

A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data

Corresponding Author: Mr Mathieu Chevalley

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Summary: The paper introduces CausalBench, a comprehensive benchmark suite for evaluating causal network inference methods on large-scale single-cell perturbation data. CausalBench includes two extensive datasets from different cell lines, novel biologically-motivated and statistical evaluation metrics, and implementations of various state-of-the-art causal inference methods. The authors use CausalBench to evaluate existing methods, revealing limitations in scalability and utilization of interventional data. They also report on a machine learning challenge that led to the development of new methods with significantly improved performance. According to the authors, the benchmark aims to accelerate progress in causal graph inference for biological networks and potentially impact drug discovery and disease understanding.

Strengths: I note below a handful of strengths of the manuscript, as I recognized them:

1. The benchmark addresses a critical need in the field of causal inference for biological networks, by providing a standardized evaluation framework using real-world data.
2. The inclusion of both biologically-based and statistical evaluation metrics provides a well-rounded assessment of method performance.
3. The large-scale nature of the datasets used (over 200,000 data points) allows for robust evaluation of method scalability and performance.
4. The benchmark's design allows for easy integration of new methods, promoting ongoing research and development.
5. The authors demonstrate the benchmark's utility through a machine learning challenge, which led to significant improvements in the state-of-the-art.

Weaknesses: Next, I mention some weaknesses:

1. The ground truth for the biological networks is not fully known, which limits the accuracy of the evaluation to some extent.
2. The benchmark focuses mostly on acyclic graph structures, which may not fully capture the complexity of biological systems that often include feedback loops.
3. The assumption of causal sufficiency may not hold in the datasets. This can be potentially impactful for the interpretation of some methods.
4. The paper does not extensively discuss the limitations of using single-cell RNA sequencing data for causal inference, such as potential confounding factors or technical biases.
5. A potential issue is the following: The reliance on existing biological databases for evaluation may perpetuate biases toward well-studied pathways and genes.

Questions:

1. Have the authors considered extending the benchmark to include multi-omics data or time-series data to capture more complex biological relationships?
2. How does the performance of the methods vary when applied to different cell types or biological contexts?
3. Maybe a constructive question, given the potential of the proposed benchmark: How do the authors plan to keep the benchmark updated as new biological knowledge and improved ground truth data become available?

Reviewer #2

(Remarks to the Author)

The authors present CausalBench, a benchmarking framework for network inference on single-cell perturbation data. In this work, they also present two statistical evaluation metrics for assessing the inferred networks.

1) The logic of the two evaluation metrics makes sense for the case of perturbation data.

However, I feel it needs more comparisons to the usual evaluation metric (use of gold standard networks). While the evaluation results using gold standard networks are also present, comparing the rank of methods in all 3 metrics would be more helpful. The Wasserstein distance and FOR rank are present in two of the tables, it would help the comparison if the F1 score rank is also reported there.

2) For biological evaluations, it might be useful to include TF binding information (at least that's what I'm used to seeing in network inference evaluations).

I don't know if there are many ChIP-seq experiments in the two cell lines used here, but it might be worth checking out resources such as ENCODE, cistrome, and ChIP-Atlas. There might be other resources too. If there are not enough data to make a ChIP-based gold standard, that should be mentioned in the discussion.

3) Would SCENIC perform better than GRNBoost?

GRNBoost was part of the SCENIC pipeline, and I assume the complete pipeline would perform better.

I think reporting the result of SCENIC on the observational data would be helpful.

On a related note, using correlation (between TFs and target genes) could be a helpful baseline too.

4) I'm not familiar with most methods compared here, but most GRN methods that I've seen provide a ranking for the inferred edges, which then can be used for creating a precision-recall curve and calculating AUPR (which is probably a better measure compared to a single precision and recall value). Was this not possible for the methods used here?

If a method does not report edge confidence, it can be run in a bootstrap or stability selection framework to calculate the confidence values. Considering the reported run times this should be feasible for many of the methods. I understand that doing this might not be in the scope of this work, but it would be worth mentioning in the discussions.

Minor points:

- It might be easier to read if you merge table 1 and table 2 (and then add rank in K562 and RPE1 as additional columns).

- When discussing related work, mentioning older works like the DREAM5 challenge (Marbach et al 2012 nature methods) could be helpful.

- There are no citations for the methods in section 4.3.4.

- While this is not related to the scope of this paper, mentioning methods that use scRNA-seq and scATAC-seq to infer regulatory networks (SCENIC+ for example) could be helpful for the readers.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I would like to thank the authors for the detailed rebuttal letter, as well as the revised version of Causalbench.

I can firmly say that the authors addressed my comments and concerns about the original version, and I believe that Causalbench can serve as a great contribution to single-cell perturbation data.

I recommend the acceptance for publication.

Reviewer #2

(Remarks to the Author)

I'm satisfied by the authors response to my comments. I have no further comments.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Authors' Response to Reviews of

CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data

Mathieu Chevalley, Yusuf H Roohani, Arash Mehrjou, Jure Leskovec, Patrick Schwab
Communications Biology

RC: Reviewers' Comment, **AR:** Authors' Response

To everyone

We would like to thank the editor and reviewers for providing valuable insights and feedback on our manuscript. In response to the reviewers' comments and suggestions, we have diligently revised the manuscript, addressing each point raised. As suggested by the reviewers, we focused on extending the discussion, analyzing the limitations of the evaluation, the benchmarked methods and the data available. To make the reading experience smoother, the edited and added parts in the revised manuscript are colored blue for better visibility.

We now respond to each reviewer's point and discuss the related revisions, additional experiments or evaluations.

Reviewer #1

RC: The benchmark focuses mostly on acyclic graph structures, which may not fully capture the complexity of biological systems that often include feedback loops. [...] The assumption of causal sufficiency may not hold in the datasets. This can be potentially impactful for the interpretation of some methods.

AR: Thank you for your insightful comment regarding the focus of the benchmark on acyclic graph structures. We concur that acyclicity and causal sufficiency are common assumptions in the causal inference literature, and these may not always hold true in single-cell RNA datasets. We would like to clarify that these are limitations of the methods being evaluated, rather than the benchmark itself. Our benchmark does not enforce that the output of the methods must be acyclic. We acknowledge that this distinction may not have been sufficiently clear in the original manuscript, and we have made revisions to address this. In particular, we have highlighted that the performance of the challenge methods may be partly attributed to their allowance for cyclic graphs. This is an important point, as it underscores the potential for methods that can handle more complex, cyclic biological systems to perform well in our benchmark.

Related revisions:

1. ll 257-260: For instance, feedback loops between genes are a well-known phenomenon in gene regulation [1, 2] that unfortunately cannot be represented by existing causal network inference methods at present that enforce acyclicity, Revised: which is the case for e.g. SparseRC and DCDI.
2. ll 264-269: Revised: We can hypothesize that those limitations are strong obstacles for causality methods, as the challenge methods, such as Mean Difference and Guanlab, which do not enforce

acyclicity and are also by design less sensitive to lack of causal sufficiency, perform much better. This is indeed an important takeaway message of our benchmarking that direct the community towards models that inherently allow feedback loops as they also empirically outperform those that impose acyclicity.

RC: The paper does not extensively discuss the limitations of using single-cell RNA sequencing data for causal inference, such as potential confounding factors or technical biases.

AR: We appreciate the importance of highlighting the challenges inherent to single-cell RNA sequencing data, as those are serious limitations for causal inference methods that should be considered by method developers. To highlight those limitations, we have moved the text of section 4.1.3 to the discussion section and extended it. This part discusses the particular distribution of gene expression data which is heavy tailed at 0, where those zeros can be both biological as well as technical artefacts. We also address batch effects as well as the confounding of asynchronous sampling.

Related revisions:

1. ll 270-279: Revised: Single-cell data presents idiosyncratic challenges that may break the common assumptions of many existing methods. Apart from the high-dimensionality in terms of number of variables and large sample size, the distribution of the gene expression present a challenge as it is highly tailed at 0 for some genes [3], with those zeros being both biological as well as technical (dropout effect [4]). Single-cell data often has much higher technical noise and variability than bulk RNA-seq, due to factors like amplification noise, low mRNA capture efficiency, and sequencing depth [5]. In addition, different cells sampled from the same batch may not be truly independent as the cells may have interacted and influenced their states. Lastly, cells in scRNAseq experiments are measured at a fixed point in time and may therefore have been sampled at various points in their developmental trajectory or cell cycle [6] - making sampling time a potential confounding factor in any analysis of scRNAseq data.
2. ll 369-377: ~~Single-cell data presents idiosyncratic challenges that may break the common assumptions of many existing methods. Apart from the high-dimensionality in terms of number of variables and large sample size, the distribution of the gene expression present a challenge as it is highly tailed at 0 for some genes . Regarding the underlying true causal graph, biological feedback loops may break the acyclicity assumption underlying the use of DAGs. In addition, different cells sampled from the same batch may not be truly independent as the cells may have interacted and influenced their states. Lastly, cells in scRNAseq experiments are measured at a fixed point in time and may therefore have been sampled at various points in their developmental trajectory or cell cycle —making sampling time a potential confounding factor in any analysis of scRNAseq data.~~

RC: A potential issue is the following: The reliance on existing biological databases for evaluation may perpetuate biases toward well-studied pathways and genes.

AR: We thank the reviewer for bringing forward this potential issue, as we agree that this is a major risk of benchmarking metrics. In our study, we address this risk by pairing statistical and biological metrics alongside each other - methods that overly perpetuate biases of existing databases are therefore penalized in the statistical evaluation and their biases are therefore made transparent. For example, only predicting known interactions from databases would perform quite poorly, as the precision and recall of those is quite low for our datasets (see table 4, that we also added to the revised manuscript). The statistical evaluation is fully dataset specific and thus harshly penalizes methods that would rely heavily on prior knowledge. Furthermore,

on the evaluation side, we ensure that the biological metric is cell-specific by only including interactions that are validated by differential expression. As can be seen in table 4, this removes a lot of interactions that are not present in the considered cell lines. We can also see that the precision is lower on RPE1 for all databases, confirming that an evaluation based solely on existing biological databases would indeed be heavily biased, and also that such biases become clearly transparent using our evaluation methodology. We also acknowledge the difficulty of constructing a ground truth for less studied cell types, and have revised the manuscript to better discuss that aspect.

1. ll 249-255: Revised: Moreover, the reported interactions in the reference databases used to construct the biological ground-truth may be highly context dependent. To remedy this, we filter out interactions that are not validated by differential expression in the benchmark datasets. This filtering ensures that the biological evaluation is cell-specific. However, we can observe that this filtering greatly reduces the size of the true positive set, highlighting the difficulty of validating gene networks against prior knowledge. This is why in our study we greatly focus on developing a sound statistical evaluation, as we consider this approach to be less biased and fully context specific.

RC: Have the authors considered extending the benchmark to include multi-omics data or time-series data to capture more complex biological relationships?

AR: We thank the reviewer for suggesting extending the benchmark. We see two ways of including the proposed data modalities. 1. As an input to the model to improve the performance on the two datasets of the benchmark. 2. As datasets to be trained and evaluated on. For the first angle, we would like to note that the methods are not restricted in terms of using additional data sources, even though in this study we focused on evaluating methods that infer gene networks in a purely data-driven way. However, using transfer learning from other datasets or incorporating prior biological knowledge are promising directions. Moreover, the implementation of the benchmark does not prohibit a method from loading additional data, as is exemplified by the inclusion of SCENIC as proposed by reviewer #2. Regarding the second angle, the main limitation is the availability of e.g. perturbational time-series data, which to our knowledge, does not publicly exist at the scale considered in this study. We would of course gladly add such a dataset when it becomes available, which would constitute an exciting new area of research. We have revised the manuscript accordingly to clarify the possibility to leverage external datasets and the future potential for the benchmark to include perturbational data of different modalities.

Related revisions:

1. ll 331-333: Revised: New data dimensions can also be explored, either included by the methods itself through the integration of prior knowledge (e.g. ATAC-seq data), or new data modalities given as input (e.g. time-series data).

RC: How does the performance of the methods vary when applied to different cell types or biological contexts?

AR: We agree that robustness to cell types or biological context is paramount for an ideal method, and that this aspect was not sufficiently analyzed and discussed in the first version of the manuscript. To address this aspect, we have added a paragraph in subsection 2.2 *Characteristics of an optimal method for this task*. In our evaluation, we indeed observed that the top-performing methods, such as Mean Difference and Guanlab, demonstrated robust performance across the two cell types. This suggests that these methods are capable of effectively handling the variability in data distribution and expression patterns across different biological

contexts. However, it's important to note that our evaluation was based on only two cell types. While these results are promising, they do not definitively prove that these methods will perform equally well across all possible cell types or biological contexts. Therefore, we believe that extending the evaluation to include more diverse cell types and biological contexts would be beneficial. This would provide a more comprehensive assessment of the methods' robustness and generalizability. As more large-scale perturbational datasets become publicly available, we plan to incorporate them into our benchmark to facilitate this extended evaluation. In the added paragraph, we discuss the limitations of evaluating the robustness of the methods with only two cell types, as well as how the performance of the evaluated methods varies across the two cell types.

Related revisions:

1. ll 211-221: [Revised:](#)

Robust performance across cell type Another important characteristic of an optimal method for this task is the robustness of the performance across biological contexts or cell types. The data distribution and expression patterns can greatly differ across cell types and measurement technologies. As such, the performance of a particular method may not translate when applied to a new dataset not included in our benchmark. However, given that CausalBench contains data from two cell types, we can already assess the methods robustness, albeit in a limited way. When comparing the results between the two cell lines, we can observe that the performance of the evaluated methods is in general consistent across the two cell types. For example, Mean Difference and Guanlab perform highly on both of them. Nevertheless, evaluating on more datasets would be necessary to confidently conclude that the methods are robust. We envision that as more large scale perturbational dataset become publicly available, their inclusion into CausalBench will make the evaluation more thorough.

