**S.T. Yau High School Science Award**

**Research Report**

**The Team**

Name of team member: Anish Mudide
School: Phillips Exeter Academy
City, Country: Exeter, USA

Name of supervising teacher: Alexander Wu
Job Title: PhD Student
School/Institution: MIT CSAIL
City, Country: Cambridge, USA

**Title of Research Report**

Neural Granger Causality on DAGs Reconstructs Gene Regulatory Networks from Single-cell Transcriptomics

**Date**

September 5, 2022

# Neural Granger Causality on DAGs Reconstructs Gene Regulatory Networks from Single-cell Transcriptomics

Anish Mudide

Under the direction of

Rohit Singh
Research Scientist
MIT CSAIL

Alexander Wu
PhD Student
MIT CSAIL

Research Science Institute
September 5, 2022

## Abstract

Uncovering causal relationships between variables is crucial for developing a global understanding of dynamical systems. Granger causal models extract the underlying causal structure by identifying variables useful in forecasting future values of other variables, thus accounting for the temporal lag between a cause and its effect. While traditional Granger causal methods model linear relationships between time series, recent work has pushed for the capture of nonlinear dynamics and long-range dependencies via sparsity-inducing regularization. However, such models assume that the dynamical system under consideration consists of a totally-ordered sequence of observations. Many real-world dynamical systems, such as cellular differentiation trajectories, possess only a partial ordering due to branching points and thus cannot be rigorously explored under current methods. In this paper, we present LagNet, a novel model architecture that extends regularized, nonlinear Granger causal inference to partially-ordered sequences of observations. We demonstrate LagNet's utility by applying it to reconstruct genetic regulatory networks from single-cell transcriptomic datasets. In a series of benchmark tests, we show that LagNet consistently outperforms established Granger causal models as well as GENIE3, a state-of-the-art regulatory network inference method.



## Summary

To study a system, researchers will often collect data from many observations. A major goal of such work is to understand the cause and effect relationships that exist within the data. Current methods for identifying these causal relationships fail to work when the system is structured as a network, as is the case for social media users and groups of cells. To address this, we developed a novel method that effectively uncovers cause and effect relationships within network-structured systems. The motivating application for our research is to develop a deeper understanding of gene regulation, the process by which special proteins called transcription factors control the production of our genes. Since many diseases arise from errors in gene regulation, further insight into these interactions is critical for developing new therapeutics.





**Keywords:** scRNA-seq, gene regulatory network, causality, deep learning

# Contents

# 1 Introduction

The expression levels of genes within a cell are intrinsically linked by causal relationships, in which a change to the expression of one gene may trigger downstream effects in other genes. Uncovering the topology of gene regulatory networks (GRNs) which describe the complete set of casual relationships is a crucial open problem in biology [1]. While causality remains a deeply philosophical notion, statistical frameworks such as Granger causality have shown promise in detecting ground-truth casual relationships from real-world datasets [2]. Often, the downstream effect of a cause does not occur instantaneously. For instance, within GRNs a transcription factor (TF) must first be translated and bind to the promoter region of a target gene (TG) before its effect on TG expression can be realized (Figure 1a). When a cause precedes its effect, the casual variable is informative for forecasting future values of the effected variable. The framework of Granger causality thus recasts causal relations as predictive relations. In particular, a time series $x$ (where $x_t$ denotes the value at time $t$) is defined to "Granger-cause" a time series $y$ if knowing past values of $x$ helps predict $y$ [3].

Modern GRN inference methods leverage Granger casual models to identify TF-TG pairs from single-cell RNA-sequencing (scRNA-seq) data, relying on the biological intuition that if the expression level of gene $a$ Granger-causes the expression level of gene $b$, we can infer that gene $a$ encodes a TF while gene $b$ is its corresponding TG [4, 5]. Ideally, such methods would operate on datasets that sample the transcriptomic state of singular cells over time (Figure 1b). However, the vast majority of available scRNA-seq datasets originate from protocols which result in cell death, preventing the temporal measure of transcriptional output. scRNA-seq instead provides a snapshot description of a cell population's state at a particular time [6]. The data is organized into an expression matrix, which contains the expression levels of every gene within each cell (Figure 1c). Despite not containing true temporal information, large-scale scRNA-seq studies capture a wide range of cell states, ranging from early progenitor

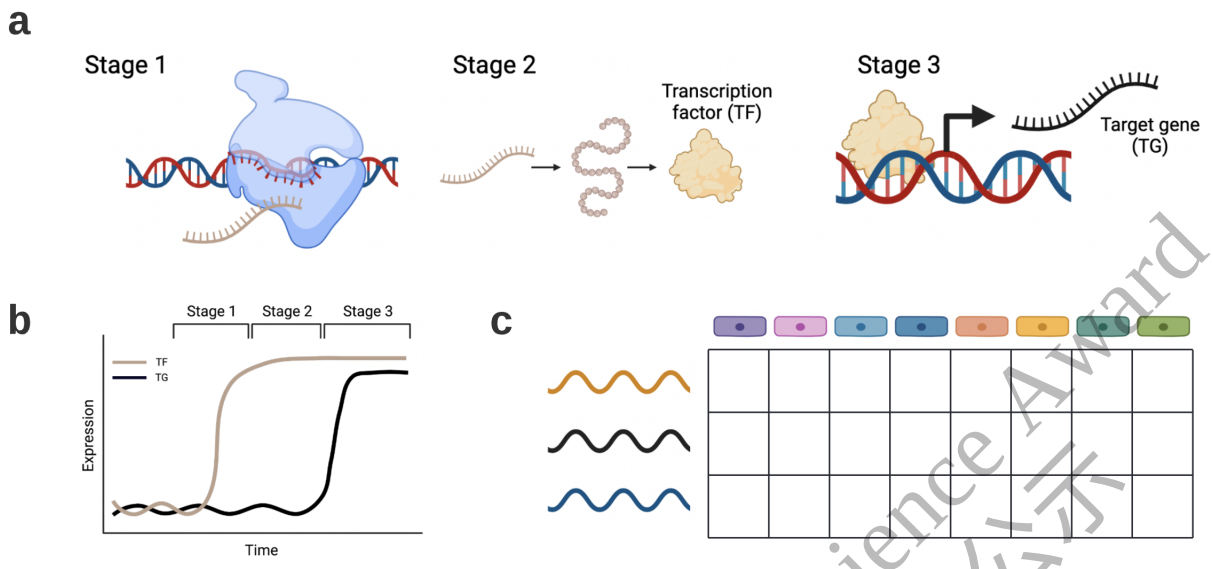


Figure 1: Overview. (a) A temporal lag separates the transcription of a transcription factor and its corresponding target gene. (b) The expression of a TF Granger-causes the expression of a TG. However, we cannot measure the transcriptional output of a cell over time because sequencing destroys the cell. (c) With scRNA-seq, we obtain a snapshot of the cell population's state, which describes the number of transcripts per gene for all cells.

cells to fully differentiated cells. Computational strategies have been developed to order these cells along a linear trajectory, where each cell is associated with an inferred "pseudotime" [7]. Cells earlier in the trajectory are assumed to be the previous states of cells with later pseudotimes. Thus, ordering cells by pseudotime estimates the evolution of a cell's expression levels over time, which enables Granger causal models to identify TF-TG pairs [8].

