Research Proposal

OCTOBER 22 2023

NEURAL NETWORK BASED GENE REGULATORY NETWORK DISCOVERY

Prepared for

Dr. Pawel Kudzia
Iryna Liubchak
Dr. Tiam Heydari

Prepared by

Muhammad Umar Ali
68477884

Table of Contents

Leveraging Deep Learning Models for Discovering Logical Relations in Gene Regulatory Networks

1 Introduction

The research project revolves around the challenges in creating executable inferred gene regulatory network (GRN) models, which are crucial for understanding gene activity and cell fate decisions. The proposed research aims to leverage deep learning to enhance the scalability and accuracy of GRN models, advancing our understanding of complex gene networks, and potentially paving the way for therapeutic cell reprogramming.

1.1 Statement of Problem

Stem cells are pluripotent cells with the potential to differentiate, or become, into any cell type in the human body, such as red blood cells or liver cells [1]. The process of differentiation and the ultimate cell fate are dictated by complex gene networks known as GRNs [2]. Understanding these networks and the rules they follow to direct cell fate decisions is a key challenge in stem cell biology. The advent of single-cell RNA sequencing (scRNA-seq) has revolutionized our ability to investigate these processes. By providing a comprehensive snapshot of gene expression in individual cells, scRNA-seq allows for a detailed exploration of the cellular and molecular mechanisms that drive stem cell differentiation. There has been work in using formal reasoning to identify the essential components of these gene networks from scRNA-seq data, which has led to successful predictions in some cases, but we still lack a scalable and comprehensive platform that can work with high-throughput data [3, 4].

1.2 Purpose of Research

Most recently, the Zandstra Lab developed a Python software package called IQCELL [5]. The IQCELL platform is designed to reconstruct these gene networks directly from scRNA-seq. This addresses the need for a comprehensive platform that can work with high-throughput data for reconstructing GRNs. However, the IQCELL platform uses the Z3 engine for its automated reasoning process to find Boolean relationships between genes [6]. Despite its efficiency in logical reasoning, the Z3 engine can face scalability challenges with larger or complex GRNs. In fact, the IQCELL limits the numbers of of total regulators in a network to six genes (four activators and two repressors). Moreover, genes in the IQCELL platform are filtered based on an automatic threshold, thus not necessarily select for genes whose expression levels vary significantly across cell developmental time. Thus, there exists a need to implement a more robust, scalable, and accurate reasoning engine for discovering logical gene relations within a GRN that can determine gene-specific thresholds.

1.3 Research Significance

The research proposed herein aims to address this gap by leveraging the power of deep learning to enhance the scalability and accuracy of GRN models. This research is significant because it not only advances our understanding of the complex gene networks that govern cell behavior, but also opens up new avenues for cell reprogramming for therapeutic purposes. Refer to Figure 1 for an image illustrating how GRNs can allows for the elucidation of cell type and state specificity, progression of dynamic trajectories, and differences between conditions. The proposed research will significantly contribute to the field of systems biology and have far-reaching implications for regenerative medicine and disease treatment.

2 Background

2.1 Stem Cell Fates & scRNA-seq

Cells regulate their gene activity through intracelluar and extracelluar signals, specifically through the involvement of transcription factors (TFs). TFs are proteins that bind to specific regions of the DNA and typically either upregulate (promote the expression) or downregulate (repress the expression) of target genes [8]. One of the reasons that gene regulation is of interest is because of the role that it plays in determining the cell fate decision of stem cells and early progenitor cells. The early principles of gene regulation were unearthed through studying developmental systems, Theodor Boveri's work, a prominent bioligist, highlighted that the blueprint of an organism's development is located inside chromosomes. Building on this foundation, Britten and Davidson proposed that the GRN is the underlying mechanism through which stem cells differentiate into specialized cells [9]. These cell fate decisions are governed by

