BEACON: Bayesian Contrastive Learning for Single-Cell Gene Regulatory Inference

Yunwei Zhao
New York University
yz5944@nyu.edu

Ankit Bhardwaj
New York University
ab9738@nyu.edu

Lakshminarayanan Subramanian
New York University
lakshmi@nyu.edu

Abstract

Gene regulatory networks (GRNs) govern cellular processes by encoding interactions between transcription factors (TFs) and their target genes. While computational GRN inference has traditionally relied on bulk RNA sequencing data, such approaches fail to capture cellular heterogeneity. Single-cell RNA sequencing (scRNA-seq) offers higher resolution but introduces additional challenges, including sparsity, technical noise, and dropout effects. Existing GRN inference methods struggle with these limitations, often yielding near-random predictions in benchmarking studies. To address this, we introduce BEACON (BayEsiAn COntrastive learNing for regulatory inference) for robust GRN edge discovery from static scRNA-seq data. BEACON learns gene representations using asymmetric dual embeddings via contrastive loss on known positive and sampled negative edges. It then employs Bayesian inference to generate calibrated probability scores for each edge. We evaluate BEACON on both simulated and real-world scRNA-seq datasets with experimentally established regulatory relationships. BEACON achieves competitive performance on experimentally-validated networks from ChIP-seq and perturbation studies. This pattern suggests it captures genuine regulatory signals rather than spurious correlations. An ablation study shows that asymmetric embeddings, contrastive learning, and Bayesian scoring are all essential to achieving this performance. These findings suggest that BEACON mitigates key scRNA-seq limitations, providing a scalable and biologically grounded solution for GRN inference. While current work treats regulatory interactions as static, BEACON provides a foundation for future conditional network modeling where regulatory relationships depend on cellular and environmental factors beyond expression alone.

1 Introduction

Gene regulatory networks (GRNs) represent the intricate web of interactions between transcription factors (TFs), genes, and other regulatory elements that control gene expression, and their existence is supported by extensive experimental and theoretical evidence [22, 7, 21, 2]. They are fundamental to our understanding of how genes are expressed, regulated, and their function within biological systems. GRNs are typically modeled as directed graphs, where nodes represent genes and edges denote regulatory influences, such as TFs modulating target gene expression. They are central to a wide range of tasks like understanding cell fate decisions, disease mechanisms, and designing therapeutic interventions [11]. However, experimental validation of GRNs is challenging, resource-intensive and time-consuming, even for individual links. As such, the full extent of GRNs in most organisms is unknown. Current experimental techniques often uncover only parts of the network, leaving gaps in the global picture.

Historically, GRN inference has relied heavily on computational methods applied to gene expression data from bulk population sequencing technologies [19, 20, 8, 28]. While these methods provide

The NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025), San Diego, California, USA, 2025.

valuable insights into transcriptional interactions, they are limited by the inability to capture cellular heterogeneity and cannot resolve spatial or temporal differences between individual cells. Bulk sequencing measures average gene expression across populations of cells, making it insufficient for studying complex systems, such as tissues at early developmental stages or those with diverse cell types. Consequently, GRNs constructed from bulk data often fail to describe specific gene relationships in individual cell types or developmental phases, and they struggle with the sparsity and large sample sizes typical of single-cell datasets.

Recent advances in biochemistry and single-cell RNA sequencing (scRNA-seq) technology allow the study of GRNs at single-cell resolution, offering unprecedented detail [33, 34]. These developments have driven the emergence of GRN inference methods specifically tailored for scRNA-seq data, addressing key limitations of bulk approaches. Despite these advances, many single-cell GRN inference methods face challenges in scalability, noise handling, and validation, limiting their broader applications in real-world settings. Furthermore, the technical limitations of scRNA-seq platforms and the inherent noise and complexity of single-cell data complicate network validation and raise questions about the reliability of computational predictions.

In a prior simulation study, Nguyen et al. [29] examined a comprehensive collection of network inference methods against their synthetic benchmark. The inference methods picked for the analysis covered Boolean model methods (Boolean Pseudotime [17], BTR [23], and SCNS [39]), differential equation methods (Inference Snapshot [30], SCODE [27], and SCOUP [26]), gene correlation methods (Empirical Bayes [4], Information Measures [3], NLNET [24], SINCERA [16], and SCENIC [1]), and correlation ensemble over pseudo-time methods (LEAP [36], SINCERITIES [31], SCIMITAR [6], and SCINGE [9]). Nguyen et al. found that most of these methods do not show strong performance on simulated data, especially as gene count and sparsity increase. Many methods produce Area under the Receiver Operating Characteristic Curve (AUCROC) values close to 0.50, indicating near-random predictions, with only a few methods performing marginally better. The results highlight the limitations of current GRN inference methods when applied to complex and large-scale datasets, particularly in the presence of sparsity, a common feature in single-cell RNA-seq data.

Inferring complete gene regulatory network (GRN) from single-cell data is highly challenging due to the significant biological and technical noise inherent in such datasets. Single-cell RNA sequencing (scRNA-seq) produces sparse, noisy data where dropouts and variability between cells complicate the reconstruction of accurate, comprehensive networks. This is due to technical dropouts where transcripts fail to be detected, limited capture efficiency, and biological heterogeneity, causing many genes having zero or near-zero counts across a large fraction of cells. Additionally, the heterogeneity of cells at various developmental stages further increases the difficulty of distinguishing true regulatory relationships from random correlations. These challenges make fully data-driven network inference methods prone to errors and limit their practical applicability. However, many organisms already have experimentally validated partial networks, offering a more reliable foundation for network inference. Incorporating this partial network information into the learning process represents a more robust problem formulation, as it allows the algorithm to anchor its predictions on known regulatory interactions. By leveraging experimentally verified GRN structures, we can enhance predictions and make network reconstruction from noisy single-cell data more feasible and biologically meaningful. Thus, we have specifically focused on the GRN completion scenario instead of complete GRN inference.

