (scRNA-seq), coupled with the exponential growth of genomic data, has expanded the boundaries of single-cell research and underscored the need for advanced computational methods to decode gene interactions and regulatory networks.[1, 2]. Gene regulatory networks (GRNs) capture these intricate interactions between genes, typically involving transcription factors (TFs) and their target genes, and play a crucial role in controlling gene expression within cells. Accurately reconstructing GRNs is fundamental to understanding various cellular processes, such as gene expression mechanisms, cellular differentiation, and research in disease pathology [3]. However, despite the promising opportunities presented by single-cell technologies, they also bring significant challenges, such as the inherent noise and complexity of scRNA-seq data [4]. Recent progress in deep learning methods offers strong solutions to these challenges by managing noisy data and combining different sources of information. These methods help uncover complex relationships between genes through feature extraction and optimization.[2, 5, 6, 7].

Several methods have been proposed to infer GRNs from single-cell data. For example, SCODE uses ordinary differential equations (ODEs) to reconstruct GRNs by treating pseudotime as time information during cell differentiation [8]. Meanwhile, GENIE3 and GRNBoost2 employ tree-based machine learning algorithms incorporated into the SCENIC pipeline to infer gene regulatory interactions [9]. These methods leverage boosting techniques to improve performance but have significant limitations. For instance, SCODE depends on pseudotime data and often oversimplifies complex biological processes by using linear ODEs, while tree-based methods, such as GENIE3 and GRNBoost2, involve high computational costs and poor scalability due to their need to segment data into multiple models iteratively.

To overcome these limitations, various deep learning-based methods have emerged [2, 5, 6, 7]. DeepSEM employs a structural equation model (SEM) combined with a beta-variational autoencoder and a neural network to infer regulatory relationships [5]. However, this method depends heavily on prior domain knowledge and SEM assumptions that may not always hold in practice. Other methods, such as GNE [6], use multilayer perceptron (MLP) architectures to infer GRNs from microarray data. These methods utilize one-hot encoded gene ID vectors, but this approach suffers from inefficiencies due to the sparsity of the resulting feature vectors. Graph-based methods, such as GENELink [7], use graph attention networks (GATs) to capture the topology of gene networks, but they often emphasize local network information at the expense of a global regulatory perspective, which can lead to suboptimal feature representation. Single-cell Gene Regulatory Embedding using Transformer or scGREAT[2] utilizes a robust transformer-based architecture to infer GRNs from single-cell transcriptomics data along with text-based BioBERT embeddings from gene names in order to overcome the local neighbourhood emphasis in GENE Link. The scGREAT model constructs the gene dictionaries of embeddings and predicts the label over all the pairs. However, the usage of hard negative sampling may make the model biased and the usage of BioBERT results in heavy computation and literature dependence.

In order to address the limitations of the state of the arts for gene network inference, GeneNet Transformer (GNT), a transformer-based deep learning model is proposed here in the spirit of GNE [6] and ScGREAT [2]. GNT leverages a standard transformer architecture with two encoding layers to capture complex dependencies and relationships between genes [10]. Using the embeddings learned from single RNA sequencing (scRNA-seq) data and interaction network data [11], GNT transforms the interaction prediction task into a link prediction task. The proposed framework for gene network inference is developed in three phases.

The first phase consists of the proposed transformer architecture for generating gene embeddings, which is developed in three layers. The first layer of the input layer contains the one-hot encoding of the gene and its corresponding expression values from scRNA-seq data. The dimension of the one-hot encoding vector is a number of genes, say N, and the number of expression values for a gene is E, say. Subsequently, both of these vectors are individually transformed to a d (say) dimensional vector through a fully connected layer following the GNE architecture [6] to reduce the dimensionality. Therefore, in the third stage, a standard Transformer architecture is used to develop gene embeddings. The transformer has two encoder layers, which take the transformed vectors of length 2d as input and generate the gene embeddings of dimension d.

Second phase generates the embeddings of the edges constituted by individual pair of gene embeddings from phase 1. In phase 3, a Support Vector Machine (SVM) classifier is trained using the edge embeddings and the ground truths of gene interactions for predicting the interactions of new gene pairs.

The thesis is organized as follows. Chapter 2 describes the related works. The research gaps in this domain is discussed in chapter 3. Chapter 4 explains the proposed method. The experimental evaluation is presented in chapter 5. Finally, we discuss the merits, limitations and future scopes of the proposed method in chapter 6.

Chapter 2

Related work

Gene regulatory networks (GRNs) capture the intricate and multi-level regulatory interactions between transcription factors (TFs) and their target genes, which are critical for understanding cellular processes and molecular functions. Advances in single-cell RNA sequencing (scRNA-seq) now allow for the inference of GRNs at a single-cell resolution, offering deeper insights into gene regulation.

The rapid progress of scRNA-seq technologies, along with the boom in genomic data, has highlighted the need for advanced computational tools to better understand gene-gene/protein-protein interactions in detail. Gene regulatory networks (GRNs), which show how molecules control each other, are key to understanding gene behavior, such as how genes are expressed in cells, and have many applications in disease research. Although single-cell technologies are powerful, they also bring major challenges, especially due to the complexity and noise in scRNA-seq data.

Deep learning-based approaches have proven effective in overcoming these challenges by handling noisy data, integrating diverse knowledge sources, and learning complex relationships through their feature extraction capabilities. While several unsupervised and self-supervised models have been proposed for GRN inference from bulk RNA-seq data, few are suitable for scRNA-seq data due to issues such as low signal-to-noise ratios and dropout rates. The increasing availability of transcription factor-DNA binding data (e.g., ChIP-seq) enables supervised GRN inference, which we approach as a graph-based link prediction problem, where the goal is to learn gene vector representations for predicting regulatory interactions.

Over recent years, numerous methods have been developed to infer GRNs, leveraging advances in machine learning, deep learning, and network theory. Each method offers unique strengths while addressing specific aspects of gene regulation, but they also exhibit limitations that hinder their broader applicability and generalization to diverse biological contexts.

SCODE[12] (Single-Cell ODE) employs ordinary differential equations (ODEs) to infer GRNs by treating pseudotime as a proxy for temporal information during the differentiation of cells. This approach assumes that the pseudotime obtained from trajectory inference methods reflects the temporal dynamics of gene expression, allowing SCODE to model changes in gene regulation over time. SCODE has shown success in modeling GRNs during cellular differentiation, particularly by leveraging the pseudotemporal ordering of cells. In dynamic or pathological contexts where gene regulation is inherently non-linear, the linear assumptions of SCODE tend to oversimplify the intricate regulatory dynamics, leading to suboptimal performance in capturing the true nature of gene expression and differentiation.

GENIE3[12] (GeNet InferencE with Ensemble of Trees) and GRNBoost2[9] are widely used machine learning algorithms for GRN inference. Both methods are based on decision trees and ensemble learning, with GENIE3 relying on random forests and GRNBoost2 using gradient boosting. These methods identify regulatory relationships by learning the importance of genes in predicting the expression of target genes, effectively iterating over the gene set and leaving one gene out at a time. This iterative approach enables the algorithms to capture potential regulatory interactions. GENIE3 and GRNBoost2 have been integrated into the SCENIC framework (Single-Cell Regulatory Network Inference and Clustering), enhancing their utility in large-scale single-cell analysis by enabling cell-type-specific GRN inference. The boosting method used in GRNBoost2 further improves the performance of tree-based approaches by refining the decision trees in successive iterations.These methods require extensive segmentation of input data and building of decision trees iteratively which is expensive for larger datasets.