Complications arise, however, because realistic cells do not evolve in a linear trajectory. Many cellular trajectories contain branching events where a cell may differentiate along one of many lineages, each with its own terminal state [9]. Consequently, the underlying differentiation structure of the cells is more aptly described as a directed acyclic graph (DAG), where an edge from node $i$ to node $j$ indicates that cell $i$ may differentiate into cell $j$. In this paper, we construct this DAG directly from the expression matrix. While ordering cells by pseudotime imposes a total ordering on the cells, the DAG structure imposes only a partial ordering, which is biologically sound as cells in separate lineages are not directly

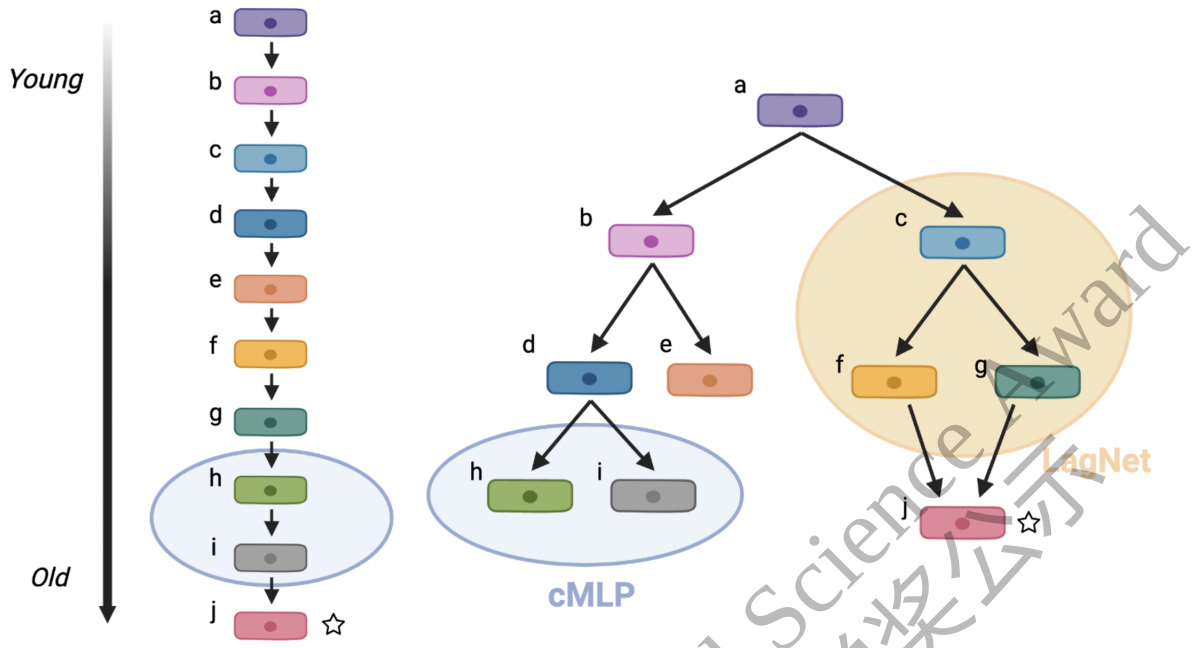


Figure 2: Current GRN inference methods determine TF-TG pairs by applying Granger causality to cells ordered by pseudotime, which disregards the differentiation structure of the cells encoded by the DAG. By taking into account the structure of the DAG, LagNet produces biologically meaningful predictions.

comparable. However, state-of-the-art Granger casual models such as the component-wise multilayer perceptron (cMLP) require that the observations obey a total ordering [10]. Figure 2 depicts a group of 10 cells (labeled $a$ through $j$) ordered by pseudotime on the left, while the true differentiation structure is shown on the right. A totally-ordered Granger causal model (e.g. cMLP) will forecast the gene expression of cell $j$ using the $L$ most recent cells; for $L = 2$, these are cells $h$ and $i$. This is illogical because cells $h$ and $i$ are part of a distinct cellular lineage and are thus far-removed from cell $j$ in the differentiation structure. In short, using totally-ordered models on DAG-structured systems ignores the underlying branching structure, necessitating new model architectures for Granger causal inference on DAGs.

In this paper, we present LagNet, a novel causal inference model that extends state-of-the-art Granger causal inference to dynamical systems structured as DAGs. LagNet's architecture combines graph convolutions that aggregate information from ancestors in the

DAG with fully connected linear layers in order to make forecasts (Figure 2). While LagNet is fully general and can be applied to any DAG-structured dynamical system, in this paper we focus specifically on its ability to reconstruct gene regulatory networks. We apply LagNet, two recently proposed Granger casual models, as well as a random forest model specialized in GRN inference to simulated scRNA-seq data with known ground truth TF-TG pairs. We find that LagNet consistently is the best-performing GRN inference model across four datasets varying in differentiation complexity. Our implementation of LagNet is available at https://github.com/amudide/LagNet.

# 2 Methods

## 2.1 Granger Causal Models

Since its proposal in 1969, Granger causality has proven to be a successful framework for inferring causal relationships within dynamical systems [11, 12]. There are two main approaches we can use to assess the existence of a Granger causal relationship between variables $x$ and $y$: ablation and invariance [3]. The ablation approach involves training two predictive models of $y$, denoted as $u$ and $u_{\bar{x}}$, where $u$ includes the full history of every variable in the system and $u_{\bar{x}}$ excludes the history of $x$. If $u$ performs significantly better than $u_{\bar{x}}$, as determined by a one-tailed F-test, then $x$ Granger-causes $y$. The invariance approach (which we adopt in this paper) involves training just one predictive model of $y$, denoted as $f$, using the full history of every variable. Then, $x$ does *not* Granger-cause $y$ if and only if the learned weights governing the interaction between $x$ and $y$ are all equal to 0. Equivalently, no causal relationship exists exactly when the prediction of $y$ is invariant to perturbations in the history of $x$. This latter approach reduces training time and allows for more direct interpretation.