Another important point to note is our focus on static data setting. While dynamic scRNA-seq can provide valuable temporal information on gene expression, it is often more challenging to obtain and analyze due to the complexity of tracking cellular states over time. Static data, on the other hand, offers a more accessible and scalable approach, as the widespread availability of static datasets also supports the development of robust computational methods tailored to noisy single-cell data, facilitating GRN inference and enabling broader applications across developmental biology, disease modeling, and drug discovery.

We propose **BEACON: Bayesian Contrastive Learning for Regulatory Inference**, a new method for GRN edge discovery on static scRNA-seq data that leverages non-linear embeddings and contrastive learning to infer regulatory interactions robustly, even in the presence of biological and technical noise. By incorporating both positive regulatory interactions from partial known networks and sampled negative edges, BEACON aims to provide a more accurate and scalable solution for GRN inference from single-cell datasets. The proposed method builds on recent developments in

machine learning to improve edge detection accuracy and offer a deeper understanding of gene regulation dynamics in heterogeneous biological systems. We further incorporate causal modeling in BEACON by integrating Bayesian frameworks and leveraging simulated and real world scRNA-seq data to systematically validate our approach. Our method aims to overcome existing limitations by explicitly modeling uncertainty and incorporating biological priors. On both simulated and real-world datasets, our method substantially outperforms state-of-the-art network inference baselines.

2 Related Work

Gene Regulatory Network inference from single-cell RNA sequencing data has emerged as a critical research direction in computational biology. The latest related methods can be coarsely classified based on their computational approaches into regression-based methods and deep learning methods [29].

Regression-based approaches model gene expression as a function of potential regulators to identify significant regulatory relationships. As a state-of-the-art regression-based method, Inferelator 3.0 [35] employs multiple regularized regression techniques to learn these relationships, including Bayesian best-subset regression, StARS-LASSO regression with stability selection, and multi-task learning regression. While this computational framework demonstrates remarkable scalability in analyzing datasets containing millions of cells, it faces limitations in capturing non-linear relationships between genes due to its reliance on linear modeling.

Recent advances in deep learning have led to new approaches for GRN inference. 3DCEMA [12] uses three-dimensional convolutional neural network (CNN) architectures to model regulatory patterns between genes. This method considers gene triplets rather than pairs to address technical artifacts in scRNA-seq data such as dropouts and noise. The use of triplet comparisons provides robustness against extreme expression values. However, processing 3D matrices introduces substantial computational overhead that impacts scalability. The method also requires careful optimization of the training process to maintain balanced representation of regulatory classes to avoid bias.

Graph neural network (GNN) architectures are also used for gene regulatory network inference by utilizing both network topology and gene expression data. GENELink [5] applies graph attention networks to transform the inference problem into a link prediction task using incomplete prior networks and gene expression data. This method demonstrates the potential of network-based approaches but faces challenges with negative sample generation and computational complexity that limit its applications. GNNLink [25] addresses these limitations by introducing a graph convolutional network-based interaction graph encoder for feature extraction from both network structure and expression data. This method achieves improved prediction accuracy and reduced training time compared to previous approaches. However, the black-box nature of deep learning models makes it difficult to interpret the biological significance of learned features and model predictions. This limitation impacts the method's utility for understanding underlying regulatory mechanisms.

3 Methodology

We propose an algorithm for gene regulatory network (GRN) inference called **BEACON: Bayesian Contrastive Learning for Regulatory Inference**, that combines probabilistic embedding learning, contrastive regularization, and a Bayesian scoring approach.

3.1 Dual-Head TF-Target Alignment

We begin with gene expression data represented as a matrix where rows correspond to genes and columns correspond to cells. We use these gene expression profiles directly as gene embeddings and employ a two-head projection model with a Soft Nearest Neighbor (SNN) Loss [13] to refine them for capturing regulatory interactions. The core objective is to bring embeddings of genes that are likely to share a regulatory relationship closer together, while pushing unrelated genes farther apart in the embedding space.

Regulatory interactions are inherently directional, so we model each gene with two distinct embeddings: a *source* embedding and a *target* embedding. The *source* embedding z_i^S encodes how gene i acts as a potential regulator, while the *target* embedding z_j^T encodes how gene j behaves as a

potential target. This two-head projection model captures the inherent asymmetry in gene regulation, where a TF may activate or repress its targets but not necessarily vice versa.

Specifically, for a gene pair (i, j) in the partial known network, we consider it a *positive pair* if i regulates j according to the prior network knowledge. During training, we project each gene i to z_i^S using a *source* projection head, and each gene j to z_j^T using a target projection head. We then measure their distance, $d(z_i^S, z_j^T)$. Given (i, j) to be a known regulatory pair, the distance is encouraged to be small; for a randomly sampled *negative pair* with no known edge, the distance is pushed larger. Unlike approaches that limit negative sampling to genes known to be involved in previously characterized regulatory interactions [35], our method considers all possible gene pairs, including those not previously associated with any regulatory context. Since the prevalence of negative edges in gene regulatory networks is extremely high, selecting negative pairs from the complete gene set is both more general and more challenging. This broader approach ensures that we are not implicitly restricting our search space to previously studied genes and interactions, thereby making our inference algorithm applicable to less-characterized systems and more robust against biases in available prior information.

We optimize a soft contrastive objective using the SNN Loss:

$$\mathcal{L}\text{SNN}(\mathbf{z}) = -\log\left(\frac{\sum(i,j) \in P \exp\left(-\frac{d(\mathbf{z}_i,\mathbf{z}j)}{T}\right)}{\sum(i,j)\exp\left(-\frac{d(\mathbf{z}_i,\mathbf{z}_j)}{T}\right)}\right), \tag{1}$$

where P denotes the set of positive pairs, $d(\cdot)$ measures the Cosine distance between embeddings, and T is a temperature parameter that controls the smoothness of the exponential weighting. By applying equation (1) to our *directional* embeddings, we can effectively preserve critical asymmetry information in regulatory relationships.