RC: [Maybe a constructive question, given the potential of the proposed benchmark: How do the authors plan to keep the benchmark updated as new biological knowledge and improved ground truth data become available?](#)

AR: we first warmly thank the reviewer for recognizing the long-term potential of CausalBench to serve as a tool to track progress in gene network inference. We acknowledge that to achieve that vision, the long-term maintenance and improvement of the benchmark needs to be carefully considered. To that extent, we have revised the discussion section to clearly state our commitment to keeping the tool up-to-date, as well the possibility for the community as a whole to contribute to it. We have thus also added a section in the README of the GitHub repository ([link](#)), inviting external contributions as well as describing a way to recognize those contributions.

Related revisions:

1. ll 327-335: [Revised:](#) We also aim for CausalBench to remain at the forefront of benchmarking of bioinformatics methods tailored to perturbational biological datasets. The modularity of its implementation makes it straightforward to add new large scale perturbational datasets as they become publicly available. The biological evaluation can also be readily updated as new or significantly expanded databases become available. New data dimensions can also be explored, either included by the methods itself through the integration of prior knowledge (e.g. ATAC-seq data), or new data modalities given

as input (e.g. time-series data). Such extensions may come from us as well as from community contributions directly to the public GitHub repository of CausalBench, as we are deeply committed to the long-term utility of our tool.

Reviewer #2

- RC:** While the evaluation results using gold standard networks are also present, comparing the rank of methods in all 3 metrics would be more helpful. The Wasserstein distance and FOR rank are present in two of the tables, it would help the comparison if the F1 score rank is also reported there.
- AR:** We thank you for the suggestion. We agree that directly comparing the two metrics was difficult given that they are separated into three tables. We have thus revised the ranking table to include a column with the biological evaluation rank. We recapitulate the new table also here (table 2).
- RC:** For biological evaluations, it might be useful to include TF binding information (at least that's what I'm used to seeing in network inference evaluations). I don't know if there are many ChIP-seq experiments in the two cell lines used here, but it might be worth checking out resources such as ENCODE, cistrome, and ChIP-Atlas. There might be other resources too. If there are not enough data to make a ChIP-based gold standard, that should be mentioned in the discussion.
- AR:** Thank you for your valuable suggestion to include transcription factor (TF) binding information in our biological evaluations. We agree that this information can provide additional insights into the performance of the network inference methods. In response to your comment, we have explored resources such as ENCODE, cistrome, and ChIP-Atlas for available ChIP-seq experiments in the two cell lines used in our study. We found that there is indeed ChIP-seq data available for the K562 cell line. We have therefore included this data in our biological evaluation, which has further enriched our understanding of the performance of the evaluated methods. For the RPE1 cell line, however, we found that ChIP-seq data is less extensively available. Given the limited data, we decided to utilize networks derived from a more extensively studied epithelial cell line, HepG2, given that RPE1 is also an epithelial cell line. This approach allowed us to construct a meaningful evaluation for the RPE1 cell line despite the limited availability of ChIP-seq data. See tables 5 and 6 for the updated results on the biological evaluation. We have mentioned this approach and its rationale in our revised manuscript. We would like to also note the evaluation of table 4, which shows the precision in terms of interactions validated by differential expression in the data for each data source. This analysis shows that the ChIPSeq network precision is lower than the other sources, indicating a high number of false positives. This observation is in line with many existing analysis of the high number of false positives of ChIP-seq data [7, 8]. We nevertheless decided to include this additional ground-truth source. We appreciate your suggestion, as it has helped us to enhance the robustness and comprehensiveness of our evaluation.

Related revisions:

1. ll 583-591: Revised: For the K562 cell line, we employed data from the ChIP-Atlas [9] and ENCODE [10] databases to build a ChIPSeq network, restricting the links to those relevant to or recorded in the K562 cell line. However, the RPE1 cell line has not been as comprehensively characterized in existing research. To address this, we utilized networks derived in a similar manner, but from a more extensively studied epithelial cell line, HepG2, given that RPE1 is also an epithelial cell line. However, we can observe that the precision of the ChIPSeq network is low, indicating a high number of false

positives (see table 4). This observation is in line with many existing analysis of the high number of false positives of ChIP-seq data [7, 8]. We nevertheless include it such as to build a comprehensive ground-truth.

RC: Would SCENIC perform better than GRNBoost? GRNBoost was part of the SCENIC pipeline, and I assume the complete pipeline would perform better. I think reporting the result of SCENIC on the observational data would be helpful. On a related note, using correlation (between TFs and target genes) could be a helpful baseline too.

AR: As per your suggestion, we have added two models to the evaluations, namely GRNBoost + TF and SCENIC. SCENIC corresponds to the full pipeline. GRNBoost + TF adds known TF names as a prior information and thus focuses on predicting TF - regulon interaction based on predictability. The performance of those two models has been included in all plots and tables (see tables 1 to 3 and 5 to 9 and figs. 1 and 2. As can be observed, the performance of those two models is quite poor. Firstly, it achieves low recall as it only focuses on one type of interaction, namely TF - regulon. Secondly, among the predicted interactions, the precision is rather poor, which indicates that correlation between a TF and a gene in the observational data is not predictive of a causal effect when intervening on the TF.

Related revisions:

1. ll 108-109: and GRNBoost, Revised: GRNBoost + TF and SCENIC [11] (tree-based GRN inference methods at different step of the SCENIC pipeline)
2. ll 132-134: Revised: However, GRNBoost + TF and SCENIC have much lower FOR, as restricting the predictions to transcription factors-regulon interactions misses a lot of interactions of different types.
3. ll 168-170: Unfortunately, many tested methods, with the exception of NOTEARS, GRNBoost Revised: (+ TF), SCENIC, DCDFG,
4. ll 513-520: GRNBoost is a GRN specific Gradient Boosting tree method, where for every gene, candidate parent gene are ranked based on their predictive power toward the expression profile of the downstream gene. As such, it acts as a feature selection method toward learning the graph. GRNBoost was identified as one of the best performing GRN method in previous observational data based benchmarks [12]. Revised: GRNBoost is a step of the SCENIC pipeline. We thus also evaluate GRNBoost + TF, which consists in restricting the putative parents to a list of known transcription factors. SCENIC additionally leverages known motif to reduce the number of false positives.

RC: I'm not familiar with most methods compared here, but most GRN methods that I've seen provide a ranking for the inferred edges, which then can be used for creating a precision-recall curve and calculating AUPR (which is probably a better measure compared to a single precision and recall value). Was this not possible for the methods used here? If a method does not report edge confidence, it can be run in a bootstrap or stability selection framework to calculate the confidence values. Considering the reported run times this should be feasible for many of the methods. I understand that doing this might not be in the scope of this work, but it would be worth mentioning in the discussions.

AR: Your point about uncertainty quantification is indeed crucial, as practitioners are highly interested in having a measure of uncertainty in the predicted interactions. As you already alluded to, we refrained from evaluation this aspect, as for most causal inference methods, there is no straightforward way of obtaining edge rankings

from those models. Most theoretically grounded methods which provides a distribution over possible graphs are based on the Bayesian framework. Unfortunately, Bayesian inference is notoriously hard to scale, making it unsuitable for the problem at hand [13, 14]. However, we would like to note that most of the challenge methods, such as Mean Difference and Guanlab, are based on ranking the edges by confidence. This is why we evaluate two versions for those methods, i.e. top 1k and top 5k interactions. We have revised the discussion accordingly to elaborate on that aspect and provide practitioners with actionable insights to obtain edge confidences.

Related revisions:

1. ll 289-298: Revised: In our study, we observed that some methods do not return edge confidence, especially from the causality literature, which is why we focused on evaluating point estimates as there is not always a theoretically sound way to obtain uncertainty estimates. This approach encourages methods to return the Pareto optimal trade-off, providing a balance between precision and recall. However, for practitioners interested in edge confidence, bootstrapping [15, 16] or stability selection [17] can be employed to calculate these values. It's important to note that the two best-performing methods in our benchmark, Mean Difference and Guanlab, are based on ranking the edges. This ranking inherently provides a measure of confidence in the predicted interactions, with higher-ranked edges being more confident. Therefore, even in the absence of explicit edge confidence values, these methods offer a form of confidence measure through their ranking system.

RC: - It might be easier to read if you merge table 1 and table 2 (and then add rank in K562 and RPE1 as additional columns). - When discussing related work, mentioning older works like the DREAM5 challenge (Marbach et al 2012 nature methods) could be helpful. - There are no citations for the methods in section 4.3.4. - While this is not related to the scope of this paper, mentioning methods that use scRNA-seq and scATAC-seq to infer regulatory networks (SCENIC+ for example) could be helpful for the readers.

AR: We thank you for the improvements suggestions and we have revised the manuscript accordingly. We merged table 1 and 2 (see table 1), discuss the DREAM5 challenge, mention methods that use scRNA-seq and scATAC-seq, and added citations in section 4.3.4.

Related revisions:

1. ll 443-451: Revised: The DREAM5 challenge [18] provided a landmark benchmark by evaluating GRN inference approaches on both observational microarray data and experimental perturbation data from *in vivo* gene knockout and knockdown studies, particularly for organisms like *E. coli* and *S. aureus*. While DREAM5 significantly advanced the field by incorporating experimental perturbations, it remained limited to bulk data and did not address the high-resolution insights enabled by large-scale, single-cell perturbational datasets. Our work, CausalBench, extends this approach by benchmarking GRN inference on expansive single-cell perturbational data, with biologically relevant metrics that offer deeper insights into causal gene regulatory mechanisms at the single-cell level.
2. ll 410-416: Revised: In recent years, several methods have also been developed to infer gene regulatory networks by integrating single-cell RNA sequencing and single-cell ATAC sequencing (scATAC-seq) data. Notably, SCENIC+ [19] is designed to construct enhancer-driven GRNs using combined or separate scRNA-seq and scATAC-seq data. Other methods have also been proposed, such as STREAM

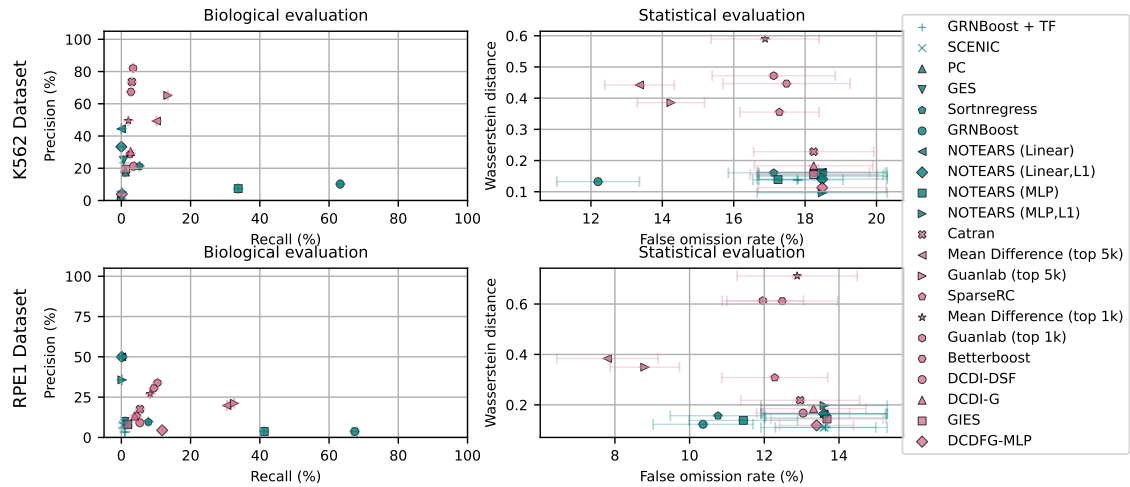


Figure 1: **Precision-Recall trade-off.** Performance comparison in terms of Precision (in %; y-axis) and Recall (in %; x-axis) in correctly identifying edges substantiated by biological interaction databases (left panels); and our own statistical evaluation using interventional information in terms of Wasserstein distance and FOR (right panels). Performance is compared across 10 different methods using observational data (green markers), and 11 different methods using interventional data (pink markers) in K562 (top panels) and RPE1 (bottom panels) cell lines. For each method, we show the mean and standard deviation from five independent runs.

[20], GLUE [21], DIRECT-NET [22], Pando [23] and scMEGA [24]. These approaches leverage the complementary information provided by gene expression and chromatin accessibility to enhance the accuracy of GRN inference.

References

- [1] Richard W Carthew. Gene regulation by micrnas. *Current opinion in genetics & development*, 16(2):203–208, 2006.
- [2] Michael Levine and Eric H Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942, 2005.
- [3] Sam Tracy, Guo-Cheng Yuan, and Ruben Dries. Rescue: imputing dropout events in single-cell rna-sequencing data. *BMC bioinformatics*, 20(1):1–11, 2019.
- [4] Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome biology*, 23(1):31, 2022.
- [5] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 11 2017.

