GLM-Prior: a nucleotide transformer model reveals prior knowledge as the driver of GRN inference performance

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

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

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

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

Abstract

Gene regulatory network inference depends on high-quality prior-knowledge, yet curated priors are often incomplete or unavailable across species and cell types. We present GLM-Prior, a genomic language model fine-tuned to predict transcription factor to target gene interactions directly from nucleotide sequence. We integrate GLM-Prior with PMF-GRN, a probabilistic matrix factorization model, to create a dual-stage training pipeline that combines sequence-derived priors with single-cell gene expression data for accurate and interpretable GRN inference. In yeast, we find that GLM-Prior outperforms motif-based and curated priors, and that downstream gene expression based inference adds little additional value, indicating that most regulatory signal is captured directly from sequence. We show that, across yeast, mouse, and human, GLM-Prior consistently improves GRN inference accuracy and generalizes across species, enabling zero-shot transfer between human and mouse without retraining. These results demonstrate that prior-knowledge construction, rather than inference algorithm complexity, is the current bottleneck in GRN modeling, and that future methods should treat expression data as a contextual modulator of a fixed regulatory scaffold, rather than a primary source of structure discovery.

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). Constructing accurate GRNs requires integrating large-scale sequencing datasets, such as single-cell RNA-seq and ATAC-seq, to infer regulatory interactions that cannot be measured experimentally (3). However, the accuracy of GRN inference depends heavily on the quality of the prior knowledge used to guide the inference algorithm (4; 5). In the context of GRNs, prior knowledge refers to an initial matrix of putative TF-gene regulatory interactions, which provides structural guidance to the inference algorithm by defining which TF-gene pairs are more likely to represent true regulatory edges. Recent methods have demonstrated that better prior knowledge, which captures biologically meaningful regulatory signals, leads to more accurate and informative GRN reconstruction, highlighting the critical role of prior knowledge in improving GRN inference quality (6; 7; 8; 9).

Prior knowledge for well-studied species is typically constructed using databases of experimentally validated TF-gene interactions (10; 11). For less-characterized organisms, prior knowledge is inferred by combining

*Corresponding author.

structural genomic data with sequencing information (12). The Inferelator 3.0's Inferelator-Prior performs TF motif enrichment within open chromatin regions to define edges between proximally bound TFs to target genes (6). CellOracle's base GRN draws edges between TFs bound within promoter regions of accessible target genes (8). SCENIC's CisTarget constructs prior knowledge by ranking enriched motifs in accessible regions using genome-wide motif scores and a hidden Markov model to predict TF-gene interactions (9). While these methods demonstrate the importance of prior knowledge in GRN inference, they also expose key limitations. Motif annotation remains incomplete (13), chromatin accessibility data is noisy and cell-type specific (14), and existing methods often fail to capture long-range regulatory interactions (15). These limitations underscore the need for a more generalizable approach to constructing high-quality prior knowledge matrices, particularly for poorly characterized organisms (4; 16).

A promising direction is to incorporate transformer-based genomic foundation models which learn regulatory logic directly from DNA sequence. Architectures such as the Nucleotide Transformer (17) have demonstrated strong performance across genomic tasks, including predicting chromatin accessibility, gene expression and transcription factor binding (18; 19; 20; 21). By learning conserved regulatory patterns directly from sequence data, transformer-based genomic models offer a compelling solution to the limitations of motif-based prior knowledge approaches (22). By leveraging attention mechanisms, they effectively capture long-range dependencies in DNA sequences (23), enabling more accurate modeling of complex regulatory interactions. When trained on large, multi-species genomic datasets, these models encode both species-specific and cross-species features, supporting generalization across cell types (18; 24) and organisms (17; 25). This opens the door to a powerful transfer learning paradigm: models trained on species with extensive regulatory data (e.g., *H. Sapiens*) can be adapted to predict interactions in related species (e.g., *M. musculus*), when curated priors are incomplete or unavailable. In this way, genomic language models offer the potential to expand prior knowledge construction to organisms and cell types for which no direct interaction data exist, enabling GRN inference in new and challenging contexts.

In this work, we present GLM-Prior, a novel approach for constructing high-quality prior knowledge to improve downstream GRN inference. GLM-Prior is built by fine-tuning the 250 million parameter Nucleotide Transformer (17), pretrained on the genomes of 850 species, into a sequence classification model for predicting TF-gene regulatory interactions. Unlike motif-based approaches that rely on the proximity of TF motifs to accessible promoter regions, GLM-Prior directly predicts interactions from paired nucleotide sequences of TF binding motifs and gene bodies. These sequences are jointly encoded by the transformer and passed through a classification head, which outputs a probability of regulatory interaction for each TF-gene pair. To construct a binary prior knowledge matrix suitable for downstream GRN inference, we apply an optimized classification threshold that converts these probabilities into discrete binary interactions. As most TF-gene pairs are non-regulatory, the model is trained using class-weighted loss and balanced sampling to mitigate class imbalance during training. Together with the Transformer's attention mechanisms, which capture long-range dependencies and complex regulatory grammar, these features enable GLM-Prior to generalize across cell types and species more effectively than motif-based approaches.

To evaluate the utility of GLM-Prior, we incorporate it into a dual-stage GRN inference pipeline. In the first stage, GLM-Prior generates a prior knowledge matrix of TF-gene interactions from nucleotide sequence. In the second stage, this matrix is used to constrain PMF-GRN (7), a probabilistic matrix factorization model that infers GRNs from single-cell gene expression data. This framework allows us to isolate and compare the contributions of prior construction and expression-based inference to overall GRN reconstruction performance.

Across experiments in *Saccharomyces cerevisiae*, *Mus musculus*, and *Homo sapiens*, we find that GLM-Prior substantially outperforms traditional motif-based priors and that downstream GRN inference yields marginal improvements over the prior alone. This saturation effect is strongest in yeast, where GRN inference adds virtually no predictive value beyond the prior, and remains pronounced in human and mouse, where gains are similarly modest. These findings suggest that GLM-Prior captures the majority of regulatory information recoverable from gene expression data, reframing the longstanding assumption that expression-based inference is the primary driver of GRN accuracy.

We investigate the generalizability of our model by evaluating three training paradigms: (1) single-species models trained and evaluated on the same organisms; (2) transfer learning, in which models trained in one species are used to predict regulatory interactions in an unseen species; and (3) multi-species training, where a single model is fine-tuned on data from multiple organisms. We find that GLM-Prior generalizes effectively between evolutionarily related species, where a model trained exclusively on human data achieves strong performance in mouse, even surpassing a mouse-specific model. In contrast, a model trained on both human and mouse data exhibits reduced performance when transferred to yeast, likely reflecting the greater evolutionary distance and divergence in regulatory grammar and TF binding motifs. While multi-species training preserves reasonable accuracy across species, we observe trade-offs in human performance, suggesting that species-specific features may be diluted in a shared representation.

These findings prompt us to reconsider how expression data should be used in GRN inference. While single-cell expression provides valuable insight into regulatory activity within a specific cellular context, it may not reliably recover the full regulatory architecture encoded in sequence, especially when technical noise, dropout rates, or biological heterogeneity obscure true interactions. Our findings suggest that expression data may serve more effectively as a means to reweigh or annotate a fixed, high-quality prior that highlights which edges are active in a given condition, rather than as a primary source of network structure discovery. In this view, GLM-Prior establishes a stable, generalizable scaffold of regulatory interactions, while expression data dynamically modulates this scaffold across cell types and states. This reframing opens the door to new inference paradigms that treat expression not as signal for edge discovery, but as context for interpreting a learned regulatory landscape.

Results

Dual-stage training pipeline overview

We develop a dual-stage training pipeline to improve the accuracy and interpretability of GRN inference by integrating sequence-derived regulatory interactions with probabilistic modeling of single-cell expression data (Figure 1). The pipeline is designed to decouple the construction of prior knowledge from the expression-driven GRN inference step, allowing each component to be independently optimized and evaluated. In this framework, the first stage generates a biologically informed prior knowledge matrix by using a fine-tuned genomic language model (GLM-Prior), while the second stage uses this matrix as structural input to the probabilistic matrix factorization model (PMF-GRN (7)) to a infer GRN.

Stage 1: Constructing sequence-informed prior knowledge with GLM-Prior

In the first stage of the pipeline, we construct a prior knowledge edge matrix representing predicted regulatory (1) and non-regulatory (0) interactions between TFs and their target genes (Figure 1A). To generate these predictions, we develop GLM-Prior, a transformer-based sequence classification model built by fine-tuning the 250 million parameter Nucleotide Transformer (17) model, which was pretrained on the genomes of 850 different species. The primary goal of GLM-Prior is to leverage sequence information to accurately predict TF-gene regulatory interactions, in order to construct an informative prior knowledge matrix.

GLM-Prior takes as input paired nucleotide sequences consisting of TF binding motifs and gene body regions, along with labels representing known regulatory (1) or non-regulatory (0) interactions. These sequences are concatenated and encoded into a shared representation using the transformer's embedding and encoding layers. This representation is passed through a classification layer that outputs the probability of a true regulatory interaction. During training, the model evaluates a range of classification thresholds and selects the threshold that maximizes the F1 score on the validation set. This optimal threshold is used to binarize the predicted probabilities into regulatory (1) or non-regulatory (0) interactions for the final prior knowledge matrix.

Since GRN datasets are often highly imbalanced, with far fewer positive (regulatory) interactions compared to negative (non-regulatory) interactions, we apply several strategies to mitigate this imbalance during training. Specifically, we apply a downsampling rate to the negative class and use a class-weighted cross-entropy loss, tuning the negative class weight through hyperparameter search. Additionally, to further ensure stable

training, we implement a custom DataLoader to create training batches with equal numbers of positive and negative examples, ensuring balanced gradient updates.

After training, GLM-Prior performs inference across all provided TF-gene pairs, generating a dense probability prior knowledge matrix. The predicted probabilities are then binarized using the optimal F1 score-derived classification threshold, generating a binary prior knowledge matrix that encodes regulatory interactions for a given organism. This sequence-based prior knowledge is then evaluated on an independent gold standard and can be passed directly into the second stage of the pipeline, enabling a modular information transfer from model-driven sequence inference to data-driven GRN inference.

