WENDY: Gene Regulatory Network Inference with Covariance Dynamics

Yue Wang[1,*], Peng Zheng[2,3], Yu-Chen Cheng[4,5,6,7], Zikun Wang[8], and Aleksandr Aravkin[9]

[1]Irving Institute for Cancer Dynamics and Department of Statistics, Columbia University, New York, NY 10027, USA
[2]Institute for Health Metrics and Evaluation, Seattle, WA 98195, USA
[3]Department of Health Metrics Sciences, University of Washington, Seattle, WA 98195, USA
[4]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[6]Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[7]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
[8]Laboratory of Genetics, The Rockefeller University, New York, NY 10065, USA
[9]Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA
[*]yw4241@columbia.edu

Abstract

Determining the structure of gene regulatory networks (GRNs) is a central problem in biology, with a variety of inference methods available for different types of data. However, for a prominent and intricate scenario with single-cell gene expression data collected post-intervention across multiple time points, where joint distributions remain unknown, there is only one known specifically developed method, which does not fully utilize the rich information contained in this data type. In response, we introduce an inference approach tailored to this challenging context: netWork infErence by covariaNce DYnamics, dubbed WENDY. The core idea of WENDY is to model the dynamics of the covariance matrix, and solve this dynamics as an optimization problem to determine the regulatory relationships. To assess its efficacy, we benchmark WENDY against alternative inference methods using synthetic data. Our findings underscore

WENDY's robust performance across diverse synthetic datasets. Moreover, we deploy WENDY to analyze three distinct experimental datasets, uncovering potential gene regulatory mechanisms.

1 Introduction

In general, a gene is transcribed into mRNA and then translated into proteins. This process, known as gene expression, commonly employs mRNA count or protein count to denote the expression level. In addition to directly changing cell phenotypes [54, 14], influencing extracellular processes [5], or even manipulating macroscopic neurological circuitry [38, 60], certain proteins can affect the transcription of other genes (mutual regulation) or their own corresponding genes (autoregulation). Genes and their regulatory relationships form a gene regulatory network (GRN).

Determining the GRN structure is a central problem in biology, as it reveals how a living organism is maintained [4], and provides control of essential biological processes [66], especially treating cancer [45, 13]. However, directly establishing the GRN using traditional technologies is extremely difficult since they cannot measure the expression levels of many genes within the same cell. Instead, numerous methods have been developed to infer the GRN structure from gene expression data. Particularly, recent advancements in single-cell RNA-sequencing technologies have made it possible to profile the whole transcriptome of single cells at large-scale. However, single-cell RNA-seq can only measure one time point because cells have to be killed during the experimental process, making it challenging to study gene regulation relationships that require multiple observations over time.

In this paper, we focus on a specific data type arising from the following setup: First, implement an intervention that affects gene expression (e.g., drugs). Then measure the expression level (generally mRNA count) of n genes for different single cells at multiple time points, and select the data from time points where the expression has not yet reached a stationary state. Since gene expression at the single-cell level is stochastic, for each time point, we obtain many samples of an n-dimensional random vector. However, since we need to kill a cell before measuring its gene expression levels, one cell can only be measured once. Thus, we measure different cells at different time points, and we do not have a joint distribution for gene expression at different time points. Although this approach has become common in recent experimental research [10], and it provides more informative data compared to most other approaches, to our knowledge, there is only one inference method developed specifically for this data type, SINCERITIES [50]. A major limitation of SINCERITIES is that it requires data from at least six time points to perform well. Additionally, for single-cell expression data of n genes over T time points, this method only extracts $n(T-1)$ numbers for further analyses, implying low data utilization efficiency.

There have been many inference methods for single-cell time series gene expression data, where the joint distribution of expression levels at different time points is known [29, 78]. Since obtaining the joint distribution of gene expression is difficult, such methods

are usually not practically applicable. There are also many inference methods for single-cell gene expression data measured at a single time point [8, 74], or bulk level gene expression data measured at multiple time points after interventions [51, 42, 67]. These data types are more common because of their low cost. Nevertheless, they provide less information compared to the data type we examine in this study, namely, time series data from single-cell gene expression. Therefore, while it is feasible to convert our considered data type into these more common forms and use corresponding inference methods, such transformations result in a significant loss of the rich information inherent in the original dataset.

In this paper, we introduce an algorithm named NetWork infErence by covariaNce DYnamics (WENDY), designed to connect single-cell gene expression data at different time points, even in the absence of knowledge about the joint distribution. The core idea behind WENDY is to compute the covariance matrices of gene expression levels at two time points and model the evolution of these covariance matrices over time. To infer the GRN, we formulate a non-convex optimization problem based on the dynamics of covariance matrices and derive a numerical solution. For a visual representation of WENDY's workflow, refer to Figure 1.

One of WENDY's key advantages is its requirement of only two time points worth of data. This feature is particularly valuable in scenarios where intervention and / or measurement ultimately result in cell death, precluding measurements at additional time points. However, if data from more time points are available, WENDY can still be applied to each pair of neighboring time points to detect potential rapid changes in the GRN during the experiment. Furthermore, for single cell expression data comprising n genes across T time points, WENDY extracts $(0.5n^2 + 1.5n)T$ numbers for further analyses, indicating significantly higher data utilization efficiency.

The paper proceeds as follows. In Section 2, we present a classification framework for gene expression data and review existing GRN inference methods. Section 3 details the WENDY method, including the mathematical gene expression model and the approach to solving the dynamics of this model. In Section 4, we evaluate WENDY and other GRN inference methods using synthetic data to compare their performance. Section 5 applies WENDY to real data sets to identify potential gene regulations and observe their evolution over time. Finally, we conclude with discussions in Section 6.

2 Data classification and literature review

2.1 A framework for data classification

There are different types of gene expression data that can be used to infer the GRN structure. Different data types correspond to different inference methods. We first present a framework for classifying related data types, modified from the framework by Wang and Wang [65]. See Table 1 for this classification framework. There are different dimensions to classify data types.

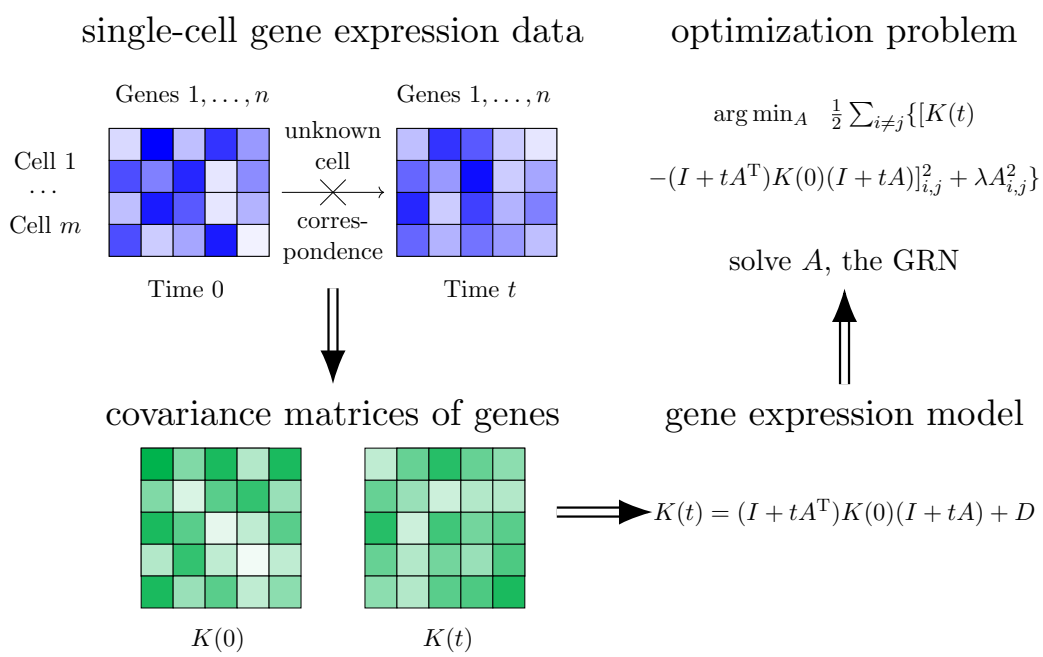

Figure 1: Workflow of the WENDY method. Given single-cell level gene expression data at two time points, where the joint distribution (cell correspondence) between two time points is unknown, first calculate the covariance matrix of gene expression for each time point. Then use the mathematical gene expression model to derive the equation of covariance matrices. Last, transform this into an optimization problem and solve the GRN numerically.

(1) We can measure the gene expression levels when the dynamics of gene expression is stationary (invariant along time), or we can add an intervention to drive the dynamics of gene expression away from stationarity, and measure the gene expression levels when they gradually return to the (possibly new) stationary state. For the intervention, we consider general interventions such as adding drugs (we cannot control which genes are affected) and specific interventions such as gene knockdown and gene knockout (we can select any genes to affect). Considering our capability to measure gene expression levels pre- and post-intervention, a specific intervention yields more informative data than a general one. Moreover, scenarios with intervention are richer in information compared to those without, where only stationary expression levels are observed.

(2) We can measure the average expression levels of many cells (bulk level) or measure the expression levels for each single cell (single-cell level). On single-cell level, gene expression is essentially stochastic, and we shall obtain a random variable for the expression level of each gene. On bulk level, the stochasticity is averaged out, and we should obtain a deterministic value for the expression level of each gene. Single-cell level measurement is more informative than bulk level measurement.

In practice, repeating the same bulk level measurement can still lead to different values, making some researchers regard such data as stochastic and apply inference methods designed for single-cell data [8]. Nevertheless, at bulk level, randomness from single cells is averaged out, and the different values from bulk level measurement can only come from systematic differences, such as different cell phenotypes or different environmental factors. Such unobserved systematic differences can affect multiple genes and make them correlated, although these genes might not have direct regulatory relations. Therefore, we do not consider bulk level data that have different values for the same measurement.

