CONSTRUCTING GENE REGULATORY NETWORK USING CHATTERJEE'S RANK CORRELATION WITH SINGLE-CELL TRANSCRIPTOMIC DATA

SHREYAN GUPTA[1], ANAMITRA CHAUDHURI[2], VISHNUVASAN RAGHURAMAN[4], YANG NI[2] AND JAMES J. CAI[1,3]

[1]*Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas, USA*

[2]*Department of Statistics and Data Sciences, University of Texas at Austin, Texas, USA*

[3]*Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA*

[4]*Department of Computer Science and Engineering, Texas A&M University, College Station, Texas, USA*

Discovering gene regulatory networks (GRNs) from single-cell RNA sequencing (scRNA-seq) data is critical for understanding cellular function, but existing methods are limited by strong theoretical assumptions or high computational complexity. We introduce a multiple testing framework for inference of GRN by using Chatterjee's rank correlation coefficient, a nonparametric measure of dependence. Our approach overcomes the limitations of traditional methods while offering a transparent, scalable, and computationally efficient alternative to recent black-box machine learning models. Crucially, we address the challenge of non-independent observations in scRNA-seq by developing a data-driven algorithm for estimating robust testing cutoffs. Furthermore, we exploit the asymmetric nature of Chatterjee's correlation to propose a new test for directed regulation, enabling the construction of biologically meaningful and directionally informed GRNs. We demonstrate that our method consistently outperforms state-of-the-art approaches in recovering true regulatory links from both simulated and real datasets, providing a powerful tool for dissecting complex GRNs.

Keywords: Gene regulatory networks, Chatterjee's correlation, single-cell RNA sequencing, directed regulation.

1. Introduction

Deciphering cellular behavior relies on understanding the intricate network of signaling pathways that orchestrate cell function. A good example is the response to growth factors, where a signal propagates through a cascade of receptor activation and phosphorylation events, ultimately altering gene expression by activating transcription factors [1, 2]. This directionality of signaling is a fundamental principle of cell biology, and accurately mapping these relationships is a key challenge [3]. Reverse engineering of gene regulatory networks (GRNs) from transcriptomic data is a powerful computational approach for understanding these gene interactions. The advent of single-cell RNA sequencing (scRNA-seq) has enabled GRN inference at a higher resolution than bulk RNA-seq [4]. This capability allows us to dissect the regulatory logic that governs cell fate, differentiation, and disease, offering deeper insights into biological systems at a cellular level.

Although several existing methods have been proposed to construct GRNs, they suffer from key limitations that restrict their effectiveness and interpretability. Traditional correlation-based methods, such as Pearson's and Spearman's correlations, rely heavily on strong assumptions, such as Gaussianity, which rarely hold in complex biological systems [5], or fail to infer the directionality of the regulation. Newer machine learning-based approaches like GRNBoost2 [6] and PCNet [7], though flexible, are computationally intensive and often function as black boxes, lacking transparency in how regulatory links are inferred [8]. To overcome these challenges,

we propose a new GRN construction method based on Chatterjee's correlation coefficient [9], a nonparametric measure of dependence. While Chatterjee's correlation has found successful applications in diverse fields, including geochemical sciences [10] and the inference of gene dependencies along scRNA-seq derived differentiation trajectories [11], its potential for deciphering directed gene-gene interactions in scRNA-seq remains largely unexplored. Our approach (a) does not require distributional assumptions, (b) captures the directionality of gene regulations, (c) can accommodate both linear and nonlinear dependencies, (d) remains interpretable through rank-based construction, (e) is computationally simple and scalable, and (f) relies only on local information—using gene pairs independently—unlike ML models, which require global modeling and thereby lose computational efficiency.

Although Chatterjee's coefficient has shown growing theoretical promise, current studies rely on the assumption of independent observations, which fails in the context of gene expression captured by scRNA-seq, where dependencies exist among cells due to shared biological origins, differentiation fates, and cellular cross-talk. To address this gap, our second contribution is an algorithm for estimating cutoff values for testing of dependency, based on synthetic or curated datasets with known ground truths. These cutoffs allow practical application of the method to real-world datasets, where we show that our approach consistently outperforms both traditional and machine learning models in recovering the true regulatory links. Finally, we use the asymmetric nature of Chatterjee's correlation to introduce a novel test for active regulation, determining whether a particular gene is functionally regulated by some other gene or vice versa. To the best of our knowledge, this directional test has not been addressed in prior literature, enabling us to construct biologically meaningful, directionally informed GRNs. Empirically, we demonstrate that our method consistently outperforms GRNBoost2 in detecting active gene regulation and delivers robust insights when applied to real single-cell datasets.

An alternative approach to constructing GRNs is via directed acyclic graphs (DAGs) [12, 13]. However, DAG learning suffers from identifiability and computational complexity due to the super-exponential growth of the DAG space with the number of genes [14]. In contrast, our framework focuses on pairwise relationships: we first identify dependent gene pairs through multiple testing procedures and subsequently test for active regulation at the pairwise level. This correlation-based strategy avoids the computational burden of global DAG search while still yielding biologically meaningful insights. Thus, while DAG-based approaches provide a principled framework for global causal discovery [15–18], our method offers a complementary perspective by prioritizing scalability and interpretability through pairwise testing. In this sense, our framework can also be utilized as a screening approach by efficiently filtering for potentially relevant gene pairs through pairwise testing, which can then be used as inputs for DAG-learning approaches, significantly narrowing the search space and thereby alleviating some of the computational and statistical burdens inherent to global DAG learning.

In the following sections, we introduce Chatterjee's correlation coefficient as a new approach for identifying gene-gene dependencies and active regulatory relationships. We first present a data-driven method for establishing cutoffs that effectively control false positive rates (FPRs), which is crucial for a fair comparison of our method against established correlation, regression, and machine learning-based approaches. We then evaluate Chatterjee's correlation against a broad spectrum of established methods for GRN construction. To make our method applicable to real-world scenarios, we propose a technique for estimating cutoffs from prior knowledge and demonstrate its transferability across different scRNA-seq datasets. Finally, we apply our approach to real single-cell data, constructing directed GRNs and accurately identifying active regulatory interactions. This work establishes a powerful tool for GRN inference that provides deeper insights into the complex regulations that govern cell fate decisions.

2. Methods

2.1. Acquisition and pre-processing of real scRNA-seq data

The real single-cell RNA sequencing (scRNA-seq) data were obtained from the Gene Expression Omnibus (GEO) database as detailed in Table 1. The data was pre-processed using the Seurat(v5) R package [19]. Quality control was performed on the raw counts. Cells were subsequently filtered to retain only those with more than 500 detected features (genes) and less than 10% mitochondrial reads. Following quality control, the data underwent

TABLE 1
scRNA-seq data sets derived from PBMCs.

	GEO Accession	Tissue Type	Cell Type	Total Cells	Total Genes
1	GSE239592	PBMC	T cells	5075	1345
2	GSM8331607	PBMC	T cells	1815	2884
3	GSM7873657	PBMC	T cells	5515	2982
4	GSM7873658	PBMC	T cells	4947	3188
5	GSM8683884	Small intestine	Epithelial cells	4017	8225
6	GSM8683885	Small intestine	Epithelial cells	3316	8435

normalization using the "LogNormalize" method with a scale factor of 10,000. Highly variable features were identified using "FindVariableFeatures", and the data was scaled using "ScaleData" using default parameters. Then, dimensionality reduction was performed by applying Principal Component Analysis (PCA) followed by Uniform Manifold Approximation and Projection (UMAP) on the first 15 principal components for visualization. Finally, cells were clustered using the Louvain algorithm with a resolution of 0.8. For the Peripheral Blood Mononuclear Cells (PBMC) datasets only, T cell clusters were extracted using high expression of three canonical gene markers: *CD3G*, *CD3E*, and *CD3D*. Genes expressed in fewer than 20% of the cells were then removed. The normalized gene expressions were stored as entries in a matrix, denoted as the gene expression matrix in the following subsection.

2.2. *Problem formulation*

Multiple testing problem We denote by $\mathbb{N}$ and $\mathbb{R}$ the set of natural and real numbers, respectively, and for any $n \in \mathbb{N}$, we denote $[n] = \{1, 2, \ldots, n\}$. Suppose that we have p many distinct genes and for any $j \in [p]$, the gene expression of the jth gene is a random variable, denoted by G_j. Due to the presence of some biological regulatory mechanism among the genes, there is naturally an underlying dependence structure between the variables $G_j, j \in [p]$. Thus, in order to infer about the gene regulatory network, it is crucial to correctly identify these dependence relationships. Moreover, the simplest yet one of the most efficient ways to comprehend any overall dependence structure is to study the degree of dependence in its fundamental units, which are the gene pairs, and classify them as either dependent or independent. Therefore, one of the popular approaches to formulate this classification problem is to consider multiple testing for their pairwise independence [20–23], which we formally state in the following. Specifically, for every $j, k \in [p]$ such that $j < k$, we consider the following test of hypothesis problem:

$$\mathsf{H}_0^{jk} : G_j \text{ and } G_k \text{ are independent} \qquad \text{vs.} \qquad \mathsf{H}_1^{jk} : G_j \text{ and } G_k \text{ are not independent.} \tag{1}$$

Furthermore, we consider the gene expressions of all p genes from n different cells which are basically n *possibly dependent* but identically distributed observations of the random vector $G = (G_1, G_2, \ldots, G_p) \in \mathbb{R}^p$, given by $G^{(i)} = (G_{i1}, G_{i2}, \ldots, G_{ip}), i \in [n]$. Moreover, we collect these observations as individual rows in our data matrix, also popularly known as the *gene expression matrix*, denoted by $\boldsymbol{G} \in \mathbb{R}^{n \times p}$, that is, the $(i, j)^{\text{th}}$ entry of $\boldsymbol{G}$ is represented by G_{ij}, $i \in [n], j \in [p]$.