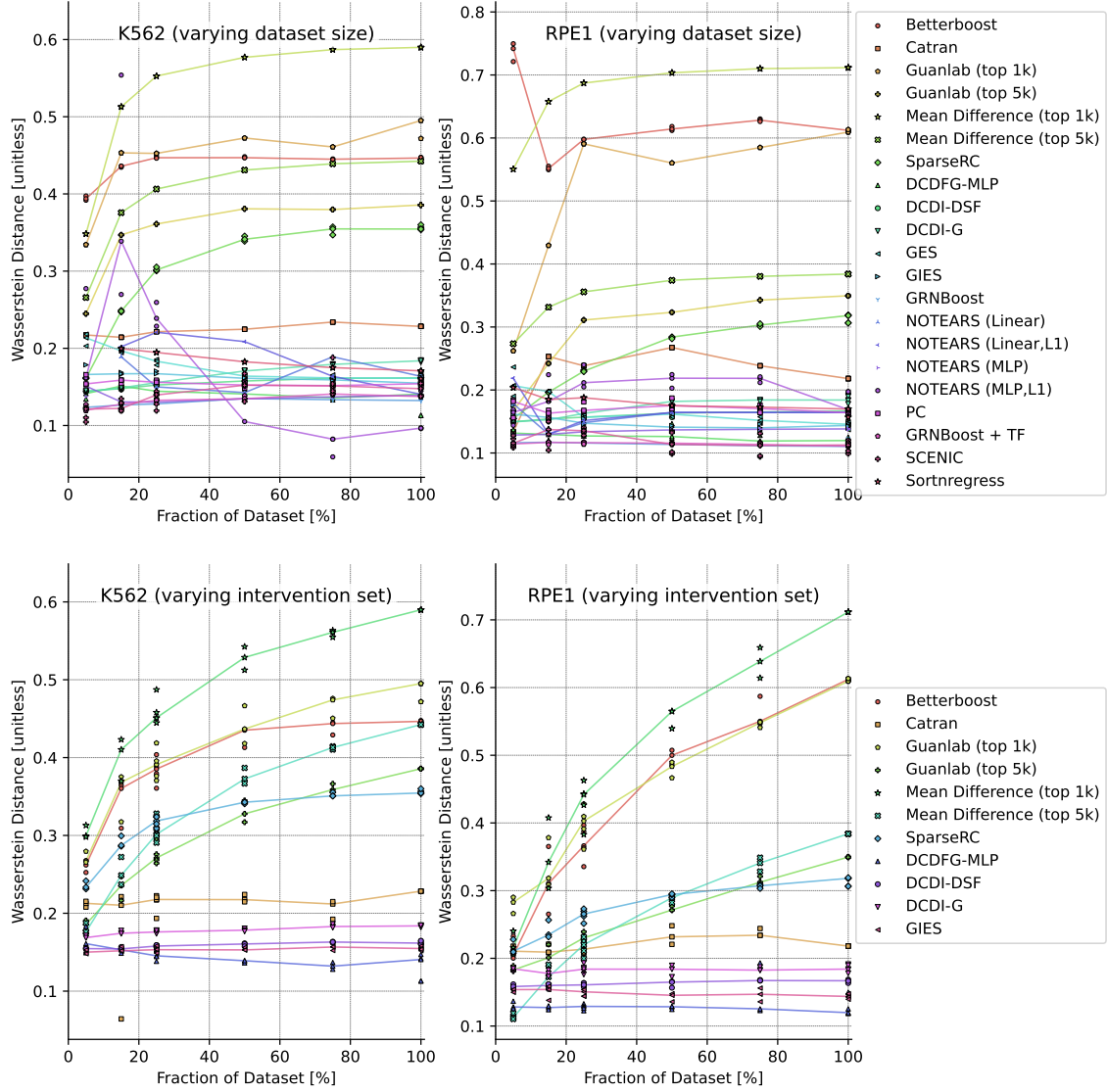


Figure 2: **Statistical evaluation of scaling characteristics.** Performance comparison in terms of Mean Wasserstein Distance (unitless; y-axis) of 10 methods for causal graph inference on observational data (top row; (see legend top right) and 11 methods on interventional scRNAseq data (centre row; see legend centre right) when varying the fraction of the full dataset size available for inference (in %; x-axis), and 4 methods on interventional data (bottom row; see legend bottom right) when varying the fraction of the full intervention set used (in %, x-axis). Markers indicate the values observed when running the respective algorithms with one of three random seeds, and colored lines indicate the median value observed across all tested random seeds for a method.

- [6] Monika S Kowalczyk, Itay Tirosh, Dirk Heckl, Tata Nageswara Rao, Atray Dixit, Brian J Haas, Rebekka K Schneider, Amy J Wagers, Benjamin L Ebert, and Aviv Regev. Single-cell rna-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research*, 25(12):1860–1872, 2015.
- [7] Dhawal Jain, Sandro Baldi, Angelika Zabel, Tobias Straub, and Peter B Becker. Active promoters give rise to false positive ‘phantom peaks’ in chip-seq experiments. *Nucleic acids research*, 43(14):6959–6968, 2015.
- [8] Joseph K Pickrell, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. False positive peaks in chip-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*, 27(15):2144–2146, 2011.
- [9] Zhaonan Zou, Tazro Ohta, Fumihito Miura, and Shinya Oki. Chip-atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating chip-seq, atac-seq and bisulfite-seq data. *Nucleic Acids Research*, page 1, 2022.
- [10] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2018.
- [11] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.
- [12] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.
- [13] Elaine Angelino, Matthew James Johnson, Ryan P Adams, et al. Patterns of scalable bayesian inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016.
- [14] Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24111–24123. Curran Associates, Inc., 2021.
- [15] David A Freedman. Bootstrapping regression models. *The annals of statistics*, 9(6):1218–1228, 1981.
- [16] Christopher Z Mooney, Robert D Duval, and Robert Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Number 95. sage, 1993.
- [17] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.
- [18] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.
- [19] Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, and Stein Aerts. Scenic+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature methods*, 20(9):1355–1367, 2023.

- [20] Yang Li, Anjun Ma, Yizhong Wang, Qi Guo, Cankun Wang, Hongjun Fu, Bingqiang Liu, and Qin Ma. Enhancer-driven gene regulatory networks inference from single-cell rna-seq and atac-seq data. *Briefings in Bioinformatics*, 25(5):bbae369, 2024.
- [21] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.
- [22] Lihua Zhang, Jing Zhang, and Qing Nie. Direct-net: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Science advances*, 8(22):eabl7393, 2022.
- [23] Jonas Simon Fleck, Sophie Martina Johanna Jansen, Damian Wollny, Fides Zenk, Makiko Seimiya, Akanksha Jain, Ryoko Okamoto, Malgorzata Santel, Zhisong He, J Gray Camp, et al. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, 621(7978):365–372, 2023.
- [24] Zhijian Li, James S Nagai, Christoph Kuppe, Rafael Kramann, and Ivan G Costa. scmega: single-cell multi-omic enhancer-based gene regulatory network inference. *Bioinformatics Advances*, 3(1):vbad003, 2023.

Table 1: Ranking of Models Based on Mean F1 Score on the biological evaluation for K562 and RPE1. The mean is over five seeds with standard deviation.

Model	K562			RPE1		
	Mean F1	Std	Rank	Mean F1	Std	Rank
Guanlab (top 5k)	0.223	0.000	1	0.257	0.000	1
Mean Difference (top 5k)	0.168	0.000	3	0.240	0.000	2
Guanlab (top 1k)	0.065	0.000	6	0.160	0.000	3
GRNBoost	0.175	0.001	2	0.069	0.000	8
Sortnregress	0.084	0.000	5	0.086	0.000	6
NOTEARS (MLP)	0.121	0.000	4	0.068	0.000	9
Betterboost	0.053	0.000	9	0.144	0.001	4
Catran	0.058	0.000	8	0.082	0.000	7
DCDI-DSF	0.060	0.001	7	0.067	0.002	10
Mean Difference (top 1k)	0.039	0.000	12	0.127	0.000	5
SparseRC	0.046	0.000	11	0.067	0.000	11
DCDI-G	0.049	0.001	10	0.059	0.003	13
GIES	0.023	0.002	13	0.030	0.005	15
PC	0.022	0.000	14	0.031	0.002	14
DCDFG-MLP	0.003	0.000	17	0.064	0.000	12
GES	0.013	0.001	15	0.019	0.002	16
GRNBoost + TF	0.004	0.000	16	0.016	0.001	17
SCENIC	0.000	0.000	20	0.009	0.002	18
NOTEARS (Linear)	0.000	0.000	18	0.002	0.000	20
NOTEARS (Linear,L1)	0.000	0.000	19	0.002	0.000	20
NOTEARS (MLP,L1)	0.000	0.000	21	0.003	0.000	19

Table 2: Model ranking. We here present a simple unbiased way of ranking the different models based on the mean Wasserstein distance and the FOR. We separate the rankings per cell type (Top table: K562, bottom table: RPE1). First, we create a preliminary ranking for each evaluation metrics. We rank by the mean score across the five different seeds. Finally, for each model, we take their average rank across the evaluation specific rankings. This ranking thus gives the same weight to each evaluation method. We also indicate the confidence interval for both metrics. The fraction of interventions indicates what ratio of the gene intervention are used for inference. For purely observational methods, this ratio is 0. For interventional methods, we run them on both high (100%) and low (25%) intervention settings. [Revised: We recapitulate the methods F1 score rank, for the observational methods and at 100% of interventions for the interventional methods.](#)

Model	Fraction Interventions	Mean Rank	Rank Wasserstein	Rank FOR	Wasserstein Distance	FOR	Rank F1 Score
Mean Difference (top 1k)	1.0	2.5	1	4	0.590 ± 0.000	0.169 ± 0.013	12.0
Mean Difference (top 5k)	1.0	3.5	5	2	0.442 ± 0.000	0.134 ± 0.009	3.0
Guanlab (top 5k)	1.0	4.5	6	3	0.386 ± 0.000	0.142 ± 0.008	1.0
Guanlab (top 1k)	1.0	4.5	2	7	0.472 ± 0.000	0.171 ± 0.015	6.0
Betterboost	1.0	8.0	4	12	0.447 ± 0.001	0.175 ± 0.016	9.0
Mean Difference (top 1k)	0.25	9.0	3	15	0.458 ± 0.015	0.179 ± 0.015	–
Guanlab (top 5k)	0.25	9.5	12	7	0.275 ± 0.011	0.171 ± 0.012	–
SparseRC	1.0	9.5	9	10	0.355 ± 0.001	0.173 ± 0.010	11.0
Mean Difference (top 5k)	0.25	10.5	11	10	0.306 ± 0.012	0.173 ± 0.012	–
SparseRC	0.25	10.5	10	11	0.321 ± 0.003	0.173 ± 0.012	–
Betterboost	0.25	12.5	8	17	0.383 ± 0.015	0.180 ± 0.014	–
Guanlab (top 1k)	0.25	12.5	7	18	0.383 ± 0.013	0.180 ± 0.015	–
Sortnregress	0.0	13.0	19	7	0.161 ± 0.000	0.171 ± 0.011	5.0
GRNBoost	0.0	15.5	30	1	0.132 ± 0.000	0.122 ± 0.010	2.0
DCDI-G	0.25	15.5	16	15	0.177 ± 0.001	0.179 ± 0.015	–
NOTEARS (MLP)	0.0	18.0	28	8	0.139 ± 0.000	0.172 ± 0.005	4.0
Catran	1.0	18.5	13	24	0.228 ± 0.000	0.182 ± 0.015	8.0
Catran	0.25	19.0	14	24	0.214 ± 0.010	0.182 ± 0.016	–
DCDI-DSF	0.25	19.0	21	17	0.158 ± 0.002	0.180 ± 0.016	–
DCDI-G	1.0	19.5	15	24	0.183 ± 0.002	0.182 ± 0.015	10.0
GRNBoost + TF	0.0	21.0	29	13	0.138 ± 0.001	0.178 ± 0.012	16.0
DCDI-DSF	1.0	21.0	18	24	0.162 ± 0.001	0.182 ± 0.014	7.0
DCDFG-MLP	0.25	22.5	26	19	0.149 ± 0.007	0.181 ± 0.016	–
GIES	1.0	23.0	22	24	0.155 ± 0.002	0.182 ± 0.016	13.0
GIES	0.25	24.5	24	25	0.153 ± 0.003	0.183 ± 0.014	–
PC	0.0	24.5	23	26	0.154 ± 0.002	0.184 ± 0.015	14.0
NOTEARS (Linear)	0.0	24.5	17	32	0.163 ± 0.000	0.185 ± 0.016	18.0
GES	0.0	26.0	20	32	0.160 ± 0.006	0.185 ± 0.016	15.0
SCENIC	0.0	28.5	25	32	0.150 ± 0.012	0.185 ± 0.016	20.0
NOTEARS (Linear,L1)	0.0	29.5	27	32	0.140 ± 0.000	0.185 ± 0.016	19.0
DCDFG-MLP	0.0	31.5	31	32	0.113 ± 0.000	0.185 ± 0.016	17.0
NOTEARS (MLP,L1)	0.0	32.0	32	32	0.097 ± 0.000	0.185 ± 0.016	21.0