(3) We can measure expression levels at one time point or multiple time points. When we measure expression level at multiple time points, one essential issue is whether we can measure the same cell multiple times. For bulk level data, this does not matter, as the data are deterministic, and whether the cells at $t+1$ are the same as the cells at t should not make a difference. However, for single-cell level data, since the measured levels are stochastic, there is an essential difference. Denote the single-cell expression level of a gene at time t as $X(t)$. If the same cell can be measured multiple times, then we have the joint probability distribution of a time series, $\mathbb{P}[X(0) = x_0, X(1) = x_1, X(2) = x_2, \ldots]$. Otherwise, we only have marginal probability distributions for each time point, $\mathbb{P}[X(0) = x_0]$, $\mathbb{P}[X(1) = x_1]$, $\mathbb{P}[X(2) = x_2]$, ..., but not the correspondence between time points, and certain quantities cannot be calculated, such as the correlation coefficients of expression levels at two time points. Time series data types are more informative than one-time data types, and joint distribution is more informative than marginal distributions.

In practice, measuring the expression levels of many genes is destructive, and we cannot measure the same cell more than once. If we only want to measure the expression level of a single gene, there are some techniques (fluorescent proteins [70], etc.) that can measure the same cell multiple times. Another approach is to measure the amount of spliced and

unspliced mRNAs, which provides both the current expression level and an approximation of its time derivative (RNA velocity [34]). This approach provides two measurements of the same cell, and some inference methods for time series data can be applied.

	One-time		Time series		
	Bulk	Single-cell	Bulk	Single-cell	
				Marginal distribution	Joint distribution
Stationary	1: No	2: Yes	3: No	4: Ditto	5: Yes
General intervention	6: No	7: Ditto	8: Yes	9: Ditto/Yes	10: Ditto
Specific intervention	11: Yes*	12: Ditto	13: Ditto	14: Ditto	15: Ditto

Table 1: Classification of data types regarding GRN inference, modified from the framework by Wang and Wang [65]. The data types are classified by different dimensions: (1) The gene expression is at stationary, or is driven away from stationary by an intervention (on general genes that we cannot choose, or specific genes that we can choose); (2) Measure at one time point or multiple time points (time series); (3) Measure the average over many genes (bulk level) or on single-cell level. When measuring at multiple time points on single-cell level, one more dimension is whether we have the joint distribution over different time points. For different data types (scenarios), we study whether the GRN structure can be inferred. For each data type, No means that the GRN structure cannot be inferred. Ditto means that there are no specific inference methods, but the GRN structure can be inferred using the same method for a less informative data type. Yes means specific inference methods exist. Scenario 11 only has an inference method that also requires the data in Scenario 1. We focus on Scenario 9, which is not as well-studied and has only one specific inference method.

2.2 Known inference methods for different data types

Given this classification framework, we can review inference methods for different data types. In this framework, there are 15 data types (scenarios). Some data types do not have enough information that can be used to infer the GRN structure. Some data types (e.g., Scenario 13) have more information than some other data types (e.g., Scenario 8), but the extra information cannot lead to new GRN inference methods. Thus for such scenarios, we can only use methods for other less informative scenarios. This approach

loses a lot of information, and therefore cannot justify the time and money spent to obtain more informative data. Some data types have extra information that allows for inference methods that work for such scenarios but not for less informative scenarios.

For bulk level data types, since we only have a single deterministic value for each gene, it is difficult to obtain the correlation between genes. For Scenarios 1, 3, 6, the GRN structure cannot be inferred. For Scenarios 8 and 13, we can regard the gene expression time series as solution trajectories of an ordinary differential equation (ODE) system. If we assume that the ODE system is linear [51] or has certain nonlinear forms [42], we can discretize the ODE system into an algebraic equation system and use regression to infer the ODE parameters. Here the ODE parameters represent the GRN. We can infer all the edges, including the directions. For Scenario 11, one can add an intervention on each gene and observe which genes (descendants of this gene in the GRN) are also affected. Such ancestor-descendant relationships can be used to partially infer the GRN structure [65]. Not all regulatory relationships (edges) can be inferred, but for a GRN with n genes, at least $n-1$ edges can be inferred.

For single-cell level one-time data types (Scenarios 2, 7, 12), there have been numerous GRN inference methods for Scenario 2, while Scenarios 7 and 12 generally do not have extra information that supports specific inference methods. For Scenario 2, most inference methods turn the GRN inference problem into a feature selection problem: select genes whose levels can be best used to predict the level of the target gene. Then such selected genes might have regulatory relationships with the target gene. The selection can be made by calculating certain quantities between the target gene and the candidate genes that measure their similarity: covariance [49], mutual information [8], or other information theory quantities [74]. Besides, one can apply regression [26], decision tree [30], or other machine learning and deep learning [57, 77, 80] methods to directly select out genes that can predict the target gene. Regularization terms (e.g., L_1 [25] and L_2 [32] regularizers) can be added to the regression to make the result sparse. Besides the idea of feature selection, another approach is to build probabilistic models (Bayesian network [2, 37], stochastic differential equation (SDE) [63], and others [41, 9]), and use likelihood to determine the most probable network. Since the number of candidate networks is large, a common solution is to apply Markov chain Monte Carlo (MCMC) to approximate the network probabilities [46, 2, 1, 24, 79]. Deterministic models, such as Boolean networks [39], can also be used. There is a well-developed platform that combines different inference methods for Scenario 2 [69]. One problem of Scenario 2 is to determine the direction of a regulatory relationship, since if the level of V_i can predict the level of V_j, then generally the level of V_j can also predict the level of V_i. To solve this problem, one can add specific interventions (Scenario 12) on V_i to see whether V_j is affected.

For single-cell level time series data types without joint distribution of different time points (Scenarios 4, 9, 14), it is common to treat Scenario 4 as Scenario 2, and treat Scenarios 9, 14 as Scenario 8, and apply corresponding inference methods. The only specific GRN inference method we know for Scenario 9 is SINCERITIES [50], which considers the

Kolmogorov–Smirnov distance between the distributions of the same gene at two time points, and then applies linear regression, similar to methods for Scenario 8.

For single-cell level time series data types with joint distribution of different time points (Scenarios 5, 10, 15), there have been numerous GRN inference methods for Scenario 5, while Scenarios 10 and 15 generally do not have extra information that support specific inference methods. For Scenario 5, most inference methods are similar to those for Scenario 2, especially those methods on regression [75], tree-based feature selection [29, 78], or more advanced machine learning tools [47, 3]. The difference is that the direction of a regulatory relationship can be determined in Scenario 5: if the level of V_i at time t can predict the level of V_j at time $t + \Delta t$, then one knows that V_i regulates V_j, not the inverse. There are also inference methods based on more complicated biological models [31].

Besides treating a more informative data type as a less informative data type, one can also use certain methods to transform a less informative data type into a more informative data type, provided there are certain assumptions about the underlying systems. For instance, from single-cell one-time data (Scenarios 2, 7, 12), one can construct the pseudotime to transform the data into time series data [55, 58], and apply corresponding inference methods [20].

Different methods need different assumptions regarding gene expression and gene regulation. For instance, some methods need the gene expression dynamics to be linear [51], and some other methods need the GRN to have no directed cycle [74]. Besides, different methods have different inference abilities: some methods can determine all edges, including the direction [29], while some other methods can only partially determine some edges, and/or cannot determine the edge direction [65].

Most GRN inference methods can only determine regulations between genes, but not autoregulation. Autoregulation inference methods generally need stronger model assumptions [71], more informative data types [22], or only produce partial results [64].

The above discussion only considers the situation of inferring GRN after obtaining all data. Another situation is to design intervention experiments, so that the GRN can be inferred with the minimal cost [15].

Readers may refer to other reviews for more details about GRN inference methods for different scenarios [7, 53, 76, 48, 65] or for other data types (besides mRNA/protein count) that can help with GRN inference, such as ChIP-seq and ATAC-seq [21, 6].

3 Novel GRN inference method

In this section, we present an algorithm of netWork infErence by covariaNce DYnamics (WENDY), that works for Scenario 9, single-cell level time series data without joint distribution of different time points, measured after general interventions. It can determine all regulatory edges including their directions, but not autoregulation. Using extra information such as DNA sequence and transcription factor binding motifs, some regulatory edges

can be excluded. Such prior knowledge can be incorporated by WENDY, but we first consider the original problem that any regulatory edge is possible (except autoregulation).

3.1 Dynamical model of covariance

After adding drugs or other interventions, we measure the expression levels of n genes at time 0, and the expression levels are n random variables: $X(0) = [X_1(0), \ldots, X_n(0)]$. Then at time t, we measure the expression levels of these n genes again to get n random variables: $X(t) = [X_1(t), \ldots, X_n(t)]$. Since we add interventions to drive the system away from stationary, $X(0)$ and $X(t)$ are not identically distributed.

We take expectations for $X(0)$ and $X(t)$ to obtain deterministic expression levels $x(0) = [x_1(0), \ldots, x_n(0)]$ and $x(t) = [x_1(t), \ldots, x_n(t)]$. For simplicity, we can assume that $x(\tau)$ satisfies a linear ODE system:

$$\mathrm{d}x(\tau)/\mathrm{d}\tau = x(\tau)A + b$$

Here A is an invertible $n \times n$ matrix, representing the GRN we want: $A_{i,j} > 0$ means gene i activates gene j; $A_{i,j} < 0$ means gene i inhibits gene j; and $A_{i,j} = 0$ means gene i does not regulate gene j directly. However, $A_{i,i}$ is the combination of degradation and possibly autoregulation of gene i, and $A_{i,i} \neq 0$ does not necessarily mean autoregulation of gene i. Besides, b is a $1 \times n$ vector, representing the base synthesis rate. Its solution is

$$x(\tau) = [x(0) + bA^{-1}]e^{\tau A} - bA^{-1}.$$

Since e^{tA} is difficult to handle in the following optimization procedure, we consider the first-order approximation

$$x(t) = x(0)(I + tA) + tb, \tag{1}$$

where I is the $n \times n$ identity matrix.