Control of false positive rate (FPR) In the multiple testing problem described in (1), we consider controlling the metric of false positive rate, defined as the expected proportion of true nulls that are incorrectly rejected. Formally, FPR is the expected ratio of the number of false positives to the total number of true null hypotheses, that is,

$$\text{FPR} = E[R], \quad \text{where} \quad R := \frac{\text{no. of false positives}}{\text{no. of true nulls}}.$$

Moreover, our focus is to design a testing procedure such that the FPR is controlled below a certain pre-specified level, which is denoted by α and is typically chosen to be 0.05 or 0.01.

2.3. Dependency measures

There are numerous existing measures of dependence which can be implemented as test statistics to the multiple testing problem formulated in (1). In this work, our main focus is to propose a testing procedure that is based on *Chatterjee's rank-based correlation coefficient*, introduced in [9], and therefore, we first formalize this correlation coefficient in our context. Later, we provide the expressions of the two popular and classical correlation coefficients: *Pearson's correlation coefficient* and *Spearman's rank correlation*, as well as two machine learning assisted and more sophisticated dependency measures arising from the recently proposed *GRNBoost2* [6] and *PCNet* [7] algorithms. All the above measures are implemented to solve our testing problem, and we further compare their performances in the subsequent sections.

To begin with, fix any arbitrary $j, k \in [p]$, and consider the n bivariate (paired) gene expression data $(G_{ij}, G_{ik}),\ i \in [n]$ observed from the jth and k^{th} genes.

Chatterjee's rank-based correlation coefficient First, we rearrange the observations as

$$(G_{(1)j}, G_{(1)k}), (G_{(2)j}, G_{(2)k}), \ldots, (G_{(n)j}, G_{(n)k})$$

such that the observations of the jth gene are ordered in an increasing fashion as

$$G_{(1)j} \leq G_{(2)j} \leq \cdots \leq G_{(n)j}.$$

If G_{ij}'s and G_{ik}'s have no ties, then there is a unique way of the above ordering. Subsequently, for every $i \in [n]$, let r_i be the rank of $G_{(i)k}$, i.e., $r_i := |\{\ell : G_{(\ell)k} \leq G_{(i)k}\}|$. Then, Chatterjee's correlation coefficient is defined as, see [9, Section 1],

$$\xi_n(j \to k) := 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}. \tag{2}$$

In the presence of ties, the above coefficient is defined as follows. If there are ties among the G_{ij}'s, then choose an increasing rearrangement as above by breaking the ties uniformly at random. Furthermore, regarding the G_{ik}'s (with the possibility of ties), we additionally define the rank-type variable l_i's as $l_i := |\{\ell : G_{(\ell)k} \geq G_{(i)k}\}|$, and define the coefficient as

$$\xi_n(j \to k) := 1 - \frac{n\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{\sum_{i=1}^{n} l_i(n - l_i)}. \tag{3}$$

It is important to note that $\xi_n(j \to k)$ is not symmetric in j and k since we first order G_{ij}'s and then based on that assign the indices $(i)k$'s to G_{ik}'s and subsequently obtain the ranks of $G_{(i)k}$'s denoted as r_i's. In case we swap j and k in the above operations, that is, start by ordering G_{ik}'s, then the resulting coefficient is denoted as $\xi_n(k \to j)$. Furthermore, to obtain a symmetric coefficient, we compute the symmetrized statistic by considering the maximum of $\xi_n(j \to k)$ and $\xi_n(k \to j)$, and denote it as $\xi_n(j, k)$, i.e.,

$$\xi_n(j, k) := \max\{\xi_n(j \to k),\ \xi_n(k \to j)\}. \tag{4}$$

The above coefficient provides us with a measure of dependence between the variables G_j and G_k, thereby estimating the degree of association or regulation between the jth and k^{th} genes. To be specific, in an asymptotic sense, when the sample size n is large enough and the cells are independent and identically distributed (iid), then $\xi_n(j, k)$ is close to 0 if the jth and k^{th} genes are independent, i.e., there are no regulations between them, and it is closer to 1, otherwise [9, Theorem 1.1]. In fact, the strength of regulation is associated with how close it is to 1, and therefore it not only serves as an estimate of the dependence, but is also useful as a test statistic for testing of independence, stated in problem (1), which we describe in the next section.

Now, we consider two traditional and well-known measures of correlation in the following, which have been extensively used in the literature for the past decades.

Pearson's correlation coefficient Pearson's correlation coefficient is defined as

$$\rho_n(j, k) := \frac{\sum_{i=1}^{n}(G_{ij} - \bar{G}_j)(G_{ik} - \bar{G}_k)}{\sqrt{\sum_{i=1}^{n}(G_{ij} - \bar{G}_j)^2}\sqrt{\sum_{i=1}^{n}(G_{ik} - \bar{G}_k)^2}}, \qquad \text{where} \quad \bar{G}_j = \frac{1}{n}\sum_{i=1}^{n} G_{ij} \ \text{ and } \ \bar{G}_k = \frac{1}{n}\sum_{i=1}^{n} G_{ik}.$$

Spearman's rank correlation coefficient First, for every $i \in [n]$, we denote by r_{ij} and r_{ik} the ranks of G_{ij} and G_{ik} among their corresponding set of observations, respectively, i.e.,

$$r_{ij} := |\{\ell \in [n] : G_{\ell j} \leq G_{ij}\}| \qquad \text{and} \qquad r_{ik} := |\{\ell \in [n] : G_{\ell k} \leq G_{ik}\}|.$$

In the presence of ties, instead of assigning them consecutive ranks, we assign them the average of the ranks they would have otherwise received. Then, Spearman's correlation coefficient is defined as

$$\psi_n(j,k) := 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)}, \qquad \text{where} \quad d_i = (r_{ij} - r_{ik}),\ i \in [n].$$

Next, we present the dependency measures that are developed in two recently proposed methods that are primarily focused on the construction of GRNs based on more advanced and complex machine learning tools.

GRNBoost2 In the case of GRNBoost2, we consider the dependency measure,

$$\gamma_n(j,k) := (j,k)^{\text{th}} \text{ entry of } W,$$

where W is the *symmetrized weight matrix* associated with the gene regulatory network that is constructed using the tree-based ensemble methods, as specified in [6].

PCNet In the case of PCNet, we consider the dependency measure,

$$\pi_n(j,k) := (j,k)^{\text{th}} \text{ entry of } W,$$

where W is the *symmetrized adjacency matrix* derived in the network construction and manifold alignment steps, as described in [7].

2.4. *Proposed testing procedure for identifying dependencies*

In this section, we develop a computationally simple and easily interpretable testing procedure that can practically utilize any of the dependency measures stated in the previous section. First, we consider Chatterjee's coefficient to design our procedure since it is the main focus of our paper, and simultaneously develop the rationale behind such a testing procedure, which will motivate us to further apply it more generally based on other dependency measures.

Testing method It has been shown in [9, Theorem 2.1] that, when G_j and G_k are independent, i.e., H_0^{jk} is true, and the observations from the cells are iid, then both $\sqrt{n}\,\xi_n(j \to k)$ and $\sqrt{n}\,\xi_n(k \to j)$ are asymptotically distributed as Normal with mean 0 and variance $2/5$, i.e.,

$$\sqrt{n}\,\xi_n(j \to k),\ \sqrt{n}\,\xi_n(k \to j) \qquad \to \ N(0, 2/5) \qquad \text{in distribution as } n \to \infty.$$

It is noteworthy that the limiting distribution is not only symmetric but also independent of the true underlying distributions of G_j and G_k, which are unknown. Thus, the limiting distribution of the symmetrized statistic $\sqrt{n}\,\xi_n(j,k)$ must also adopt such a favorable property, but its specific parameter values are yet to be characterized in the statistical literature even under the scenario that the cell observations are assumed to be iid.

Nevertheless, in the spirit of the above results, to leverage the transparency and the distribution-free nature of the test statistics, we propose the testing procedure that,

$$\text{for every } j,k \in [p],\ j < k, \qquad \text{reject the null hypothesis} \quad \mathsf{H}_0^{jk} \ \text{ if } \ |\xi_n(j,k)| > C_{\xi,n}, \tag{5}$$

where $C_{\xi,n}$ is some positive testing cutoff which must be determined appropriately, such that the FPR is controlled by the required level. However, since any rigorous characterization of the limiting behavior or at least the limiting variance of $\xi_n(j,k)$ is yet absent in the existing literature, and more importantly, the cell observations are quite possibly dependent in practice, we need to estimate $C_{\xi,n}$ separately from synthetic or pre-existing gene expression databases, where it is a priori known which gene pairs are independent. Subsequently, we use the estimated cutoff to apply the testing procedures on real datasets to identify the independent gene pairs, as shown in Section 3.3.