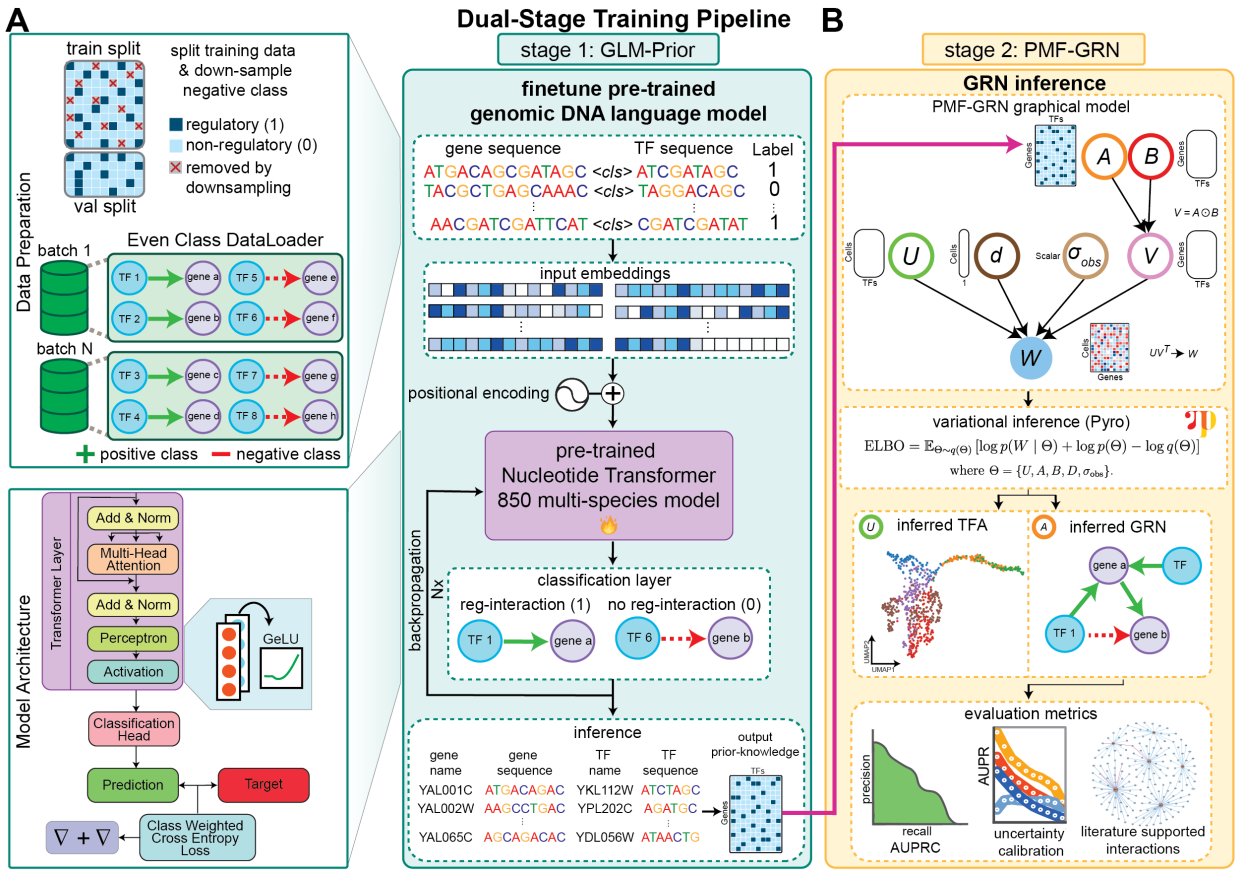

Figure 1: Schematic of the dual-stage training pipeline. (**A**) GLM-Prior is trained on concatenated TF binding motifs and gene sequences labeled as interacting or non-interacting. Input data is split into training and validation sets, where the training set is downsampled and placed into even class batches using a custom DataLoader. These input sequences are embedded and encoded by the transformer architecture, and passed through a classification head to predict the probability of a regulatory interaction. After training, inference is performed on all TF-gene pairs to generate a prior knowledge matrix. (**B**) Inferred prior knowledge is passed to PMF-GRN to provide a structural constraint to the probabilistic graphical model during inference. PMF-GRN infers transcription factor activity and a GRN for the input cell-line dataset. The resulting GRN is then evaluated using AUPRC and uncertainty calibration with an independent gold standard.

Stage 2: Refining GRN inference from single-cell data with PMF-GRN

In the second stage of the pipeline, the prior knowledge matrix predicted by GLM-Prior serves as a structured input to our previously developed model, PMF-GRN (7) (Figure 1B). PMF-GRN leverages this sequence-informed prior to enhance the inference of GRNs from single-cell gene expression data. It does so by using probabilistic matrix factorization, which decomposes the observed expression into latent factors representing

transcription factor activity and regulatory connectivity. Incorporating the prior knowledge matrix as a structural constraint biases the factorization toward biologically plausible regulatory relationships, thereby grounding the inferred GRN in sequence-based evidence.

To further tailor the inferred GRN to specific cellular or condition-specific contexts, PMF-GRN integrates single-cell expression data from distinct cell types or conditions. This integration yields a context-specific GRN, reflecting both the regulatory patterns suggested by GLM-Prior and the context provided by the single-cell data. Additionally, PMF-GRN not only produces point estimates for each TF-gene interaction, but also infers posterior variance, quantifying the model's uncertainty regarding each predicted interaction. These uncertainty estimates can be used for interpretation when gold standards are incomplete or unavailable.

PMF-GRN's flexibility allows it to incorporate multiple prior knowledge matrices and gene expression datasets into a unified GRN. This is achieved by first constructing independent GRNs for each combination of prior knowledge and expression data. These intermediate GRNs are then averaged after inference, resulting in an integrated network that captures both shared regulatory features and condition-specific variations.

This dual-stage training pipeline combines the predictive accuracy of large-scale sequence models with the interpretability and flexibility of probabilistic modeling. As a result, it produces GRNs that effectively integrate diverse biological information, leading to a more robust and context-aware representation of gene regulation.

GLM-Prior outperforms motif-based priors by learning regulatory logic directly from sequence in yeast

To evaluate the performance of our dual-stage training pipeline, we benchmarked GLM-Prior in yeast (*S. cerevisiae*), a model organism with extensively validated transcriptional regulatory networks. Yeast serves as a standard benchmark for GRN inference due to the availability of high-quality single-cell expression datasets and experimentally validated TF-gene interactions.

We evaluate the quality of prior knowledge generated by GLM-Prior by training the model on TF and gene sequences from yeast, using interaction labels curated from the YEASTRACT database (11). To address the substantial class imbalance between regulatory (positive) and non-regulatory (negative) interactions, we implement a set of custom features to our model. First, we split our input dataset into a 99% training and 1% validation set. We then apply a downsampling strategy to reduce the number of negative examples while maintaining their diversity, as defined in Equation 1. Next, to ensure balanced learning signals during training, we implement a custom DataLoader (Equation 2) that generates batches with equal numbers of positive and negative examples. Each input consists of a paired TF binding motif and gene sequence, processed into a shared representation by the transformer's embedding and encoding layers. The output is then passed through a classification layer that estimates the probability of a regulatory interaction using a class-weighted cross-entropy loss function (Equation 6).

We observe substantial variability in model performance across different hyperparameter configurations, underscoring the importance of careful tuning. To identify the optimal settings, we perform a hyperparameter sweep over class weights and downsampling rates (Figure 2A). We conduct the sweep over one training epoch and evaluate on the held-out validation set. A histogram of F1 scores across runs reveals the best-performing configuration uses class weights $[0.7, 1.0]$, for the negative and positive classes in the loss function, respectively, and a negative class downsampling rate of 0.4 on the input training dataset. These hyperparameters are then selected for full model training.

With these optimal hyperparameters, we then train GLM-Prior for 10 full epochs. The final model demonstrates strong predictive performance across key classification metrics (Figure 2B). GLM-Prior achieves a negative class recall of 1.0, a negative class precision of 0.99, and a receiver operating characteristic area under the curve (ROC-AUC) of 0.97. The optimal classification threshold is identified at 0.94, where predictions ≥ 0.94 are considered regulatory (1), and < 0.94 are considered non-regulatory (0). This threshold yields a positive class precision of 0.88, a positive class recall of 0.44, and an F1-score of 0.64. These results indicate that GLM-Prior effectively distinguishes true regulatory interactions, achieving high specificity while retaining precision for true positives, an essential property for constructing a reliable prior knowledge matrix.

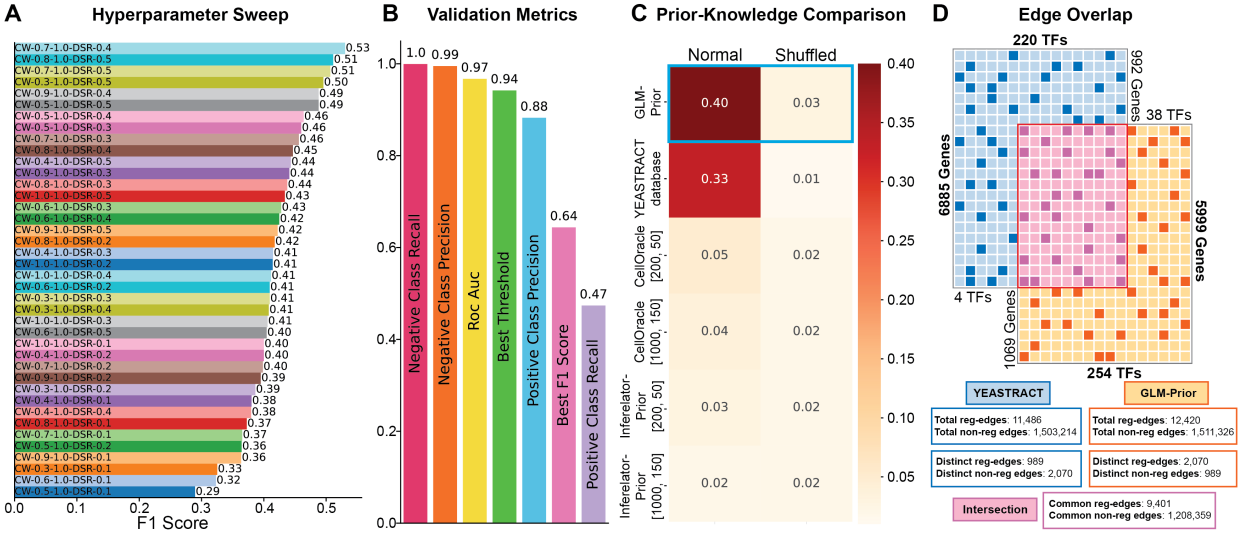

Figure 2: Evaluating GLM-Prior against motif-based and database-derived prior knowledge in yeast. (**A**) A hyperparameter sweep evaluating F1 scores across class weights and downsampling rates on a 1% held-out validation set. (**B**) Final model validation performance metrics following full training (across 10 epochs) with optimal hyperparameters. (**C**) prior knowledge benchmarking using AUPRC against an independent literature-curated gold standard. GLM-Prior outperforms database prior knowledge, a shuffled control, and motif-based priors across two windowing strategies. (**D**) Structural comparison between YEASTRACT and GLM-Prior matrices. Schematic Venn diagram shows the overlap in predicted positive interactions. GLM-Prior identifies $2,070$ novel regulatory interactions not present in YEASTRACT and reclassifies 989 edges, while preserving a large subset of shared interactions.