Inspired by this deterministic relation, we can write down a similar trajectory/cell-wise relation for the random variables:

$$X(t) = X(0)(I + tA) + tb + X(0) \odot \epsilon(t), \tag{2}$$

where $\epsilon(t) = [\epsilon_1(t), \ldots, \epsilon_n(t)]$ is an n-dimensional normal random noise with mean $(0, \ldots, 0)$ and covariance matrix G, and $\odot$ is the entrywise (Hadamard) product: $X(0) \odot \epsilon(t) = [X_1(0)\epsilon_1(t), \ldots, X_n(0)\epsilon_n(t)]$. Here we assume that G is diagonal, meaning that noise terms $\epsilon_i(t), \epsilon_j(t)$ for different genes are independent, but the diagonal elements of G are unknown. Besides, $\epsilon(t)$ and $X(0)$ are also independent. However, we do not know the joint distribution of $X(0)$ and $X(t)$, meaning that we do not know which sample of $X(0)$ corresponds to which sample of $X(t)$, and this trajectory/cell-wise relation does not help. Instead, we can assume that some statistical quantities on both sides should be equal.

If we take expectation on Eq. 2, it returns to Eq. 1. For any A, we can find the value of b to make Eq. 1 hold. Thus we cannot solve A from Eq. 1. Instead, we can consider the covariance matrices of $X(0)$ and $X(t)$, denoted as $K(0)$ and $K(t)$. We have

$$\begin{aligned}
K(t) =& \mathbb{E}\{[X(t)^{\mathrm{T}} - x(t)^{\mathrm{T}}][X(t) - x(t)]\} \\
=& \mathbb{E}\{(I + tA^{\mathrm{T}})[X(0)^{\mathrm{T}} - x(0)^{\mathrm{T}}][X(0) - x(0)](I + tA)\} \\
&+ \mathbb{E}\{(I + tA^{\mathrm{T}})[X(0)^{\mathrm{T}} - x(0)^{\mathrm{T}}][X(0) \odot \epsilon(t)]\} \\
&+ \mathbb{E}\{[X(0) \odot \epsilon(t)]^{\mathrm{T}}[X(0) - x(0)](I + tA)\} \\
&+ \mathbb{E}\{[X(0) \odot \epsilon(t)]^{\mathrm{T}}[X(0) \odot \epsilon(t)]\} \\
=& (I + tA^{\mathrm{T}})\mathbb{E}\{[X(0)^{\mathrm{T}} - x(0)^{\mathrm{T}}][X(0) - x(0)]\}(I + tA) \\
&+ \mathbb{E}\{(I + tA^{\mathrm{T}})[X(0)^{\mathrm{T}} - x(0)^{\mathrm{T}}][X(0) \odot \mathbb{E}\epsilon(t)]\} \\
&+ \mathbb{E}\{[X(0) \odot \mathbb{E}\epsilon(t)]^{\mathrm{T}}[X(0) - x(0)](I + tA)\} \\
&+ \mathbb{E}[X(0)^{\mathrm{T}}X(0)] \odot \mathbb{E}[\epsilon(t)^{\mathrm{T}}\epsilon(t)] \\
=& (I + tA^{\mathrm{T}})K(0)(I + tA) + D.
\end{aligned} \tag{3}$$

Here D is diagonal, with $D_{i,i} = \mathbb{E}[X_i(0)^2]G_{i,i}$.

Given $K(0)$ and $K(t)$, we cannot solve A directly from Eq. 3, even if we set $D = 0$. Assume $K(0)$ and $K(t)$ are invertible. As covariance matrices, they are positive-definite and have Cholesky decomposition $K(0) = L_0^{\mathrm{T}}L_0$ and $K(t) = L_1^{\mathrm{T}}L_1$ with upper-triangular L_0 and L_1. Then for any orthonormal matrix O with $O^{\mathrm{T}}O = I$, $A = (L_0^{-1}OL_1 - I)/t$ is a solution of Eq. 3 with $D = 0$. Thus Eq. 3 has infinitely many solutions in this case, and we need to add some conditions to obtain a unique A.

3.2 Optimization formulation for covariance dynamics

Assume we measure the expression levels of n genes for m single cells. When $m < n$, which is common in reality, if we directly calculate the covariance matrix, it will always be degenerate (non-invertible). Therefore, we need to apply a specific method, called graphical lasso, that estimates the covariance matrix K in this case, where K is invertible, and the inverse of K is sparse [23].

For a $1 \times n$-dimensional random vector $N = [N_1, \ldots, N_n]$ with invertible covariance matrix K, there is a result that $K_{i,j}^{-1} = 0$ if and only if the partial Pearson correlation coefficient satisfies [68]

$$\rho_{i,j|1,\ldots,i-1,i+1,\ldots,j-1,j+1,\ldots,n} = 0.$$

Therefore, if N is multivariate normal, then $K_{i,j}^{-1} = 0$ if and only if N_i and N_j are independent conditioned on other variables.

By assuming the expression levels of n genes satisfy a multivariate normal distribution, there is a GRN inference method [49]: gene i and gene j have a direct regulatory relation

(direction unknown) if and only if $K^{-1}_{i,j} \neq 0$. Nevertheless, this method assumes that the gene expression is in stationary.

For the data set we consider, the intervention might change the dynamics, and the gene expression is not in stationary. Therefore, $K^{-1}_{i,j} \neq 0$ might just mean that gene i and gene j has a direct regulatory relation before the intervention, not necessarily implying $A_{i,j} \neq 0$. However, the inverse should be true: if $K(0)^{-1}_{i,j} = 0$ and $K(t)^{-1}_{i,j} = 0$, then gene i and gene j should have no direct regulatory relation, whether before intervention or after intervention. Define $\mathcal{C} = \{(i,j) \mid K(0)^{-1}_{i,j} = 0 \text{ and } K(t)^{-1}_{i,j} = 0\}$. Then $A_{i,j} = 0$ if $(i,j) \in \mathcal{C}$. Since $K(0)^{-1}$ and $K(t)^{-1}$ are symmetric, $\mathcal{C}$ is also symmetric: $(i,j) \in \mathcal{C}$ implies $(j,i) \in \mathcal{C}$. Besides, since $K(0)^{-1}$ and $K(t)^{-1}$ are sparse, $\mathcal{C}$ contains most edges.

Certain data, such as DNA sequence and transcription factor binding motifs and ATAC-seq data, can provide prior knowledge that gene i cannot regulate gene j, meaning that $A_{i,j} = 0$. We denote the set of such forbidden edges as $\mathcal{F}$. Notice that $\mathcal{F}$ might not be symmetric.

From the data, after estimating the covariance matrices, we obtain invertible covariance matrices $K(0)$ and $K(t)$, where $K(0)^{-1}$ and $K(t)^{-1}$ are sparse. Now we have

$$K(t) - (I + tA^{\mathrm{T}})K(0)(I + tA) = D. \tag{4}$$

Since the diagonal matrix D is unknown, we only want to match off-diagonal elements of $K(t)$ and $(I + tA^{\mathrm{T}})K(0)(I + tA)$. Therefore, we need to solve A from Eq. 4 regardless of diagonal elements, so that $(I + tA)_{i,j} = 0$ whenever $(i,j) \in \mathcal{C} \cup \mathcal{F}$. Under this restriction, there might not be a solution. Instead, we can minimize the matching error and turn it into an optimization problem:

$$\min_A f_\lambda(A) := \frac{1}{2} \sum_{i \neq j} \{[K(t) - (I + tA^{\mathrm{T}})K(0)(I + tA)]_{i,j}\}^2 + \lambda A^2_{i,j}, \tag{5}$$

where $A_{i,j} = 0$ whenever $(i,j) \in \mathcal{C} \cup \mathcal{F}$ or $i = j$, while $\lambda \geq 0$ is a predetermined constant. The constraints are handled by only optimizing over the nonzero edges, and thus (5) simplifies to a (possibly) regularized nonlinear least squares problem. We set $\lambda = 0$ in the following simulations, but allow users to adjust λ if necessary.

Since $\mathcal{C} \cup \mathcal{F}$ contains most edges, the final A is sparse, which is biologically favorable, since we do not want a very dense GRN. If calculated $A_{i,j} > 0$ or $A_{i,j} < 0$ $(i \neq j)$, we claim that gene i regulates (activates/inhibits) gene j. The diagonal elements of A do not provide information about gene regulation, as we cannot distinguish between autoregulation and normal gene expression.

We use the BFGS algorithm to minimize (5). The overall algorithm is presented in Algorithm 1. For the WENDY method, step 4 of Algorithm 1 is much faster than step 2, since the solver terminates after a constant number of iterations. Graphical lasso has time complexity $\mathcal{O}(n^3)$ [23], which does not depend on the cell number m. Therefore, the

overall time complexity of WENDY method is $\mathcal{O}(n^3)$. In Section 4, we will see that in practice, the time cost of WENDY increases with n, but not m.

We present the workflow of WENDY in Algorithm 1. See

`https://github.com/zhengp0/genet`

and

`https://github.com/YueWangMathbio/WENDY`

for the Python implementation of WENDY.

Algorithm 1: Detailed workflow of WENDY method.

---

1. **Input**:

   Single-cell gene expression data at $T = 0$ (for m_1 cells and n genes) and at $T = t$ (for m_2 cells and the same n genes), both measured after general interventions. Prior knowledge of forbidden edges, $\mathcal{F}$

2. **Call** graphical lasso method to calculate the covariance matrix $K(0)$ for expression data at $T = 0$, so that $K(0)$ is invertible, and $K(0)^{-1}$ is sparse. Also calculate the covariance matrix $K(t)$ for expression data at $T = t$

3. **Construct** $\mathcal{C} = \{(i, j) \mid K(0)^{-1}_{i,j} = 0 \text{ and } K(t)^{-1}_{i,j} = 0\}$.

4. **Call** BFGS solver for the optimization problem

$$\arg\min_{A} \ \frac{1}{2} \sum_{i \neq j} \{[K(t) - (I + tA^{\mathrm{T}})K(0)(I + tA)]^2_{i,j} + \lambda A^2_{i,j}\}$$

   with constraints $A_{i,j} = 0$, $\forall (i, j) \in \mathcal{C} \cup \mathcal{F}$

5. **Output**:

   The GRN matrix A. $A_{i,j} > 0$ or $A_{i,j} < 0$ means gene i activates/inhibits gene j. However, $A_{i,i} \neq 0$ does not necessarily mean autoregulation of gene i

---

3.3 Theoretical comparison with other methods

To infer GRN structure in Scenario 9, besides WENDY, we have three other options: (i) apply SINCERITIES method; (ii) calculate the average gene expression levels over all cells to transform the data into Scenario 8, and apply corresponding methods; (iii) only consider the data at one time point, which is Scenario 2, and apply corresponding methods.