DeepSEM[13] (Deep Structural Equation Modeling) is a deep learning approach designed to overcome some of the limitations of traditional statistical models for GRN inference. It integrates a structural equation model (SEM) with a beta-variational autoencoder (β-VAE) to capture the underlying causal structure of gene regulatory interactions. DeepSEM models the relationships between transcription factors (TFs) and target genes by assuming that there is a latent causal structure driving gene expression patterns. The β-VAE helps in learning latent representations of the regulatory relationships while accounting for noise in the data. DeepSEM has the advantage of integrating causal modeling with deep learning, offering a more flexible framework for inferring complex regulatory networks. DeepSEM's reliance on prior domain knowledge to define the causal structure of gene regulatory interactions restricts its generalizability.This dependency on domain-specific knowledge limits the model's ability to infer GRNs in novel or under-explored biological contexts, thereby reducing its overall utility.

GNE[6] is a deep learning-based method that utilizes a multilayer perceptron (MLP) architecture for GRN inference. The key feature of GNE is its use of one-hot encoding to represent gene identities, capturing the topological relationships between genes in the regulatory network. By employing MLPs, GNE attempts to model the complex, nonlinear relationships in gene expression data, particularly when applied to microarray data. The MLP architecture used in GNE may not be well-suited for modelling the complex, non-linear relationships inherent in GRNs, further limiting its effectiveness in accurately capturing regulatory dynamics at scale.

CNNC[5] (Convolutional Neural Network for Co-expression) is a convolutional neural network (CNN)-based model designed to infer GRNs by transforming co-expression data into images. The model first computes co-occurrence values between genes and then normalizes these values into a probability distribution function. The resulting normalized co-expression data is converted into pixel values, effectively creating image representations of gene-gene interactions. CNNC applies convolutional filters to these images to identify patterns of co-expression that correspond to regulatory relationships. Although this method leverages the power of CNNs, which have been highly successful in image processing tasks, transforming transcriptomic data into image-based data for each gene pair introduces a substantial computational burden, making the method less scalable for large datasets.The process of converting transcriptomic data into image-based representations for each gene pair introduces substantial computational complexity, making the method less scalable for large datasets. This computational overhead, combined with the fact that CNNC does not fully utilize end-to-end deep learning capabilities, reduces the method's overall efficiency and applicability in large-scale GRN inference tasks.

GENELink[7] is a deep learning framework that utilizes graph attention networks (GATs) to infer GRNs. GATs extend the capabilities of traditional graph convolutional networks by assigning different attention weights to neighboring nodes in the graph, allowing GENELink to capture the importance of individual regulatory relationships. GENELink uses interaction data to construct gene networks and then applies graph convolution and attention mechanisms to model the regulatory interactions between TFs and target genes. GENELink emphasizes the importance of nearby relationships within the graph by focusing on local network structures. However, this local emphasis presents a limitation in its ability to capture global regulatory dynamics across the entire network.GENELink's reliance on high-quality node features and sparse ground truth networks can hinder its performance in contexts where the available data is incomplete or unreliable, particularly for novel or poorly characterized genes.

scGREAT[2] (Single-cell Gene Regulatory Embedding using Transformer) is a state-of-the-art method that utilizes a transformer-based architecture to infer GRNs from single-cell transcriptomics data. The model builds upon recent advances in natural language processing by employing a transformer to learn regulatory interactions between TFs and target genes. scGREAT uses gene expression data and biotext information (e.g., BioBERT embeddings) to enhance the accuracy of its predictions. Unlike traditional methods that rely solely on gene expression data, scGREAT incorporates textual information from biomedical literature, allowing it to capture context-specific regulatory relationships. The transformer architecture enables scGREAT to learn complex, long-range dependencies between genes, making it highly effective for GRN inference.

In conclusion, while each method offers valuable insights into gene regulatory networks, their limitations—ranging from computational inefficiency and scalability issues to dependency on ground truth data and assumptions about regulatory relationships—highlight the need for further innovation in GRN inference

Chapter 3

Research Gap

Despite the significant advancements in gene regulatory network (GRN) inference, each method exhibits specific limitations that constrain its effectiveness and scalability in certain biological contexts. Addressing these limitations is critical for enhancing the accuracy, efficiency, and generalizability of GRN models, particularly as single-cell transcriptomics and multi-omics data grow in complexity and size.

The SCODE model [8], which relies on ordinary differential equations (ODEs) and pseudotime data, faces challenges primarily due to its dependence on accurate pseudotime information [2]. The DeepSEM model [5] relies on prior domain knowledge to define the causal structure of gene regulatory interactions, which restricts its generalizability, particularly in biological systems where the regulatory mechanisms are not well understood. The SEM component assumes a predefined causal structure, and these assumptions may not always hold in practice, especially in complex or poorly characterized systems [2]. GNE model [6] encounters limitations due to its use of multi layer perceptron (MLP) technqiue for generating embeddings [2]. MLP is a computationally expensive technique and many advanced techniques have been developed in this spirit. GENELink technqiue [7] focus on local network interactions, which is unable to capture global regulatory dynamics across the entire network [2]. scGREAT's [2] reliance on BioBERT embeddings for gene representation introduces challenges when handling newly discovered genes for which there is limited or no literature. Moreover, it employs hard negative sampling of non interacting pairs, which can introduce bias, as the model may overfit to non-interacting pairs, reducing its ability to accurately distinguish between true regulatory interactions and non-interactions.

To our knowledge, there is no study in this direction to address these limitations of the existing techniques for gene network inference.

Chapter 4

GeneNet Transformer Model

In this spirit, a transformer-based model is proposed, which is named as **GeneNetTransfomer Model**. The aim is to learn a lower-dimensional embedding for genes. These lower-dimensional representations of genes can be utilised for downstream tasks like link prediction. The proposed framework for gene network inference is developed into the following three phases:

- Phase 1 consists of a feature transformation network followed by the GeneNet Transformer architecture for generating gene embeddings.
- Second phase generates the embeddings of the edges constituted by individual pair of gene embeddings from phase 1.
- In phase 3, a Support Vector Machine (SVM) classifier is trained using the edge embeddings and the ground truths of gene interactions for predicting the interactions of new gene pairs.

4.1 Phase 1: GeneNet Transformer

In the initial step in feature representation involves processing the data through two distinct types of embeddings: one-hot encoding of genes and normalized expression values obtained from single-cell RNA sequencing (scRNA-seq) data. The one-hot encoding of each gene is transformed into a d-dimensional vector, denoted as $V_i^{(\mathrm{id})}$, through a fully connected layer. The weights associated with this transformation are represented by the matrix $W^{(\mathrm{id})}$. Similarly, the normalized expression values from the scRNA-seq data for each gene are also mapped to a d-dimensional vector, denoted as $V_i^{(\mathrm{att})}$, using a separate fully connected layer. The weights for the attribute vector transformation of all genes are stored in the matrix $W^{(\mathrm{att})}$. After the sum of multiplication of $V_i^{(\mathrm{att})}$ with $W^{(\mathrm{a})}$ and $V_i^{(\mathrm{id})}$ with $W^{(\mathrm{b})}$ we get a d-dimensional vector. This process ensures that both the structural information (from one-hot encoding) and the attribute information (from expression values) are effectively reduced to a common d-dimensional feature space, facilitating integration and subsequent analysis.