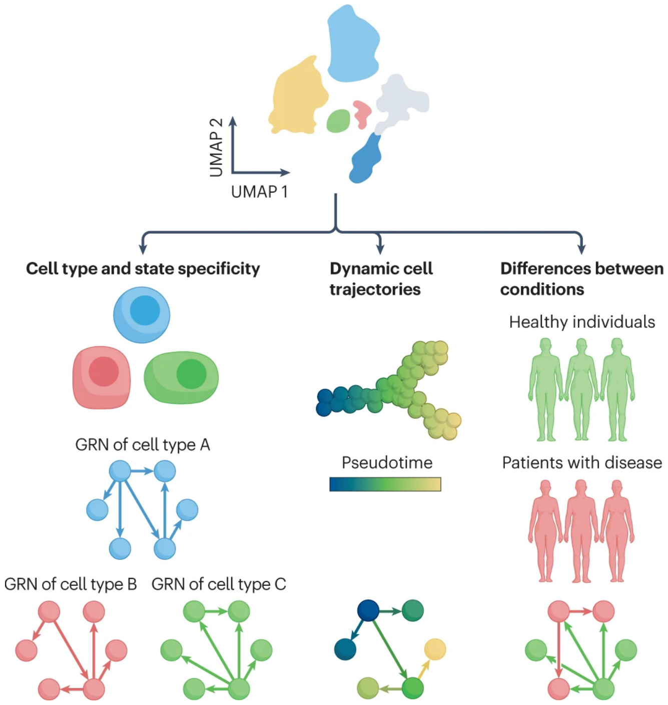

Figure 1. GRNs generated from scRNA-seq allows understanding of cell type and state specificity, progression of dynamic trajectories and differences between conditions. Illustration from Badia et al [7].

a dense array of interacting TFs such that the regulated genes themselves form a complex network called GRNs [2]. The information contained within GRNs holds the key to not only understanding how cells currently differentiate into various cell types, but also how we can reprogram cells for therapeutic purposes.

Single-cell RNA sequencing (scRNA-seq) is a powerful technique that analyzes mRNA expression in individual cells, providing detailed insights into the gene activity of each cell. By capturing mRNA expression at the single-cell level, scRNA-seq reveals the unique genetic signatures of individual cells within a complex biological system. We can now use scRNA-seq to investigate how stem cells make critical decisions about their future roles by examining the complex networks of interacting genes, the GRN. Many gold standard methods often require several gene perturbation experiments to piece together enough information to construct a GRN model from scRNA-seq data, which is labour intensive, expensive and time-consuming [10].

The first gene regulatory model involved a process of genetic perturbation, measurement, and model development, introducing the Boolean paradigm where genes are considered either "on" (expressed) or "off" (not expressed) [11]. Building on this, Dunn et al harnessed computational models to simplify intricate GRNs, leading to an efficient model that accurately forecasts stem cell behavior [12]. The model was further streamlined by identifying and eliminating five nonessential components. When compared with open-source microarray and chromatin immunoprecipitation sequencing data, these models proved superior as those derived solely from perturbation data fell short in satisfying experimental observations. Thus, demonstrating the power of leveraging large amounts of scRNA-seq data and computational methods as a powerful tool to create gene regulatory models.

2.2 Computational Efforts for Gene Regulatory Models

GRNs have been a central concept in biology for several decades, but it is only with the advent of high-throughput technologies such as scRNA-seq that we have begun to appreciate their complexity. Over the years, our understanding of GRNs has evolved significantly. GRN's dynamism and adaptability are what enable GRNs to guide the intricate process of cell differentiation.

Computational and mathematical modeling have played an indispensable role in deciphering GRNs. Boolean models, for instance, represent gene expression in binary terms (either "on" or "off"), capturing the essential dynamics of GRNs and enabling predictions of cell behavior. Boolean models start by binarizing the expression data to construct the state space for the model, in which each cell represents a vector of gene expression [13]. Refer to Figure 2 for the generic pipeline process of generating a boolean model. In the traditional Boolean modeling of gene regulatory networks, gene expression is discretized into two states: "on" or "off", based on a fixed threshold. This binary representation, although

