Causal Discovery from Single-Cell Perturbation Data

Abstract

Our summer research project focuses on the problem of causal discovery (structure learning) from single-cell perturbation data. Inferring causality from purely observational data is challenging. Perturb-seq provides a vast amount of interventional data without the need for costly or time-consuming scientific interventions, allowing us to explore the application of causal discovery methods to Perturb-seq datasets to estimate causal interactions and gain valuable insights into the regulatory relationships among genes. By modeling the gene regulatory network as directed acyclic graphs (DAGs) in causal Bayesian networks, we leverage graphical models to infer causal influences between genes.

Our project introduces a two-step approach for causal discovery from single-cell perturbation data. We apply normalization methods and address the confound variable sequencing depth by performing Poisson regression. Then we take the transformed working residuals and execute GIES method to find plausible causal hypotheses. We conduct numerical experiments on current Perturb-seq datasets to evaluate the identifiability and accuracy of our method. Despite challenges such as zero-inflation, we apply data imputation techniques such as MAGIC and scImpute, and modify the original model to account for zero-inflated data. Through these efforts, we aim to enhance the recovery of signals and improve the overall performance of our causal discovery method.

Contents

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized the field of genomics by enabling the analysis of gene expression at the individual cell level, providing larger sample size and more underlying information than bulk sequencing measurements. Applying causal discovery methods to current available single-cell datasets can estimate the causal interactions and give valuable insights into regulatory relationship among genes, providing precious clues for investigating gene regulations and molecular interactions.

In causal discovery of scRNA-seq data, the structure of gene regulatory network can be modelled as graphical models, for example, directed acyclic graphs (for short, DAGs) in causal Bayesian networks, where variables represent relative genes and directed edges represent causal influences. However, causal inference from purely observational data is considered as a task. Although the causal influences between variables reveal the conditional independence (Markov properties) and determine the skeleton of DAGs, the direction of edges in general can not be identified by the Markov property for the observational distribution. Interventions can help overcome such limitations (Hauser and Bühlmann, 2012).

While scientific interventions are crucial for acquiring causal knowledge, they can be costly, time-consuming, or even unethical. Single-cell perturb-seq has recently emerged and drastically transformed this field of research, enabling the simultaneous measurement of tens of thousands of interventions, molecules, and cells in a single experiment (Dixit et al., 2016; Replogle et al., 2022). Current available single-cell perturbation datasets are summarized by Green et al. (2022), providing an information resource for causal discovery with interventional data.

In this project, we propose a method for causal discovery (structure learning) from single-cell perturbation data called two-step approach. Numerical experiments on the single-cell perturbation datasets are conducted to support the identifiability and accuracy of our method. Although we are faced with several challenges, such as zero-inflation, we have tried recovering signals by applying data imputation methods ,and revising the original model to be zero-inflated in order to address these problems.

2 Basic concepts of Causal Structure Learning

We formulate the model as a structural causal model (Pearl 2009). In particular, we consider a linear structural equation model (SEM) (e.g., Wright 1921) for a p-dimensional random variable $X = (X_1, \cdots, X_p)^t$ under random noise ε: $X \leftarrow BX + \varepsilon$; where B is a $p \times p$ matrix.

We use a directed graph G to represent the underlying causal interaction in the SEM, where each variable X_k is represented by a node k, $k = 1, \cdots, p$, and there is an edge from node k to node j $(k \neq j)$ if and only if $B_{j,k} \neq 0$, which means X_k is a direct cause of X_j. The graph G is called acyclic if it does not contain a cycle. Some assumptions should be considered ,like causal sufficiency (There are no latent or confound variables) and Gaussianity of the noise distribution (The random noise ε follows Gaussian distributions or t-distributions with various degrees of freedom). The aim is to identify the graphical model and causal relationships among the variables.

3 Two-step Approach

Our research focuses on single-cell genetic perturbation/intervention datasets of RNA or protein. These datasets are typically presented as a count matrix denoted by $X = (X_{ik})$, where X_{ik} represents the measurement of the k-th gene or protein expression in the i-th cell. The perturbation information can be summarized by a binary matrix $Z = (Z_{ik})$, where $Z_{ik} = 1$ indicates that the k-th gene or protein is perturbed in the i-th cell.

For the count data matrix X, a common modeling assumption for read count data without biological variability is that the k-th gene takes up a certain fraction p_k of the total amount n_i of counts in the i-th cell (Townes et al., 2019). Given the expected value $\mu_{ik} = n_i \cdot p_k$, for each gene or protein k, the count data X_{ik} are modelled as Poisson samples without zero-inflation, i.e

$$X_{ik} \sim \text{Poisson}(\mu_{ik}) = \text{Poisson}(n_i \cdot p_k)$$

The total amount n_i of the i-th cell can be approximated by its sequencing depth. To further get the observational distribution information, we need to eliminate the influence of sequencing depth.

Step 1 In order to address the primary confounding variable, sequencing depth, we perform a Poisson regression with log(sequencing depth) as the intercept, separately for each gene or protein. Then we use the working residuals from the iteratively reweighted least squares fit . Some normalization methods, such as log-transformation, variance-stabilizing transformation, and other gene-expression based methods can be applied to the working residuals.