Model	Fraction Interventions	Mean Rank	Rank Wasserstein	Rank FOR	Wasserstein Distance	FOR	Rank F1 Score
Mean Difference (top 5k)	1.0	3.5	6	1	0.384 ± 0.000	0.078 ± 0.012	2.0
Guanlab (top 1k)	1.0	4.5	2	7	0.613 ± 0.000	0.120 ± 0.010	3.0
Guanlab (top 5k)	1.0	5.0	8	2	0.350 ± 0.000	0.088 ± 0.008	1.0
Betterboost	1.0	6.5	3	10	0.611 ± 0.001	0.125 ± 0.013	4.0
Guanlab (top 1k)	0.25	7.0	5	9	0.399 ± 0.021	0.123 ± 0.007	–
Mean Difference (top 1k)	1.0	8.0	1	15	0.712 ± 0.000	0.129 ± 0.014	5.0
SparseRC	0.25	8.0	10	6	0.265 ± 0.007	0.118 ± 0.009	–
SparseRC	1.0	9.0	9	9	0.308 ± 0.004	0.123 ± 0.012	11.0
Betterboost	0.25	9.5	7	12	0.361 ± 0.019	0.127 ± 0.011	–
Mean Difference (top 1k)	0.25	10.0	4	16	0.431 ± 0.026	0.129 ± 0.012	–
Mean Difference (top 5k)	0.25	12.0	13	11	0.218 ± 0.011	0.126 ± 0.008	–
Guanlab (top 5k)	0.25	12.0	11	13	0.228 ± 0.009	0.128 ± 0.009	–
Sortnregress	0.0	13.5	23	4	0.157 ± 0.000	0.108 ± 0.011	6.0
Catran	1.0	14.5	12	17	0.218 ± 0.000	0.130 ± 0.014	7.0
GRNBoost	0.0	16.0	29	3	0.123 ± 0.000	0.104 ± 0.012	8.0
NOTEARS (MLP)	0.0	16.5	28	5	0.138 ± 0.000	0.114 ± 0.013	9.0
Catran	0.25	17.0	14	20	0.215 ± 0.007	0.132 ± 0.013	–
DCDI-DSF	0.25	18.0	22	14	0.161 ± 0.002	0.128 ± 0.009	–
DCDI-G	0.25	18.0	17	19	0.182 ± 0.006	0.130 ± 0.009	–
DCDI-DSF	1.0	18.5	18	19	0.167 ± 0.003	0.130 ± 0.011	10.0
DCDI-G	1.0	19.0	16	22	0.185 ± 0.006	0.133 ± 0.012	13.0
NOTEARS (MLP,L1)	0.0	22.0	15	29	0.197 ± 0.000	0.136 ± 0.015	19.0
PC	0.0	23.0	19	27	0.166 ± 0.003	0.136 ± 0.014	14.0
DCDFG-MLP	0.25	23.5	26	21	0.145 ± 0.022	0.132 ± 0.010	–
NOTEARS (Linear)	0.0	24.0	21	27	0.164 ± 0.000	0.136 ± 0.015	20.5
GIES	0.25	24.0	24	24	0.151 ± 0.006	0.135 ± 0.015	–
NOTEARS (Linear,L1)	0.0	25.0	21	29	0.164 ± 0.000	0.136 ± 0.015	20.5
DCDFG-MLP	0.0	26.5	30	23	0.119 ± 0.000	0.134 ± 0.009	12.0
GES	0.0	28.0	25	31	0.148 ± 0.008	0.136 ± 0.014	16.0
SCENIC	0.0	29.5	32	27	0.110 ± 0.008	0.136 ± 0.015	18.0
GIES	1.0	29.5	27	32	0.144 ± 0.003	0.137 ± 0.013	15.0
GRNBoost + TF	0.0	31.0	31	31	0.111 ± 0.001	0.136 ± 0.012	17.0

Table 3: **Model characteristics.** Table summarizing the characteristics of the test baselines. Graph scaling refers to the ability to infer on the complete set of genes. A method is deemed to have successfully achieved sample scaling and intervention scaling if its performance improves by at least 10% on both datasets. This comparison is made between settings where only a subset, specifically 25% of datapoints or perturbations, is used and settings where 100% of them are utilized. As such, only SparseRC fulfills all the criteria.

Model	Graph scaling	Sample scaling	Intervention scaling
PC	×	×	–
GES	×	×	–
GIES	×	×	×
NOTEARS	✓	×	–
DCDI-G	×	✓	×
DCDI-DSF	×	×	×
DCDI-FG	✓	×	×
GRNBoost	✓	×	–
GRNBoost + TF	✓	×	–
SCENIC	✓	×	–
Sortnregress	✓	×	–
Betterboost	✓	×	✓
Catran	✓	×	×
Guanlab	✓	×	✓
Mean Difference	✓	×	✓
SparseRC	✓	✓	✓

Table 4: Precision of each database in terms of percentage of links that are validated by differential expression in the datasets. CORUM, being the most curated database, has the highest precision. For all sources, the precision is much lower on RPE1 than on K562, confirming the difficulty of constructing a ground-truth for less studied cell lines. PPI (N) and PPI(P) surprisingly have similar precisions, even though PPI (N) contains much more interactions that are supposedly less certain.

Data source	Precision RPE1	Precision K562
ChiPSeq	12.92%	16.42%
PPI (N)	17.17%	25.43%
PPI (P)	17.49%	30.08%
CORUM	19.78%	39.18%

Table 5: **Biological evaluation results on RPE1.** Performance results on RPE1 Dataset for the biological evaluation, for each evidence type. Pooled correspond to pooling all the extracted databases into one network for evaluation.

Model	Precision					Recall				
	Pooled	CORUM	ChipSeq	PPI (N)	PPI (P)	Pooled	CORUM	ChipSeq	PPI (N)	PPI (P)
Betterboost	0.31±0.00	0.02±0.00	0.02±0.00	0.30±0.00	0.08±0.00	0.09±0.00	0.04±0.00	0.01±0.00	0.04±0.00	0.04±0.00
Catran	0.17±0.00	0.01±0.00	0.01±0.00	0.38±0.00	0.11±0.00	0.05±0.00	0.03±0.00	0.00±0.00	0.05±0.00	0.05±0.00
Guanlab (top 1k)	0.34±0.00	0.02±0.00	0.02±0.00	0.33±0.00	0.08±0.00	0.10±0.00	0.05±0.00	0.01±0.00	0.04±0.00	0.04±0.00
Guanlab (top 5k)	0.21±0.00	0.01±0.00	0.02±0.00	0.24±0.00	0.06±0.00	0.33±0.00	0.18±0.00	0.04±0.00	0.14±0.00	0.14±0.00
Mean Difference (top 1k)	0.27±0.00	0.02±0.00	0.01±0.00	0.27±0.00	0.07±0.00	0.08±0.00	0.04±0.00	0.01±0.00	0.03±0.00	0.03±0.00
Mean Difference (top 5k)	0.20±0.00	0.02±0.00	0.01±0.00	0.22±0.00	0.05±0.00	0.31±0.00	0.20±0.00	0.04±0.00	0.13±0.00	0.13±0.00
SparseRC	0.13±0.00	0.02±0.00	0.01±0.00	0.27±0.00	0.08±0.00	0.04±0.00	0.04±0.00	0.01±0.00	0.03±0.00	0.04±0.00
DCDFG-MLP	0.04±0.00	0.01±0.00	0.03±0.00	0.18±0.00	0.04±0.00	0.12±0.00	0.20±0.00	0.15±0.00	0.19±0.00	0.17±0.00
DCDI-DSF	0.09±0.00	0.01±0.00	0.01±0.00	0.20±0.01	0.06±0.00	0.05±0.00	0.05±0.01	0.01±0.00	0.05±0.00	0.05±0.00
DCDI-G	0.13±0.01	0.02±0.00	0.01±0.00	0.26±0.00	0.08±0.00	0.04±0.00	0.04±0.01	0.00±0.00	0.03±0.00	0.04±0.00
GIES	0.08±0.01	0.02±0.00	0.01±0.00	0.29±0.01	0.08±0.01	0.02±0.00	0.04±0.01	0.01±0.00	0.03±0.00	0.03±0.00
GES	0.10±0.01	0.02±0.01	0.01±0.01	0.36±0.03	0.11±0.02	0.01±0.00	0.02±0.01	0.00±0.00	0.01±0.00	0.02±0.00
GRNBoost	0.04±0.00	0.01±0.00	0.01±0.00	0.09±0.00	0.02±0.00	0.67±0.01	0.84±0.01	0.43±0.01	0.65±0.00	0.66±0.01
GRNBoost + TF	0.03±0.00	0.00±0.00	0.10±0.01	0.19±0.00	0.04±0.00	0.01±0.00	0.01±0.00	0.06±0.00	0.02±0.00	0.02±0.00
SCENIC	0.08±0.01	0.00±0.00	0.28±0.03	0.30±0.02	0.04±0.01	0.00±0.00	0.00±0.00	0.03±0.00	0.01±0.00	0.00±0.00
NOTEARS (Linear)	0.50±0.00	0.12±0.00	0.00±0.00	0.88±0.00	0.38±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (Linear,L1)	0.50±0.00	0.12±0.00	0.00±0.00	0.88±0.00	0.38±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (MLP)	0.04±0.00	0.01±0.00	0.01±0.00	0.14±0.00	0.03±0.00	0.41±0.00	0.80±0.00	0.24±0.00	0.61±0.00	0.62±0.00
NOTEARS (MLPL1)	0.36±0.00	0.14±0.00	0.00±0.00	1.00±0.00	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
PC	0.09±0.01	0.02±0.01	0.01±0.00	0.28±0.01	0.08±0.01	0.02±0.00	0.04±0.01	0.01±0.00	0.02±0.00	0.03±0.00
Sortnregress	0.10±0.00	0.02±0.00	0.02±0.00	0.29±0.00	0.08±0.00	0.08±0.00	0.12±0.00	0.02±0.00	0.09±0.00	0.10±0.00

Table 6: **Biological evaluation results on K562.** Performance results on K562 Dataset for the biological evaluation, for each evidence type. Pooled correspond to pooling all the extracted databases into one network for evaluation.

Model	Precision					Recall				
	Pooled	CORUM	ChipSeq	PPI (N)	PPI (P)	Pooled	CORUM	ChipSeq	PPI (N)	PPI (P)
Betterboost	0.67±0.01	0.27±0.00	0.00±0.00	0.62±0.01	0.43±0.00	0.03±0.00	0.06±0.00	0.00±0.00	0.01±0.00	0.03±0.00
Catran	0.74±0.00	0.88±0.00	0.00±0.00	0.99±0.00	0.91±0.00	0.03±0.00	0.21±0.00	0.00±0.00	0.02±0.00	0.07±0.00
Guanlab (top 1k)	0.82±0.00	0.39±0.00	0.00±0.00	0.75±0.00	0.52±0.00	0.03±0.00	0.09±0.00	0.00±0.00	0.02±0.00	0.04±0.00
Guanlab (top 5k)	0.65±0.00	0.18±0.00	0.00±0.00	0.62±0.00	0.32±0.00	0.13±0.00	0.21±0.00	0.01±0.00	0.07±0.00	0.12±0.00
Mean Difference (top 1k)	0.49±0.00	0.13±0.00	0.01±0.00	0.49±0.00	0.22±0.00	0.02±0.00	0.03±0.00	0.00±0.00	0.01±0.00	0.02±0.00
Mean Difference (top 5k)	0.49±0.00	0.12±0.00	0.01±0.00	0.48±0.00	0.23±0.00	0.10±0.00	0.14±0.00	0.01±0.00	0.05±0.00	0.08±0.00
SparseRC	0.29±0.00	0.19±0.00	0.00±0.00	0.45±0.00	0.26±0.00	0.03±0.00	0.09±0.00	0.00±0.00	0.02±0.00	0.04±0.00
DCDFG-MLP	0.04±0.00	0.00±0.00	0.00±0.00	0.39±0.00	0.07±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.00	0.00±0.00
DCDI-DSF	0.21±0.00	0.07±0.00	0.00±0.00	0.39±0.00	0.14±0.00	0.04±0.00	0.06±0.00	0.00±0.00	0.03±0.00	0.04±0.00
DCDI-G	0.30±0.01	0.12±0.00	0.01±0.00	0.44±0.01	0.18±0.00	0.03±0.00	0.06±0.00	0.00±0.00	0.02±0.00	0.03±0.00
GIES	0.19±0.01	0.11±0.00	0.01±0.00	0.52±0.01	0.19±0.01	0.01±0.00	0.04±0.00	0.00±0.00	0.02±0.00	0.02±0.00
GES	0.25±0.02	0.18±0.01	0.00±0.00	0.59±0.02	0.26±0.01	0.01±0.00	0.03±0.00	0.00±0.00	0.01±0.00	0.01±0.00
GRNBoost	0.10±0.00	0.02±0.00	0.01±0.00	0.21±0.00	0.07±0.00	0.63±0.00	0.81±0.01	0.32±0.01	0.70±0.00	0.71±0.00
GRNBoost + TF	0.05±0.00	0.01±0.00	0.05±0.01	0.15±0.01	0.02±0.00	0.00±0.00	0.00±0.00	0.02±0.00	0.00±0.00	0.00±0.00
SCENIC	0.03±0.02	0.01±0.02	0.00±0.00	0.36±0.04	0.02±0.02	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (Linear)	0.44±0.00	0.44±0.00	0.00±0.00	0.89±0.00	0.56±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (Linear,L1)	0.33±0.00	0.50±0.00	0.00±0.00	0.83±0.00	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (MLP)	0.07±0.00	0.02±0.00	0.01±0.00	0.28±0.00	0.08±0.00	0.34±0.00	0.55±0.00	0.30±0.00	0.66±0.00	0.62±0.00
NOTEARS (MLP,L1)	0.00±0.00	0.00±0.00	0.00±0.00	0.80±0.00	0.20±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
PC	0.18±0.00	0.10±0.00	0.01±0.00	0.46±0.02	0.16±0.01	0.01±0.00	0.04±0.00	0.00±0.00	0.02±0.00	0.02±0.00
Sortnregress	0.21±0.00	0.15±0.00	0.00±0.00	0.54±0.00	0.24±0.00	0.05±0.00	0.22±0.00	0.01±0.00	0.07±0.00	0.10±0.00

Table 7: **Partition sizes.** Partition sizes used for each model. -1 means that the graph was not partitioned.