To evaluate the quality of the predicted prior knowledge matrix generated by GLM-Prior, we compare it to a curated database prior (YEASTRACT) (11) as well as two widely used motif-based methods (Inferelator-Prior (6) and CellOracle's base GRN (8)) (Figure 2C). To fairly compare each approach, we compute the area under the precision-recall curve (AUPRC) of each generated prior knowledge matrix against an independent gold standard of literature curated TF-gene regulatory interactions (26). GLM-Prior achieves an AUPRC of 0.40, significantly outperforming both a shuffled negative control (AUPRC = 0.03) and the YEASTRACT prior (AUPRC = 0.33), corresponding to a 21.2% performance improvement over the well-established database. Additionally, GLM-prior substantially outperforms the two motif-based methods. For each motif-based method, we construct prior knowledge matrices using two TF-gene assignment strategies. In the first configuration, we define the regulatory assignment window as 200bp downstream and 50bp upstream of the gene promoter. Under this setup, CellOracle's base GRN and Inferelator-Prior achieves AUPRCs of 0.05 and 0.03, respectively. In the second configuration, using a broader window of $1,000$bp downstream and 150bp upstream, CellOracle's base GRN reaches an AUPRC of 0.04, while Inferelator-Prior falls to 0.02. These results demonstrate that GLM-Prior produces a state-of-the-art prior knowledge matrix, significantly outperforming proximity-based motif assignment methods and highlighting the advantage of sequence-based inference.

To complement the quantitative benchmarking results, we further examine the differences between the YEASTRACT-derived prior and the GLM-Prior matrix by comparing their respective sets of regulatory edges. The YEASTRACT prior includes $11,486$ positive and $1,503,214$ negative interactions across 220 TFs and 6885 genes (described in Appendix A.1). However, because GLM-Prior requires nucleotide sequences for both TFs and genes, the final sequence-filtered dataset was restricted to 254 TFs (from the CisBP database (27)) and $5,999$ genes with available DNA sequences. After filtering, 466 positive interactions present in YEASTRACT are excluded due to missing sequence information.

Using the remaining sequence-compatible TF-gene pairs, GLM-Prior infers $12,420$ positive and $1,511,326$ negative interactions. Of these, 2070 interactions that are labeled as non-regulatory (0) in YEASTRACT are

flipped to regulatory (1) by GLM-Prior, and 989 interactions are flipped from regulatory to non-regulatory, reflecting a combination of refinement and reassignment. A schematic Venn diagram (Figure 2D) illustrates the overlap between the two priors, highlighting a substantial number of shared interactions as well as novel regulatory links introduced by GLM-Prior.

These findings demonstrate that GLM-Prior preserves many biologically validated interactions while introducing new, sequence-informed edges, and reclassifying others. This suggests that GLM-Prior not only generalizes beyond motif-based and curated prior knowledge but also actively refines the structure of prior knowledge matrices, ultimately producing a more comprehensive and data-driven foundation for downstream GRN inference.

Benchmarking GLM-Prior in yeast reveals limited additive value of GRN inference

Following the first stage of our pipeline, in which GLM-Prior generates a sequence-informed prior knowledge matrix for yeast, we evaluate its impact on downstream GRN inference using the second stage of our dual-stage framework. Specifically, we apply PMF-GRN (7) to two single-cell gene expression datasets from yeast, GSE125162 (Y1) (28) and GSE144820 (Y2) (29). We benchmark GLM-Prior against the YEASTRACT-derived prior knowledge by comparing GRN inference performance across both datasets (Figure 3A). As in previous evaluations, we use AUPRC as the primary metric, with performance assessed against a literature-curated gold standard (26). Dashed lines in Figure 3A indicate the AUPRC of each prior alone, allowing us to quantify whether GRN inference offers any additive benefit.

First, we apply PMF-GRN to each dataset independently to assess how the choice of prior affects GRN inference across different single-cell datasets. Using the GLM-Prior, we observe that PMF-GRN was not able to improve upon the initial prior, achieving AUPRCs of 0.39 for Y1 and 0.36 for Y2. This result suggests that gene expression data contributes minimal information beyond what is already captured by sequence, despite the model's capacity to integrate both sources. Rather than reflecting a limitation of the inference method, this finding underscores the fact that expression data, especially at single-cell resolution, is often too sparse, noisy, or context-specific to reveal the full regulatory architecture encoded in sequence. In contrast, when using the YEASTRACT prior, PMF-GRN improves performance from the prior's baseline, achieving AUPRCs of 0.35 ($\approx$ 6.06% performance increase) and 0.34 ($\approx$ 3.03%) for Y1 and Y2, respectively. This improvement reflects that inference can add value when the prior is weaker or incomplete. Together, these findings support a reframing of GRN inference: expression data may be best understood as a mechanism for reweighting or modulating a high-quality prior, rather than discovering novel structure *de novo*.

Building on these results, we next examine whether incorporating additional expression data could enhance GRN inference beyond what is already captured by sequence. To test this, we apply PMF-GRN separately to Y1 and Y2 using the same GLM-Prior, and average the resulting GRNs. This combined-task inference approach yielded a slight improvement in AUPRC to 0.40, but crucially, only achieved performance equal to that of GLM-Prior alone. Here, we find that adding more expression data does not uncover new regulatory structure beyond what is already encoded in sequence. Only when we combine both the expression datasets and the prior knowledge sources, integrating GLM-Prior with YEASTRACT, does GRN inference provide a modest performance increase, achieving an AUPRC of 0.42. This suggests that such gains likely stem from the complementary nature of the priors rather than novel information extracted from expression data. These findings reinforce that while combining priors may be beneficial, expression data contributes little additional value when a high-quality prior is already available.

These findings further highlight that GLM-Prior provides a substantially stronger foundation than traditional motif-based priors, and that the capacity of GRN inference to improve upon this foundation is limited. In nearly all settings, inference with PMF-GRN does not outperform the GLM-Prior itself, indicating that the predictive signal is already embedded in sequence. The slight performance gain observed when integrating complementary priors underscores the value of diverse sources of prior knowledge, but also reveals that current inference methods may be nearing a ceiling in what they can extract from expression data alone. In this light, single-cell expression may be better understood not as a driver of novel edge discovery, but as a context-dependent signal that modulates an already-established regulatory scaffold.

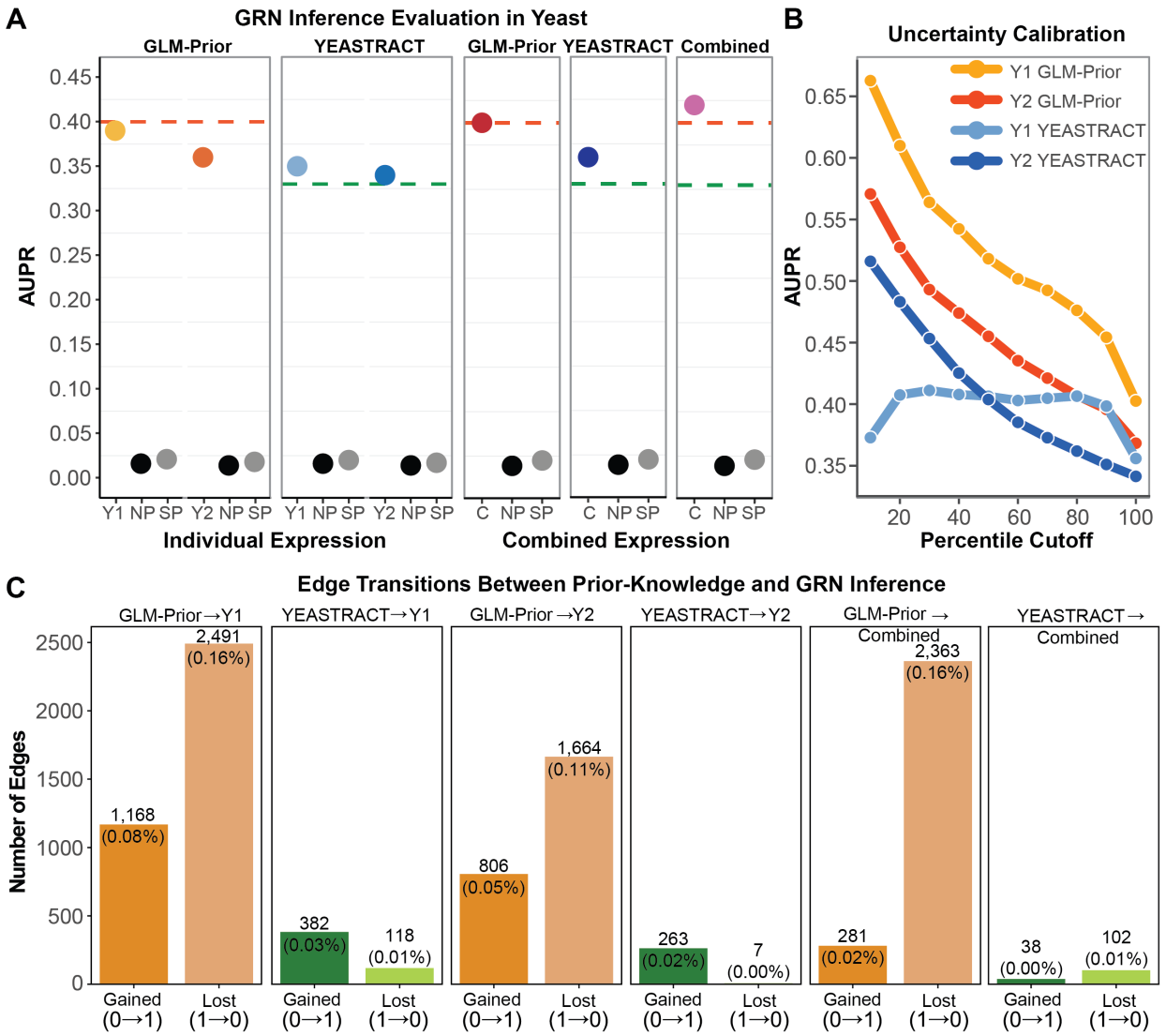

Figure 3: Evaluating GRN inference performance and uncertainty calibration with GLM-Prior in yeast. (**A**) AUPRC performance of GRNs inferred with PMF-GRN across two single-cell expression datasets (Y1 and Y2), using GLM-Prior and YEASTRACT as prior knowledge. Each prior was evaluated individually on Y1 and Y2, with no-prior (NP) and shuffled-prior (SP) controls, shown as black and gray dots respectively. Dashed lines represent the AUPR of the input prior knowledge to demonstrate whether GRN inference has any additive benefit. Additional conditions include combined expression of Y1 and Y2 datasets (C) with each prior and an integrated prior combining YEASTRACT and GLM-Prior (combined). GLM-Prior consistently outperforms YEASTRACT across all settings. (**B**) Uncertainty calibration plots for GRNs inferred on Y1 and Y2 using PMF-GRN with GLM-Prior and YEASTRACT priors. Interactions binned by posterior variance show higher predictive accuracy in lower-variance bins, with GLM-Prior yielding more confident predictions overall. (**C**) Edge transitions between prior knowledge matrices and GRN inference demonstrates GRN inference prunes high-quality priors (GLM-Prior) or contributes edges to less comprehensive priors (YEASTRACT).