Step 2 After Step 1, we assume the transformed data to be multivariate Gaussian-distributed and use it to find plausible causal hypotheses. With corresponding perturbation matrix Z, causal discovery from such interventional data follows the score-based greedy search, Greedy Equivalence Interventional Search method (GIES), proposed by Hauser and Bühlmann (2012, 2015). The implementation is available as the gies function in package pcalg (this approach defines the score function based on Gaussian-distributed data).

Overview of GIES method Greedy interventional equivalence search method (Hauser and Bühlmann, 2012) is a technique used in structure learning for causal Bayesian networks. This method is an extended work of the Greedy Equivalence Search (GES) algorithm proposed by Chickering (2002) with data from known interventions. GES greedily maximizes some score function for given data over essential graphs, such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC).

While GES is a score-based optimization algorithm working on the space of observational essential graphs, GIES establishes a criterion for two DAGs being Markov equivalent under a given intervention setting, defining the so-called interventional Markov equivalence class and interventional Markov essential graphs. Since these two characterizations are similar to each other, the greedy search algorithm can be generalised. In the local searching part, instead of optimizing the score function in two phases: forward and backward, GIES execute the forward, backward and turning phase repeatedly in this order until none of them can augment the score function any more. Ultimately, the target graphical object of GIES is a interventional Markov equivalence class, which is a subclass of the Markov equivalence class of the underlying DAG and can be seen as a partially directed acyclic graph (PDAG).

4 Numerical experiments

We have performed two-step approach on the current Perturb-seq datasets to support our methodology. These datasets are collected and summarized by Green et al. (2022), also at http://projects.sanderlab.org/scperturb/.

Attributes of Current Perturb-seq datasets Green et al. (2022) collect a set of 44 publicly available single-cell perturbation-response datasets with molecular readouts, including transcriptomics, proteomics and epigenomics. The cells in the dataset are from Homo sapiens and Mus musculus, and most of them have using cell lines derived from various disease (leukemia is the most common disease,appearing in 23 datasets and including chronic myelogenous leukemia, acute monocytic leukemia, acute T cell leukemia and etc) such as A549 (lung adenocarcinoma), K562 (chronic myelogenous leukemia), and MCF7 (mammary adenocarcinoma).

33 datasets in this resource were perturbed using CRISPR and 8 datasets perturbed with drugs. Besides, 32 datasets measure scRNA-seq exclusively, there are 5 datasets measuring scATAC-seq and 1 measuring ATAC and protein simultaneously. Different perturbations act at different layers in the hierarchy of gene expression and protein production. Perturbations included in scPerturb include CRISPR-cas9, which acts directly on the genome, using indels to induce frameshift mutations which effectively knock out one or multiple specified genes; CRISPR a, which activates transcription of a target gene; CRISPRi, which blocks transcription of targeted genes; CRISPR-cas13, which cleaves targeted mRNAs and promotes their degradation; cytokines, that bind cell surface receptors; and small molecules perturb various cellular mechanisms. Single cell measurements probe the response to perturbation, also at different layers of gene expression: scATAC-seq directly probes chromatin state; scRNA-seq measures mRNA; and protein count data currently is typically obtained via antibodies bound to proteins.

CRISPR datasets tend to have more perturbations than drug datasets because they are easier to scale up using multiplexing, with a corresponding smaller number of cells per perturbation. By far the most detailed CRISPR dataset is from a recently published study which perturbed 9867 genes in human cells (ReplogleWeissman2022). Containing >2.5 million cells, this dataset is the largest in our database, with the number of cells each gene is detected in significantly higher than in other datasets.

Highlights on the numerical experiments Experiments conducted on different Perturb-seq datasets and variations in the selection of gene subsets can lead to divergent numerical results. Since in the GIES method, known interventions applied to variables is required. This necessitates the identification of targeted genes in the perturbation within each cell. Consequently, datasets using drug perturbations may not be well-suited for our setting, as is the case with Perturb-seq datasets using cytokine perturbations.

In order to effectively utilising the perturb information, it is desirable to have a higher overlapping between the genes of interest and the target genes in perturbation. To put it simple,we should select a subset of genes of interest that includes several target genes involved in the perturbation.

We can also consider selecting genes of interest from the original research papers associated with the Perturb-seq datasets. Typically, these target genes identified for perturbation align with primary or essential genes within the studied biochemical reactions or regulatory networks of interest.

Furthermore, we have encountered challenges relating to memory storage during the execution of numerical experiments. Datasets including an extensive number of genes and cells necessitate substantial storage capacity, exemplified by the ”ReplogleWeissman2022-K562-gwps.h5ad” dataset, which occupies a memory storage space of 8.8 GB. Performing tasks such as reading, processing, and conducting numerical analyses on these sizable datasets can be time-consuming and even unfeasible .To address this issue, one can explore approaches for accelerating computation or resorting to sub-sampling strategies.