Model name	partition size
PC	30
GES	30
GIES	30
NOTEARS (Linear)	-1
NOTEARS (Linear, L1)	-1
NOTEARS (MLP)	-1
NOTEARS (MLP, L1)	-1
DCDI-DSF	50
DCDI-G	50
DCDFG-MLP	-1
GRNBoost (+ TF)	-1
SCENIC	-1
Sortnregress	-1
Mean Difference	-1
Guanlab	-1
SparseRC	-1
Betterboost	-1
Catran	-1

Table 8: **Run time K562.** Run time in wall clock hours for each method on the K562 cell line.

Model	Run time (hours)
SparseRC	0.004
GRNBoost + TF	0.012
SCENIC	0.046
Sortnregress	0.087
Mean Difference (top 5k)	0.094
Mean Difference (top 1k)	0.094
GRNBoost	0.097
Catran	0.197
GIES	0.824
DCDFG-MLP	1.087
DCDI-DSF	1.575
PC	1.672
NOTEARS (MLP,L1)	1.705
Guanlab (top 5k)	2.921
Guanlab (top 1k)	3.113
GES	3.148
NOTEARS (Linear,L1)	4.974
DCDI-G	5.043
NOTEARS (Linear)	6.297
NOTEARS (MLP)	10.778
Betterboost	11.934

Table 9: **Run time RPE1.** Run time in wall clock hours for each method on the RPE1 cell line.

Model	Run time (hours)
GRNBoost + TF	0.011
SparseRC	0.018
GRNBoost	0.031
Sortnregress	0.035
Mean Difference (top 1k)	0.040
Mean Difference (top 5k)	0.041
SCENIC	0.050
Catran	0.166
DCDI-DSF	0.493
DCDFG-MLP	0.550
Guanlab (top 1k)	0.933
Guanlab (top 5k)	0.955
NOTEARS (MLP,L1)	1.421
GIES	1.450
NOTEARS (Linear,L1)	1.557
NOTEARS (Linear)	1.680
DCDI-G	1.955
GES	2.207
Betterboost	3.736
PC	4.686
NOTEARS (MLP)	12.217

Authors' Response to Reviews of

CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data

Mathieu Chevalley, Yusuf H Roohani, Arash Mehrjou, Jure Leskovec, Patrick Schwab
Communications Biology

RC: Reviewers' Comment, **AR:** Authors' Response

To everyone

We would like to thank the editor and reviewers for providing valuable insights and feedback on our manuscript. In response to the reviewers' comments and suggestions, we have diligently revised the manuscript, addressing each point raised. As suggested by the reviewers, we focused on extending the discussion, analyzing the limitations of the evaluation, the benchmarked methods and the data available. To make the reading experience smoother, the edited and added parts in the revised manuscript are colored blue for better visibility.

We now respond to each reviewer's point and discuss the related revisions, additional experiments or evaluations.

Reviewer #1

RC: The benchmark focuses mostly on acyclic graph structures, which may not fully capture the complexity of biological systems that often include feedback loops. [...] The assumption of causal sufficiency may not hold in the datasets. This can be potentially impactful for the interpretation of some methods.

AR: Thank you for your insightful comment regarding the focus of the benchmark on acyclic graph structures. We concur that acyclicity and causal sufficiency are common assumptions in the causal inference literature, and these may not always hold true in single-cell RNA datasets. We would like to clarify that these are limitations of the methods being evaluated, rather than the benchmark itself. Our benchmark does not enforce that the output of the methods must be acyclic. We acknowledge that this distinction may not have been sufficiently clear in the original manuscript, and we have made revisions to address this. In particular, we have highlighted that the performance of the challenge methods may be partly attributed to their allowance for cyclic graphs. This is an important point, as it underscores the potential for methods that can handle more complex, cyclic biological systems to perform well in our benchmark.

Related revisions:

1. ll 257-260: For instance, feedback loops between genes are a well-known phenomenon in gene regulation [1, 2] that unfortunately cannot be represented by existing causal network inference methods at present that enforce acyclicity, Revised: which is the case for e.g. SparseRC and DCDI.
2. ll 264-269: Revised: We can hypothesize that those limitations are strong obstacles for causality methods, as the challenge methods, such as Mean Difference and Guanlab, which do not enforce

acyclicity and are also by design less sensitive to lack of causal sufficiency, perform much better. This is indeed an important takeaway message of our benchmarking that direct the community towards models that inherently allow feedback loops as they also empirically outperform those that impose acyclicity.

RC: The paper does not extensively discuss the limitations of using single-cell RNA sequencing data for causal inference, such as potential confounding factors or technical biases.

AR: We appreciate the importance of highlighting the challenges inherent to single-cell RNA sequencing data, as those are serious limitations for causal inference methods that should be considered by method developers. To highlight those limitations, we have moved the text of section 4.1.3 to the discussion section and extended it. This part discusses the particular distribution of gene expression data which is heavy tailed at 0, where those zeros can be both biological as well as technical artefacts. We also address batch effects as well as the confounding of asynchronous sampling.

Related revisions:

1. ll 270-279: Revised: Single-cell data presents idiosyncratic challenges that may break the common assumptions of many existing methods. Apart from the high-dimensionality in terms of number of variables and large sample size, the distribution of the gene expression present a challenge as it is highly tailed at 0 for some genes [3], with those zeros being both biological as well as technical (dropout effect [4]). Single-cell data often has much higher technical noise and variability than bulk RNA-seq, due to factors like amplification noise, low mRNA capture efficiency, and sequencing depth [5]. In addition, different cells sampled from the same batch may not be truly independent as the cells may have interacted and influenced their states. Lastly, cells in scRNAseq experiments are measured at a fixed point in time and may therefore have been sampled at various points in their developmental trajectory or cell cycle [6] - making sampling time a potential confounding factor in any analysis of scRNAseq data.
2. ll 369-377: ~~Single-cell data presents idiosyncratic challenges that may break the common assumptions of many existing methods. Apart from the high-dimensionality in terms of number of variables and large sample size, the distribution of the gene expression present a challenge as it is highly tailed at 0 for some genes . Regarding the underlying true causal graph, biological feedback loops may break the acyclicity assumption underlying the use of DAGs. In addition, different cells sampled from the same batch may not be truly independent as the cells may have interacted and influenced their states. Lastly, cells in scRNAseq experiments are measured at a fixed point in time and may therefore have been sampled at various points in their developmental trajectory or cell cycle —making sampling time a potential confounding factor in any analysis of scRNAseq data.~~

RC: A potential issue is the following: The reliance on existing biological databases for evaluation may perpetuate biases toward well-studied pathways and genes.

AR: We thank the reviewer for bringing forward this potential issue, as we agree that this is a major risk of benchmarking metrics. In our study, we address this risk by pairing statistical and biological metrics alongside each other - methods that overly perpetuate biases of existing databases are therefore penalized in the statistical evaluation and their biases are therefore made transparent. For example, only predicting known interactions from databases would perform quite poorly, as the precision and recall of those is quite low for our datasets (see table 4, that we also added to the revised manuscript). The statistical evaluation is fully dataset specific and thus harshly penalizes methods that would rely heavily on prior knowledge. Furthermore,

on the evaluation side, we ensure that the biological metric is cell-specific by only including interactions that are validated by differential expression. As can be seen in table 4, this removes a lot of interactions that are not present in the considered cell lines. We can also see that the precision is lower on RPE1 for all databases, confirming that an evaluation based solely on existing biological databases would indeed be heavily biased, and also that such biases become clearly transparent using our evaluation methodology. We also acknowledge the difficulty of constructing a ground truth for less studied cell types, and have revised the manuscript to better discuss that aspect.

1. ll 249-255: Revised: Moreover, the reported interactions in the reference databases used to construct the biological ground-truth may be highly context dependent. To remedy this, we filter out interactions that are not validated by differential expression in the benchmark datasets. This filtering ensures that the biological evaluation is cell-specific. However, we can observe that this filtering greatly reduces the size of the true positive set, highlighting the difficulty of validating gene networks against prior knowledge. This is why in our study we greatly focus on developing a sound statistical evaluation, as we consider this approach to be less biased and fully context specific.

RC: Have the authors considered extending the benchmark to include multi-omics data or time-series data to capture more complex biological relationships?

AR: We thank the reviewer for suggesting extending the benchmark. We see two ways of including the proposed data modalities. 1. As an input to the model to improve the performance on the two datasets of the benchmark. 2. As datasets to be trained and evaluated on. For the first angle, we would like to note that the methods are not restricted in terms of using additional data sources, even though in this study we focused on evaluating methods that infer gene networks in a purely data-driven way. However, using transfer learning from other datasets or incorporating prior biological knowledge are promising directions. Moreover, the implementation of the benchmark does not prohibit a method from loading additional data, as is exemplified by the inclusion of SCENIC as proposed by reviewer #2. Regarding the second angle, the main limitation is the availability of e.g. perturbational time-series data, which to our knowledge, does not publicly exist at the scale considered in this study. We would of course gladly add such a dataset when it becomes available, which would constitute an exciting new area of research. We have revised the manuscript accordingly to clarify the possibility to leverage external datasets and the future potential for the benchmark to include perturbational data of different modalities.

Related revisions:

1. ll 331-333: Revised: New data dimensions can also be explored, either included by the methods itself through the integration of prior knowledge (e.g. ATAC-seq data), or new data modalities given as input (e.g. time-series data).

RC: How does the performance of the methods vary when applied to different cell types or biological contexts?

AR: We agree that robustness to cell types or biological context is paramount for an ideal method, and that this aspect was not sufficiently analyzed and discussed in the first version of the manuscript. To address this aspect, we have added a paragraph in subsection 2.2 *Characteristics of an optimal method for this task*. In our evaluation, we indeed observed that the top-performing methods, such as Mean Difference and Guanlab, demonstrated robust performance across the two cell types. This suggests that these methods are capable of effectively handling the variability in data distribution and expression patterns across different biological

contexts. However, it's important to note that our evaluation was based on only two cell types. While these results are promising, they do not definitively prove that these methods will perform equally well across all possible cell types or biological contexts. Therefore, we believe that extending the evaluation to include more diverse cell types and biological contexts would be beneficial. This would provide a more comprehensive assessment of the methods' robustness and generalizability. As more large-scale perturbational datasets become publicly available, we plan to incorporate them into our benchmark to facilitate this extended evaluation. In the added paragraph, we discuss the limitations of evaluating the robustness of the methods with only two cell types, as well as how the performance of the evaluated methods varies across the two cell types.

Related revisions:

1. ll 211-221: [Revised:](#)

Robust performance across cell type Another important characteristic of an optimal method for this task is the robustness of the performance across biological contexts or cell types. The data distribution and expression patterns can greatly differ across cell types and measurement technologies. As such, the performance of a particular method may not translate when applied to a new dataset not included in our benchmark. However, given that CausalBench contains data from two cell types, we can already assess the methods robustness, albeit in a limited way. When comparing the results between the two cell lines, we can observe that the performance of the evaluated methods is in general consistent across the two cell types. For example, Mean Difference and Guanlab perform highly on both of them. Nevertheless, evaluating on more datasets would be necessary to confidently conclude that the methods are robust. We envision that as more large scale perturbational dataset become publicly available, their inclusion into CausalBench will make the evaluation more thorough.

RC: [Maybe a constructive question, given the potential of the proposed benchmark: How do the authors plan to keep the benchmark updated as new biological knowledge and improved ground truth data become available?](#)

AR: we first warmly thank the reviewer for recognizing the long-term potential of CausalBench to serve as a tool to track progress in gene network inference. We acknowledge that to achieve that vision, the long-term maintenance and improvement of the benchmark needs to be carefully considered. To that extent, we have revised the discussion section to clearly state our commitment to keeping the tool up-to-date, as well the possibility for the community as a whole to contribute to it. We have thus also added a section in the README of the GitHub repository ([link](#)), inviting external contributions as well as describing a way to recognize those contributions.

Related revisions:

1. ll 327-335: [Revised:](#) We also aim for CausalBench to remain at the forefront of benchmarking of bioinformatics methods tailored to perturbational biological datasets. The modularity of its implementation makes it straightforward to add new large scale perturbational datasets as they become publicly available. The biological evaluation can also be readily updated as new or significantly expanded databases become available. New data dimensions can also be explored, either included by the methods itself through the integration of prior knowledge (e.g. ATAC-seq data), or new data modalities given

as input (e.g. time-series data). Such extensions may come from us as well as from community contributions directly to the public GitHub repository of CausalBench, as we are deeply committed to the long-term utility of our tool.

