Gene Regulatory Network Inference from Pre-trained Single-Cell Transcriptomics Transformer with Joint Graph Learning

Sindhura Kommu [1] **Yizhi Wang** [2] **Yue Wang** [2] **Xuan Wang** [1]

Abstract

Inferring gene regulatory networks (GRNs) from single-cell RNA sequencing (scRNA-seq) data is a complex challenge that requires capturing the intricate relationships between genes and their regulatory interactions. In this study, we tackle this challenge by leveraging the single-cell BERT-based pre-trained transformer model (scBERT), trained on extensive unlabeled scRNA-seq data, to augment structured biological knowledge from existing GRNs. We introduce a novel joint graph learning approach scTransNet that combines the rich contextual representations learned by pre-trained single-cell language models with the structured knowledge encoded in GRNs using graph neural networks (GNNs). By integrating these two modalities, our approach effectively reasons over both the gene expression level constraints provided by the scRNA-seq data and the structured biological knowledge inherent in GRNs. We evaluate scTransNet on human cell benchmark datasets from the BEELINE study with cell type-specific ground truth networks. The results demonstrate superior performance over current state-of-the-art baselines, offering a deeper understanding of cellular regulatory mechanisms.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed the exploration of gene expression patterns at the individual cell level (Jovic et al., 2022), offering an unprecedented opportunity to unravel the intricate regulatory mechanisms governing cellular identity and function (Pratapa et al., 2020). One such promising application is the inference of gene regulatory networks (GRNs) which represent the complex interplay between transcription factors (TFs) and their downstream target genes (Akers & Murali, 2021; Cramer, 2019). A precise understanding of GRNs is crucial for understanding cellular processes, molecular functions, and ultimately, developing effective therapeutic interventions (Biswas et al., 2021).

However, inferring GRNs from scRNA-seq data is challenging due to cell heterogeneity (Wagner et al., 2016), cell cycle effects (Buettner et al., 2015), and high sparsity caused by dropout events (Kharchenko et al., 2014), which can impact accuracy and robustness. Additionally, the availability of labeled scRNA-seq data corresponding to a GRN is limited, making it challenging to train models from scratch. Traditional unsupervised or self-supervised models, while not reliant on label information, often struggle to effectively handle the noise, dropouts, high sparsity, and high dimensionality characteristics of scRNA-seq data (Moerman et al., 2019; Matsumoto et al., 2017; Zeng et al., 2023). Supervised methods are also proposed for GRN reconstruction (Zhao et al., 2022; Shu et al., 2022; KC et al., 2019; Chen & Liu, 2022a) but struggle to handle batch effects and fail to leverage latent gene-gene interaction information effectively limiting their generalization capabilities.

Recent advancements in large language models (LLMs) and the pre-training followed by fine-tuning paradigm (Devlin et al., 2019; OpenAI, 2023) have significantly contributed to the development of transformer-based architectures tailored for scRNA-seq data analysis (Yang et al., 2022; Cui et al., 2024; Chen et al., 2023; Theodoris et al., 2023). These models effectively leverage vast amounts of unlabeled scRNA-seq data to learn contextual representations and capture intricate latent interactions between genes. To address the limitations of the current methods, we effectively leverage one of these large-scale pre-trained transformer models, namely scBERT (Yang et al., 2022), which has been pre-trained on large-scale unlabelled scRNA-seq data to learn domain-irrelevant gene expression patterns and interactions from the whole genome expression. By fine-tuning scBERT on user specific scRNA-seq datasets, we can mitigate batch effects and capture latent gene-gene interactions for down-

[1]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, United States [2]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA, United States. Correspondence to: Sindhura Kommu <sindhura@vt.edu>, Yizhi Wang <yzwang@vt.edu>, Yue Wang <yuewang@vt.edu>, Xuan Wang <xuanw@vt.edu>.

ICML 2024 AI for Science workshop, Vienna, Austria. PMLR 235, 2024.

stream tasks.

We propose an innovative knowledge-aware supervised GRN inference framework, scTransNet (see Figure 1), which integrates pre-trained single-cell language models with structured knowledge of GRNs. Our approach combines gene representations learned from scBERT with graph representations derived from the corresponding GRNs, creating a unified context-aware and knowledge-aware representation (Feng et al., 2020). This joint learning approach enables us to surpass the accuracy of current state-of-the-art methods in supervised GRN inference. By harnessing the power of pre-trained transformer models and incorporating biological knowledge from diverse data sources, such as gene expression data and gene regulatory networks, our approach paves the way for more precise and robust GRN inference. Ultimately, this methodology offers deeper insights into cellular regulatory mechanisms, advancing our understanding of gene regulation.

2. Related Work

Several methods have been developed to infer GRNs from scRNA-seq data, broadly categorized into unsupervised and supervised methods.

Unsupervised methods primarily include information theory-based, model-based, and machine learning-based approaches. Information theory-based methods, such as mutual information (MI) (Margolin et al., 2006), Pearson correlation coefficient (PCC) (Salleh et al., 2015; Raza & Jaiswal, 2013), and partial information decomposition and context (PIDC) (Chan et al., 2017), conduct correlation analyses under the assumption that the strength of the correlation between genes is is positively correlated with the likelihood of regulation between them. Model-based approaches, such as SCODE (Matsumoto et al., 2017), involve fitting gene expression profiles to models that describe gene relationships, which are then used to reconstruct GRNs (Shu et al., 2021; Tsai et al., 2020).