Although we have applied two-step approach in many Perturb-seq datasets and gene subsets of interest, in the following analysis and discussion, we mainly focus on the PapalexiSatija2021 datasets and ReplogleWeissman2022 datasets for the abundant perturbation target genes and the reasonable data size. For the selection of gene subset,since the cells in PapalexiSatija2021 are with acute monocytic leukemia disease(a kind of Acute myeloid leukemia), while cells in ReplogleWeissman2022-K562 are with Chronic myeloid leukemia, we aim to investigate the causal interactions of the genes in the pathways of Acute myeloid leukemia and Chronic myeloid leukemia.

Normalization In the data processing part, uniform pre-processing and quality control pipelines had already been applied on the Perturb-seq datasets (Green et al., 2022). The count numbers of the cells are in a suitable range, and for quality control, the percent of mitocondrial genes in cells are reasonable and there is no need to filter those cells (For example, in PapalexiSatija2021 dataset, the minimal number of count data in cells is 2969, and the percent of mitocondrial genes is under 10% for each cell). To drop the low-expressed genes, we have filtered those genes which express in less than 5% of the total cells.

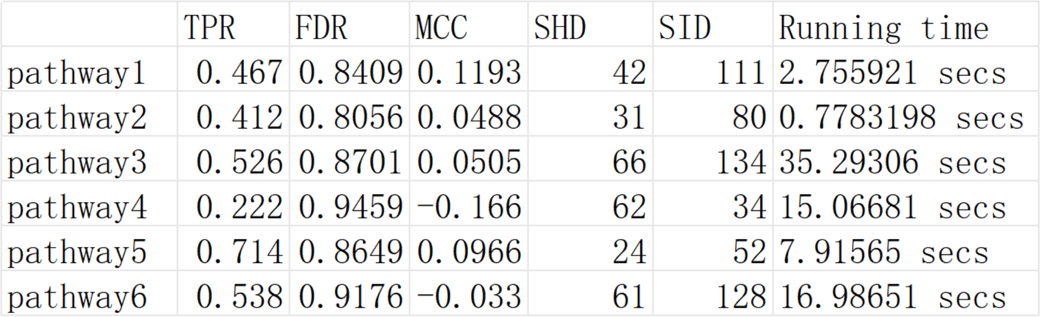
We consider some transformation-based methods for normalization. One widely-used technology is log transformation, using $\log(x)$ or $\log(x+1)$ as the transformation function. Besides, Normalizing the count data X by sequencing depths followed by one of the square-root transformations has been advocated for scRNA-seq data processing (Lause et al., 2021). For Poisson data, the square root transformation is approximately variance-stabilizing, as well as its modifications. Based on the performance, we choose the Freeman-

Tukey transformation $\sqrt{x}+\sqrt{x+1}$ (Freeman, Tukey; 1950). To fit in the definition domain, we will make a transformation by $x+1$ before Freeman-Tukey transformation.

Discussion We have chosen two networks of acute monocytic leukemia in Papalexi eccite RNA dataset, four networks of Chronic myeloid leukemia in ReplogleWeissman2022-K562-essential dataset. Papalexi eccite RNA dataset contains 18649 genes and 20729 cells, While ReplogleWeissman2022-K562-essential dataset contains 8563 genes and 310385 cells.

The true MAGs and true DAGs for each of the networks are available in KEGG Pathway Database: https://www.genome.jp/kegg/. The numerical performance of two-step approach is evaluated in terms of estimation accuracy of directed edges and coefficients. For the accuracy of estimated directed edges, we employ the true positive rate (TPR) and false discovery rate (FDR) as the evaluation metric. To evaluate the closeness of the estimated and true DAG, we report the normalized structural Hamming distance (SHD) (Tsamardinos et al., 2006), which measures the smallest number of edge insertions, deletions, and flips to convert the estimated DAG into the truth. We also report structural intervention distance (Peters et al., 2015) to evaluate the different causal inerence statements and different intervention distributions. For overall accuracy of the estimated DAG structure, we use the Matthews correlation coefficient (MCC) as an overall evaluation metric, which is also considered in Yuan et al. (2019).

Note that a good estimation is implied with small values of FDR, SHD, SID, but large values of TPR and MCC. Here are the numerical results (See Table 1):



	TPR	FDR	MCC	SHD	SID	Running time
pathway1	0.467	0.8409	0.1193	42	111	2.755921 secs
pathway2	0.412	0.8056	0.0488	31	80	0.7783198 secs
pathway3	0.526	0.8701	0.0505	66	134	35.29306 secs
pathway4	0.222	0.9459	-0.166	62	34	15.06681 secs
pathway5	0.714	0.8649	0.0966	24	52	7.91565 secs
pathway6	0.538	0.9176	-0.033	61	128	16.98651 secs

Table 1

From the numerical results, we can infer that two-step approach doesn't perform well on the datasets. Although the TPR is rather high, the FDR is much higher than our expectations. Most importantly, MCC is extremely low and even in some cases, MCC is lower than zero, which means that the inference performance of Two-step approach is worse than a random guess. SHD and SID are also high in some cases.

This can be explained by the estimated graph of two-step approach. Two-step approach tends to generate a rather dense causal graph and the direction of some edges remains undetected or even reverse (see Figure 1-12).

