GLM-Prior: A Genomic Language Model for Sequence-Derived Prior Knowledge in GRN Inference

Claudia Skok Gibbs
Center for Data Science,
New York University
csg337@nyu.edu

Angelica Chen
Center for Data Science,
New York University
ac5968@nyu.edu

Richard Bonneau
Prescient Design,
Genentech
bonneau.richard@gene.com

Kyunghyun Cho *
Center for Data Science,
New York University
& Prescient Design,
Genentech
kyunghyun.cho@nyu.edu

Abstract

Gene regulatory network (GRN) inference relies on high-quality prior knowledge, which are often incomplete or unavailable, particularly for complex organisms and diverse cell types. We present GLM-Prior, a genomic language model that fine-tunes the pretrained Nucleotide Transformer to learn transcription factor to target gene regulatory interactions from nucleotide sequence, yielding a sequence-derived prior for downstream GRN inference. In yeast, GLM-Prior outperforms motif-based and curated prior knowledge. When trained on general interaction data in human or mouse, GLM-Prior recovers cell line-specific regulatory structure and enables zero-shot transfer between species. Across settings, adding expression-based inference provides only modest improvements, indicating that most recoverable regulatory structure is capture by sequence features learned by GLM-Prior. These results support sequence-derived prior knowledge as a strong basis for GRN inference, with expression data used primarily to refine and contextualize a fixed regulatory scaffold.

1 Introduction

Gene regulatory networks (GRNs) map the transcriptional relationships between transcription factors (TFs) and their target genes, providing a framework for understanding cellular function and gene expression control in cells [1, 2]. Accurate GRN inference depends heavily on prior knowledge, which describes an initial set of putative TF-gene interactions that guides the inference process [3, 4]. Prior knowledge for well-studied species is typically constructed using databases of experimentally validated TF-gene interactions [5, 6]. For less-characterized organisms, prior knowledge is typically inferred using motif-based and accessibility-driven approaches, which combine structural genomic data with sequencing information to identify putative regulatory sites [7]. However, these methods are limited by incomplete annotations, noisy data, and an inability to capture long-range regulation [8–10].

Large-scale genomic foundation models offer a powerful alternative to motif-based methods by learning regulatory logic directly from DNA sequence [11–14]. Transformer-based architectures, such

*Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

as the Nucleotide Transformer [15], use attention mechanisms to capture long-range dependencies and encode both species-specific and cross-species regulatory patterns. While these models are pretrained on massive genomic corpora to learn general-purpose sequence representations, they can be fine-tuned using TF–target gene interaction data to accurately predict regulatory relationships. Once fine-tuned, these models can generalize to contexts where curated priors are unavailable, including zero-shot transfer across species or prediction within specific cell types [11, 16, 15, 17].

In this work, we present GLM-Prior, a genomic language model obtained by fine-tuning the pre-trained 250M-parameter Nucleotide Transformer [15] to predict regulatory interactions from paired nucleotide sequences of TF binding motifs and gene bodies. Pretrained on the genomes of 850 species, the model jointly encodes each TF-gene sequence pair and passes the resulting representation through a classification head to estimate the probability of a regulatory interaction. Unlike motif-based methods that rely on proximity assumptions, GLM-Prior leverages the transformer's capacity to model long-range dependencies and complex regulatory grammar, enabling improved generalization across cell types and species.

In yeast, GLM-Prior outperforms both motif-based and curated priors, recovering most regulatory interactions directly from nucleotide sequence. Incorporating expression data during downstream GRN inference yields only marginal improvements, indicating that the sequence-derived prior already captures the majority of meaningful structure. In human and mouse, models trained on general species-level interactions accurately reconstruct cell line–specific regulatory edges and enable zero-shot transfer from human to mouse without retraining. Together, these results suggest that high-quality, sequence-derived prior knowledge from genomic language models can form a robust foundation for GRN inference, with expression data serving primarily to modulate this scaffold in a context-specific manner rather than define its structure.

2 Methods

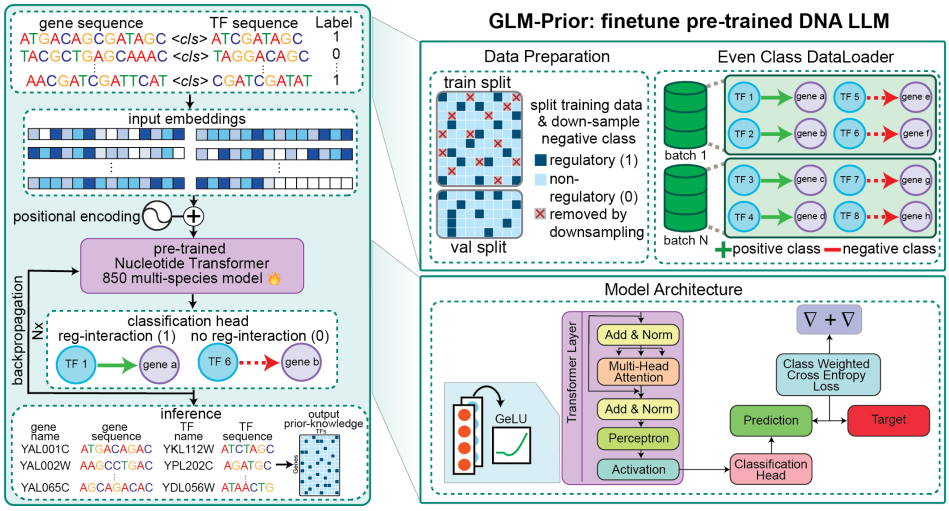

Figure 1: GLM-Prior processes concatenated TF motif and gene sequences to learn regulatory interactions from nucleotide sequence. The model is trained on balanced, downsampled batches to predict interaction probabilities, generating sequence-derived prior knowledge for downstream GRN inference.

2.1 Model Overview

GLM-Prior is a genomic language model built by fine-tuning the 250M-parameter Nucleotide Transformer [15] to predict transcription factor–target gene regulatory interactions directly from DNA sequence (Figure 1). The model uses a transformer encoder architecture to process two distinct sequence types: TF motif sequences from the CisBP database [18] and gene body sequences derived from genome annotations in a GTF file. Each TF-gene input pair is formed by concatenating their respective sequences, which are then tokenized and passed through the transformer encoder. The (`<cls>`) token embedding from the final hidden layer is passed through a classification head to predict whether each TF-gene pair represents a true regulatory interaction (positive label) or a non-regulatory interaction (negative label), producing logits $z \in \mathbb{R}^2$ with class probabilities:

$$P(y = 1|x_{TF}, x_{gene}) = \frac{\exp(z_+)}{\exp(z_+) + \exp(z_-)}. \tag{1}$$

Training labels are derived from experimentally validated TF-gene interaction databases (e.g., YEASTRACT [6], STRING [19–22] and TRRUST [23, 24]). Due to substantial class imbalance in the input dataset, we apply downsampling to the negative class:

$$N_{sampled} = \lfloor r \cdot N_- \rfloor, \tag{2}$$

where r is the downsampling rate. We additionally construct balanced batches:

$$B = \{(x_i^+, x_j^-) : x_i^+ \in X_+, x_j^- \in X_-\}, \tag{3}$$

with positive examples sampled with replacement and negative examples cycled without replacement. The model is trained using a class-weighted binary cross-entropy loss to address label imbalance, where w_+ and w_- are class-specific weights:

$$\mathcal{L} = -w_+ y \log p - w_-(1 - y) \log(1 - p), \quad \text{where} \quad p = \frac{\exp(z_+)}{\exp(z_+) + \exp(z_-)}. \tag{4}$$

We fix $w_+ = 1.0$ and tune w_- via hyperparameter search. After training, the optimal classification threshold t^* is selected to maximize validation F1 score:

$$t^* = \arg\max F_1(t). \tag{5}$$

This threshold is then used to binarize predicted probabilities into a prior knowledge matrix of regulatory (1) and non-regulatory (0) interactions for downstream GRN inference.