In the same vein as in (5), we adopt a general testing policy which can incorporate the other dependency measures, resulting in the four testing procedures listed below,

$$\text{(6)} \qquad \text{for every } j,k \in [p],\ j<k, \qquad \text{reject } \mathsf{H}_0^{jk} \qquad \text{if} \quad |y_n(j,k)| > C_{y,n}, \qquad y \in \{\rho,\psi,\gamma,\pi\},$$

where the testing cutoffs $C_{y,n},\ y \in \{\rho,\psi,\gamma,\pi\}$, similarly to $C_{\xi,n}$, depend on the number of observations n and the corresponding limiting distributions of the coefficients. Although for the traditional measures, there are relevant asymptotic theories regarding their limiting distributions, their validity primarily depends on the crucial assumption that the cell observations must be iid, which does not hold in practice. Therefore, instead of using asymptotics or theoretical results, these testing cutoffs are similarly estimated from data, not only to avoid any error due to asymptotic approximations but also to perform a fair and more precise comparison with the procedure proposed in (5), as shown later in Section 3.1.

2.5. *Testing for active regulation*

In this section, we develop a novel testing framework that further incorporates whether there is any *active regulation* present between two dependent genes. To be specific, for every $j,k \in [p]$, such that the jth and k^{th} genes are dependent, we say that G_k is actively regulated by G_j if we can express G_k as $f_{kj}(G_j)$ plus some possible, low-fluctuating noise term, for some nontrivial function $f_{kj}(\cdot)$. In other words, G_j serves as an independent variable whereas G_k is dependent on G_j through some regulation function $f_{kj}(\cdot)$. Moreover, if either of them is actively regulated by the other, we say that there is an active regulation present between these two genes. However, it is important to emphasize here that G_j and G_k can still be dependent without any active regulation between them, for example, when both are actively regulated simultaneously (co-regulated) by the expression of some third gene. We illustrate the notion of active regulation in Fig. 2a, clearly distinguishing them from the cases of mere dependence.

Multiple testing problem Formally, in this context, we consider the following test of hypothesis problem: for every $j,k \in [p]$, $j<k$ such that the G_j and G_k are dependent,

$$\text{(7)} \qquad \bar{\mathsf{H}}_0^{jk}:\ \text{No active regulation between } G_j \text{ and } G_k \qquad \text{vs.} \qquad \bar{\mathsf{H}}_1^{jk}:\ G_j \text{ and } G_k \text{ are actively regulated.}$$

In this way, in addition to testing for dependence, it is of further interest to check whether one gene actively regulates the other, and to the best of our knowledge, formulation of such a problem has not yet been considered in the past literature.

Testing procedure To propose a solution to the testing problem in (7), first, it is important to consider a dependency measure that can suitably capture the notion of active regulation. More specifically, for every $j,k \in [p]$ such that G_j and G_k are dependent, we must measure the extent G_k is actively regulated by G_j, i.e., to what degree G_k can be expressed in terms of G_j through some regulation function. The construction of Chatterjee's correlation coefficient $\xi_n(j \to k)$, as defined in (2), or (3), suggests that it can be utilized as an appropriate measure in this context, and in fact, following [9, Theorem 1.1] it converges to 1 if and only if $G_k = f_{kj}(G_j)$ for some measurable function $f_{kj}(\cdot)$, i.e., G_k is actively regulated by G_j. Therefore, if $\xi_n(k \to j)$ is significantly larger than $\xi_n(j \to k)$, it serves as strong evidence that G_k is actively regulated by G_j whereas the degree of opposite regulation is comparatively weak. In fact, if both $\xi_n(k \to j)$ and $\xi_n(j \to k)$ are simultaneously large, i.e., close to each other, then it suggests that they are co-regulated by some third gene contributing to higher values in both directions. Thus, based on the above discussion and following the rationales stated in Section 2.4, we propose the following testing procedure,

$$\text{(8)} \qquad \text{for every } j,k \in [p],\ j<k, \quad \text{reject} \quad \bar{\mathsf{H}}_0^{jk} \quad \text{if} \qquad \bar{\xi}_n(j,k) := \Big|\xi_n(j \to k) - \xi_n(k \to j)\Big| > C_{\bar{\xi},n},$$

where $C_{\bar{\xi},n}$ is some positive constant, which must be appropriately determined such that the FPR is controlled by the required level. However, since any characterization of the limiting behavior of $(\xi_n(j \to k) - \xi_n(k \to j))$ is not present in the theoretical literature even for iid observations, we need to estimate $C_{\bar{\xi},n}$ from simulated gene expression datasets, where the active regulations among the gene pairs are a priori specified.

It is noteworthy that GRNBoost2 also supplies us with a directed (asymmetric) dependency measure provided by the elements of the asymmetric importance matrix W evaluated at the tree construction step. Therefore, in a similar vein, we propose the following testing procedure,

$$\text{(9)} \qquad \text{for every } j,k \in [p],\ j<k, \quad \text{reject} \quad \bar{\mathsf{H}}_0^{jk} \quad \text{if} \qquad \bar{\gamma}_n(j,k) := \Big|\gamma_n(j \to k) - \gamma_n(k \to j)\Big| > C_{\bar{\gamma},n},$$

which we later compare with (8) in Section 3.2.

2.6. *Data-driven calculation of the testing cutoff*

In the last two sections, we proposed several testing procedures, described in (5), (6), (8), and (9), for which, as previously indicated, it is necessary to determine the associated testing cutoffs in a data-driven way. In addition, it is important to emphasize that in our experiments we observed that using theoretical or asymptotic cutoffs, based on the impractical assumption of independence between the cell observations, may lead us to false results by either rejecting a significant proportion of true null hypotheses (anti-conservative or invalid tests) or not rejecting true alternatives (overly conservative and low-powered tests). To address this, in this section, we propose a general algorithm to estimate the cutoff for any such testing procedure based on some existing or simulated dataset, where it is known a priori the true null hypotheses.

First, fix any testing procedure that is supposed to be constructed based on n cell observations, and for which the cutoff needs to be estimated. In this regard, we consider some pre-existing or synthetic gene expression dataset, in which the number of cells is larger than n, and it is known a priori the knowledge of all true null hypotheses, for example, the identity of all independent gene pairs in problem (1), or identity of actively regulated pairs in problem (7). Furthermore, we denote the total number of cells by $n' \geq n$, and the total number of true nulls by N_0. Now, we follow the steps below.

1. Resample from the dataset a subsample of size n. In case n is reasonably smaller than n', we resample without replacement; otherwise, especially when $n = n'$, we resample with replacement, as in the case of Bootstrap.
2. Fix some $C > 0$ as the testing cutoff and apply the testing procedure to the subsample. Moreover, let the total number of resulting false positives be denoted as $N_{\text{fp}}(C)$.
3. Record the following ratio:

$$R(C) := \frac{N_{\text{fp}}(C)}{N_0},$$

   which clearly serves as an unbiased estimate of the FPR.
4. Since we desire to control the FPR below level α, find the maximum value as the testing cutoff C for which $R(C)$ is kept below α, and denote it as the optimal testing cutoff $\tilde{C}$, i.e.,

$$\tilde{C} := \max\{C : R(C) \leq \alpha\}.$$

   A grid search or some other suitable method is typically considered to find the above maxima.
5. Repeat the above steps 1-4 for B many iterations, and for every $b \in [B]$, specifically denote the optimal testing cutoff, estimated in step 4, as $\tilde{C}_b$.
6. Denote by C_* the median of all calculated optimal cutoffs $\tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_B$, i.e.,

$$C_* := \text{median}\{\tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_B\}.$$

   Consider C_* as the final testing cutoff, which clearly serves as an estimate of the true cutoff value. It is important to emphasize that the median is particularly chosen in the above to obtain a more robust estimate of the testing cutoff. More specifically, it prevents us from accidentally selecting a very large or small value of the final cutoff, which may occur as we estimate it purely from the observational data.

We formalize the above method in Algorithm 1, described below.

