Exact Statistical Tests for Gene Regulatory Network Discovery from Single-Cell RNA Sequencing

Anonymous Author(s)
Affiliation
Address
email

Abstract

Gene regulatory networks encode causal relationships between transcription factors and target genes, but inferring these networks from single-cell RNA sequencing data faces extreme sparsity and class imbalance challenges. We present a framework using exact statistical tests to evaluate whether predicted regulatory edges are enriched above background rates in the top-ranked predictions where experimental validation would focus. This approach moves beyond global metrics to assess performance where it matters for practical discovery. Using our scoring method, we demonstrate strong performance across two evaluations. Curated positive edges receive mean posterior probability 0.908 versus 0.0054 for random negatives. Across 44 BEELINE benchmark datasets, we achieve mean ROC-AUC 0.926 and mean precision 39.9% in the top 100 predictions (47-fold improvement over random selection). Enrichment tests confirm statistical significance on all 44 datasets. These results show that exact statistical tests provide actionable evidence for network discovery, offering practical guidance for experimental validation while maintaining statistical rigor for structure learning from noisy single-cell data.

1 Introduction

Gene regulatory networks (GRNs) represent the causal mechanisms through which transcription factors control gene expression, determining cellular identity, development, and disease progression [1, 3, 9, 10]. These networks consist of directed graphs where edges indicate regulatory relationships from transcription factors to target genes [5]. Understanding these regulatory circuits is essential for addressing fundamental questions in biology, from cell fate determination to therapeutic intervention design.

Current evaluation approaches for GRN prediction rely heavily on global metrics such as area under the ROC curve (AUROC) and area under the precision-recall curve (AUPRC). While these metrics enable method comparison, they fail to address the practical question: given limited validation resources, are the highest-ranked predictions enriched for true regulatory relationships? This question becomes critical under extreme class imbalance, where high AUROC values may still correspond to poor precision in the top predictions that researchers would actually test.

We study exact statistical tests that quantify enrichment in top-ranked predictions. Our approach treats GRN inference as causal structure learning where prior biological knowledge guides the search through possible edges. We employ Fisher's exact test to determine whether predicted edges concentrate significantly above background rates in regions where experimental validation would focus. We demonstrate this framework using a Bayesian scoring method for edge ranking, though the evaluation approach applies to any ranking method.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Related Work

Early GRN inference methods leveraged bulk RNA sequencing, employing techniques from correlation analysis to probabilistic models, with approaches like WGCNA identifying co-expression modules but unable to distinguish direct from indirect relationships [8], while information-theoretic methods such as ARACNE used mutual information to detect non-linear dependencies [4], and regression frameworks including GENIE3 employed random forests to rank interactions [7]. The transition to single-cell data required new algorithms to handle sparsity and noise, leading to methods like Inferelator 3.0 which combines Bayesian regression with stability selection and scales to millions of cells despite limitations in capturing non-linear relationships [14], while deep learning approaches including 3DCEMA's three-dimensional convolution and graph neural networks like GENELink and GNNLink emerged to model complex dependencies though with varying computational costs and data requirements [2, 6, 11]. Comprehensive benchmarking studies have revealed that performance degrades substantially with increasing sparsity and gene count, with many methods approaching random performance under realistic noise conditions, motivating our focus on evaluation metrics that capture performance where experimental validation occurs rather than averaged global statistics [12, 13].

3 Problem and Methods

Let $\mathcal{V} = \{v_1, v_2, \ldots, v_G\}$ denote the set of genes. The true gene regulatory network $H \subseteq \mathcal{V} \times \mathcal{V}$ is a directed graph where edge $(v_i, v_j) \in H$ indicates that transcription factor v_i regulates target gene v_j. Let $G \subset H$ denote experimentally validated edges available as prior knowledge. These come from ChIP-seq experiments, genetic perturbations, and literature curation.

Given single-cell expression data $X \in \mathbb{R}^{N \times G}$ with N cells, we compute for each candidate edge $e = (v_i, v_j)$ a score $P(e \in H \mid X, G)$ representing the probability that the edge is a true regulatory relationship. These scores induce a ranking over candidate edges. The evaluation challenge is assessing this ranking when ground truth exists only for a sparse subset with typical prevalence below 1%.