This vector is fed forward into two transformer encoder layer. As the data passes through the layers of the Transformer, the multi-head self-attention mechanism dynamically weighs the interactions between nodes, combining both the identity-based and attribute-based embeddings. The output of the second Transformer Encoder layer, therefore, captures both the relational information (neighbourhood dependencies) and biological expression values, preserving both the structural identity of nodes (through attention mechanisms) and their biological attributes (through attribute embedding). The final output layer of transformer encoder is set to be d dimensional. This vector is transformed into a probability vector with the matrix $W^{(\mathrm{out})}$. Elements in this vector represent the conditional probability of that gene connected to all the other genes.

To train the model, the predicted logits are computed from the output of the last Transformer Encoder layer using a fully connected output layer:

$$\hat{y} = \mathrm{softmax}(W_o Z_{\mathrm{Layer2}} + b_o)$$

where $\hat{y} \in \mathbb{R}^C$ is the predicted probability distribution over C classes. The model is trained using Cross-Entropy Loss:

$$L = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

Output layer of second transformer encoder represents a rich latent space embedding that encodes neighbourhood information alongside expression data, which is then used to predict gene interactions or relationships in the output layer. The final predicted values are compared with ground truth values (gene interaction labels) using a loss function such as cross-entropy. This process ensures that the model simultaneously learns from the structure of the network (neighbourhood information) and the underlying biological data (expression values), effectively capturing the complexity of gene interactions.

4.2 Phase 2: Edge Embedding Generation

The given edge list is partitioned into training, validation, and test sets with a ratio of $8:1:1$. To create a balanced dataset, an equal number of negative interactions are introduced and concatenated with the positive interactions. This results in a final dataset that is balanced for model training. Gene embeddings of dimension d are obtained from the GNT model, where n embeddings are generated. For an edge connecting Gene G_i and G_j, we create an edge embedding by the **Hadamard product** of their respective embeddings, i.e. the elementwise multiplication of gene embeddings of Gene G_i and G_j. The edge embeddings are also d dimensional.

4.3 Phase 3: Training SVM for Gene Network Inference

Subsequently, using the same training data split utilized for training the GNT model, an SVM classifier is trained over the edges to predict the interaction labels between edge embeddings. The final output is the label predicted by the SVM classifier over the embeddings of edges from the test dataset, i.e. 1 in case of a true interaction or 0 in the case of no/missing interaction.

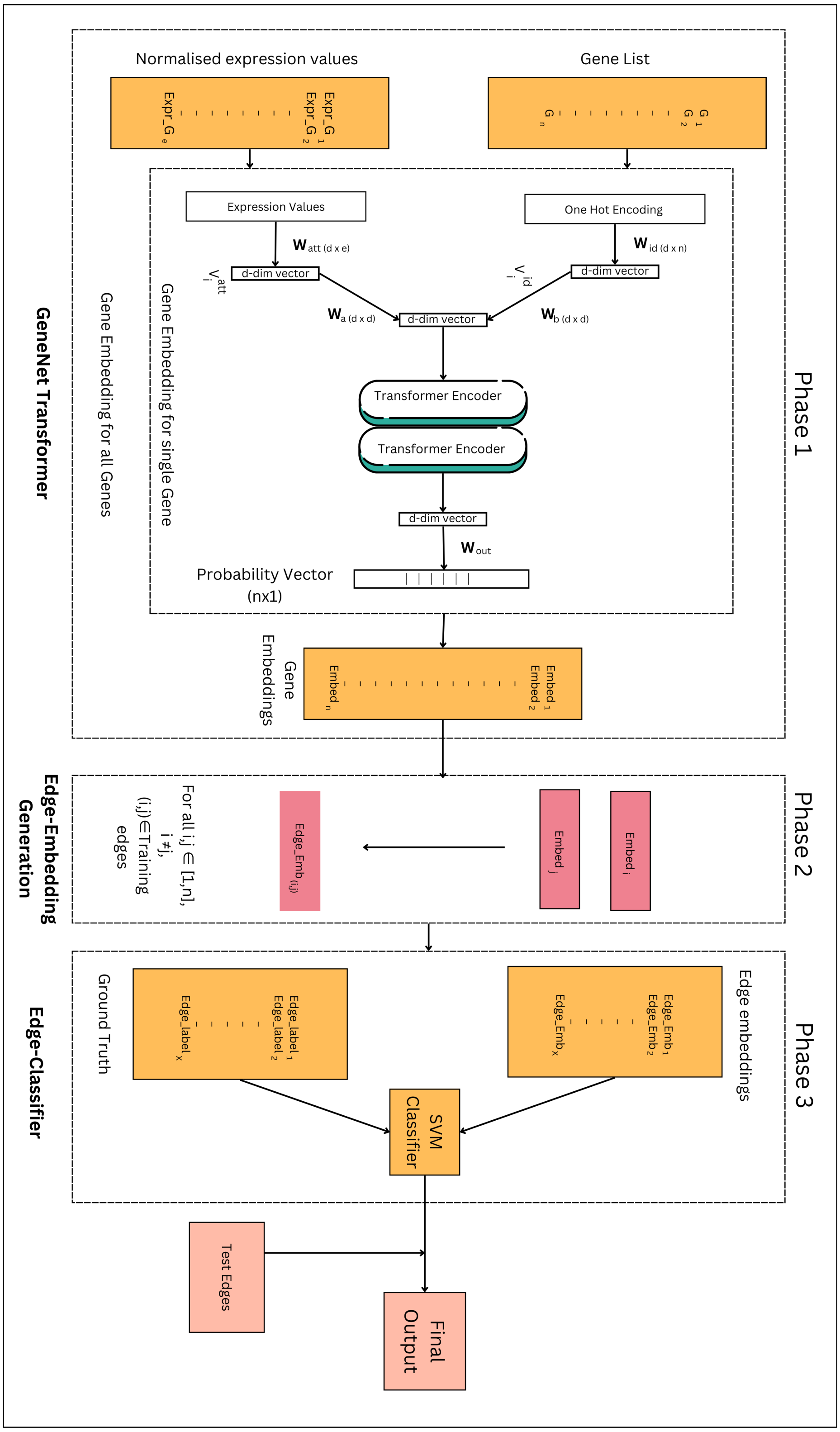

Figure 4.1: **Proposed Methodology**

Chapter 5

Experimental Evaluation

5.1 Overview of Datasets

We assess the performance of GNT on seven distinct cell types derived from scRNA-seq datasets provided by the BEELINE benchmark [11]. These include: (i) human embryonic stem cells (hESC), (ii) human mature hepatocytes (hHEP), (iii) mouse dendritic cells (mDC), (iv) mouse embryonic stem cells (mESC), (v) mouse hematopoietic stem cells of the erythroid lineage (mHSC-E), (vi) mouse hematopoietic stem cells of the granulocyte-monocyte lineage (mHSC-GM), and (vii) mouse hematopoietic stem cells of the lymphoid lineage (mHSC-L)[7, 2]. Each of these datasets is paired with three distinct ground-truth networks sourced from functional interaction data in the STRING database[14], non-specific ChIP-seq data[15, 16, 17], and cell-type-specific ChIP-seq data [18, 19, 20].