Algorithm 1 Cutoff estimation method for any testing procedure with sample size n

```
Consider a dataset with sample size n' ≥ n
Set  B ← desired no. of resampling iteration, e.g., typically 50 or 100
Set  α ← required level of FPR, e.g., typically 0.05 or 0.01
Set  N_0 ← set of all true null hypotheses
Set  N_0 ← size of N_0
Set  b ← 1
Set  c.val ← an increasing grid of possible cutoff values, e.g., the sequence (start: 0, end: 2, gap: 0.01)
Set  C̃ ← a vector of size B with all element being 0
    while b ≤ B do
        if n' < n then
            take a subsample of size n without replacement
        else take a subsample of size n with replacement (Bootstrap)
        end if
        ℓ ← 0
        N_fp ← 0
        while (N_fp/N_0) ≤ α do
            ℓ ← ℓ + 1
            C ← cval_ℓ        (Comment: the ℓ^th value of cval is denoted by cval_ℓ)
            j ← 1
            while j ≤ p − 1 do
                k ← j + 1
                while k ≤ p do
                    calculate y_n(j, k)        (Comment: in this work we have y ∈ {ξ, ρ, ψ, γ, π, ξ̄, γ̄})
                    if |y_n(j, k)| > C and (j, k) ∈ N_0 then
                        N_fp ← N_fp + 1
                    end if
                    k ← k + 1
                end while
                j ← j + 1
            end while
        end while
        C̃_b ← cval_{ℓ−1}        (Comment: storing the estimated cutoff as the b^th value of C̃)
        b ← b + 1
    end while
    return  C_{y,n} = median of {C̃_1, C̃_2, ..., C̃_B}
```

2.7. *Evaluation of testing cutoffs to apply on real data*

To apply our proposed testing methods to identify gene-gene dependencies in a target gene-expression dataset, we first need to determine appropriate testing cutoffs. This is particularly crucial because the inherent cell-cell signaling and other biological interactions create a dependency among cells (samples), which violates the assumption of independence upon which theoretical or asymptotic cutoffs are based in the literature. Thus, cutoffs must be estimated based on pre-existing databases, if available, or synthetic datasets, otherwise. However, since the nature of cell dependencies varies across different biological scenarios, a single set of cutoffs isn't universally applicable. Therefore, we propose a context-specific approach. This means that whenever a relevant database is available, we tailor the cutoffs to the specific biological context of the dataset being analyzed.

Testing for identifying dependency For the proposed procedures in (5) and (6), which consider testing of independence through the problem in (1), we consider the human protein-protein interaction (PPi) network from the STRING database [24], a comprehensive collection of text-mined and experimentally determined functional associations, as an estimated ground truth. We kept only interactions with a combined score greater than 400.

Given that the number of cells in the target dataset is n, we chose a published reference dataset from the GEO database that is of size $n' \geq n$, which consists of cells from the same tissue in the same experimental condition as

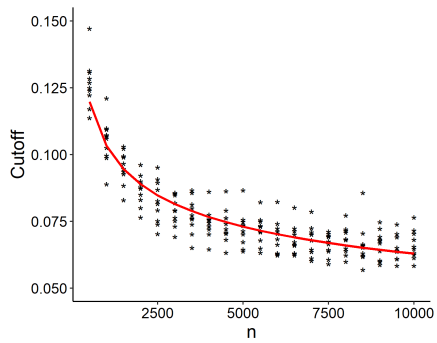

FIG 1. *Cutoff selection function for active regulation*

the target dataset. Since the cutoff for Chatterjee's correlation is dependent on n, we estimated the cutoffs using Algorithm 1, over $B = 50$ iterations. For our specific testing case, we randomly sampled 1800 cells from the reference dataset and selected 1080 cells for training and 720 cells for testing. We randomly sampled 720 cells from all the target datasets as well. We also randomly selected 300 genes that were expressed across all datasets under consideration. The computed cutoff from the reference dataset was then applied to the target dataset to identify the significantly dependent gene pairs.

Testing for active regulation Since, to the best of our knowledge, there is no database available in real-world settings with the ground truth of active regulations being clearly specified, we prescribe a cutoff evaluation function using simulated datasets. To be specific, we consider several choices for the number of cell observations, as $n \in \{500k : k \in [20]\}$. Furthermore, for each n, we further simulate 10 separate datasets in each of which active regulations between a select gene pairs are imposed through four representative functions, presented in Section 3.1.1, and the cutoff $\bar{C}_{\xi,n}$ is estimated following Algorithm 1 with $n' = n$ and $B = 50$. Since in the statistical literature the testing cutoffs are generally some decreasing function of n, specifically of $O(n^{-\alpha})$ for some $\alpha > 0$, we accordingly fit the curve $g_\xi(n) = c_1 n^{-c_2}$, where $c_1, c_2 > 0$, to all the estimates of $\bar{C}_{\xi,n}$ (shown as the points in Fig. 1) by means of least squares. As a result, we obtain estimates of c_1 and c_2 that prompts us to use $\hat{g}_\xi(n)$ as the cutoff $\bar{C}_{\xi,n}$, where

$$\hat{g}_\xi(n) = 0.455\, n^{-0.215} \tag{10}$$

that is marked in red in Fig. 1. We use the above in Section 3.3 to evaluate the cutoff for the testing procedure (8) to apply it on real data. Similarly, we estimate the corresponding function to be used to evaluate the cutoff for (9), employed in the comparative study of Section 3.2.

2.8. *Performance metrics*

For any testing procedure, we evaluate its testing and learning performances in terms of some standard performance metrics. Specifically, we evaluate their testing performance by using the following performance metrics:

$$\text{true positive rate:} \qquad \text{TPR} = \frac{\text{no. of true positives}}{\text{no. of true alternatives}},$$

$$\text{true negative rate:} \qquad \text{TNR} = \frac{\text{no. of true negatives}}{\text{no. of true nulls}},$$

$$\text{balanced accuracy:} \qquad \text{BA} = \frac{\text{TPR} + \text{TNR}}{2}.$$

Besides, we evaluate the performance of their classification by using AUROC, the area under the ROC curve that plots TPR vs FPR. Furthermore, we simultaneously determine their learning performance by computing

$$\text{root mean square error:} \qquad \text{RMSE} = \sqrt{\frac{2}{p(p-1)} \sum_{1 \le j < k \le p} (\tilde{y}_n(j,k) - \mathbb{1}(j,k))^2},$$

where $y \in \{\xi, \rho, \psi, \gamma, \pi, \bar{\xi}, \bar{\gamma}\}$, and $\tilde{y}_n(j,k)$ is the scaled version of $y_n(j,k)$ by letting

$$\tilde{y}_n(j,k) = |y_n(j,k)| / \max_{j,k} |y_n(j,k)|,$$

that is, $\tilde{y}_n(j,k) \in [0,1]$. Accordingly, if testing dependency, that is, $y \notin \{\bar{\xi}, \bar{\gamma}\}$, $\mathbb{1}(j,k)$ denotes the indicator function that takes value 1 if and only if j^{th} and k^{th} genes are not independent, i.e., $\mathbb{1}(j,k) := \mathbb{1}\{G_j \text{ and } G_k \text{ are not independent}\}$, and otherwise, we let $\mathbb{1}(j,k) := \mathbb{1}\{G_j \text{ and } G_k \text{ are actively regulated}\}$.

3. Results

3.1. Chatterjee's Correlation Accurately Infers Gene Dependencies in Complex Simulated Data

The high cellular resolution of gene expression captured by scRNA-seq is particularly advantageous for reverse engineering gene networks based on co-expression patterns. Consequently, scRNA-seq enables the investigation of gene co-expression networks on an intra-sample basis. The substantial number of cells yielded by scRNA-seq, as well as diverse patterns of gene-gene interaction, favors the application of non-parametric statistical coefficients, such as Chatterjee's correlation coefficient, and we investigate its efficiency in the following studies. Specifically, we conduct two separate studies: first, we perform our experiments on simulated data, and second, we apply our methods to real scRNA-seq data.

3.1.1. First study: experiments on synthetic data

To comprehensively evaluate the performance of the proposed methods under a range of different gene interaction patterns, we simulated scRNA-seq data incorporating both linear and non-linear gene interactions, as they can exhibit both types of relationships. Our simulation employed a Poisson approximation and accounted for characteristic scRNA-seq noise, including Gaussian noise and expression dropouts, and four distinct types of gene-gene interactions are employed, represented in Fig. 2a.

Simulation of scRNA-seq data with gene-gene interactions We simulated scRNA-seq count data with predefined gene-gene interactions, including both linear and non-linear relationships. For each simulated dataset, we generated data for 100 genes and 1000 cells. Gene expression was modeled using a Poisson approximation.

Specifically, for each gene j and cell i, a theoretical mean expression level was calculated as the product of a gene-specific mean and a cell-specific size factor. Gene-specific means were drawn from a Gamma distribution with shape parameter $1/\sigma$ and scale parameter $\mu \times \sigma$, where μ was set to 15, and σ was set to 0.5. Here μ and σ represent the mean expression and dispersion of gene j, respectively. Cell-specific size factors were drawn from a Gamma distribution with shape 1 and scale 1. To approximate Poisson-distributed counts, we used Bernoulli trials via the binomial distribution. The number of trials was set to a maximum of 1000, and ten times the theoretical mean expression, and the probability of success was set to the theoretical mean divided by the number of trials. This approach approximates a Poisson distribution. A dropout rate of 0.3 was introduced by randomly setting counts to zero with a probability of 0.3.

To introduce gene-gene interactions, we selected 50 gene pairs, each of which consists of a target gene and a source gene. The expression of the target gene was determined by the expression of the source gene as

$$\max\{f(x) + \epsilon,\ 0\},$$

where x represents gene expression counts of the source gene, and ϵ represents a normally distributed noise variable with a standard deviation of 0.2. For the signal function $f(\cdot)$ we consider the following four forms:

1. **Linear**: $f(x) = 2x$
2. **Parabolic**: $f(x) = -0.02\,x^2 + x + 5$
3. **Exponential**: $f(x) = e^{5x/\theta}, \quad \theta =$ maximum expression of the source gene
4. **Sinusoidal**: $f(x) = 5\sin(x/15) + 10$

After applying the gene relationship through the above functions, the resulting expression values for the target gene were converted back to count data using the same Poisson approximation method described above. The gene-gene interactions thus simulated are depicted in Fig. 2a. Ten independent datasets were simulated, each with a unique random seed. Linear and exponential/log relationships were classified as non-directed (no active regulation), while parabolic and sinusoidal relationships were classified as directed (active regulation).