In the original formulation of Granger causality, every variable is modelled as a linear com-

bination of the system's variable histories. While this traditional casual model successfully captures simple dynamics, real-world datasets often conform to nonlinear and long-range interactions. To identify causal relationships in these settings, recent work has focused on developing new variations of the Granger causal model based on deep learning architectures [10, 13]. Tank et al. [10] introduce a regularized multilayer perceptron (cMLP) and a long short-term memory network (cLSTM) that model nonlinear relationships while simultaneously determining the lag of each putative causal relationship. Marcinkevics and Vogt [13] extend self-explaining neural networks to multivariate time series data in order to determine whether a causal relationship induces a positive or negative effect.

In general, given a dynamical system with $N$ observations and $g$ variables, Granger causal inference involves training $g$ models $f_1, f_2, \ldots, f_g$. $f_j$ models variable $j$ as a function of the previous $L$ observations:

$$x_{tj} = f_j\big(x_{(t-L):(t-1)1}, x_{(t-L):(t-1)2}, \ldots, x_{(t-L):(t-1)g}\big) + e_{tj}.$$

Here, $x_{(t-L):(t-1)k} = (x_{(t-L)k}, \ldots, x_{(t-1)k})$, $x_{tj}$ denotes the value of the variable $j$ at observation $t$ and $e_{tj}$ denotes an error term [10]. Each pair of variables $(i, j)$ has an associated weight matrix $\mathbf{W}_{ij}$ that defines how variable $j$ depends on past lags of variable $i$. Variable $i$ is inferred to Granger-cause $j$ if $|\mathbf{W}_{ij}| \neq 0$, meaning that $f_j$ is not invariant to $x_{(t-L):(t-1)i}$. During training, regularization can be applied to each weight matrix $\mathbf{W}_{ij}$ to assist in achieving exact zeros for the non-causal relationships. In addition, $\mathbf{W}_{ij}^l$ refers to the vector that defines the interaction between $x_{(t-l)i}$ and $x_{tj}$. By applying additional regularization terms on each $\mathbf{W}_{ij}^l$, we can automatically detect the relevant lags of each putative causal relationship [10].

## 2.2 LagNet

In the above formulation, $x_t$ is defined to temporally precede $x_{t+1}$, which means that a total ordering on the $N$ observations is required. Thus, dynamical systems which consist of branching points, such as cellular differentiation trajectories and Twitter retweet networks,

cannot be successfully modelled under current architectures. To address this, LagNet extends the nonlinearity and automatic lag selection of modern Granger causal methods to DAG-structured dynamical systems which possess only a partial ordering over their observations.

LagNet takes in two inputs, $\mathbf{A} \in \mathbb{R}^{N \times N}$ and $\mathbf{X} \in \mathbb{R}^{N \times g}$, and produces one output $\mathbf{GC} \in \mathbb{R}^{g \times g}$. $\mathbf{A}$ is the adjacency matrix of the DAG, $\mathbf{X}$ is the feature matrix which describes the values of the $g$ variables over the $N$ observations, and $\mathbf{GC}$ is the adjacency matrix of the inferred causal graph, where $GC_{ij} = 1$ if variable $i$ Granger-causes variable $j$ and $GC_{ij} = 0$ otherwise. We precompute a modified matrix $\mathbf{A'}$ which is the result after normalizing the sum of each row in $\mathbf{A}^T$ to 1. If a row of $\mathbf{A}^T$ consists of all zeros, the row is unaltered.

To infer causal relationships, we train $g$ separate models $f_1, f_2, \ldots, f_g$. We propose $f_j$ to be a multilayer neural network that models variable $j$ as a nonlinear function of ancestors within the DAG. The key differentiating factor of our model architecture is the first hidden layer, which takes on the form

$$\mathbf{h}^{(1)} = \sigma\Big(\sum_{\ell=1}^{L} (\mathbf{A}^T)^\ell \mathbf{X}\mathbf{W}^{1,\ell} + \mathbf{b}_1\Big).$$

Here, $(\mathbf{A'})^\ell$ represents the $\ell$th power of $\mathbf{A'}$, $\mathbf{W}^{1,\ell}$ is a learned weight matrix and $\mathbf{b_1}$ is a learned bias term. For each $1 \leq l \leq L$, $(\mathbf{A'})^\ell \mathbf{X}\mathbf{W}^{1,\ell}$ aggregates information from ancestors $\ell$ steps backwards in the DAG defined by $\mathbf{A}$ (Figure 3a). The row-normalization of $\mathbf{A'}$ allows for predictions to be uniform despite varying indegrees. Let $d$ be the number of hidden units per hidden layer. Then, we have $\mathbf{h}^{(1)} \in \mathbb{R}^{N \times d}$, $\mathbf{W}^{1,\ell} \in \mathbb{R}^{g \times d}$ and $\mathbf{b}_1 \in \mathbb{R}^{N \times d}$. Note that, as the $\mathbf{A'}$ and $\mathbf{X}$ matrices are fixed, we can pre-compute $(\mathbf{A'})^\ell \mathbf{X} = \mathbf{A'}\big((\mathbf{A'})^{\ell-1}\mathbf{X}\big)$ inductively for $1 \leq \ell \leq L$. $\sigma$ is a nonlinear activation function.

In a $K$ layer model, the hidden layers $\mathbf{h}^{(k)}$ for $1 < k < K$ are given by

$$\mathbf{h}^{(k)} = \sigma\big(\mathbf{h}^{(k-1)}\mathbf{W}^k + \mathbf{b}_k\big),$$

where $\mathbf{h}^{(k)} \in \mathbb{R}^{N \times d}$, $\mathbf{W}^k \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_k \in \mathbb{R}^{N \times d}$ (Figure 3b). Finally, the output $f_j \in \mathbb{R}^N$ of

the autoregressive model is given by

$$c\big(\mathbf{h}^{(K-1)}\mathbf{W}^K + \mathbf{b}_K\big),$$

where $\mathbf{W}^K \in \mathbb{R}^{d\times 1}$, $\mathbf{b}^K \in \mathbb{R}^{N\times 1}$ and $c$ is an optional element-wise decoder function that maps the real numbers to a domain-specific output. While we limit $c$ to the identity function, future work could explore using $c(x) = e^x$ for gene expression data, which is always non-negative. The only constraint on the weight matrices is that each of the $\mathbf{b}_i$ matrices must have constant columns so that the bias term is uniform across observations.