3 Experimental Results

3.1 GLM-Prior Constructs Generalizable and High-Quality Priors Across Species

We first benchmark GLM-Prior in *S. cerevisiae*, comparing it to the curated YEASTRACT prior [6] and two motif-based methods, Inferelator-Prior [25] and CellOracle's base GRN [26] (Figure 2A). For motif-based methods, priors are constructed by scanning for TF motifs within fixed windows around gene promoters. Each prior is compared to a[] gold standard of literature curated interactions [27]. GLM-Prior achieves an AUPRC of 0.40, outperforming YEASTRACT (0.33) by 21.2%, and far exceeding Inferelator-Prior $(0.03 - 0.02)$ and CellOracle's base GRN $(0.05 - 0.04)$, depending on window size. In addition to higher accuracy, GLM-Prior introduces 2,070 novel edges and reclassifies 989 existing edges from the YEASTRACT database, demonstrating its ability to refine and expand curated priors using sequence alone.

To assess generalization to complex systems, we train separate models in human and mouse using training labels from the STRING [19–22] and TRRUST [23, 24] databases. To demonstrate that GLM-Prior can predict cell-line specific priors when trained on general species-level interactions, we task the model with inferring TF-target gene edges specific to human and mouse embryonic stem cells (hESCs and mESCs), respectively (Figure 2B), using BEELINE [28] ChIP-seq interactions for evaluation. Despite training on general data, GLM-Prior accurately recovers cell line-specific regulatory structure, achieving AUPRCs of 0.24 in human and 0.22 in mouse. Further, a human-trained model transfers successfully to mouse (AUPRC $= 0.23$), outperforming a mouse-trained model (0.22), while mouse-to-human transfer achieves a lower AUPRC of 0.17, likely due to the smaller mouse training dataset (Figure 2C).

These findings demonstrate that GLM-Prior produces high-quality, sequence-derived priors that generalize across cell lines and closely related species, and can be robustly evaluated against independent gold standards even in the absence of cell-line-specific training data.

3.2 Prior Quality Determines GRN Inference Performance

We evaluate the effect of prior quality on GRN inference in *S. cerevisiae* by performing a cross-method evaluation that pairs three GRN inference models, PMF-GRN [29], the Inferelator 3.0 [25],

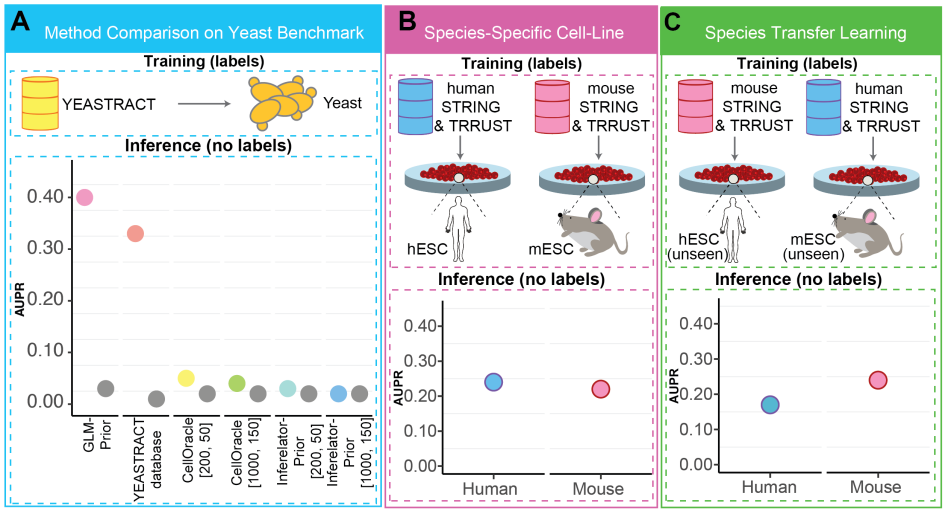

Figure 2: (**A**) AUPRC comparison of GLM-Prior, YEASTRACT, and motif-based priors in yeast. Grey dots represent shuffled controls. (**B**) GLM-Prior performance in hESC and mESC using species-specific models. (**C**) Cross-species transfer learning of GLM-Prior from human to mouse and vice versa.

and CellOracle [26], with each of their corresponding prior construction approaches, GLM-Prior, Inferelator-Prior, and CellOracle's base GRN. Although each method is typically designed with its own prior construction strategy, we systematically evaluate all combinations to disentangle the effects of prior construction and inference algorithm. All inferred GRNs are compared to a common gold standard [27] described in Section 3.1.

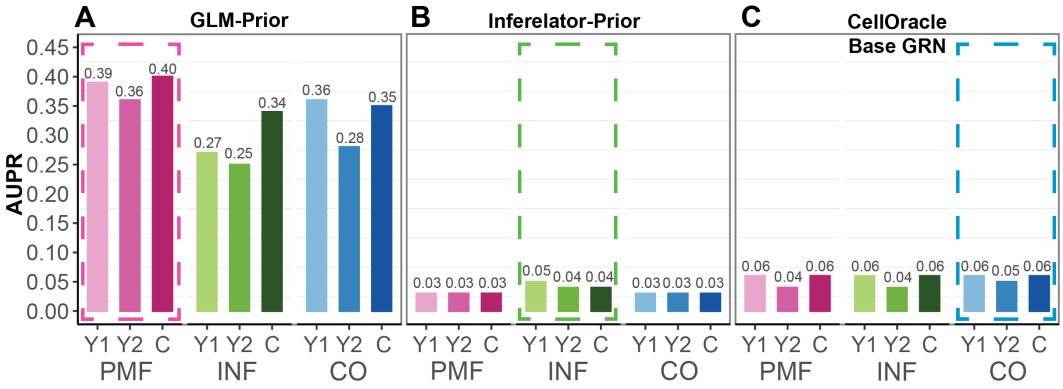

Figure 3: (**A**) GRN inference using GLM-Prior across three GRN models (PMF-GRN, Inferelator, CellOracle). (**B**) Performance using prior knowledge constructed by Inferelator-Prior. (**C**) Performance using prior knowledge constructed by CellOracle's base GRN.

Across yeast expression datasets Y1 (GSE125162 [30]), Y2 (GSE144820 [31]), and C (combined Y1 and Y2), GLM-Prior consistently enables the highest performance regardless of inference algorithm, with PMF-GRN and GLM-Prior achieving up to 0.40 AUPRC (Figure 3A). In contrast, motif-based priors (Inferelator-Prior (Figure 3B) and CellOracle's base GRN (Figure 3C)) performing poorly in all cases, with AUPRCs ≤ 0.06.

Importantly, adding expression-driven inference on top of GLM-Prior does not consistently improve performance and sometimes reduces it. This finding suggests that genomic language models pre-trained on nucleotide sequences, finetuned to predict TF-target gene interactions, capture a high quality scaffold of regulatory interactions. GRN inference with expression data then serves to modulate this scaffold to reflect the cell-type and condition-specific context captured by the experimental assay.

4 Conclusion and Future Work

GLM-Prior enables accurate and generalizable prior construction for GRN inference by leveraging large-scale genomic pretraining and finetuning on TF–gene sequence pairs. It outperforms motif-based and curated priors in yeast, and accurately reconstructs cell line–specific regulatory structure in human and mouse. GLM-Prior additionally supports zero-shot cross-species transfer, offering a scalable solution for investigating understudied organisms. Further, when combined with single-cell expression data for GRN inference in yeast, little to no performance improvement is observed, highlighting that performance in GRN inference is dominated by the quality of the prior, rather than the choice of inference algorithm. Future work will expand the application of genomic language models to incorporate additional regulatory modalities and contexts, enabling broader and more nuanced modeling of gene regulation.

Declarations

Acknowledgments

We thank the members of the Bonneau Lab and Cho Lab for insightful discussions and feedback on this manuscript. We also thank the staff of the NYU IT High Performance Computing and Flatiron Institute Scientific Computing Core. We are additionally grateful to Yanis Bahroun, Daniel Berenberg, Sarah Robinson, and Sabrina Mielke for insightful discussions related to this work.

Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research. This work was also supported by the Samsung Advanced Institute of Technology (under the project Next Generation Deep Learning: From Pattern Recognition to AI) and the National Science Foundation (under NSF Award 1922658). CSG is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2234660 and NSF Award 1922658. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Availability of Data and Materials

Code for GLM-Prior is available at `https://github.com/cskokgibbs/GLM-Prior`. Instructions on how to process nucleotide sequences for genes and TFs are available in the GLM-Prior GitHub repository, in the folder '`create_sequence_datasets`'. Code for previously published PMF-GRN is available at `https://github.com/nyu-dl/pmf-grn`. Expression datasets used in this manuscript have the following accessions: Yeast: GSE125162 (Y1) [30] and GSE144820 (Y2) [31]; Human embryonic stem cells (derived from BEELINE [28]) GSE75748 [32]; Mouse embryonic stem cells (derived from BEELINE [28]) GSE98664 [33]. The YEASTRACT prior-knowledge was derived from the YEASTRACT database [6]. The gold standard for the yeast datasets was obtained from [27]. The prior-knowledge matrices created from the BEELINE benchmarks [28] can be found at `https://zenodo.org/records/7682713`. Models for experiments from yeast, human and mouse can be found at `https://huggingface.co/cskokgibbs` as well as their corresponding tokenized datasets.

References

[1] Pau Badia-i Mompel, Lorna Wessels, Sophia Müller-Dott, Rémi Trimbour, Ricardo O Ramirez Flores, Ricard Argelaguet, and Julio Saez-Rodriguez. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, 24(11):739–754, 2023.

[2] Daniel Kim, Andy Tran, Hani Jieun Kim, Yingxin Lin, Jean Yee Hwa Yang, and Pengyi Yang. Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *NPJ Systems Biology and Applications*, 9(1):51, 2023.

[3] Marco Stock, Corinna Losert, Matteo Zambon, Niclas Popp, Gabriele Lubatti, Eva Hörmanseder, Matthias Heinig, and Antonio Scialdone. Leveraging prior knowledge to infer gene regulatory networks from single-cell rna-sequencing data. *Molecular Systems Biology*, pages 1–17, 2025.

[4] Sunnie Grace McCalla, Alireza Fotuhi Siahpirani, Jiaxin Li, Saptarshi Pyne, Matthew Stone, Viswesh Periyasamy, Junha Shin, and Sushmita Roy. Identifying strengths and weaknesses of methods for computational network inference from single-cell rna-seq data. *G3: Genes, Genomes, Genetics*, 13(3):jkad004, 2023.

[5] Natalie de Souza. The encode project. *Nature methods*, 9(11):1046–1046, 2012.

[6] Miguel C Teixeira, Pedro T Monteiro, Margarida Palma, Catarina Costa, Cláudia P Godinho, Pedro Pais, Mafalda Cavalheiro, Miguel Antunes, Alexandre Lemos, Tiago Pedreira, et al. Yeastract: an upgraded database for the analysis of transcription regulatory networks in saccharomyces cerevisiae. *Nucleic acids research*, 46(D1):D348–D353, 2018.

[7] Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194430, 2020.

[8] Sachi Inukai, Kian Hong Kock, and Martha L Bulyk. Transcription factor–dna binding: beyond binding site motifs. *Current opinion in genetics & development*, 43:110–119, 2017.

[9] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

[10] Jens Uwe Loers and Vanessa Vermeirssen. A single-cell multimodal view on gene regulatory network inference from transcriptomics and chromatin accessibility data. *Briefings in Bioinformatics*, 25(5):bbae382, 2024.

[11] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[12] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[13] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

[14] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, pages 1–13, 2025.

[15] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pages 1–11, 2024.

[16] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

[17] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

[18] Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158 (6):1431–1443, 2014.

[19] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

[20] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl_1):D561–D568, 2010.

[21] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[22] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

[23] Heonjong Han, Hongseok Shim, Donghyun Shin, Jung Eun Shim, Yunhee Ko, Junha Shin, Hanhae Kim, Ara Cho, Eiru Kim, Tak Lee, et al. Trrust: a reference database of human transcriptional regulatory interactions. *Scientific reports*, 5(1):11432, 2015.

[24] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.

[25] Claudia Skok Gibbs, Christopher A Jackson, Giuseppe-Antonio Saldi, Andreas Tjärnberg, Aashna Shah, Aaron Watters, Nicholas De Veaux, Konstantine Tchourine, Ren Yi, Tymor Hamamsy, et al. High-performance single-cell gene regulatory network inference at scale: the inferelator 3.0. *Bioinformatics*, 38(9):2519–2528, 2022.

[26] Kenji Kamimoto, Blerta Stringa, Christy M Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2023.

[27] Konstantine Tchourine, Christine Vogel, and Richard Bonneau. Condition-specific modeling of biophysical parameters advances inference of regulatory networks. *Cell reports*, 23(2):376–388, 2018.

[28] 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.

[29] Claudia Skok Gibbs, Omar Mahmood, Richard Bonneau, and Kyunghyun Cho. Pmf-grn: a variational inference approach to single-cell gene regulatory network inference using probabilistic matrix factorization. *Genome biology*, 25(1):88, 2024.

[30] Christopher A Jackson, Dayanne M Castro, Giuseppe-Antonio Saldi, Richard Bonneau, and David Gresham. Gene regulatory network reconstruction using single-cell rna sequencing of barcoded genotypes in diverse environments. *elife*, 9:e51254, 2020.

[31] Abbas Jariani, Lieselotte Vermeersch, Bram Cerulus, Gemma Perez-Samper, Karin Voordeckers, Thomas Van Brussel, Bernard Thienpont, Diether Lambrechts, and Kevin J Verstrepen. A new protocol for single-cell rna-seq reveals stochastic gene expression during lag phase in budding yeast. *elife*, 9:e55320, 2020.

[32] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17:1–20, 2016.

[33] Tetsutaro Hayashi, Haruka Ozaki, Yohei Sasagawa, Mana Umeda, Hiroki Danno, and Itoshi Nikaido. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature communications*, 9(1):619, 2018.

[34] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*, 2020.

[35] Jennifer J Van Oudenhove, Rodrigo A Grandy, Prachi N Ghule, Roxana Del Rio, Jane B Lian, Janet L Stein, Sayyed K Zaidi, and Gary S Stein. Lineage-specific early differentiation of human embryonic stem cells requires a g2 cell cycle pause. *Stem Cells*, 34(7):1765–1775, 2016.

[36] Tomoaki Hishida, Yuriko Nozaki, Yutaka Nakachi, Yosuke Mizuno, Yasushi Okazaki, Masatsugu Ema, Satoru Takahashi, Masazumi Nishimoto, and Akihiko Okuda. Indefinite self-renewal of escs through myc/max transcriptional complex-independent mechanisms. *Cell stem cell*, 9 (1):37–49, 2011.

Appendix

A Extended Methods

A.0.1 Training procedure

We trained the model on 4 H100 GPUs using PyTorch's Distributed Data Parallel (DDP) framework [34] to enable efficient multi-GPU scaling. The training process was distributed across GPUs to accelerate computation and ensure consistent gradient updates. We used a per-device batch size of 32 and set the gradient accumulation steps to 32, resulting in an effective batch size of 4096. The model was optimized using Adam with a learning rate of 10^{-5}. Training spanned 10 epochs, using optimal hyperparameters selected through a sweep over the negative class weight (w_{-}) and downsampling rate (see Appendix 4 for more details). We selected the configuration that achieved the highest F1 score on the validation set for final training.

After training, the model was used to infer a prior-knowledge matrix of TF-gene regulatory interactions from a list of gene-TF sequence pairs without labels. This matrix then served as input for downstream GRN inference, where GRN inference can further tailor these language model derived interactions using cell-type, cell-line, or condition-specific expression data.