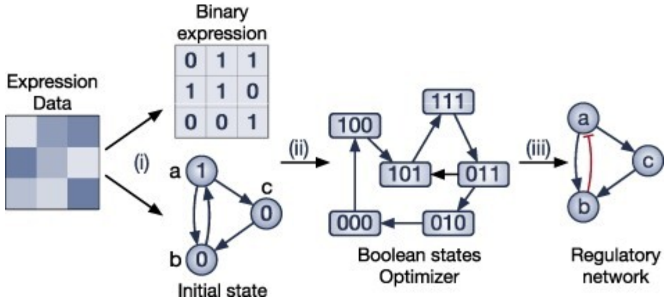

Figure 2. Typical Boolean models have the following pipeline: (i) convert gene data into binary to create initial boolean states, (ii) adjust these states based on the binary data, and (iii) generate the executable GRN. Illustration taken from Nguyen et al [13].

useful for capturing the essential dynamics of gene regulation, may oversimplify the complex reality of gene expression, which often involves a continuum of states rather than a simple dichotomy. Additionally, the fixed threshold used for discretization may not accurately reflect the unique characteristics and context-dependency of individual genes.

Two popular and state of the art Boolean models are GENIE3 and GRNBoost2 [14, 15]. These are techniques focus on ranking potential regulatory links between genes based on a Random Forest (RG) and Gradient Boosting Machine (GBM) respectively. Both methods process gene expression data and use machine learning techniques to predict gene interactions. GENIE3 uses a RF algorithm to assign scores to potential gene links; GRNBoost2 uses a GBM for the same purpose. GENIE3 constructs separate random for models for each gene, while GRNBoost2 uses a single GBM model for all genes, which can be more efficient.

While these models are good at performing inferences, they fail to be able to predict gene perturbations, find trajectories, or be scalable to a large number of genes. Which is why platforms such as IQCELL is such a powerful platform for end-to-end analysis of scRNA-seq [5]. However, there exists only a handful papers on the application of neural network, or deep learning methods, on inferring GRNs [16, 17, 18]. Thus, there is a knowledge deficit in the area of applying neural networks for the inference of GRNs.

3 Research Objective

The objectives of this research project are to:

1. Use artificial neural networks (ANN) to infer gene regulatory networks (GRN) from candidate activator and repressor genes, acquired from downstream IQCELL analysis via scRNA-seq data, to create executable boolean GRNs
2. Integrate the ANN models into a graph neural network (GNN) based approach for GRN inference
3. Validate the ANN and GNN model performance by running inference on datasets from the Zandstra lab

4 Description of Proposed Research

4.1 Experimental Methodology

We propose to use a neural network architecture for GRN reasoning and integrate it with IQCELL such that the whole system can be trained end-to-end through backpropagation to infer Boolean GRNs. We break this process into four modules: (1) making an AND gate network for candidate activator genes, (2) making an OR gate network for candidate repressor genes, (3) combining both networks, and (4) expanding the network to several cells to form a graph neural network (GNN).

We plan to use several datasets to train, validate, and assess the neural networks, namely: (i) synthetic scRNA-seq data, (ii) early mouse T-cell and red blood cell development data, and (iii) experimental lab data. This data includes information on gene expression levels in individual cells at different time points or conditions, which we can use to find gene expression at different development checkpoints called pseudo-time.

4.1.1 Logical Modelling using Neural Networks

The first step in our research project will be curate a synthetic dataset that will act as our standard for building a mock version of our neural network. The purpose of this is to construct simple GRNs that can easily be explained. Our dataset will follow a general multi-input regulatory function for N number of total genes:

$$\frac{dY}{dt} = f(X_1,\ldots,X_N,Y) - \alpha Y = \frac{\prod \beta_a \left(\frac{X_a}{K_a}\right)^H a}{1 + \prod \beta_a \left(\frac{X_a}{K_a}\right)^H a + \frac{1}{N_r}\sum \beta_r \left(\frac{X_r}{K_r}\right)^H r} - \alpha Y \tag{1}$$

The gene expression curve for a given gene X_i, where $i \in [1,N]$, will be randomly generated and then binarized so that it resembles the output of the IQCELL function. The resulting curves will go from being a continuous and differentiable bell-shaped curve to a non-differentiable square-wave like curve. H is the Hill equation, and $H(X)$ will fit the random expression information to a Hill model as described in Equation 1. Then we explicitly encode a regulatory relationship between genes X_A and X_C on gene X_B, for example Equation 3 will encode an AND relation between the two genes.