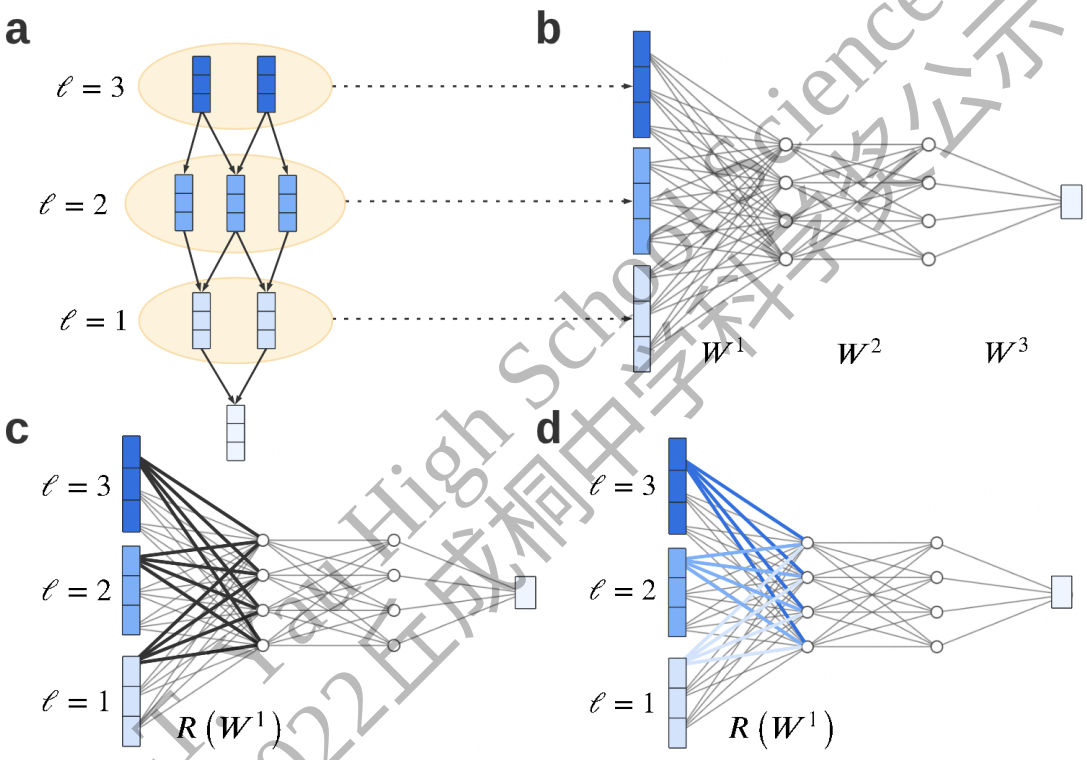


Figure 3: LagNet architecture: $N = 8$, $g = 3$, $L = 3$, $d = 4$ and $K = 3$. (a) LagNet predicts a node's features based on ancestors in the DAG. (b) Predictions are made by passing aggregated information from each lag through a feed-forward neural network. (c) If the weights associated with variable $i$ (highlighted in black for $i = 1$) are all equal to 0, then variable $i$ does *not* Granger-cause variable $j$. (d) The lag-selection and hierarchical penalties further penalize groups of weights associated with particular lags.

To infer causality, we concatenate the $\mathbf{W}^{1,\ell}$ matrices into $\mathbf{W}^1 \in \mathbb{R}^{L \times g \times d}$. Let $\mathbf{W}_i^1 \in \mathbb{R}^{L \times d}$ denote the weights that govern the interaction between variable $i$ and variable $j$ (Figure 3c). Time series $i$ does *not* Granger-cause time series $j$ exactly when $\|\mathbf{W}_i^1\|_F = 0$, where $||\cdot||_F$ denotes the Frobenius matrix norm. Similarly, let $\mathbf{W}_i^{1,l} \in \mathbb{R}^d$ denote the weights that govern the interaction between variable $i$ and variable $j$ at a particular lag $l$. If variable $i$ Granger-causes $j$, then $l$ is not a relevant lag exactly when $\left\|\mathbf{W}_i^{1,l}\right\|_2 = 0$, where $||\cdot||_2$ denotes the $L_2$ norm.

We wish to simultaneously reduce the prediction error of each model $f_j$ while encouraging exact zeros within the $\mathbf{W}_i^1$ matrices in order to induce sparsity in the causal graph. To achieve this, our objective function $J$ is defined to be the sum of the MSE loss and a regularization term applied to $\mathbf{W}^1$, which is weighted by $\lambda$:

$$J = \text{MSE}(f_j, \mathbf{X}_{:j}) + \lambda R(\mathbf{W}^1).$$

Here, $\mathbf{X}_{:j} \in \mathbb{R}^N$ denotes the value of variable $j$ over all $N$ observations.

We implement three distinct regularization functions defined by Tank et al. [10]. The group regularization penalizes all weights within $\mathbf{W}^1$ symmetrically:

$$R(\mathbf{W}^1) = \sum_{i=1}^{g} \left\|\mathbf{W}_i^1\right\|_F.$$

The lag-selection regularization aids in automatic lag detection by merging weights from the same lag together (Figure 3d):

$$R(\mathbf{W}^1) = \sum_{i=1}^{g} \left( \left\|\mathbf{W}_i^1\right\|_F + \sum_{l=1}^{L} \left\|\mathbf{W}_i^{1,l}\right\|_2 \right).$$

Finally, the hierarchical regularization also merges weights from the same lag, but penalizes longer lags more than shorter lags (Figure 3d):

$$R(\mathbf{W}^1) = \sum_{i=1}^{g} \left( \sum_{l=1}^{L} \left\|(\mathbf{W}_i^{1,l}, \ldots, \mathbf{W}_i^{1,L})\right\|_F \right).$$

Traditional gradient descent algorithms such as Adam [14] and stochastic gradient descent often fail to converge the learned weights to exact zeros, hindering causal detection. We thus

optimize the objective function $J$ via proximal gradient descent, a specialized algorithm designed to induce sparsity given a regularized objective function [15]. We train the models $f_1, f_2, \ldots, f_g$ using all $N$ observations. The models are never shown the ground truth causal graph $\mathbf{GC}^*$, thus preventing overfitting from occurring. During evaluation, $\mathbf{GC}^*$ is compared to $\mathbf{GC}$, which is inferred from the learned weight matrices of $f_1, f_2, \ldots, f_g$. Across all LagNet experiments, we set $d = 100$, $K = 2$ and $L = 5$. We use ReLU [16] for our activation function $\sigma$, the hierarchical regularization penalty, and the identity function for $c$.