3.2 Bayesian Edge Scoring

For each potential edge $e \in E(H \backslash G)$, where H is the complete network and G is the known partial network, we aim to estimate its posterior probability:

$$P(e \in H|x, G) \propto P(x|e, G)P(e|G)$$

where x represents our processed data, $P(e|G)$ is the *prior*, and $P(x|e, G)$ is the *likelihood*.

Based on the assumption that the distance between gene embeddings reflects the likelihood of regulatory interactions, we define a likelihood function:

$$P(x|e, G) \approx f(d(z_i, z_j)) \ = \exp(-\alpha d(z_i, z_j))$$

where f is a decreasing function and α is a hyperparamter that controls the decay. $P(e|G)$ is prior probability of edge e existing, derived from the partial known network structure. We set prior for positive edges to be 1, negative edges to be 0, and unknown edges to be the average positive rate over known network, formulated as:

$$P(e|G) = \frac{\sum_{i,j} \mathbb{I}(G_{ij} = 1)}{\sum_{i,j} \mathbb{I}(G_{ij} = 1 \vee G_{ij} = 0)}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

After computing the unnormalized score $score(e) = \exp(-\alpha \cdot d(z_i, z_j)) \times P(e|G)$, we normalize over all candidate edges in $H \backslash G$ to produce a valid posterior probability in $[0, 1]$ that provides relative probability distribution over the unknown edges:

$$P(e \in H|x, G) = \frac{\text{score}(e)}{\sum_{e' \in \{H \backslash G\}} \text{score}(e')}$$

The naive likelihood function assumes a fixed parametric form for the relationship between embedding distances and regulatory interactions. The model's performance depends on hyperparameters such as α. In addition, the average density assigned to unknown network by the prior function can bias

predictions toward the known network. To overcome these limitations, we use Gaussian Process (GP) classification, a nonparametric model that can approximate a flexible likelihood function $f(x)$ directly from data. For each edge (i, j), we construct a feature vector $x_{ij} = [z_i; z_j]$ by concatenating the embeddings of nodes i and j. The GP model defines a latent function:

$$f(x_{ij}) \sim \mathcal{GP}(0, k(x_{ij}, x_{i'j'}))$$

We employ the radial basis function (RBF) kernel:

$$k(x_{ij}, x_{i'j'}) = \exp(-\gamma||x_{ij} - x_{i'j'}||^2)$$

which measures similarity between feature vectors while capturing nonlinear relationships. The likelihood function is then modeled as:

$$P(y_{ij} = 1|f(x_{ij})) = \sigma(f(x_{ij}))$$
$$P(y_{ij} = 0|f(x_{ij})) = 1 - \sigma(f(x_{ij}))$$

where $\sigma(\cdot)$ is the sigmoid function. Due to the intractable nature of the exact posterior in GP classification, we define a Variational Evidence Lower Bound (ELBO) as our optimization function:

$$\text{ELBO} = \mathbb{E}_{q(f)} \left[\sum_{(i,j)\in D} \log P(y_{ij}|f(x_{ij})) \right] - KL\left[q(f)||p(f)\right]$$

where $q(f)$ denotes the variational distribution approximating the true posterior $p(f|X, y)$ and $KL[.||.]$ denotes Kullback-Leibler divergence. This ELBO provides a tractable objective that balances the model's fit to the data with the complexity of the approximating distribution.

After training, the GP model learning a direct mapping from embeddings to edge probabilities:

$$P(e \in H|x, G) \approx P(y_{ij} = 1|f(x_{ij}))$$

This section presents the experimental evaluation of BEACON on both simulated and real scRNA-seq datasets. We follow benchmarking practices [32] by assessing performance on in silico SERGIO-simulated data as well as real scRNA-seq data. The evaluation is designed to achieve two main goals. We first quantitatively evaluate BEACON's performance against several state-of-the-art GRN inference methods on common benchmarks. We then analyze the contribution of each BEACON module by comparing the full model with its variants.

4 Experimental Results

4.1 Baselines

We compare BEACON against several representative baseline methods, spanning three categories of GRN inference baselines: (1) one regression-based method Inferelator 3.0, (2) one CNN-based method 3DCEMA, (3) two GNN-based methods including GENELink and GNNLink. For each method, we run three times with different seeds and average their performance scores. For Inferelator 3.0, we run three times with each of the three regression algorithms they implemented and then get the averaged performance. This can lead to a fair comparison with their method. All baseline methods were either obtained from the authors' released code, with hyperparameters set according to the respective papers or tuned on our validation sets for fairness.

4.2 Datasets

We first simulate "ground-truth" gene expression data as described in §A.1.1. We then add "technical noise" to the ground truth data, mimicking the nature of measurement errors attributed to scRNA-seq technology, as described in §A.1.2. For our experiments, we reproduce three independent steady state simulation datasets (DS1, DS2, and DS3) as described in the SERGIO paper. The properties of these datasets is given in Table 5. The Gene Regulatory Networks used for the generation of these datasets are a subset of the experimentally validated networks for E. coli and S. cerevisiae.

We did extensive evaluation of BEACON on multiple real single-cell expression datasets from the BEELINE benchmark [32]. The real scRNA-seq datasets are taken from the BEELINE collection,

Table 1: Edge prediction AUROC (%) on test set for different methods on SERGIO-simulated datasets.