For an expression level data set with n genes at T time points, SINCERITIES only extracts $n(T-1)$ values, and then applies linear regression. If we transform the data into Scenario 8, we can only obtain nT values before fitting to an ODE system. If we abandon data at other time points to switch to Scenario 2, we lose the temporal information. In comparison, WENDY extracts $(0.5n^2+1.5n)T$ values from such data before proceeding to the next step. Therefore, WENDY can extract more information from data.

SINCERITIES requires at least $T=4$ time points to work, and requires at least $T=6$ time points to perform well. Methods for Scenario 8 need at least $T=3$ time points to work, and also more time points to work well. In comparison, WENDY only needs $T=2$ time points. This can be explained by information theory: the GRN (without autoregulation) has n^2-n independent values. WENDY uses data at $T=2$ time points to extract n^2+3n values; SINCERITIES needs data at $T=n$ time points to extract n^2-n values; methods for Scenario 8 need data at $T=n-1$ time points to extract n^2-n values.

SINCERITIES and methods for Scenario 8 require more time points. This not only increases the cost, but also requires that the GRN does not change among such time points. This might not hold in reality, such as in the early development of embryos, where the regulatory strength might change rapidly. Also, when studying the effect of drugs on gene expression, the regulatory effect of drugs will gradually disappear. Instead, since WENDY only needs two time points, when there are data from more time points, we can apply WENDY for each pair of adjacent time points, and study how the GRN evolves along time.

The theoretical foundation of some methods for Scenario 2 requires that the gene expression is at stationary. For Scenario 9, after adding the intervention, we need to wait for the gene expression to return to stationary state. However, in some experiments, the intervention (e.g., adding certain drugs) can drive the gene expression too far away from stationary, which leads to the death of cells before returning to stationary. Therefore, WENDY is a better solution to this transient scenario.

4 Performance evaluation on synthetic data

In this section, we use synthetic data sets to test the performance of WENDY and four other GRN inference methods that work for different data types: GENIE3, dynGENIE3, NonlinearODEs, SINCERITIES.

4.1 Methods and measurements

GENIE3 [30] works for Scenario 2, single-cell level one-time expression data at stationary. The idea is to infer the level of one gene by the levels of other genes using random forest or extra trees. All edges can be inferred, but sometimes the direction is hard to determine.

dynGENIE3 [29] is the revised version of GENIE3 that works for time series data, such as Scenario 5 (and 10), single-cell level time series expression data at stationary

or after general interventions. The idea is basically the same as GENIE3, but here it infers the level of one gene at a later time by the levels of other genes at an earlier time. Therefore, all edges including the directions can be determined. dynGENIE3 requires the correspondence of cells at different time points. This means that it cannot be applied to Scenario 9 data directly. Therefore, dynGENIE3 is included just for completeness, not as a main comparison target.

NonlinearODEs [42] works for Scenario 8, bulk level time series expression data measured after general interventions. The idea is to fit the data to a nonlinear ODE model. It uses XGBoost to determine the importance score of each edge. There is a tunable parameter α that can be chosen manually or automatically, and we use the *from_data* mode to determine the parameter α automatically. All edges including the directions can be determined.

SINCERITIES [50] works for Scenario 9, single-cell level time series expression data measured after general interventions, where the joint distribution at different time points is unknown. The idea is to calculate the distance between the distributions of the same gene at two time points, and then applies linear regression. All edges including the directions can be determined.

Our WENDY method also works for Scenario 9. All edges including the directions can be determined. For the data sets in this section, there is no extra biological knowledge about regulations. Therefore, the set of forbidden edges $\mathcal{F} = \emptyset$.

Since there are not many real data sets with known gold standard GRN, the common practice is to compare the performance of different GRN inference methods on synthetic data [50, 42, 29], since the gold standard GRN is known. The data are generally generated by numerically simulating an SDE system. We will test different inference methods on two data sets: DREAM4 and SINC.

Each inference method obtains a calculated GRN matrix A', whose entries take values in $\mathbb{R}$. To generate the synthetic data, there is a true GRN matrix A, which can take values $1/0/-1$. To evaluate the inference result A' with the true GRN matrix A, we calculate two measurements, AUROC and AUPR [50]. AUROC is the area under the curve of true positive rate versus false positive rate, and AUPR is the area under the curve of precision versus recall. They evaluate how the inferred GRN fits the true GRN from different perspectives. Since $A'_{i,i} \neq 0$ does not necessarily mean autoregulation, we do not compare the diagonal elements of A and A'. See Algorithm 2 for the calculation procedure of such quantities.

4.2 DREAM4 data sets

The most common synthetic data set used to evaluate GRN inference methods is DREAM4 – In Silico Network Challenge [43]. It has multiple challenges, each with multiple data types. We will consider the two in silico challenges and use only the time series data set. There are 5 GRNs with 10 genes, each accompanied by 5 stochastic trajectories at 21 time points.

Algorithm 2: Calculation procedure of AUROC and AUPR.

1. **Input**: true GRN matrix A and calculated GRN matrix A', both with size $n \times n$. Here $A_{i,j}$ can take values $1/0/-1$, and $A'_{i,j} \in \mathbb{R}$

2. **Initiation**: true positive (TP) is 0; true negative (TN) is the number of 0 in all off-diagonal elements of A; false positive (FP) is 0; false negative (FN) is the number of $1/-1$ in all off-diagonal elements of A; TPR_0=FPR_0=Precision_0=Recall_0=0

3. **Sort** the off-diagonal elements of A' by its abstract value in descending order as an array, denoted as Val

4. **For** index $k = 1, \ldots, n^2 - n$:

   $A'_{i,j} = \text{Val}_k$

   If ($A'_{i,j} \geq 0$ and $A_{i,j} = 1$) or ($A'_{i,j} < 0$ and $A_{i,j} = -1$):

   TP=TP+1; FN=FN-1

   Elif ($A'_{i,j} < 0$ and $A_{i,j} = 1$) or ($A'_{i,j} \geq 0$ and $A_{i,j} = -1$):

   FP=FP+1; FN=FN-1

   Else ($A_{i,j} = 0$):

   FP=FP+1; TN=TN-1

   End of if

   Calculate TPR_k=TP/(TP+FN); FPR_k=FP/(FP+TN); Precision_k=TP/(TP+FP); Recall_k=TP/(TP+FN)=TPR_k

5. **Calculate** AUROC, the area under the curve of TPR vs. FPR; calculate AUPR, the area under the curve of Precision vs. Recall

6. **Output**: AUROC and AUPR

There are also 5 GRNs with 100 genes, each accompanied by 10 stochastic trajectories at 21 time points. The time points are equally distributed: 0, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000. Each GRN is represented by a matrix A, where $A_{i,j} = 1$ or $A_{i,j} = 0$, which means that gene i can or cannot regulate gene j. Also, $A_{i,i} = 0$ for each gene i. These two data sets are denoted as DREAM4 (10 genes) and DREAM4 (100 genes). DREAM4 data are generated by GeneNetWeaver software [56], which integrates some unknown SDE system.

We apply WENDY, GENIE3, dynGENIE3, NonlinearODEs, and SINCERITIES methods to DREAM4 data sets. To test the performance of different methods on DREAM4 data, we compare different settings of the same method and choose the best one. Specifically, we find that for methods that can use data from multiple time points, it is not always good to use data from all 21 time points. Instead, data from some time points should be abandoned. Here we list the best settings for each method on each data set:
(1) For WENDY, we regard DREAM4 data as Scenario 9 data. For DREAM4 data (10 genes), we use the data at $t = 450$ and $t = 850$ without the cell correspondence between different time points. For DREAM4 data (100 genes), we use the data at $t = 300$ and $t = 550$ without the cell correspondence between different time points.
(2) For SINCERITIES, we regard DREAM4 data as Scenario 9 data by ignoring the cell correspondence between different time points. For DREAM4 data (10 genes), we use the data at 6 time points: t =(100, 150, 200, 250, 300, 350). For DREAM4 data (100 genes), we use the data at 5 time points: t =(750, 800, 850, 900, 950).
(3) For NonlinearODEs, we regard DREAM4 data as Scenario 8 data. For DREAM4 data (10 genes), we use the data at 3 time points: t =(300, 350, 400). For DREAM4 data (100 genes), we use the data at 8 time points: t =(200, 250, 300, 350, 400, 450, 500, 550).
(4) For GENIE3, we regard DREAM4 data as Scenario 2 data by considering only one time point. For DREAM4 data (10 genes), we use the data at t =350. For DREAM4 data (100 genes), we use the data at t =450.
(5) For dynGENIE3, we regard DREAM4 data as Scenario 10 data. For DREAM4 data (10 genes), we use the data at 13 time points: t =(0, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600). For DREAM4 data (100 genes), we use the data at 18 time points: t =(0, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850).

See Table 2 for the results, where the AUROC and AUPR are averaged over all 5 GRNs in the same data set. We can see that on average, WENDY is slightly better than NonlinearODEs and GENIE3, which are slightly better than SINCERITIES, while dynGENIE3 is significantly better than all other methods. This is not surprising, since dynGENIE3 can utilize the cell correspondence (joint distribution) of different time points, which is not realistic, but contains more information.

Besides, some results are lower than those values reported in corresponding papers [30, 29, 42], since we only use the time series data, not combining with other data types, and we do not manually fine-tune the parameters accordingly.

One problem of the DREAM4 data is that each GRN only generates 5 or 10 stochastic

Table 2: AUROC and AUPR of different methods on DREAM4 data sets

		WENDY	SINCE-RITIES	Nonline-arODEs	GENIE3	dynG-ENIE3
DREAM4	AUROC	0.64	0.61	0.64	0.63	0.76
(10 genes)	AUPR	0.29	0.21	0.28	0.24	0.46
DREAM4	AUROC	0.62	0.55	0.58	0.64	0.74
(100 genes)	AUPR	0.05	0.03	0.03	0.04	0.14
	Total	1.60	1.40	1.53	1.55	2.10

trajectories (each corresponds to a measured cell population). This fits with the mainstream of bulk level data in the early 2010s, when it was difficult to repeat the measurement many times. Nevertheless, with the development of single-cell RNA sequencing, it is easier to obtain single-cell data from thousands of cells. Therefore, we want to generate single-cell expression data over more cells and test the performance of inference methods under different cell numbers.