## 2.3 Comparison with GrID-Net

To the best of our knowledge, the only other Granger causal model that takes into account the DAG structure of a system is GrID-Net, an approach based on graph neural networks [3]. GrID-Net adopts the ablation strategy: for each pair of variables $(i, j)$, GrID-Net will train a reduced model using only variable $j$ and a full model using variables $i$ and $j$. The models are then compared via the in-sample loss; if the full model performs significantly better, $i$ is inferred to Granger-cause $j$. By considering only bivariate relationships, GrID-Net fails to capture the system-wide dynamics modelled by LagNet. Moreover, GrID-Net applies no regularization to model weights, does not quantify the lag of causal relationships and overall, lacks the interpretability provided by LagNet. In this paper, we also demonstrate that GrID-Net fails to consistently produce meaningful inferences in the GRN reconstruction task.

## 2.4 DAG Construction

Given a scRNA-seq dataset with $N$ cells and $g$ genes, we aim to construct a DAG (with adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$) on the cells that preserves the underlying differentiation trajectories. We want $A_{ij} = 1$ if cell $j$ is a logical subsequent cell state of cell $i$, and $A_{ij} = 0$

otherwise.

To build this DAG, we first construct a graph $G$ that connects each cell to its $k$-nearest neighbors based on latent space representations generated from principal component analysis (PCA). Next, we infer a pseudotime value for each cell using the diffusion pseudotime algorithm [7]. Finally, we orient each edge $e \in G$ in the direction of increasing pseudotime. This preserves the underlying differentiation structure while ensuring that the constructed graph is acyclic.

We note that dynamical systems with totally ordered sequences of observations can also be modelled by LagNet. In this case, the adjacency matrix satisfies $A_{ij} = 1$ if $i + 1 = j$ and $A_{ij} = 0$ otherwise.

# 3 Results

We evaluate LagNet on simulated datasets by comparing LagNet's inferred Granger causal relationships to the ground truth causal graph. We present two major sets of results. First, we apply LagNet to multivariate time series data, for which there exists a total ordering on the observations. Despite being designed for partially ordered sequences, LagNet still performs on-par with state-of-the-art Granger casual models that assume a total ordering. Second, we apply LagNet to reconstruct gene regulatory networks from simulated scRNA-seq datasets. We find that LagNet outperforms GrID-Net [3], cMLP [10] and GENIE3 [17] across four benchmark tasks.

## 3.1 Multivariate Time Series

The Lorenz-96 model [18] is used as a standard benchmark task for Granger casual models due to the nonlinear nature of the underlying causal relationships. The simulated dynamical system consists of $g$ variables, where variable $i$ is Granger-caused by variables $i - 2$, $i - 1$

and $i+1$, and variable indices outside of $[0, g-1]$ are taken modulo $g$. Each variable obeys the differential equation

$$\frac{dx_{ti}}{dt} = \big(x_{t(i+1)} - x_{t(i-2)}\big)x_{t(i-1)} - x_{ti} + F,$$

where higher values of $F$ result in greater nonlinearity [10]. Figure 4 illustrates examples of simulated time series.

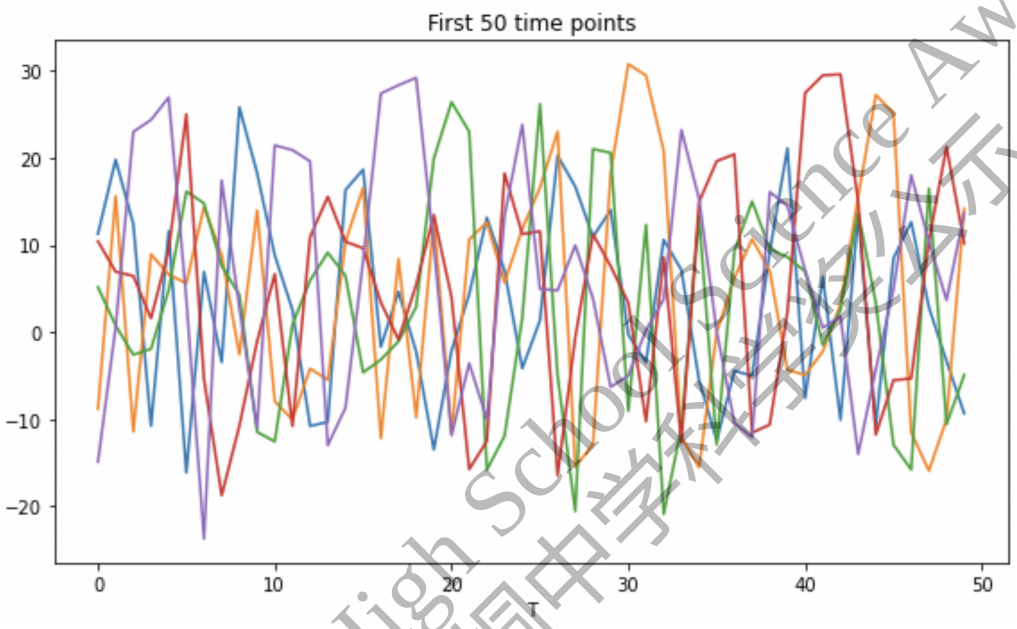


Figure 4: First 50 data points of 5 simulated Lorenz-96 time series (F = 40).

We run LagNet on 10 simulated time series, each with 1000 observations. We compare the inferred Granger casual graph with the ground truth interactions defined by the differential equation. Figure 5 visually compares the adjacency matrices of the ground truth and predicted casual graphs, where the relative intensity of each entry $(i, j)$ in the predicted matrix corresponds to the Frobenius norm of the learned weight matrix governing the interaction between variables $i$ and $j$. We find that LagNet perfectly reconstructs the relative intensities of each entry. However, LagNet faces difficulty in achieving exact zeros for non-casual relationships, despite our optimization of the objective function via proximal gradient descent. It is thus more practical to infer a Granger casual relationship between two variables if the

interaction intensity is above a certain threshold $s$.

Marcinkevičs and Vogt [13] rigorously benchmark six different Granger casual models on simulated Lorenz-96 data ($F = 40$, $N = 500$ observations, $g = 20$ time series). We apply LagNet to the benchmark dataset to compare its performance with current top-performing models (Table 1). While cMLP performs the best in terms of both AUROC and AUPRC, LagNet performs on par, placing second in terms of AUROC and third in terms of AUPRC. Thus, in generalizing to DAGs, LagNet does not compromise its ability to infer casual relationships within totally-ordered datasets.

