GLM-Prior: a genomic language model for transferable sequence-derived priors in gene regulatory network inference

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–target gene interactions directly from nucleotide sequence. We integrate GLM-Prior with PMF-GRN, a probabilistic matrix factorization model, to create a dual-stage pipeline that combines sequence-derived priors with single-cell gene expression data for GRN inference. Across six human, mouse, and yeast cell lines, GLM-Prior performance scales with positive label abundance and diverse transcription factor coverage, achieving strong accuracy in well-annotated mammalian contexts. We evaluate single-species, species-transfer learning, and multi-species training paradigms and show that GLM-Prior generalizes to held-out gene and TF sequences, enabling experiment-agnostic prior construction in previously unprofiled contexts. Furthermore, comparisons to accessibility-based priors across multiple GRN inference methods show that GLM-Prior provides the most robust priors in mammalian cell lines. Together, our results demonstrate that prior construction, rather than the choice of GRN inference algorithm, is the primary determinant of GRN inference performance, and establish GLM-Prior as a framework for building high-quality, experiment-agnostic priors that can be deployed even in understudied or experimentally inaccessible systems.

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).

*Corresponding author.

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 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). Further, these existing approaches cannot create prior knowledge matrices which generalize to other species. 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.

We investigate the generalizability of GLM-Prior by evaluating three training paradigms: (1) single-species models trained and evaluated on the same organism; (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. Across these settings, performance scales with abundance and diversity of positive class labels and TF coverage in the training data. Human and mouse cell lines with rich label sets yield higher AUPRCs than yeast, where extreme class imbalance and sparse labels limit generalization. Transfer learning between evolutionarily related species (human and mouse) maintains or modestly improves performance relative to single-species models, demonstrating that GLM-Prior captures conserved regulatory sequence features that can be reused across lineages. In contrast, transfer learning between mammals and yeast does not exceed near-chance performance, reflecting a deeper divergence in regulatory grammar. Multi-species training on human, mouse, and yeast preserves accuracy comparable to the best single-species or transfer models in most cell lines, with only mild trade-offs in

some human contexts, and produces a domain-stable prior that can be applied to new systems without species-specific retraining.

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.

Building on this, we benchmark GLM-Prior against accessibility-based priors (Inferelator-Prior (6), and CellOracle's baseGRN (8)), and integrate each prior with it's downstream GRN inference algorithm (PMF-GRN (7), the Inferelator 3.0 (6), and CellOracle (8)) across yeast, mouse, and human. GLM-Prior consistently performs at or above chance in all six cell lines and provides the strongest prior in four of six contexts (hESC, mESC, mDC, and mHSC). In contrast, accessibility-based priors remain competitive in settings dominated by proximal regulatory logic, such as yeast, but struggle to generalize above chance performance in complex mammalian cell lines. When these priors are passed into downstream GRN inference, GRN inference typically yields modest gains when the prior is already strong, and larger improvements are observed only when the prior is weak or inaccurate. Critically, the identity of the best-performing GRN in each cell line almost always matches the identity of the best prior, and no GRN inference method is uniformly superior across priors or species. Together, these results indicate that prior construction, rather than the choice of GRN inference algorithm, is the primary driver of GRN reconstruction performance.

Together, our findings recast GRN inference as a problem primarily limited by the quality of prior knowledge rather than by expression-based modeling. GLM-Prior provides a scalable, species-aware framework for constructing priors, while downstream GRN methods act mainly to refine or contextualize this scaffold. In this work, we characterize GLM-Prior's generalization across cell types and species, quantify its advantages over accessibility-based priors, and dissect how different GRN inference algorithms interact with these priors to shape the final inferred networks.

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 positive regulatory (1) and negative 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 positive or negative 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, selecting the threshold that maximizes the F1 score on the validation set. This optimal threshold is used to binarize the predicted probabilities into positive or negative interactions for the final prior knowledge matrix.

Since GRN datasets are often highly imbalanced, with far fewer positive interactions compared to negative 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.

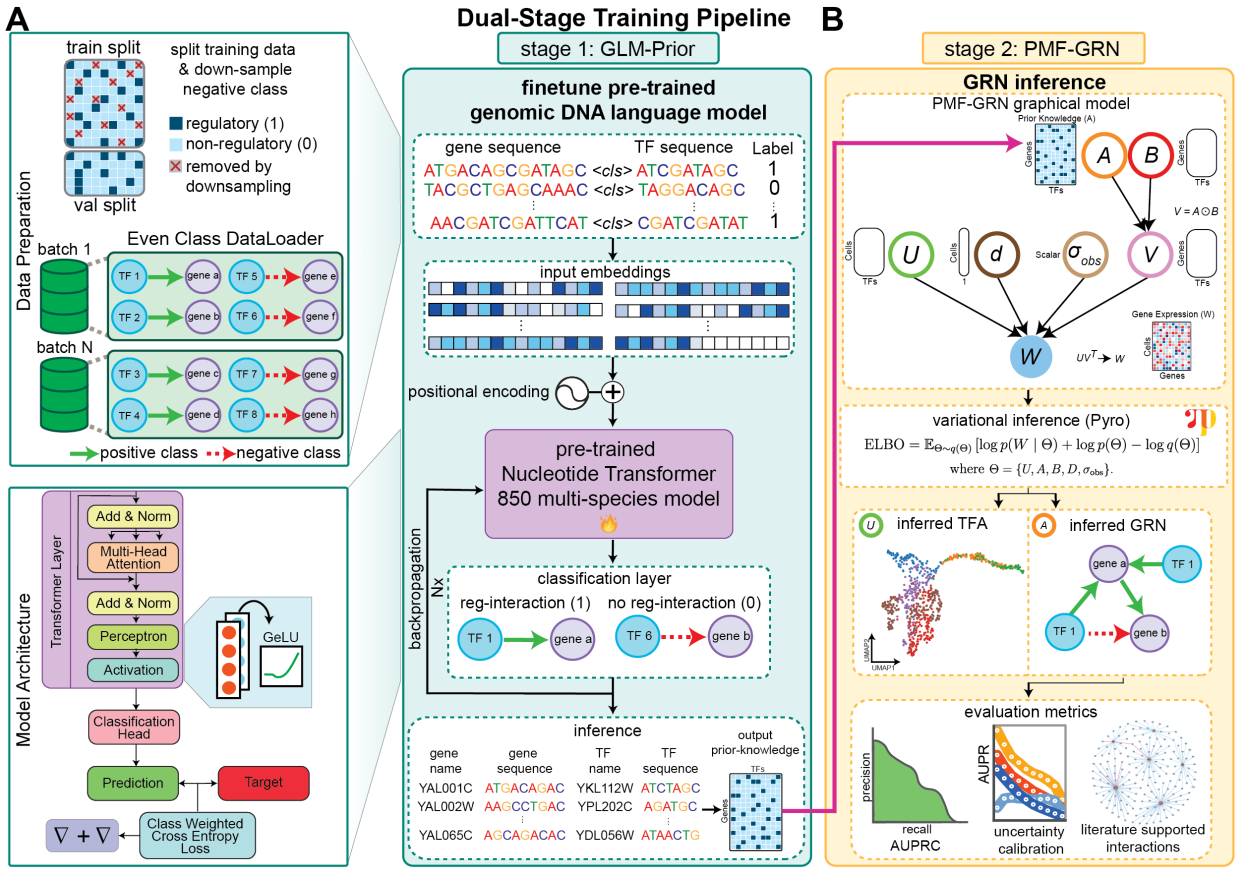

Figure 1: Schematic of the dual-stage training pipeline. (**A**) GLM-Prior is trained on concatenated TF binding motifs and gene sequences labeled as interacting or non-interacting. Input data is split into training and validation sets, where the training set is downsampled and placed into even class 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.