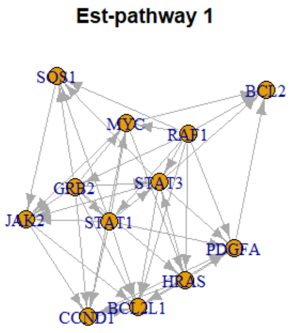

Est-pathway 1

Figure 1

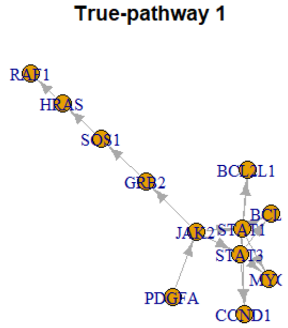

True-pathway 1

Figure 2

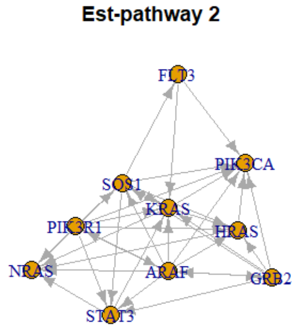

Est-pathway 2

Figure 3

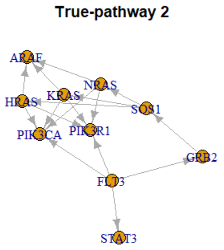

True-pathway 2

Figure 4

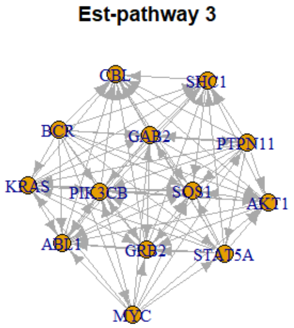

Est-pathway 3

Figure 5

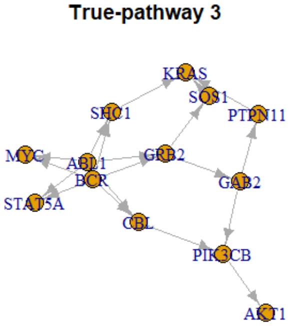

True-pathway 3

Figure 6

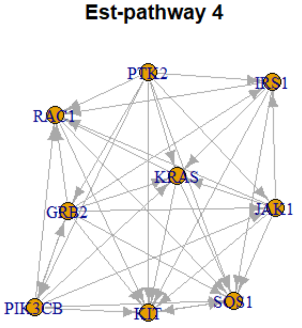

Est-pathway 4

Figure 7

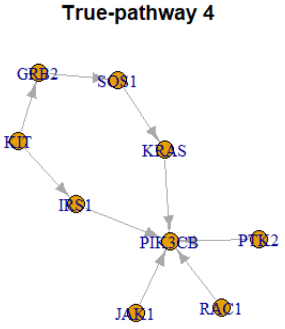

True-pathway 4

Figure 8

Hence, naturally we will consider making some constrain for the max degree of the variables to enhance the sparsity (This is also consistent with biological knowledge in gene regulatory networks, each gene is only expected to be regulated by a small number of other genes), by setting a maximum as one half of the total number of variables or one in three of the total number of variables. We hope that making constrain on the degrees of nodes is able to offset the over-density of GIES somehow. However, the FDR ,TPR, MCC of two-step approach is even worse in some cases(See Table 2).

Note that we use the pathway information to determine the network of genes, however, genes in the pathway may participate in other biochemical reaction and have other causal relationship. Except this reason, several