METHODS	DS1	DS2	DS3
INFERELATOR 3.0	96.79±0.17	95.86±0.14	95.95±0.05
3DCEMA	99.03±0.29	83.81±0.66	OOM
GNNLINK	91.67 ±0.94	74.46±5.70	91.12±0.44
GENELINK	99.10±0.11	96.68±0.18	89.06±0.11
BEACON *(Ours)*	97.24±0.10	97.15±0.59	94.85±0.26

which includes diverse cell types (e.g., human embryonic stem cells, mouse dendritic cells, various mouse hematopoietic lineages). These scRNA-seq expression datasets and the ground truth networks are completely independent, coming from different experimental sources [32]. For each real dataset, we use the corresponding reference GRN provided by the benchmark – these reference networks are constructed from experimental evidence such as protein–DNA binding (ChIP-seq) or perturbation studies. In particular, the reference GRNs encompass four types of regulatory interaction evidence: known functional interactions (e.g., STRING database), loss-of-function/gain-of-function (LOF/GOF) perturbation data, non-specific ChIP-seq binding networks, and cell-type-specific ChIP-seq networks. This variety allows us to test BEACON under different ground-truth conditions, from broadly conserved networks (STRING, LOF/GOF) to context-specific regulatory networks (cell-type-specific interactions). As shown in Table 7, these networks are extremely sparse, with densities below 2.4%, meaning over 97% of possible edges are negative. The properties of these datasets is given in Table 6 and Table 7.

We evaluate two gene set sizes: TFs+500 and TFs+1000, where we include all significantly varying TFs plus either 500 or 1000 most varying genes. For mHSC, we evaluate three lineages separately.

Detailed discussions on dataset selection and processing are presented in Appendix A.

4.3 Implementation

We implemented BEACON using PyTorch for model components. Gene expression profiles are used directly as gene embeddings without dimensionality reduction. We evaluate using AUROC and AUPRC. AUPRC is more informative for sparse networks as it directly measures precision-recall trade-offs [32].

We partition edges into 80% training, 10% validation, and 10% test sets. We sample negative edges uniformly from all non-edges.

4.4 Simulated Datasets

Tables 1 and 2 summarize performance on simulated datasets. BEACON demonstrates stable and consistent performance across all three datasets, with relatively small standard deviations indicating reliable results across different runs. In contrast, several baseline methods exhibit high variance - for example, Inferelator 3.0 shows standard deviations up to 16.27 on AUPRC for DS2, and GENELINK displays substantial instability across runs. While 3DCEMA achieves competitive scores on smaller datasets, it fails to scale to DS3 due to memory limitations.

The AUPRC results reveal systematic performance degradation correlated with network sparsity. BEACON maintains reasonable performance on DS1 but experiences substantial decline on DS2 and DS3. This degradation correlates directly with network density. The sparsity-dependent performance suggests that BEACON's Bayesian scoring framework requires sufficient positive examples to calibrate probability estimates effectively. Extremely sparse networks challenge this calibration, resulting in reduced precision at specific recall thresholds. Further investigation into density-adaptive scoring methods could address this limitation.

4.5 Real-world Benchmark

Figures 1a and 1b summarize the performance of BEACON and all baselines across real-world datasets. BEACON demonstrates competitive performance across multiple benchmark settings,

Table 2: Edge prediction AUPRC (%) on test set for different methods on SERGIO-simulated datasets.

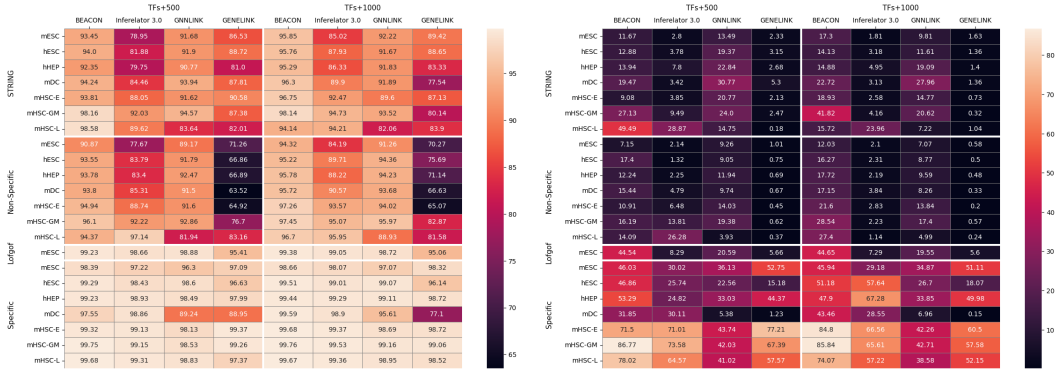
METHODS	DS1	DS2	DS3
INFERELATOR 3.0	57.73±9.27	42.46±16.27	57.14±3.40
3DCEMA	85.03±1.94	5.09±0.21	OOM
GNNLINK	20.65±0.82	5.99±0.95	22.83±0.93
GENELINK	82.30±2.08	10.99±0.58	1.36±0.50
BEACON *(Ours)*	51.55±1.26	25.84±2.01	11.59 ±0.89

Table 3: Ablation test measured by AUROC (%) on SERGIO-simulated datasets.

METHODS	DS1	DS2	DS3
BEACON	**97.24±0.10**	**97.15±0.59**	**94.85±0.26**
W/O ASYMMETRIC	87.88±1.94	93.37±0.55	94.76±0.32
W/O SNN	50.26±0.00	69.79±0.01	84.74±0.13
W/O GP SCORING	10.41±0.64	33.93±0.51	32.86±0.69
W/O ALL	36.21±0.00	41.15±0.00	47.25±0.00

particularly on STRING and non-specific ChIP-seq networks. For STRING networks capturing functional interactions and LOF/GOF networks from perturbation experiments, BEACON achieves competitive scores. These results indicate BEACON effectively recovers conserved regulatory relationships. On cell-type-specific networks, all methods face greater challenges due to context-dependent regulation and increased noise. This performance variation aligns with our design, as BEACON's Bayesian framework and contrastive learning are well-suited for capturing patterns that generalize across cell types.

The AUPRC results show greater variation than AUROC, which is expected given the extreme sparsity of these networks. In such imbalanced settings, small prediction differences substantially affect precision-recall metrics. Despite this challenge, BEACON maintains competitive performance across diverse evaluation scenarios. The consistent performance on STRING and non-specific networks suggests BEACON effectively balances between capturing direct and indirect regulatory relationships.