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. Although we can use the probabilities directly, we binarize the prior knowledge matrix to be in line with existing methods and for more direct comparison. We then either use this sequence-based

prior knowledge as is and evaluate on an independent gold standard, or can pass it into the second stage of the pipeline, enabling a modular information transfer from model-driven sequence inference to data-driven GRN inference.

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 knowledge to enhance the inference of GRNs from single-cell gene expression data. It does so by using probabilistic matrix factorization, which decomposes the observed gene expression into latent factors representing transcription factor activity (TFA) and TF-target gene regulatory relationships.

Concretely, PMF-GRN models the expression matrix W using the likelihood function $p(W|U, V = A \odot B, \sigma_{obs}, d)$, where $\mathbb{E}[W] \approx d \odot UV^\top$, with per-cell sequencing depth d and observation noise σ_{obs}. Here, U is a nonnegative cells by TFs matrix of TF activities, and V is a genes by TFs matrix, which is further factorized as $V = A \odot B$. $A \in (0, 1)$ is defined as the degree of existence of a TF-target gene interaction and $B \in \mathbb{R}$ is its signed effect. Each latent variable, $U, A, B, d,, \sigma_{obs}$ is assumed independent *a priori*. In this formulation, the GLM-Prior predicted prior knowledge matrix provides the hyperparameters for the logistic-normal prior on A, anchoring factors to specific TFs and mitigating column-label identifiability inherent to matrix factorization.

Using variational inference, PMF-GRN estimates the posterior distribution of all latent variables. The posterior mean of A defines the inferred GRN, while the posterior variance provides an edge-level uncertainty score. These uncertainty estimates enable confidence-weighted interpretation of TF-gene edges and are especially valuable when gold standard regulatory annotations are incomplete or unavailable.

The combination of GLM-Prior and PMF-GRN is motivated by their complementary strengths. GLM-Prior uses nucleotide sequence to construct a biologically grounded scaffold of regulatory edges, capturing both local motif-driven interactions and potential long-range dependencies. This scaffold can be generated as a general prior matrix or tailored to specific cell lines or conditions. However, as GLM-Prior has not seen any experimental data, its predictions require refinement using real observations. PMF-GRN provides this refinement by integrating single-cell expression data within a principled generative framework, updating edge probabilities based on observed regulatory patterns while maintaining consistency with the sequence-derived prior. Moreover, like most GRN inference algorithms, PMF-GRN relies on prior knowledge to obtain a meaningful posterior GRN; without an informative prior, factor identifiability issues limit interpretability. GLM-Prior resolves this limitation by anchoring TF-gene relationships, while PMF-GRN tailors the scaffold to the data, quantifies uncertainty, and produces context-specific networks. Together, this integration balances predictive power from large-scale genomic language models with the interpretability and flexibility of probabilistic modeling, yielding GRNs that are both grounded in sequence evidence and informed by experimental observations.

GLM-Prior's generalization scales with data composition across species and cell types

We assess GLM-Prior by benchmarking across six cell lines spanning yeast, human, and mouse (Figure 2). For each organism, we create gene-TF sequence pairs by extracting nucleotide sequences for genes from GTF annotations and TF binding motifs from CisBP (26). For each gene-TF pair, we gather available regulatory labels from the YEASTRACT database (11) for yeast, and the STRING (27; 28; 29; 30) and TRRUST (31; 32) databases for human and mouse. To assess out-of-distribution performance, we evaluate on held-out test sets constructed from well-established reference labels such as the literature curated gold standard from (33) for yeast, as well as cell-line specific reference networks from BEELINE (34) for human and mouse. Before training each cell-line specific model, we ensure strict independence between training and test sets at the level of gene-TF sequence pairs and their labels, eliminating sequence and label leakage. This design ensures that performance reflects the model's ability to generalize to unseen sequences and correctly predict their regulatory interactions.