Machine learning-based unsupervised methods, like GENIE3 (Huynh-Thu et al., 2010) and GRNBoost2 (Moerman et al., 2019), utilize tree-based algorithms to infer GRNs. These methods are integrated into tools like SCENIC (Aibar et al., 2017; Van de Sande et al., 2020), employing tree rules to learn regulatory relationships by iteratively excluding one gene at a time to determine its associations with other genes. Despite not requiring labeled data, these unsupervised methods often struggle with the noise, dropouts, high sparsity, and high dimensionality typical of scRNA-seq data. Additionally, the computational expense and scalability issues of these tree-based methods, due to the necessity of segmenting input data and iteratively establishing multiple models, present further challenges for large datasets.

Supervised methods, including DGRNS (Zhao et al., 2022), convolutional neural network for co-expression (CNNC) (Yuan & Bar-Joseph, 2019), and DeepDRIM (Chen et al., 2021), have been developed to address the increasing scale and inherent complexity of scRNA-seq data. Compared with unsupervised learning, supervised models are capable of detecting much more subtle differences between positive and negative pairs (Yuan & Bar-Joseph, 2019).

DGRNS (Zhao et al., 2022) combines recurrent neural networks (RNNs) for extracting temporal features and convolutional neural networks (CNNs) for extracting spatial features to infer GRNs. CNNC (Yuan & Bar-Joseph, 2019) converts the identification of gene regulation into an image classification task by transforming the expression values of gene pairs into histograms and using a CNN for classification. However, the performance of CNNC (Yuan & Bar-Joseph, 2019) is hindered by the issue of transitive interactions. To address this, DeepDRIM (Chen et al., 2021) considers the information from neighboring genes and converts TF–gene pairs and neighboring genes into histograms as additional inputs, thereby reducing the occurrence of transitive interactions to some extent. Despite their success there exist certain limitations to the employment of CNN model-based approaches for GRN reconstruction. First of all, the generation of image data not only gives rise to unanticipated noise but also conceals certain original data features. Additionally, this process is time-consuming, and since it changes the format of scRNA-seq data, the predictions made by these CNN-based computational approaches cannot be wholly explained.

In addition to CNN-based methods, there are also other approaches such as GNE (Kc et al., 2019) and GRN-Transformer (Shu et al., 2022). GNE (gene network embedding) (Kc et al., 2019) is a deep learning method based on multilayer perceptron (MLP) for GRN inference applied to microarray data. It utilizes one-hot gene ID vectors from the gene topology to capture topological information, which is often inefficient due to the highly sparse nature of the resulting one-hot feature vector. GRN-Transformer (Shu et al., 2022) constructs a weakly supervised learning framework based on axial transformer to infer cell-type-specific GRNs from scRNA-seq data and generic GRNs derived from the bulk data.

More recently, graph neural networks (GNNs) (Wu et al., 2020), which are effective in capturing the topology of gene networks, have been introduced into GRN prediction methods. For instance, GENELink (Chen & Liu, 2022b) treats GRN inference as a link prediction problem and uses graph attention networks to predict the probability of interaction between two gene nodes. However, existing methods often suffer from limitations such as improper handling of batch effects, difficulty in leveraging latent gene-gene interaction

information, and making simplistic assumptions, which can impair their generalization and robustness.

3. Approach

As shown in (Figure 1), our approach contains four parts: BERT encoding, Attentive Pooling, GRN encoding with GNNs and Output layer. The input scRNA-seq datasets are processed into a cell-by-gene matrix, $\mathbf{X} \in R^{N \times T}$, where each element represents the read count of an RNA molecule. Specifically, for scRNA-seq data, the element denotes the RNA abundance for gene $t \in \{0, 1, ..., T\}$ in cell $n \in \{0, 1, ..., N\}$. In subsequent sections, we will refer to this matrix as the raw count matrix. Let us denote the sequence of gene tokens as $\{g_1, ..., g_T\}$, where T is the total number of genes.

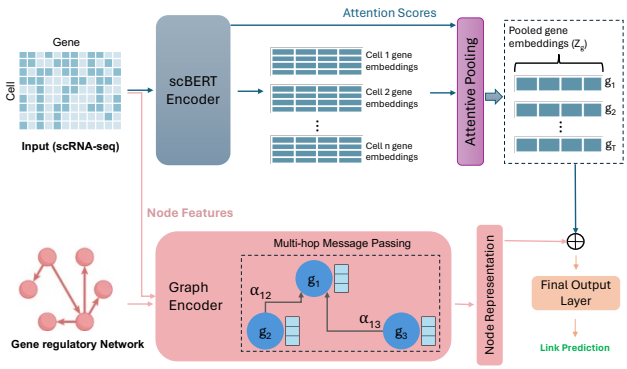

Figure 1. **Overview of scTransNet framework for supervised GRN inference** with BERT Encoding Layer (top left; Section 3.1), Attentive Pooling (top right; Section 3.2), GRN encoding with GNNs (bottom left; Section 3.3) and Final Output layer (bottom right; Section 3.4) . It augments the output from graph encoder (for knowledge understanding) with scBERT encoder (for contextual understanding) to infer regulatory interdependencies between genes.