(a) AUROC (%) on BEELINE datasets **(b)** AUPRC (%) on BEELINE datasets

Figure 1: Performance of BEACON vs. baselines across four reference networks.

4.6 Ablation Studies

BEACON combines three modules: a directional projection to capture regulatory asymmetry, an SNN loss to refine embeddings, and Bayesian scoring for edge-level posterior inference. We systematically remove each module to assess its contribution; results in Tables 3 and 4 show that omitting any part degrades performance. Note that, different from other entries, w/o All means we do not use any module mentioned above but directly use a straightforward $-||z_i - z_j||$ distance measure to see if closely embedded gene pairs correspond to real regulatory relationships. Without directional projection, AUROC drops significantly on DS1 from 97.24% to 87.88%, confirming its importance for capturing regulatory asymmetry. Removing SNN loss causes substantial degradation across all

Table 4: Ablation test measured by AUPRC (%) on SERGIO-simulated datasets.

METHODS	DS1	DS2	DS3
BEACON	**51.55±1.26**	**25.84±2.01**	**11.59±0.89**
W/O ASYMMETRIC	43.66±0.63	16.85±0.81	26.85±1.52
W/O SNN	51.31±0.00	2.03±0.00	24.23±2.55
W/O GP SCORING	1.38±0.00	0.50±0.00	0.41±0.00
W/O ALL	1.86±0.00	0.64±0.00	0.56±0.00

datasets. The Bayesian scoring component proves most critical, with its removal causing AUROC to drop below 35% on all datasets. These results confirm that each component addresses a specific challenge in single-cell GRN inference. While we observe performance variation across simulated dataset scales DS1 to DS3, BEACON maintains consistent competitive performance on real-world datasets as shown in Figures 1a and 1b. This suggests the scale-dependent patterns in simulated data may reflect limitations of synthetic data generation rather than fundamental method limitations.

5 Discussions & Conclusion

In this paper, we have introduced BEACON: Bayesian Contrastive Learning for Regulatory Inference, a new method for gene regulatory network (GRN) inference from static scRNA-seq data. BEACON leverages non-linear embeddings, contrastive learning, and Bayesian causal modeling to infer regulatory interactions with improved robustness to biological and technical noise. By integrating partial known networks as positive interactions and sampled negatives, BEACON enhances edge discovery accuracy while explicitly modeling uncertainty.

BEACON demonstrates competitive performance across diverse evaluation scenarios. On real-world datasets, BEACON achieves competitive results across multiple network types. Performance is particularly notable on mHSC lineages, demonstrating BEACON's applicability to lineage-specific differentiation contexts. The method maintains competitive performance across different dataset configurations, from TFs+500 to TFs+1000 genes.

A key observation is the variability across different ground-truth datasets. Cell-type-specific and LOF/GOF networks yield higher absolute scores but also show greater variation among methods. STRING and non-specific ChIP-seq networks show more consistent but lower scores across methods. This pattern likely reflects the different natures of these networks. Cell-type-specific networks have higher edge densities and capture context-dependent regulation. The broader networks capture conserved but sparser regulatory patterns.

Our ablation studies confirm that each BEACON component contributes meaningfully to performance. The Bayesian scoring module proves particularly critical. The asymmetric embeddings and SNN loss work together to capture directional regulatory relationships. This integrated architecture enables BEACON to balance between direct and indirect regulatory interactions.

Limitations: Despite the competitive performance, the results highlight the ongoing challenges in single-cell GRN inference. The drop in performance for specific datasets suggests that regulatory relationships inferred from population-level methods may not always translate well into fine-grained, context-specific interactions. The complexity of cell-state transitions, sparsity in scRNA-seq data, and technical noise remain obstacles in achieving consistently high accuracy across all benchmarks. Future work will focus on further improving interpretability of inferred networks, integrating additional regulatory modalities (e.g., chromatin accessibility, epigenetics), and optimizing the method for scalability across larger datasets. Additionally, exploring hybrid models that combine statistical methods with deep learning frameworks may further enhance predictive accuracy, particularly in complex, cell-type-specific GRNs.

References

[1] 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.

[2] 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.

[3] Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3):251–267, 2017.

[4] Thalia E Chan, Ananth V Pallaseni, Ann C Babtie, Kirsten R McEwen, and Michael PH Stumpf. Empirical bayes meets information theoretical network reconstruction from single cell data. *BioRxiv*, page 264853, 2018.

[5] 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`.

[6] Pablo Cordero and Joshua M Stuart. Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories. In *Pacific Symposium on Biocomputing 2017*, pages 576–587. World Scientific, 2017.

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

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

[9] Atul Deshpande, Li-Fang Chu, Ron Stewart, and Anthony Gitter. Network inference with granger causality ensembles on single-cell transcriptomic data. *BioRxiv*, page 534834, 2019.

[10] Payam Dibaeinia and Saurabh Sinha. Sergio: A single-cell expression simulator guided by gene regulatory networks. *Cell Systems*, 11(3):252–271.e11, 2020. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2020.08.003. URL `https://www.sciencedirect.com/science/article/pii/S2405471220302878`.

[11] 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.

[12] 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`.

[13] Nicholas Frosst, Nicolas Papernot, and Geoffrey Hinton. Analyzing and improving representations with the soft nearest neighbor loss. In *International Conference on Machine Learning*, pages 2012–2020. PMLR, 2019.

[14] Collin Giguere, Harsh Vardhan Dubey, Vishal Kumar Sarsani, Hachem Saddiki, Shai He, and Patrick Flaherty. Scsim: Jointly simulating correlated single-cell and bulk next-generation dna sequencing data. *BMC Bioinformatics*, 21(1):215, May 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-03550-1. URL `https://doi.org/10.1186/s12859-020-03550-1`.