A.0.2 Performance and Evaluation

We evaluated model performance using standard binary classification metrics, with a focus on metrics that remain robust under class imbalance. Specifically, we report precision, recall, F1 score, area under the receiver operating characteristic curve (AUC-ROC), area under the precision-recall curve (AUPRC), and Matthews correlation coefficient (MCC).

Precision and recall were computed separately for the positive and negative classes to assess the model's ability to minimize false positives and false negatives, respectively. Let TP, FP, FN and TN denote the true positives, false positives, false negatives and true negatives. Then,

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}. \tag{6}$$

The F1 score, which represents the harmonic mean of precision and recall, was used as the primary metric for model selection and hyperparameter optimization. It can be computed as,

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{7}$$

To account for class imbalance and provide a threshold-independent measure of performance, we also computed AUC-ROC and AUPRC. the ROC curve plots true positive rate (TPR) against false positive rate (FPR), defined as:

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN}, \quad \text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}. \tag{8}$$

While AUC-ROC captures the model's general discrimintative ability, AUPRC is more informative in imbalanced settings, as it directly reflects the trade-off between precision and recall. We used AUPRC to benchmark model predictions against curated gold standard datasets of TF-gene interactions.

To determine the optimal classification threshold, we performed a grid search over the predicted positive class probabilities. The threshold t^* that maximized the F1 score on the validation set was selected for final inference,

$$t^* = F_1(t). \tag{9}$$

Finally, we report the Matthews correlation coefficient (MCC), a balanced measure of classification quality that incorporates all four confusion matrix components,

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{10}$$

MCC ranges from -1 to $+1$, where $+1$ indicates perfect predictions, 0 indicates random predictions, and -1 indicates incorrect predictions. MCC remains informative even when classes are highly imbalanced, making it a useful complement to F1 and AUPRC.

A.1 GRN inference with PMF-GRN

We performed GRN inference using our previously published method, PMF-GRN (Probabilistic Matrix Factorization for Gene Regulatory Network inference) [29] to infer the regulatory interactions between TFs and their target genes. The goal of PMF-GRN is to decompose an observed gene expression matrix into latent factors that represent TF activity and regulatory interactions between TFs and their target genes. These latent factors capture the underlying GRN structure, which cannot be measured directly from gene expression data alone. Further details regarding the PMF-GRN model and the inference strategy used to obtain GRNs can be found in [25].

Using PMF-GRN, we perform inference independently on each single-cell dataset to obtain dataset-specific GRNs. These inferred networks are then combined post-inference using a simple averaging strategy to produce a consensus GRN,

$$\text{GRN}_{\text{Consensus}} = \frac{1}{N} \sum_{i=1}^{N} \text{GRN}_i, \tag{11}$$

where N is the number of datasets and GRN_i is the inferred network for dataset i. This consensus GRN captures regulatory interactions that are consistently inferred across datasets, while preserving dataset-specific networks for context-specific analyses.

B Extended Results

B.1 Prior quality determines the additive value of expression data in GRN inference for human and mouse embryonic stem cells

Having established that GLM-Prior effectively captures species and cell-type specific regulatory structure directly from sequence, we next test whether the limited additive value of expression-based inference observed in yeast extends to more complex systems, focusing on human and mouse embryonic stem cells (hESCs and mESCS). Specifically, we explore hESCs and mESCs, using GLM-Prior to construct a prior knowledge matrix from nucleotide sequences, followed by GRN inference with PMF-GRN on paired single-cell expression data.

In the human setting, GLM-Prior is trained on TF–gene interaction pairs using non-cell line specific regulatory edges from the STRING and TRRUST databases. To evaluate its ability to generalize to an unseen cell type, we task the model with predicting regulatory interactions in hESCs using nucleotide sequence pairs derived from genes and TFs in a held-out hESC ChIP-seq dataset. GLM-Prior achieves an AUPRC of 0.24 (Figure 4A), indicating that it successfully generalizes beyond curated databases and recovers context-specific regulatory structure from sequence alone.

To evaluate whether expression data offers additive value beyond this sequence-derived prior, we apply PMF-GRN to a single-cell RNA-seq time course of hESC differentiation [32], spanning six time points from pluripotency through early lineage commitment. Using the GLM-Prior as input, GRNs inferred across these timepoints yield only marginal performance gains, with AUPRCs ranging from 0.22 to 0.24 for individual time points (Appendix Table 4) and 0.27 when inferred on the pooled dataset (Figure 4A). This modest 12.5% improvement suggests that most of the regulatory signal recoverable from expression data was already captured by the prior.

To determine whether these predictions are nonetheless meaningful, we assess the calibration of PMF-GRN's uncertainty estimates. Specifically, we ask whether lower posterior variance, used by PMF-GRN to quantify uncertainty over each edge, correspond to more accurate predictions. We find that variance estimates are well-calibrated, with edges with lower posterior variance demonstrating higher precision (Figure 4B). This suggests that even when expression data adds little in terms of overall edge recovery, model uncertainty offers interpretability by flagging high-confidence, cell-type specific predictions.

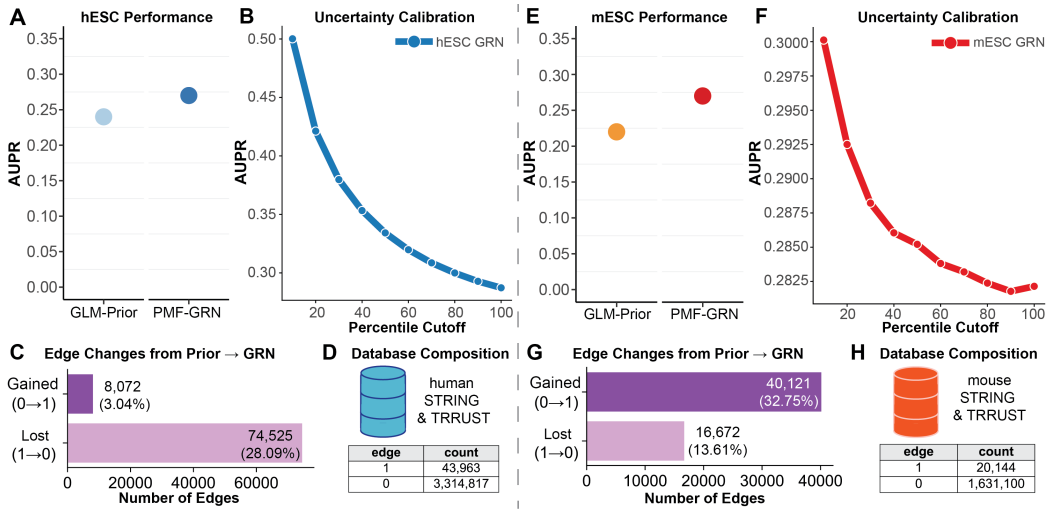

Figure 4: Performance analysis of GLM-Prior and PMF-GRN in human and mouse ESCs. (**A**) AUPRC scores in hESCs show modest gains from GRN inference over the sequence-derived prior. (**B**) Uncertainty calibration in hESCs confirms that lower posterior variance predicts higher precision. (**C**) GRN inference in hESCs prunes the prior, removing 74,525 edges while adding only 8,072. (**D**) Human training data includes 43,963 positive interactions, supporting strong prior construction. (**E**) AUPRC scores in mESCs show larger gains from GRN inference, suggesting greater value from expression. (**F**) Uncertainty calibration in mESCs remains well-aligned with predictive confidence. (**G**) GRN inference in mESCs expands the prior, adding 40,121 edges and removing 16,672. (**H**) Mouse training data includes only 20,144 positive interactions, explaining the weaker prior and increased reliance on expression.