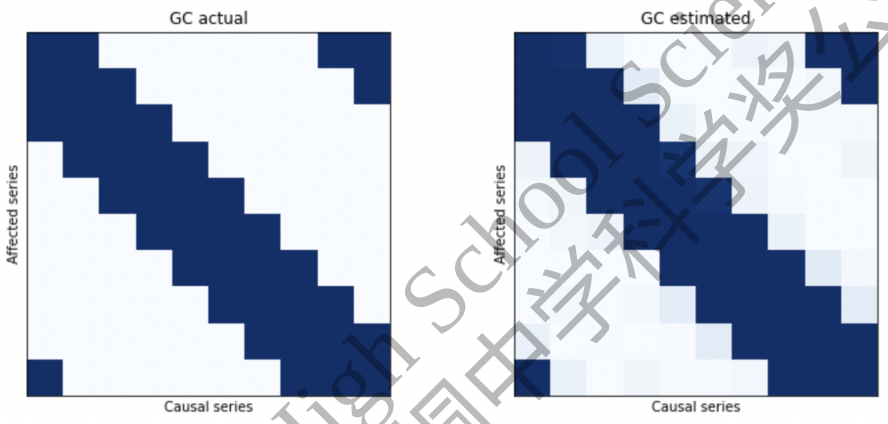


Figure 5: Lorenz-96 ground truth versus predicted casual graphs, visualized as adjacency matrices.

| Model | VAR | cMLP | cLSTM | TCDF | eSRU | GVAR | LagNet (ours) |
|---|---|---|---|---|---|---|---|
| AUROC (± SD) | 0.745 (± 0.047) | **0.979 (± 0.016)** | 0.661 (± 0.038) | 0.679 (± 0.021) | 0.934 (± 0.021) | 0.970 (± 0.009) | 0.975 (± 0.015) |
| AUPRC (± SD) | 0.474 (± 0.036) | **0.956 (± 0.033)** | 0.385 (± 0.063) | 0.314 (± 0.050) | 0.834 (± 0.033) | 0.916 (± 0.024) | 0.908 (± 0.041) |

Table 1: Comparison of various Granger casual models on simulated Lorenz-96 data. AUROC and AUPRC are calculated by sweeping the threshold value $s$. Data shown is the mean and standard deviation values across 5 replicates. [13]

While LagNet is designed for nonlinear interactions, we show that LagNet is robust

to linear dynamics as well. Using vector autoregression (VAR), we simulate 10 time series across 1000 observations. Figure 6 depicts how VAR time series evolve over time. Applying LagNet to the simulated data leads to an almost perfect (98% accuracy) reconstruction of the underlying causal graph (Figure 7). Once again, we observe that LagNet struggles to converge to exact zeros for certain non-causal interactions.

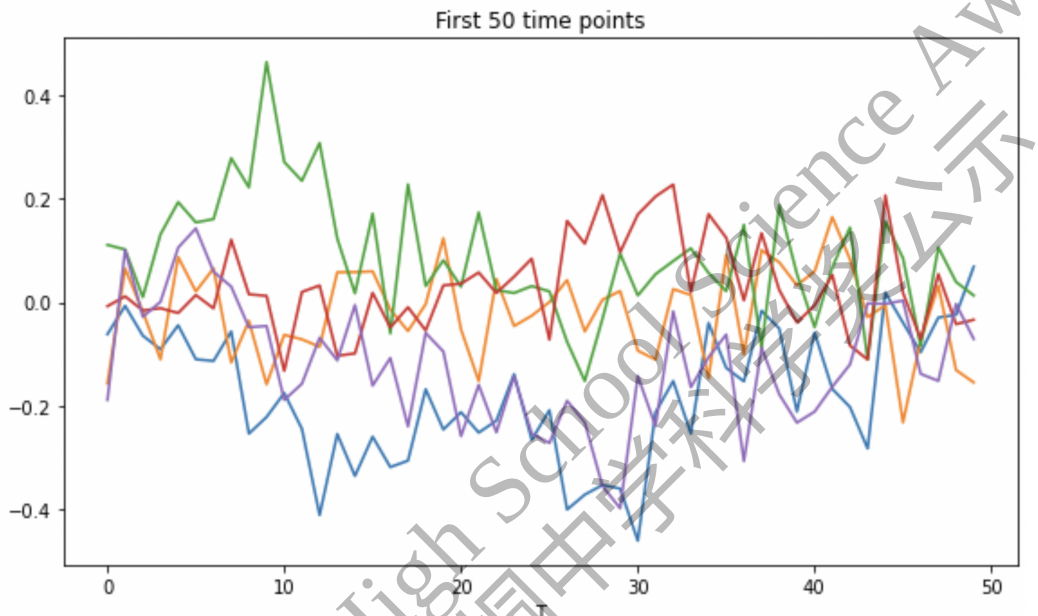


Figure 6: First 50 data points of 5 simulated VAR time series.

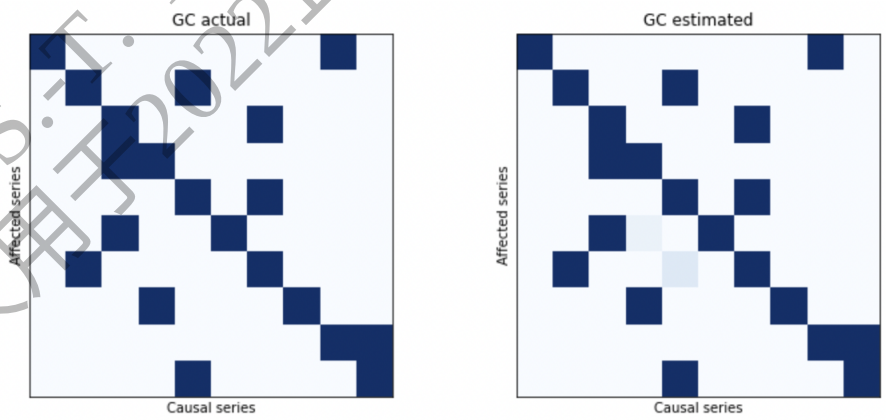


Figure 7: VAR ground truth versus predicted casual graphs, visualized as adjacency matrices.

## 3.2 Gene Regulatory Network Reconstruction

SERGIO [19] is a computational tool that simulates scRNA-seq datasets which reflect the dynamics of a user-defined GRN. Datasets generated by SERGIO are realistic and allow for the rapid assessment of GRN inference tools since the underlying GRN is known. We assess LagNet's ability to reconstruct gene regulatory networks by applying it to four benchmark datasets. We additionally evaluate the performance of cMLP [10], GrID-Net [3] and GENIE3 [17] on the same datasets in order to provide a comparison. We run GrID-Net and GENIE3 using the default parameters. To apply cMLP, we first order the cells by pseudotime, as determined by the diffusion pseudotime algorithm [7].