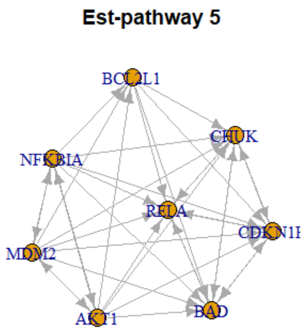


Figure 9

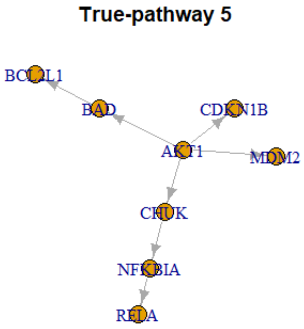


Figure 10

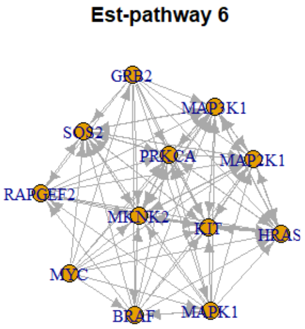


Figure 11

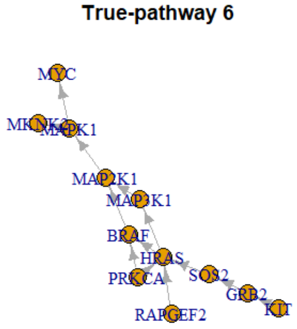


Figure 12

Pathway1						
maxdegree constrain?	TPR	FDR	MCC	SHD	SID	Running time
no constrain	**0.467**	0.841	0.119	42	111	2.755921 secs
p/2	0.4	**0.818**	**0.139**	34	112	2.084231 secs
p/3	0.133	0.909	−0.02	**31**	**101**	1.690966 secs
Pathway2						
no constrain	**0.412**	0.806	0.049	31	80	0.7783198 secs
p/2	0.294	**0.762**	**0.093**	**24**	79	0.457504 secs
p/3	0.118	0.857	−0.03	24	**70**	0.543267 secs
Pathway3						
no constrain	**0.526**	**0.87**	**0.051**	66	134	35.29306 secs
p/2	0.105	0.946	−0.1	49	150	1.058039 mins
p/3	0	1	−0.14	**41**	**148**	56.29149 secs
Pathway4						
no constrain	0.222	0.946	−0.17	34	62	15.06681 secs
p/2	**0.444**	**0.765**	**0.204**	17	58	15.89638 secs
p/3	0.222	0.833	0.074	**16**	**39**	15.80896 secs
Pathway5						
no constrain	0.714	0.865	0.097	24	52	7.91565 secs
p/2	0.286	0.889	0.003	16	47	7.826938 secs
p/3	0.286	0.8	0.125	**11**	**19**	8.275424 secs
Pathway6						
no constrain	**0.538**	**0.918**	**−0.03**	61	128	16.98651 secs
p/2	0.077	0.972	−0.13	44	114	16.32711 secs
p/3	0.077	0.957	−0.07	**33**	**109**	12.76076 secs

Table 2

plausible factors could potentially account for this phenomenon. Firstly, due to the exclusion of poorly-expressed genes during the gene subset selection process, this filtering step may introduce the confound or latent variables that may influence the observed relationships. Secondly, the score function employed in the implementation of the Greedy Equivalence Search (GIES) method is based on the assumption of a Gaussian distribution. However, the transformed residual data distribution might not necessarily conform to this assumption, resulting in a loss of identifiability and potentially affecting the accuracy of causal inference. Furthermore, in certain datasets, the number of overlapping genes between the target genes and the genes under study may be insufficient. Consequently, the available perturbation information cannot be fully utilized, leading to limited or biased conclusions about the underlying causal relationships within the dataset. Additionally, since the GIES method adopts a greedy search strategy, which may encounter challenges in escaping local optima. As a result, the method may fail to recover the most accurate and globally optimal results, impacting the reliability of the inferred causal relationships.

In summary, several factors should be considered when interpreting the observed phenomenon. These include the potential presence of confounding variables, deviations from the assumed Gaussian distribution, limitations arising from the number of overlapping genes, and the inherent limitations of the greedy search employed by the GIES method.

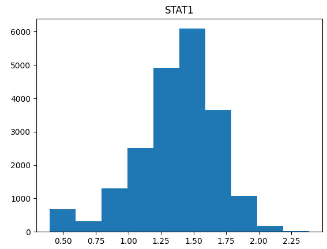


Figure 13: STAT

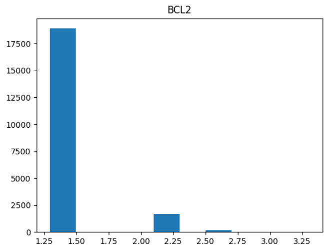


Figure 14: BCL2

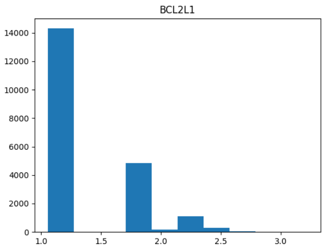


Figure 15: BCL2L1

Zero inflation There are over 90% of zeroes in the Papalexi eccite RNA dataset and ReplogleWeissman2022-K562-essential dataset. Excessive zeroes in the datasets can account for the first and second factor. Firstly, excessive zeroes in the gene expression matrix imply poorly-expressed genes, which are filtered in the preprocessing and can possibly become the latent factors. Secondly, in Poisson regression, we have

$$\text{working residual} = \frac{y_{\text{expected}} - y_{\text{estimates}}}{y_{\text{estimates}}}$$

where y_{expected} is the true value while $y_{\text{estimates}}$ is the estimated value in the model. When $y_{\text{expected}} = 0$, working residual = -1, however, working residual

≥ -1 naturally. Hence, in our model, excessive zeroes in the gene expression matrix can directly lead to the deviation from Gaussion distribution, which contradicts with the original assumption. According to the numerical results, only working residuals of genes which express in over 90% cells perform like Gaussian distributed (see Figure 4-6, where STAT gene expresses in over 80% of total cells, BCL2 and BCL2L1 gene express in less than 50% of total cells). However, in scRNA-seq datasets, this condition is usually hard to be satisfied, I have filtered the genes expressing in over 90% of total cells, the number of those genes is lower than 2% of total cells. Also, it turned out these cells are normally the essentially functional cells , such as Mitochondrial ribosomal protein or Aerobic respiratory enzyme.

Several alternative types of residuals from the generalized linear model (GLM) were explored in our analysis, including response residuals, Pearson residuals, and deviance residuals. However, it was observed that the presence of zeros in the original datasets introduced undesired effects. To mitigate the issue of zero inflation, we employed various methods.