To better understand the role of expression data in shaping the GRN, we compare the inferred network to the GLM-Prior and quantify structural changes. PMF-GRN removes $74,525$ edges (28.09% of all possible edges), while adding only $8,072$ new edges (3.04%) (Figure 4C). This gain-loss asymmetry confirms that expression data does not drive novel edge discovery, but instead acts as a selective filter by pruning edges not active in hESCs and preserving those supported by context-specific expression. Rather than constructing the GRN *de novo*, expression reweights a scaffold learned from sequence, tailoring it to the relevant regulatory state.

This filtering behavior reflects the strength of the prior knowledge, whereby GLM-Prior was trained on a large, well-annotated human regulatory dataset containing $43,963$ positive and 3.3 million negative TF-gene pairs from STRING and TRRUST (Figure 4D). This abundance of positive examples enables the model to learn generalizable regulatory principles directly from sequence. As a result, GRN inference does not need to discover new structure, but instead refine the scaffold to reflect the hESC context. These results support our broader conclusion that when a strong sequence-informed prior is available, expression data serves to modulate the network, not construct it.

We next apply GLM-Prior and PMF-GRN to mouse embryonic stem cells (mESCs) to assess whether the patterns observed in human generalize to other species. As in the human system, we began by evaluating GLM-Prior trained on TF-gene interaction pairs from the STRING and TRRUST databases. When tested against the held-out ChIP-seq reference labels, GLM-Prior achieves an AUPRC of 0.22 in mESCs (Figure 4E). GRN inference with PMF-GRN and paired single-cell expression data improves AUPRC to 0.27, a 22.7% performance increase over the prior. This exceeds the smaller gains observed in hESC and suggests that expression data plays a more substantial

role in recovering regulatory structure in the mouse setting. To determine whether this additive performance reflected meaningful model confidence, we assess the PMF-GRNs uncertainty estimate calibration. As in the human setting, posterior variance is well-calibrated, with edges with lower posterior variance demonstrating higher precision (Figure 4F). This confirms that PMF-GRN not only improves predictive accuracy in mESCs but also provides reliable confidence estimates, providing a layer of interpretability to the inferred GRN.

We then examine how expression data reshaped the GRN relative to the prior. In contrast to human, where expression primarily filters the prior, GRN inference in mouse introduces substantially more edges than it removes, $40,121$ new edges (32.75% of all possible edges) are added, while only $16,672$ (13.61%) are pruned (Figure 4G). This gain-loss asymmetry indicates that in mESCs, expression expands the network, recovering interactions not predicted from sequence alone. This expansion is consistent with the more limited training data used to construct the mouse GLM-Prior, which included $20,144$ positive TF-gene pairs compared to $43,963$ in human (Figure 4H). With less annotated regulatory information available, the sequence-derived prior in mouse left more room for expression to contribute additively.

Together, these results reinforce our central claim that the role of expression data in GRN inference depends on the quality and completeness of the prior. When the prior is strong, as in yeast and human, expression primarily tailors the scaffold to cell-type context, often by pruning unsupported edges. When the prior is weaker, as in mouse, expression plays a more constructive role by expanding the network to capture structure absent from sequence alone. Rather than serving as a general source of regulatory structure, expression is best understood as a context-aware modifier of a sequence-derived regulatory scaffold.

C Single-Species Experiment Details

C.1 Yeast data processing and experiments

To train our GLM-Prior model in yeast, we first obtained all $5,999$ gene body nucleotide sequences from the ENSEMBL *S. cerevisiae* (R64-1-1.UTR.gtf) genome. We obtained 212 TF sequences from the CisBP database [18] under *S. cerevisiae*.

Validation Metrics for Yeast Single-Species GLM-Prior

Metric	Score
Best F1 Score	0.64
ROC AUC	0.97
Best Classification Threshold	0.94
Positive Class Precision	0.88
Positive Class Recall	0.47
Negative Class Precision	0.99
Negative Class Recall	1.00
AUPRC (vs. gold standard [27])	0.40

Table 1: Validation performance of the GLM-Prior model trained on yeast. Evaluation was conducted on a held-out validation set that assess positive and negative class contributions during training (top) and AUPRC against an independent gold standard for the final inferred prior-knowledge matrix (bottom).

Next, to pair these gene and TF nucleotide sequences with interaction labels, we used the YEASTRACT database of interactions [6] ($6,885$ genes by 220 TFs). From these 220 YEASTRACT TFs, 46 did not have sequences associated with them from CisBP. Due to this large portion of data loss (20%), we used the promoter regions of the target genes for each missing TF as a proxy for it's binding sequence. We defined the promoter sequence following YEASTRACT's definition of 1000bp upstream or downstream of the gene TSS (depending on the strand orientation). This resulted in an input dataset containing $5,999$ genes by 254 TFs.

A hyperparameter sweep over 1 epoch of training using different class-weights and downsampling rates for the negative class revealed $[0.7, 1.0]$ to be the optimal class-weights and 0.4 to be the optimal

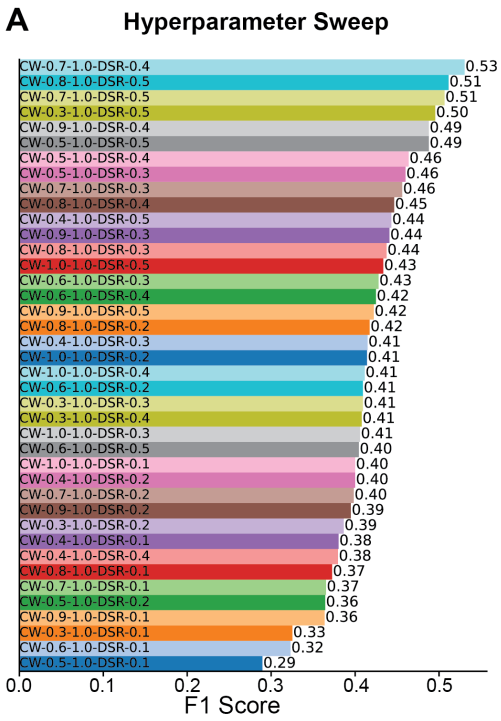

Figure 5: (**A**) A hyperparameter sweep performed on a 99% train and 1% validation split across class weights and downsampling rates reveals substantial variability in model performance across different configurations.

negative class downsampling rate (Figure 5A). These hyperparameters were used during final training over 10 epochs obtaining the validation metrics on the held-out 1% of training data found in Table 1.

Following training of the single-species yeast GLM-Prior model, we ran inference on all genes and TF sequences from the YEASTRACT database and evaluated our inferred prior-knowledge matrix using AUPRC on an independent gold standard [27].

Prior-knowledge datasets for CellOracle and Inferelator-Prior were obtained from [25] without further modification. These priors were constructed using motifs from CisBP, the same motifs used in GLM-Prior to ensure compatibility and fairness during evaluation. YEASTRACT derived prior-knowledge was downloaded from the database without further modification [6]. The reference gold standard was taken from [27] without further modification.

Single cell gene expression datasets were obtained from GSE125162 ($38,225$ cells by $6,763$ genes) [30] and GSE144820 ($6,118$ cells by $6,763$ genes) [31] without modification. The combined dataset was created by concatenating GSE125162 and GSE144820 on the cells axis ($44,343$ cells by $6,763$ genes). We performed GRN inference using PMF-GRN [29] and comparable methods such as the Inferelator 3.0 [25] and CellOracle [26] across datasets Y1, Y2, and the combined data (Table 2).

C.2 Human data processing and experiments

To train our GLM-Prior model in human, we first obtained hg38 gene body nucleotide sequences from ENSEMBL (RCh38.113.gtf). Due to the lengthy nature of human genes, and the inherent limitations of context length in large language models, we filtered our gene sequences to retain all sequences for a gene body $\leq 12,000$ nucleotides in length. This provided us with a list of $23,533$ genes. We obtained TF binding motif sequences from the CisBP database [18], under *H. sapiens*.