Reviewer #2

- RC:** While the evaluation results using gold standard networks are also present, comparing the rank of methods in all 3 metrics would be more helpful. The Wasserstein distance and FOR rank are present in two of the tables, it would help the comparison if the F1 score rank is also reported there.
- AR:** We thank you for the suggestion. We agree that directly comparing the two metrics was difficult given that they are separated into three tables. We have thus revised the ranking table to include a column with the biological evaluation rank. We recapitulate the new table also here (table 2).
- RC:** For biological evaluations, it might be useful to include TF binding information (at least that's what I'm used to seeing in network inference evaluations). I don't know if there are many ChIP-seq experiments in the two cell lines used here, but it might be worth checking out resources such as ENCODE, cistrome, and ChIP-Atlas. There might be other resources too. If there are not enough data to make a ChIP-based gold standard, that should be mentioned in the discussion.
- AR:** Thank you for your valuable suggestion to include transcription factor (TF) binding information in our biological evaluations. We agree that this information can provide additional insights into the performance of the network inference methods. In response to your comment, we have explored resources such as ENCODE, cistrome, and ChIP-Atlas for available ChIP-seq experiments in the two cell lines used in our study. We found that there is indeed ChIP-seq data available for the K562 cell line. We have therefore included this data in our biological evaluation, which has further enriched our understanding of the performance of the evaluated methods. For the RPE1 cell line, however, we found that ChIP-seq data is less extensively available. Given the limited data, we decided to utilize networks derived from a more extensively studied epithelial cell line, HepG2, given that RPE1 is also an epithelial cell line. This approach allowed us to construct a meaningful evaluation for the RPE1 cell line despite the limited availability of ChIP-seq data. See tables 5 and 6 for the updated results on the biological evaluation. We have mentioned this approach and its rationale in our revised manuscript. We would like to also note the evaluation of table 4, which shows the precision in terms of interactions validated by differential expression in the data for each data source. This analysis shows that the ChIPSeq network precision is lower than the other sources, indicating a high number of false positives. This observation is in line with many existing analysis of the high number of false positives of ChIP-seq data [7, 8]. We nevertheless decided to include this additional ground-truth source. We appreciate your suggestion, as it has helped us to enhance the robustness and comprehensiveness of our evaluation.

Related revisions:

1. ll 583-591: Revised: For the K562 cell line, we employed data from the ChIP-Atlas [9] and ENCODE [10] databases to build a ChIPSeq network, restricting the links to those relevant to or recorded in the K562 cell line. However, the RPE1 cell line has not been as comprehensively characterized in existing research. To address this, we utilized networks derived in a similar manner, but from a more extensively studied epithelial cell line, HepG2, given that RPE1 is also an epithelial cell line. However, we can observe that the precision of the ChIPSeq network is low, indicating a high number of false

positives (see table 4). This observation is in line with many existing analysis of the high number of false positives of ChIP-seq data [7, 8]. We nevertheless include it such as to build a comprehensive ground-truth.

RC: Would SCENIC perform better than GRNBoost? GRNBoost was part of the SCENIC pipeline, and I assume the complete pipeline would perform better. I think reporting the result of SCENIC on the observational data would be helpful. On a related note, using correlation (between TFs and target genes) could be a helpful baseline too.

AR: As per your suggestion, we have added two models to the evaluations, namely GRNBoost + TF and SCENIC. SCENIC corresponds to the full pipeline. GRNBoost + TF adds known TF names as a prior information and thus focuses on predicting TF - regulon interaction based on predictability. The performance of those two models has been included in all plots and tables (see tables 1 to 3 and 5 to 9 and figs. 1 and 2. As can be observed, the performance of those two models is quite poor. Firstly, it achieves low recall as it only focuses on one type of interaction, namely TF - regulon. Secondly, among the predicted interactions, the precision is rather poor, which indicates that correlation between a TF and a gene in the observational data is not predictive of a causal effect when intervening on the TF.

Related revisions:

1. ll 108-109: and GRNBoost, Revised: GRNBoost + TF and SCENIC [11] (tree-based GRN inference methods at different step of the SCENIC pipeline)
2. ll 132-134: Revised: However, GRNBoost + TF and SCENIC have much lower FOR, as restricting the predictions to transcription factors-regulon interactions misses a lot of interactions of different types.
3. ll 168-170: Unfortunately, many tested methods, with the exception of NOTEARS, GRNBoost Revised: (+ TF), SCENIC, DCDFG,
4. ll 513-520: GRNBoost is a GRN specific Gradient Boosting tree method, where for every gene, candidate parent gene are ranked based on their predictive power toward the expression profile of the downstream gene. As such, it acts as a feature selection method toward learning the graph. GRNBoost was identified as one of the best performing GRN method in previous observational data based benchmarks [12]. Revised: GRNBoost is a step of the SCENIC pipeline. We thus also evaluate GRNBoost + TF, which consists in restricting the putative parents to a list of known transcription factors. SCENIC additionally leverages known motif to reduce the number of false positives.

RC: I'm not familiar with most methods compared here, but most GRN methods that I've seen provide a ranking for the inferred edges, which then can be used for creating a precision-recall curve and calculating AUPR (which is probably a better measure compared to a single precision and recall value). Was this not possible for the methods used here? If a method does not report edge confidence, it can be run in a bootstrap or stability selection framework to calculate the confidence values. Considering the reported run times this should be feasible for many of the methods. I understand that doing this might not be in the scope of this work, but it would be worth mentioning in the discussions.

AR: Your point about uncertainty quantification is indeed crucial, as practitioners are highly interested in having a measure of uncertainty in the predicted interactions. As you already alluded to, we refrained from evaluation this aspect, as for most causal inference methods, there is no straightforward way of obtaining edge rankings

from those models. Most theoretically grounded methods which provides a distribution over possible graphs are based on the Bayesian framework. Unfortunately, Bayesian inference is notoriously hard to scale, making it unsuitable for the problem at hand [13, 14]. However, we would like to note that most of the challenge methods, such as Mean Difference and Guanlab, are based on ranking the edges by confidence. This is why we evaluate two versions for those methods, i.e. top 1k and top 5k interactions. We have revised the discussion accordingly to elaborate on that aspect and provide practitioners with actionable insights to obtain edge confidences.

Related revisions:

1. ll 289-298: Revised: In our study, we observed that some methods do not return edge confidence, especially from the causality literature, which is why we focused on evaluating point estimates as there is not always a theoretically sound way to obtain uncertainty estimates. This approach encourages methods to return the Pareto optimal trade-off, providing a balance between precision and recall. However, for practitioners interested in edge confidence, bootstrapping [15, 16] or stability selection [17] can be employed to calculate these values. It's important to note that the two best-performing methods in our benchmark, Mean Difference and Guanlab, are based on ranking the edges. This ranking inherently provides a measure of confidence in the predicted interactions, with higher-ranked edges being more confident. Therefore, even in the absence of explicit edge confidence values, these methods offer a form of confidence measure through their ranking system.

RC: - It might be easier to read if you merge table 1 and table 2 (and then add rank in K562 and RPE1 as additional columns). - When discussing related work, mentioning older works like the DREAM5 challenge (Marbach et al 2012 nature methods) could be helpful. - There are no citations for the methods in section 4.3.4. - While this is not related to the scope of this paper, mentioning methods that use scRNA-seq and scATAC-seq to infer regulatory networks (SCENIC+ for example) could be helpful for the readers.

AR: We thank you for the improvements suggestions and we have revised the manuscript accordingly. We merged table 1 and 2 (see table 1), discuss the DREAM5 challenge, mention methods that use scRNA-seq and scATAC-seq, and added citations in section 4.3.4.

Related revisions:

1. ll 443-451: Revised: The DREAM5 challenge [18] provided a landmark benchmark by evaluating GRN inference approaches on both observational microarray data and experimental perturbation data from *in vivo* gene knockout and knockdown studies, particularly for organisms like *E. coli* and *S. aureus*. While DREAM5 significantly advanced the field by incorporating experimental perturbations, it remained limited to bulk data and did not address the high-resolution insights enabled by large-scale, single-cell perturbational datasets. Our work, CausalBench, extends this approach by benchmarking GRN inference on expansive single-cell perturbational data, with biologically relevant metrics that offer deeper insights into causal gene regulatory mechanisms at the single-cell level.
2. ll 410-416: Revised: In recent years, several methods have also been developed to infer gene regulatory networks by integrating single-cell RNA sequencing and single-cell ATAC sequencing (scATAC-seq) data. Notably, SCENIC+ [19] is designed to construct enhancer-driven GRNs using combined or separate scRNA-seq and scATAC-seq data. Other methods have also been proposed, such as STREAM

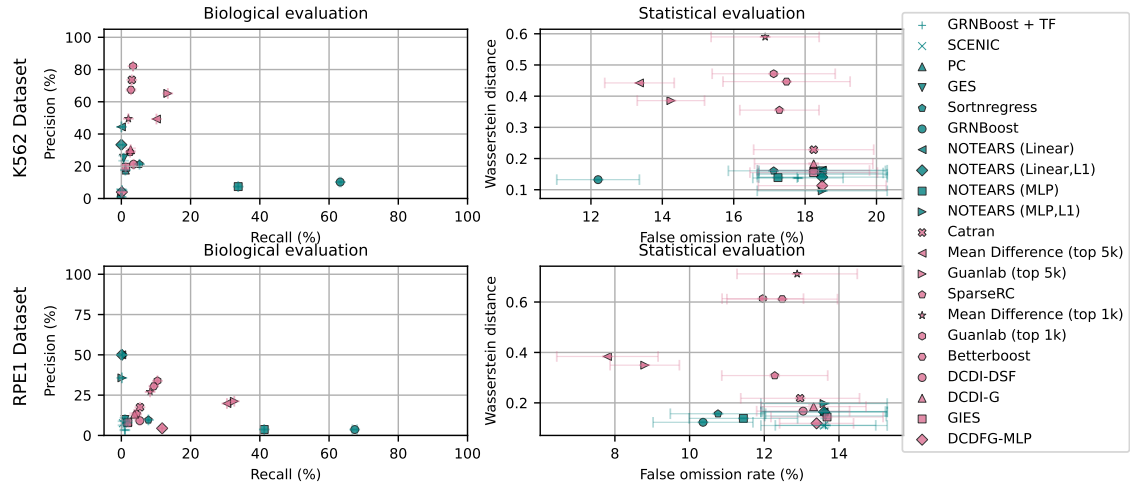


Figure 1: **Precision-Recall trade-off.** Performance comparison in terms of Precision (in %; y-axis) and Recall (in %; x-axis) in correctly identifying edges substantiated by biological interaction databases (left panels); and our own statistical evaluation using interventional information in terms of Wasserstein distance and FOR (right panels). Performance is compared across 10 different methods using observational data (green markers), and 11 different methods using interventional data (pink markers) in K562 (top panels) and RPE1 (bottom panels) cell lines. For each method, we show the mean and standard deviation from five independent runs.

[20], GLUE [21], DIRECT-NET [22], Pando [23] and scMEGA [24]. These approaches leverage the complementary information provided by gene expression and chromatin accessibility to enhance the accuracy of GRN inference.

References

- [1] Richard W Carthew. Gene regulation by micrnas. *Current opinion in genetics & development*, 16(2):203–208, 2006.
- [2] Michael Levine and Eric H Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942, 2005.
- [3] Sam Tracy, Guo-Cheng Yuan, and Ruben Dries. Rescue: imputing dropout events in single-cell rna-sequencing data. *BMC bioinformatics*, 20(1):1–11, 2019.
- [4] Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome biology*, 23(1):31, 2022.
- [5] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 11 2017.

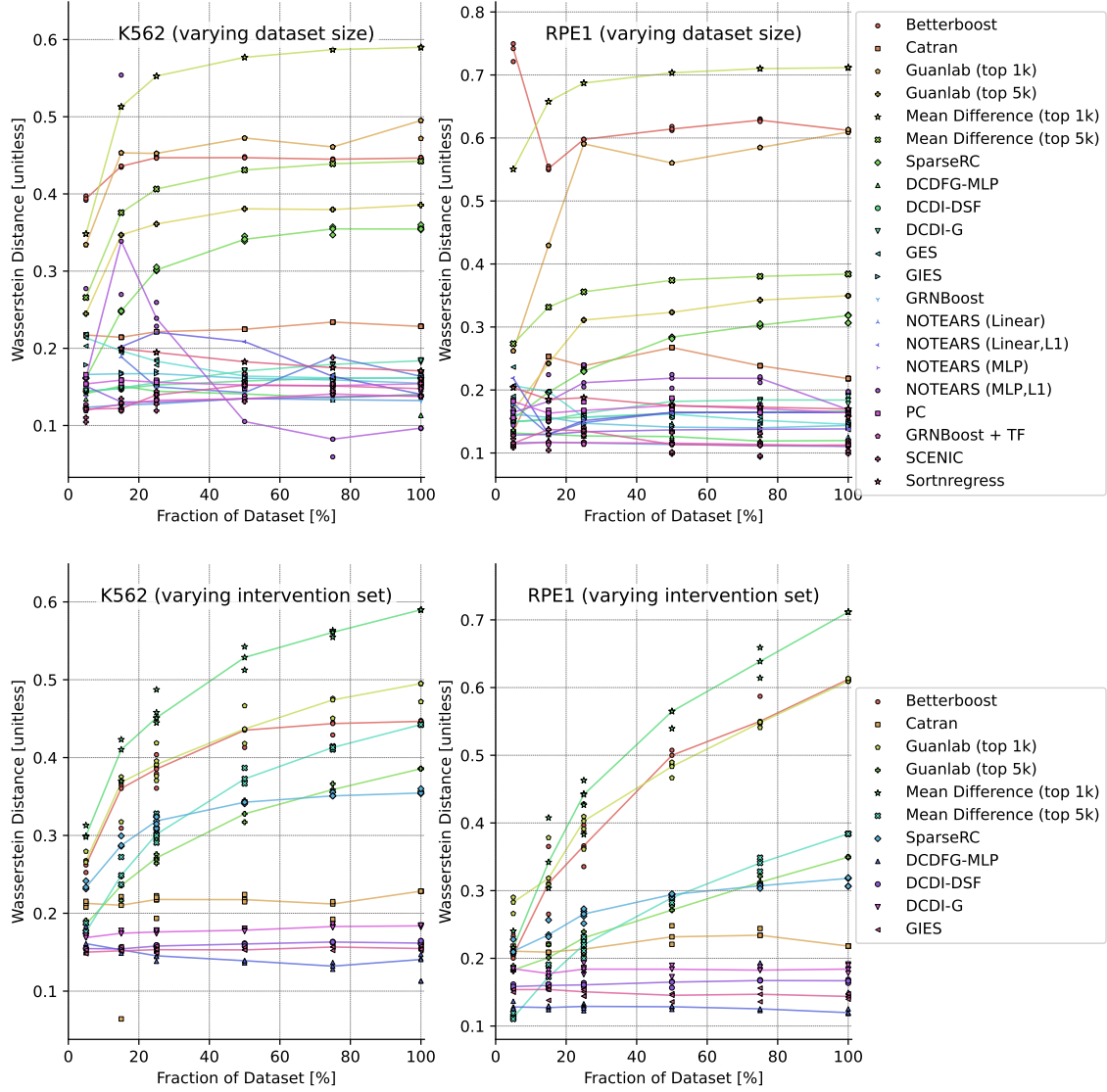


Figure 2: **Statistical evaluation of scaling characteristics.** Performance comparison in terms of Mean Wasserstein Distance (unitless; y-axis) of 10 methods for causal graph inference on observational data (top row; (see legend top right) and 11 methods on interventional scRNAseq data (centre row; see legend centre right) when varying the fraction of the full dataset size available for inference (in %; x-axis), and 4 methods on interventional data (bottom row; see legend bottom right) when varying the fraction of the full intervention set used (in %, x-axis). Markers indicate the values observed when running the respective algorithms with one of three random seeds, and colored lines indicate the median value observed across all tested random seeds for a method.