3.1. BERT Encoding Layer

(Yang et al., 2022; Cui et al., 2024; Chen et al., 2023; Theodoris et al., 2023) show that pre-trained transformer models have a strong understanding of gene-gene interactions across cells and have achieved state-of-the-art results on a variety of single-cell processing tasks. We use scBERT (Yang et al., 2022) as the backbone, which is a successful pre-trained model with the advantage of capturing long-distance dependency as it uses Performer (Choromanski et al., 2022) to improve the scalability of the model to tolerate over 16,000 gene inputs.

The scBERT model adopts the advanced paradigm of BERT and tailors the architecture to solve single-cell data analysis. The connections of this model with BERT are given as follows. First, scBERT follows BERT's revolutionary method to conduct self-supervised pre-training (Devlin et al., 2019) and uses Transformer as the model backbone (Choromanski et al., 2022). Second, the design of embeddings in scBERT is similar to BERT in some aspects while having unique features to leverage gene knowledge. From this perspective, the gene expression embedding could be viewed as the token embedding of BERT. As shuffling the columns of the input does not change its meaning (like the extension of BERT to understand tabular data with TaBERT (Yin et al., 2020)), absolute positions are meaningless for gene. Instead gene2vec is used to produce gene embeddings, which could be viewed as relative embeddings (Du et al., 2019) that capture the semantic similarities between any of two genes. Third, Transformer with global receptive field could effectively learn global representation and long-range dependency without absolute position information, achieving excellent performance on non-sequential data (such as images, tables) (Parmar et al., 2018; Yin et al., 2020).

In spite of the gene embedding, there is also a challenge on how to utilize the transcription level of each gene, which is actually a single continuous variable. The gene expression could also be considered as the occurrence of each gene that has already been well-documented in a biological system. Drawing from bag-of-words (Zhang et al., 2010) insight, the conventionally used term-frequency-analysis method is applied that discretizes the continuous expression variables by binning, and converts them into 200-dimensional vectors, which are then used as token embeddings for the scBERT model.

For each token g_t in a cell, we construct its input representation as:

$$h_t^0 = emb_{gene2vec}(g_t) + emb_{expr}(g_t) \quad (1)$$

where $emb_{gene2vec}(g_t)$ represents gene2vec embedding (Du et al., 2019) of gene g_t analogous to position embedding in BERT and $emb_{expr}(g_t)$ represents expression embedding of the gene expression of g_t analogous to token embedding in BERT.

Such input representations are then fed into L successive Transformer encoder blocks, i.e.,

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

so as to generate deep, context-aware representations for genes. The final hidden states $\{h_t^L\}_{t=1}^T$ are taken as the output of this layer (Devlin et al., 2019; Vaswani et al., 2023).

3.2. Attentive Pooling

After extracting the BERT encodings we further utilize the attention scores across cells from the model to select the

most representative cells for pooling of each gene representation. For each input gene token g_t we get the embeddings for all cells denoted as $\{h^L_{t(n)}\}^N_{n=1}$, where N is the number of cells.

The quadratic computational complexity of the BERT model, with the Transformer as its foundational unit, does not scale efficiently for long sequences. Given that the number of genes in scRNA-seq data can exceed 20,000, this limitation becomes significant. To address this issue, scBERT employs a matrix decomposition variant of the Transformer, known as Performer (Choromanski et al., 2022), to handle longer sequence lengths. In a regular Transformer, the dot-product attention mechanism maps Q, K, and V, which are the encoded representations of the input queries, keys, and values for each unit. The bidirectional attention matrix is formulated as follows:

$$\begin{aligned} Att(Q,K,V) = D^{-1}(QK^T)V, \\ D = diag(QK^T 1_L) \end{aligned} \tag{3}$$

where $Q = W_q X$, $K = W_K X$, $V = W_V X$ are linear transformations of the input X; W_Q, W_K and W_V are the weight matrices as parameters; 1_L is the all-ones vector of length L; and diag(.) is a diagonal matrix with the input vector as the diagonal.

The attention matrix in Performer is described as follows:

$$\begin{aligned} \hat{Att}(Q,K,V) = \hat{D}^{-1}(Q^{'}((K^{'})^T V)), \\ \hat{D} = diag(Q^{'}((K^{'})^T 1_L)) \end{aligned} \tag{4}$$

where $Q^{'} = \phi(Q)$, $K^{'} = \phi(K)$, and the function $\phi(x)$ is defined as:

$$\phi(X) = \frac{c}{\sqrt{m}} f(\omega^T X) \tag{5}$$

where c is a positive constant, ω is a random feature matrix, and m is the dimensionality of the matrix.

The attention weights can be obtained from equation 3, modified by replacing V with V^0, where V^0 contains one-hot indicators for each position index. All the attention matrices are integrated into one matrix by taking an element-wise average across all attention matrices in multi-head multi-layer Performers. In this average attention matrix for each cell, $A(i,j)$ represents how much attention from gene i was paid to gene j. To focus on the importance of genes to each cell n, the attention matrix is summed along the columns into an attention-sum vector a_n, and its length is equal to the number of genes. These attention scores of gene g_t are obtained across cells and normalized denoted as $\{a^n_t\}^N_{n=1}$

These normalized scores are used for weighted aggregation of gene embeddings across cells. We aggregate each cell's gene representations together into one gene-level cell embedding such that the updated matrix is of the form $Z \in R^{T \times d}$, where d is the dimension of the output gene embedding.

$$Z_g[t] = \oplus^N_{i=1} h^L_{t(n)} \cdot a^n_t \tag{6}$$

3.3. GRN encoding with GNNs

In this module, we use raw count matrix as the features of the genes. Subsequently, we utilize graph convolutional network (GCN)-based interaction graph encoders to learn gene features by leveraging the underlying structure of the gene interaction graph.