Experiments and results To begin with, we randomly divide each dataset into training and test sets, and use the training part to estimate the testing cutoffs by Algorithm 1 for all testing procedures stated in (5) and (6) such that they all control the same desired level of FPR when applied on the test data. To be specific, we let the number of cells in the training data be $n_{\text{train}} = 600$ and that in the test data be $n_{\text{test}} = 400$. Subsequently, we apply Algorithm 1 with $n' = n_{\text{train}}$, $n = n_{\text{test}}$, $B = 50$ and $\alpha = 0.05$, to estimate the cutoffs.

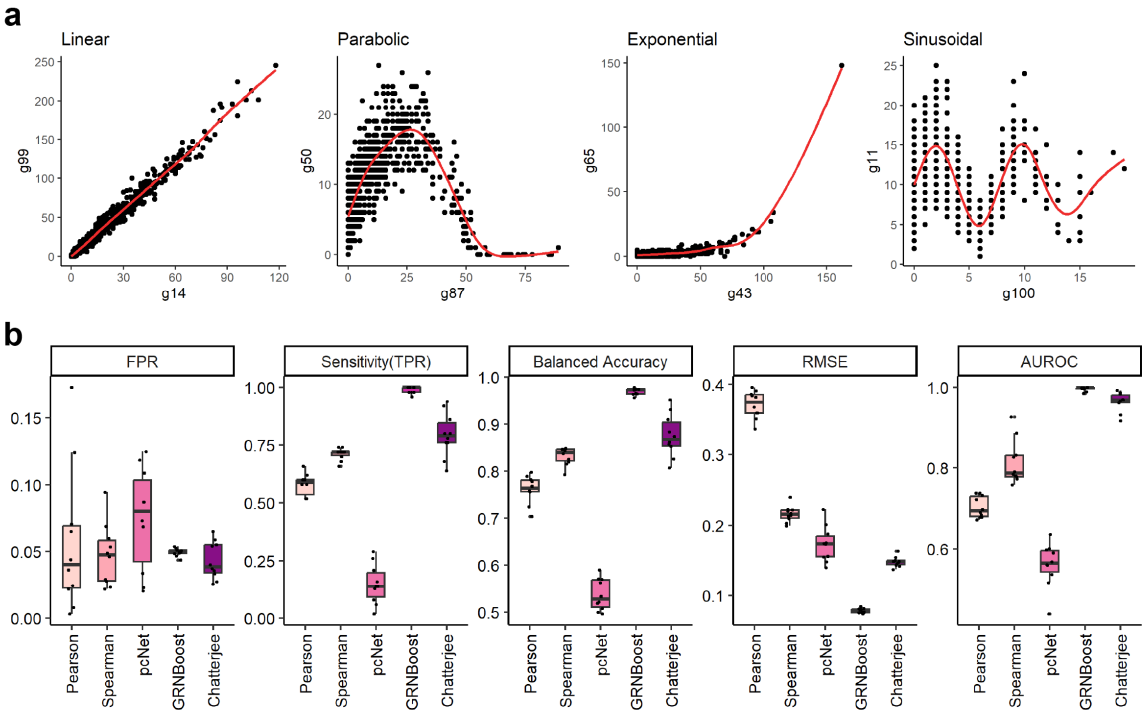

FIG 2. *The plots in* (a) *describe various dependence patterns in Study 1, and the performances are summarized in* (b).

Then we applied the computed cutoff to the test data. Our proposed cutoff estimation procedure demonstrated effective control of FPR, maintaining it at approximately 0.05 across all evaluated methods, see Fig. 2b. This provides a robust foundation for comparing performance metrics.

For sensitivity and balanced accuracy, the performance of Chatterjee's correlation was higher than both traditional correlation measures (Pearson, Spearman) and the PC regression-based PCNet, and in fact, competitive with random-forest-based GRNBoost2 (Supplementary Table 1). In terms of RMSE, the proposed method significantly outperformed both traditional correlation measures (Pearson, Spearman) and the PC regression-based PCNet. However, GRNBoost exhibited a lower RMSE, which is likely because Chatterjee's correlation assesses local, pair-wise relationships, while GRNBoost benefits from incorporating broader data context.

Crucially, when evaluating the AUROC, the performance of the proposed method was comparable to that of GRNBoost, with differences on the order of 10^{-2}. These results demonstrate that a computationally efficient, rank-based dependency measure can not only provide a significant improvement over traditional correlation metrics but also achieve performance levels on par with more complex and computationally intensive learning-based approaches.

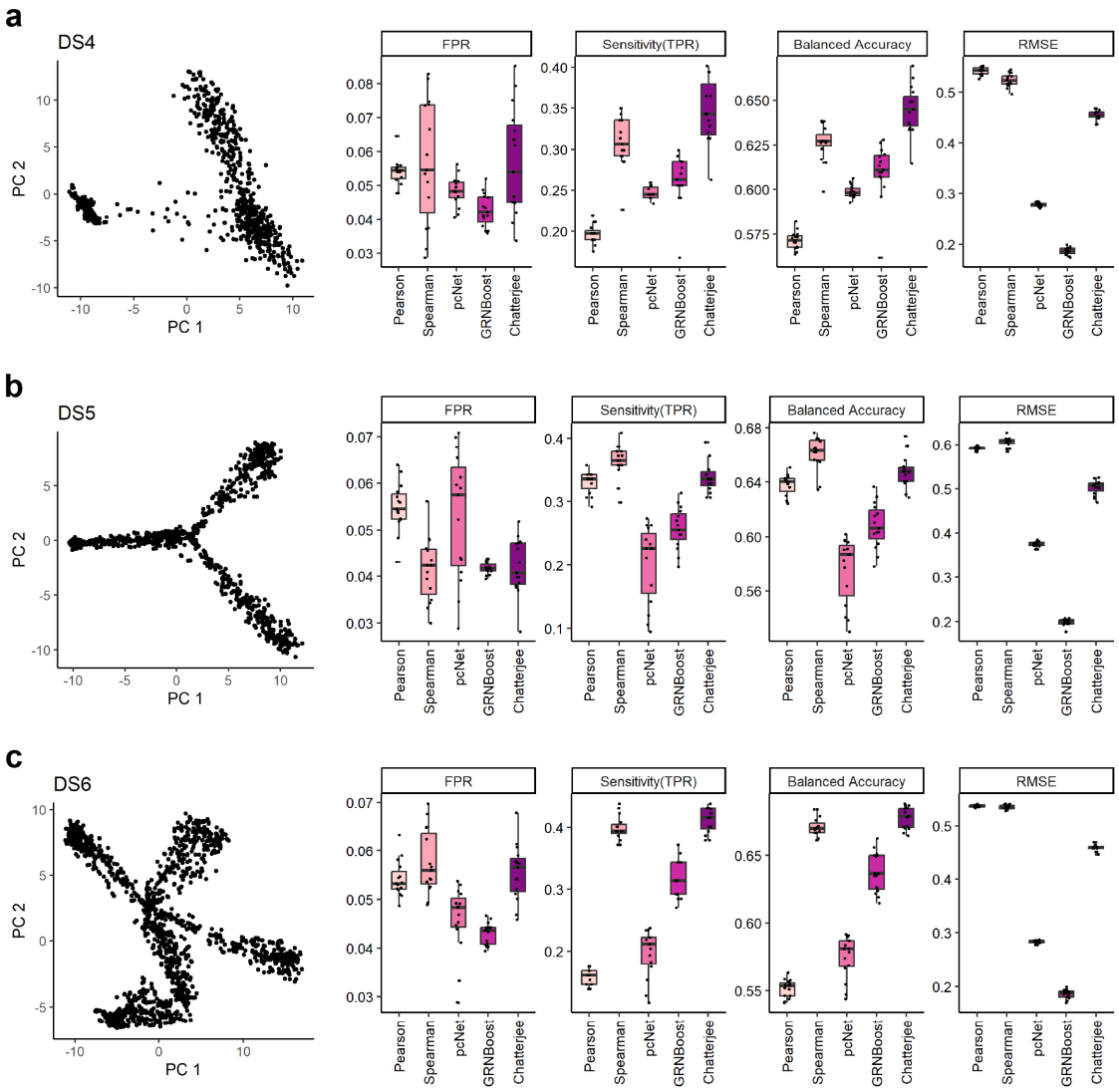

FIG 3. *Benchmark of GRN construction methods in simulated scRNA-seq datasets using SERGIO. Each row presents the analysis of a simulated scRNA-seq dataset: the left plot displays the variability explained by the first two principal components, while the right plot summarizes method performance.*

3.1.2. Second study: experiments on SERGIO simulated data

For this study, we used datasets generated by SERGIO [25], a single-cell gene expression simulator guided by GRNs. This choice was made because it allowed us to obtain a ground-truth GRN that is essential for both cutoff selection and subsequent performance testing, and also allowed us to simulate cell differentiation trajectories, a feature of real scRNA-seq data. Following a similar approach to our previous study, we first applied Algorithm 1 to select the testing cutoffs for all five GRN construction methods under consideration. We then evaluated their testing and learning performances through a rigorous comparative study.