$$X_i = H\left(N(\mu,\sigma^2)\right), \tag{2}$$

$$X_B = X_A \cdot X_C \tag{3}$$

We can then concatenate all gene expression vectors and create a matrix G where each row is X_i^T, the i-th gene expression curve. With the curated dataset, we can now construct a simple model to that will learn to predict target gene X_k and the learn the expression threshold value for a gene, θ_i, for when a gene is turned "on". Learning θ_i will introduce a novel discretization strategy that allows each gene to have its unique threshold based on the target gene. This is achieved by using sigmoid functions that can be learned for each target gene separately. This approach provides a more nuanced and accurate representation of gene expression, as it can capture the specific regulatory relationships between individual genes.

We can predict the regulatory relationship of a set of genes by designing a simple neural network that emulates the function of AND and OR gates. The AND network will be built for candidate activator genes, indicating that all inputs must be present for a gene to be activated. On the other hand, the OR gate network will be built for candidate repressor genes, suggesting that the presence of any input will lead to the gene being repressed. Refer to Figure 3 to see the architecture for a network to learn the AND relation between a set of genes.

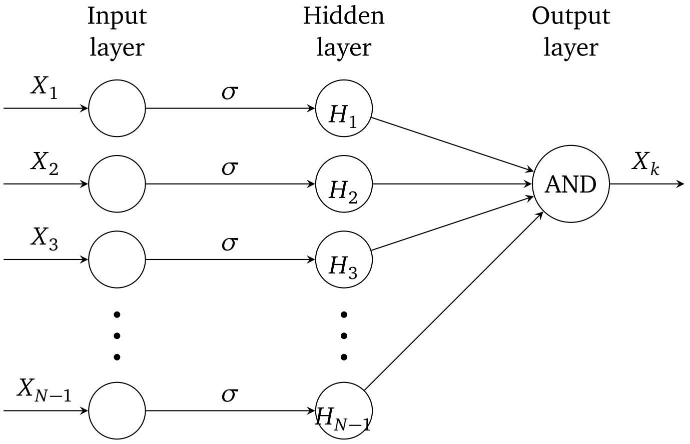

Figure 3. Neural network architecture for the AND model that predicts the gene expression of gene X_i from all other genes in network. The network also learns the thresholds of gene expression for switching "on".

The output of the resulting network can be described as:

$$\hat{X}_k = \prod \left(\sigma\left(W^T G\right)\right) \tag{4}$$

Where W is the weights matrix of the learned expression threshold value. Using this network as a basis, we can create a similar model for the OR network.

In terms of a neural network model, the logical rule for the regulation of a gene X_k can be described as follows. Let's denote the output of the AND network as $AND(H_1, H_2, \ldots, H_N)$ and the output of the OR network as $OR(H_1, H_2, \ldots, H_N)$, where $H_1, H_2, \ldots, H_N$ represent the activations of the neurons in the hidden layer. The goal is to add a final layer which can be represented as a function f, which takes as input the outputs of the AND and OR networks.

The final layer can be represented as a function f that combines the outputs of the `AND` and `OR` networks to produce the output of the final layer. In our specific requirements of the model, f is defined as $f(A,R) = A \wedge \neg R$. This model can represent a wide variety of logical rules for the regulation of a gene. The weights in the `AND` and `OR` networks can be learned from data, allowing the model to adapt to different regulatory rules.

