Simulation-based Benchmarking for Causal Structure Learning in Gene Perturbation Experiments

Luka Kovačević[1] **Izzy Newsham**[1] **Sach Mukherjee**[1,2] **John Whittaker**[1]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[2]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

1 INTRODUCTION

Causal structure learning (CSL) refers to the task of learning causal relationships between variables in the form of a directed graph. Recent years have seen considerable theoretical and methodological advances in CSL, including formulation via continuous optimization [Zheng et al., 2018, Lippe et al., 2021, Lopez et al., 2022]. Despite these developments, CSL in scientific applications remains challenging, due to problem dimension, noise, data limitations, latent variables, and a lack of ground-truth causal knowledge.

The application of CSL to problems in biomedicine yields many of such problems. Advances in molecular biology have enabled gene-knockout experiments at an unprecedented scale [Dixit et al., 2016] with measurements of tens of thousand of variables under thousands of interventions, and recent experiments spanning millions of cells [Replogle et al., 2022]. These developments promise to transform biomedical discovery by enabling causal experiments at genome-wide scale. However, it remains challenging to understand performance, making it difficult to establish reliable CSL-based workflows.

Related Work. The need to better understand CSL performance characteristics has motivated a line of work in benchmarking methods [Eigenmann et al., 2020, Chevalley et al., 2022]. Single-cell Expression of Genes In-silico (SERGIO) was developed by Dibaeinia and Sinha [2020] as an approach based on stochastic differential equations. CausalBench [Chevalley et al., 2022] has also been developed as a framework for evaluating structure learners directly from real-world data.

Contributions. We propose CausalRegNet, an approach for generating realistic synthetic data from an underlying directed acyclic graph (DAG) at the scale now seen in contemporary experiments and data acquisition pipelines. Our work has several novel contributions: (i) We show how structural causal models can be used to generate realistic data; (ii) We put forward an interpretable and user-friendly approach to calibrating simulators; (iii) We develop CausalRegNet, a simulator motivated by large-scale interventional experiments in biology, which can be matched to real-world gene perturbation screens such as Replogle et al. [2022].

2 BACKGROUND

Definition 2.1 (Structural Causal Model; Peters et al. [2017])**.** A *structural causal model* (SCM) $\mathfrak{C} = (S, \mathcal{P}_N)$ is a collection S of d structural assignments corresponding to each $X_j \in \mathbf{X} = \{X_1, \ldots, X_d\}$,

$$X_j := f_j(X_{\text{pa}(j)}, N_j), \qquad j = 1, \ldots, d, \tag{1}$$

where $X_{\text{pa}(j)} \subseteq \mathbf{X} \backslash \{X_j\}$ are *parents* of X_j, $\mathcal{P}_N$ is a joint distribution over noise variables, which we require to be jointly independent.

The causal relationship between X_j and $X_{\text{pa}(j)}$ can be represented via a causal graph. The causal graph $\mathcal{G} = (V, E)$ is composed of vertices $V = [d]$, each corresponding to a random variable in $\mathbf{X}$, and edges $E = \{i \to j : X_i \in X_{\text{pa}(j)}\}$. We assume that $\mathcal{G}$ has no paths of the form $j \to \ldots \to i$, that is, $\mathcal{G}$ is acyclic and there are no paths along directed edges that lead from a child to one of its parents.

Definition 2.2 (Additive Noise Model; Hoyer et al. [2008])**.** An SCM with structural assignments of the form,

$$X_j := f_j(X_{\text{pa}(j)}) + N_j \tag{2}$$

where $f_j(\cdot)$ is an arbitrary function and the noise variables N_j are jointly independent, is an additive noise model (ANM). If $f_j(\cdot)$ is linear and the noise variables are distributed as $N_j \sim \mathcal{N}(0, \sigma_j^2)$ this gives the family of linear Gaussian ANMs.

3 METHODOLOGY

We begin by defining the *desiderata* for CausalRegNet, which are used to guide development of a simulator for CSL

in genomics data: (i) interpretability of simulator parameters; (ii) scalability to the size of real-world experiments; (iii) non-varsortability as defined by Reisach et al. [2021]; (iv) matched distributional properties across empirical and synthetic data.

For each node j, the random variable associated with this node is distributed according to,

$$X_j \sim \text{NegBin}(\mu_j, \sigma_j^2), \tag{3}$$

where μ_j is the mean expression and σ_j^2 is the variance. The parameters are defined as

$$\mu_j = \mu_j^0 \cdot f_j^{\text{reg}}(X_{\text{pa}(j)}; \Theta_j), \quad \text{and,} \tag{4}$$

$$\sigma_j^2 = \mu_j(1 + \mu_j/\theta_j), \tag{5}$$

where μ_j^0 is the observational mean and $f_j^{\text{reg}}(\cdot)$ represents the regulatory effect of the parents of X_j on X_j. The mean-variance relationship imposed via σ_j^2 and the multiplicative regulatory effect reflect established domain knowledge [Love et al., 2014], satisfying desiderata (i).