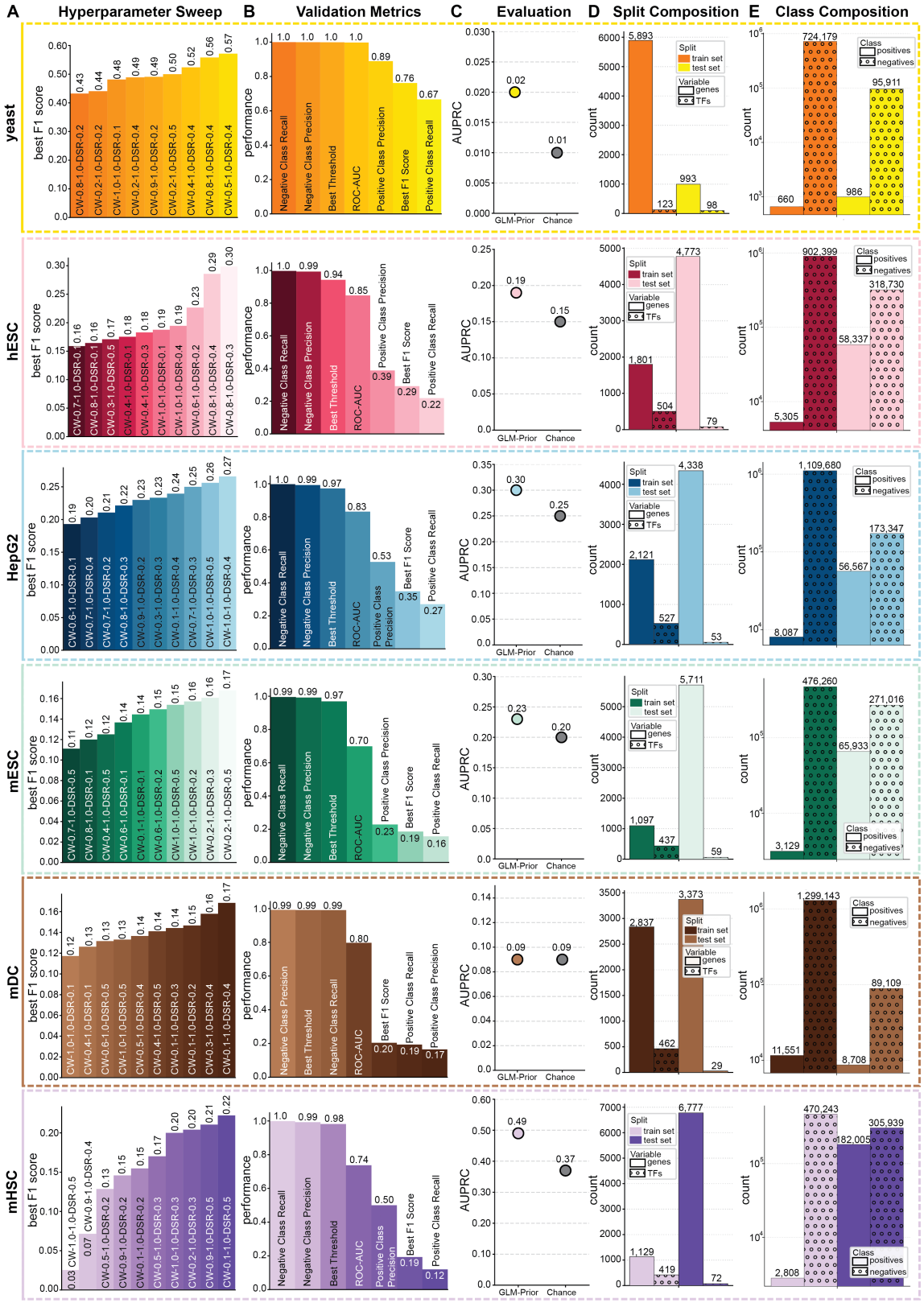

Figure 2: GLM-Prior model performance across yeast, hESC, HepG2, mESC, mDC, and mHSC cell lines. (**A**) One-epoch hyperparameter sweep over class-weights and negative-class downsampling rates, evaluated by F1-score. (**B**) Validation metrics during 10 epoch training using the best hyperparameter configuration. (**C**) Test AUPRC on held-out gold standards (with chance performance shown for reference in gray). (**D**) Train and test split composition (number of genes and TFs). (**E**) Class composition (number of positive and negative labels) in train and test sets.

To evaluate model performance within each cell line, we implement a multi-stage benchmarking procedure (Figure 2A-E). In Figure 2A, we perform a one-epoch hyperparameter sweep over combinations of class weights in the loss function and downsampling rates for the negative class. We identify the configuration that yields the highest F1-score and use these parameters for full model training. In Figure 2B, we train the model for ten epochs using this optimal configuration and monitor performance on a held-out validation set. Following training, we apply the model to all gene-TF pairs, including those unseen during training, and evaluate our predictions against the corresponding held-out set's gold standard labels (Figure 2C). To interpret what drives the resulting performance, we examine the number of genes and TFs represented in the training and test sets for each cell line (Figure 2D), providing context for possible domain shifts between splits. Finally, we quantify the class composition of positive and negative examples in the training and test sets (Figure 2E), allowing us to relate generalization behavior to the underlying data balance.

Across the six evaluated cell lines, GLM-Prior demonstrates variation in predictive performance, highlighting the influence of data composition, particularly the number of positive regulatory interactions, on generalization (Figure 2E). In yeast, GLM-Prior achieves a near-chance AUPRC of 0.02 compared to a 0.01 baseline, despite strong within-training metrics (Figure 2B). This limited generalization likely stems from the extreme class imbalance in the training data, where only 660 positive interactions are available among $724,179$ negatives. Although the test set contains a moderate number of positives (986), the small number of training examples and high-performing validation metrics suggest that the model overfits the limited regulatory patterns seen during training and fails to extrapolate to novel gene-TF combinations.

In human embryonic stem cells (hESC), performance improves modestly, with an AUPRC of 0.19 versus a 0.15 chance baseline. Here, GLM-Prior benefits from a larger pool of positive training examples $(5,305)$ and a more balanced test set ($53,337$ positives and $318,730$ negatives). However, TF coverage drops from 504 in training to 79 in testing, implying that while model generalizes across genes, reduced TF diversity in the test set limits broader generalization.

In HepG2 hepatocytes, GLM-Prior achieves a stronger performance (AUPRC = 0.30 versus 0.25 chance) supported by both a higher number of positive training examples $(8,087)$ and a well-represented test set ($56,567$ positives). Despite the test set containing fewer TFs (53 versus 527 in training), the model still generalizes effectively, indicating that it learns transferable sequence features that extend across TFs in this cell line. This performance, coupled with the relatively large and balanced training data, suggest that sufficient label diversity allows the model to learn regulatory sequence logic that is not confined to specific TF identities.

In mouse embryonic stem cells (mESC), GLM-Prior achieves an AUPRC of 0.23 versus 0.20 chance, showing moderate but consistent generalization. The training set includes $3,129$ positives among $476,260$ negatives while the test set is considerably more balanced ($65,933$ positives, $271,015$ negatives), covering $5,711$ genes unseen during training. The broad gene coverage supports cross-gene generalization, though limited positive training examples likely cap performance relative to the human cell lines.

In mouse dendritic cells (mDC), performance remains at chance (AUPRC = 0.09, despite a substantial number of training examples $11,551$). The sharp reduction in TF coverage from 462 in training to only 29 in testing suggests that the test TFs may regulate distinct, cell-type specific programs absent from training.

Finally, mouse hematopoietic stem cells (mHSC) exhibit the strongest generalization (AUPRC = 0.49 versus 0.37 chance). Although the training set contains a modest number of positives $(2,808)$, the test set is both large ($182,005$ positives) and diverse ($6,777$ genes, 72 TFs), providing a robust evaluation. The high AUPRC gain indicates that GLM-Prior captures sequence-level regulatory features that transfer effectively to unseen gene-TF pairs in this cell line.

Together, these results demonstrate diverse composition of gene-TF relationships in training data is a critical factor behind GLM-Prior's generalization. In particular, generalization is most affected by the abundance and diversity of positive interactions and TFs in training. We interpret the above chance performance gains as the model's ability to learn transferable sequence regulatory patterns when sufficient label diversity is available.

GLM-Prior supports successful species-transfer learning and multi-species training

Following our six cell line evaluation of GLM-Prior, we next task the model to predict prior knowledge in a species unseen during training (Figure 3). Building on the previous section, which established that generalization depends strongly on data composition, we now ask whether the model can transfer learned regulatory logic across species. Such cross-species transfer would represent a major advantage over existing approaches, enabling the construction of informative priors for understudied organisms using models trained on evolutionarily similar species. To test this, we perform two complementary experiments: A) transfer learning, where models trained in one species are applied to another, and B) multi-species training, where a single model is trained sequentially on human, mouse, and yeast data. In all cases, the held-out evaluation sets are fully disjoint from training data, and overlapping gene and TF identifiers are removed to prevent cross-species leakage.

In Figure 3A, we provide a schematic for the transfer learning setup. Three transfer learning configurations are evaluated, (*i*) a model trained jointly on human and mouse data, evaluated on yeast, (*ii*) a mouse-only model evaluated on human cell lines (hESC and HepG2), and (*iii*) a human-only model evaluated on mouse cell lines (mESC, mDC, and mHSC). Across all six cell lines, we find that transfer learning maintains or modestly improves performance relative to single-species models.