As mentioned, once the two separate models exist for `AND` as well as `OR`, we can concatenate them together. In order to integrate back these models back into IQCELL, we simply feed the `AND` model candidate activator genes A_i and the `OR` model candidate repressor genes R_i. Then we add our final layer to the model using the f function. This will give us a general logical rule for the regulation of a gene G, this can be described as:

$$\hat{X}_k = \left(\prod_{i=0}^{N-1} A_i\right) \cdot \neg\sigma\left(\frac{1}{N-1}\sum_{i=0}^{N-1} R_i\right) \tag{5}$$

Lastly, in order to ensure that our model only learns the most relevant genes when learning regulatory gene relationships, we must introduce sparsity regularization into the loss function. Ideally, we would like to use $\mathbf{L_0}$ regularization, which penalizes non-zero weights, thus driving the solution towards sparsity. However, $\mathbf{L_0}$ regularization is a combinatorics problem that cannot be solved with gradient descent and thus cannot be used for optimizing neural networks [19]. Thus, in this research project we will use a smooth approximation of a $\mathbf{L_q}$ regularization where $0 < q < 1$.

$$\mathscr{L}(W) = \left\|\hat{X}_k - X_k\right\|_2^2 + \frac{\lambda}{2}\|W\|_q \tag{6}$$

4.1.2 Graph Neural Network Integration

After the construction of a neural network engine to reason GRN relations, the next step is to expand into a graph neural network (GNN) that will also take reason the regulatory relationships between different genes. By expanding our network to use a GNN to learn and represent complex regulatory relationships between genes, we are taking advantage of the inherently graph-like structure of GRNs. In fact, this will allow us to incorporate more genes in our GRN model, thereby increasing the accuracy of its predictive capabilities. The GNN will have the following components:

- **Node features:** The node features will be the gene expression levels. Each node will thus represent a gene, and its features will be a vector of the gene expression levels under various conditions.
- **Edge features:** The edge features will represent the regulatory interactions between genes. These could be activation or repression events, encoded as binary features.
- **Message passing:** The message passing mechanism will be used to propagate information across the network. This will allow the model to capture the dependencies between genes.

The GNN will be trained to predict the gene expression levels of individual genes, given the gene expression levels of its regulatory genes and the interactions between them.

Formally, the graph neural network can be represented as a function $g : \mathbb{R}^{N\times D} \rightarrow \mathbb{R}^N$, where N is the number of genes and D is the dimensionality of the gene expression levels. The function g maps the gene expression levels of all genes to the predicted gene expression levels of the target gene. The function g can be trained to minimize the difference between the predicted and actual gene expression levels, as measured by a suitable loss function.

To transform the `AND` and `OR` networks into a GNN, we will employ a unified learning approach. In this framework, the GNN will be adapted to represent both the `AND` and `OR` networks, predicting gene expression levels and the regulatory relationships between genes simultaneously. This comprehensive training approach will enable the GNN to encapsulate the functionalities of the `AND` and `OR` networks, allowing for an intricate interplay between them during the learning process, thereby resulting in a more accurate and robust model.

Where g is the GNN, and $AND(H_1, H_2, \ldots, H_N)$ and $OR(H_1, H_2, \ldots, H_N)$ are the outputs of the `AND` and `OR` networks, respectively. The GNN's update function to predict the expression level of a gene X_k at time $t+1$ can be described as follows:

$$X_k^{t+1} = g\left(f\left(X_i^t, W_{ki}\right)\right) \tag{7}$$

Here, X_k^{t+1} represents the predicted expression level of gene X_k at time $t+1$. The function g will perform the sum operation $\sum_{i \in \mathcal{N}(k)}$ and aggregate all the messages from all neighboring genes. $\mathcal{N}(k)$ represents the set of neighboring genes that regulate gene k. $f\left(X_i^t, W_i^k\right)$ is a function that computes a message based on the current state of gene i at time t, X_i^t, and the weight W_i that represents the regulatory interaction (edge) from gene i to gene k. The output of the function g gives the updated state of gene k at time $t+1$, which represents its predicted expression level.

To ensure the model only learns the most relevant interactions and dependencies between genes, we will also incorporate sparsity regularization into the loss function of the GNN, in a similar way as we did for the `AND` and `OR` networks. This will help to ensure that the simplest model is found and improve the interpretability of the model.