- [6] Monika S Kowalczyk, Itay Tirosh, Dirk Heckl, Tata Nageswara Rao, Atray Dixit, Brian J Haas, Rebekka K Schneider, Amy J Wagers, Benjamin L Ebert, and Aviv Regev. Single-cell rna-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research*, 25(12):1860–1872, 2015.
- [7] Dhawal Jain, Sandro Baldi, Angelika Zabel, Tobias Straub, and Peter B Becker. Active promoters give rise to false positive ‘phantom peaks’ in chip-seq experiments. *Nucleic acids research*, 43(14):6959–6968, 2015.
- [8] Joseph K Pickrell, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. False positive peaks in chip-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*, 27(15):2144–2146, 2011.
- [9] Zhaonan Zou, Tazro Ohta, Fumihito Miura, and Shinya Oki. Chip-atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating chip-seq, atac-seq and bisulfite-seq data. *Nucleic Acids Research*, page 1, 2022.
- [10] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2018.
- [11] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.
- [12] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.
- [13] Elaine Angelino, Matthew James Johnson, Ryan P Adams, et al. Patterns of scalable bayesian inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016.
- [14] Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24111–24123. Curran Associates, Inc., 2021.
- [15] David A Freedman. Bootstrapping regression models. *The annals of statistics*, 9(6):1218–1228, 1981.
- [16] Christopher Z Mooney, Robert D Duval, and Robert Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Number 95. sage, 1993.
- [17] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.
- [18] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.
- [19] Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, and Stein Aerts. Scenic+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature methods*, 20(9):1355–1367, 2023.

- [20] Yang Li, Anjun Ma, Yizhong Wang, Qi Guo, Cankun Wang, Hongjun Fu, Bingqiang Liu, and Qin Ma. Enhancer-driven gene regulatory networks inference from single-cell rna-seq and atac-seq data. *Briefings in Bioinformatics*, 25(5):bbae369, 2024.
- [21] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.
- [22] Lihua Zhang, Jing Zhang, and Qing Nie. Direct-net: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Science advances*, 8(22):eabl7393, 2022.
- [23] Jonas Simon Fleck, Sophie Martina Johanna Jansen, Damian Wollny, Fides Zenk, Makiko Seimiya, Akanksha Jain, Ryoko Okamoto, Malgorzata Santel, Zhisong He, J Gray Camp, et al. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, 621(7978):365–372, 2023.
- [24] Zhijian Li, James S Nagai, Christoph Kuppe, Rafael Kramann, and Ivan G Costa. scmega: single-cell multi-omic enhancer-based gene regulatory network inference. *Bioinformatics Advances*, 3(1):vbad003, 2023.

Table 1: Ranking of Models Based on Mean F1 Score on the biological evaluation for K562 and RPE1. The mean is over five seeds with standard deviation.

Model	K562			RPE1		
	Mean F1	Std	Rank	Mean F1	Std	Rank
Guanlab (top 5k)	0.223	0.000	1	0.257	0.000	1
Mean Difference (top 5k)	0.168	0.000	3	0.240	0.000	2
Guanlab (top 1k)	0.065	0.000	6	0.160	0.000	3
GRNBoost	0.175	0.001	2	0.069	0.000	8
Sortnregress	0.084	0.000	5	0.086	0.000	6
NOTEARS (MLP)	0.121	0.000	4	0.068	0.000	9
Betterboost	0.053	0.000	9	0.144	0.001	4
Catran	0.058	0.000	8	0.082	0.000	7
DCDI-DSF	0.060	0.001	7	0.067	0.002	10
Mean Difference (top 1k)	0.039	0.000	12	0.127	0.000	5
SparseRC	0.046	0.000	11	0.067	0.000	11
DCDI-G	0.049	0.001	10	0.059	0.003	13
GIES	0.023	0.002	13	0.030	0.005	15
PC	0.022	0.000	14	0.031	0.002	14
DCDFG-MLP	0.003	0.000	17	0.064	0.000	12
GES	0.013	0.001	15	0.019	0.002	16
GRNBoost + TF	0.004	0.000	16	0.016	0.001	17
SCENIC	0.000	0.000	20	0.009	0.002	18
NOTEARS (Linear)	0.000	0.000	18	0.002	0.000	20
NOTEARS (Linear,L1)	0.000	0.000	19	0.002	0.000	20
NOTEARS (MLP,L1)	0.000	0.000	21	0.003	0.000	19

Table 2: Model ranking. We here present a simple unbiased way of ranking the different models based on the mean Wasserstein distance and the FOR. We separate the rankings per cell type (Top table: K562, bottom table: RPE1). First, we create a preliminary ranking for each evaluation metrics. We rank by the mean score across the five different seeds. Finally, for each model, we take their average rank across the evaluation specific rankings. This ranking thus gives the same weight to each evaluation method. We also indicate the confidence interval for both metrics. The fraction of interventions indicates what ratio of the gene intervention are used for inference. For purely observational methods, this ratio is 0. For interventional methods, we run them on both high (100%) and low (25%) intervention settings. [Revised: We recapitulate the methods F1 score rank, for the observational methods and at 100% of interventions for the interventional methods.](#)

Model	Fraction Interventions	Mean Rank	Rank Wasserstein	Rank FOR	Wasserstein Distance	FOR	Rank F1 Score
Mean Difference (top 1k)	1.0	2.5	1	4	0.590 ± 0.000	0.169 ± 0.013	12.0
Mean Difference (top 5k)	1.0	3.5	5	2	0.442 ± 0.000	0.134 ± 0.009	3.0
Guanlab (top 5k)	1.0	4.5	6	3	0.386 ± 0.000	0.142 ± 0.008	1.0
Guanlab (top 1k)	1.0	4.5	2	7	0.472 ± 0.000	0.171 ± 0.015	6.0
Betterboost	1.0	8.0	4	12	0.447 ± 0.001	0.175 ± 0.016	9.0
Mean Difference (top 1k)	0.25	9.0	3	15	0.458 ± 0.015	0.179 ± 0.015	–
Guanlab (top 5k)	0.25	9.5	12	7	0.275 ± 0.011	0.171 ± 0.012	–
SparseRC	1.0	9.5	9	10	0.355 ± 0.001	0.173 ± 0.010	11.0
Mean Difference (top 5k)	0.25	10.5	11	10	0.306 ± 0.012	0.173 ± 0.012	–
SparseRC	0.25	10.5	10	11	0.321 ± 0.003	0.173 ± 0.012	–
Betterboost	0.25	12.5	8	17	0.383 ± 0.015	0.180 ± 0.014	–
Guanlab (top 1k)	0.25	12.5	7	18	0.383 ± 0.013	0.180 ± 0.015	–
Sortnregress	0.0	13.0	19	7	0.161 ± 0.000	0.171 ± 0.011	5.0
GRNBoost	0.0	15.5	30	1	0.132 ± 0.000	0.122 ± 0.010	2.0
DCDI-G	0.25	15.5	16	15	0.177 ± 0.001	0.179 ± 0.015	–
NOTEARS (MLP)	0.0	18.0	28	8	0.139 ± 0.000	0.172 ± 0.005	4.0
Catran	1.0	18.5	13	24	0.228 ± 0.000	0.182 ± 0.015	8.0
Catran	0.25	19.0	14	24	0.214 ± 0.010	0.182 ± 0.016	–
DCDI-DSF	0.25	19.0	21	17	0.158 ± 0.002	0.180 ± 0.016	–
DCDI-G	1.0	19.5	15	24	0.183 ± 0.002	0.182 ± 0.015	10.0
GRNBoost + TF	0.0	21.0	29	13	0.138 ± 0.001	0.178 ± 0.012	16.0
DCDI-DSF	1.0	21.0	18	24	0.162 ± 0.001	0.182 ± 0.014	7.0
DCDFG-MLP	0.25	22.5	26	19	0.149 ± 0.007	0.181 ± 0.016	–
GIES	1.0	23.0	22	24	0.155 ± 0.002	0.182 ± 0.016	13.0
GIES	0.25	24.5	24	25	0.153 ± 0.003	0.183 ± 0.014	–
PC	0.0	24.5	23	26	0.154 ± 0.002	0.184 ± 0.015	14.0
NOTEARS (Linear)	0.0	24.5	17	32	0.163 ± 0.000	0.185 ± 0.016	18.0
GES	0.0	26.0	20	32	0.160 ± 0.006	0.185 ± 0.016	15.0
SCENIC	0.0	28.5	25	32	0.150 ± 0.012	0.185 ± 0.016	20.0
NOTEARS (Linear,L1)	0.0	29.5	27	32	0.140 ± 0.000	0.185 ± 0.016	19.0
DCDFG-MLP	0.0	31.5	31	32	0.113 ± 0.000	0.185 ± 0.016	17.0
NOTEARS (MLP,L1)	0.0	32.0	32	32	0.097 ± 0.000	0.185 ± 0.016	21.0

Model	Fraction Interventions	Mean Rank	Rank Wasserstein	Rank FOR	Wasserstein Distance	FOR	Rank F1 Score
Mean Difference (top 5k)	1.0	3.5	6	1	0.384 ± 0.000	0.078 ± 0.012	2.0
Guanlab (top 1k)	1.0	4.5	2	7	0.613 ± 0.000	0.120 ± 0.010	3.0
Guanlab (top 5k)	1.0	5.0	8	2	0.350 ± 0.000	0.088 ± 0.008	1.0
Betterboost	1.0	6.5	3	10	0.611 ± 0.001	0.125 ± 0.013	4.0
Guanlab (top 1k)	0.25	7.0	5	9	0.399 ± 0.021	0.123 ± 0.007	–
Mean Difference (top 1k)	1.0	8.0	1	15	0.712 ± 0.000	0.129 ± 0.014	5.0
SparseRC	0.25	8.0	10	6	0.265 ± 0.007	0.118 ± 0.009	–
SparseRC	1.0	9.0	9	9	0.308 ± 0.004	0.123 ± 0.012	11.0
Betterboost	0.25	9.5	7	12	0.361 ± 0.019	0.127 ± 0.011	–
Mean Difference (top 1k)	0.25	10.0	4	16	0.431 ± 0.026	0.129 ± 0.012	–
Mean Difference (top 5k)	0.25	12.0	13	11	0.218 ± 0.011	0.126 ± 0.008	–
Guanlab (top 5k)	0.25	12.0	11	13	0.228 ± 0.009	0.128 ± 0.009	–
Sortnregress	0.0	13.5	23	4	0.157 ± 0.000	0.108 ± 0.011	6.0
Catran	1.0	14.5	12	17	0.218 ± 0.000	0.130 ± 0.014	7.0
GRNBoost	0.0	16.0	29	3	0.123 ± 0.000	0.104 ± 0.012	8.0
NOTEARS (MLP)	0.0	16.5	28	5	0.138 ± 0.000	0.114 ± 0.013	9.0
Catran	0.25	17.0	14	20	0.215 ± 0.007	0.132 ± 0.013	–
DCDI-DSF	0.25	18.0	22	14	0.161 ± 0.002	0.128 ± 0.009	–
DCDI-G	0.25	18.0	17	19	0.182 ± 0.006	0.130 ± 0.009	–
DCDI-DSF	1.0	18.5	18	19	0.167 ± 0.003	0.130 ± 0.011	10.0
DCDI-G	1.0	19.0	16	22	0.185 ± 0.006	0.133 ± 0.012	13.0
NOTEARS (MLP,L1)	0.0	22.0	15	29	0.197 ± 0.000	0.136 ± 0.015	19.0
PC	0.0	23.0	19	27	0.166 ± 0.003	0.136 ± 0.014	14.0
DCDFG-MLP	0.25	23.5	26	21	0.145 ± 0.022	0.132 ± 0.010	–
NOTEARS (Linear)	0.0	24.0	21	27	0.164 ± 0.000	0.136 ± 0.015	20.5
GIES	0.25	24.0	24	24	0.151 ± 0.006	0.135 ± 0.015	–
NOTEARS (Linear,L1)	0.0	25.0	21	29	0.164 ± 0.000	0.136 ± 0.015	20.5
DCDFG-MLP	0.0	26.5	30	23	0.119 ± 0.000	0.134 ± 0.009	12.0
GES	0.0	28.0	25	31	0.148 ± 0.008	0.136 ± 0.014	16.0
SCENIC	0.0	29.5	32	27	0.110 ± 0.008	0.136 ± 0.015	18.0
GIES	1.0	29.5	27	32	0.144 ± 0.003	0.137 ± 0.013	15.0
GRNBoost + TF	0.0	31.0	31	31	0.111 ± 0.001	0.136 ± 0.012	17.0

Table 3: **Model characteristics.** Table summarizing the characteristics of the test baselines. Graph scaling refers to the ability to infer on the complete set of genes. A method is deemed to have successfully achieved sample scaling and intervention scaling if its performance improves by at least 10% on both datasets. This comparison is made between settings where only a subset, specifically 25% of datapoints or perturbations, is used and settings where 100% of them are utilized. As such, only SparseRC fulfills all the criteria.