In yeast, performance remains near chance (AUPRC = 0.02 vs 0.01), consistent with deep evolutionary divergence between eukaryotic and mammalian regulatory syntax. In contrast, cross-species transfer between human and mouse demonstrates clear success, where hESC maintains an AUPRC = 0.19 and HepG2 improves to 0.31. Simiarly, human-trained models applied to mouse cell lines achieve stable performance comparable to single-species results. These results demonstrate that GLM-Prior captures conserved sequence-level features that transfer across closely related species, enabling accurate prior construction without species-specific retraining.

In Figure 3B, we present the multi-species model trained sequentially on human, mouse, and yeast data and evaluated on all six cell lines. The multi-species model achieves comparable AUPRCs relative to both single-species and species-transfer settings (yeast = 0.02, hESC = 0.19, HepG2 = 0.29, mESC = 0.23, mDC = 0.09, and mHSC = 0.49).

To interpret these trends, Figure 3C-D summarize the training compositions and normalized AUPRC (Figure 3E) across single, transfer, and multi-species setups. Across each cell line, transfer learning consistently expands the number of genes, TFs, and positive labels seen during training. In some settings, such as with HepG2, this broader coverage allows for a modest normalized AUPRC increase (0.08 vs 0.07 single-species). In others, transfer learning maintains single-species performance. In contrast, mHSC shows a slight decrease in performance (0.14 species-transfer vs 0.19 single-species) despite larger training breadth. This drop suggests that the regulatory logic active in mouse HSCs is relatively cell-type and species-specific and therefore not well represented in the context-agnostic human interaction data used for species-transfer training, limiting the model's ability to fully recover mHSC-specific edges.

Notably, the multi-species model maintains similar performance to the single-species models with only slight performance drops in hESC and HepG2. Here, we demonstrate that multi-species training yields a robust, domain-stable prior that generalizes across lineages. This provides a practical route to build prior knowledge for poorly understood systems or cell lines when a well-diversified training set is desired, offering broad regulatory coverage without assuming a species specific match.

Together, these results demonstrate that GLM-Prior achieves successful cross-species transfer, a property not attainable with accessibility-based prior knowledge approaches which depend on species-specific data. Applying a model trained in one organism directly to another enables scalable, sequence-based prior construction in under-characterized systems. In parallel, multi-species training yields a robust, lineage-agnostic prior that maintains stable performance across contexts. When the target cell line is poorly characterized or its closest species match is uncertain, the multi-species model presents a practical default for constructing reliable prior knowledge.

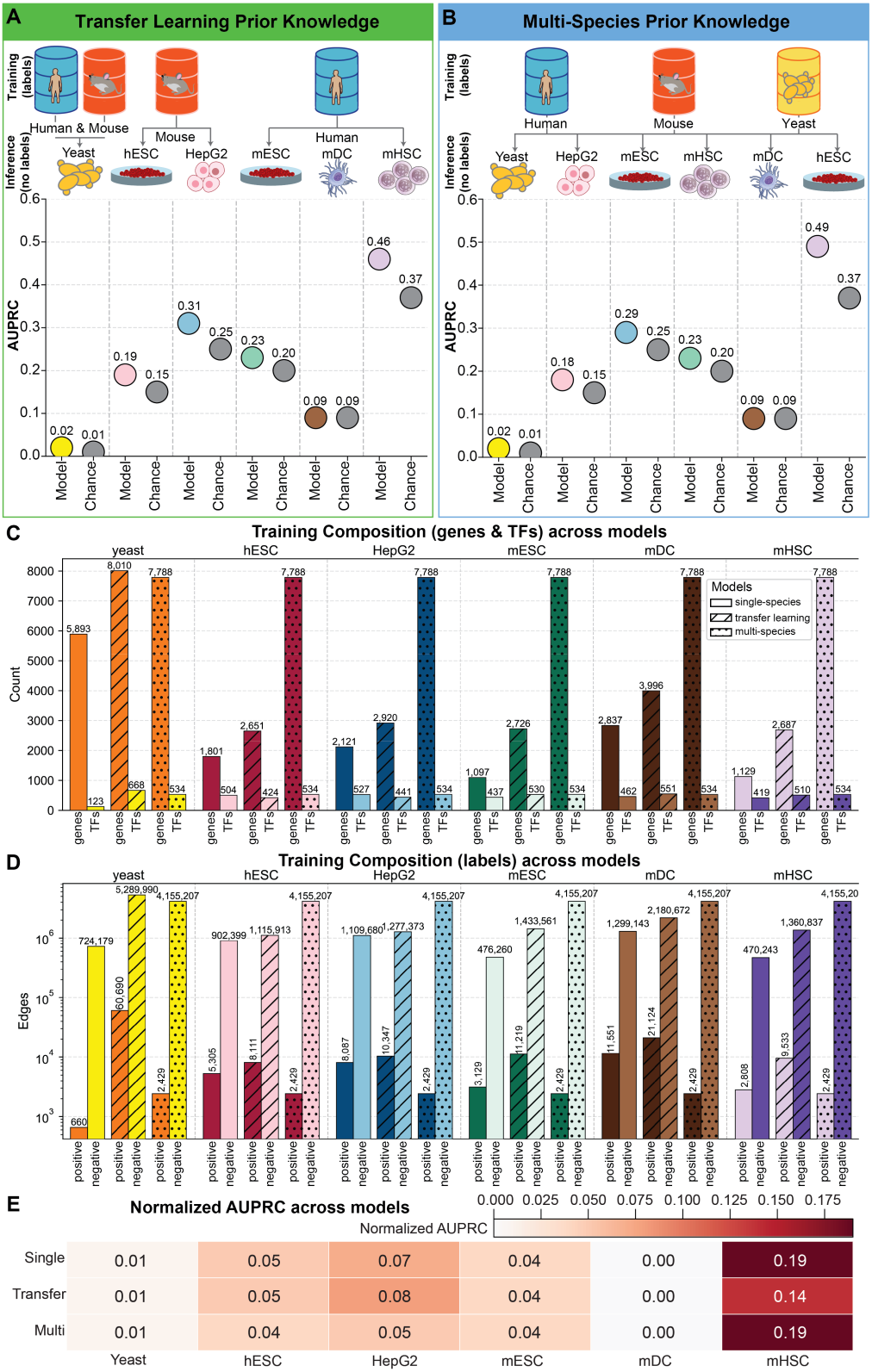

Figure 3: Transfer learning and multi-species training with GLM-Prior. (**A**) Schematic and AUPRC results of the species-transfer learning setup, showing three configurations: (*i*) a model trained jointly on human and mouse data evaluated on yeast, (*ii*) a mouse-only model evaluated on human cell lines, and (*iii*) a human-only model evaluated on mouse cell lines. (**B**) Schematic and AUPRC results of the multi-species model trained sequentially on human, mouse and yeast data, evaluated across six cell lines. (**C**) Comparison of training composition (number of genes and TFs) for single-species, species-transfer, and multi-species models. (**D**) Class composition (positive and negative labels) in the corresponding training sets. (**E**) Normalized AUPRC across single-species, species-transfer, and multi-species models for each cell line.