Let us denote the prior network as $G = \{V, E\}$, where V is the set of nodes and E is the set of edges. To perform the reasoning on this prior gene regulatory network G, our GNN module builds on the graph attention framework (GAT) (Velickovic et al., 2017), which induces node representations via iterative message passing between neighbors on the graph. In each layer of this GNN, the current representation of the node embeddings $\{v^{l-1}_1, ..., v^{l-1}_T\}$ is fed into the layer to perform a round of information propagation between nodes in the graph and yield pre-fused node embeddings for each node:

$$\begin{aligned} \{\tilde{v_1}^l, ..., \tilde{v_T}^l\} = GNN(\{v^{l-1}_1, ..., v^{l-1}_T\}) \\ for \;\; l = 1, ..., M \end{aligned} \tag{7}$$

Specifically, for each layer l, we update the representation $\tilde{v_t}^l$ of each node by

$$\tilde{v_t}^l = f_n\Big(\sum_{v_s \in \eta_{v_t} \cup \{v_t\}} \alpha_{st} \mathbf{m}_{st}\Big) + {v_t}^{l-1} \tag{8}$$

where η_{v_t} represents the neighborhood of an arbitrary node v_t, m_{st} denotes the message one of its neighbors v_s passes to v_t, α_{st} is an attention weight that scales the message m_{st}, and f_n is a 2-layer MLP. The messages m_{st} between nodes allow entity information from a node to affect the model's representation of its neighbors, and are computed in the following manner:

$$\mathbf{r}_{st} = f_r(\tilde{\mathbf{r}}_{st}, \mathbf{u}_s, \mathbf{u}_t) \tag{9}$$

$$\mathbf{m}_{st} = f_m(\mathbf{v}^{(l-1)}_s, \mathbf{u}_s, \mathbf{r}_{st}) \tag{10}$$

where $\mathbf{u}_s$, $\mathbf{u}_t$ are node type embeddings, $\tilde{\mathbf{r}}_{st}$ is a relationembedding for the relation connecting v_s and v_t, f_r is a 2-layer MLP, and f_m is a linear transformation. The attention weights α_{st} scale the contribution of each neighbor's message by its importance, and are computed as follows:

$$\mathbf{q}_s = f_q(\mathbf{v}^{(l-1)}_s, \mathbf{u}_s) \tag{11}$$

$$\mathbf{k}_t = f_k(\mathbf{v}^{(l-1)}_t, \mathbf{u}_t, \mathbf{r}_{st}) \tag{12}$$

$$\alpha_{st} = \frac{\exp(\gamma_{st})}{\sum_{v_s \in \eta_{v_t} \cup \{v_t\}} \exp(\gamma_{st})}, \gamma_{st} = \frac{\mathbf{q}_s^\intercal \mathbf{k}_t}{\sqrt{D}} \quad (13)$$

where f_q and f_k are linear transformations and $\mathbf{u}_s$, $\mathbf{u}_t$, $\mathbf{r}_{st}$ are defined the same as above.

3.4. Final Output Layer

In the final output layer, we concatenate the input gene representations Z_g from the BERT encoding layer with the graph representation of each gene from GNN to get the final gene embedding.

We input these final embeddings of pairwise genes i and j into two channels with the same structure. Each channel is composed of MLPs to further encode representations to low-dimensional vectors which serve for downstream similarity measurement or causal inference between genes.

4. Experimental Setup

4.1. Benchmark scRNA-seq datasets

The performance of scTransNet is evaluated on two human cell types using single-cell RNA-sequencing (scRNA-seq) datasets from the BEELINE study (Pratapa et al., 2020): human embryonic stem cells (hESC (Yuan & Bar-Joseph, 2019)) and human mature hepatocytes (hHEP (Camp et al., 2017)). The cell-type-specific ChIP-seq ground-truth networks are used as a reference for these datasets. The scRNA-seq datasets are preprocessed following the approach described in the (Pratapa et al., 2020), focusing on inferring interactions outgoing from transcription factors (TFs). The most significantly varying genes are selected, including all TFs with a corrected P-value (Bonferroni method) of variance below 0.01. Specifically, 500 and 1000 of the most varying genes are chosen for gene regulatory network (GRN) inference. The scRNA-seq datasets can be accessed from the Gene Expression Omnibus (GEO) with accession numbers GSE81252 (hHEP) and GSE75748 (hESC). The evaluation compares the inferred gene regulatory networks to known ChIP-seq ground-truth networks specific to these cell types.