The four datasets each contain 100 genes and 300 cells per cell type. The underlying GRN used to simulate the data is a network found in *E. coli* consisting of 10 TFs and 137 TF-TG pairs [19]. Dataset 1 contains three cell types arranged in a linear trajectory, dataset 2 has four cell types with a bifurcation event, dataset 3 has six cell types with a trifurcation event, and dataset 4 has seven cell types with both bifurcation and trifurcation events (Figure 8).

We evaluate the predicted causal graphs of LagNet, GrID-Net, cMLP and GENIE3 on the four benchmark datasets in terms of AUROC and AUPRC. The baseline AUPRC for a random model is 0.0137. Across all four datasets, we find that LagNet is the top performer as measured by both AUROC and AUPRC (Figure 8). GENIE3 consistently outperforms the two other Granger causal models. While cMLP displayed unsurpassed performance on the totally-ordered Lorenz-96 dataset, it fails to compete with LagNet in this DAG-structured system. Despite taking the DAG structure into account, GrID-Net performs poorly, especially on datasets 2 and 4. The success of LagNet on these datasets demonstrates the power of incorporating domain knowledge and regularization when building interpretable models.

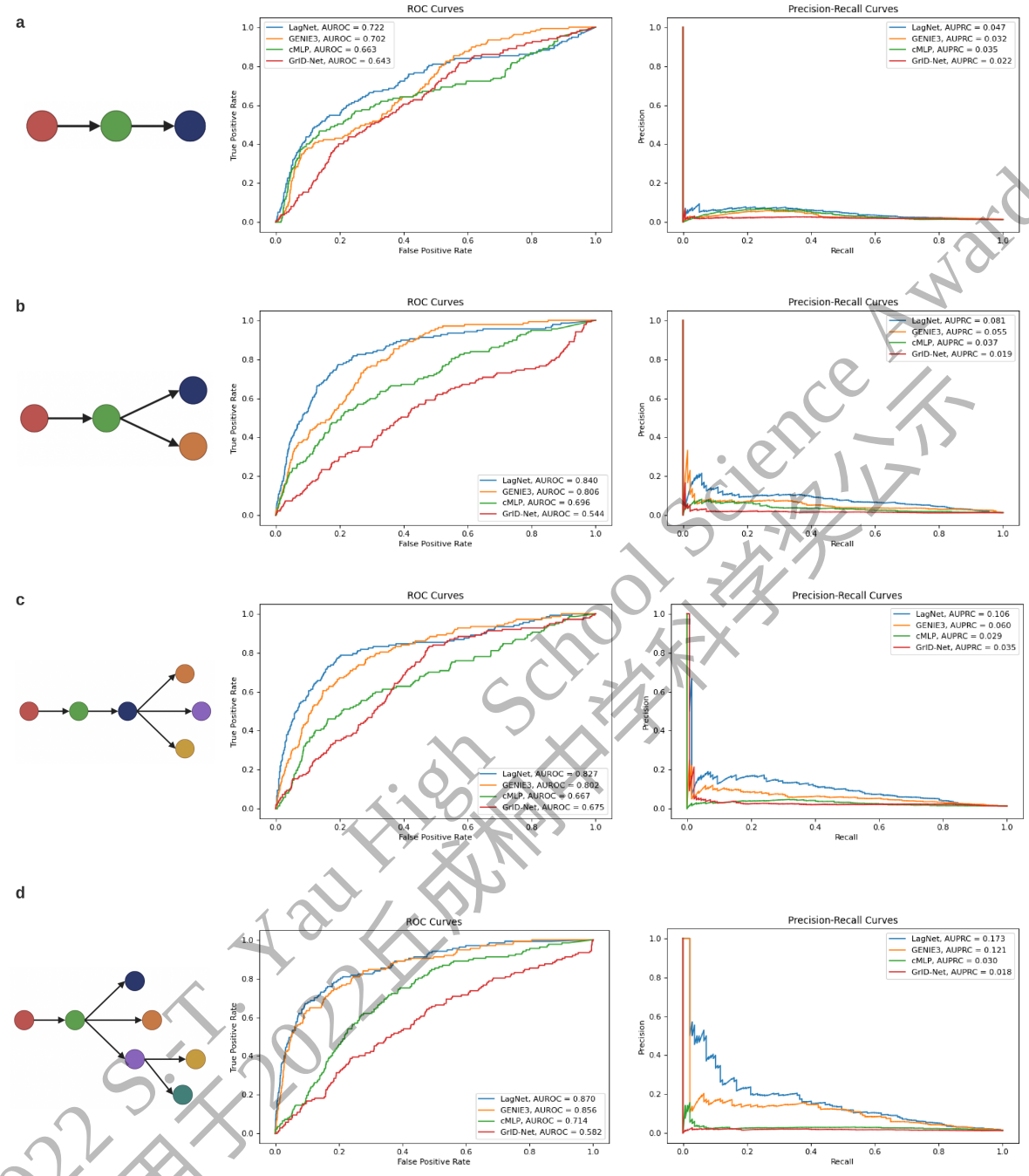




Figure 8: Comparison of Granger causal (LagNet, GrID-Net, cMLP) and random forest models (GENIE3) models on the gene regulatory network inference task. Across four benchmark datasets varying in differentiation complexity, LagNet performs the best in terms of both AUROC and AUPRC.

# 4 Discussion

The modern deep learning toolbox is hyper-focused on data structured as sequences (e.g. audio, text) and grids (e.g. images). On the other hand, the study of neural networks that operate on graphs was first introduced by Scarselli et al. in 2008 and is still in its early stages [20]. This poses a major challenge because a great wealth of pressing problems, such as drug design and protein folding, can be formulated using graphs [21]. In fact, any sequence or grid can be re-casted as a graph, meaning that further research into graph-based machine learning is a step in the right direction towards general intelligence. In this paper, we make meaningful contributions to the study of graphs by generalizing state-of-the-art causality inference to directed acyclic graphs.

In addition, the vast majority of machine learning architectures are optimized for prediction, whether that be in the form of classification, regression or forecasting. Such models often involve complex sets of dependencies and vast sets of parameters that yield them difficult to interpret. Instead of optimizing predictive power, we focus on an interpretable architecture that reveals the underlying structure within data in the form of causal relationships. We achieve this through combining a biologically-informed representation with harsh penalties that force the model to extract only the relevant information. Incorporating regularization additionally allows our method to generalize well to high-dimensional settings.

