# GSK.AI CAUSALBENCH CHALLENGE (ICLR 2023) REPORT SUBMISSION – BETTERBOOST

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The introduction of large-scale, genome-wide, single-cell perturbation datasets provides the chance to learn a full gene regulatory network in the relevant cell line. However, existing gene regulatory network inference methods either fail to scale or do not explicitly leverage the interventional nature of this data. In this work, we propose an algorithm that builds upon GRNBoost by adding an additional step that complements its performance in the presence of labeled, single-gene interventional data. Our method, BetterBoost, significantly outperforms the baseline on provided single-cell perturbation datasets when non-zero fractions of labeled interventions are available, demonstrating the efficacy of our approach for inferring gene regulatory networks from large-scale single-cell perturbation datasets.

## 1 INTRODUCTION

The introduction of large-scale, genome-wide, single-cell perturbation datasets (Replogle et al., 2022; Dixit et al., 2016) provides the chance to learn a full gene regulatory network. However, existing methods for gene regulatory network inference either fail to scale (Brouillard et al., 2020; Sethuraman et al., 2023) or do not explicitly leverage the interventional nature of this data (Moerman et al., 2019; Passemiers et al., 2022). Methods that fail to scale often have algorithmic complexity issues, such as those that arise when computing the exponential of a large matrix. Conversely, some methods that can handle over 10,000 genes (Moerman et al., 2019; Passemiers et al., 2022) treat the dataset as observational, overlooking the valuable interventional data. Although incorporating interventional data can improve the predictive power of models that treat the dataset as observational, such models fail to exploit causal inference principles that can help identify causal relationships. For example, recent works (Peters et al., 2015; Arjovsky et al., 2019) have leveraged the invariance of the conditional distribution $P(\boldsymbol{x}|\mathrm{Pa}(\boldsymbol{x}))$ for causal discovery. Here, $\mathrm{Pa}(\boldsymbol{x})$ denotes the set of direct causal parents of $\boldsymbol{x}$.

Among the scalable models that do not incorporate interventional data, we found that GRNBoost (Moerman et al., 2019) performed the best. GRNBoost defines the target gene's parents as the target's most predictive genes using a prediction importance score $G_{i,j}$ from gene $i$ to gene $j$. We adapted the GRNBoost score $G_{i,j}$ into a score $B_{i,j}$ in our proposed method, BetterBoost, which leverages interventional data in complement to observational data. The score $B_{i,j}$ reduces to $G_{i,j}$ when only observational data is available and improves as more interventional data becomes available.

BetterBoost assumes that if the dataset was generated by a causal model, the observed data's joint distribution can be factorized as:

$$p(\boldsymbol{x}_1 \dots \boldsymbol{x}_G) = \prod_{i=1}^{G} p(\boldsymbol{x}_i|\mathrm{Pa}(\boldsymbol{x}_i)). \tag{1}$$

If a candidate gene is a parent of the target, it will be a good predictor for the target, as GRNBoost assumes. But with labeled, interventional data, one can attempt to identify the true causal parents of a given observed variable $\boldsymbol{x}_i$ by looking at the effects of interventions on the candidate parents of $\boldsymbol{x}_i$. In particular, in a sample where a candidate parent gene is knocked down, the perturbed gene will only remain a good predictor for the target gene if it is a true causal parent of the target. Hence, if knocking down a candidate gene leads to a statistically significant prediction of the target gene,

it indicates strong evidence of a causal relationship directed from the candidate parent to the target gene. We leverage the impact of knocking down candidate genes in the prediction importance score of BetterBoost.

We find that BetterBoost performs significantly better than leading methods GRNBoost (Passemiers et al., 2022) and DCDI (Brouillard et al., 2020) on provided sample data according to the challenge metric, average Wasserstein distance. Below, we detail the proposed method and go over the preliminary results of BetterBoost and relevant baselines on sample datasets.

## 2 METHODS

In this section, we restate the objective of the challenge and detail the algorithm, BetterBoost.

### 2.1 OBJECTIVE

The considered single-cell perturbational datasets each consist of a matrix of UMI counts per cell, $\boldsymbol{X} \in \mathbb{Z}^{+^{N \times G}}$, and associated interventional labels, $\boldsymbol{s} \in \{\text{unperturbed}, \text{unlabeled}, 1, \ldots, G\}^N$, for each cell. Note the interventions can only affect at most one gene, which can be achieved via high-precision CRISPRi technology (Larson et al., 2013). We denote the fraction of genes $g \in [G]$ with labeled interventional data as $\rho$.

Since ground truth causal network data does not exist for these datasets, a proposed causal graph is evaluated by the average Wasserstein distance which is defined as follows: for each edge in the inferred causal graph $(i, j) \in \hat{\mathcal{G}}$, the Wasserstein distance is computed between the distribution of $X_j$ in the unperturbed data and in the subset of data where $X_i$ is perturbed. Therefore, the average Wasserstein distance can be written as:

$$d(\hat{\mathcal{G}}) := \frac{1}{|\hat{\mathcal{G}}|} \sum_{(i,j) \in \hat{\mathcal{G}}} W_1(p(\boldsymbol{x}_j | \boldsymbol{s} = \text{unperturbed}), p(\boldsymbol{x}_j | \boldsymbol{s} = i)) \tag{2}$$

where $W_1$ denotes the first Wasserstein distance between two distributions.

The space of valid causal graphs, $\hat{\mathcal{G}}$ is constrained to $\{\hat{\mathcal{G}} : |\hat{\mathcal{G}}| \geq 1000\}$, but otherwise can include cycles and disconnected components.