4.2. Implementation and Training details

Data Preparation We utilized the benchmark networks that containing labeled directed regulatory dependencies between gene pairs. These dependencies were classified as positive samples (labeled 1) if present in the network, and negative samples (labeled 0) if absent. Due to the inherent network density, the number of negative samples significantly outnumbered positive samples. To address the class imbalance, known transcription factor (TF)-gene pairs are split into training (2/3), and test (1/3) sets. Positive training samples were randomly selected from the known TF-gene pairs. Moreover, 10% of TF-gene pairs are randomly selected from training samples for validation. The remaining positive pairs formed the positive test set. Negative samples were generated using the following strategies: 1) Unlabeled interactions: All unobserved TF-gene interactions outside the labeled files were considered negative instances. 2) Hard negative sampling: To enhance model learning during training, we employed a uniformly random negative sampling strategy within the training set. This involved creating "hard negative samples" by pairing each positive sample (g1, g2) with a negative sample (g1, g3), where both share the same gene g1. This approach injects more discriminative information and accelerates training. 3) Information leakage prevention: Negative test samples were randomly selected from the remaining negative instances after generating the training and validation sets. This ensured no information leakage from the test set to the training process. The positive-to-negative sample ratio in each dataset was adjusted to reflect the network density i.e.

$$\frac{Positive}{Negative} = \frac{NetworkDensity}{1 - NetworkDensity} \quad (14)$$

Model Training To account for the class imbalance, we adopted two performance metrics: Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC). The supervised model was trained for 100 iterations with a learning rate of 0.003. The Graph Neural Network (GNN) architecture comprised two layers with hidden layer sizes of 256 and 128 units, respectively.

Evaluation All reported results are based solely on predictions from the held-out test set. To ensure a fair comparison, identical training and validation sets were utilized for all evaluated supervised methods. This approach eliminates potential bias introduced by different data splits.

4.3. Baseline Methods

To assess the effectiveness of our model in predicting GRNs, we compare our model scTransNet against the existing baseline methods commonly used for inferring GRNs, as follows:

- GNNLink (Mao et al., 2023) is a graph neural network model that uses a GCN-based interaction graph encoder to capture gene expression patterns.
- GENELink (Chen & Liu, 2022b) proposes a graph attention network (GAT) approach to infer potential GRNs by leveraging the graph structure of gene regulatory interactions.
- GNE (gene network embedding) (Kc et al., 2019) proposes a multilayer perceptron (MLP) approach to encode

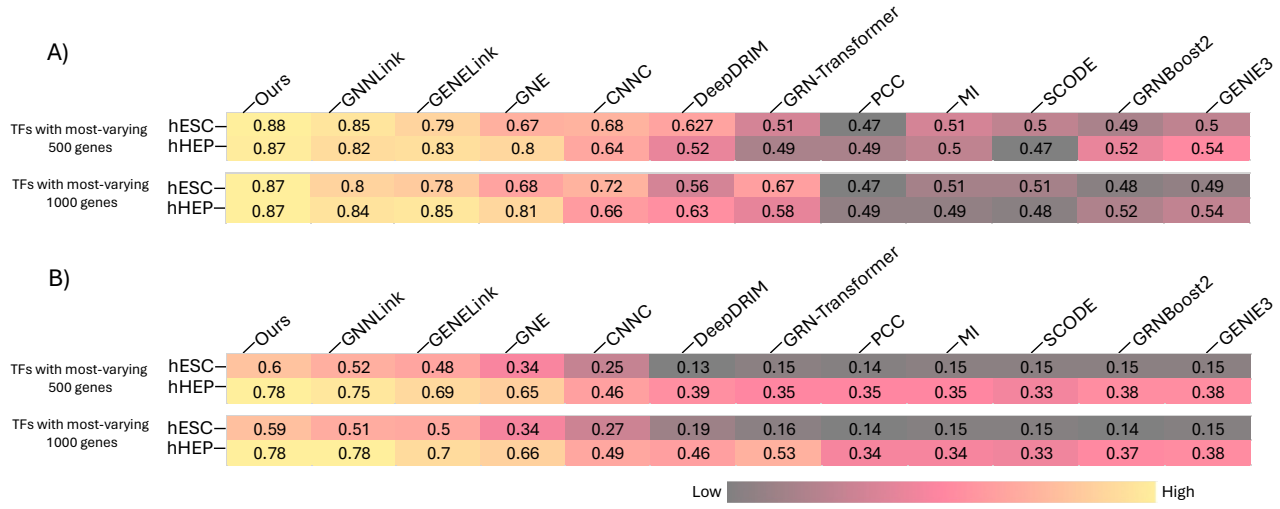

Figure 2. Summary of the GRN prediction performance of scTransNet in the (A) AUROC metric (top) (B) and the AUPRC metric (bottom). Our evaluation is conducted on two human single-cell RNA sequencing (scRNA-seq) datasets, with a cell-type-specific ground-truth network. The scRNA-seq datasets consist of significantly varying transcription factors (TFs) and the 500 or 1000 most-varying genes.

both gene expression profiles and network topology for predicting gene dependencies.

- CNNC (Yuan & Bar-Joseph, 2019) proposes inferring GRNs using deep convolutional neural networks (CNNs).
- DeepDRIM (Chen et al., 2021) 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 (Shu et al., 2022) 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) (Salleh et al., 2015; Raza & Jaiswal, 2013) is a traditional statistical method for measuring the linear correlation between two variables, often used as a baseline for GRN inference.
- Mutual information (MI) (Margolin et al., 2006) is an information-theoretic measure of the mutual dependence between two random variables, also used as a baseline for GRN inference.
- SCODE (Matsumoto et al., 2017) is a computational method for inferring GRNs from single-cell RNA-seq data using a Bayesian framework.
- GRNBoost2 (Moerman et al., 2019) is a gradient boosting-based method for GRN inference.
- GENIE3 (Huynh-Thu et al., 2010) 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.