Integrating GLM-Prior into PMF-GRN provides contextual GRN inference and edge refinement

To evaluate how sequence-derived regulatory prior knowledge interacts with expression-based GRN inference, we next used each cell line-specific GLM-Prior matrix as input to PMF-GRN. Whereas Stage 1 (Section 1) directly predicts TF-gene interactions from sequence, Stage 2 (Section 1) treats these predictions as prior knowledge and refines the network using single cell expression data. For each cell line, we paired GLM-Prior with matched single cell datasets, for example, GSE125162 (35) and GSE144820 (36) for yeast, and the corresponding expression datasets provided in BEELINE (34) for the human and mouse cell lines. This experiment is designed to assess not only whether PMF-GRN improves predictive accuracy, but also how much the network must be altered when inference is initialized from an already information-rich prior.

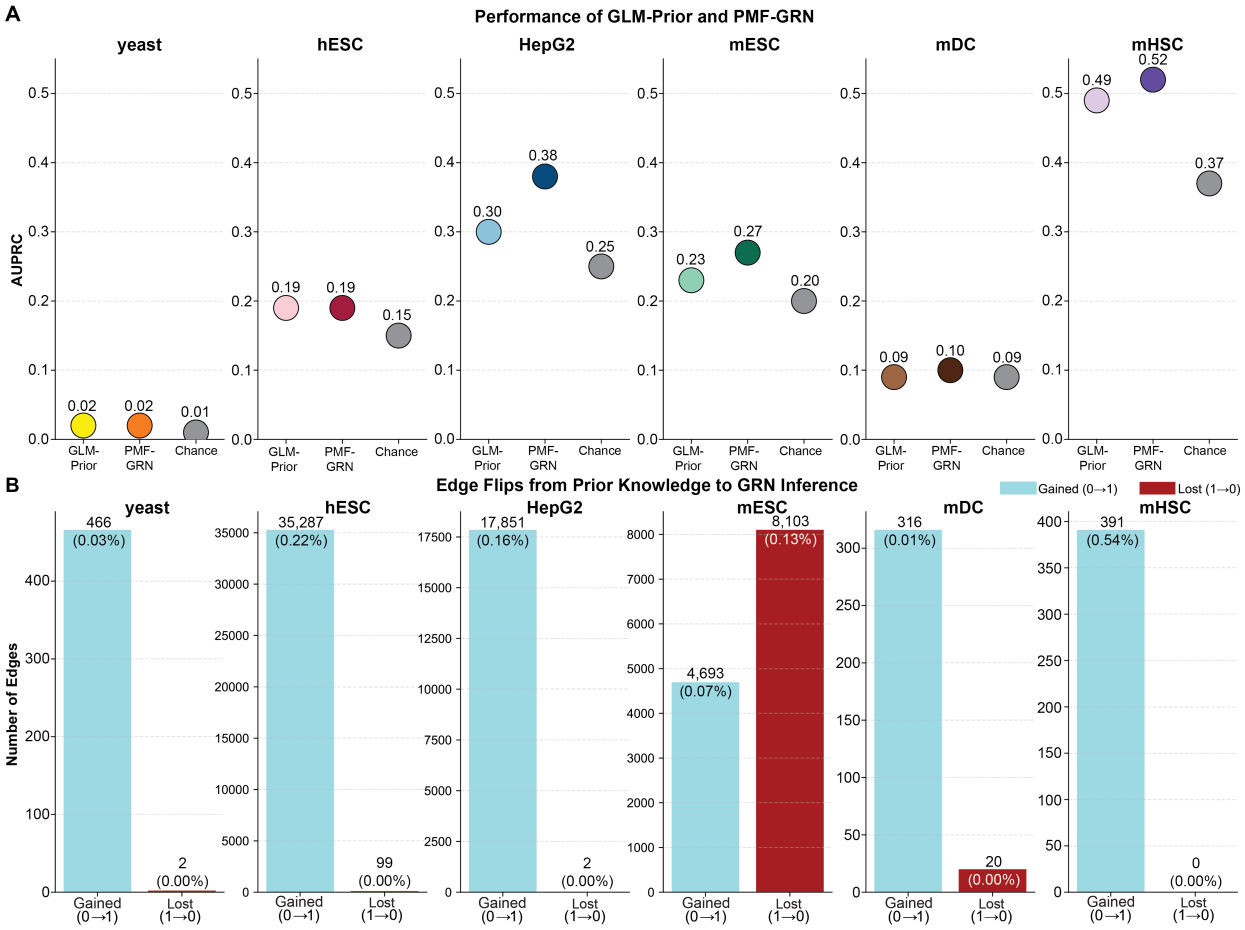

Figure 4: Integration of GLM-Prior with PMF-GRN for downstream GRN inference. (**A**) AUPRC comparison of GLM-Prior alone versus integrated with PMF-GRN across six cell line contexts. (**B**) Edge flip analysis quantifying changes made to the prior knowledge interaction matrix during GRN inference. Blue bars indicate edges gained and red bars indicate edges removed by GRN inference.

Across all datasets, integrating GLM-Prior into PMF-GRN results in equal or improved AUPRC relative to the prior alone (Figure 4A). The magnitude of improvement varies by species, where performance remains unchanged in yeast (0.02) and hESC (0.19), while clear gains are observed in HepG2 (0.30 prior $\rightarrow$ 0.38 GRN), mESC (0.23 prior $\rightarrow$ 0.27 GRN), and mHSC (0.49 prior $\rightarrow$ 0.52 GRN). These improvements occur despite PMF-GRN modifying only a small fraction of the GLM-Prior edge matrix (Figure 4B), with added edges ranging from 0.03% in yeast to 0.54% in mHSC, and edge removals remaining minimal across all contexts. Notably, modifications are strongly biased toward edge additions rather than deletions, indicating

that PMF-GRN treats the GLM-Prior as a stable scaffold and makes targeted expansions rather than large-scale refinement.

Interestingly, the extent of structural changes does not strictly track with performance improvement. For example, hESC and mESC exhibit tens of thousands of added or removed edges, yet show minimal differences in AUPRC. This mismatch suggests that many of the changes introduced by PMF-GRN occur outside the held-out evaluation portion of the network, which only cover a subset of TF-gene interactions. Conversely, in yeast, where GLM-Prior is close to chance (AUPRC = 0.02) and the available training interactions are sparse, PMF-GRN fails to recover meaningful signal despite access to two large single-cell expression datasets, indicating that expression-driven inference alone cannot compensate for a weak or uninformative prior.

These findings indicate that when a sequence-derived prior already captures substantial regulatory signal, the role of expression-based inference shifts from discovery to contextual refinement. Rather than constructing the network from expression patterns alone, PMF-GRN behaves conservatively, preserving high-confidence structure, adding plausible missing edges, and performing limited pruning. In this dual-stage framework, the strength of GLM-Prior largely determines the achievable performance ceiling, while PMF-GRN contributes selective improvements where transcriptomic evidence provides additional support; when the prior is poor, as in yeast, expression-based GRN inference has limited capacity to recover accurate regulatory structure.