One approach involved applying data imputation techniques for scRNA datasets to recover the underlying signal within the gene regulatory network. By imputing missing values, we aimed to address the challenge posed by excessive zeros and enhance the accuracy of downstream analyses.

Another strategy involved revising the original modeling framework. For instance, instead of relying on a Poisson model assumption, alternative distributions such as hyper-Poisson or negative binomial were considered for modeling the count data. These distributions account for overdispersion and are better suited to handle data exhibiting zero inflation characteristics.

By incorporating these methodological adjustments, we aimed to improve the robustness and reliability of our analyses, overcoming the limitations associated with zero-inflated data and enabling more accurate inference and interpretation of the underlying biological processes within the studied system.

5 Data imputation methods

Zero-inflated distributions of gene expression data is considered as a type of noise. While the zeros in the data matrix can be true biologically, technical failure to detect the signal is commonly observed in single-cell data and is referred to as dropout for single-cell gene expression experiments (Kharchenko, Silberstein and Scadden, 2014, Zhu et al., 2018). Hence, there are a increasing number of data imputation methods proposed in order to recover the signal from scRNA sequencing data accurately and robustly (Patruno et al., 2021). We consider applying these methods to recover the count data before we implement two-step approach on the datasets.

Prior work The current data imputation methods for single-cell sequencing data can be roughly divided into several categories, including data smoothing methods, such as MAGIC (Dijk et al., 2018); Deep learning-based methods, such as DeepImpute (Arisdakessian et al., 2019); Matrix factorization methods, such as scRMD (Chen et al., 2020); Model-based method, such as scImpute (Li et al., 2018). We have implemented MAGIC and scImpute on the datasets.

MAGIC (Markov Affinity-based Graph Imputation of Cells): MAGIC is a graph-based imputation method specifically designed for single-cell data. Firstly It uses techniques like k-nearest neighbors (k-NN) or mutual information to construct a graph representation capturing cell-to-cell similarities .Then, an affinity matrix is defined by applying a Gaussian kernel on the principal components of the graph. Lastly it uses diffusion processes to propagate information and impute missing values. MAGIC incorporates local neighborhood structure to impute missing values accurately.

scImpute is a Bayesian framework that models the dropout events in scRNA-seq data explicitly. Firstly it uses PCA or other cluster methods to divide the subpopulations of cells. Then it models genes in each subpopulation with a gamma-normal mixture mode to infer the probability of drop-out.In the end, it uses a linear combination of the gene expression in the cell subpopulation as the imputation value for those highly probable dropout events.

However, the numerical results may be not very ideal.In our numerical experiments on PapalexiSatija2021 datasets, scImpute failed to give the recovery results in nearly 24 hours. According to Li et al., (2018), scImpute completes computation in seconds when applied to a dataset with 10,000 genes and 100 cells, running with 10 cores. The memory requirement for this data set is around 2G. Importantly, the running time mostly depends on the number of processors and the number of cells in the scRNA-seq data. We set the number of processors to be 10 while the numbers of cells is over the 100 or 1000 times of 100. Li et al., (2018) suggest that, When the number of cells is extremely large, a filtering step on the cells can save the computation time. However, the filtering step may cause the loss of information in the datasets.

In the implementation of MAGIC, when the number of cells is extremely large, the algorithm can not give an answer due to the overwhelmed memory storage and excessive running time. For the two pathways in PapalexiSatija2021 datasets, although the zeroes in the original count matrix are recovered after transformation, the distribution of working residuals data are still not very Gaussian-distributed. Besides, We have applied two-step approach on the after-MAGIC datasets and found that the accuracy of two-step approach is refined after MAGIC to some extent. However, some metrics, such as FDR, MCC, TPR are not surely better than numerical results in the experiments where no imputation is performed, which means that MAGIC may not have a significant effect in recovering the interactions among genes in our method.

Discussion Perturb-seq datasets are different from the general scRNA datasets in underlying structure. Ignoring the perturbation information will make the imputation methods biased and inaccurate. For example, when the perturbation uses knock-out technology, the expression of the target gene and a series of downstream genes in the pathway will become extremely low, hence the resulting zeros are biologically true values and are not suitable for imputation and recovery in this situation.

Also, building the metrics of ”similarity bewteen cells” is a significant step in MAGIC and other data smoothing methods since the imputation is completed by borrowing information from those similar cells. We should also take the perturbation information into consideration in designing such a metric.

6 Zero-inflated Graphical model

Prior work Although negative binomial or zero-inflated models may work for the zero-inflated datasets, extending such models for causal inference can be difficult since most model are designed for a single variable (gene) , but causal structure learning concerens the joint distribution of multiple variables (genes). We consider building such models based on the prior work in zero-inflated graphical model.