The regulatory effect function takes a sigmoidal form,

$$f_j^{\text{reg}}(X_{\text{pa}(j)}; \Theta_j) = \frac{\alpha_j}{1 + \exp\left\{-\gamma_j(w(x_{\text{pa}(j)}) + b_j)\right\}}, \tag{6}$$

where α_j is the maximum regulatory effect of $X_{\text{pa}(j)}$ on X_i, and $w(x_{\text{pa}(j)})$ is any function that aggregates the regulatory effect of parents.

For simplicity, here we use $w(x_{\text{pa}(j)}) = \sum_{i \in \text{pa}(j)} w_{ij} \cdot x_i/\mu_i^0$, where $W \in \mathbb{R}^{n \times n}$ is a weighted adjacency matrix that defines the strength of relationship between X_i and X_j and $[W]_{ij} = w_{ij}$ is the edge weight for the edge going from X_i to X_j.

The parameters γ_j and b_j are calibrated according to the aggregation function $w(\cdot)$, maximum regulatory effect $\alpha_j > 1$, and minimum regulatory effect $0 < \beta_j < 1$. To do this, we define the following value conditions:

I. **Baseline Expression Condition (BEC):**

$$\text{if } \forall i \in \text{pa}(j), X_i = 0, \text{ then } f_j^{\text{reg}}(X_{\text{pa}(j)}) = \beta_j,$$

II. **Observational Expression Condition (OEC):**

$$\text{if } \forall i \in \text{pa}(j), X_i = \mu_i^0 \text{ then } f_j^{\text{reg}}(X_{\text{pa}(j)}) = 1.$$

In other words, we first have that when the expression of each of the parents of node j are at zero, the regulatory effect will be at the baseline β_j. Conversely, when each of the parents of j are expressed at the observational mean, then the regulatory effect is equal to 1, implying $\mu_j = \mu_j^0 \cdot 1 = \mu_j^0$. That is, X_j is at its observational mean.

These value conditions allow us to freely specify the aggregation function $w(\cdot)$ and calibrate the regulatory effect function to guarantee maximum and minimum effects.These conditions are sufficient to yield closed-form solutions for γ_j and b_j given the parameters $\{\alpha_j, \beta_j, w(\cdot)\}$.

4 EXPERIMENTS

Figure 1a shows the time taken to simulate data from graphs of size 3 to 10,000. The scalability of CausalRegNet is shown to be comparable to the simple Gaussian ANM. For the same graphs, SERGIO takes upwards of 2.78 hours. To examine the varsortability of data generated by each simulator we generated data from causal chain and causal graph structures. The results for the causal chain and causal graph simulations are shown in Figure 1b-c, respectively, with CausalRegNet outperforming SERGIO and the Gaussian ANM. Hence, CausalRegNet satisfies desiderata (ii) and (iii).

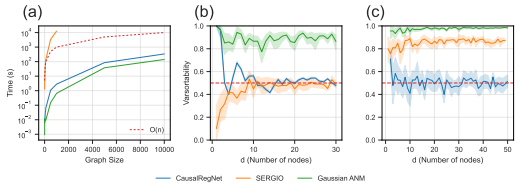

Figure 1: (a) Simulation time. Mean varsortability of data generated from (b) a causal chain and (c) a causal graph structure by each simulator with 95% confidence intervals.

We extract 100 cancer-related genes from the Replogle et al. [2022] dataset and fit negative binomial distributions to each gene that are then used to simulate via CausalRegNet. Figure 2b shows the distribution of Wasserstein distances (WDs) between the synthetic and real marginal expression distributions when compared to the true expression distributions (green) and a random gene (purple). The distribution of interventional effects is shown to be realistic in Figures 2c-d. CausalRegNet therefore satisfies desiderata (iv).

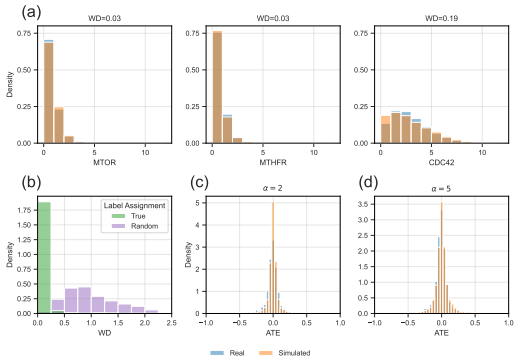

Figure 2: (a) Comparison between real and simulated data in DAG with 3 nodes. (b) Comparison between WD for shuffled and true gene labels on synthetic distributions. The distribution of interventional effects with (c) $\alpha_j = 2$ and (d) $\alpha_j = 5$ for each node.

Furthermore, in the full paper, we present a simulation study to examine the performance of various CSL algorithms using CausalRegNet as our synthetic data generator.

Acknowledgements

LK acknowledges the support of an Ivan D Jankovic Studentship at Clare Hall, University of Cambridge. This work was funded by UKRI Programme Grant MC_UU_0000218. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. *arXiv preprint arXiv:2210.17283*, 2022.

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

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

Marco Eigenmann, Sach Mukherjee, and Marloes Maathuis. Evaluation of causal structure learning algorithms via risk estimation. In *Conference on Uncertainty in Artificial Intelligence*, pages 151–160. PMLR, 2020.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. *arXiv preprint arXiv:2107.10483*, 2021.

Romain Lopez, Jan-Christian Hütter, Jonathan Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35:19290–19303, 2022.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.