The preprocessing of each scRNA-seq dataset was performed following the methodology outlined by Pratapa et al. (2020), with gene regulatory network (GRN) inference restricted to interactions originating from transcription factors (TFs). For GRN inference, we selected the top 500 and 1000 most highly variable genes in conjunction with TFs whose variance showed a Bonferroni-corrected P-value below 0.01, as recommended by Pratapa et al. (2020). The scRNA-seq datasets for the seven cell types are available through the Gene Expression Omnibus, with the accession numbers GSE81252 (hHEP), GSE75748 (hESC), GSE98664 (mESC), GSE48968 (mDC), and GSE81682 (mHSC)[7]. All single-cell datasets, along with four types of ground-truth networks, can be accessed at https://doi.org/10.5281/zenodo.3378975.[11, 7]

5.2 Data Preprocessing

The process of data splitting and the introduction of negative pairs in the model training is a crucial aspect of preparing the dataset for effective and unbiased machine learning. In this framework, the adjacency matrix representing gene interactions is first converted into an edge list that contains pairs of interacting genes, denoted as positive pairs. The process is followed by the careful sampling of negative pairs, i.e., pairs of genes that do not exhibit known interactions, and the partitioning of these positive and negative pairs into training, validation, and test sets. The primary objective is to ensure that the model is exposed to a balanced set of interactions (positive) and non-interactions (negative), thereby improving its ability to generalize beyond the training data.

The data splitting process begins by converting the adjacency matrix A, representing the undirected graph of gene interactions, into an edge list. This adjacency matrix $A \in \mathbb{R}^{N \times N}$, where N is the number of genes, consists of binary entries such that:

$$A_{ij} = \begin{cases} 1, & \text{if gene i interacts with gene j} \\ 0, & \text{otherwise} \end{cases}$$

Self-loops, represented by diagonal elements, are removed, ensuring that the adjacency matrix only reflects interactions between distinct genes.

5.2.1 Sampling Negative Pairs

Since the majority of potential gene pairs do not interact, negative pairs (i.e., non-interactions) must be explicitly sampled. Negative pairs are generated by randomly selecting pairs of genes from the adjacency matrix where no interaction is observed, ensuring no overlap with the positive pairs. These negative pairs represent the absence of interaction between two genes and serve as a counterbalance to the positive pairs.

$$\text{Negative Pairs} = \{(i, j) \mid A_{ij} = 0\}$$

5.2.2 Balancing Positive and Negative Pairs

In most real-world biological networks, positive interactions are sparse, whereas the number of potential non-interactions is exceedingly large. To address this imbalance, an equal number of negative pairs is sampled to match the number of positive pairs. By doing so, the dataset becomes balanced, with an equal number of interacting (positive) and non-interacting (negative) pairs. This is critical to prevent the model from becoming biased towards predicting the majority class, which would predominantly be negative pairs in an imbalanced setting. The sampled negative pairs and the positive pairs are combined to form the full dataset.

5.3 Train, Validation, and Test Split

Once the positive and negative pairs are defined, they are split into training, validation, and test sets. The typical split allocates 80% of the data to the training set, 10% to the validation set, and the remaining 10% to the test set. This ensures that the model is trained on a sufficient amount of data while preserving enough data for validation and testing.

$$\text{Training Set} = \{(i, j, k) \mid k \in \{0, 1\}, \text{ for 80\% of data}\}$$

$$\text{Validation Set} = \{(i, j, k) \mid k \in \{0, 1\}, \text{ for 10\% of data}\}$$

$$\text{Test Set} = \{(i, j, k) \mid k \in \{0, 1\}, \text{ for 10\% of data}\}$$

Here, (i, j, k) refers to the gene pair (i, j) with label k, where $k = 1$ indicates a positive pair (interaction), and $k = 0$ represents a negative pair (non-interaction).

5.4 Experimental settings

5.4.1 libraries

In developing the GNT model, several key libraries were utilized:

- NumPy: Used for efficient numerical computation and array manipulation, particularly for handling high-dimensional gene expression data and tensor operations in the model.
- Pandas: Employed for data management and preprocessing, enabling the organization and transformation of gene expression and interaction datasets.
- PyTorch: Provided the deep learning framework for constructing and training the transformer architecture, supporting dynamic computation graphs and GPU acceleration.

- Transformers: Integrated to implement pre-built transformer models, enabling the GNT model to capture complex dependencies between transcription factors and target genes.

- Scikit-learn: Used for splitting datasets, computing evaluation metrics, and applying cross-validation to ensure robust model performance.

- SciPy: Supplemented numerical optimization and statistical analysis, particularly in fine-tuning model parameters and conducting significance testing.

- NetworkX: Utilized for visualizing and analyzing the topology of the inferred gene regulatory networks, representing genes and interactions as nodes and edges.

5.4.2 Parameters

The parameters used in the GNT (Gene Network Transformer) model define key aspects of the architecture, training process, and optimization, ensuring effective learning of gene regulatory interactions. Each parameter is carefully selected to balance model complexity, computational efficiency, and the accuracy of inferred gene embeddings.

- id embedding size (128): This parameter specifies the dimensionality of the embeddings used to represent gene identities. A size of 128 allows the model to learn compact, informative representations of genes based on their one-hot encoded identities.

- attr embedding size (128): This parameter controls the size of the embeddings for gene attributes, such as biological features derived from gene expression data. Similar to the identity embeddings, a dimension is set to 128.

- representation size (128): This parameter defines the dimensionality of the final gene representations output by the transformer encoder.

- alpha (1): The alpha parameter is a weighting of attr embedding vector over id embedding vector. We can modulate the contribution of expression data to the training and final embedding.

- n neg samples (10): This parameter specifies the number of negative samples generated for each positive interaction during training. Negative sampling helps the model learn to distinguish true gene regulatory relationships from random or spurious connections by presenting non-interacting gene pairs.

- epoch (30): The number of epochs defines how many times the entire dataset is passed through the model during training.
- batch size (256): This parameter determines the number of samples processed in a single forward and backward pass during training. A batch size of 256 strikes a balance between computational efficiency and model convergence.
- learning rate (0.002): The learning rate controls the step size at each iteration while optimizing the model parameters. A learning rate of 0.002 ensures steady convergence without overshooting the optimal solution.

These parameters are crucial for the GNT model's ability to learn low-dimensional, meaningful representations of genes and accurately infer transcription factor-target gene interactions. It was found that lower learning rate might enhance feature learning in datasets with low number of genes. But the results given in the subsequent sections uses the parameters set as mentioned above.