To ensure that the observed improvements are driven by the quality of prior knowledge and not random variation, we evaluate two baseline conditions. In the first, we shuffle the gene labels in the prior matrix to disrupt any meaningful regulatory structure. In the second, we remove the prior entirely by zeroing out the matrix. Both the shuffled-prior (SP) and no-prior (NP) baselines yield AUPRC scores near zero, confirming that the predictive performance of PMF-GRN stems from biological signal encoded in both GLM-Prior and YEASTRACT rather than chance.

In addition to these point estimates, PMF-GRN provides posterior variance for each TF-gene interaction, offering a measure of uncertainty. To assess whether these uncertainty estimates reflect predictive reliability, we bin interactions by posterior variance, sorting from low to high, and evaluating AUPRC within each bin, following the procedure in (7). As expected, lower-variance bins exhibit higher predictive accuracy, confirming that posterior variance serves as a reliable proxy for model confidence (Figure 3B). GRNs inferred using GLM-Prior consistently demonstrate lower posterior variance than those inferred with YEASTRACT, indicating that the model is more confident in its predictions when guided by a strong, sequence-informed prior. While YEASTRACT-based GRNs also show relatively low variance, their values are consistently higher, mirroring the observed gap in AUPRC. These findings demonstrate that GLM-Prior not only yields more accurate GRN predictions, but also produces lower uncertainty estimates, suggesting that most of the regulatory information extracted from expression data is already captured by the sequence-derived prior.

To further investigate how inference modifies the structure of the prior, we directly compare the binarized edge sets of each inferred GRN to its corresponding prior and quantify the number of edges gained (from 0 to 1) and lost (from 1 to 0) after GRN inference (Figure 3C). This analysis reveals that when using GLM-Prior, GRN inference consistently removes more edges than it adds. For example, in Y1, $2,491$ edges are removed while only $1,168$ are added (0.16% and 0.08% of all possible edges, respectively), with similar patterns in Y2 and the combined-task model. This pruning effect is strongest in the combined-expression setting, where the number of added edges reduces to 281, suggesting that integrating more expression data does not uncover new interactions, but instead serves to tighten the prior. In contrast, inference starting from the weaker YEASTRACT prior leads to more edge gains than losses, consistent with the notion that expression data contributes more when the prior is incomplete. These edge transitions provide structural evidence, independent of performance metrics, that when a high-quality prior is available, expression data acts primarily to filter or refine the network rather than discover new regulatory structure.

Together, these findings suggest a critical shift in the role of expression data with GRN modeling. Rather than serving as the primary driver of edge discovery, single-cell expression more often functions as a context-specific signal that modulates an already established regulatory scaffold. In high-quality prior settings, such as GLM-Prior, inference mainly suppresses unsupported edges or reweights them, potentially reflecting biologically meaningful refinement such as cell-state-specific silencing. As GRN modeling progresses, future methods may need to look beyond edge-centric inference and focus instead on capturing how regulatory programs shift across time, perturbations, and cell-types, using expression not to discover structure, but to annotate its dynamic use.

Cross-method benchmarking reveals the dominant role of prior quality over inference algorithm choice

Having shown that GLM-Prior alone captures the majority of regulatory information recoverable from single-cell expression data in yeast, we next ask whether this pattern holds across different GRN inference algorithms. To address this, we compare our dual-stage training pipeline to two widely used methods, the Inferelator 3.0 (6) and CellOracle (8), both of which integrate motif-based prior construction with regression-based inference. While these methods are typically evaluated as unified pipelines, it remains unclear whether their performance is driven by the prior, the inference algorithm, or their combination. To disentangle these effects, we perform a systematic cross-evaluation in which each inference algorithm (PMF-GRN, Inferelator, and CellOracle) is paired with each available prior knowledge matrix (GLM-Prior, Inferelator-Prior, CellOracle's Base GRN, and YEASTRACT). This design enables us to isolate the effects of prior construction from those of the inference model and assess how their interaction impacts GRN recovery (Figure 4).

Across the three sequence and motif-based priors evaluated (GLM-Prior, Inferelator-Prior, and CellOracle's base GRN), the Y1 dataset (n cells = $38,225$) (Figure 4A) reveals that GLM-Prior consistently enables the

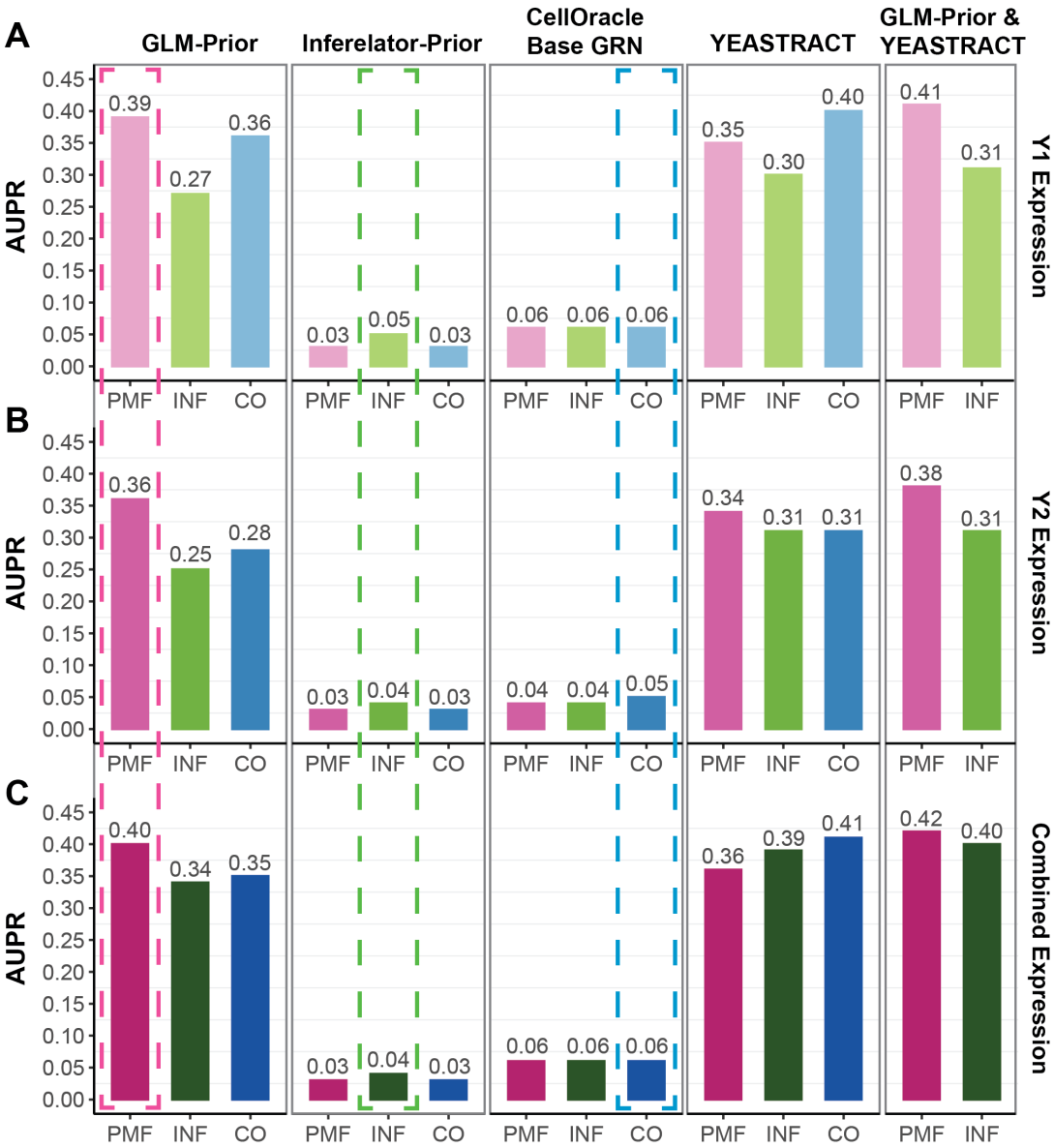

Figure 4: Cross-evaluation of GRN inference methods and prior knowledge matrices in yeast. (**A**) Performance on the Y1 dataset ($n = 38,225$ cells), showing AUPRC values for all combinations of GRN inference algorithms, PMF-GRN (PMF), Inferelator (INF), and CellOracle (CO), paired with four prior knowledge matrices, GLM-Prior, YEASTRACT, Inferelator-Prior, CellOracle Base GRN. Combined priors (GLM-Prior and YEASTRACT) are also evaluated for PMF-GRN and the Inferelator. (**B**) Performance on Y2 dataset ($n = 6,118$ cells). (**C**) Performance on the combined Y1 and Y2 datasets ($n = 44,343$ cells). Across all panels, PMF-GRN paired with the combined GLM-Prior and YEASTRACT achieves the highest overall AUPRC, highlighting the complementary value of integrating sequence-derived and curated regulatory information. Dashed boxes indicate each method's original end-to-end pipeline configuration (i.e., its original prior paired with its inference algorithm), providing a baseline for comparing standard method performance across all datasets and prior configurations.

highest performance across all inference algorithms. PMF-GRN achieves the highest AUPRC (0.39), followed by CellOracle (0.36) and the Inferelator (0.27). In contrast, both motif-based priors result in uniformly poor performance across all algorithms, with AUPRCs never exceeding 0.06. These results demonstrate that GLM-Prior provides a robust, sequence-informed regulatory scaffold that generalizes across methods, underscoring that the dominant factor in GRN inference performance is the quality of the prior, not the algorithm used to interpret it.

Despite the strength of the prior, GRN inference on top of GLM-Prior does not improve AUPRC and, in some cases, slightly reduces performance relative to the prior alone. This suggests that existing inference methods, which optimize for alignment with expression covariance, may filter out sequence-derived interactions from GLM-Prior that are not strongly supported by gene expression patterns. Rather than revealing new interactions, expression may act as a selective filter, reinforcing only the context-specific subset of regulatory relationships visible in a particular dataset. This filtering behavior, observed across methods, highlights the broader limitation of current GRN inference frameworks: when prior quality is high, the marginal value of expression-driven inference becomes constrained not by model capacity, but by the incompleteness and sparsity of expression data itself.