4.2 Expected Results

By employing neural networks in the inference of GRNs from scRNA-seq data, we anticipate the ability to accurately predict the regulatory interactions between genes within cells. This method should replicate the GRN performance derived from the current IQCELL platform. The new method should be able to: infer 74% of causal gene interactions of the reported gene interactions in early T-cell development & erythropoiesis [5], show agreement for all perturbations with experimental studies, and in early T-cell development, out of 18 perturbations done in the IQCELL paper, make prediction that are similar in 17 cases [5]. Moreover, we hope that the inferences on the in-house experimental scRNA-seq also yields similar results found wiht the early T-cell development & erythropoiesis scRNA-seq datasets.

The utilization of neural networks also offers the advantage of scalability. The new model should be able to handle a larger number of genes across various cells compared to traditional methods. This scalability is crucial as it allows for a broader and more comprehensive exploration of potential GRNs, accommodating the complexity and diversity inherent in biological systems.

We hope that the enhanced scalability and integration capabilities offered by neural networks will not only match the predictive performance of IQCELL but also provide new avenues for improving the accuracy and utility of GRN prediction.

5 References

[1] Wojciech Zakrzewski, Maciej Dobrzyński, Maria Szymonowicz, and Zbigniew Rybak. Stem cells: past, present, and future. *Stem cell research & therapy*, 10(1):1–22, 2019.

[2] Stefan Semrau and Alexander van Oudenaarden. Studying lineage decision-making in vitro: emerging concepts and novel tools. *Annual review of cell and developmental biology*, 31:317–345, 2015.

[3] Fiona K Hamey, Sonia Nestorowa, Sarah J Kinston, David G Kent, Nicola K Wilson, and Berthold Göttgens. Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proceedings of the National Academy of Sciences*, 114(23):5822–5829, 2017.

[4] Nir Piterman, V Moignard, S Woodhouse, L Haghverdi, J Lilly, Y Tanaka, A Wilkinson, F Buettner, I Macaulay, W Jawaid, et al. Decoding the regulatory network for blood development from single cell gene expression measurements. *Nature Biotechnology*, 2015.

[5] Tiam Heydari, Matthew A. Langley, Cynthia L Fisher, Daniel Aguilar-Hidalgo, Shreya Shukla, Ayako Yachie-Kinoshita, Michael Hughes, Kelly M. McNagny, and Peter W Zandstra. Iqcell: A platform for predicting the effect of gene perturbations on developmental trajectories using single-cell rna-seq data. *PLoS computational biology*, 18(2):e1009907, 2022.

[6] Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008.

[7] Pau Badia-i Mompel, Lorna Wessels, Sophia Müller-Dott, Rémi Trimbour, Ricardo O Ramirez Flores, Ricard Argelaguet, and Julio Saez-Rodriguez. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, pages 1–16, 2023.

[8] Seungsoo Kim and Joanna Wysocka. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell*, 2023.

[9] Roy J Britten and Eric H Davidson. Gene regulation for higher cells: A theory: New facts regarding the organization of the genome provide clues to the nature of gene regulation. *Science*, 165(3891):349–357, 1969.

[10] Isabelle S Peter, Emmanuel Faure, and Eric H Davidson. Predictive computation of genomic logic processing functions in embryonic development. *Proceedings of the National Academy of Sciences*, 109(41):16434–16442, 2012.

[11] Stuart A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22 (3):437–467, 1969.

[12] S-J Dunn, Graziano Martello, Boyan Yordanov, Stephen Emmott, and AG Smith. Defining an essential transcription factor program for naive pluripotency. *Science*, 344(6188):1156–1160, 2014.

[13] Hung Nguyen, Duc Tran, Bang Tran, Bahadir Pehlivan, and Tin Nguyen. A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. *Briefings in bioinformatics*, 22(3):bbaa190, 2021.

[14] 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.

[15] 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.

[16] Yanglan Gan, Xin Hu, Guobing Zou, Cairong Yan, and Guangwei Xu. Inferring gene regulatory networks from single-cell transcriptomic data using bidirectional rnn. *Frontiers in Oncology*, 12:899825, 2022.

[17] 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.

[18] Juexin Wang, Anjun Ma, Qin Ma, Dong Xu, and Trupti Joshi. Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks. *Computational and structural biotechnology journal*, 18:3335–3343, 2020.

[19] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.