5.5 Results and Analysis

We evaluated the AUC-ROC and AUPRC of GNT over 7 cell types and 3 network types, with the most varying 500 and 1000 genes and compared with the methods scGREAT[2], GENELink[7], GNE[6], DeepSEM[5], Pearson's correlation coefficient, Mutual information [5], SCODE[8], GRNBoost2[9], GENIE3[12] epoch: 30, batch size: 256, learning rate: 0.002
The results of the Gene Network Transformer (GNT) model across various datasets and transcription factor (TF) ranges demonstrate its significant outperformance over existing methods, particularly scGREAT, the current state-of-the-art (SOTA) model. In an interaction network generated from cell-type specific ChIP-Seq data, with TF+500 most varying genes, GNT shows remarkable success across all seven cell types. The average AUC-ROC score of GNT in this scenario is 0.98, which surpasses scGREAT's average score of 0.89 by a margin of 0.09. The average AUPRC of GNT over seven cell types is 0.97, whereas the AUPRC of SOTA is 0.76. This substantial increase reflects GNT's capability to infer more accurate gene regulatory interactions across diverse biological environments. Table 5.1 and Tabel 5.2

show that GNT outperforms the SOTA model in all datasets, highlighting its robustness and scalability for TF+500 gene regulatory networks.

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.50	0.54	0.50	0.50	0.52	0.53	0.52
GRNBoost2	0.49	0.52	0.52	0.53	0.53	0.50	0.52
SCODE	0.50	0.47	0.53	0.51	0.52	0.53	0.45
MI	0.51	0.50	0.55	0.53	0.52	0.49	0.51
PCC	0.47	0.49	0.54	0.51	0.49	0.54	0.55
DeepSEM	0.58	0.55	0.51	0.50	0.51	0.53	0.54
GNE	0.67	0.80	0.52	0.81	0.82	0.83	0.77
GENELink	0.82	0.84	0.71	0.88	0.87	0.89	0.83
scGREAT	0.89	0.91	0.81	0.94	0.93	0.93	0.88
GNT (proposed methodology)	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**	**0.98**	**0.98**

Table 5.1: AUC-ROC: Cell-type Specific ChIP-seq (TFs with most varying 500 genes)

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.15	0.39	0.05	0.31	0.56	0.53	0.50
GRNBoost2	0.15	0.38	0.06	0.32	0.57	0.52	0.50
SCODE	0.15	0.33	0.05	0.32	0.59	0.55	0.44
MI	0.15	0.35	0.05	0.33	0.57	0.50	0.49
PCC	0.14	0.35	0.06	0.31	0.56	0.53	0.52
DeepSEM	0.19	0.40	0.05	0.31	0.56	0.52	0.53
GNE	0.34	0.65	0.06	0.64	0.80	0.78	0.70
GENELink	0.50	0.70	0.11	0.75	0.89	0.89	0.83
scGREAT	0.63	0.86	0.21	0.89	0.95	0.94	0.88
GNT (proposed methodology)	**0.97**	**0.98**	**0.96**	**0.97**	**0.98**	**0.98**	**0.97**

Table 5.2: AUPRC: cell type Specific ChIP-seq (TFs with most varying 500 genes)

Further analysis using cell-type specific ChIP-Seq data with TF+1000 most varying genes reinforces GNT's superior performance. The average

AUC-ROC score of GNT remains at 0.98, which is 0.08 higher than the 0.90 average AUC-ROC of scGREAT. The average AUPRC of GNT is 0.97, whereas the average AUPRC of SOTA is 0.76. we can see a gain of 0.2. As summarized in Table 5.3 and Table 5.4, GNT consistently outperforms the state-of-the-art across all seven cell types.

When applied to non-specific ChIP-Seq data with TF+500 most varying genes, GNT shows its strength, outperforming scGREAT in all datasets except for mESC. The average AUC-ROC score of GNT across all cell types is 0.94, compared to scGREAT's 0.90, representing a 0.04 improvement. The average APRC of GNT is 0.93, whereas the AUPRC of scGREAT is 0.32. As presented in Table 5.5 and Table 5.6, this improvement indicates that GNT is particularly effective in more generalized interaction networks, where specificity may be reduced. The only exception, mESC, suggests that some cell-type-specific factors may influence GNT's performance in certain datasets, but the overall trend remains positive.

The performance of GNT is even more pronounced when Non-Specific ChIP-Seq data with TF+1000 most varying genes is considered. As shown in Table 5.7, GNT achieves an average AUC-ROC of 0.95 across the seven datasets, outperforming scGREAT's 0.89 by 0.06. As shown in Table 5.8, the average AUPRC of GNT is 0.93; meanwhile, the average AUPRC of scGREAT is 0.31. We can see a significantly low AUPRC of the other networks.

In contrast, when applied to STRING datasets with both TF+500 and TF+1000 most varying genes, GNT's performance is comparable to scGREAT, rather than significantly superior. Table 5.9 shows that for TF+500 genes, the average AUC-ROC of GNT is 0.94, while scGREAT achieves a score of 0.93. We can see a gain of 0.3 in average AUPRC in Table 5.10. The average AUPRC of GNT is 0.94 whereas that of scGREAT is 0.60.

Similarly, in Tables 5.7 and 5.9, for TF+1000 genes, GNT's average AUC-ROC is 0.93, only marginally below scGREAT's 0.945. However, we have to note that this result was achieved after integrating bioBERT embeddings. If we drop the BioBERT from scGREAT architecture, it will lose its benefit over GNT. The average AUPRC of GNT is 0.94, whereas that of scGREAT is 0.60. These results indicate that, while GNT performs well, scGREAT retains a slight advantage in this dataset type, only in the case of AUC-ROC, particularly in STRING interaction networks. However, the difference is insignificant, and GNT remains competitive across all datasets.

In Tables 5.5 and 5.7, we observe a slight performance gap between

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.49	0.54	0.52	0.50	0.50	0.51	0.52
GRNBoost2	0.48	0.52	0.53	0.53	0.51	0.49	0.53
SCODE	0.51	0.48	0.53	0.52	0.53	0.53	0.45
MI	0.51	0.49	0.57	0.55	0.49	0.50	0.54
PCC	0.47	0.49	0.54	0.49	0.48	0.54	0.55
DeepSEM	0.58	0.55	0.50	0.51	0.54	0.53	0.57
GNE	0.68	0.81	0.52	0.82	0.84	0.84	0.77
GENELink	0.83	0.85	0.74	0.90	0.90	0.90	0.84
scGREAT	0.89	0.91	0.84	0.95	0.94	0.94	0.89
GNT (proposed methodology)	**0.98**	**0.98**	**0.97**	**0.98**	**0.98**	**0.98**	**0.98**

Table 5.3: AUC-ROC: Cell-type Specific ChIP-seq (TFs with most varying 1000 genes)