4.3 SINC data sets

Pinna et al. [52] and Papili Gao et al. [50] use the following SDE to generate synthetic data:

$$\mathrm{d}X_j(t) = V\left\{\beta\prod_{i=1}^{n}\left[1 + A_{i,j}\frac{X_i(t)}{X_i(t)+1}\right] - \theta X_j(t)\right\}\mathrm{d}t + \sigma X_j(t)\mathrm{d}W_j(t). \tag{6}$$

Here $X_i(t)$ is the expression level of gene i at time t, and $W_j(t)$ is a standard Brownian motion, independent with other $W_i(\tau)$. $A_{i,j}$ describes the GRN. $V = 30$, $\beta = 1$, $\theta = 0.2$, $\sigma = 0.1$. When $t \to \infty$, each $X_i(t)$ will converge to the stationary value and fluctuate slightly around it [73, 11, 12].

The following equation is the first-order approximation of Eq. 6 when $X_i(t)$ is small:

$$\mathrm{d}X_j(t) = V\left\{\beta\left[1 + \sum_{i=1}^{n} A_{i,j}X_i(t)\right] - \theta X_j(t)\right\}\mathrm{d}t + \sigma X_j(t)\mathrm{d}W_j(t), \tag{7}$$

Also, Eq. 2 is the discretization of Eq. 7: A, b_i, $F_{i,i}$ in Eq. 2 correspond to $V\beta A - V\theta I$, $V\beta$, $\sigma^2 t$ in Eq. 7. These facts provide the theoretical support that WENDY (derived from Eq. 2) can work for data generated by Eq. 6, since we only care about off-diagonal elements of A, which represent mutual regulations between different genes. For other gene regulation mechanisms, we can also use first-order approximation and discretization to obtain Eq. 2 or similar forms. Therefore, WENDY should be applicable to different gene regulation mechanisms.

For Eq. 6, we simulate the system with the Euler-Maruyama method [33] for time $t \in [0, 1]$ with time step $\Delta t = 0.01$. This means treating $\mathrm{d}X_j(t)$ as $X_j(t + \Delta t) - X_j(t)$, $\mathrm{d}t$ as Δt, and $\mathrm{d}W_j(t)$ as a normal random variable $\mathcal{N}(0, \Delta t)$.

For the GRN matrix A, we use the 40 networks in Papili Gao et al.'s paper [50]: 10 *Escherichia coli* networks with 10 genes, 10 *E. coli* networks with 20 genes, 10 *Saccharomyces cerevisiae* (yeast) networks with 10 genes, 10 yeast networks with 20 genes. Each network has $A_{i,j} = 1/0/-1$, and $A_{i,i} = 0$ for each i.

For each group of data, we run the simulation m times, where m represents the number of cells/trajectories measured in reality. We set $m = 10$, $m = 30$, and $m = 100$ to test the performance of different methods under different m. The initial state $X_i(0)$ is independently and uniformly sampled in $[0, 1)$. For each network in Papili Gao et al.'s paper, and each value of m, we generate 100 groups of data. The same as Papili Gao et al.'s paper, the simulation finishes at $t = 3.0$. We record the expression levels at the following 11 time points: $t =$(0.0, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0). For the data sets generated by Eq. 6 on GRNs with 10 or 20 genes, we name them SINC (10 or 20 genes) data. Since the initial state is generally different from the stationary state, SINC data should be regarded as Scenario 10 data (single-cell level time series data under general interventions, where the joint distribution of different time points is known).

To test the performance of different methods on SINC data, we compare different settings of the same method and choose the best one. Specifically, we find that for methods that can use data from multiple time points, it is not always good to use data from all 11 time points. Instead, data from some time points should be abandoned. Here we list the best settings for each method on each data set:

(1) For WENDY, we regard SINC data as Scenario 9 data. For SINC data (10 genes), we use the data at $t = 0.0$ and $t = 3.0$ without the cell correspondence between different time points. For SINC data (20 genes), we use the data at $t = 2.7$ and $t = 3.0$.

(2) For SINCERITIES, we regard SINC data as Scenario 9 data by ignoring the cell correspondence between different time points. For SINC data (10 genes), we use the data at 8 time points: $t =$(0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0). For SINC data (100 genes), we use the data at 9 time points: $t =$(0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0). This fits with Papili Gao et al.'s paper, as they also abandon the data for $t \leq 0.5$.

(3) For NonlinearODEs, we regard SINC data as Scenario 8 data by taking average of gene expression levels over different cells. For SINC data (10 genes), we use the data at 10 time points: $t =$(0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0). For SINC data (100 genes), we use the data at 8 time points: $t =$(0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4).

(4) For GENIE3, we regard SINC data as Scenario 2 data by considering only one time point. For both SINC data (10 genes) and SINC data (100 genes), we use the data at $t = 0.3$.

(5) For dynGENIE3, we regard SINC data as Scenario 10 data. For both SINC data (10 genes) and SINC data (100 genes), we use the data at 10 time points: $t =$(0.0, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7).

We apply WENDY, SINCERITIES, NonlinearODEs, GENIE3, and dynGENIE3 to SINC data sets and compare the corresponding AUROC and AUPR. Each AUROC or AUPR value is averaged over 2000 simulations (20 GRNs, each with 100 groups of data). See Table 3 for the results. We can see that WENDY is significantly better than all other methods. Here dynGENIE3 has the worst performance, possibly because it cannot determine the sign (activation/inhibition) of a regulation.

For different values of cell/trajectory number m, we can see that the performance metrics generally increase with m, meaning that more cells provide more information.

Table 3: AUROC and AUPR of different methods on SINC data sets

		WENDY	SINCERITIES	NonlinearODEs	GENIE3	dynGENIE3
SINC $m = 10$	AUROC	0.59	0.44	0.47	0.46	0.43
(10 genes)	AUPR	0.14	0.11	0.11	0.11	0.10
SINC $m = 10$	AUROC	0.76	0.42	0.51	0.52	0.40
(20 genes)	AUPR	0.16	0.05	0.08	0.08	0.04
SINC $m = 30$	AUROC	0.67	0.46	0.49	0.50	0.44
(10 genes)	AUPR	0.19	0.12	0.12	0.15	0.10
SINC $m = 30$	AUROC	0.81	0.47	0.53	0.47	0.40
(20 genes)	AUPR	0.21	0.07	0.08	0.07	0.04
SINC $m = 100$	AUROC	0.69	0.48	0.50	0.55	0.42
(10 genes)	AUPR	0.22	0.13	0.12	0.19	0.10
SINC $m = 100$	AUROC	0.82	0.48	0.54	0.54	0.39
(20 genes)	AUPR	0.23	0.08	0.08	0.09	0.05
	Total	5.49	3.31	3.63	3.73	2.91

Notice that WENDY is not always better in all settings. For example, if we only have the data at $t = 0.0$ and $t = 0.1$ for SINC data with $m = 100$, then WENDY has AUROC 0.77 and AUPR 0.34 for SINC (10 genes) data, and AUROC 0.83 and AUPR 0.29 for SINC (20 genes) data, while GENIE3 has AUROC 0.77 and AUPR 0.32 for SINC (10 genes) data, and AUROC 0.92 and AUPR 0.50 for SINC (20 genes) data. Besides, for the setting in Table 3, if we reduce m to 5, then the performance of WENDY is also no better than GENIE3.

We also test the running time of each inference method. For SINC (10 genes) and SINC (20 genes) data sets, we measure the execution time (averaged over different GRNs) of each algorithm for $m = 10$, $m = 30$, and $m = 100$ cells. See Table 4 for the running time (in the form of mean $\pm$ standard deviation) in different settings. We can see that for each algorithm, the time cost increases with the gene number n. In addition, WENDY, SINCERITIES and non-linearODEs are insensitive to the cell number m, while the time costs of GENIE3 and dynGENIE3 increase significantly with m. When m and n are large,

WENDY is roughly the same fast as SINCERITIES and NonlinearODEs, but much faster than GENIE3 and dynGENIE3. Therefore, WENDY has a satisfactory speed.

Table 4: Computational time for different algorithms. We consider the SINC data of $n = 10/20$ genes over $m = 10$, $m = 30$, or $m = 100$ cells. For each situation, we use 20 different GRNs, and present the computational time as mean $\pm$ standard deviation. All time costs are in seconds, and are measured on a desktop with Intel i7-13700 CPU.

		WENDY	SINCE-RITIES	Nonline-arODEs	GENIE3	dynG-ENIE3
SINC	$m = 10$	0.22 ± 0.04	0.09 ± 0.01	0.19 ± 0.02	4.27 ± 0.06	7.28 ± 0.08
$n = 10$	$m = 30$	0.15 ± 0.03	0.11 ± 0.01	0.18 ± 0.02	4.83 ± 0.05	15.86 ± 0.17
genes	$m = 100$	0.14 ± 0.03	0.12 ± 0.01	0.19 ± 0.03	7.32 ± 0.09	51.79 ± 0.63
SINC	$m = 10$	0.38 ± 0.10	0.42 ± 0.01	0.35 ± 0.03	8.55 ± 0.05	16.24 ± 0.13
$n = 20$	$m = 30$	0.34 ± 0.03	0.45 ± 0.01	0.36 ± 0.08	9.95 ± 0.07	38.21 ± 0.28
genes	$m = 100$	0.27 ± 0.02	0.64 ± 0.01	0.34 ± 0.04	16.32 ± 0.17	131.89 ± 0.76

5 Results of WENDY method on experimental data

In this section, we apply WENDY to three experimental data sets in Scenario 9 and analyze how GRN evolves over time. The same as Matsumoto et al. [44], we study the following three data sets:

mESC data set: Single-cell expression levels of mouse embryonic stem cells, measured at $t = 0, 12, 24, 48, 72$h [27]. There are totally 456 cells measured.