Model	Graph scaling	Sample scaling	Intervention scaling
PC	×	×	–
GES	×	×	–
GIES	×	×	×
NOTEARS	✓	×	–
DCDI-G	×	✓	×
DCDI-DSF	×	×	×
DCDI-FG	✓	×	×
GRNBoost	✓	×	–
GRNBoost + TF	✓	×	–
SCENIC	✓	×	–
Sortnregress	✓	×	–
Betterboost	✓	×	✓
Catran	✓	×	×
Guanlab	✓	×	✓
Mean Difference	✓	×	✓
SparseRC	✓	✓	✓

Table 4: Precision of each database in terms of percentage of links that are validated by differential expression in the datasets. CORUM, being the most curated database, has the highest precision. For all sources, the precision is much lower on RPE1 than on K562, confirming the difficulty of constructing a ground-truth for less studied cell lines. PPI (N) and PPI(P) surprisingly have similar precisions, even though PPI (N) contains much more interactions that are supposedly less certain.

Data source	Precision RPE1	Precision K562
ChiPSeq	12.92%	16.42%
PPI (N)	17.17%	25.43%
PPI (P)	17.49%	30.08%
CORUM	19.78%	39.18%

Table 5: **Biological evaluation results on RPE1.** Performance results on RPE1 Dataset for the biological evaluation, for each evidence type. Pooled correspond to pooling all the extracted databases into one network for evaluation.

Model	Precision					Recall				
	Pooled	CORUM	ChipSeq	PPI (N)	PPI (P)	Pooled	CORUM	ChipSeq	PPI (N)	PPI (P)
Betterboost	0.31±0.00	0.02±0.00	0.02±0.00	0.30±0.00	0.08±0.00	0.09±0.00	0.04±0.00	0.01±0.00	0.04±0.00	0.04±0.00
Catran	0.17±0.00	0.01±0.00	0.01±0.00	0.38±0.00	0.11±0.00	0.05±0.00	0.03±0.00	0.00±0.00	0.05±0.00	0.05±0.00
Guanlab (top 1k)	0.34±0.00	0.02±0.00	0.02±0.00	0.33±0.00	0.08±0.00	0.10±0.00	0.05±0.00	0.01±0.00	0.04±0.00	0.04±0.00
Guanlab (top 5k)	0.21±0.00	0.01±0.00	0.02±0.00	0.24±0.00	0.06±0.00	0.33±0.00	0.18±0.00	0.04±0.00	0.14±0.00	0.14±0.00
Mean Difference (top 1k)	0.27±0.00	0.02±0.00	0.01±0.00	0.27±0.00	0.07±0.00	0.08±0.00	0.04±0.00	0.01±0.00	0.03±0.00	0.03±0.00
Mean Difference (top 5k)	0.20±0.00	0.02±0.00	0.01±0.00	0.22±0.00	0.05±0.00	0.31±0.00	0.20±0.00	0.04±0.00	0.13±0.00	0.13±0.00
SparseRC	0.13±0.00	0.02±0.00	0.01±0.00	0.27±0.00	0.08±0.00	0.04±0.00	0.04±0.00	0.01±0.00	0.03±0.00	0.04±0.00
DCDFG-MLP	0.04±0.00	0.01±0.00	0.03±0.00	0.18±0.00	0.04±0.00	0.12±0.00	0.20±0.00	0.15±0.00	0.19±0.00	0.17±0.00
DCDI-DSF	0.09±0.00	0.01±0.00	0.01±0.00	0.20±0.01	0.06±0.00	0.05±0.00	0.05±0.01	0.01±0.00	0.05±0.00	0.05±0.00
DCDI-G	0.13±0.01	0.02±0.00	0.01±0.00	0.26±0.00	0.08±0.00	0.04±0.00	0.04±0.01	0.00±0.00	0.03±0.00	0.04±0.00
GIES	0.08±0.01	0.02±0.00	0.01±0.00	0.29±0.01	0.08±0.01	0.02±0.00	0.04±0.01	0.01±0.00	0.03±0.00	0.03±0.00
GES	0.10±0.01	0.02±0.01	0.01±0.01	0.36±0.03	0.11±0.02	0.01±0.00	0.02±0.01	0.00±0.00	0.01±0.00	0.02±0.00
GRNBoost	0.04±0.00	0.01±0.00	0.01±0.00	0.09±0.00	0.02±0.00	0.67±0.01	0.84±0.01	0.43±0.01	0.65±0.00	0.66±0.01
GRNBoost + TF	0.03±0.00	0.00±0.00	0.10±0.01	0.19±0.00	0.04±0.00	0.01±0.00	0.01±0.00	0.06±0.00	0.02±0.00	0.02±0.00
SCENIC	0.08±0.01	0.00±0.00	0.28±0.03	0.30±0.02	0.04±0.01	0.00±0.00	0.00±0.00	0.03±0.00	0.01±0.00	0.00±0.00
NOTEARS (Linear)	0.50±0.00	0.12±0.00	0.00±0.00	0.88±0.00	0.38±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (Linear,L1)	0.50±0.00	0.12±0.00	0.00±0.00	0.88±0.00	0.38±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (MLP)	0.04±0.00	0.01±0.00	0.01±0.00	0.14±0.00	0.03±0.00	0.41±0.00	0.80±0.00	0.24±0.00	0.61±0.00	0.62±0.00
NOTEARS (MLPL1)	0.36±0.00	0.14±0.00	0.00±0.00	1.00±0.00	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
PC	0.09±0.01	0.02±0.01	0.01±0.00	0.28±0.01	0.08±0.01	0.02±0.00	0.04±0.01	0.01±0.00	0.02±0.00	0.03±0.00
Sortnregress	0.10±0.00	0.02±0.00	0.02±0.00	0.29±0.00	0.08±0.00	0.08±0.00	0.12±0.00	0.02±0.00	0.09±0.00	0.10±0.00

Table 6: **Biological evaluation results on K562.** Performance results on K562 Dataset for the biological evaluation, for each evidence type. Pooled correspond to pooling all the extracted databases into one network for evaluation.

Model	Precision					Recall				
	Pooled	CORUM	ChipSeq	PPI (N)	PPI (P)	Pooled	CORUM	ChipSeq	PPI (N)	PPI (P)
Betterboost	0.67±0.01	0.27±0.00	0.00±0.00	0.62±0.01	0.43±0.00	0.03±0.00	0.06±0.00	0.00±0.00	0.01±0.00	0.03±0.00
Catran	0.74±0.00	0.88±0.00	0.00±0.00	0.99±0.00	0.91±0.00	0.03±0.00	0.21±0.00	0.00±0.00	0.02±0.00	0.07±0.00
Guanlab (top 1k)	0.82±0.00	0.39±0.00	0.00±0.00	0.75±0.00	0.52±0.00	0.03±0.00	0.09±0.00	0.00±0.00	0.02±0.00	0.04±0.00
Guanlab (top 5k)	0.65±0.00	0.18±0.00	0.00±0.00	0.62±0.00	0.32±0.00	0.13±0.00	0.21±0.00	0.01±0.00	0.07±0.00	0.12±0.00
Mean Difference (top 1k)	0.49±0.00	0.13±0.00	0.01±0.00	0.49±0.00	0.22±0.00	0.02±0.00	0.03±0.00	0.00±0.00	0.01±0.00	0.02±0.00
Mean Difference (top 5k)	0.49±0.00	0.12±0.00	0.01±0.00	0.48±0.00	0.23±0.00	0.10±0.00	0.14±0.00	0.01±0.00	0.05±0.00	0.08±0.00
SparseRC	0.29±0.00	0.19±0.00	0.00±0.00	0.45±0.00	0.26±0.00	0.03±0.00	0.09±0.00	0.00±0.00	0.02±0.00	0.04±0.00
DCDFG-MLP	0.04±0.00	0.00±0.00	0.00±0.00	0.39±0.00	0.07±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.00	0.00±0.00
DCDI-DSF	0.21±0.00	0.07±0.00	0.00±0.00	0.39±0.00	0.14±0.00	0.04±0.00	0.06±0.00	0.00±0.00	0.03±0.00	0.04±0.00
DCDI-G	0.30±0.01	0.12±0.00	0.01±0.00	0.44±0.01	0.18±0.00	0.03±0.00	0.06±0.00	0.00±0.00	0.02±0.00	0.03±0.00
GIES	0.19±0.01	0.11±0.00	0.01±0.00	0.52±0.01	0.19±0.01	0.01±0.00	0.04±0.00	0.00±0.00	0.02±0.00	0.02±0.00
GES	0.25±0.02	0.18±0.01	0.00±0.00	0.59±0.02	0.26±0.01	0.01±0.00	0.03±0.00	0.00±0.00	0.01±0.00	0.01±0.00
GRNBoost	0.10±0.00	0.02±0.00	0.01±0.00	0.21±0.00	0.07±0.00	0.63±0.00	0.81±0.01	0.32±0.01	0.70±0.00	0.71±0.00
GRNBoost + TF	0.05±0.00	0.01±0.00	0.05±0.01	0.15±0.01	0.02±0.00	0.00±0.00	0.00±0.00	0.02±0.00	0.00±0.00	0.00±0.00
SCENIC	0.03±0.02	0.01±0.02	0.00±0.00	0.36±0.04	0.02±0.02	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (Linear)	0.44±0.00	0.44±0.00	0.00±0.00	0.89±0.00	0.56±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (Linear,L1)	0.33±0.00	0.50±0.00	0.00±0.00	0.83±0.00	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOTEARS (MLP)	0.07±0.00	0.02±0.00	0.01±0.00	0.28±0.00	0.08±0.00	0.34±0.00	0.55±0.00	0.30±0.00	0.66±0.00	0.62±0.00
NOTEARS (MLP,L1)	0.00±0.00	0.00±0.00	0.00±0.00	0.80±0.00	0.20±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
PC	0.18±0.00	0.10±0.00	0.01±0.00	0.46±0.02	0.16±0.01	0.01±0.00	0.04±0.00	0.00±0.00	0.02±0.00	0.02±0.00
Sortnregress	0.21±0.00	0.15±0.00	0.00±0.00	0.54±0.00	0.24±0.00	0.05±0.00	0.22±0.00	0.01±0.00	0.07±0.00	0.10±0.00

Table 7: **Partition sizes.** Partition sizes used for each model. -1 means that the graph was not partitioned.

Model name	partition size
PC	30
GES	30
GIES	30
NOTEARS (Linear)	-1
NOTEARS (Linear, L1)	-1
NOTEARS (MLP)	-1
NOTEARS (MLP, L1)	-1
DCDI-DSF	50
DCDI-G	50
DCDFG-MLP	-1
GRNBoost (+ TF)	-1
SCENIC	-1
Sortnregress	-1
Mean Difference	-1
Guanlab	-1
SparseRC	-1
Betterboost	-1
Catran	-1

Table 8: **Run time K562.** Run time in wall clock hours for each method on the K562 cell line.

Model	Run time (hours)
SparseRC	0.004
GRNBoost + TF	0.012
SCENIC	0.046
Sortnregress	0.087
Mean Difference (top 5k)	0.094
Mean Difference (top 1k)	0.094
GRNBoost	0.097
Catran	0.197
GIES	0.824
DCDFG-MLP	1.087
DCDI-DSF	1.575
PC	1.672
NOTEARS (MLP,L1)	1.705
Guanlab (top 5k)	2.921
Guanlab (top 1k)	3.113
GES	3.148
NOTEARS (Linear,L1)	4.974
DCDI-G	5.043
NOTEARS (Linear)	6.297
NOTEARS (MLP)	10.778
Betterboost	11.934

Table 9: **Run time RPE1.** Run time in wall clock hours for each method on the RPE1 cell line.

Model	Run time (hours)
GRNBoost + TF	0.011
SparseRC	0.018
GRNBoost	0.031
Sortnregress	0.035
Mean Difference (top 1k)	0.040
Mean Difference (top 5k)	0.041
SCENIC	0.050
Catran	0.166
DCDI-DSF	0.493
DCDFG-MLP	0.550
Guanlab (top 1k)	0.933
Guanlab (top 5k)	0.955
NOTEARS (MLP,L1)	1.421
GIES	1.450
NOTEARS (Linear,L1)	1.557
NOTEARS (Linear)	1.680
DCDI-G	1.955
GES	2.207
Betterboost	3.736
PC	4.686
NOTEARS (MLP)	12.217