Next, to pair these gene and TF nucleotide sequences with interaction labels, we followed the BEELINE benchmarking protocol [28], obtaining labels from their human reference datasets from

Expression Input	Input Prior-Knowledge	PMF-GRN	Inferelator 3.0	CellOracle
GSE125162 (Y1)	GLM-Prior	0.39	0.27	0.36
	YEASTRACT	0.35	0.30	0.40
	Inferelator-Prior	0.03	0.05	0.03
	CellOracle Base GRN	0.06	0.06	0.06
	GLM-Prior + YEASTRACT	0.41	0.31	NA
GSE144820 (Y2)	GLM-Prior	0.36	0.25	0.28
	YEASTRACT	0.34	0.31	0.31
	Inferelator-Prior	0.03	0.04	0.03
	CellOracle Base GRN	0.04	0.04	0.05
	GLM-Prior + YEASTRACT	0.38	0.31	NA
Y1 + Y2 Combined	GLM-Prior	0.40	0.34	0.35
	YEASTRACT	0.36	0.39	0.41
	Inferelator-Prior	0.03	0.04	0.03
	CellOracle Base GRN	0.06	0.06	0.06
	GLM-Prior + YEASTRACT	0.42	0.40	NA

Table 2: Yeast GRN inference performance (AUPRC) across different prior-knowledge sources and methods (PMF-GRN, Inferelator 3.0, CellOracle), in these three yeast gene expression datasets: GSE125162 (Y1), GSE144820 (Y2), and their combined expression input, evaluated against an independent gold standard [27].

the STRING [19–22] and TRRUST [23, 24] databases. This provided us with $43,963$ positive interactions and $3,314,817$ negative interactions.

Due to the large number of negative interactions and the length of gene sequences, it was necessary to implement a $1:1$ downsampling strategy to prevent memory issues during training. For this reason, we trained GLM-Prior using all positive examples as well as a paired number of random samples of negative class examples. A hyperparameter sweep over the class weights for negative and positive classes revealed $[0.3, 1.0]$ to be the optimal class weights. These weights were used during final training and achieved the validation metrics found in Table 3.

Following training of the human single-species GLM-Prior model on STRING and TRRUST database labels, we ran inference using the genes and TFs associated with the BEELINE hESC reference ChIP-seq network. This provided us with $4,773$ genes and 79 TFs. After inference, the model predictions were evaluated using AUPRC with the corresponding labels from the hESC reference ChIP-seq network.

Validation Metrics for Human Single-Species GLM-Prior

Metric	Score
Best F1 Score	0.77
ROC AUC	0.84
Best Classification Threshold	0.67
Positive Class Precision	0.68
Positive Class Recall	0.90
Negative Class Precision	0.85
Negative Class Recall	0.59
AUPRC (vs. hESC reference network)	0.24

Table 3: Validation performance of the GLM-Prior model trained on human. Evaluation was conducted on a held-out set using validation metrics that consider the contribution of the positive and negative classes on performance (top) and AUPRC against an hESC ChIP-seq-derived reference network (bottom) after inference of the prior-knowledge matrix.

After prior-knowledge generation, GRN inference was run on six timepoint hESC single-cell expression datasets (GSE75748) [32], as similarly done in the BEELINE benchmarking framework. Single-cell expression datasets were separated by their respective timepoints for each timepoint GRN,

and combined post inference using our post-inference averaging strategy for our "Consensus" GRN. Additionally, we learned one GRN without any tasks, "No Tasks", to compare performance (Table 4).

AUPRC Results for hESC GRNs Compared to ChIP-seq Reference

GRN	Number of Cells	AUPRC
00h	92	0.22
12h	102	0.24
24h	66	0.24
36h	172	0.22
72h	138	0.23
96h	188	0.23
Consensus	758	0.24
No Tasks	758	0.27

Table 4: AUPRC performance for GRNs inferred at each developmental timepoint in hESCs, as well as a consensus model and a model trained without timepoint supervision ("No Tasks"). Evaluation is based on overlap with an hESC ChIP-seq-derived reference network.

C.3 Analysis of timepoint GRNs and transcription factor activity in hESCs

To complement the main evaluation of the dual-stage training pipeline in hESCs (Section B.1), we present supplementary analyses that examine the dynamic regulatory structure inferred by PMF-GRN. These include both timepoint-specific GRNs and transcription factor activity (TFA) trajectories, and compact GRN visualizations centered on stage-specific marker genes. While not central to our core methodological evaluation, these analyses illustrate the biological interpretability enabled by the posterior distributions produced during GRN inference.

We begin by evaluating the sequence-informed prior generated by GLM-Prior. The model was trained on human TF-gene nucleotide sequence pairs using curated interaction labels from the STRING and TRRUST databases. As shown in Appendix Figure 6A, GLM-Prior achieves strong validation metrics, including a positive-class recall of 0.90, negative-class precision of 0.85, ROC-AUC of 0.84, and an F1 score of 0.77 using an optimal classification threshold of 0.67. These results demonstrate that GLM-Prior effectively captures regulatory sequence logic from external reference datasets.

In addition to the pooled (no-task) GRN analysis (described in Section Prior quality determines the additive value of expression data in GRN inference for human and mouse embryonic stem cells), we evaluated the performance of PMF-GRN when inferring separate, timepoint-specific GRNs across the hESC differentiation time course. This task-specific decomposition allowed us to assess whether modeling regulatory programs at finer temporal resolution provides any performance advantage over pooled modeling. Across six timepoints (00h, 12h, 24h, 36h, 72h, 96h), PMF-GRN achieved consistent AUPRCs ranging from 0.22 to 0.24 when evaluated against the hESC ChIP-seq reference network (Appendix Table 4). The highest performance was observed at 12h and 24h (AUPRC = 0.24), which may correspond to early differentiation events as cell begin to exit the pluripotent state. However, these timepoint-specific models did not outperform the pooled no-task GRN (AUPRC = 0.27), suggesting that the increased statistical power of pooled inference outweighs the potential benefits of task-specific decomposition when cell observations are limited.

In parallel with recovering regulatory edges, PMF-GRN infers latent TFA for each cell, offering a powerful lens to explore the dynamic regulatory landscape underlying cell state transitions. To investigate how TFA evolves over time, we visualized the inferred TFA matrix using UMAP, projecting cells into a low-dimension space based on their TF activity profiles (Appendix Figure 6B). The resulting UMAP embeddings recapitulate the developmental progression of the hESC time course, where 00h cells form a distinct cluster from which two trajectories emerge: one leading through 12h and 24h, and the other progressing through 36h, 72h and 96h. This bifurcation suggests early lineage commitment events by 24h, consistent with exit from pluripotency and the onset of differentiation [35].

To further explore the biological relevance of these TFA profiles, we examined the activity dynamics of individual TFs. We identified MAX as the most highly active TF across the full dataset and plotted