MEF data set: Single-cell expression levels of mouse embryonic fibroblast cells, measured at $t = 0, 2, 5, 20, 22$d [59]. There are totally 373 cells measured.

hESC data set: Single-cell expression levels of human embryonic stem cells, measured at $t = 0, 12, 24, 36, 72, 96$h [16]. There are totally 758 cells measured.

In all these experiments, the cells are actively replicating and differentiating, during which the cell types and the GRNs might change along time.

For each data set, Matsumoto et al. [44] calculated the mean expression level of each gene at each time point, and then calculated the variance of the mean expression levels at different time points for each gene. Using this approach, they selected out top 100 highly varying genes, which might actively regulate each other. Given the data of these 100 highly varying genes, we further select out genes that express in at least 95% of the cells. The reason is that when the expression data have too many 0s, graphical lasso might fail to converge, which fails WENDY. For such genes, we apply WENDY for the data of each neighboring pair of time points and calculate the corresponding GRN. We illustrate how top 15 strong regulations (edges) evolve along time.

For mESC data set, we select 18 genes: PHB, DNMT3B, TRP53, PARP1, TBX3,

KLF6, DNMT3L, ELF3, KLF4, ZFHX3, HMGA2, POU5F1, CREB3, ID1, SALL4, RERE, NFXL1, DNMT3A. See Fig. 2 for the GRNs at different time points. We can see that initially DNMT3L (catalytically inactive regulatory factor of DNA methyltransferases) [62] and DNMT3B (establishes DNA methylation patterns in embryos) [36] are major targets of regulation. As time increases, many regulatory effects show a weakening in the latter three time points, relative to the initial state of the GRNs.

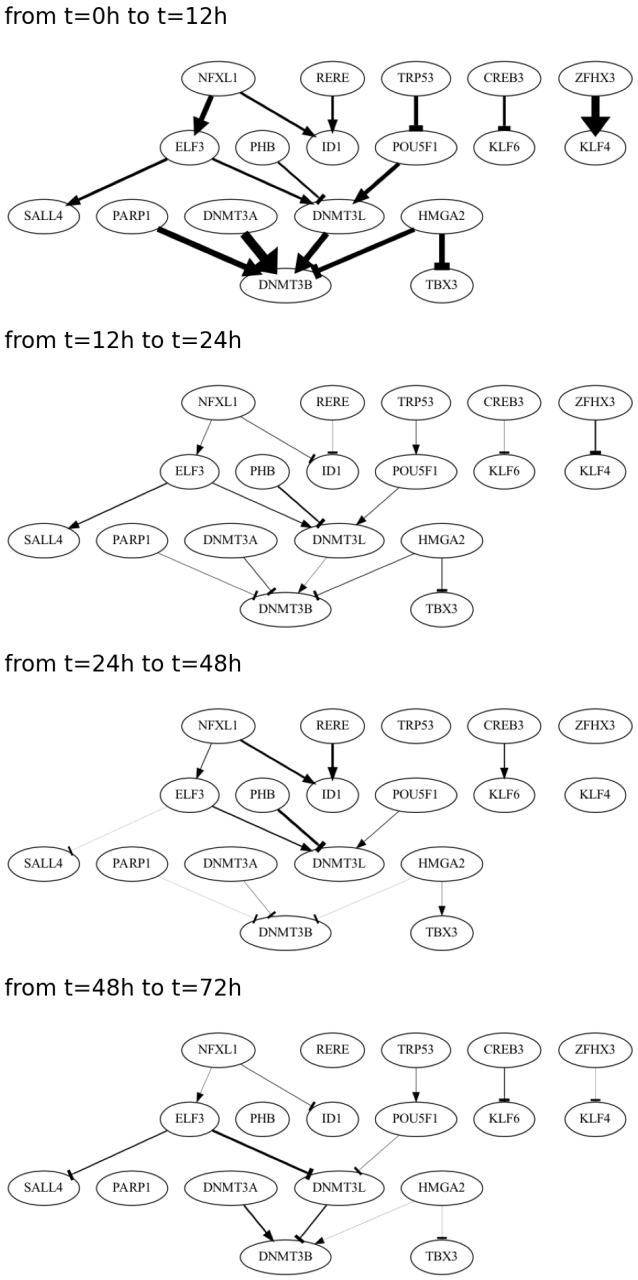

Figure 2: The GRNs inferred by the WENDY method on mESC data set. For each pair of neighboring time points, we illustrate the GRN, where sharp arrow $\rightarrow$ means activation, and flat arrow $\dashv$ means inhibition. The width of arrow represents the regulation strength.

For MEF data set, we select 9 genes: NFIA, MATR3, TAX1BP3, GATAD1, RUNX1T1, PCBP2, ID2, TCF4, SFPQ. See Fig. 3 for the GRNs at different time points. We can see that initially the major gene in regulation is GATAD1 (regulates sequence-specific DNA binding activity and chromatin remodeling) [35], and then the major gene gradually becomes ID2 (enables transcription factor binding activity and transcription regulator inhibitor activity) [40], and lastly the major gene switches to RUNX1T1 (regulates DNA binding activity) [72]. As time increases, the regulation strength transitions from the left side to the right side of the GRNs during the first three time points. However, towards the end of the process, the regulation generally weakens.

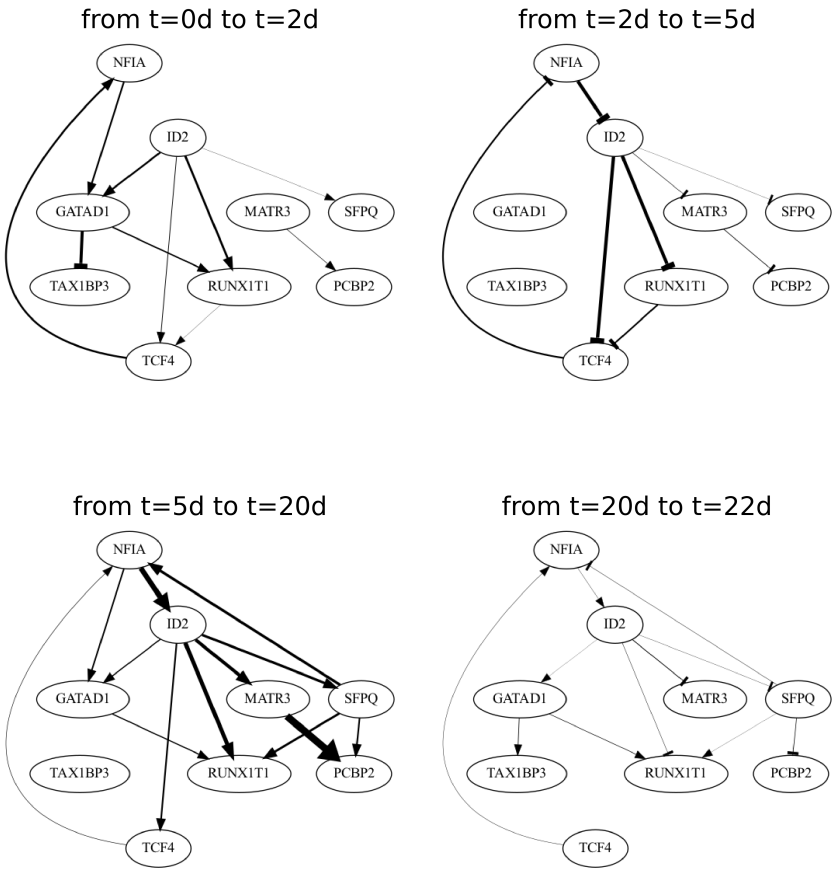

Figure 3: The GRNs inferred by the WENDY method on MEF data set. For each pair of neighboring time points, we illustrate the GRN, where sharp arrow $\rightarrow$ means activation, and flat arrow $\dashv$ means inhibition. The width of arrow represents the regulation strength.

For hESC data set, we select 11 genes: SOX11, TULP4, SMAD2, ZFP42, AEBP2,

ARID4B, ZFX, CEBPZ, MIER1, BBX, TERF1. See Fig. 4 for the GRNs at different time points. We can see that initially the major source of regulation is CEBPZ (DNA-binding transcriptional activator) [28], and the major targets of regulation are SOX11 (regulates embryonic development and determines cell fate) [19] and AEBP2 (enables transcription coregulator activity) [17]. After a silent period during 24–36h, AEBP2 exerts stronger regulations on its downstream genes.