We observe a similar pattern with the YEASTRACT-derived prior. All three inference methods perform well with YEASTRACT, achieving AUPRCs of 0.35 (PMF-GRN), 0.30 (Inferelator), and 0.40 (CellOracle). The stronger performance of the Inferelator and CellOracle with YEASTRACT compared to GLM-Prior likely reflects the curated nature of YEASTRACT, which provides a set of experimentally validated interactions that are more easily recoverable by expression-driven models. In contrast, GLM-Prior includes additional sequence-derived interactions that may not exhibit strong statistical support in gene expression data. Inference methods that prioritize expression covariance, particularly regression-based approaches, tend to filter out these edges, limiting their ability to leverage prior information not directly supported by the expression snapshot. This filtering behavior is a fundamental limitation of current GRN inference frameworks, when the prior is high quality, expression data serves less as a discovery signal and more as a context-specific filter, potentially discarding valid regulatory edges.

Despite the filtering limitations of expression-driven inference, combining GLM-Prior with YEASTRACT offers a practical strategy to improve GRN recovery by leveraging the complementary strengths of both priors. Using a task-specific integration strategy, where GRNs are inferred separately on Y1 using each prior and then merged post-inference by averaging, PMF-GRN and the Inferelator achieve the highest overall AUPRCs of 0.41 and 0.31, respectively. These additive gains suggest that GLM-Prior and YEASTRACT capture distinct but complementary aspects of regulatory structure. In particular, the improvement likely reflects the coverage gap in GLM-Prior, where 886 genes present in the YEASTRACT prior lack associated sequence information and were therefore not modeled. Together, these findings emphasize that while expression data has limited power to discover new edges, combining diverse sources of prior knowledge, both curated and sequence-derived, yields a more complete and biologically grounded representation of regulatory structure.

A similar pattern emerges in the Y2 dataset ($6,118$ cells) (Figure 4B). PMF-GRN again achieves the highest AUPRC with GLM-Prior (0.36), followed by CellOracle (0.28) and the Inferelator (0.25). The slight decrease in performance relative to Y1 likely reflects the smaller dataset size and reduced statistical power in the Y2 expression data. Despite this, motif-based priors again fail to achieve competitive performance across all methods, with AUPRCs never exceeding 0.05. These results reinforce that the quality of the prior, rather than the inference algorithm choice or dataset size, is the primary determinant of GRN reconstruction performance.

Importantly, when combining Y1 and Y2 expression datasets ($44,343$ cells) (Figure 4C), GLM-Prior again supports the highest overall AUPRCs across methods, with 0.40 for PMF-GRN, 0.35 for CellOracle, and 0.34 for the Inferelator. As in the individual datasets, motif-based priors remain noncompetitive with AUPRCs not exceeding 0.06, while YEASTRACT continues to achieve strong performance (0.36 for PMF-GRN, 0.39 for Inferelator, and 0.41 for CellOracle). Notably, integrating both the expression datasets and prior sources (GLM-Prior and YEASTRACT) in a task-specific manner yields the best overall AUPRC across all configurations tested for PMF-GRN (0.42) and the Inferelator (0.40). These results suggest that the

performance gain stems from integrating complementary sources of prior knowledge, rather than from the addition of expression data.

Taken together, these results reinforce our conclusion that high-quality priors, such as those generated by GLM-Prior, resolve much of the regulatory structure required for GRN reconstruction on their own, substantially limiting the value of expression-based inference. Current inference frameworks, particularly those relying on regression, offer little consistent benefit on top of such priors and may even degrade performance by pruning biologically plausible but expression-invisible interactions. The modest gains observed through integration of GLM-Prior with curated databases like YEASTRACT highlight that expanding and diversifying prior knowledge is a more effective strategy for improving GRN inference than refining the inference algorithms themselves. Future progress will require models that treat expression data not as the source of regulatory discovery, but as a dynamic readout that reweights or contextualizes an underlying sequence-derived scaffold.

Cross-species evaluation highlights the strengths and limitations of GLM-Prior for generalizable prior construction

Building on our findings in yeast, we next evaluate whether GLM-Prior can generalize to more complex mammalian systems. Specifically, we ask whether regulatory logic learned from species-level interaction data can be transferred across cell types or even across species to construct meaningful prior knowledge networks using nucleotide sequence as input.

We first train GLM-Prior separately in human and mouse using TF-gene interaction pairs from the STRING (30; 31; 32; 33) and TRRUST (34; 35) databases. We adopt the same benchmark datasets as used in the BEELINE framework (36) for consistency. To evaluate generalization, we test these models on held-out, cell type-specific ChIP-seq networks derived from human and mouse embryonic stem cells (hESCs and mESCs), which are never seen during training. GLM-Prior is provided only with the nucleotide sequences of TFs and target genes and tasked with predicting regulatory interactions *de novo*.

Across these single-species experiments, GLM-Prior achieves AUPRCs of 0.24 in human and 0.22 in mouse, demonstrating its ability to generalize from species-level interaction data to cell type-specific regulatory contexts (Figure 5A). Importantly, these results reinforce our central claim that a well-constructed, sequence-informed prior can recover biologically meaningful regulatory structure, independent of cell-type or matched expression data. This highlights the ability of GLM-Prior to model generalizable regulatory features directly from nucleotide sequence alone.

To evaluate the extent to which GLM-Prior generalizes across species, we next evaluate its performance in a cross-species transfer learning setting. We train GLM-Prior exclusively on human TF-gene interactions and apply it to predict regulatory interactions in mESC, using held-out ChIP-seq data for evaluation. Remarkably, despite never being exposed to mouse training data, the human-trained model achieves an AUPRC of 0.23, slightly outperforming the mouse-specific model (0.22) (Figure 5B). This result highlights GLM-Prior's ability to leverage shared regulatory features that generalize across species. Such transferability likely reflects both sequence homology and conservation of core regulatory programs between human and mouse, particularly in early developmental contexts like embryonic stem cells (37; 38; 39). By learning directly from sequence, rather than species-specific annotations, GLM-Prior captures regulatory signals that move beyond species boundaries, enabling accurate prior construction even in the absence of organism-specific training data.

We then evaluate the reverse transfer, training GLM-Prior on mouse and testing it on human. While performance decreases slightly to 0.17, this reduction likely reflects the limited scale of mouse training data ($20,144$ positive interactions) compared to the much larger human dataset ($43,963$ positive examples), rather than a fundamental limitation in cross-species generalization. Taken together, these results highlight GLM-Prior's ability to transfer regulatory logic across species boundaries when evolutionary conservation is high, while also emphasizing the critical role of training data scale in enabling effective generalization.

To further test the limits of cross-species generalization, we apply a GLM-Prior model trained on both human and mouse to yeast. Despite leveraging data from two mammalian species during training, the model

performs poorly in yeast, achieving an AUPRC of 0.02 (Figure 5B), far below the yeast-specific model (0.40). This sharp performance drop highlights the biological limits of transfer learning in regulatory genomics. Regulatory architectures in yeast differ fundamentally from those in mammals, including differences in sequence grammar, TF binding preferences, and genome organization. These results demonstrate that while GLM-Prior can generalize across closely related species, effective prior construction still requires species-specific modeling when regulatory systems diverge across deep evolutionary distances.

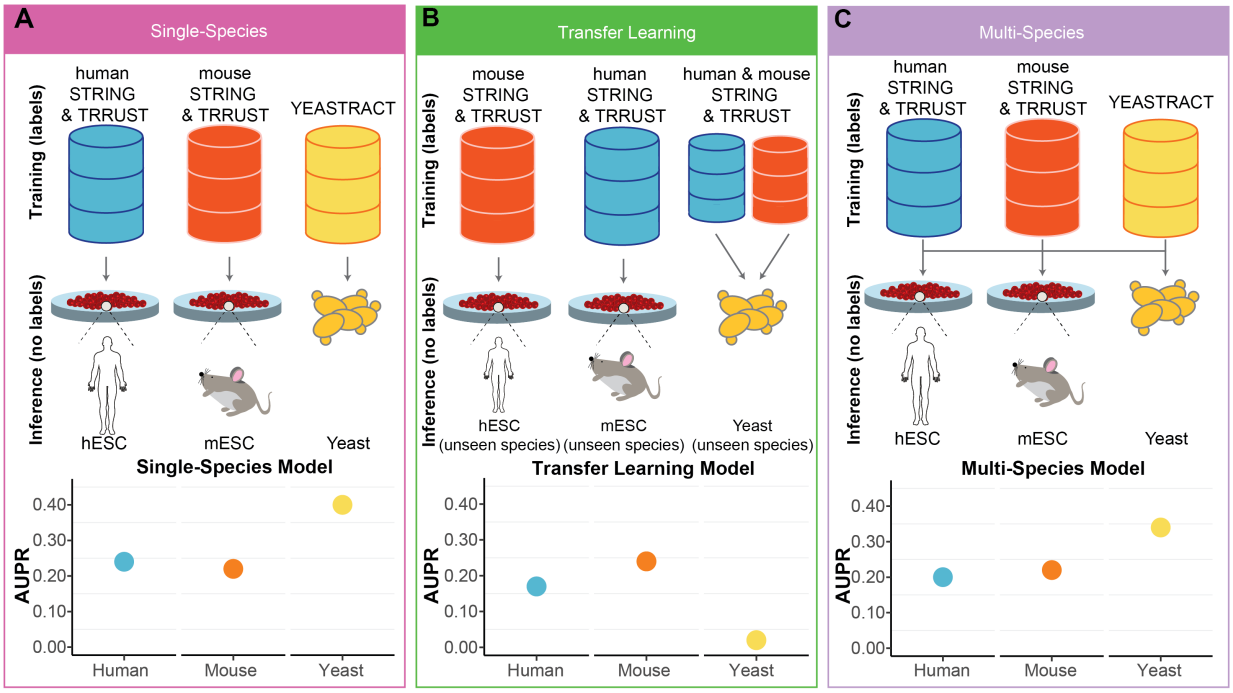

Figure 5: Overview of species-specific, transfer-learning and multi-species training paradigms for prior knowledge construction with GLM-Prior. (**A**) Single-species training: GLM-Prior is fine-tuned on species-specific databases and evaluated on held-out ChIP-seq datasets (human, mouse) or literature curated gold standard (yeast). Strong AUPRCs are achieved in yeast and moderate AUPRCs in human and mouse, demonstrating the ability of GLM-Prior to generalize to cell-type specific priors from species-level training data. (**B**) Transfer learning: A model trained on human generalizes well to mouse, while a model trained on mouse performs slightly worse in human, likely due to the smaller scale of the mouse training set. A model trained on human and mouse fails to generalize to yeast, highlighting the limits of cross-species transfer when regulatory architectures diverge. (**C**) Multi-species training: Sequential fine-tuning on human, mouse, and yeast maintains strong performance in yeast and mouse but results in reduced accuracy in human, suggesting a trade-off between generalization and species-specific retention, potentially due to order-dependent feature overwriting during sequence training.