[15] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1): 297–306, 2000.

[16] Minzhe Guo, Hui Wang, S Steven Potter, Jeffrey A Whitsett, and Yan Xu. Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology*, 11(11):e1004575, 2015.

[17] Fiona K Hamey, Sonia Nestorowa, Sarah J Kinston, David G Kent, Nicola K Wilson, and Berthold Göttgens. Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proceedings of the National Academy of Sciences*, 114(23):5822–5829, 2017.

[18] https://github.com/PayamDiba/SERGIO.

[19] 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.

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

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

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

[23] Chee Yee Lim, Huange Wang, Steven Woodhouse, Nir Piterman, Lorenz Wernisch, Jasmin Fisher, and Berthold Göttgens. Btr: training asynchronous boolean models using single-cell expression data. *BMC bioinformatics*, 17:1–18, 2016.

[24] Haodong Liu, Peng Li, Mengyao Zhu, Xiaofei Wang, Jianwei Lu, and Tianwei Yu. Nonlinear network reconstruction from gene expression data using marginal dependencies measured by dcol. *PloS one*, 11(7):e0158247, 2016.

[25] 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`.

[26] Hirotaka Matsumoto and Hisanori Kiryu. Scoup: a probabilistic model based on the ornstein–uhlenbeck process to analyze single-cell expression data during differentiation. *BMC bioinformatics*, 17:1–16, 2016.

[27] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

[28] Fantine Mordelet and Jean-Philippe Vert. Sirene: supervised inference of regulatory networks. *Bioinformatics*, 24(16):i76–i82, 2008.

[29] 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.

[30] Andrea Ocone, Laleh Haghverdi, Nikola S Mueller, and Fabian J Theis. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12): i89–i96, 2015.

[31] Nan Papili Gao, SM Minhaz Ud-Dean, Olivier Gandrillon, and Rudiyanto Gunawan. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, 2018.

[32] 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.

[33] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

[34] C Wyatt Shields Iv, Catherine D Reyes, and Gabriel P López. Microfluidic cell sorting: a review of the advances in the separation of cells from debulking to rare cell isolation. *Lab on a Chip*, 15(5):1230–1249, 2015.

[35] 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.

[36] Alicia T Specht and Jun Li. Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering. *Bioinformatics*, 33(5):764–766, 2017.

[37] Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Jessica Li. scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biology*, 22(1):163, May 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02367-2. URL https://doi.org/10.1186/s13059-021-02367-2.

[38] Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: Power analysis for bulk and single cell rna-seq experiments. *bioRxiv*, 2017. doi: 10.1101/117150. URL https://www.biorxiv.org/content/early/2017/06/26/117150.

[39] Steven Woodhouse, Nir Piterman, Christoph M Wintersteiger, Berthold Göttgens, and Jasmin Fisher. Scns: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC systems biology*, 12:1–7, 2018.

[40] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[41] Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell rna sequencing. *Nature communications*, 10(1):2611, 2019.

A Technical Appendices and Supplementary Material

A.1 Synthetic Data Simulation

To benchmark the performance of network inference methods, one common strategy is to work with simulated synthetic data, as the "ground truth" regulatory network is generally hard to obtain. Some notable scRNA-seq simulators include Splatter [40], scDesign2 [37], SymSim [41], powsimR [38], SCSIM [14], etc, but none of the above simulators account for GRNs in their data generation process. As such, they are not suitable for benchmarking the performance of zero-imputation techniques on the network inference task. The SERGIO [10] simulator is one of the few that provide us with a mechanism for generating synthetic scRNA-seq data by modeling GRNs, transcriptional dynamics, and technical noise.

SERGIO begins by taking a gene regulatory network as input. In this network nodes represent genes and edges represent regulatory interactions between genes, which can be either activation or repression. The simulator requires two input files, the first: a comma-separated file where each row represents a target gene, the IDs of its regulators, the interaction strengths for each gene pair K (with positive values for activation and negative for repression), and the Hill coefficients for each pair. Master regulators are excluded from this file, but provided in the second, a comma-separated file of master regulators where each row contains the regulator's ID followed by its production rates for each cell type.

A.1.1 Clean Data Simulation

SERGIO models the dynamics of the concentration of genes using systems of stochastic differential equations that are derived from the chemical Langevin equation (CLE) [15]. For every cell, the ground-truth mRNA concentration is modeled as

$$\frac{dx_i}{dt} = P_i(t) - \lambda_i x_i(t) + q_i(\sqrt{P_i(t)}\alpha + \sqrt{\lambda_i x_i(t)}\beta)$$

$$P_i(t) = \sum_{j \in R_i} p_{ij}(t) + b_i(t)$$

$$p_{ij}(t) = K_{ij}\frac{x_j(t)^{n_{ij}}}{h_{ij}^{n_{ij}} + x_j(t)^{n_{ij}}}, \text{if j is an activator of i}$$

$$p_{ij}(t) = K_{ij}(1 - \frac{x_j(t)^{n_{ij}}}{h_{ij}^{n_{ij}} + x_j(t)^{n_{ij}}}), \text{if j is an repressor of i}$$

where,

$$\begin{aligned}
x_i(t) &\rightarrow \text{Expression value of gene i} \\
P_i(t) &\rightarrow \text{Production rate of gene i} \\
\lambda_i &\rightarrow \text{Decay rate of gene i} \\
q_i &\rightarrow \text{Noise amplitude in the transcription for gene i} \\
\alpha, \beta &\rightarrow \text{Independent Gaussian white noise processes} \\
R_i &\rightarrow \text{Regulator set of gene i} \\
b_i &\rightarrow \text{Basal production rate of gene i} \\
p_{ij}(t) &\rightarrow \text{Regulatory effect of gene j on gene i} \\
K_{ij} &\rightarrow \text{Maximum contribution of gene j on gene i} \\
n_{ij} &\rightarrow \text{Hill's coefficient} \\
h_{ij} &\rightarrow \text{Regulator concentration for half response}
\end{aligned}$$

In this formulation, a nonzero K_{ij} indicates that gene j regulates gene i, corresponding to a directed edge in the GRN from j to i. The sign and magnitude of K_{ij}, together with whether j acts as an activator or repressor, determine the direction and strength of this regulatory relationship. If $K_{ij} = 0$, no edge exists between gene j and gene i in the regulatory network.