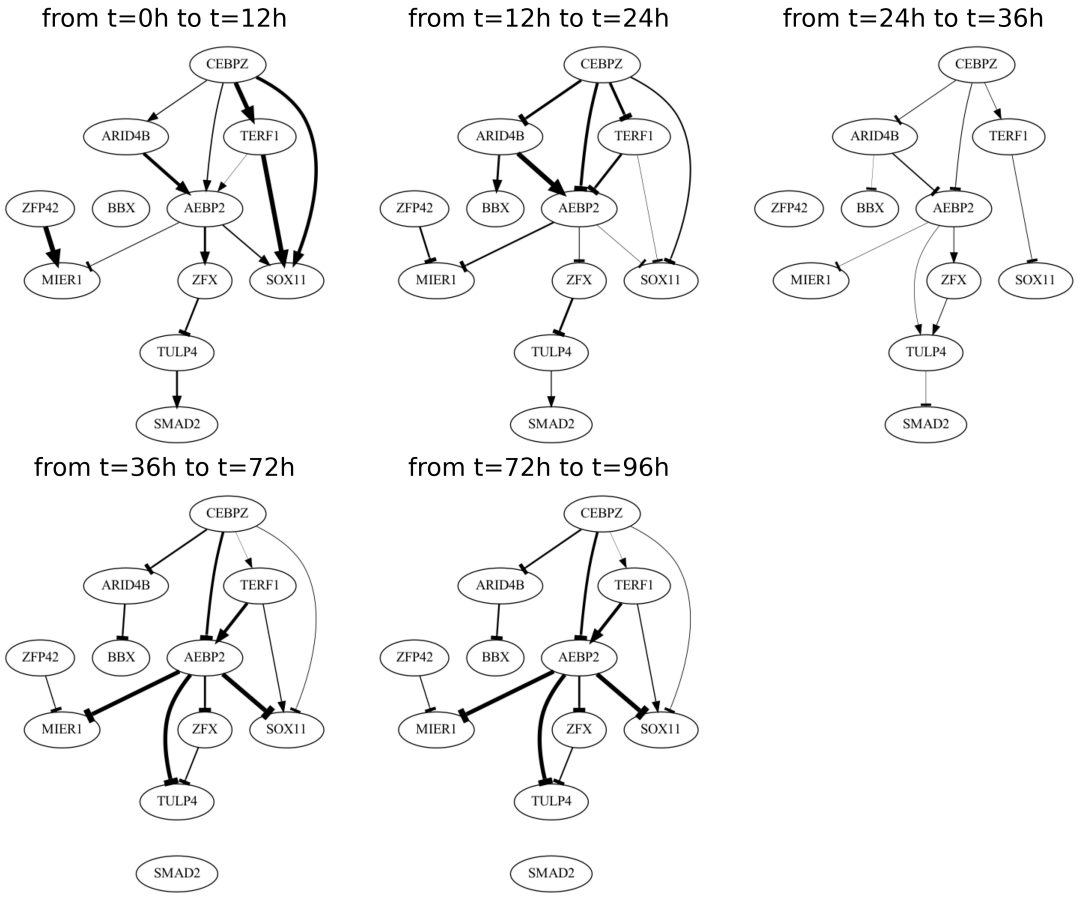

Figure 4: The GRNs inferred by the WENDY method on hESC data set. For each pair of neighboring time points, we illustrate the GRN, where sharp arrow $\rightarrow$ means activation, and flat arrow $\dashv$ means inhibition. The width of arrow represents the regulation strength.

In sum, we can see that the center of gene regulation can gradually shift during development, and the overall strength of regulations decreases along time in general. However, for inferred regulatory relations between two genes, it is always possible that these two

genes do not regulate each other, but are both regulated by a third unmeasured gene (or other environmental factors).

6 Discussion

In this paper, we address the GRN inference problem for single-cell time series gene expression data following general interventions, where the joint distribution of different time points is unknown. Although this type of data is common in recent experiments, there are few GRN inference methods that fully utilize the information contained in the data. Therefore, we introduce WENDY, a GRN inference method developed for single-cell gene expression data spanning two time points after an intervention. This method is capable of inferring all mutual regulatory relations, including direction.

Similar to most other GRN inference methods, WENDY cannot infer autoregulation. One potential future direction is to develop an autoregulation inference method inspired by the principles of WENDY. For instance, we can choose a reasonable nonlinear gene expression model that allows autoregulation, and study whether autoregulation can make a difference for the dynamics of covariance matrices.

In interventional experiments, it is customary to measure expression levels before intervention as the control group. These control data align with Scenario 2 data, enabling the application of corresponding inference methods. Therefore, a comparison of GRNs before and after intervention can elucidate the effect of the intervention on gene regulation.

As of 2024, single-cell level gene expression measurement remains relatively insensitive. It is common to miss all mRNAs of one gene in a single cell, resulting in experimental data with many zeros. This characteristic may impede graphical lasso, and consequently, WENDY. A potential future direction involves developing more robust inference methods capable of handling data with numerous missing values.

Although WENDY aims for data type 9, other types of data can help to improve the results of WENDY. For instance, one can use a CRISPR gene knockout study to verify regulations inferred by WENDY. Besides, advancements in biotechnology introduce new types of data, which are not in our classification framework, but might be applicable to GRN inference, and can enhance WENDY's capabilities through incorporation.

In our tests, we observed that WENDY performs better for data in earlier time points when dealing with a small number of genes (10 or fewer). Conversely, for data collected long after intervention, where gene expression approaches a new steady state, the inference results are less reliable compared to earlier time points. One possible reason is that after enough time, the covariance matrix approaches its steady state, and the small changes of covariances might be covered by random noise. Therefore, when applying WENDY to experimental data, caution should be exercised with results obtained several days after the intervention.

Theoretically, the GRN matrix A calculated by WENDY is not symmetric, allowing

determination of the directions of the regulatory relations ($i \to j$ or $j \to i$). However, our simulations indicate that $A_{i,j}$ and $A_{j,i}$ generally exhibit proximity, resulting in a significant decrease in AUPR. A prospective avenue involves developing a novel solver capable of producing highly asymmetric results.

In our WENDY implementation, we employ the standard graphical lasso algorithm to determine the inverse of the covariance matrix. Other algorithms with similar purposes [61, 18] can also be considered for integration into WENDY.

Acknowledgments

YW would like to thank Joseph Zhou for helpful discussions.

Data and code availability

The solver used in WENDY method is in

`https://github.com/zhengp0/genet`

Main function of WENDY method (including a tutorial), and other data and code files used in this paper are in

`https://github.com/YueWangMathbio/WENDY`

References

[1] Aalto, A., Viitasaari, L., Ilmonen, P., Mombaerts, L., and Gonçalves, J. Gene regulatory network inference from sparsely sampled noisy data. *Nature communications 11*, 1 (2020), 3493.

[2] Agostinho, N. B., Machado, K. S., and Werhli, A. V. Inference of regulatory networks with a convergence improved mcmc sampler. *BMC bioinformatics 16* (2015), 1–10.

[3] Atanackovic, L., Tong, A., Hartford, J., Lee, L. J., Wang, B., and Bengio, Y. Dyngfn: Bayesian dynamic causal discovery using generative flow networks. *arXiv preprint arXiv:2302.04178* (2023).

[4] Axelrod, S., Li, X., Lincoln, S., Terceros, A., O'Neil, J., Wang, Z., Nguyen, A., and Young, M. W. A role for the drosophila blood-brain barrier in the regulation of sleep. *Available at SSRN 3895317* (2021).

[5] Axelrod, S., Li, X., Sun, Y., Lincoln, S., Terceros, A., O'Neil, J., Wang, Z., Nguyen, A., Vora, A., Spicer, C., et al. The drosophila blood–brain barrier regulates sleep via moody g protein-coupled receptor signaling. *Proceedings of the National Academy of Sciences 120*, 42 (2023), e2309331120.

[6] Badia-i Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R. O., Argelaguet, R., and Saez-Rodriguez, J. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics* (2023), 1–16.

[7] Barbosa, S., Niebel, B., Wolf, S., Mauch, K., and Takors, R. A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints. *Biosystems 174* (2018), 37–48.

[8] Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. Reverse engineering of regulatory networks in human b cells. *Nature genetics 37*, 4 (2005), 382–390.

[9] Burdziak, C., Zhao, C. J., Haviv, D., Alonso-Curbelo, D., Lowe, S. W., and Pe'er, D. sckinetics: inference of regulatory velocity with single-cell transcriptomics data. *Bioinformatics 39*, Supplement_1 (2023), i394–i403.

[10] Chakraborty, P., George, J. T., Woodward, W. A., Levine, H., and Jolly, M. K. Gene expression profiles of inflammatory breast cancer reveal high heterogeneity across the epithelial-hybrid-mesenchymal spectrum. *Translational oncology 14*, 4 (2021), 101026.

[11] Cheng, Y.-C., and Qian, H. Stochastic limit-cycle oscillations of a nonlinear system under random perturbations. *Journal of Statistical Physics 182*, 3 (2021), 47.

[12] Cheng, Y.-C., Qian, H., and Zhu, Y. Asymptotic behavior of a sequence of conditional probability distributions and the canonical ensemble. In *Annales Henri Poincaré* (2021), vol. 22, Springer, pp. 1561–1627.

[13] Cheng, Y.-C., Stein, S., Nardone, A., Liu, W., Ma, W., Cohen, G., Guarducci, C., McDonald, T. O., Jeselsohn, R., and Michor, F. Mathematical modeling identifies optimum palbociclib-fulvestrant dose administration schedules for the treatment of patients with estrogen receptor–positive breast cancer. *Cancer Research Communications 3*, 11 (2023), 2331–2344.

[14] Cheng, Y.-C., Zhang, Y., Tripathi, S., BV, H., Jolly, M. K., Schiebinger, G., Levine, H., McDonald, T. O., and Michor, F. Reconstruction of single cell lineage trajectories and identification of diversity in fates during the epithelial-to-mesenchymal transition. *bioRxiv* (2023), 2023–09.

[15] Cho, H., Berger, B., and Peng, J. Reconstructing causal biological networks through active learning. *PloS one 11*, 3 (2016), e0150611.

[16] Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., Choi, J., Kendziorski, C., Stewart, R., and Thomson, J. A. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology 17* (2016), 1–20.

[17] Ciferri, C., Lander, G. C., Maiolica, A., Herzog, F., Aebersold, R., and Nogales, E. Molecular architecture of human polycomb repressive complex 2. *elife 1* (2012), e00005.

[18] Danaher, P., Wang, P., and Witten, D. M. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*, 2 (2014), 373–397.

[19] Decaesteker, B., Louwagie, A., Loontiens, S., De Vloed, F., Bekaert, S.-L., Roels, J., Vanhauwaert, S., De Brouwer, S., Sanders, E., Berezovskaya, A., et al. Sox11 regulates swi/snf complex components as member of the adrenergic neuroblastoma core regulatory circuitry. *Nature communications 14*, 1 (2023), 1267.

[20] Deshpande, A., Chu, L.-F., Stewart, R., and Gitter, A. Network inference with granger causality ensembles on single-cell transcriptomics. *Cell reports 38*, 6 (2022).

[21] Erbe, R., Gore, J., Gemmill, K., Gaykalova, D. A., and Fertig, E. J. The use of machine learning to discover regulatory networks controlling biological systems. *Molecular cell 82*, 2 (2022), 260–273.

[22] Feigelman, J., Ganscha, S., Hastreiter, S., Schwarzfischer, M., Filipczyk, A., Schroeder, T., Theis, F. J., Marr, C., and Claassen, M. Analysis of cell lineage trees by exact bayesian inference identifies negative autoregulation of nanog in mouse embryonic stem cells. *Cell systems 3*, 5 (2016), 480–490.

[23] Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*, 3 (2008), 432–441.

[24] Ghosh, S., and Zhong, C. Convergence rate of a collapsed gibbs sampler for crossed random effects models. *arXiv preprint arXiv:2109.02849* (2021).