Finally, we evaluate whether a single, generalizable GLM-Prior model could be trained across multiple species. To test this, we sequentially fine-tune GLM-Prior on the human, mouse, and yeast datasets, aiming to capture regulatory features shared across eukaryotic systems. The resulting multi-species model achieves AUPRCs of 0.20 in human, 0.22 in mouse, and 0.34 in yeast (Figure 5C). While performance in mouse and yeast remains strong, the human AUPRC drops compared to the single-species model (0.20 vs 0.24). This asymmetry may reflect the added complexity of human gene regulation, which relies on distal enhancers (40), combinatorial TF binding (41), and context-specific mechanisms that simpler systems like yeast do not exhibit. These findings suggest that while multi-species training can capture broadly conserved regulatory features, preserving high-resolution, species-specific performance may require more targeted strategies.

Nevertheless, the ability of the multi-species model to maintain strong performance in yeast and mouse suggests that cross-species training can successfully capture broadly conserved regulatory principles, particularly

in simpler or more evolutionarily stable systems. These results illustrate both the promise and the limitations of multi-species modeling, highlighting the challenge of balancing generalization with the preservation of species-specific regulatory precision, particularly when regulatory architectures diverge substantially across evolution.

Taken together, these findings demonstrate that GLM-Prior not only generalizes across species and cell types, but also effectively transfers regulatory logic between closely related organisms. However, the failure to generalize to yeast and the trade-offs observed in the multi-species training underscore that species-specific modeling remains essential when regulatory architectures diverge. More broadly, these results reinforce our central conclusion that improving the construction of biologically grounded priors is currently the most impactful strategy for advancing GRN modeling. While GLM-Prior provides a scalable and generalizable foundation for prior construction, realizing the full potential of sequence-informed GRN inference will require the development of next generation frameworks that move beyond linear assumptions to flexibly integrate both conserved and species-specific logic.

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 apply our dual-stage training pipeline to 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 6A), 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 (42), 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 6A). 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 6B). 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.

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 6C). 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 6D). 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 our dual-stage pipeline 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 6E). 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 6F). 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.

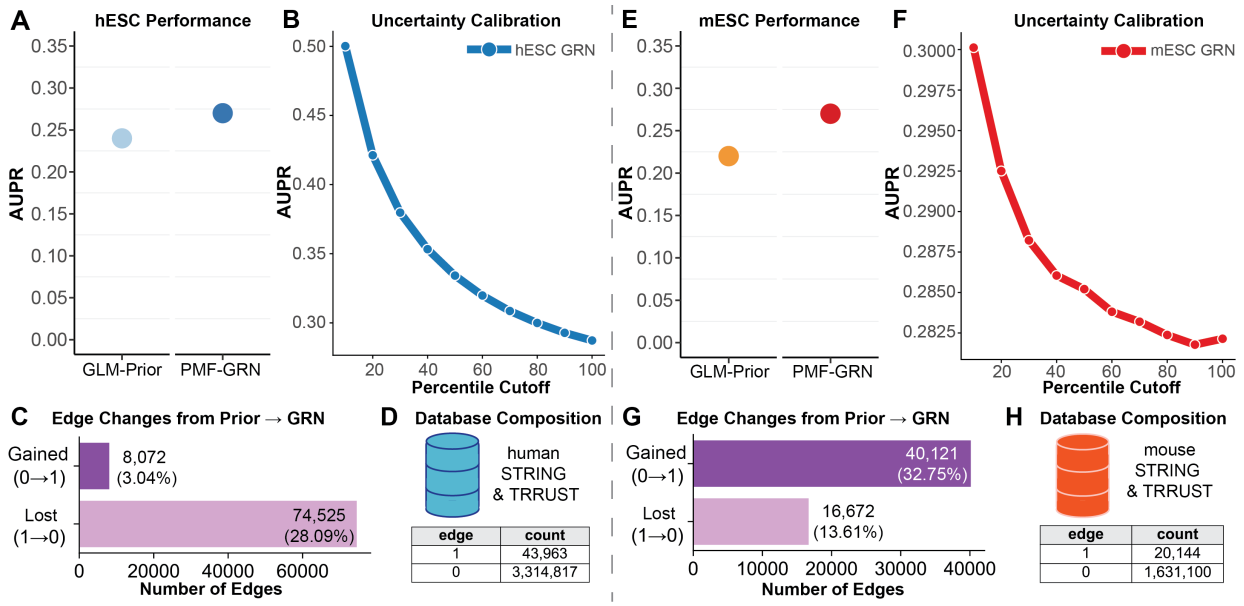

Figure 6: Performance analysis of GLM-Prior and PMF-GRN in human and mouse ESCs. (**A**) AUPRC scores in hESCs show modest gains from GRN inference over the sequence-derived prior. (**B**) Uncertainty calibration in hESCs confirms that lower posterior variance predicts higher precision. (**C**) GRN inference 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.

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 6G). 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 6H). 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.

Beyond edge-level improvements, our dual-stage framework also provides interpretability advantages. Across both species, PMF-GRNs posterior variance estimates are well-calibrated, offering a principled way to asses confidence in individual edges. This enables expression data to support network refinement and edge prioritization, even when structural contributions are modest. Additionally, model outputs such as transcription factor activity trajectories and marker gene-centered subnetworks (Appendix Figures A1B-D and A2) demonstrate that expression data can enhance the biological relevance of inferred networks by contextualizing regulatory usage over time, further supporting its role as a dynamic annotation layer over a sequence-derived scaffold.

Discussion

Accurately reconstructing GRNs remains a central challenge in genomics, particularly in complex systems where direct experimental measurements are incomplete or unavailable. In this work, we present GLM-Prior, a genomic language model fine-tuned to predict TF-gene regulatory interactions directly from DNA sequence. By integrating GLM-Prior with PMF-GRN, a probabilistic matrix factorization framework, we establish a dual-stage training pipeline that decouples prior-knowledge construction from expression-based inference. This separation allows both components to be independently optimized, enabling us to systematically evaluate their individual and combined contributions to GRN reconstruction.

Across three eukaryotic species, we show that the dominant driver of GRN inference performance is the quality of the prior, not the downstream inference algorithm. In yeast, GLM-Prior substantially outperforms traditional motif-based priors, achieving the highest AUPRC and recovering biologically meaningful regulatory edges. However, when paired with expression data in PMF-GRN, the additive performance gains are marginal, suggesting that GLM-Prior alone captures most of the regulatory information recoverable from single-cell expression. This saturation effect is particularly pronounced in yeast, but generalizes to more complex systems, including human and mouse.

Cross-method benchmarking further reveals that GLM-Prior improves the performance of multiple GRN inference algorithms, not just PMF-GRN. Nevertheless, these gains are largely determined by the prior itself, with inference providing limited additional benefit. This observation challenges the prevailing view that GRN inference is primarily a problem of modeling expression data, and instead shifts the focus toward the construction of accurate prior knowledge as the key determinant of inference performance.

We further demonstrated that GLM-Prior generalizes effectively across closely related species, with models trained in human successfully transferring to mouse and even out-performing a mouse-specific model, despite never being trained on mouse data. This demonstrates that GLM-Prior can capture regulatory features that are conserved at the sequence level, such as binding site preferences of shared transcription factors, enabling accurate prediction of regulatory interactions even in species where no direct training data is available. However, this generalization fails to extend to more distantly related organisms like yeast, where regulatory grammar, TF binding motifs, and genome organization diverge substantially.

In contrast to transfer learning, multi-species training, where GLM-Prior is sequentially fine-tuned on human, mouse, and yeast, captures some shared regulatory logic but comes with notable trade-offs. While performance remains strong in yeast and mouse, it is reduced in human, likely due to catastrophic forgetting or the dilution of species-specific features during sequential training. These results suggest that while transfer learning can be highly effective between evolutionary neighbors, multi-species modeling requires careful

balancing to preserve organism-specific regulatory programs, particularly in complex systems like human where distal enhancers and combinatorial regulation play dominant roles.

Applying our dual-stage pipeline to human and mouse embryonic stem cells reveals that the additive value of GRN inference is highly dependent on the quality of the sequence-derived prior. In hESCs, where GLM-Prior was trained on a large and well-annotated regulatory dataset, PMF-GRN provided only marginal gains, with expression data primarily pruning rather than expanding the network. In contrast, in mESCs, where the prior was trained on a smaller, less comprehensive dataset, expression data contributed more substantially, expanding the network by introducing novel edges absent from the prior. These findings reinforce our broader conclusion that prior construction, not expression-based inference, is the primary driver of GRN reconstruction performance.

This work establishes GLM-Prior as a scalable, generalizable framework for constructing high-quality regulatory priors directly from nucleotide sequence. By moving beyond motif-based methods, GLM-Prior leverages large-scale transformer models to capture non-linear regulatory features that generalize across cell types and species. Across yeast, mouse, and human, our results show that GRN reconstruction performance is driven primarily by the quality of the prior, rather than the specific inference algorithm.

While our dual-stage pipeline with PMF-GRN demonstrates that sequence-informed priors can support biologically interpretable GRNs and well-calibrated uncertainty estimates, our results reveal that current inference frameworks provide only marginal additive benefit when strong priors are available. This suggests that GLM-Prior substantially alleviates the long-standing bottleneck of constructing high-quality priors in GRN modeling, shifting the challenge to how we leverage these priors in downstream inference.

Our findings further reveal a critical limitation in current GRN inference frameworks. Although these methods incorporate priors, they ultimately fit to the expression data, often filtering out regulatory edges from the prior that are not strongly supported by gene expression covariance. This behavior likely reflects the incomplete and noisy nature of single-cell data, which may fail to capture the full regulatory complexity encoded in the prior due to technical noise, dropout events, or biological variability. As a result, inference methods risk discarding biologically meaningful regulatory edges simply because they are not evident in the expression snapshot, limiting their ability to fully leverage the prior's information.