These methods represent a diverse range of approaches, including traditional statistical methods, machine learning techniques, and deep learning models, for inferring gene regulatory networks from various types of data, such as bulk and single-cell RNA-seq, as well as incorporating additional information like network topology and chromatin accessibility.

5. Results

5.1. Performance on benchmark datasets

The results (see Figure 2) demonstrate that scTransNet outperforms state-of-the-art baseline methods across all four benchmark datasets, achieving superior performance in terms of both AUROC and AUPRC evaluation metrics. Notably, scTransNet's AUROC values are approximately 5.4% and 7.4% higher on average compared to the second-best methods, namely GNNLink (Mao et al., 2023) and GENELink (Chen & Liu, 2022b), respectively. Similarly, scTransNet's AUPRC values show an impressive improvement of approximately 7.4% and 16% on average over GNNLink and GENELink, respectively.

To gain further insights, we analyzed scTransNet's final gene regulatory network (GRN) predictions and compared them with those from GENELink. Our analysis revealed

Table 1. Comparison of average AUROC and AUPRC evaluation metrics on human benchmark datasets, validating the roles of the GNN encoder, scBERT encoder, and Attentive Pooling using the cell-type-specific ChIP-seq network for our proposed method.

	w/o GNN encoder		w/o scBERT encoder		w/o Attentive Pooling		scTransNet	
Dataset	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
hESC	0.842	0.544	0.853	0.572	0.860	0.569	**0.880**	**0.595**
hHEP	0.830	0.725	0.854	0.753	0.862	0.683	**0.870**	**0.780**

that scTransNet effectively captured all the gene regulatory interactions predicted by GENELink. This finding suggests that by incorporating joint learning, scTransNet does not introduce additional noise to the predictive power of the graph representations. Instead, it enhances the predictive capability through the scBERT encoder in its architecture.

Figure 3 provides a visualization of a partial subgraph of the ground truth GRN, highlighting the predictions made by scTransNet that were not captured by GENELink, which solely relies on graphs for predicting gene-gene interactions. Additionally, the figure visualizes the ground truth labels that scTransNet failed to capture. In summary, the comparative analysis demonstrates that scTransNet effectively captures all the regulatory interactions predicted by GENELink while leveraging joint learning to improve predictive performance. The visualization illustrates the additional interactions scTransNet could predict beyond GENELink, as well as the ground truth interactions it missed, providing insights into the strengths and limitations of the proposed method.

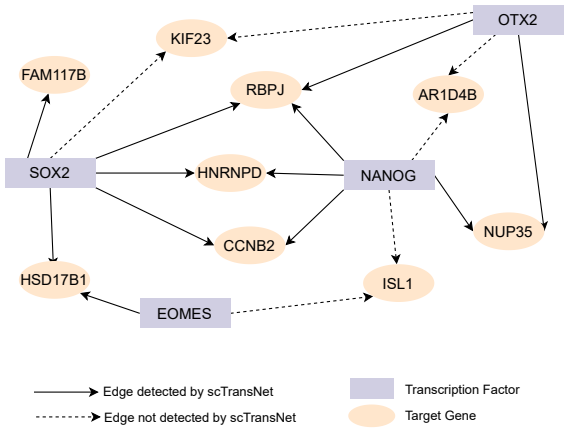

Figure 3. GRN prediction performance of scTransNet on a partial ground truth subgraph. Solid line edges depict ground truth regulatory interactions correctly predicted by scTransNet but missed by the baseline GENELink method, which relies solely on graph representations. Notably, scTransNet effectively identified all regulatory links predicted by GENELink (not visualized). Dotted line edges represent ground truth interactions that scTransNet failed to capture reveal its limitations and providing insights for further improvement. Overall, this highlights scTransNet's strength in leveraging joint learning to uncover additional true regulatory interactions beyond graphs.

5.2. Discussion and Ablations

To evaluate the effectiveness of jointly learning from pre-trained scRNA-seq language models (Yang et al., 2022), which capture rich contextual representations, and Gene Regulatory Networks (GRNs), which encode structured biological knowledge, we compare the average Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) metrics with and without these encoders across the four human cell type benchmark datasets (Pratapa et al., 2020). The average AUROC and AUPRC scores are calculated across both the TFs+500 highly variable genes and TFs+1000 highly variable genes datasets for each human cell data type (i.e, hESC (human embryonic stem cells) and hHEP (human mature hepatocytes)). Additionally, we validate the importance of incorporating Attentive Pooling (Section 3.2) by contrasting the results when using average pooling of gene embeddings across cells instead of attentive pooling. Consistent parameter settings are employed across all four human cell benchmark datasets, with Cell-type-specific ChIP-seq network data serving as the ground truth.