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.15	0.38	0.05	0.31	0.54	0.53	0.48
GRNBoost2	0.14	0.37	0.05	0.32	0.54	0.52	0.48
SCODE	0.15	0.33	0.005	0.32	0.58	0.56	0.42
MI	0.15	0.34	0.05	0.34	0.53	0.51	0.49
PCC	0.14	0.34	0.05	0.34	0.53	0.51	0.49
DeepSEM	0.19	0.41	0.05	0.31	0.56	0.54	0.52
GNE	0.34	0.66	0.05	0.65	0.81	0.81	0.68
GENELink	0.50	0.71	0.12	0.76	0.90	0.91	0.81
scGREAT	0.64	0.86	0.18	0.90	0.95	0.95	0.88
GNT (proposed methodology)	**0.98**	**0.98**	**0.92**	**0.97**	**0.98**	**0.98**	**0.98**

Table 5.4: AUPRC: cell type Specific ChIP-seq (TFs with most varying 1000 genes)

GNT and scGREAT in STRING datasets with both TF+500 and TF+1000 genes. For TF+500 genes, GNT's average AUC-ROC is 0.94, while scGREAT achieves 0.945. For TF+1000 genes, GNT scores 0.93, slightly below scGREAT's 0.945. Despite this, the performance of GNT remains comparable,

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.51	0.51	0.55	0.55	0.61	0.66	0.69
GRNBoost2	0.52	0.53	0.52	0.54	0.61	0.64	0.67
SCODE	0.51	0.53	0.48	0.51	0.50	0.52	0.60
MI	0.48	0.48	0.47	0.55	0.57	0.61	0.65
PCC	0.53	0.57	0.47	0.55	0.58	0.61	0.65
DeepSEM	0.55	0.57	0.57	0.55	0.58	0.60	0.63
GNE	0.66	0.69	0.67	0.65	0.53	0.56	0.64
GENELink	0.85	0.87	0.89	0.90	0.86	0.85	0.80
scGREAT	0.90	0.91	0.93	**0.93**	0.88	0.88	0.83
GNT (proposed methodology)	**0.92**	**0.93**	**0.95**	0.92	**0.949**	**0.979**	**0.942**

Table 5.5: AUC-ROC: NonSpecific ChIP-seq (TFs with most varying 500 genes)

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.09	0.09	0.04	0.07	0.15	0.17	0.10
GRNBoost2	0.09	0.09	0.04	0.07	0.15	0.17	0.10
SCODE	0.09	0.09	0.04	0.07	0.15	0.17	0.10
MI	0.04	0.05	0.04	0.07	0.12	0.23	0.10
PCC	0.04	0.05	0.04	0.07	0.15	0.23	0.10
DeepSEM	0.09	0.09	0.09	0.07	0.19	0.19	0.10
GNE	0.09	0.09	0.06	0.07	0.07	0.06	0.04
GENELink	0.17	0.18	0.30	0.21	0.29	0.31	0.09
scGREAT	0.25	0.29	0.44	0.35	0.34	0.35	0.23
GNT (proposed methodology)	**0.92**	**0.92**	**0.93**	**0.90**	**0.94**	**0.98**	**0.94**

Table 5.6: AUPRC: NonSpecific ChIP-seq (TFs with most varying 500 genes)

showing that it is able to keep up with scGREAT even in datasets where STRING interaction networks are utilized.

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.51	0.49	0.48	0.56	0.61	0.68	0.68
GRNBoost2	0.53	0.51	0.48	0.55	0.62	0.68	0.67
SCODE	0.54	0.53	0.45	0.52	0.53	0.55	0.58
MI	0.51	0.48	0.45	0.54	0.62	0.72	0.68
PCC	0.54	0.55	0.45	0.57	0.57	0.64	0.65
DeepSEM	0.56	0.57	0.52	0.56	0.57	0.59	0.62
GNE	0.67	0.65	0.62	0.69	0.54	0.60	0.61
GENELink	0.85	0.86	0.88	0.89	0.85	0.83	0.73
scGREAT	0.90	0.91	0.93	0.93	0.89	0.88	0.81
GNT (proposed methodology)	**0.95**	**0.95**	**0.96**	**0.94**	**0.95**	**0.96**	**0.91**

Table 5.7: AUC-ROC: NonSpecific ChIP-seq (TFs with most varying 1000 genes)

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.06	0.04	0.05	0.07	0.13	0.21	0.10
GRNBoost2	0.06	0.04	0.05	0.07	0.15	0.19	0.11
SCODE	0.06	0.04	0.05	0.03	0.05	0.07	0.06
MI	0.06	0.09	0.05	0.07	0.15	0.26	0.11
PCC	0.06	0.09	0.05	0.07	0.15	0.24	0.11
DeepSEM	0.12	0.09	0.08	0.07	0.18	0.19	0.11
GNE	0.12	0.09	0.08	0.07	0.05	0.07	0.04
GENELink	0.19	0.17	0.29	0.16	0.28	0.40	0.09
scGREAT	0.23	0.25	0.40	0.35	0.30	0.44	0.19
GNT (proposed methodology)	**0.94**	**0.93**	**0.93**	**0.93**	**0.95**	**0.95**	**0.93**

Table 5.8: AUPRC: NonSpecific ChIP-seq (TFs with most varying 1000 genes)

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.65	0.64	0.64	0.64	0.69	0.78	0.73
GRNBoost2	0.62	0.61	0.57	0.61	0.68	0.78	0.74
SCODE	0.44	0.46	0.50	0.51	0.47	0.54	0.68
MI	0.65	0.62	0.51	0.67	0.65	0.72	0.82
PCC	0.61	0.70	0.54	0.64	0.72	0.81	0.74
DeepSEM	0.63	0.63	0.62	0.63	0.67	0.74	0.68
GNE	0.78	0.78	0.83	0.80	0.65	0.74	0.76
GENELink	0.91	0.92	0.94	0.93	0.90	0.91	0.82
scGREAT	**0.95**	**0.96**	0.96	**0.96**	0.94	**0.94**	0.85
GNT (proposed methodology)	0.93	0.95	**0.97**	0.94	**0.95**	0.89	**0.97**

Table 5.9: AUC-ROC: STRING (TFs with most varying 500 genes)

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.10	0.13	0.12	0.13	0.26	0.41	0.33
GRNBoost2	0.10	0.10	0.12	0.13	0.26	0.38	0.40
SCODE	0.05	0.04	0.10	0.09	0.06	0.06	0.13
MI	0.17	0.13	0.07	0.13	0.13	0.14	0.13
PCC	0.07	0.15	0.10	0.13	0.28	0.51	0.38
DeepSEM	0.12	0.10	0.17	0.16	0.34	0.43	0.42
GNE	0.12	0.10	0.22	0.16	0.09	0.11	0.08
GENELink	0.40	0.54	0.56	0.53	0.40	0.49	0.20
scGREAT	0.61	0.68	0.74	0.67	0.57	0.59	0.38
GNT (proposed methodology)	**0.94**	**0.95**	**0.97**	**0.95**	**0.95**	**0.90**	**0.98**

Table 5.10: AUPRC: STRING (TFs with most varying 500 genes)

5.6 Introduction of Negative Pairs and Their Role

The introduction of negative pairs is a key step in binary classification problems such as this, where the task is to distinguish between interacting and non-interacting gene pairs. By generating negative pairs and balancing them