We test whether our approach can identify true regulatory relationships from single-cell data through four hypotheses. The first hypothesis examines whether true regulatory edges concentrate in the highest-ranked predictions beyond what random chance would produce, which we evaluate using Fisher's exact test comparing the top-ranked set against the remainder. The second hypothesis tests whether the scoring method assigns meaningfully different probabilities to true regulatory edges versus non-edges, enabling practical threshold selection. The third hypothesis assesses whether the approach generalizes across different cell types, organisms, and experimental conditions rather than overfitting to specific datasets. The fourth hypothesis evaluates whether the precision in top predictions is sufficient to guide laboratory validation efforts, providing substantial improvement over random selection.

scRNA-seq Data Representation. We construct gene representations through dimensionality reduction followed by supervised refinement. Given expression matrix $X \in \mathbb{R}^{N \times G}$, we compute the gene covariance matrix and extract its top r principal components. For gene i with expression vector $x_i \in \mathbb{R}^N$:

$$g_i = W^\top (x_i - \bar{x})$$

where $W \in \mathbb{R}^{N \times r}$ contains the top eigenvectors and $\bar{x}$ is the mean expression. We set $r = 64$ based on variance explained analysis.

Directional Embedding for Regulatory Relationships. Regulatory relationships have inherent directionality - transcription factors regulate targets, not vice versa. We model this through separate transformations for regulator and target roles. With learned matrices $A, B \in \mathbb{R}^{d \times r}$:

$$z_i^S = A \cdot g_i \quad \text{(regulator)}$$

$$z_j^T = B \cdot g_j \quad \text{(target)}$$

Similarity between potential regulator i and target j uses cosine distance:

$$d(z_i^S, z_j^T) = 1 - \frac{(z_i^S)^\top z_j^T}{\|z_i^S\|_2 \|z_j^T\|_2}$$

Contrastive Loss. We refine embeddings using the Soft Nearest Neighbor loss. Given positive edges P from curated network G and sampled negative edges N:

$$\mathcal{L}_{\text{SNN}} = -\log \frac{\sum_{(i,j)\in P} \exp[-d(z_i^S, z_j^T)/T]}{\sum_{(u,v)\in P\cup N} \exp[-d(z_u^S, z_v^T)/T]}$$

where temperature T controls focus on hard examples.

Bayesian Edge Scoring. For candidate edge $e = (i, j)$, we compute a posterior combining distance-based likelihood with prior knowledge:

$$L(e) = \exp[-\alpha \cdot d(z_i^S, z_j^T)]$$

$$P(e \in H \mid X, G) = \frac{L(e) \cdot \pi(e)}{\sum_{e' \in \mathcal{U}} L(e') \cdot \pi(e')}$$

where $\pi(e) = \bar{\pi}$ is the observed positive rate in G.

Scoring with Nonparametric Models. We also implement a Gaussian Process classifier for comparison. For each edge, we concatenate features $x_{ij} = [z_i^S; z_j^T; \delta]$ where δ indicates direction. We use a radial basis kernel and optimize the variational evidence lower bound for scalability.

4 Experiments

BEELINE benchmark. We first evaluate on the complete BEELINE benchmark comprising 44 datasets from diverse biological systems. Each dataset pairs with one of four reference network types: STRING interactions, non-specific ChIP-seq, cell-type-specific ChIP-seq, or genetic perturbations. These references vary in quality and completeness, providing a robust generalization test.

Table 1 summarizes performance by reference type. We achieve mean ROC-AUC 0.926 across all datasets, demonstrating strong global ranking. Perturbation-based networks provide clearest signal (ROC-AUC 0.993), while cell-type-specific references prove most challenging (0.853), likely due to condition-specific regulation not captured in expression data.

Table 1: Performance across reference network types. Values show mean ± standard deviation. Enrichment indicates fraction of datasets with Fisher's exact test $p < 0.001$ in top 100 predictions.

Reference Type	Datasets	ROC-AUC	PR-AUC	Precision@100	Enrichment
STRING	14	0.956 ± 0.020	0.207 ± 0.118	0.267 ± 0.124	14/14
Non-Specific ChIP	14	0.950 ± 0.018	0.161 ± 0.057	0.272 ± 0.098	14/14
Cell-Type-Specific	14	0.853 ± 0.220	0.508 ± 0.208	0.439 ± 0.187	14/14
Perturbation	2	0.993 ± 0.001	0.445 ± 0.022	0.635 ± 0.106	2/2
Overall	44	0.926 ± 0.124	0.289 ± 0.200	0.335 ± 0.178	44/44

Performance in Top Predictions. For practical application, performance in the top predictions matters most since researchers can only validate a limited number of edges. Table 2 shows results at three thresholds corresponding to typical validation budgets.