We observe that LagNet more accurately predicts the ground truth regulatory network when given larger amounts of data. While the maximum number of cells in our benchmark datasets was roughly 2000, modern scRNA-seq pipelines easily profile tens of thousands of cells [22]. These datasets often contain many cell types and complex differentiation patterns. We project that LagNet will continue to excel over other methods in these real-world conditions. Additionally, further iterations of LagNet have the potential to strengthen GRN reconstruction. We notice that LagNet often fails to converge non-causal weight matrices to

exact zeros, even within linear settings. Thus, GRN inference requires selecting a threshold value $s$, which in it of itself is a difficult task. Marcinkevičs and Vogt [13] resolve this issue in the context of totally-ordered Granger casual models by using a stability-based procedure. In their method, two casual graphs are inferred: $\mathbf{GC}_1$ from the unaltered dataset and $\mathbf{GC}_2$ from the time-reversed dataset. A perfect model would yield $\mathbf{GC}_1^T = \mathbf{GC}_2$. Thus, the best threshold corresponds to the value of $s$ that results in the best agreement between $\mathbf{GC}_1^T$ and $\mathbf{GC}_2$. We propose to extend this technique to DAGs, where the time-reversed dataset is constructed by reversing all the edges of the DAG.

LagNet distinguishes itself from GrID-Net through its ability to determine the relevant lags of an inferred Granger causal relation. When applied to single-cell gene expression data, LagNet quantifies the lag between the transcription of a transcription factor and its target gene, providing deeper mechanistic insight into gene regulation. This is especially relevant for the elucidation of gene regulatory cascades, in which gene $i$ regulates gene $i+1$ for $1 \leq i < g$. Future work could investigate the detected lags for interactions between genes part of the same regulatory cascade.

Furthermore, LagNet has potential applications in the modelling of *in silico* perturbations. Once we infer the causal graph $\mathbf{GC}$, we can train a sparsified model of each gene using only the inferred casual relationships. In this scenario, GRN inference acts analogously to feature selection. After training is complete, we can perform knockout experiments where we zero out a gene in a cell early in the trajectory and use our model to predict what the downstream effects would be. Being able to conduct genetic perturbations *in silico* opens the door to unprecedented large-scale experiments that could reveal undiscovered gene functions.

# 5 Conclusion

In this paper, we generalize Granger causal models to systems structured as DAGs while retaining interpretability. Our method, LagNet, successfully evaluates putative causal relationships by examining the learned weight matrices of trained neural networks. LagNet incorporates sparsity-inducing regularization which aids in removing noise and selecting only genuine relationships. We demonstrate that LagNet performs on par with current methods in totally-ordered settings, and consistently outperforms in partially-ordered settings. Our major result is that LagNet can be applied to successfully resolve the topology of gene regulatory networks.

Our work sets the stage for a range of follow-up research, some of which is theoretical and some of which is experimental. Future theoretical work includes designing the automatic selection of the threshold value $s$ used to infer causality, as well as creating novel optimization methods that better converge weights to exact zeros. Promising directions for new experimental results involve applying LagNet to recover gene regulatory cascades or to perform perturbation experiments *in silico*.

While we focus specifically on applying LagNet to single-cell expression data, LagNet is fully generalizable to all DAG-structured systems. Directed acyclic graphs occur naturally in a wide variety of real-world settings, including citation networks, Twitter retweet networks and evolution. Future applications of LagNet could reveal insights into the spread of misinformation and genomic mutations.



# 6 Acknowledgments

I am grateful for the continuous support I have received from my mentors Dr. Rohit Singh and Alexander Wu of MIT CSAIL. I also thank Samuel Sledzieski, a PhD student in the Computation and Biology group at MIT CSAIL. I would also like to thank Professor

Bonnie Berger, the head of the Computation and Biology group at MIT CSAIL. I also thank my RSI tutor Barış Ekim for insightful feedback and discussions throughout the research process. I would also like to thank Alec Dewulf, Franklyn Wang and Viney Kumar for providing comments on previous drafts of this paper. I also thank Ms. Sharon Vidal, Mr. Wilco Groenhuysen, Dr. Leonard S. Schleifer, Dr. and Mrs. Shiby Thomas, Mr. Munir Javeri and Mr. Wes S. Beebee, Jr., as well as Illumina, Novocure and Regeneron Pharmaceuticals for sponsoring my participation in the Research Science Institute (RSI). Finally, I would like to thank the Massachusetts Institute of Technology and the Center for Excellence in Education for giving me the opportunity to conduct research at RSI.

# References


[1] D. Seçilmiş, T. Hillerton, D. Morgan, A. Tjärnberg, S. Nelander, T. E. Nordling, and E. L. Sonnhammer. Uncovering cancer gene regulation by accurate regulatory network inference from uninformative data. *NPJ systems biology and applications*, 6(1):1–8, 2020.

[2] S. L. Bressler and A. K. Seth. Wiener–granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, 2011.

[3] A. P. Wu, R. Singh, and B. Berger. Granger causal inference on dags identifies genomic loci regulating transcription. In *International Conference on Learning Representations*, 2021.

[4] Y. Zhang, X. Chang, and X. Liu. Inference of gene regulatory networks using pseudo-time series data. *Bioinformatics*, 37(16):2423–2431, 2021.

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

[6] A. Ocone, L. Haghverdi, N. S. Mueller, and F. J. Theis. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12):i89–i96, 2015.

[7] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848, 2016.

[8] A. Deshpande, L.-F. Chu, R. Stewart, and A. Gitter. Network inference with granger causality ensembles on single-cell transcriptomic data. *BioRxiv*, page 534834, 2021.

[9] L. Deconinck, R. Cannoodt, W. Saelens, B. Deplancke, and Y. Saeys. Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology*, 27:100344, 2021.

[10] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. B. Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[11] C. Hiemstra and J. D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

[12] A. K. Seth, A. B. Barrett, and L. Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

[13] R. Marcinkevičs and J. E. Vogt. Interpretable models for granger causality using self-explaining neural networks. *arXiv preprint arXiv:2101.07600*, 2021.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

[16] K. Hara, D. Saito, and H. Shouno. Analysis of function of rectified linear unit used in deep learning. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2015.

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

[18] A. Karimi and M. R. Paul. Extensive chaos in the lorenz-96 model. *Chaos: An interdisciplinary journal of nonlinear science*, 20(4):043105, 2010.

[19] P. Dibaeinia and S. Sinha. Sergio: a single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 11(3):252–271, 2020.

[20] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[21] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[22] J. C. Melms, J. Biermann, H. Huang, Y. Wang, A. Nair, S. Tagore, I. Katsyv, A. F. Rendeiro, A. D. Amin, D. Schapiro, et al. A molecular single-cell lung atlas of lethal covid-19. *Nature*, 595(7865):114–119, 2021.