There have been a few recent BNs that are fully identifiable for observational zero-inflated data. Using Hurdle conditional distributions (McDavid et al., 2016), Yu et al. (2020) proposed fully identifiable BNs (ZiDAG) for zero-inflated Gaussian data under a natural and practical assumption. To estimate the underlying DAG, they apply greedy search (including exhaustive search and greedy search) and the score function is based on BIC. Choi et al. (2020) developed zero-inflated Poisson BNs for observational zero-inflated count data, introducing a fully Bayesian inference approach which exploits the parallel tempering Markov chain Monte Carlo algorithm to efficiently explore the multi-modal network space. To further accounting for over dispersion in the count data, Choi et al. (2023) propose a new zero-inflated generalized hypergeometric directed acyclic graph (ZiG-DAG) model for inference of causal structure from purely observational zero-inflated count data. They exploit a broad family of generalized hypergeometric probability distributions and are useful for modeling various types of zero-inflated count data with great flexibility, where score-based algorithms are developed for causal structure learning.

Combine GIES with Hurdle Normal Model Currently, We consider combining GIES with Hurdle Normal Model and fully identifiable BNs for zero-inflated Gaussian data proposed by Yu et al. (2020).

Hurdle model In Greedy Interventional Equivalence Search , the score function should be score-equivalent for each interventional Markov equivalence class (Here, interventional Markov equivalence class stands for interventional-equivalent DAGs), and decomposable (see Definition 1).

Definition 1 A score function S is called **decomposable** if for each DAG D, S can be written as a sum

$$S(D;\tau,X)=\sum_{i=1}^{p}s(i,\mathrm{pa}_D(i);\tau,X)$$

where the local score s depends on X only via $X_{.i}$ and $X_{.\mathrm{pa}_D}$, with $X_{.i}$ denoting the i^{th} column of X and $X_{.\mathrm{pa}_D}$ the submatrix of X corresponding to the columns with index in $\mathrm{pa}_D(i).(\tau,X)$ is the interventional data.

In Gaussian Causal Models, we always assume Gaussian intervention variables. In this case, not only the observational density f is Gaussian, but also the interventional densities $f(x \mid \mathrm{do}_D(X_I = U_I))$. An interventional data set $(\tau, X$ consists of n independent but not identically distributed Gaussian samples.

Usually, in GIES, We use BIC as score function (the consistency has been proved by Hauser and Bühlmann , 2015):

$$S(D;\tau,X):=\sup\{l_D(B,\sigma^2;\tau,X)\mid B\in B(D),\sigma^2\in\mathbb{R}^p_{>0}\}-\frac{k_D}{2}\log(n)$$

,

where l_D denotes the log-likelihood of density :

$$l_D(B, \sigma^2; \tau, X) = \sum_{i=1}^{n}[\sum_{j \notin T^{(i)}} logf(X_j^{(i)} \mid X_{\text{pa}_D(j)}^{(i)}) + \sum_{j \in T^{(i)}} log\hat{f}(X_j^{(i)})]$$

where B is the weight matrix, for each i, $T^{(i)}$ denotes the interventional target under which the sample $X^{(i)}$ was produced. The joint product density of X_j in the probability space of interventional variables, called level density, is denoted by $\hat{f}$, which can be seen as a constant in the optimization of score function.

Then we consider the score based function in ZiDAG model.Different from the assumption that $f(X_j^{(i)} \mid X_{\text{pa}_D(j)}^{(i)})$ is Gaussian, the density follows a Hurdle normal conditional(here we use (p, μ, σ^2)-Hurdle conditionals):

when $X_j^{(i)} = 0$,

$$log\ f(X_j^{(i)} \mid X_{\text{pa}_D(j)}^{(i)}) = log\{1 - p_{X_j}(X_{\text{pa}_D(j)}^{(i)})\}$$

when $X_j^{(i)} \neq 0$,

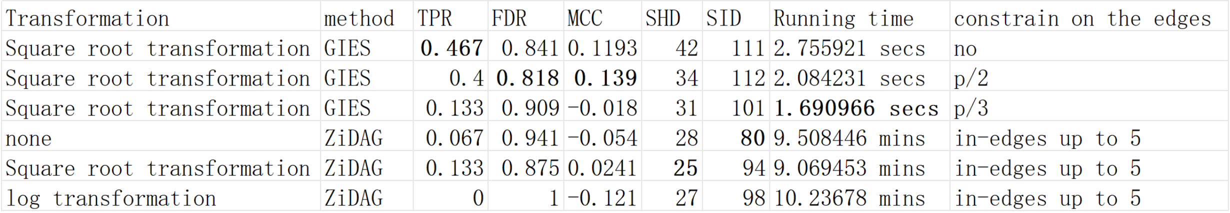
$$\begin{aligned}\log\ f(X_j^{(i)} \mid X_{\text{pa}_D(j)}^{(i)}) = log\ p_{X_j}(X_{\text{pa}_D(j)}^{(i)}) - \frac{1}{2}log(2\pi\sigma_{X_j}^2) - \\ \{X_j^{(i)} - \mu_{X_j}(X_{\text{pa}_D(j)}^{(i)})\}^2/(2\sigma_{X_j}^2)\end{aligned}$$

where p_{X_j}, μ_{X_j} are polynomials in $X_{\text{pa}_D(j)}$ and their indicators.

In the ZiDAG implementation, we use l_2 regularization and select its tuning parameter using BIC. We also assume the highest degree of Hurdle polynomials and select the degree by optimising BIC simultaneously over the degree and penalty.