Effect of Graph Neural Network Component: The results demonstrate the significant impact of incorporating the Graph Neural Network (GNN) encoder component in the proposed method. With the GNN encoder, the average AUROC value across all the human cell type datasets is 87.5%, and the average AUPRC value is 68.7%. In contrast, without the GNN encoder, the average AUROC drops to 83.6%, and the average AUPRC decreases to 63.4%. The inclusion of the GNN encoder leads to an improvement of 4.6% in the average AUROC and a notable 8.3% increase in the average AUPRC. These results highlight the consistent performance enhancement provided by the GNN encoder across both AUROC and AUPRC metrics for the human cell type benchmark datasets. The GNN encoder plays a crucial role in the architecture as the task is formulated as a supervised Gene Regulatory Network (GRN) inference problem, aiming to identify potential gene regulatory dependencies

given prior knowledge of the GRN. The GNN models the regulatory interactions as a graph, learning node representations that effectively encode the network topology and gene interdependencies present in the GRN, which serves as the primary source of biological knowledge. The results in Table 1 justify the use of this structural graph representation for understanding the complex regulatory networks in single-cell transcriptomics data.

Effect of Pre-trained Single-Cell Transcriptomics Transformer: The removal of the scBERT encoder also leads to a drop in performance, with the average AUROC decreasing from 87.5% to 85.3%, and the average AUPRC declining from 68.7% to 66.2% across both cell types (see Table 1). The inclusion of scBERT representations improves the AUROC by 2.6% and the AUPRC by 3.8%. While the improvement is less significant compared to the GNN encoder, this is expected as the contextual representations from scRNA-seq data are learned through pre-training on millions of unlabeled single cells and then fine-tuned for the specific cell type. In addition to rich contextual representations, scBERT captures long-range dependencies between genes by leveraging self-attention mechanisms and pretraining on large-scale unlabeled scRNA-seq data (Pratapa et al., 2020). This comprehensive understanding of gene-gene interactions and semantic relationships allows for effective modeling of complex, non-linear gene regulatory patterns that extend beyond immediate neighbors in the gene regulatory network.

The contextual representations learned by the pre-trained Transformer facilitate the identification of intricate regulatory relationships that might be overlooked by traditional methods focused on local neighborhoods or predefined gene sets. The ability to capture global context and long-range dependencies is a key advantage of pre-trained single-cell Transformer models for deciphering the intricate gene regulatory mechanisms governing cellular states and identities. The improvement shown in Table 1 justifies the effectiveness of this approach.

Effect of Attentive Pooling Mechanism: The impact of incorporating Attentive Pooling is evaluated by comparing the AUROC and AUPRC metrics with and without attentive pooling across four datasets. As shown in Table 1, the inclusion of attentive pooling results in a slight improvement, with a 1.6% increase in the average AUROC and a 9.6% increase in the average AUPRC. While the improvement is not significant, the experiments confirm that attentive pooling offers some support for the gene regulation task. We believe that the significance of attentive pooling will be more pronounced when scaling the method to larger datasets. The cell type data is sparse and of low quality. However, the attention weights learned from scBERT (Pratapa et al., 2020) demonstrate that the marker genes are automatically learned for each cell. Consequently, attentive pooling helps to effectively focus on high-quality cell data by removing noise. By employing an attentive pooling mechanism, scTransNet selectively focuses on the most informative cells for each gene, mitigating noise and filtering out irrelevant information, thereby enhancing the quality of the input data used for GRN inference.

6. Conclusion and Future Work

In this work, we propose scTransNet, a joint graph learning inference framework that integrates prior knowledge from known Gene Regulatory Networks (GRNs) with contextual representations learned by pre-trained single-cell transcriptomics Transformers. Our approach aims to effectively boost GRN prediction by leveraging the complementary strengths of structured biological knowledge and rich contextual representations. We evaluate our method on four human cell scRNA-seq benchmark datasets and demonstrate consistent improvements over current baselines in predicting gene-gene regulatory interactions. Our framework comprises four key modules: a GNN encoder to capture the network topology from known GRNs, a scBERT encoder to learn contextual representations from scRNA-seq data, an Attentive Pooling mechanism to focus on informative cells, and a Final Output layer for prediction. The synergistic combination of these modules is verified to be effective in accurately inferring gene regulatory dependencies.

Moving forward, we plan to incorporate the knowledge integration process directly into the fine-tuning of the Transformer model, aiming to fuse information across layers more effectively. Additionally, we will evaluate our approach on various other datasets, including simulated datasets, to further validate its robustness and generalizability. Beyond GRN inference, we intend to investigate the advantages of jointly learning single-cell Transformers and structured biological knowledge for other cell-related tasks. These tasks include cell type annotation, identifying echo archetypes (Luca et al., 2021), and enhancing the interpretability of single-cell models. By leveraging the complementary strengths of contextual representations and structured knowledge, we aim to advance the understanding and analysis of complex cellular processes and regulatory mechanisms.

Acknowledgements

Our work is sponsored by the NSF NAIRR Pilot and PSC Neocortex, Commonwealth Cyber Initiative, Children's National Hospital, Fralin Biomedical Research Institute (Virginia Tech), Sanghani Center for AI and Data Analytics (Virginia Tech), Virginia Tech Innovation Campus, and a generous gift from the Amazon + Virginia Tech Center for Efficient and Robust Machine Learning.

Impact Statement

"This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

References

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.

Akers, K. and Murali, T. Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*, 26:87–97, 2021.