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.66	0.67	0.62	0.64	0.72	0.79	0.77
GRNBoost2	0.63	0.63	0.56	0.61	0.71	0.80	0.78
SCODE	0.45	0.48	0.80	0.55	0.49	0.54	0.72
MI	0.67	0.65	0.52	0.67	0.68	0.72	0.85
PCC	0.65	0.73	0.56	0.64	0.78	0.85	0.75
DeepSEM	0.65	0.65	0.59	0.63	0.67	0.73	0.74
GNE	0.78	0.80	0.81	0.83	0.67	0.73	0.77
GENELink	0.92	0.94	0.93	0.94	0.92	0.93	0.85
scGREAT	**0.95**	**0.96**	**0.97**	**0.96**	0.95	0.93	**0.90**
GNT (proposed methodology)	0.94	0.95	**0.97**	0.95	**0.96**	**0.94**	0.81

Table 5.11: AUC-ROC: STRING (TFs with most varying 1000 genes)

Method	**hESC**	**hHEP**	**mDC**	**mESC**	**mHSC-E**	**mHSC-GM**	**mHSC-L**
GENIE3	0.13	0.18	0.12	0.09	0.22	0.45	0.39
GRNBoost2	0.13	0.15	0.10	0.09	0.24	0.43	0.40
SCODE	0.06	0.07	0.07	0.05	0.04	0.07	0.21
MI	0.19	0.18	0.07	0.07	0.09	0.13	0.15
PCC	0.09	0.22	0.10	0.09	0.28	0.59	0.39
DeepSEM	0.16	0.15	0.12	0.12	0.28	0.43	0.41
GNE	0.16	0.18	0.19	0.12	0.07	0.11	0.10
GENELink	0.40	0.53	0.58	0.52	0.40	0.45	0.29
scGREAT	0.60	0.65	0.75	0.64	0.89	0.62	0.33
GNT (proposed methodology)	**0.95**	**0.94**	**0.97**	**0.96**	**0.96**	**0.95**	**0.72**

Table 5.12: AUPRC: STRING (TFs with most varying 1000 genes)

with the positive pairs, the model is provided with a more comprehensive learning scenario, preventing it from overfitting to the more frequent negative class. Compared to *hard negative sampling* (HNS), our uniformly random negative sampling approach presents several key advantages, particularly in the context of biological networks and gene interaction prediction.

While HNS focuses on making all the unknown interactions with label 0, providing more discriminative information, this strategy can introduce a form of bias that may not always generalize well to unseen data. The rationale behind HNS is to make the model more sensitive to subtle differences between positive and negative samples, which can indeed accelerate the training process and improve model convergence. However, in complex biological systems, where the boundaries between interacting and non-interacting gene pairs are often uncertain or noisy, relying heavily on hard negatives might lead the model to overfit to specific challenging examples. This overfitting could limit its ability to perform well on less distinct, real-world negative pairs that might not share the same nuanced characteristics as the hard negatives.

Our uniformly random negative sampling strategy avoids these potential pitfalls by providing a more diverse and representative selection of negative samples. This ensures that the model is not disproportionately influenced by a subset of difficult negative pairs, which could skew its learning towards specialized cases that may not be reflective of broader biological interactions. By selecting negative pairs randomly, we enable the model to learn a more generalized distinction between interacting and non-interacting genes across a wider range of cases, leading to better overall generalization.

Furthermore, uniformly random negative sampling helps in maintaining computational efficiency. HNS requires additional computational resources to identify and maintain the set of "hard" negatives, which can increase the complexity of the training process, especially in large-scale biological datasets. Our approach, by contrast, reduces this overhead, allowing for a more straightforward and scalable training procedure while still maintaining robust performance. This is particularly beneficial in biological research, where datasets often involve thousands of genes and interactions, necessitating computationally efficient solutions.

In conclusion, while HNS can offer benefits in certain scenarios by introducing more challenging learning tasks, our random sampling approach is better suited for biological interaction prediction. It fosters broader generalization, avoids overfitting, and remains computationally efficient, making it more appropriate for handling the complexity and variability inherent in biological networks. This might be the reason for the low value of AUPRC of scGREAT even after using transformer architecture.

5.7 Novel Interactions

5.7.1 False Positives

During the training and testing of the Gene Network Transformer (GNT) model, coupled with the Support Vector Machine (SVM) classifier, the performance of the model was evaluated on a test set of gene interaction pairs. In this test, true gene interactions were labeled as 1 (indicating a positive interaction), and non-interacting gene pairs were labeled as 0 (indicating a negative or no interaction). Despite the overall success of the model, the classifier produced a number of false positives—gene pairs that were labeled as 0 (non-interacting) but were predicted as 1 (interacting) by the classifier.

5.7.2 literature validation

In order to investigate the validity of these false positives, we conducted an extensive search through the PUBMED repository to find any existing literature that could provide evidence of interactions for these gene pairs. For several of these falsely classified interactions, we were able to identify literature-supported evidence confirming their interactions, which had not been included in the training dataset.

These findings are significant, as they suggest that the model might be capable of uncovering previously unrecognized or poorly characterized gene interactions. The literature-supported interactions, originally classified as false positives, point towards the possibility that the model has predictive power beyond the known interactions provided in the training set.

5.7.3 Evidence

During our exploration of the cell-type specific interaction network for the mESC cell type (with TF+1000 most varying genes), we sought to validate the predicted interactions using existing literature. We performed a comprehensive search of the PubMed repository for articles mentioning "Mus musculus" and identified several gene pairs for which we found direct or indirect evidence of interaction or comparative studies in biological experiments.

Some of these gene pairs were either shown to interact directly, affect each other's regulation, or be used in comparative studies. The following table (Table 5.7) summarizes these newly validated interactions:

From this analysis, we identified several previously unrecognized interactions or regulatory relationships between gene pairs that had been classified as non-interacting by the model but were predicted as interacting. Upon reviewing the literature, we found supporting evidence for these interactions. For instance, the pair RUNX3 and EZH2 was validated by evidence indicating that EZH2 negatively regulates RUNX3, as highlighted in [21]. Another example is the interaction between MYCN and LIFR, where the literature demonstrates the involvement of MYCN in regulating LIFR as part of a signaling pathway, supported by [22]]. These findings enhance the credibility of the model's predictions and suggest that it may uncover interactions that are not well-documented or underrepresented in current biological datasets.

By validating these gene pairs through existing research, we reinforce the hypothesis that the model may identify potential novel regulatory interactions, offering insights that can guide further experimental studies. This highlights the model's predictive power in generating biologically meaningful hypotheses for further investigation.