Hence, combining these two methods together can be practical by adapting the score function in GIES method and use the greedy search. Although ZiDAG proves the identifiability of the exact graph while GIES method can only recover a PDAG, the searching space of GIES is reduced to a interventional Markov equivalence class space, and the running time through search can decline a lot. Irrespective of the learning class, structure learning represents an NP-hard problem where the number of possible graphs grows super-exponentially with the number of variables. This is a problem because computational complexity increases both with the number of the variables and the sample size. This can possibly limit current large or dense networks which tend to require large sample sizes to achieve reasonable structure learning accuracy.

We have performed the ZiDAG on the control data in the Perturb-seq datasets, and compared its performance with GIES on the Perturb-seq datasets (see Table 3). From the table we can infer that square root transformation is



Transformation	method	TPR	FDR	MCC	SHD	SID	Running time	constrain on the edges
Square root transformation	GIES	**0.467**	0.841	0.1193	42	111	2.755921 secs	no
Square root transformation	GIES	0.4	**0.818**	**0.139**	34	112	2.084231 secs	p/2
Square root transformation	GIES	0.133	0.909	-0.018	31	101	**1.690966 secs**	p/3
none	ZiDAG	0.067	0.941	-0.054	28	**80**	9.508446 mins	in-edges up to 5
Square root transformation	ZiDAG	0.133	0.875	0.0241	**25**	94	9.069453 mins	in-edges up to 5
log transformation	ZiDAG	0	1	-0.121	27	98	10.23678 mins	in-edges up to 5

Table 3

recommended for the data preprocessing .For the accuracy of this method, we can see that ZiDAG performs worse than GIES in terms of FDR, TPR, MCC, but better in SHD and SID. Besides, from the perspective of running time, ZiDAG normally takes much longer than GIES. The numerical results is not consistent with our assumption, and the possible reason can be that, mutual interaction of genes may not be fully identified through pathway information.

Since we are still working on combining these two methods and conducting numerical experiments, we will update the application on real data and numerical analysis later.

7 Discussion

In this project, we propose a method for causal discovery (structure learning) from single-cell perturbation data called two-step approach. We have conducted numerical experiments on current Perturb-seq datasets to test the identifiability and accuracy of our method. The attributes of existing Perturb-seq datasets and recommendation for selection of gene subsets are discussed. We have tried recovering signals by applying data imputation methods ,and revising the original model to be zero-inflated in order to address these problems in order to improve the robustness and reliability of our method and overcome the limitations associated with zero-inflated data. Although our work is more of an exploratory progress than a completed project, we still hope to provide some information for the modeling, processing and analysing of Perturb-seq data.

8 Referfence

Freeman MF, Tukey JW. Transformations related to the angular and the square root. Ann Math Stat. 1950;21(4): 607–11

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., et al. “Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens.” Cell, 167(7):1853–1866 (2016).

Green, T. D., Peidli, S., Shen, C., Gross, T., Min, J., Garda, S., Taylor-King, J. P., Marks, D. S., Luna, A., Bl¨uthgen, N., et al. “scPerturb: Information resource for harmonized single-cell perturbation data.” In NeurIPS 2022 Workshop on Learning Meaningful Representations of Life (2022).

Hauser, A. and B¨uhlmann, P. “Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs.” Journal of Machine Learning Research, 13(79):2409–2464 (2012).

—. “Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs.” Journal of the Royal Statistical Society: Series B, 77(1):291–318 (2015).

Lause, J., Berens, P., and Kobak, D. “Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data.” Genome biology, 22(258) (2021).

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., et al. “Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq.” Cell, 185(14):2559–2575 (2022).

Dalerba P, Kalisky T, Sahoo D, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat Biotechnol 2011;29(12):1120.

Zhu L, Lei J, Devlin B, et al. A unified statistical framework for single cell and bulk RNA sequencing data. Ann Appl Stat 2018;12(1):609.

Van Dijk D, Sharma R, Nainys J, et al. Recovering gene interactions from single-cell data using data diffusion. Cell 2018;174(3):716–29.

Arisdakessian, C., Poirion, O., Yunits, B. et al. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biol 20, 211 (2019).

Li, W.V., Li, J.J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun 9, 997 (2018)

Chen C, Wu C, Wu L, Wang X, Deng M, Xi R. scRMD: imputation for single cell RNA-seq data via robust matrix decomposition. Bioinformatics. 2020

McDavid, Andrew, Gottardo, Raphael , Simon, Noah, Drton, Mathias. (2016). Graphical Models for Zero-Inflated Single Cell Gene Expression. The Annals of Applied Statistics. 13. 10.1214/18-AOAS1213.

Yu S , Drton M , Shojaie A .Directed Graphical Models and Causal Discovery for Zero-Inflated Data[J]. 2020.DOI:10.48550/arXiv.2004.04150.

Choi J , Chapkin R S , Ni Y .Bayesian Causal Structural Learning with Zero-Inflated Poisson Bayesian Networks[C]//Neural Information Processing Systems.2020.

Choi J , Ni Y .Model-Based Causal Discovery for Zero-Inflated Count Data. Journal of Machine Learning Research. 2023.