### 2.2 ALGORITHM

We found GRNBoost to work the best in the observational case, i.e. no labeled interventional data, but fail to improve on this metric after adding strictly more information in the form of intervention labels. Thus, we developed a simple procedure for leveraging any available intervention labels. As previously mentioned, we assume that the true causal graph, $\mathcal{G}$ is a directed, acyclic graph (DAG), and therefore the joint distribution factorizes as in Equation 1. To identify if gene $j \in [G]$ is a strong candidate parent gene for a given target gene $i \in [G]$, we look if $j$ is predictive of the target gene $i$ in the dataset formed by observational data and the interventional data on gene $j$. For a true causal parent, we expect that when $j$ is knocked down, there will be a statistically significant shift in the distribution of observed UMIs of gene $i$ between observational and interventional data. Since we held no priors on the nature of causal effects, we chose to use the Kolmogorov-Smirnov (KS) test (Massey, 1951) to test these distributional shifts between observational and interventional data. Additionally, we used the Benjamini-Hochberg procedure to correct the p-values for multiple testing (Benjamini & Hochberg, 1995).

To formulate the new score used by BetterBoost to rank the impact of gene $i$ on gene $j$, we write $G_{i,j}$ the predictive score of gene $i$ on gene $j$ computed by GRNBoost, and $p_{i,j}$ the Benjamini-Hochberg corrected KS test p-value of the impact of knocking down gene $i$ on gene $j$. If no interventional data was available on $i$, we set all p-values $p_{i,*}$ to $0.05$, as to neither strongly accept nor reject hypotheses for these interactions. We then define the score $B_{i,j} = (-p_{i,j}, G_{i,j})$ that we sort from larger to smaller (in lexicographic order).

For some desired number of edges, $K$, BetterBoost returns the $K_B := \min(K, |\{(i, j) : B_{i,j}[0] \geq -0.05\}|)$ candidate edges with the smallest $H$ score and acceptable p-values. The $K_B$ candidate

Table 1: Average Wasserstein Distance of Methods on RPE1 Perturb-seq dataset

| **Method** | $\rho = 0$ | $\rho = 0.25$ | $\rho = 0.5$ | $\rho = 0.75$ | $\rho = 1.0$ |
|---|---|---|---|---|---|
| DCDI | 0.126 | 0.126 | 0.127 | 0.125 | 0.130 |
| GRNBoost | 0.115 | 0.106 | 0.106 | 0.106 | 0.106 |
| GRNBoost-1000 | **0.151** | 0.147 | 0.146 | 0.146 | 0.145 |
| BetterBoost | **0.151** | **0.398** | **0.531** | **0.599** | **0.636** |

edges will have the smallest p-values for the KS test up to 0.05, which can include gene pairs where no interventional data and hence no p-value was available. Since the p-values of these gene pairs were set to $0.05$, this ranking will favor in practice the edges of pairs with small p-values (obtained from combined interventional and observation data) followed by the edges with the highest GRNBoost scores $G_{i,j}$ (from observational data only). Typically, this results in more of the final edges being chosen by p-value than by GRNBoost score as more labeled interventional data becomes available.

## 3 RESULTS

We compared BetterBoost to the two suggested baseline methods, GRNBoost and DCDI, on the RPE1 perturbational data from (Replogle et al., 2022). The methods were evaluated with varying fractions of available labeled interventional data, ranging from $0.25$ to $1.0$. In order to comply with the challenge requirements, we choose to return $K = 1000$ edges for the challenge. By default, GRNBoost returns all edges with non-zero importance, so we additionally tested against a variant of GRNBoost that only returns the $1000$ top importance edges.

We found that for every fraction of labeled interventional data, $\rho$ considered, BetterBoost improved significantly on the average Wasserstein metric. Additionally, we found that the improvement in the metric correlated perfectly with $\rho$ as shown in Table 1.

*Remark: We haven't tuned DCDI; the reported results are from running the provided baseline.*

## 4 DISCUSSION

Our proposed method, BetterBoost, utilizes labeled interventional data to identify the true causal parents of a given observed variable by looking at the effects of interventions on candidate parents. BetterBoost significantly outperforms leading methods GRNBoost and DCDI on provided sample data according to the challenge metric, average Wasserstein distance. In conclusion, our results suggest that BetterBoost is a promising gene regulatory network inference method.

BetterBoost can be extended for future work to consider the invariance property of causal relationships mentioned previously. Currently, if a chain of strong causal effects exists, $\boldsymbol{x}_i \to \boldsymbol{x}_j \to \boldsymbol{x}_k$, BetterBoost will likely assign an edge from $\boldsymbol{x}_i \to \boldsymbol{x}_k$. However, if the interventional data on $\boldsymbol{x}_j$ is present and labeled, one can identify that an edge does not exist between $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$. This scenario also exposes a shortcoming of the average Wasserstein metric, which would not penalize the presence of such an edge in the inferred graph.

## REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. July 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300, 1995. doi: http://dx.doi.org/10.2307/2346101. URL `http://dx.doi.org/10.2307/2346101`.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. July 2020.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable Single-Cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, December 2016.

Matthew H Larson, Luke A Gilbert, Xiaowo Wang, Wendell A Lim, Jonathan S Weissman, and Lei S Qi. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.*, 8(11):2180–2196, November 2013.

F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.

Antoine Passemiers, Yves Moreau, and Daniele Raimondi. Fast and accurate inference of gene regulatory networks through robust precision matrix estimation. *Bioinformatics*, 38(10):2802–2809, May 2022.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. January 2015.

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, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L Bonnar, Marco Jost, Thomas M Norman, and Jonathan S Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. May 2022.

Muralikrishnna G Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. NODAGS-Flow: Nonlinear cyclic causal structure learning. January 2023.