Comparing prior construction strategies across yeast, mouse, and human cell lines

Next, we assess how GLM-Prior compares to widely used accessibility-based prior construction approaches, evaluating performance relative to CellOracle and Inferelator priors across six cell line contexts. Unlike GLM-Prior, which leverages sequence-level information to predict TF-gene regulatory edges, both CellOracle and Inferelator derive putative interactions from chromatin accessibility and motif enrichment. Importantly, these accessibility-based priors are constructed using ATAC-seq measured in the matched cell line, whereas GLM-Prior is trained on context-agnostic interaction resources (STRING and TRRUST) and evaluated without explicitly conditioning on cell-line specific chromatin state. This comparison therefore evaluates whether sequence-derived priors can match or exceed the performance of accessibility-based approaches across diverse biological settings.

Across all six cell lines, GLM-Prior performs at or above chance, exceeding the chance accuracy in five of the six contexts, demonstrating robust generalization across species and cell types (Figure 5A). In comparison, Inferelator-Prior performs above chance in only two out of six cell lines, and CellOracle's Prior achieves above-chance performance in three out of six, indicating less consistent cross-context behavior. Yeast represents the main exception, where CellOracle produces the highest-performing prior (AUPRC 0.05 versus 0.03 for Inferelator-Prior and 0.02 for GLM-Prior), highlighting the strengths of accessibility-based priors in simple eukaryotic systems where regulatory interactions are predominantly promoter-proximal and well captured by open chromatin-based assignment. In this setting, GLM-Prior is further disadvantaged by the limited number of positive regulatory interactions (660) available for model training (as seen in Figure 3D).

A broader comparison across species using normalized AUPRC values further illustrates these patterns (5B), where normalized AUPRC is defined as (AUPRC−chance)/(1−chance). While GLM-Prior provides the most consistent gains overall, CellOracle's prior outperforms all other approaches in HepG2 (0.36 versus 0.30 for GLM-Prior and 0.24 for Inferelator-Prior). This comparatively strong performance in a mammalian cell line can be explained by alignment between the construction of CellOracle's prior and the BEELINE reference network used for evaluation. BEELINE assembles reference networks using ChIP-seq supported TF-gene edges curated from ENCODE (10), ChIP-ATLAS (37) and ESCAPE (38) databases, where HepG2 is among the most densely profiled ENCODE cell line. As a result, the HepG2 reference likely contains a large and internally consistent set of TF occupancy-supported edges that are recoverable by accessibility-conditioned, proximity-based linking, precisely the evidence CellOracle prioritize when constructing its prior from motifs in accessible regions. By contrast, Inferelator-Prior follows a similar accessibility and motif-based strategy but imposes stronger sparsity which may remove a substantial number of true TF-gene edges in this densely profiled HepG2 context and thereby reduce its alignment with the BEELINE reference network.

In contrast, GLM-Prior is the top-performing method in the four remaining cell lines: hESC, mESC, mDC, and mHSC. In hESC, all methods exceed chance, with GLM-Prior achieving the highest AUPRC (0.19

versus 0.17 for Inferelator-Prior and 0.16 for CellOracle's prior). In mESC and mHSC, the advantage of GLM-Prior becomes more pronounced, as accessibility-based priors fall below chance while GLM-Prior remains predictive. These results suggest that in several mammalian contexts, sequence-based priors may better capture distal enhancer regulation and long-range TF-gene interactions that proximity-based methods miss.

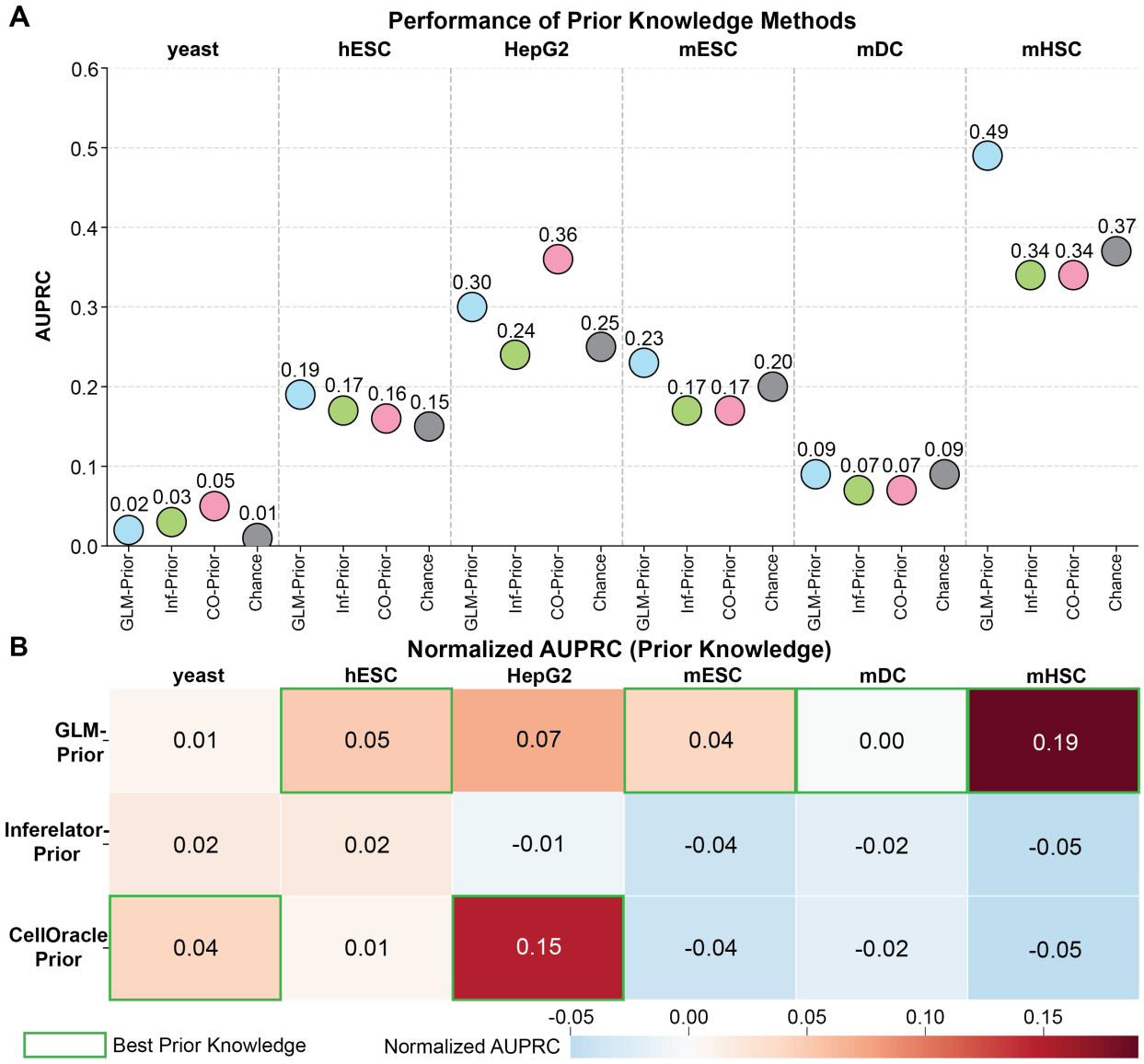

Figure 5: Comparison of prior construction strategies across species. (**A**) AUPRC of GLM-Prior, Inferelator-Prior and CellOracle Prior across six cell lines, with chance performance (gray) provided for each context. (**B**) Normalized AUPRC values with the best-performing prior (green box) highlighted in each cell line.