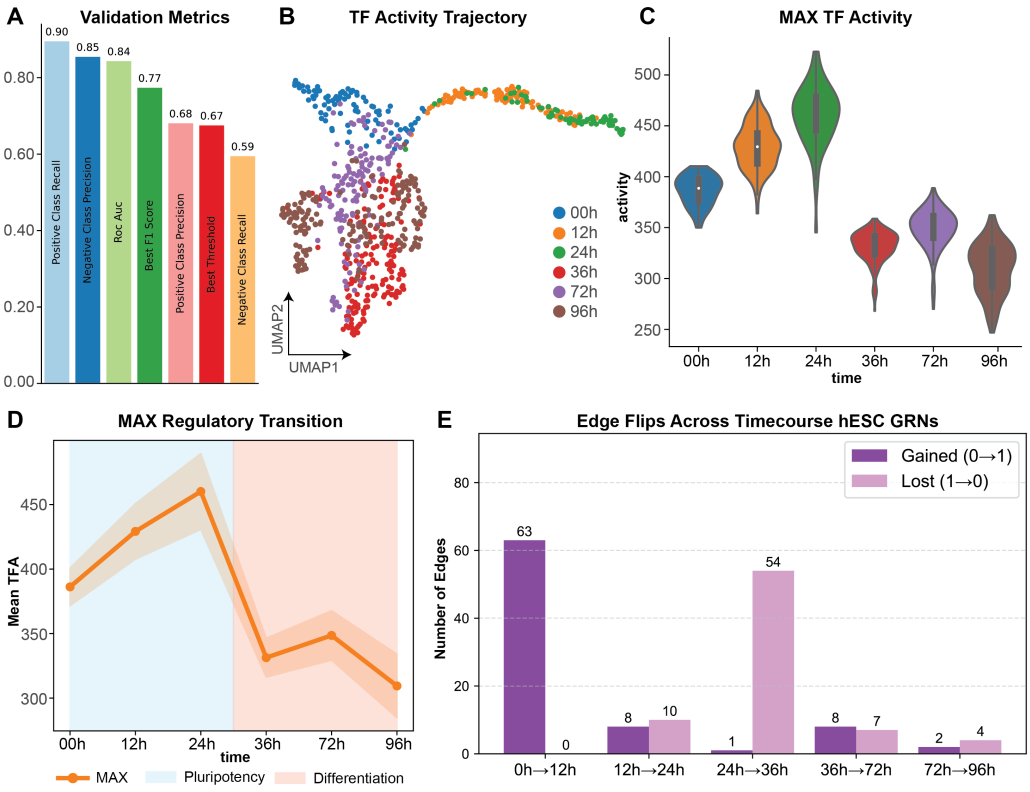

Figure 6: Performance and interpretability of the dual-stage GLM-Prior and PMF-GRN pipeline applied to human ESCs. (**A**) Validation metrics for GLM-Prior trained on human TF-gene interactions from STRING and TRRUST. (**B**) UMAP projection of TFA reveals temporal structure and bifurcating developmental trajectories. (**C**) Violin plot of MAX TFA across time points, showing increased activity from 00h to 36h followed by a sharp drop at 36h. (**D**) Mean TFA of MAX across the time course highlights a transition point between 24h and 36h, suggesting a role for MAX in pluripotency and an activity decline associated with differentiation onset. (**E** Number of edge flips between time-resolved GRNs.)

its activity distribution over time (Appendix Figure 6C). MAX demonstrated a rise in activity from 00h to 24h, followed by a sharp drop at 36h, and more moderate changes at 72h and 96h. To visualize this temporal trend, we plotted the mean TFA of MAX across time points, revealing a clear inflection point between 24h and 36h that aligns with the transition from the pluripotent state to early lineage specification (Appendix Figure 6D).

This pattern is particularly intriguing given that MAX is known to partner with other transcriptional regulators such as MYC and plays a key role in controlling proliferation, chromatin accessibility, and pluripotency [36]. The observed role in MAX activity during the early stages of the time course likely reflects its involvement in maintaining or priming the pluripotent state, while the drop post-24h may signify the onset of lineage-specific roles as cells diverge in fate. This regulatory transition is evident solely from TFA profiles, further underscoring the interpretability and resolution provided by the dual-stage training pipeline.

To further investigate the temporal dynamics captured by these timepoint-specific GRNs, we analyzed edge turnover between consecutive GRNs (Figure 6E). This revealed a punctuated pattern of regulatory modeling across the differentiation trajectory. From 00h to 12h, the network expanded sharply with 63 new edges and no losses, indicating broad activation of regulatory programs as cells exit pluripotency. Subsequent transitions showed increased pruning, with 12h to 24h involving moderate edge refinement (8 edge gains and 10 losses), while 24h to 36h featured substantial contraction (1 gain and 54 losses), consistent with a shift toward lineage commitment. From 36h onward, the network stablized, showing more balanced turnover (8 gains and 7 losses from 36h to 72h, and only

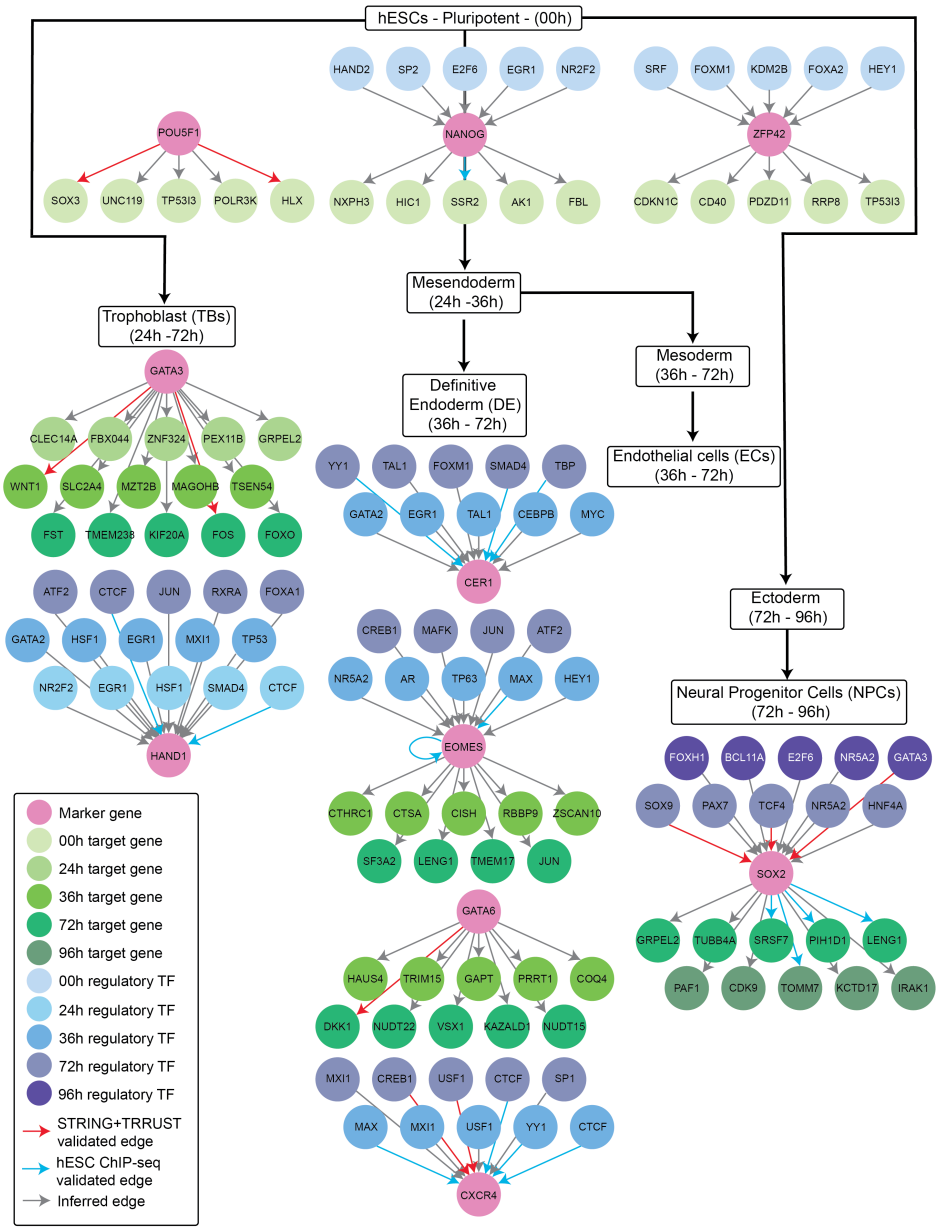

Figure 7: Marker gene GRN visualizations revealing stage-specific regulatory programs across hESC differentiation. Compact GRNs highlight top five regulators (TFs) (blue) and/or targets (genes) (green) of different marker genes (pink) across developmental trajectories in hESCs. Color gradient illustrates increasing time from light to dark (00h to 96h). Inferred edges are grey, edges supported by STRING and TRRUST are red, while edges supported by the reference hESC ChIP-seq network are light blue.

2 gains and 4 losses from 72h to 96h). These patterns reflect dynamic GRN reprogramming over time, with an early wave of activation followed by selective refinement and eventual stabilization, consistent with expected transcriptional transitions during stem cell differentiation.