These findings raise a central question: what is the true role of expression data in GRN inference when strong sequence-derived priors are available? While expression offers context-specific insight into regulatory activity, it does not consistently recover the underlying regulatory topology. Our results suggest that expression data may be better suited to contextualize or reweight a latent regulatory scaffold, rather than define it outright. Going forward, inference frameworks must move beyond edge-centric recovery toward models that treat expression as a dynamic signal that modulates a fixed encoded network. In this reframing, GLM-Prior shifts the challenge in GRN modeling from discovering structure to interpreting regulatory activity across conditions, opening the door to more generalizable and biologically plausible models of gene regulation.

Online Methods

The GLM-Prior model

We developed GLM-Prior, a genomic language model fine-tuned to predict transcription factor (TF)-target gene regulatory interactions directly from sequence data. Specifically, we fine-tuned the 250 million parameter Nucleotide-Transformer model (17), originally pre-trained on the genomes of 850 species, to infer binary interaction matrices between TFs and their target genes using their nucleotide sequences. These predicted interactions serve as a prior-knowledge for downstream GRN inference, introducing biologically grounded constraints that guide the structure of the inferred networks.

The model uses a transformer encoder architecture to process two distinct sequence types, TF motif sequences from the CisBP database (27) and gene body sequences derived from genome annotations in a GTF file. We append a classification head to the transformer encoder to predict whether each TF-gene pair represents a true regulatory interaction (positive label) or non-regulatory interaction (negative label). Training labels are obtained from experimentally validated interaction databases such as STRING (30; 31; 32; 33) and TRRUST

(34; 35), supplying high confidence positive and negative examples of regulatory interactions. This design allows the model to leverage both sequence-level regulatory motifs encoded in the pre-trained transformer and database-derived TF-gene interactions to improve the predictive performance and generalization across genomic contexts.

Each training example is constructed by concatenating the nucleotide sequence of a TF with that of its candidate target gene, separated by a special classification token (`<cls>`). This composite sequence is tokenized, during which a model-specific (`<cls>`) token is automatically prepended to the input. The tokenized sequence is passed through the pre-trained transformer, which produces contextual embeddings across the full input. The embedding corresponding to the prepended (`<cls>`) token is fed into the classification head composed of a dropout layer, a linear projection, a tanh activation, a second dropout, and a final linear layer that maps to two logits representing the binary interaction label. This setup allows the model to jointly learn representations of TF-binding specificity and gene regulatory potential in a unified sequence-to-interaction framework.

Dataset construction and preprocessing

To train the model, we constructed a dataset of TF-gene pairs where positive examples were derived from a curated database of experimentally validated interactions (YEASTRACT (11), STRING (30; 31; 32; 33) and TRRUST (34; 35)), and negative examples were sampled from a combination of all remaining TF-gene sequence pairs. This dataset was split into a 99% training and 1% validation set.

Given the substantial class imbalance between positive and negative samples, we implemented a downsampling strategy to reduce the number of negative samples in the training set. This preserved the diversity of negative samples, while preventing the model from overfitting to the negative class. Specifically, the number of retained negative samples after downsampling can be defined as,

$$N_{sampled} = \lfloor r \cdot N_{-} \rfloor, \tag{1}$$

where $r \in (0, 1)$ is the downsampling rate, and N_{-} is the total number of negative examples in the dataset.

To further address class imbalance during training, we designed a custom DataLoader that performs even class batch sampling. Each batch was constructed to contain an equal number of positive and negative samples, ensuring a balanced signal during training. Specifically, each batch is defined as,

$$B = \{(x_i^+, x_j^-) : x_i^+ \in X_+, x_j^- \in X_-\}, \tag{2}$$

where positive examples x_i^+ were sampled uniformly with replacement from X_+ (the positive class) and negative examples x_j^- were cycled through without replacement:

$$x_i^+ \sim X_+, \quad x_j^- \sim X_- \quad \text{with} \quad j \bmod N_-. \tag{3}$$

This strategy ensures that each negative example was seen exactly once during training while positive examples were reused as needed to maintain class balance. Even-class batching was critical for stabilizing training dynamics and improving the model's sensitivity to true positive interactions without inflating the false-negative rate.

Model Architecture and Classification

The model architecture builds on the pre-trained Nucleotide Transformer by retaining its embedding and transformer encoding layers. A classification head is applied to the final hidden state corresponding to the

prepended (<cls>) token to compute logits for a binary classification task. This classification head outputs a two-dimensional logit vector $z \in \mathbb{R}^2$ representing unnormalized scores for the positive and negative classes,

$$z = W \cdot h + b, \tag{4}$$

where h is the transformed hidden state of the (<cls>) token following the intermediate projection and non-linearity, and W, b are the weights and bias of the final linear layer. The components of z are denoted z_+ and z_-, corresponding to the logits for the positive and negative classes, respectively. Class probabilities are computed using a softmax function,

$$P(y = 1|x_{TF}, x_{gene}) = \frac{\exp(z_+)}{\exp(z_+) + \exp(z_-)}, \tag{5}$$

Training used a class-weighted cross-entropy loss function to account for residual class imbalance after downsampling. The loss assigns a fixed weight of 1.0 to positive examples, while the negative class weight w_- is tuned through hyperparameter search to optimize the balance between precision and recall. The loss is computed as,

$$\mathcal{L} = -w_+ y \log p - w_-(1-y)\log(1-p), \quad \text{where} \quad p = \frac{\exp(z_+)}{\exp(z_+) + \exp(z_-)}. \tag{6}$$

Here, $y \in \{0, 1\}$ is the true label, z_+ and z_- are the predicted logits for the positive and negative classes, respectively. w_- is the down-weighted negative class weight, varied during hyperparameter search, and the positive class weight w_+ is fixed at 1.0. This approach maintains sensitivity to positive predictions while mitigating bias toward the negative class.

Training procedure

We trained the model on 4 H100 GPUs using PyTorch's Distributed Data Parallel (DDP) framework (43) 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 1 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.

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,

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}. \tag{7}$$

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,

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

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:

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN}, \quad \text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}. \tag{9}$$

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,

$$t^* = \arg\max F_1(t). \tag{10}$$

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

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{11}$$

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.

GRN inference with PMF-GRN

We performed GRN inference using our previously published method, PMF-GRN (Probabilistic Matrix Factorization for Gene Regulatory Network inference) (7) 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 (6).

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,

$$\mathrm{GRN}_{\mathrm{Consensus}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{GRN}_i, \tag{12}$$

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.

To evaluate the relationship between posterior uncertainty and predictive accuracy, we rank all TF-gene interaction pairs by their posterior variance, constructing 10 cumulative bins corresponding to the variances in increments of 10%. For each bin k, we use the posterior point estimates of the interactions within the bin and compute the AUPRC using a gold standard set of validated TF-gene interactions $\mathcal{G}$ as the reference. All interactions in each bin are included in the evaluation, regardless of whether they appear in the gold standard.

Let $\mathcal{B}_k$ be the set of posterior point estimates in the k% of variances. Then the AUPRC for bin k is given by:

$$\mathrm{AUPRC}_k = \mathrm{AUPRC}(\mathcal{B}_k; \mathcal{G}), \tag{13}$$

where the AUPRC is computed by comparing the predicted scores in $\mathcal{B}_k$ to labels derived from $\mathcal{G}$. The cumulative binning strategy ensures that each bin contains a sufficient number of interactions for stable AUPRC estimation.

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.

Author Contributions

CSG, AC, and KC jointly contributed to the conceptualization, design, and implementation of the project. CSG carried out the GLM-Prior experiments, dataset curation, validation, formal analysis, and visualizations. CSG and KC prepared the original draft. KC and RB provided supervision, project administration, and funding acquisition.

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) (28) and GSE144820 (Y2) (29); Human embryonic stem cells (derived from BEELINE (36)) GSE75748 (42); Mouse embryonic stem cells (derived from BEELINE (36)) GSE98664 (44). The YEASTRACT prior-knowledge was derived from the YEASTRACT database (11). The gold standard for the yeast datasets was obtained from (26). The prior-knowledge matrices created from the BEELINE benchmarks (36) can be found here: (link Zenodo). Models for experiments from yeast, human and mouse can be found at https://huggingface.co/cskokgibbs as well as their corresponding tokenized datasets.

Competing Interests

The authors declare that they have no competing interests.

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] Mark WEJ Fiers, Liesbeth Minnoye, Sara Aibar, Carmen Bravo González-Blas, Zeynep Kalender Atak, and Stein Aerts. Mapping gene regulatory networks from single-cell omics data. *Briefings in functional genomics*, 17(4):246–254, 2018.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Bram Van de Sande, Christopher Flerin, Kristofer Davie, Maxime De Waegeneer, Gert Hulselmans, Sara Aibar, Ruth Seurinck, Wouter Saelens, Robrecht Cannoodt, Quentin Rouchon, et al. A scalable scenic workflow for single-cell gene regulatory network analysis. *Nature protocols*, 15(7):2247–2276, 2020.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Sara Barbosa, Bastian Niebel, Sebastian Wolf, Klaus Mauch, and Ralf Takors. A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints. *Biosystems*, 174:37–48, 2018.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Micaela E Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J Theis, Alan Moses, and Bo Wang. Transformers and genome language models. *Nature Machine Intelligence*, pages 1–17, 2025.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Lingxiang Zhu, Gary D Swergold, and Michael F Seldin. Examination of sequence homology between human chromosome 20 and the mouse genome: intense conservation of many genomic elements. *Human genetics*, 113:60–70, 2003.

[38] Yong Cheng, Zhihai Ma, Bong-Hyun Kim, Weisheng Wu, Philip Cayting, Alan P Boyle, Vasavi Sundaram, Xiaoyun Xing, Nergiz Dogan, Jingjing Li, et al. Principles of regulatory information conservation between mouse and human. *Nature*, 515(7527):371–375, 2014.

[39] Soo Bin Kwon and Jason Ernst. Learning a genome-wide score of human–mouse conservation at the functional genomics level. *Nature communications*, 12(1):2495, 2021.

[40] Candice J Coppola, Ryne C. Ramaker, and Eric M Mendenhall. Identification and function of enhancers in the human genome. *Human molecular genetics*, 25(R2):R190–R197, 2016.

[41] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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 Single-Species Experiments

A.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 (27) under *S. cerevisiae*.

Next, to pair these gene and TF nucleotide sequences with interaction labels, we used the YEASTRACT database of interactions (11) (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 negative class downsampling rate. 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 (26).

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 (26))	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).

Prior-knowledge datasets for CellOracle and Inferelator-Prior were obtained from (6) 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 (11). The reference gold standard was taken from (26) without further modification.

Single cell gene expression datasets were obtained from GSE125162 (38,225 cells by 6,763 genes) (28) and GSE144820 (6,118 cells by 6,763 genes) (29) 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 (7) and comparable methods such as the Inferelator 3.0 (6) and CellOracle (8) across datasets Y1, Y2, and the combined data (Table 2).

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 (26).