These comparisons indicate that accessibility-driven approaches such as CellOracle remain highly effective in contexts dominated by proximal regulatory logic, whereas GLM-Prior provides a more robust and transferable strategy across diverse mammalian systems. The complementary performance profiles across species highlight that no single prior construction approach is universally optimal; instead, GLM-Prior offers a scalable, sequencing technology-independent alternative that is better suited to capturing complex long-range regulatory architecture and transferring information across closely related species and cell lines.

Paired prior-GRN strategies reveal prior-dependent gains from expression-based inference

Following our comparative analysis of prior construction approaches, we next asked how each complete prior-GRN pipeline performs across the six cell lines (Figure 6). GLM-Prior and PMF-GRN form our dual-stage pipeline, combining a finetuned sequence language model for prior construction with a probabilistic matrix factorization model for GRN inference from single cell expression. CellOracle constructs a prior ("CellOracle baseGRN") using motif enrichment within accessible promoter regions and pairs it with a Bayesian ridge regression model for GRN inference. Similarly, Inferelator-Prior uses chromatin accessibility and TF motif enrichment to build a prior edge matrix that is then refined using regularized regression during GRN inference. By comparing these three paired pipelines, GLM-Prior + PMF-GRN, Inferelator-Prior + Inferelator, and CellOracle prior + CellOracle, we can ask two related questions: (*i*) how well each method converts its own prior into a high-performing GRN, and (*ii*) to what extent expression-based GRN inference can compensate for poor priors versus refining already strong ones.

To interpret these patterns, we consider the absolute AUPRCs in Figure 6A together with the corresponding ΔAUPRC values in Figure 6B. In yeast and mDC, all priors are at or only slightly above chance, and GRN inference yields minimal gains ($\Delta\text{AUPRC} \leq 0.02$), indicating that expression-based models struggle to recover meaningful signal when the starting network is near-random. In hESC and mHSC, GLM-Prior already achieves above-chance performance, and PMF-GRN adds only small improvements, whereas Inferelator and CellOracle modestly refine their weaker priors. These contexts illustrate that when priors are already strong, GRN inference primarily fine-tunes edge weights rather than substantially reshaping the network.

By contrast, HepG2 and mESC are the settings where GRN inference provides the largest benefits. Here, priors span a range from near- or slightly below-chance (Inferelator-Prior in HepG2; Inferelator-Prior and CellOracle prior in mESC) to moderately above-chance (GLM-Prior and CellOracle prior in HepG2; GLM-Prior in mESC), leaving partially specified structure that expression data can refine. In HepG2, GLM-Prior + PMF-GRN and Inferelator-Prior + Inferelator show sizeable increases in AUPRC, while CellOracle begins from the strongest prior and improves only modestly. In mESC, regression-based inference on top of weak or near-chance priors yields the single largest observed ΔAUPRC for CellOracle and substantial gains for Inferelator, whereas PMF-GRN makes only a small but consistent improvement over an already above-chance GLM-Prior. Taken together, these six cell lines reveal a coherent pattern: the magnitude of GRN-inference gains is strongly prior-dependent, with the largest boosts arising when priors are of intermediate quality (informative but incomplete), limited performance improvements when priors are very weak, and only modest refinements when priors are already strong.

We next normalize AUPRC by the chance level in each species (as in Section 1) to directly compare performance above (or below) chance across methods and cell lines (Figure 6C). In this view, GLM-Prior provides the best prior in four of six cell lines (hESC, mESC, mDC, mHSC), while CellOracle prior yields the strongest priors in yeast and HepG2. We observe that in four mammalian cell lines (HepG2, mESC, mDC, and mHSC), Inferelator-Prior rarely exceeds random-chance performance and typically remains close to or below chance. Similarly, in all three mouse cell lines, CellOracle's prior does not construct a prior above chance. Here, complex and long-range regulatory interactions may not be sufficiently captured by these proximity-based approaches, highlighting a key limitation of accessibility-based prior knowledge in complex mammalian cell lines.

In addition, we find that the best GRNs largely track the best priors, with PMF-GRN producing the top-performing GRN in three of six cell lines, CellOracle in two, and Inferelator in one (Figure 6C). Notably, there are instances where GRN inference can partially compensate for a weaker prior. In mESC, CellOracle's prior is below chance (normalized AUPRC $= -0.04$), yet its regression-based GRN achieves the highest normalized performance across all methods (0.21), outperforming GLM-Prior + PMF-GRN despite starting from an inferior prior. This behavior is clarified by examining how each GRN model modulates its prior in terms of edge flips (Figure 6D).

PMF-GRN primarily improves performance by adding edges to the prior, effectively expanding the regulatory network in a data-informed way while leaving most high-confidence edges intact. Inferelator both adds and removes edges, suggesting a balance between edge discovery and denoising. In contrast, CellOracle's regression-based design only allows removal of edges: it can shrink edge weights to zero but cannot introduce