Table 5: Properties of different datasets used for our experiments.

Dataset	# Cells	# Genes	# Cell Types	# Regulators	# Edges	# Density	Species
DS1	2700	100	9	10	258	0.0258	E. coli
DS2	2700	400	9	37	1155	0.0072	S. cerevisiae
DS3	2700	1200	9	127	2713	0.0019	E. coli

The above set of equations is defined for all genes, making the simulator a discrete dynamical system of the system of stochastic differential equations. The details of the simulator can be found in [10], and the implementation is publicly available at the GitHub repository [18].

A.1.2 Noise Simulation

The SERGIO simulation environment allows us to simulate three different noise models, corresponding to different aspects of sequencing pipelines [40]. These include

Outlier Genes: Each gene is marked as an outlier based on a probability set by the user. If a gene is an outlier, its expression levels in all cells are multiplied by a factor drawn from a log-normal distribution; if not, the expression remains unchanged.

Library Size: For each cell, a library size parameter is drawn from a log-normal distribution. The expression values of all genes in the cell are then scaled by a constant factor so that the total depth of the cell matches the sampled library size.

Dropouts: To introduce dropouts to the simulated data, a probability is initially assigned to the expression of each gene in each of the simulated cells not being a dropout. This probability is modeled as a logistic function of the expression of the gene in that cell, so that a high expression value is less likely to be zeroed out. Subsequently, this probability is used as the parameter of a Bernoulli distribution from which a binary variable is sampled to indicate whether the gene is not a dropout in the cell.

UMI Count Conversion: In steady-state simulations, UMI (Unique Molecular Identifier) counts are generated by sampling from a Poisson distribution, where the mean corresponds to the gene's simulated expression level in each cell. This introduces noise, as the UMI conversion process inherently adds stochasticity, reflecting biological and technical variability in the captured gene expression.

A.1.3 Dataset Details

Each simulated dataset has a clean and a noisy version, where noise is later added to the dataset. For the noisy version of each of the simulated datasets, we configured SERGIO to introduce technical noise to an extent that matches the published real scRNA-seq dataset on mouse cerebral cortex generated by Illumina HiSeq, 2000 sequencing methodology. The details of the noise iteration process, and comparison between datasets can be found in the SERGIO paper [10].

One critical challenge in GRN inference is the extreme sparsity of regulatory interactions. As shown in Table 5, the density of positive edges ranges from only 2.58% in DS1 to 0.19% in DS3, making the network inference task essentially a needle-in-haystack problem. This imbalance between positive and negative edges possesses risks in proper GRN inference. However, with careful consideration in both model design and evaluation metrics, the payoff is a more accurate and robust method for GRN inference, which could have substantial impact in computational biology.

A.1.4 Network Splitting

In many real-world scenarios, we possess partial knowledge of the GRN (e.g., curated TF–target pairs). Let H be the complete ground-truth network for a given dataset, and let $G \subseteq H$ be the known subset of edges. We randomly sample G by permuting all edges in H and selecting a fraction sample ratio (default 0.9). The edges in G are further partitioned into training and validation sets ($8:1$ ratio), while the remaining edges in $H \backslash G$ form our test set. This sampling resulted in a distribution of edges across train, validation, and test sets in the approximate ratio of $8:1:1$.

This setup simulates the practical condition where a researcher has partial, curated regulatory links but aims to discover unknown interactions. Table 5 details the number of positive (true regulatory) vs. negative (non-regulatory) edges in each split.

A.2 Real World Data

A.2.1 Data Splitting

For each dataset, we partition the available data into train/valid/test splits in a consistent manner, as mentioned in section A.1.4. We maintain the same ratio of positive to negative edges across all splits of each dataset. To ensure fairness, we fix the random seed for splitting and apply the same strategy across all datasets, so that differences in performance reflect model capability rather than data selection. The validation set is used for hyper-parameter tuning and early stopping, while the held-out test set is used only during final evaluations.

Table 6: The statistics of single-cell transcriptomic datasets with TFs and 500 (1000) most-varying genes.

Cell Type	Cells	TFs	Genes
hESC	758	410 (410)	910 (1410)
hHEP	425	448 (449)	948 (1448)
mDC	383	323 (323)	821 (1321)
mESC	421	627 (628)	1120 (1620)
mHSC-E	1071	205 (209)	704 (1204)
mHSC-GM	889	135 (136)	632 (1132)
mHSC-L	847	61 (61)	560 (692)

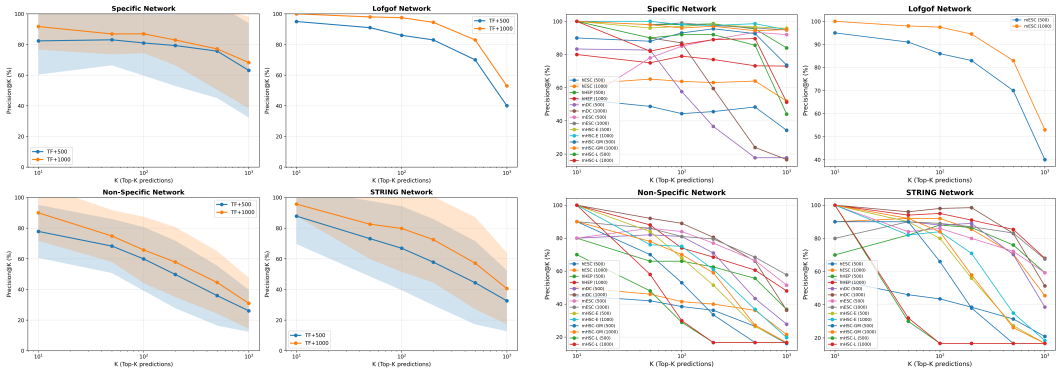
Table 7: Densities of four ground-truth networks (STRING, Non-specific ChIP-seq (Non-specific), Cell-type-specific ChIP-seq (Specific), LOF/GOF) for the same datasets. Parentheses indicate values for 1000 genes. A dash (–) indicates no data for that network/cell type.