As shown in Fig. 3, our data-driven cutoff estimation procedure successfully controlled FPR at approximately 0.05 for all methods, enabling a ground for fair comparative analysis (Supplementary Tables 2,3,4).

In DS5 (Fig. 3b), where the cell trajectory was largely linear, traditional measures performed well in terms of balanced accuracy as they rely on the linearity assumptions. Nevertheless, the sensitivity and balanced accuracy of the procedure with Chatterjee's correlation are comparable. However, in datasets with complex non-linear

patterns, such as DS4 and DS6 (Fig. 3a,c), Chatterjee's correlation-based method consistently exhibited the highest balanced accuracy. Notably, in the most complex non-linear cell trajectory case (DS6), its performance was significantly superior to both GRNBoost and PCNet. These results indicate that Chatterjee's correlation even outperforms more computationally expensive learning-based approaches for identifying complex dependencies.

Regarding RMSE, our method showed lower values than traditional correlation-based approaches. As expected, ML-based methods like PCNet and GRNBoost, which leverage global data, outperformed the local, pair-wise inference methods. Overall, the sensitivity and balanced accuracy of our procedure were consistently superior to other methods across most of the tested scenarios.

3.1.3. Remarks

From the above studies, we conclude that, for identifying the dependent gene pairs to decipher the underlying gene regulatory network, the proposed multiple-testing procedure based on Chatterjee's correlation coefficient can perform significantly better than the traditional and commonly used measures. Furthermore, with this dependency measure, it is possible to perform the testing with much lower computational cost than the more recent, complex, and learning-based methods, without even compromising on the accuracy of classification. In fact, in most cases, the performance is better, or at least as good as the existing methods, to a significant extent. Therefore, in the rest of the paper, we focus on the testing procedure (5) based on Chatterjee's correlation coefficient, and apply it on a real gene expression dataset in Section 3.3 to recover the underlying gene regulatory network for further biological discoveries.

3.2. Chatterjee's correlation accurately identifies active regulations in simulated scRNA-seq data

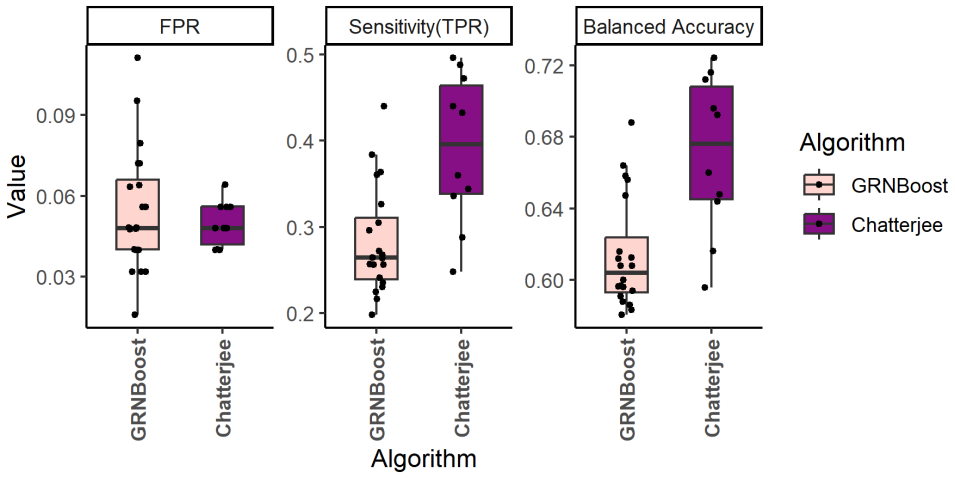

FIG 4. *Chatterjee's correlation accurately identifies actively regulated gene pairs, outperforming GRNBoost2.*

In order to apply the proposed testing procedures (8) and (9) for active regulation, and to rigorously assess and compare their efficiency, we simulated synthetic scRNA-seq count data with predefined gene-gene interactions, following a similar approach previously presented in Section 3.1.1. However, this simulation created datasets with 500 genes and 1000 cells, including 250 labeled gene-gene interactions. These interactions were classified as either "no active regulation" or "active regulation," and encompassed both linear and non-linear relationships.

For each of the simulated datasets, we randomly partitioned the cells into a training set of size $n_{\text{train}} = 600$ and a test set of size $n_{\text{test}} = 400$, composed of the remaining cells. The training set, as usual, was used to determine the cutoff values for the testing procedure stated in (8) and (9) for detecting active regulation by Algorithm 1,

with $n' = n_{\text{train}}$, $n = n_{\text{test}}$, $B = 100$ and $\alpha = 0.05$, on the known dependent interactions defined during the simulation. The test set was used for the subsequent assessment and comparison of the methods.

Next, we applied the computed cutoff to the asymmetric GRNs obtained from the test set using both GRNBoost and Chatterjee's correlation. The active regulations were identified by applying their respective cutoffs, which were previously computed. Our proposed cutoff estimation procedure demonstrated effective control of the FPR, maintaining it at approximately 0.05 across all evaluated methods, see Fig. 4, providing a robust and impartial foundation for comparing the procedures (Supplementary Table 5). Furthermore, our analysis discovers a remarkable phenomenon that, in terms of both sensitivity and balanced accuracy, the performance of Chatterjee's correlation was significantly higher than GRNBoost2. This strongly demonstrates that, in addition to enjoying computational advantages due to its rank-based construction, Chatterjee's correlation performs significantly better than more intricate and computationally expensive learning-based approaches in discovering active regulation among the dependent genes. This arises from the fact that Chatterjee's correlation leverages the ranks of the observations to directly assess whether, under sufficient sample size, one gene can be expressed as a non-constant function of another, for example, see [9, Theorem 1.1] in the case of iid observations. In contrast, less transparent machine learning methods typically seek to learn the global dependency structure of all variables simultaneously, which obscures local functional relationships and thus limits their ability to capture gene regulation at an atomic or mechanistic level. Based on the above analysis, we only focus on Chatterjee's correlation for testing active regulation between the genes while applying our methods to real data.

3.3. Application on real data

Data set Given the dynamic and context-specific nature of biological systems, we investigated the transferability of our proposed cutoff selection method across similar biological datasets from distinct donors and experimental conditions. For this purpose, we utilized eight different scRNA-seq gene expression datasets. Specifically, six of the datasets were sourced from PBMCs, while two were derived from small intestine epithelial cells, as detailed in Table 1. The respective GRNs for these datasets were constructed using Chatterjee's correlation, as previously described. To establish a ground truth for our validation, we obtained an estimated network of protein-protein interactions from the STRING database [26], version 12.

Testing method Our initial objective was to classify gene pairs as either independent or dependent. To achieve this, we frame the problem as a multiple hypothesis testing scenario, as described in (1). We employ Chatterjee's correlation coefficient to implement the testing procedure detailed in (5). For our analysis, we selected 300 genes that were expressed in at least 20% of the cells across all six datasets and retained only these for subsequent testing (Supplementary Table 6).

To estimate the optimal cutoff value, we used dataset GSE239592, with a significance level of 0.05, over 50 iterations. In each iteration, we sampled 1,800 cells and then partitioned this subset into a training set (60%) and a test set (40%). The training data, consisting of 1080 cells, was sampled iteratively to 720 cells to determine the respective cutoff. The final step involved applying this cutoff, derived from dataset GSE239592, to classify dependent interactions in the GRNs from the remaining PBMC datasets.

Results Our analysis of the healthy human PBMC datasets revealed the robustness and transferability of the proposed cutoff selection method. As shown in Fig. 5a, we first trained our cutoff selection function on the GSE239592 dataset. This selection successfully controlled FPR in the corresponding test set, maintaining it at a level of approximately 0.05. A key finding was the successful transferability of this cutoff across multiple datasets. The cutoff determined from GSE239592 was applied to three additional PBMC datasets (GSM8331607, GSM7873657, and GSM7873658), and in all cases, it effectively maintained the FPR near our desired threshold of 0.05.