Gene Pair	**PubMed ID**	**Evidence from Literature**
RUNX3, EZH2	PMC5216711	"For instance, INK4B-ARF-INK4A, p57, bone morphogenetic protein receptor 1B, MyoD and RUNX3 are all negatively regulated by EZH2, which is critical for tumor cell proliferation and aggressiveness."[21]
MYCN, LIFR	PMC8427239	"MiR-9 is regulated by PDGFR, MYC/MYCN, miR-7/c-Myc signal and promotes metastasis via targeting STARD13, E-cadherin, FOXI1, CYP4Z1, LIFR, PTEN and DUSP14 signal pathway."[22]
CTCF, ARID5B	PMC5337971	"This interval, anchored by CTCF binding sites, forms a 'loop domain' which is expected to bring two regions of RUNX3 binding, separated by a linear distance of around 60Kb, into physical contact close to the TSS of ARID5B."[23]
FOXO3, REST	PMC6907729	"RE1-silencing complex (REST), a major neuronal gene repressor in non-neuronal cells, and the aging-associated TF FOXO3 play important roles in controlling neuronal gene expression and show differential activity between fetal and adult/old fibroblasts, resulting in decreased conversion efficacy in aged starting cells."[24]
ETS1, REST	PMC8096796	"REST promotes ETS1-dependent vascular growth in medulloblastoma. Interestingly, REST elevation is also associated with increased expression of vascular endothelial growth factor receptor-1 (VEGFR1) and the proangiogenic transcription factor, E26 oncogene homolog 1 (ETS1), in CGNPs of RESTTG mice compared with cells from WT cerebellar."[25]

Table 5.13: Literature-supported validation of predicted gene interactions in mESC (TF+1000 genes)

Chapter 6

Conclusion

The primary objective of this research was to infer Gene Regulatory Networks (GRNs) with a focus on transcription factors (TFs) and their target genes. GRNs represent the regulatory relationships between various molecular entities, such as transcription factors and their downstream target genes, which ultimately control gene expression and cellular behaviour. The goal was to develop a model capable of predicting novel TF-target interactions, improving the understanding of regulatory mechanisms, and advancing the discovery of new biological pathways.

The Gene Network Transformer (GNT) model was developed with the intention of addressing limitations found in existing GRN inference methods, such as scGREAT, GENELink, GNE, and other methods. These traditional methods often rely on simplistic assumptions or computationally expensive strategies, which hinder their scalability and accuracy.The GNT model builds on the strengths of transformer-based architectures by learning low-dimensional embeddings for genes, which represent their interactions and regulatory potential. Unlike scGREAT, which uses text-based embeddings (e.g., BioBERT embeddings), GNT directly learns embeddings from gene expression data, making it more flexible and applicable to datasets where text-based embeddings are unavailable or insufficient. GNT also avoids biases introduced by strategies like hard negative sampling, which are employed in other models but can skew predictions in certain scenarios.

GNT consistently outperformed existing models in various experimental datasets, particularly in cell-type specific and non-specific ChIP-Seq datasets, as well as in STRING interaction networks, demonstrating its robustness in capturing gene regulatory interactions. The model's AUC-ROC scores in almost all test cases, both with TF+500 and TF+1000 most varying genes, consistently surpassed the state-of-the-art model scGREAT by a significant margin.

This approach is particularly advantageous because it removes the need for prior dependence over domain knowledge and can infer relationships directly from the data, making GNT applicable to a wide variety of datasets. GNT successfully predicted TF-target interactions across a variety of datasets. The model identified known interactions with high accuracy and uncovered previously unrecognized interactions that were later validated through literature searches.

One of the key strengths of the GNT model is its ability to discover novel gene interactions that were not explicitly annotated in the training datasets. As noted during the evaluation process, several false positives—gene pairs predicted to interact despite being labelled as non-interacting—were later validated through extensive searches in the PubMed repository. For example, gene pairs such as RUNX3 and EZH2[21], and MYCN and LIFR[22], initially classified as false positives, were confirmed through literature evidence, showcasing GNT's potential to predict biologically relevant interactions that are not well-documented. These findings suggest that GNT can be used to hypothesize new regulatory mechanisms, offering valuable insights for experimental validation and further research. The evidence of novel interactions discovered using the GNT model reinforces its utility in advancing biological research. These interactions, many of which were previously unknown or underreported, provide new avenues for experimental validation.

This model does not strictly infer causality in gene interactions. There were plenty of interactions within false positives, which indicated analogous behaviours between gene products in a biological pathway. By scanning literature, we found direct and indirect regulatory effects between several of these false positive pairs, indicating that the GNT model has the potential to identify hidden regulatory connections that might otherwise go unnoticed.

References

[1] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

[2] Yuchen Wang, Xingjian Chen, Zetian Zheng, Lei Huang, Weidun Xie, Fuzhou Wang, Zhaolei Zhang, and Ka-Chun Wong. scgreat: Transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics. *Iscience*, 27(4), 2024.

[3] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[4] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10):1491–1498, 2015.

[5] 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.

[6] 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.

[7] 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.

[8] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

[9] 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.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need.(nips), 2017. *arXiv preprint arXiv:1706.03762*, 10:S0140525X16001837, 2017.

[11] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

[12] 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.

[13] Hantao Shu, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng, and Jianzhu Ma. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, 1(7):491–501, 2021.

[14] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

[15] Luz Garcia-Alonso, Christian H Holland, Mahmoud M Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, 29(8):1363–1375, 2019.

[16] Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015:bav095, 2015.

[17] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. Trrust v2: an expanded reference database of human and

mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.

[18] ENCODE Project Consortium et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[19] Shinya Oki, Tazro Ohta, Go Shioi, Hideki Hatanaka, Osamu Ogasawara, Yoshihiro Okuda, Hideya Kawaji, Ryo Nakaki, Jun Sese, and Chikara Meno. Ch ip-atlas: a data-mining suite powered by full integration of public ch ip-seq data. *EMBO reports*, 19(12):e46255, 2018.

[20] Huilei Xu, Caroline Baroukh, Ruth Dannenfelser, Edward Y Chen, Christopher M Tan, Yan Kou, Yujin E Kim, Ihor R Lemischka, and Avi Ma'ayan. Escape: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database*, 2013:bat045, 2013.

[21] Bo Pang, Xiang-Rong Zheng, Jing-xia Tian, Tai-hong Gao, Guang-yan Gu, Rui Zhang, Yi-Bing Fu, Qi Pang, Xin-Gang Li, and Qian Liu. Ezh2 promotes metabolic reprogramming in glioblastomas through epigenetic repression of eaf2-hif1α signaling. *Oncotarget*, 7(29):45134, 2016.

[22] Yichen Liu, Qiong Zhao, Tao Xi, Lufeng Zheng, and Xiaoman Li. Microrna-9 as a paradoxical but critical regulator of cancer metastasis: Implications in personalized medicine. *Genes & Diseases*, 8(6):759–768, 2021.

[23] James B Studd, Jayaram Vijayakrishnan, Minjun Yang, Gabriele Migliorini, Kajsa Paulsson, and Richard S Houlston. Genetic and regulatory mechanism of susceptibility to high-hyperdiploid acute lymphoblastic leukaemia at 10q21. 2. *Nature communications*, 8(1):14616, 2017.

[24] Larissa Traxler, Frank Edenhofer, and Jerome Mertens. Next-generation disease modeling with direct conversion: a new path to old neurons. *FEBS letters*, 593(23):3316–3337, 2019.

[25] Shavali Shaik, Shinji Maegawa, Amanda R Haltom, Feng Wang, Xue Xiao, Tara Dobson, Ajay Sharma, Yanwen Yang, Jyothishmathi Swaminathan, Vikas Kundra, et al. Rest promotes ets1-dependent vascular growth in medulloblastoma. *Molecular oncology*, 15(5):1486–1506, 2021.