[25] Gustafsson, M., Hornquist, M., and Lombardi, A. Constructing and analyzing a large-scale gene-to-gene regulatory network lasso-constrained inference and biological validation. *IEEE/ACM Transactions on computational biology and bioinformatics 2*, 3 (2005), 254–261.

[26] Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology 6*, 1 (2012), 1–17.

[27] Hayashi, T., Ozaki, H., Sasagawa, Y., Umeda, M., Danno, H., and Nikaido, I. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature communications 9*, 1 (2018), 619.

[28] Herold, T., Metzeler, K. H., Vosberg, S., Hartmann, L., Röllig, C., Stölzel, F., Schneider, S., Hubmann, M., Zellmeier, E., Ksienzyk, B., et al. Isolated trisomy 13 defines a homogeneous aml subgroup with high frequency of mutations in spliceosome genes and poor prognosis. *Blood, The Journal of the American Society of Hematology 124*, 8 (2014), 1304–1311.

[29] Huynh-Thu, V. A., and Geurts, P. dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. *Scientific reports 8*, 1 (2018), 3384.

[30] 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 (2010), e12776.

[31] Huynh-Thu, V. A., and Sanguinetti, G. Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics 31*, 10 (2015), 1614–1622.

[32] Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. Dissecting cell identity via network inference and in silico gene perturbation. *Nature 614*, 7949 (2023), 742–751.

[33] Kloeden, P. E., Platen, E., Kloeden, P. E., and Platen, E. *Stochastic differential equations*. Springer, 1992.

[34] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. Rna velocity of single cells. *Nature 560*, 7719 (2018), 494–498.

[35] Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., and Kingston, R. E. Characterization of the pirna complex from rat testes. *Science 313*, 5785 (2006), 363–367.

[36] Lauria, A., Meng, G., Proserpio, V., Rapelli, S., Maldotti, M., Polignano, I. L., Anselmi, F., Incarnato, D., Krepelova, A., Donna, D., et al. Dnmt3b supports meso-endoderm differentiation from mouse embryonic stem cells. *Nature communications 14*, 1 (2023), 367.

[37] Lee, H.-C., Danieletto, M., Miotto, R., Cherng, S. T., and Dudley, J. T. Scaling structural learning with no-bears to infer causal transcriptome networks. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020* (2019), World Scientific, pp. 391–402.

[38] Li, W., Wang, Z., Syed, S., Lyu, C., Lincoln, S., O'Neil, J., Nguyen, A. D., Feng, I., and Young, M. W. Chronic social isolation signals starvation and reduces sleep in drosophila. *Nature 597*, 7875 (2021), 239–244.

[39] Lim, C. Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., and Göttgens, B. Btr: training asynchronous boolean models using single-cell expression data. *BMC bioinformatics 17*, 1 (2016), 1–18.

[40] Liu, F., Chen, S., Yu, Y., Huang, C., Chen, H., Wang, L., Zhang, W., Wu, J., and Ye, Y. Inhibitor of dna binding 2 knockdown inhibits the growth and liver metastasis of colorectal cancer. *Gene 819* (2022), 146240.

[41] Liu, F., Zhang, S.-W., Guo, W.-F., Wei, Z.-G., and Chen, L. Inference of gene regulatory network based on local bayesian networks. *PLoS computational biology 12*, 8 (2016), e1005024.

[42] Ma, B., Fang, M., and Jiao, X. Inference of gene regulatory networks based on nonlinear ordinary differential equations. *Bioinformatics 36*, 19 (2020), 4885–4893.

[43] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., et al. Wisdom of crowds for robust gene network inference. *Nature methods 9*, 8 (2012), 796–804.

[44] 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 (2017), 2314–2321.

[45] McDonald, T. O., Cheng, Y.-C., Graser, C., Nicol, P. B., Temko, D., and Michor, F. Computational approaches to modelling and optimizing cancer treatment. *Nature Reviews Bioengineering 1*, 10 (2023), 695–711.

[46] Morrissey, E. R., Juárez, M. A., Denby, K. J., and Burroughs, N. J. On reverse engineering of gene interaction networks using time course data with repeated measurements. *Bioinformatics 26*, 18 (2010), 2305–2312.

[47] Nauta, M., Bucur, D., and Seifert, C. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction 1*, 1 (2019), 19.

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

[49] Nouri, N., Gaglia, G., Mattoo, H., de Rinaldis, E., and Savova, V. Comparative analysis of association networks using single-cell rna sequencing data reveals perturbation-relevant gene signatures. *bioRxiv* (2023), 2023–09.

[50] Papili Gao, N., Ud-Dean, S. M., Gandrillon, O., and Gunawan, R. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics 34*, 2 (2018), 258–266.

[51] Perrin, B.-E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d'Alché Buc, F. Gene networks inference using dynamic bayesian networks. *Bioinformatics 19*, 2 (2003), 138–148.

[52] Pinna, A., Soranzo, N., and De La Fuente, A. From knockouts to networks: establishing direct cause-effect relationships through graph analysis. *PloS one 5*, 10 (2010), e12912.

[53] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods 17*, 2 (2020), 147–154.

[54] Qian, H., and Cheng, Y.-C. Counting single cells and computing their heterogeneity: from phenotypic frequencies to mean value of a quantitative biomarker. *Quantitative Biology 8*, 2 (2020), 172–176.

[55] Reid, J. E., and Wernisch, L. Pseudotime estimation: deconfounding single cell time series. *Bioinformatics 32*, 19 (2016), 2973–2980.

[56] Schaffter, T., Marbach, D., and Floreano, D. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics 27*, 16 (2011), 2263–2270.

[57] Shrivastava, H., Zhang, X., Song, L., and Aluru, S. Grnular: A deep learning framework for recovering single-cell gene regulatory networks. *Journal of Computational Biology 29*, 1 (2022), 27–44.

[58] Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics 19* (2018), 1–16.

[59] Treutlein, B., Lee, Q. Y., Camp, J. G., Mall, M., Koh, W., Shariati, S. A. M., Sim, S., Neff, N. F., Skotheim, J. M., Wernig, M., et al. Dissecting

direct reprogramming from fibroblast to neuron using single-cell rna-seq. *Nature 534*, 7607 (2016), 391–395.

[60] Vijayan, V., Wang, Z., Chandra, V., Chakravorty, A., Li, R., Sarbanes, S. L., Akhlaghpour, H., and Maimon, G. An internal expectation guides drosophila egg-laying decisions. *Science Advances 8*, 43 (2022), eabn3852.

[61] Vinci, G., Dasarathy, G., and Allen, G. I. Graph quilting: graphical model selection from partially observed covariances. *arXiv preprint arXiv:1912.05573* (2019).

[62] Vlachogiannis, G., Niederhuth, C. E., Tuna, S., Stathopoulou, A., Viiri, K., de Rooij, D. G., Jenner, R. G., Schmitz, R. J., and Ooi, S. K. The dnmt3l add domain controls cytosine methylation establishment during spermatogenesis. *Cell reports 10*, 6 (2015), 944–956.

[63] Wang, L., Trasanidis, N., Wu, T., Dong, G., Hu, M., Bauer, D. E., and Pinello, L. Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nature Methods 20*, 9 (2023), 1368–1378.

[64] Wang, Y., and He, S. Inference on autoregulation in gene expression with variance-to-mean ratio. *Journal of Mathematical Biology 86*, 5 (2023), 87.

[65] Wang, Y., and Wang, Z. Inference on the structure of gene regulatory networks. *Journal of Theoretical Biology 539* (2022), 111055.

[66] Wang, Z. *Identification of Gene Expression Changes in Sleep Mutants Associated With Reduced Longevity in Drosophila*. PhD thesis, The Rockefeller University, 2020.

[67] Wang, Z., Lincoln, S., Nguyen, A. D., Li, W., and Young, M. W. Chronic sleep loss disrupts rhythmic gene expression in drosophila. *Frontiers in Physiology 13* (2022), 1048751.

[68] Wasserman, L. *All of statistics: a concise course in statistical inference*, vol. 26. Springer, 2004.

[69] Wen, Y., Huang, J., Guo, S., Elyahu, Y., Monsonego, A., Zhang, H., Ding, Y., and Zhu, H. Applying causal discovery to single-cell analyses using causalcell. *Elife 12* (2023), e81464.

[70] Wu, B., Piatkevich, K. D., Lionnet, T., Singer, R. H., and Verkhusha, V. V. Modern fluorescent proteins and imaging technologies to study gene expression, nuclear localization, and dynamics. *Current opinion in cell biology 23*, 3 (2011), 310–317.

[71] Xing, B., and Van Der Laan, M. J. A causal inference approach for constructing transcriptional regulatory networks. *Bioinformatics 21*, 21 (2005), 4007–4013.

[72] Yan, M., Ahn, E.-Y., Hiebert, S. W., and Zhang, D.-E. Runx1/aml1 dna-binding domain and eto/mtg8 nhr2-dimerization domain are critical to aml1-eto9a leukemogenesis. *Blood, The Journal of the American Society of Hematology 113*, 4 (2009), 883–886.

[73] Yang, Y.-J., and Cheng, Y.-C. Potentials of continuous markov processes and random perturbations. *Journal of Physics A: Mathematical and Theoretical 54*, 19 (2021), 195001.

[74] Zhang, X., Zhao, X.-M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.-K., Liu, Z.-P., and Chen, L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics 28*, 1 (2012), 98–104.

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

[76] Zhao, M., He, W., Tang, J., Zou, Q., and Guo, F. A comprehensive overview and critical evaluation of gene regulatory network inference technologies. *Briefings in Bioinformatics 22*, 5 (2021), bbab009.

[77] 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 (2022), bbab568.

[78] Zheng, R., Li, M., Chen, X., Wu, F.-X., Pan, Y., and Wang, J. Bixgboost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. *Bioinformatics 35*, 11 (2019), 1893–1900.

[79] Zhong, C. Mallows permutation models with l^1 and l^2 distances i: hit and run algorithms and mixing times. *arXiv preprint arXiv:2112.13456* (2021).

[80] Zhong, C., and Tibshirani, R. Survival analysis as a classification problem. *arXiv preprint arXiv:1909.11171* (2019).