Biswas, S., Manicka, S., Hoel, E., and Levin, M. Gene regulatory networks exhibit several kinds of memory: quantification of memory in biological and random transcriptional networks. *iScience*, 24(3):102131, March 2021.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33 (2):155–160, 2015.

Camp, J. G., Sekine, K., Gerber, T., Loeffler-Wirth, H., Binder, H., Gac, M., Kanton, S., Kageyama, J., Damm, G., Seehofer, D., et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature*, 546(7659):533–538, 2017.

Chan, T. E., Stumpf, M. P., and Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3): 251–267, 2017.

Chen, G. and Liu, Z.-P. Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data. *Bioinformatics*, 38(19):4522–4529, 08 2022a. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac559. URL https://doi.org/10.1093/bioinformatics/btac559.

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

Chen, J., Cheong, C., Lan, L., Zhou, X., Liu, J., Lyu, A., Cheung, W. K., and Zhang, L. 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.

Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., and Han, J.-D. J. Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1): 223, Jan 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-35923-4. URL https://doi.org/10.1038/s41467-023-35923-4.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., and Weller, A. Rethinking attention with performers, 2022.

Cramer, P. Organization and regulation of gene transcription. *Nature*, 573(7772):45–54, 2019.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. 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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., and Zhi, D. Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20:7–15, 2019.

Feng, Y., Chen, X., Lin, B. Y., Wang, P., Yan, J., and Ren, X. Scalable multi-hop relational reasoning for knowledge-aware question answering, 2020.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., and Luo, Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.*, 12(3):e694, March 2022.

KC, K., Li, R., Cui, F., Yu, Q., and Haake, A. R. Gne: a deep learning framework for gene network inference by aggregating biological information. *BMC Systems Biology*, 13(2):38, Apr 2019. doi: 10.1186/s12918-019-0694-y. URL https://doi.org/10.1186/s12918-019-0694-y.

Kc, K., Li, R., Cui, F., Yu, Q., and Haake, A. R. Gne: a deep learning framework for gene network inference by aggregating biological information. *BMC systems biology*, 13:1–14, 2019.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

Luca, B. A., Steen, C. B., Matusiak, M., Azizi, A., Varma, S., Zhu, C., Przybyl, J., Espín-Pérez, A., Diehn, M., Alizadeh, A. A., van de Rijn, M., Gentles, A. J., and Newman, A. M. Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell*, 184(21):5482–5496.e28, 2021. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2021.09.014. URL https://www.sciencedirect.com/science/article/pii/S0092867421010618.

Mao, G., Pang, Z., Zuo, K., Wang, Q., Pei, X., Chen, X., and Liu, J. 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.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., and andCalifano, A. Aracne: Analgorithm forthereconstructionofgeneregulatorynetworksina mammaliancellularcontext. *BMCbioinformatics*, 7, 2006.

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S., Ko, S. B., Gouda, N., Hayashi, T., and Nikaido, I. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., and Aerts, S. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.

OpenAI, R. Gpt-4 technical report. *ArXiv*, 2303, 2023.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. Image transformer. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4055–4064. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/parmar18a.html.

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. 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.

Raza, K. and Jaiswal, R. Reconstruction and analysis of cancer-specific gene regulatory networks from gene expression profiles. *arXiv preprint arXiv:1305.5750*, 2013.

Salleh, F. H. M., Arif, S. M., Zainudin, S., and Firdaus-Raih, M. Reconstructing gene regulatory networks from knock-out data using gaussian noise model and pearson correlation coefficient. *Computational biology and chemistry*, 59:3–14, 2015.

Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., and Ma, J. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, 1 (7):491–501, 2021.

Shu, H., Ding, F., Zhou, J., Xue, Y., Zhao, D., Zeng, J., and Ma, J. Boosting single-cell gene regulatory network reconstruction via bulk-cell transcriptomic data. *Briefings in Bioinformatics*, 23(5):bbac389, 2022.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. 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.

Tsai, M.-J., Wang, J.-R., Ho, S.-J., Shu, L.-S., Huang, W.-L., and Ho, S.-Y. Grema: modelling of emulated gene regulatory networks with confidence levels based on evolutionary intelligence to cope with the underdetermined problem. *Bioinformatics*, 36(12):3833–3840, 2020.

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., et al. A scalable scenic workflow for single-cell gene regulatory network analysis. *Nature protocols*, 15(7):2247–2276, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio', P., and Bengio, Y. Graph attention networks. *ArXiv*, abs/1710.10903, 2017. URL https://api.semanticscholar.org/CorpusID:3292002.

Wagner, A., Regev, A., and Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145–1160, 2016.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., and Yao, J. 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.

Yin, P., Neubig, G., Yih, W.-t., and Riedel, S. TaBERT: Pretraining for joint understanding of textual and tabular data. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8413–8426, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.745. URL https://aclanthology.org/2020.acl-main.745.

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

Zeng, Y., He, Y., Zheng, R., and Li, M. Inferring single-cell gene regulatory network by non-redundant mutual information. *Briefings in Bioinformatics*, 24(5):bbad326, 2023.

Zhang, Y., Jin, R., and Zhou, Z.-H. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52, 2010.

Zhao, M., He, W., Tang, J., Zou, Q., and Guo, F. A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. *Briefings in bioinformatics*, 23(2):bbab568, 2022.