A.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 (27), under *H. sapiens*.

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.

Next, to pair these gene and TF nucleotide sequences with interaction labels, we followed the BEELINE benchmarking protocol (36), obtaining labels from their human reference datasets from the STRING (30; 31; 32; 33) and TRRUST (34; 35) 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.

After prior-knowledge generation, GRN inference was run on six timepoint hESC single-cell expression datasets (GSE75748) (42), 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.

A.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 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 A1A, 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.

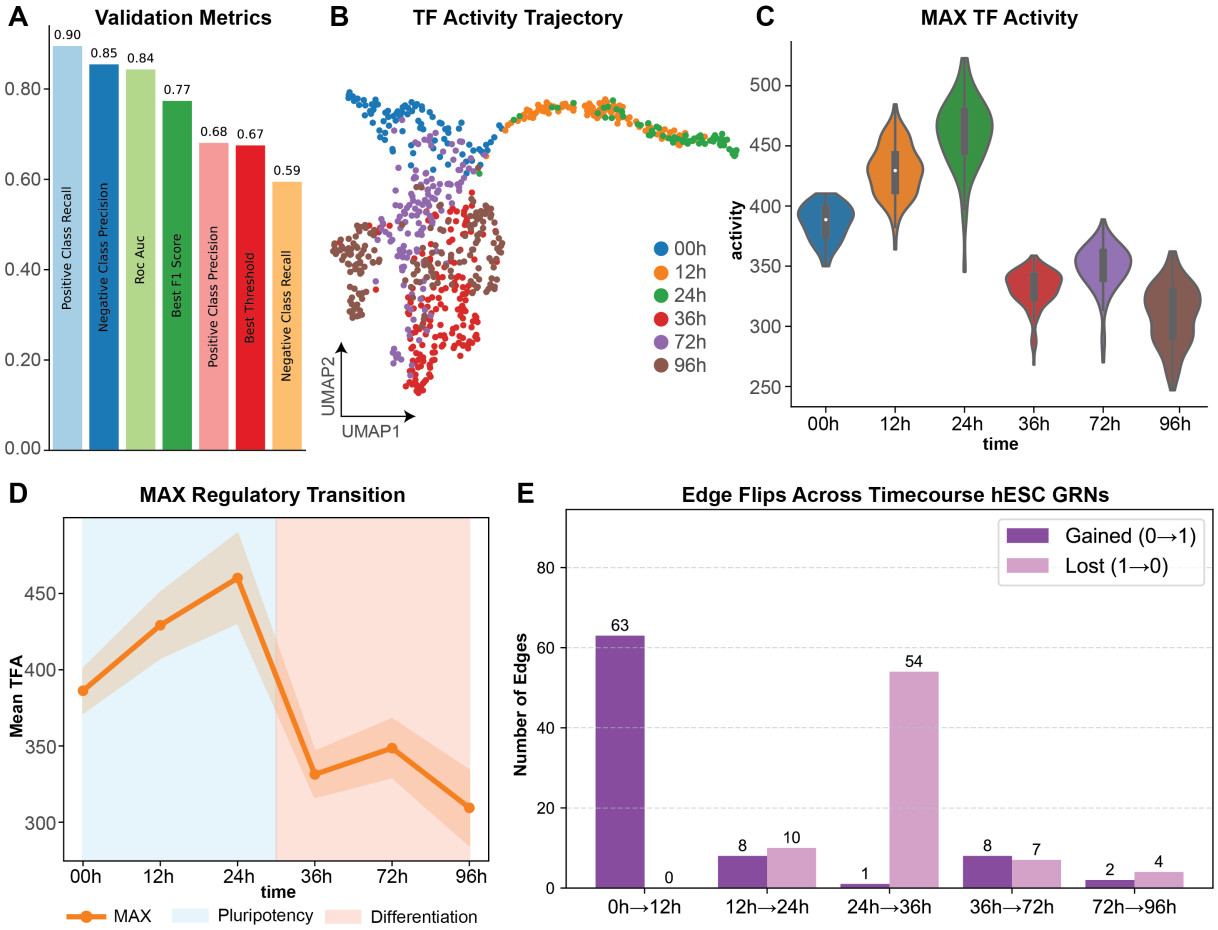

Figure A1: Performance and interpretability of the dual-stage GLM-Prior and PMF-GRN pipeline applied to human ESCs. (**A**) Validation metrics for GLM-Prior trained on human TF-gene interactions from STRING and TRRUST. (**B**) UMAP projection of TFA reveals temporal structure and bifurcating developmental 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.)

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 A1B). 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 (45).

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 its

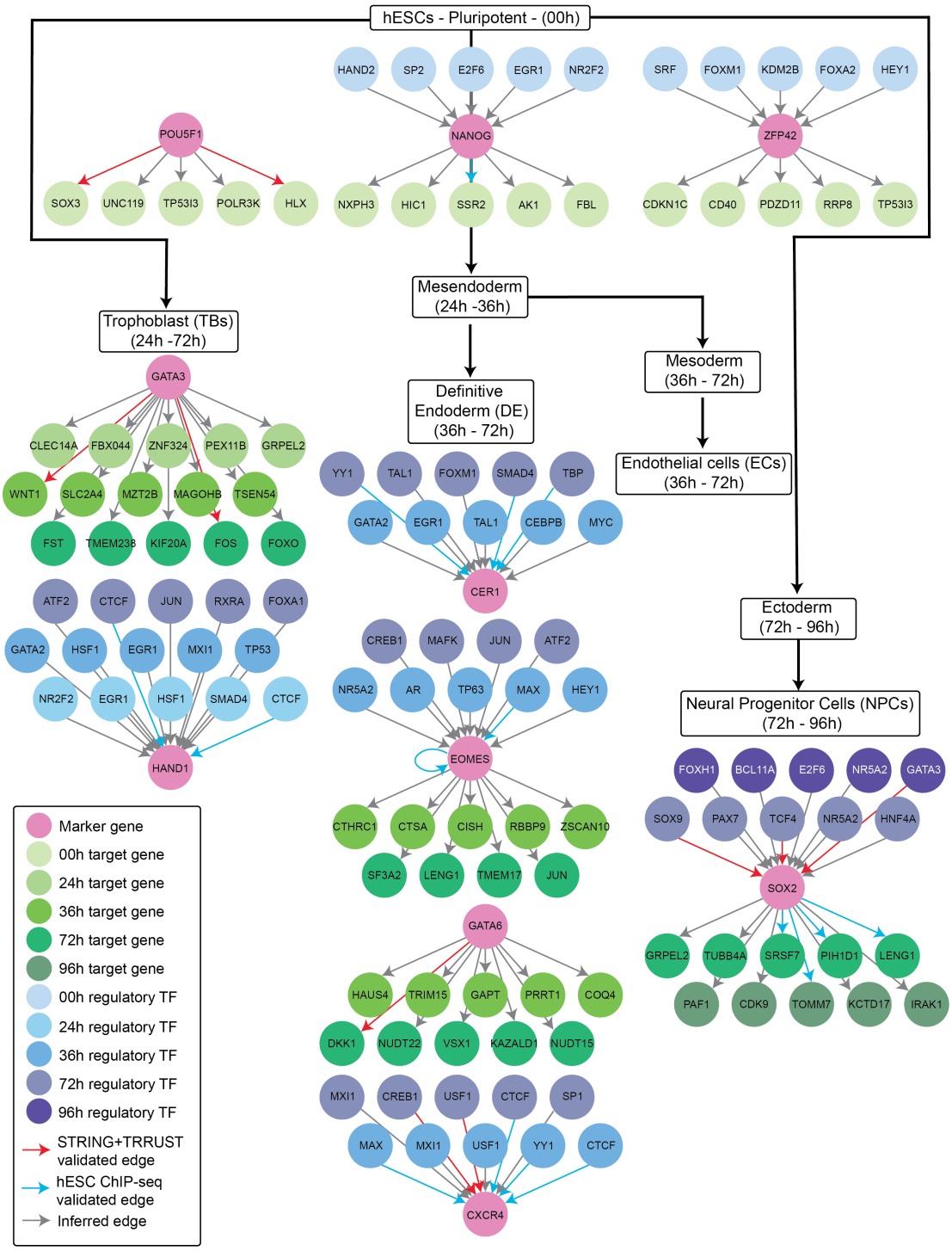

Figure A2: Marker gene GRN visualizations revealing stage-specific regulatory programs across hESC differentiation. Compact GRNs highlight top five regulators (TFs) (blue) and/or targets (genes) (green) of different marker genes (pink) across developmental trajectories in hESCs. Color gradient illustrates 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.

activity distribution over time (Appendix Figure A1C). 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 A1D).

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 (46). 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 A1E). 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 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 (42). 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 A2).

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.

A.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 (27), under *M. musculus*.

To pair gene and TF nucleotide sequences with interaction labels, we followed the BEELINE benchmarking protocol (36), using mouse reference datasets from STRING (30; 31; 32; 33) and TRRUST (34; 35) 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.

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.

Using the predicted prior matrix, we performed GRN inference on five timepoint-specific mESC single-cell expression datasets from GSE98664 (44), 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.

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.

A.5 Single-species training runtime and batch statistics

Training batch composition statistics for each species-specific GLM-Prior model are summarized in Appendix Figure A3A. 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.

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 A3: 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.

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.

Appendix Figure A3A 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.

B Transfer-Learning Experiments

Transferring knowledge from human to mouse

To transfer knowledge from human to mouse, we took the single-species trained human model described in Section A.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.

Transferring knowledge from mouse to human

To transfer knowledge from human to mouse, we took the single-species trained mouse model described in Section A.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.

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 A.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 A.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 (26), 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.

C Multi-Species Experiments

To train a multi-species model, we took the single-species trained human model described in Appendix A.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 (Appendix A.4). Following this, we then finetuned the model one final time using the yeast gene (n=5,999) and TF (n=254) nucleotide sequences associated with the YEASTRACT database, using the optimal class-weights and downsampling rate obtained in the yeast single-species model (Appendix A.1).

Following the consecutive training of the multi-species model, we ran inference in human, mouse and yeast respectively. For human inference, we again used the gene and TF sequences associated with hESC (described in Appendix A.2). For mouse inference, we again used the gene and TF sequences associated with mESC (described in Appendix A.4). For yeast inference, we again used the gene and TF sequences associated with YEASTRACT (described in Appendix A.1). Results for this multi-species model inference can be found in Table 8.

AUPRC Results for Multi-Species Model Inference

Inference Dataset	AUPRC
Human	0.20
Mouse	0.22
Yeast	0.34

Table 8: AUPRC scores for multi-species model inference in human, mouse, and yeast. GRNs were inferred using a unified model trained jointly on all three species, and evaluated against species-specific reference networks.