Furthermore, we assessed the relationship between our selected cutoff and optimal performance, as measured by the F1 score. As depicted in Fig. 5b, for each dataset, our chosen cutoff value is consistently located at or very near the peak of the F1-score curve. This demonstrates that our method not only accurately controls the FPR but also identifies a cutoff that maximizes the balance between precision and recall, effectively capturing

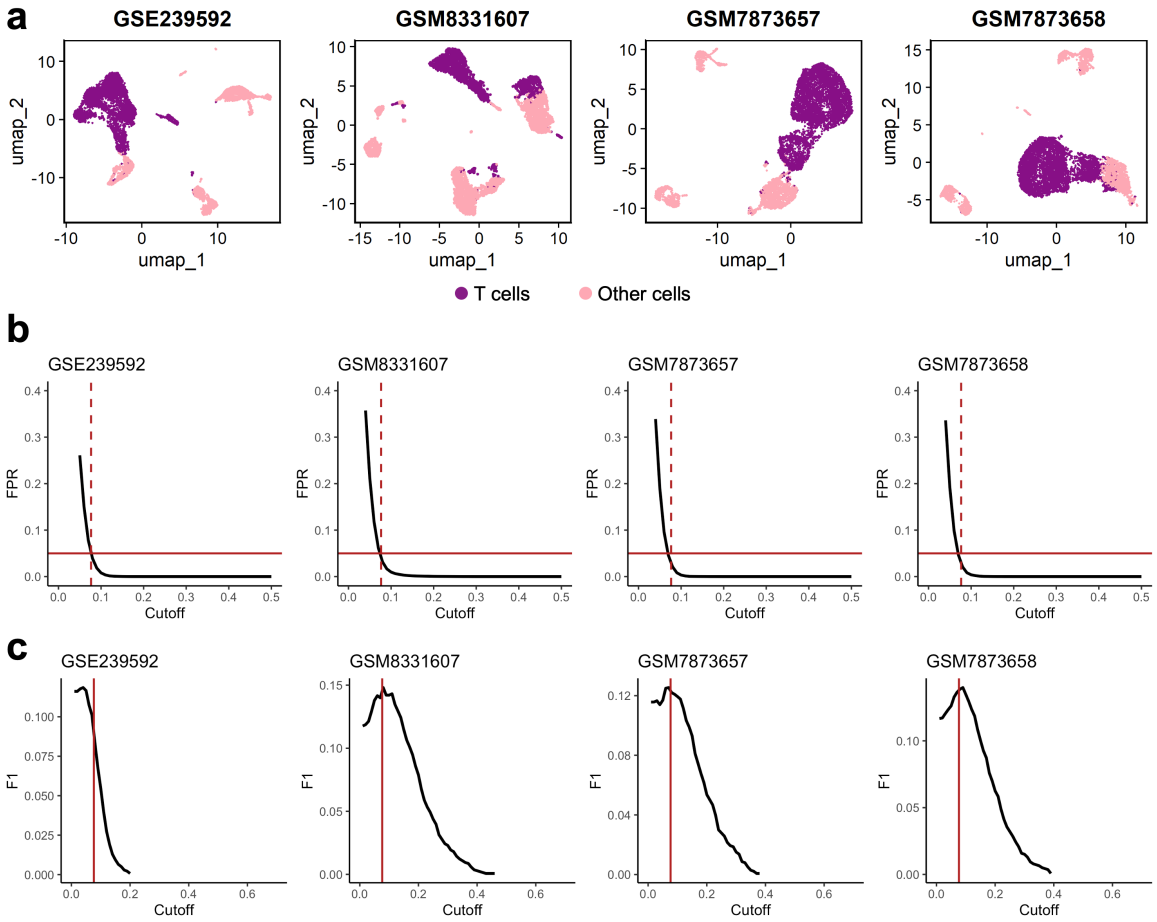

FIG 5. ***Cutoff selection along with Chatterjee's correlation coefficient is for scRNA-seq data from PBMC samples.*** *(a) FPR associated with various cut-off values for respective GRNs constructed from Chatterjee's correlation. A red vertical line depicts the computed cutoff value, and the horizontal red line depicts the FPR of 0.95. (b) F1 score associated with various cut-off values for respective GRNs constructed from Chatterjee's correlation. A red vertical line depicts the computed cutoff value.*

the most relevant regulatory interactions. This dual performance highlights the utility of our method for practical applications where both statistical rigor and predictive accuracy are important.

To investigate the transferability of the cutoff, we analyzed two publicly available Epithelial cell datasets (GSM8683884 and GSM8683885) derived from the small intestine (Fig. 6a). These datasets represent a morphologically and functionally distinct tissue compared to the T cells from the PBMC datasets used for initial cutoff determination. We preserved the same set of genes used in the initial PBMC cutoff selection for consistency.

We applied the previously determined PBMC cutoff (red line) to these epithelial cell datasets. As shown in Fig. 6b, this cutoff did not effectively control the FPR to the target level of 5%. Instead, we then computed a new optimal cutoff from the GSM8683884 dataset (green line) and applied it to the other epithelial cell dataset, which successfully controlled the FPR at the 5% level. The F1 score also highlighted the need for a tissue-specific cutoff. Fig. 6c shows that the PBMC cutoff (red line) did not correspond to the peak F1 score for either epithelial cell dataset. The highest F1 scores for the epithelial cells were instead achieved with their own tissue-specific cutoffs (green lines). These findings underscore that an optimal cutoff for identifying gene-gene interactions is tissue-specific and cannot be directly transferred between different tissue types due to fundamental differences in gene expression and cellular communication.

Biological interpretation for active regulation

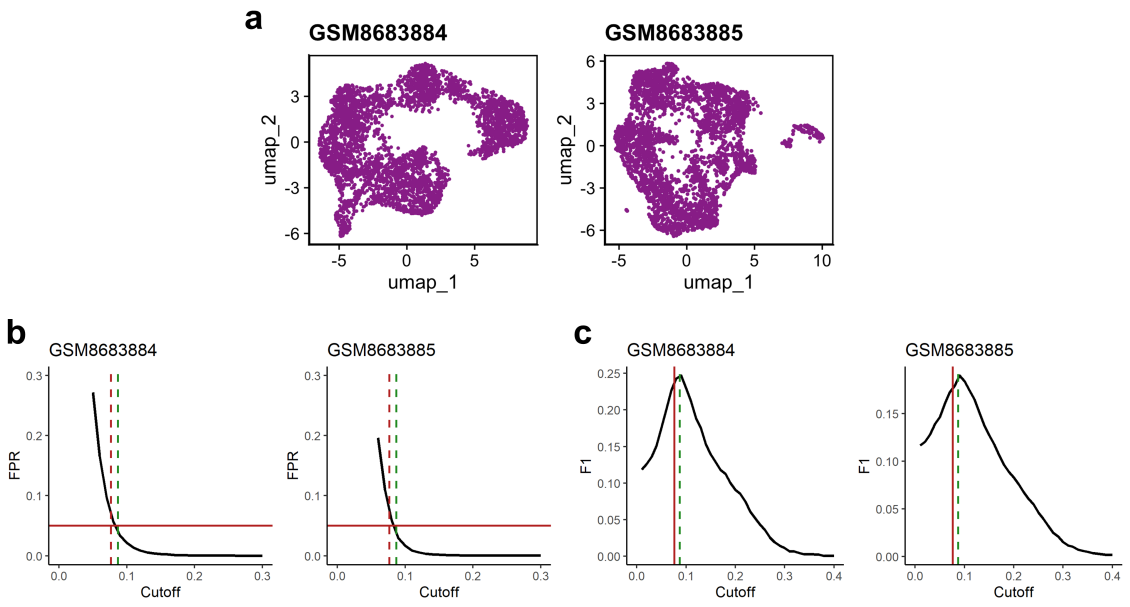

FIG 6. ***Cutoff selection along with Chatterjee's correlation coefficient is for scRNA-seq data from small intestine samples.*** *(a) UMAPs of scRNA-seq data from the respective GEO dataset. (b) FPR associated with various cut-off values for respective GRNs constructed from Chatterjee's correlation. The red vertical line depicts the computed cutoff value from PBMC data, the green vertical line depicts the cutoff specific to the small intestine data, and the horizontal red line depicts the FPR of 0.05. (b) F1 score associated with various cut-off values for respective GRNs constructed from Chatterjee's correlation. A red vertical line depicts the computed cutoff value, and the green vertical line depicts the cutoff specific to the small intestine data.*

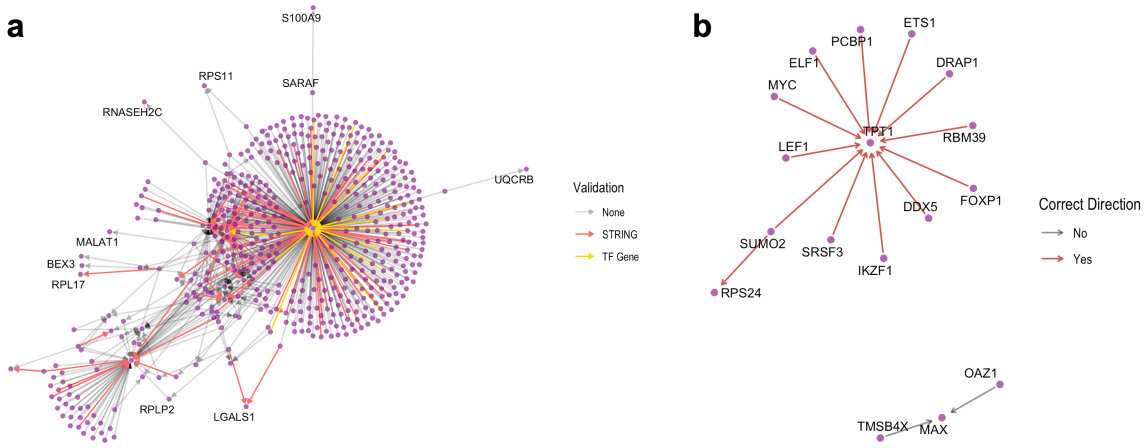

FIG 7. ***Real data application to identify active regulation in GSE239592 dataset*** *(a) All significant actively regulated interactions identified. Edges are colored to indicate validation: red for interactions found in the STRING database and yellow for those in the TFLink database. (b) Focused sub-network showing only the regulatory interactions present in the TFLink database. Black arrows denote interactions where the inferred direction (transcription factor to target gene) is correct, while red arrows indicate an incorrect direction.*

Results Next, we aimed to identify the actively regulated gene pairs, considering the problem in (7). As we have shown in Section 3.2 that Chatterjee's correlation outperforms GRNBoost2 in terms of identifying the active regulations, we focus on the former, and use the procedure in (8), and to evaluate the corresponding cutoff, we use the function given in (10).