Table 2: Performance at different numbers of top predictions across 44 datasets. Lift measures fold-improvement over random selection. Hit rate shows fraction of datasets with at least one true positive. Fisher counts are two-sided tests of enrichment in top-k vs rest (per-dataset, $p < 0.001$).

Metric	Top 100	Top 500	Top 1000
Mean Precision	0.399	0.298	0.240
Mean Recall	0.170	0.440	0.589
Mean Lift	51.876×	37.224×	29.262×
Hit Rate	1.00	1.00	1.00
Fisher $p < 0.001$	44/44	44/44	44/44

In the top 100 predictions, mean precision reaches 39.9% - a 47-fold improvement over the 0.7% background rate. All datasets contain at least one true positive in their top 100, enabling discovery even with limited resources. Fisher's exact test confirms significant enrichment on all 44 datasets.

Nonparametric Model Validation. To verify that enrichment is not model-specific, we tested a Gaussian Process classifier on a challenging subset with 74,539 edges and 0.61% prevalence. The GP achieves ROC-AUC 0.796 and identifies 7 true positives in the top 100 (precision 7.0%, 11.5-fold lift). This confirms enrichment persists across different architectures.

Crohn Disease Dataset. We next perform detailed analysis on a Crohn disease dataset with 8,076 cells and 27,289 genes. This dataset presents unique challenges: extreme sparsity (>90% zeros), disease-altered regulation, and minimal validation data (27 total curated edges). We use 25 for training, one for validation, and hold out one for testing.

The held-out regulatory edge achieves rank one across all possible gene pairs. The 25 training edges receive mean posterior 0.908 (standard deviation 0.044), while random non-edges show mean 0.0054 (standard deviation 0.0003) - a 168-fold difference demonstrating clear discrimination.

For enrichment analysis, we find 24 of 26 available curated edges in the top 1000 predictions.

5 Discussion

Our results demonstrate that meaningful regulatory structure can be recovered from single-cell expression despite extreme sparsity and class imbalance. The ability to rank true edges at the top of massive search spaces and achieve significant enrichment across diverse datasets indicates the learned representations capture genuine biological signal rather than spurious correlations.

The enrichment results directly inform experimental design. With 39.9% precision in top 100 predictions, researchers can expect approximately one-third of tested edges to validate, compared to less than 1% for random selection. This 47-fold improvement translates to substantial resource savings. The 100% hit rate further suggests that even limited validation efforts will likely yield discoveries.

The primary limitation is data availability - most GRN references remain incomplete and few large-scale datasets combine comprehensive experimental validation with scRNA-seq measurements. Additionally, while expression correlation suggests regulatory relationships, it cannot prove causation without interventional data. Performance also depends on prior network quality, though strong results across diverse references suggest robustness.

Future research could extend this framework by incorporating time-series measurements to test causal precedence, integrating multi-modal data like chromatin accessibility, and developing uncertainty quantification through Bayesian deep learning. As single-cell technologies advance toward higher resolution and multi-modal measurements, principled integration of causal discovery methods with rigorous statistical evaluation will become increasingly important.

6 Conclusion

We presented a framework using exact statistical tests to evaluate gene regulatory network inference from single-cell RNA sequencing. By focusing on enrichment in top predictions where experimental validation occurs, we demonstrate that meaningful regulatory structure can be recovered despite extreme sparsity and class imbalance.

Perfect ranking of held-out edges, extreme enrichment significance, and consistent performance across 44 datasets validate our approach. The 47-fold improvement in validation efficiency at top 100 predictions provides immediate practical value. Exact tests that quantify enrichment in top-ranked predictions give clear answers to the questions researchers ask and complement global metrics. This framework provides a foundation for rigorous causal structure discovery from observational single-cell data when evaluation aligns with practical scientific objectives.

References

[1] Claudia Angelini and Valerio Costa. Understanding gene regulatory mechanisms by integrating chip-seq and rna-seq data: statistical solutions to biological problems. *Frontiers in cell and developmental biology*, 2:51, 2014.