To further contextualize the inferred GRNs, we turned to the original study from which the hESC single-cell dataset was derived [32]. In this work, the authors defined a curated set of marker genes representative of distinct stages of pluripotency and lineage commitment. These markers were shown to exhibit temporally appropriate expression patterns, making them ideal anchor points for investigating stage-specific regulatory dynamics. Inspired by this, we constructed compact GRN visualizations centered on these individual marker genes. For each selected marker, we extracted the top five predicted regulators (TFs) and/or top five predicted targets from the timepoints that aligned with the marker. For each predicted edge, we highlighted whether this edge had support from the STRING or TRRUST database, or the corresponding hESC ChIP-seq reference network from BEELINE. This allowed us to create interpretable sub-networks aligned with known developmental transitions (Appendix Figure 7).

This marker-centered GRN analysis provides a framework to trace the dynamic regulatory roles of key genes throughout hESCs progression. The selected marker genes highlight transitions from the pluripotent state through bifurcating trajectories into trophoblast (TBs), mesendoderm, or ectoderm lineages, and further into specialized fates such as definitive endoderm (ED), mesoderm, endothelial cells (ECs), or neural progenitor cells (NPCs). By anchoring our GRNs to these well-established marker genes, we were able to map not only the timing of regulatory transitions but also the identities of candidate TFs driving them.

Together, these analyses highlight the power of GLM-Prior and PMF-GRN to recover biologically meaningful, temporally resolved GRNs and TF activity dynamics from single-cell data. In the context of human ESCs, this enables a systems-level view of how regulatory programs are rewired over developmental time, offering mechanistic insight into critical transitions such as pluripotency exit and early lineage commitment.

C.4 Mouse data processing and experiments

To train our GLM-Prior model in mouse, we first obtained the mm10 gene body nucleotide sequences from ENSEMBL (GRCm39.113.gtf). Similarly to the single-species human experiments, we again filtered the length of our mouse genes to retain all sequences for a gene body $\leq 12,000$ nucleotides in length. This provided us with a list of $37,755$ genes. We obtained TF binding motif sequences from the CisBP database [18], under *M. musculus*.

To pair gene and TF nucleotide sequences with interaction labels, we followed the BEELINE benchmarking protocol [28], using mouse reference datasets from STRING [19–22] and TRRUST [23, 24] databases. This resulted in a training set comprising $5,326$ genes, 491 TFs, $27,909$ positive examples, and $2,587,157$ negative examples.

As in the human setup, we applied a $1:1$ downsampling strategy to balance positive and negative classes and manage memory constraints. A hyperparameter sweep over class weights yielded an optimal configuration of $[0.8, 1.0]$ for negative and positive classes, respectively. These weights were used during final training, achieving the validation metrics shown in Table 5.

Following training, we used GLM-Prior to generate predictions for the $5,711$ genes and 59 TFs present in the BEELINE mESC ChIP-seq reference network. These predictions were evaluated using AUPRC against the same ChIP-seq derived labels, providing a performance estimate of the learned prior-knowledge matrix.

Using the predicted prior matrix, we performed GRN inference on five timepoint-specific mESC single-cell expression datasets from GSE98664 [33], following the BEELINE framework. Each GRN was inferred using cells from a single timepoint and subsequently aggregated using our post-inference averaging strategy to form a "Consensus" GRN. Additionally, a pooled "No Tasks" GRN was learned from all cells without timepoint supervision. AUPRC results for each inferred GRN are shown in Table 6.

Validation Metrics for Mouse Single-Species GLM-Prior

Metric	Score
Best F1 Score	0.73
ROC AUC	0.76
Best Classification Threshold	0.47
Positive Class Precision	0.61
Positive Class Recall	0.89
Negative Class Precision	0.78
Negative Class Recall	0.41
AUPRC (vs. mESC reference network)	0.22

Table 5: Validation performance of the GLM-Prior model trained on mouse. Metrics are computed on a held-out validation set, as well as against an mESC ChIP-seq-derived reference network (bottom row) to compute AUPRC of the final predicted prior-knowledge matrix.

AUPRC Results for mESC GRNs Compared to ChIP-seq Reference

GRN	Number of Cells	AUPRC
00h	90	0.27
12h	68	0.27
24h	90	0.27
48h	82	0.27
72h	91	0.27
Consensus	421	0.27
No Tasks	421	0.27

Table 6: Area under the precision-recall curve (AUPRC) for GRNs inferred at each developmental timepoint in mESCs, as well as a consensus model and a model trained without timepoint supervision ("No Tasks"). Evaluation is based on overlap with a mESC ChIP-seq-derived reference network.

C.5 Single-species training runtime and batch statistics

Training batch composition statistics for each species-specific GLM-Prior model are summarized in Appendix Figure 8A. All models were trained for 10 epochs using distributed data parallel (DDP) across 4 GPUs. Training time varied substantially across species, with yeast requiring approximately 80 hours, human approximately 16 hours, and mouse approximately 3 hours. These difference reflect both dataset size and batching constrained imposed by sequence length and class balance.

Unlike the human and mouse datasets, which used a $1:1$ ratio of positive to negative examples to avoid out-of-memory (OOM) issues caused by long gene context lengths, the yeast dataset used a 0.4 downsampling rate for negative examples. This was feasible due to the shorter averaage length of yeast gene sequences, allowing for deeper batching and more efficient memory use. As a result, yeast training involved significantly more batches per epoch, contributing to longer overall runtime.

A

	Yeast	Human	Mouse
Positives	68685	416222	67369
Negatives	166489	416222	67369
Batches/Epoch	2602	6504	1053
Total Batches	26020	65040	10530
Total Pos Seen	1664890	4162220	673690
Total Neg Seen	1664890	4162220	673690
Pos Reuse Ratio	24.2×	10×	10×

Single-Species Training Over 10 Epochs

Figure 8: Training batch statistics across datasets. (**A**) Summary table of training batch statistics, including the number of positive and negative examples, batches per epoch, and total batches.

Appendix Figure 8A details per-epoch batch statistics, including the number of positive and negative examples, batches per epoch, and total batches across 10 epochs. These illustrate the impact of sequence length and sampling strategy on data reuse and training efficiency across species.

D Transfer Learning Experiment Details

D.1 Transferring knowledge from human to mouse

To transfer knowledge from human to mouse, we took the single-species trained human model described in Section C.2 and used it to run inference on the $5,711$ genes and 59 TFs in mESC. We evaluate the predictions using AUPRC on the reference labels from the mESC ChIP-seq experiment from BEELINE.

D.2 Transferring knowledge from mouse to human

To transfer knowledge from human to mouse, we took the single-species trained mouse model described in Section C.4 and used it to run inference on the $4,773$ genes and 79 TFs in hESC. We evaluate the predictions using AUPRC on the reference labels from the hESC ChIP-seq experiment from BEELINE.

D.3 Transferring knowledge from human and mouse to yeast

To transfer knowledge from human to mouse, we took the single-species trained human model described in Section C.2, and further finetuned this model using the mouse gene (n=$5,326$) and TF (n=491) nucleotide sequences associated with the STRING and TRRUST databases, and the optimal class weights obtained during hyperparameter search in the mouse single-species model (Section C.4). After consecutively training our model on human and then mouse, we run inference on the $5,999$ genes and 254 TFs from YEASTRACT. We evaluate the predictions using AUPRC on the independent gold standard [27], as done in the yeast single-species model.

Results for the transfer learning experiments can be found in Table 7.

AUPRC Results from Cross-Species Transfer Learning

Training Dataset	Inference Dataset	AUPRC
Human	Mouse	0.23
Mouse	Human	0.17
Human and Mouse	Yeast	0.02

Table 7: AUPRC scores for cross-species transfer learning. Each row indicates the species used to train the GLM-Prior model and the target species on which GRN inference was performed. Evaluation was conducted against species-specific ChIP-seq or curated reference datasets.