Cell Type	Density			
	STRING	**Non-specific**	**Specific**	**LOF/GOF**
hESC	0.0051 (0.0026)	0.0042 (0.0023)	0.0055 (0.0036)	–
hHEP	0.0084 (0.0043)	0.0046 (0.0026)	0.0111 (0.0074)	–
mDC	0.0071 (0.0034)	0.0045 (0.0022)	0.0011 (0.0007)	–
mESC	0.0062 (0.0032)	0.0055 (0.0031)	0.0236 (0.0163)	0.0033 (0.0022)
mHSC-E	0.0028 (0.0013)	0.0029 (0.0014)	0.0233 (0.0152)	–
mHSC-GM	0.0019 (0.0010)	0.0019 (0.0011)	0.0184 (0.0110)	–
mHSC-L	0.0004 (0.0003)	0.0009 (0.0007)	0.0140 (0.0108)	–

A.3 Experimental Validation

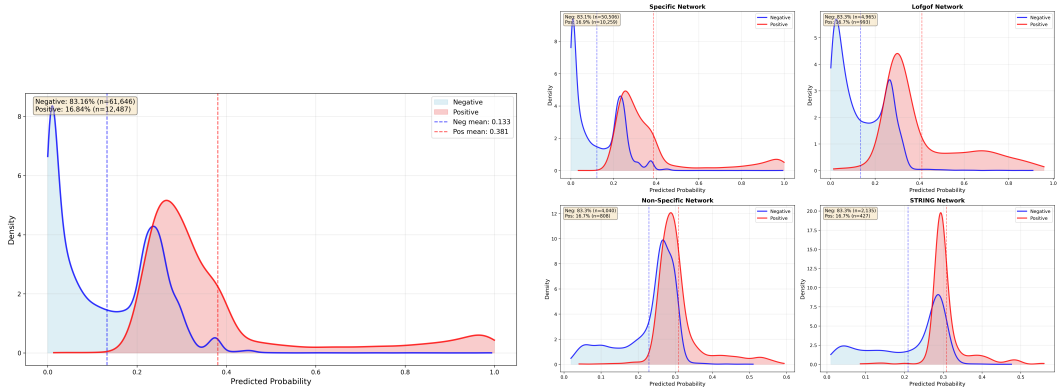
To assess BEACON's utility for experimental validation, we evaluated Precision@K across K $\in$ {10, 50, 100, 200, 500, 1000}, simulating scenarios where researchers validate top predictions in the lab. As shown in Figure 2a, performance follows a clear hierarchy tied to ground-truth quality: LOF/GOF networks achieve near-perfect precision exceeding 95% at K=10 and sustain over 80% even at K=200, while Specific ChIP-seq networks show comparable strength with most datasets exceeding 90% at K=50. By contrast, STRING and Non-Specific networks decline more steeply, falling below 50% precision by K=200-500.

This hierarchy reflects the inherent trustworthiness of each reference network type. LOF/GOF captures causal genetic interactions validated through loss-of-function screens, representing the gold standard for regulatory relationships. Specific ChIP-seq provides direct cell-type matched TF binding evidence. Meanwhile, STRING aggregates heterogeneous evidence across studies and organisms, and Non-Specific ChIP-seq lacks cell-type resolution, both introducing noise. BEACON's better performance on higher-quality networks indicates the model learns genuine biological patterns rather than noise in the reference networks.



(a) Precision@K curves across network types **(b)** Dataset-specific precision curves

Figure 2: Precision@K performance for experimental validation. **(a)** LOF/GOF and Specific networks maintain high precision at large K values, while STRING and Non-Specific networks decline rapidly. **(b)** Individual dataset curves show consistent performance within network categories across cell types.



(a) Overall probability distribution **(b)** Distribution by network category

Figure 3: Predicted probability distributions for positive and negative samples. **(a)** BEACON achieves 2.8-fold separation between true edges (mean=0.379) and non-edges (mean=0.133). **(b)** Separation quality correlates with ground-truth reliability: LOF/GOF shows strongest separation, followed by Specific ChIP-seq, STRING, and Non-Specific networks.

Beyond network-level trends, dataset-level analysis in Figure 2a reveals BEACON maintains consistent performance across diverse cell types. Within each network category, precision curves cluster tightly, typically varying by less than 10% across cell types at any given K. Notably, the 1000-gene configuration consistently outperforms the 500-gene configuration across all network types, indicating that BEACON consistently benefits from increased gene coverage.

A.4 Probability Calibration

Figure 3 shows the predicted probability distributions for positive and negative samples. Across all networks, BEACON achieves clear separation with mean probabilities of 0.379 for true edges versus 0.133 for non-edges, a 2.8-fold difference. The negative distribution shows a sharp peak near low probabilities, while the positive distribution displays a broader shape with a long tail extending toward higher probabilities, indicating many true edges achieve high-confidence scores suitable for experimental prioritization.

Category-specific analysis (Figure 3) reveals that separation quality directly correlates with ground-truth reliability. BEACON achieves the highest separation between positive and negative mean probabilities for LOF/GOF networks at 0.276 and Specific ChIP-seq networks at 0.265. In contrast, STRING and Non-Specific ChIP-seq networks show minimal separation at 0.099 and 0.080 respec-

tively. This pattern indicates that BEACON's predictive performance depends more on the quality of training labels than on model limitations, with higher-quality reference networks enabling better predictions.