[2] Guangyi Chen and Zhiping Liu. Graph attention network for link prediction of gene regulations from single-cell rna-sequencing data. *Bioinformatics*, 2022. URL `https://api.semanticscholar.org/CorpusID:251540037`.

[3] James E Darnell Jr. Variety in the level of gene control in eukaryotic cells. *Nature*, 297(5865): 365–371, 1982.

[4] Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717–729, 2010.

[5] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, 2:38, 2014.

[6] Yue Fan and Xiuli Ma. Gene regulatory network inference using 3d convolutional neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):99–106, May 2021. doi: 10.1609/aaai.v35i1.16082. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16082`.

[7] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[8] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9:1–13, 2008.

[9] David S Latchman. Eukaryotic transcription factors. *Biochemical journal*, 270(2):281, 1990.

[10] David S Latchman. Inhibitory transcription factors. *The international journal of biochemistry & cell biology*, 28(9):965–974, 1996.

[11] Guo Mao, Zhengbin Pang, Ke Zuo, Qinglin Wang, Xiangdong Pei, Xinhai Chen, and Jie Liu. Predicting gene regulatory links from single-cell rna-seq data using graph neural networks. *Briefings in Bioinformatics*, 24, 2023. URL `https://api.semanticscholar.org/CorpusID:265307839`.

[12] Hung Nguyen, Duc Tran, Bang Tran, Bahadir Pehlivan, and Tin Nguyen. A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. *Briefings in bioinformatics*, 22(3):bbaa190, 2021.

[13] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and T M Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, 2020.

[14] Claudia Skok Gibbs, Christopher A Jackson, Giuseppe-Antonio Saldi, Andreas Tjärnberg, Aashna Shah, Aaron Watters, Nicholas De Veaux, Konstantine Tchourine, Ren Yi, Tymor Hamamsy, Dayanne M Castro, Nicholas Carriero, Bram L Gorissen, David Gresham, Emily R Miraldi, and Richard Bonneau. High-performance single-cell gene regulatory network inference at scale: the inferelator 3.0. *Bioinformatics*, 38(9):2519–2528, 02 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac117. URL `https://doi.org/10.1093/bioinformatics/btac117`.

A Extended Results

A.1 Detailed Performance

Table 3 shows results for representative datasets from each reference type.

Table 3: Performance on representative datasets. P@100 is precision in top 100 predictions.

Dataset	Reference	ROC-AUC	PR-AUC	P@100	Lift@100
mESC	Perturbation	0.994	0.467	0.740	121.3
mHSC-E	STRING	0.977	0.325	0.380	62.3
hESC	Non-Specific	0.968	0.218	0.310	50.8
mDC	Cell-Specific	0.892	0.683	0.650	58.2

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that the paper presents a framework using exact statistical tests for evaluating GRN inference, with specific performance metrics reported that match the experimental results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] Justification: The Discussion section explicitly addresses limitations including incomplete GRN references, the inability to prove causation without interventional data, and dependence on prior network quality.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not present theoretical results or proofs; it is an empirical study using established statistical tests.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the methods, specifies the BEELINE benchmark datasets used, provides hyperparameters in the appendix, and states that code will be released upon acceptance.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper uses publicly available BEELINE datasets and the Crohn disease dataset (GEO accession GSE134809), with a commitment to release implementation code upon acceptance.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: The appendix provides implementation details including embedding dimensions (d=128), temperature (T=0.1), decay parameter (α=5.0), learning rate (10≙3), and training epochs (100).

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper extensively reports Fisher's exact test p-values for enrichment analysis and provides standard deviations for performance metrics across datasets in Table 1.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix specifies that experiments run on a single NVIDIA A100 GPU, with typical runtime of 3 minutes for representation learning and 12 minutes for edge scoring, using less than 8GB memory.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available biological data and aims to advance understanding of gene regulation for medical applications, with no apparent ethical concerns.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive impacts for understanding disease mechanisms and therapeutic design. As foundational biological research with no direct negative applications, extensive discussion of negative impacts is not warranted.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper presents methods for biological network inference that pose no apparent risk for misuse requiring safeguards.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites the BEELINE benchmark framework and provides GEO accession numbers for public datasets, though specific license terms could be made more explicit.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

Justification: No such assets are introduced.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects involved.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects involved.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used as part of the core methodology for gene regulatory network inference or statistical testing presented in this research.