For this analysis, we used the GSE239592 dataset, which contains 1,345 genes. We randomly sampled 4,000 cells for GRN construction. The STRING database served as our ground truth for undirected interactions, while

the TFLink database [27] served as our ground truth for directed interactions. In this study, we only preserved links from the STRING dataset with a combined score greater than 400. We first partitioned the data into a 60-40 train-test split. The training set was used to estimate a cutoff value that corresponds to a 0.05 FPR. This cutoff was then applied to the test set to identify significantly dependent interactions. Finally, we used the function in (10) to identify the directed, actively regulated interactions among these dependent links (Fig. 7a).

Applying a data-driven cutoff derived from the training set, we identified 25,254 undirected dependent interactions in the test set. The proposed cutoff function (10) was then applied as a filter, yielding 779 active regulatory interactions. For validation, we compared these interactions to established biological databases. Of the 779 identified interactions, 83 were validated within the STRING database, and 15 were confirmed as known transcription factor-gene (TF-Gene) interactions in the TFLink database. To further assess biological relevance and directionality, we examined the 15 confirmed TF-Gene links (Fig. 7b). We hypothesized that for a truly regulatory interaction ($TF \to gene$), Chatterjee's coefficient would be greater in the forward direction. This analysis revealed that 13 of these 15 interactions (86.67%) were correctly directed, demonstrating that Chatterjee's correlation coefficient can accurately identify active regulatory relationships, particularly when non-linear patterns are present.

4. Discussion

In this study, we have demonstrated the effectiveness and efficiency of using Chatterjee's correlation coefficient for constructing GRNs from scRNA-seq data. Our findings suggest that this method offers a robust and accurate alternative to traditional approaches. Through a series of simulation studies, we first established that Chatterjee's correlation coefficient is more effective at capturing relationships within GRNs compared to other established correlation coefficients. We then used a more complex simulation framework, SERGIO, to model intricate gene interactions. In this more challenging setting, Chatterjee's correlation not only surpasses traditional correlation methods but also shows promising behavior to outperform more advanced machine learning algorithms, including regression tree-based [28, 29] and principal component regression models [30].

A key advantage of our approach is its ability to discern active regulation within GRNs. By using the non-symmetric form of Chatterjee's correlation, we were able to identify directional regulatory relationships. We tested this by simulating GRNs with both active and linear interactions, and our method proved to be more accurate and interpretable than GRNBoost. This enhanced explainability is crucial for researchers seeking to understand the underlying mechanisms of gene regulation.

Our proposed cut-off selection function, while effective, is currently reliant on a well-curated ground-truth network. In this study, we used the STRING database to provide this ground truth for our real-data application. We acknowledge that the performance of this approach is tied to the quality of the reference network. However, the STRING database is continuously updated, reflecting the latest advances in biological knowledge. We believe that as these databases become more comprehensive and accurate, our method will become even more effective and scalable, offering a more robust way to identify significant regulatory interactions in gene networks. A notable characteristic of Chatterjee's correlation is that it successfully captures directionality in non-linear and functional dependency, which is abundant in practice. In the case of linear dependencies, the coefficient exhibits symmetry, thereby refraining from assigning a direction of regulation while reflecting its efficacy in detecting non-linear dependencies. Consequently, our findings underscore the utility of this method for accurately identifying and directing active regulatory interactions in complex biological networks that exhibit non-linear dependencies.

5. Data Availability

The sources of data sets used in this article can be found in Table 1. Simulated datasets using SERGIO were obtained from https://github.com/PayamDiba/SERGIO. The ground truth PPI network was obtained from the STRING database v12.0 (https://string-db.org/), and the ground truth TF-Gene dataset was obtained from the TFLink database (https://tflink.net/). No new data was generated to support this research.

6. Code Availability

All code used to generate the results in this study is available at https://github.com/Xenon8778/scXiNet.

7. References

1. Blinov, M. L., Faeder, J. R., Goldstein, B. & Hlavacek, W. S. A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Biosystems* **83,** 136–151 (2006).
2. Waters, K. M. *et al.* Network analysis of epidermal growth factor signaling using integrated genomic, proteomic and phosphorylation data. *PloS one* **7,** e34515 (2012).
3. Blencowe, M. *et al.* Network modeling of single-cell omics data: challenges, opportunities, and progresses. *Emerging topics in life sciences* **3,** 379–398 (2019).
4. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine* **9,** 75 (2017).
5. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology* **8,** 717–729 (2010).
6. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35,** 2159–2161 (2019).
7. Osorio, D., Zhong, Y., Li, G., Huang, J. Z. & Cai, J. J. scTenifoldNet: a machine learning workflow for constructing and comparing transcriptome-wide gene regulatory networks from single-cell data. *Patterns* **1** (2020).
8. Kang, Y., Thieffry, D. & Cantini, L. Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. *Frontiers in genetics* **12,** 617282 (2021).
9. Chatterjee, S. A new coefficient of correlation. *Journal of the American Statistical Association* **116,** 2009–2022 (2021).
10. Sadeghi, B. Chatterjee Correlation Coefficient: A robust alternative for classic correlation methods in geochemical studies-(including "TripleCpy" Python package). *Ore Geology Reviews* **146,** 104954 (2022).
11. Suo, C. *et al.* Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. *Nature Biotechnology* **42,** 40–51 (2024).
12. Spirtes, P., Glymour, C. & Scheines, R. *Causation, prediction, and search* (MIT press, 2001).
13. Glymour, C., Zhang, K. & Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in genetics* **10,** 524 (2019).
14. Kelly, J., Berzuini, C., Keavney, B., Tomaszewski, M. & Guo, H. A review of causal discovery methods for molecular network analysis. *Molecular genetics & genomic medicine* **10,** e2055 (2022).
15. Choi, J., Chapkin, R. & Ni, Y. Bayesian causal structural learning with zero-inflated Poisson Bayesian networks. *Advances in Neural Information Processing Systems* **33,** 5887–5897 (2020).
16. Choi, J. & Ni, Y. Model-based causal discovery for zero-inflated count data. *Journal of Machine Learning Research* **24,** 1–32 (2023).
17. Choi, J., Chapkin, R. S. & Ni, Y. Bayesian differential causal directed acyclic graphs for observational zero-inflated counts with an application to two-sample single-cell data. *The Annals of Applied Statistics* **19,** 1908 (2025).
18. Chaudhuri, A., Bhattacharya, A. & Ni, Y. Consistent DAG selection for Bayesian causal discovery under general error distributions. *arXiv preprint arXiv:2508.00993* (2025).
19. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology* **42,** 293–304 (2024).
20. Sun, W. & Tony Cai, T. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **71,** 393–424 (2009).
21. Leek, J. T. & Storey, J. D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105,** 18718–18723 (2008).
22. Chaudhuri, A. & Fellouris, G. *Sequential detection and isolation of a correlated pair* in *2020 IEEE International Symposium on Information Theory (ISIT)* (2020), 1141–1146.

23. Chaudhuri, A. & Fellouris, G. Joint sequential detection and isolation for dependent data streams. *The Annals of Statistics* **52,** 1899–1926 (2024).
24. Mering, C. v. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic acids research* **31,** 258–261 (2003).
25. Dibaeinia, P. & Sinha, S. SERGIO: a single-cell expression simulator guided by gene regulatory networks. *Cell systems* **11,** 252–271 (2020).
26. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51,** D638–D646 (2023).
27. Liska, O. *et al.* TFLink: an integrated gateway to access transcription factor–target gene interactions for multiple species. *Database* **2022,** baac083 (2022).
28. Breiman, L., Friedman, J. & Stone, C. J. *Classification and regression trees* (Chapman and Hall/CRC, 2017).
29. Chaudhuri, A. & Chatterjee, S. A cross-validation framework for signal denoising with applications to trend filtering, dyadic CART and beyond. *The Annals of Statistics* **51,** 1534–1560 (2023).
30. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed. ISBN: 978-0-387-84857-0 (Springer, New York, NY, 2009).

8. Acknowledgments

This research was funded by the Cancer Prevention and Research Institute of Texas (RP230204 to J.J.C) and the U.S. Department of Defense (GW200026 to J.J.C). The research of A.C. and Y.N. was supported by the National Institute of Health R01 (GM148974) and the National Science Foundation (DMS-2112943).

9. Author Contributions

S.G.: Conceptualization, Methodology, Visualization, Software, Formal Analysis, Writing - Original Draft. A.C.: Conceptualization, Methodology, Formal Analysis, Writing - Original Draft. V.R.: Software, Formal Analysis. Y.N.: Writing - Review & Editing. J.J.C: Conceptualization, Supervision, Writing - Review & Editing, Resources.

10. Ethics Declarations

10.1. Competing Interests

The authors declare no competing interests.