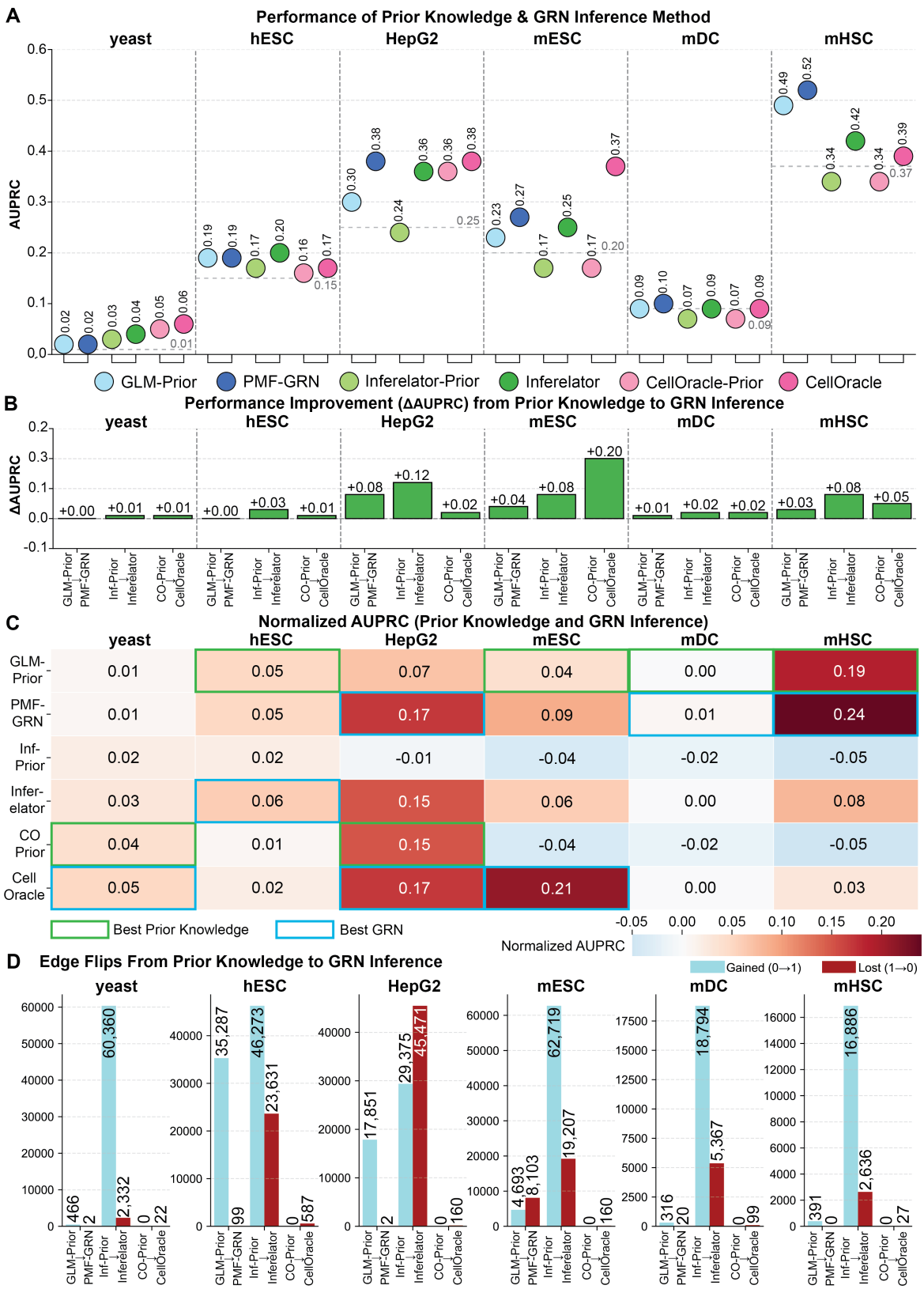

Figure 6: Paired prior-GRN strategies reveal prior dependent gains from expression-based GRN inference across six cell lines. (**A**) AUPRC of each prior (GLM-Prior, Inferelator-Prior, and CellOracle's prior) and its corresponding GRN (PMF-GRN, Inferelator, CellOracle) across six cell lines, with gray dashed lines indicating the chance level in each species. (**B**) Change in AUPRC from prior to GRN (ΔAUPRC) for each paired method, quantifying the added value of expression-based inference. (**C**) Normalized AUPRC performance for priors and GRNs in each cell line. Green boxes highlight the best prior and blue boxes highlight the best GRN per cell line. (**D**) Edge flips between prior and GRN for each method demonstrating how GRN inference methods modulate their priors.

new edges, discovered through expression data, that were absent from the accessibility-derived prior. The large performance gains observed for CellOracle in settings such as mESC therefore arise from aggressive pruning of false-positive edges in a dense, noisy prior, rather than from discovery of novel regulatory interactions. This pruning behavior can make a weak but overly dense prior appear substantially stronger after GRN inference, yet also highlights an important limitation: when single-cell expression is used only to filter a fixed network, not to propose new edges, it cannot expand the regulatory scaffold beyond what is already encoded in the prior.

Together, these results demonstrate that while GRN inference can substantially improve performance in some settings, particularly when priors are informative but incomplete, the overall ranking of methods is largely dictated by prior quality. Expression-based GRN inference models mostly refine and re-weight the structure supplied by the prior, rather than overturning it, underscoring the central role of prior construction in determining GRN reconstruction performance.

Cross-method comparison disentangles prior and GRN inference contributions to performance

To disentangle the contribution of prior construction and GRN inference from the GRN methods themselves, we next performed a full cross-comparison in which each GRN inference method is applied to each prior across all six cell lines (Figure 7). Specifically, we combine GLM-Prior, Inferelator-Prior, and CellOracle baseGRN with PMF-GRN, Inferelator, and CellOracle, yielding nine prior-GRN combinations per cell line. In Figure 7A, we show the resulting AUPRC values, along with species-specific chance levels (gray dashed line), and the baseline performance of the prior (colored dashed line) used for GRN inference. In Figure 7B, we report the same results normalized relative to chance, where values above zero indicate performance above random. We focus our interpretation on the normalized AUPRC values, which allows for easier comparison across species and methods, and refer to the raw AUPRCs in Figure 7A where relevant.

Across species, the dominant pattern is that the identity of the best GRN for a given cell line is largely determined by the underlying prior, not by the choice of GRN inference method. In yeast, for example, any GRN method paired with CellOracle's prior attains the same normalized AUPRC (0.05), whereas all combinations using GLM-Prior or Inferelator-Prior remain at or below 0.03. Once CellOracle's prior is selected, the choice among PMF-GRN, Inferelator, or CellOracle has no quantifiable impact, indicating that the prior itself, rather than the choice of downstream inference algorithm, is the bottleneck of GRN inference performance.

Similar patterns appear in the mammalian cell lines. In hESC, GLM-Prior and Inferelator-Prior yield the strongest priors (0.19 and 0.17 AUPRC, respectively), and nearly all GRNs inferred from these priors achieve competitive performance. In HepG2, CellOracle's prior is best (AUPRC = 0.36), and the two highest performing GRNs are obtained by pairing this prior with PMF-GRN or CellOracle, demonstrating again that performance is largely determined by the prior rather than the choice of inference algorithm. An analogous effect is observed in mHSC, where GLM-Prior combined with PMF-GRN yields the highest performing combination (AUPRC = 0.24), while all GRNs built on Inferelator-Prior or CellOracle's prior plateau at 0.08 or lower. These results support a consistent conclusion, the quality and structure of the prior largely dictate the achievable performance, and the GRN inference methods primarily modulate performance within a range set by the prior.

The full cross-comparison design also allows us to ask whether any GRN inference method is inherently superior across priors and species. We find that across PMF-GRN, Inferelator, and CellOracle, the overarching evidences suggests no GRN inference method is superior. Each approach wins within different regimes defined by the combination of species and prior. PMF-GRN delivers the highest normalized performance in HepG2 when paired with CellOracle's prior (0.20) and in mHSC when paired with GLM-Prior (0.24). PMF-GRN also performs well in yeast when paired with CellOracle's prior, and mESC when paired with GLM-Prior or Inferelator-Prior. The Inferelator is best in hESC when paired with Inferelator-Prior (0.06) and in mDC when paired with CellOracle's prior (0.02), and remains competitive in HepG2 across multiple priors. CellOracle is uniquely optimal in mESC when paired with CellOracle's prior (0.21), substantially outperforming all other prior-GRN combinations in that species. The ranking of each GRN inference method thus changes with both prior and cell line, consistent with our claim that GRN methods are context-dependent refiners

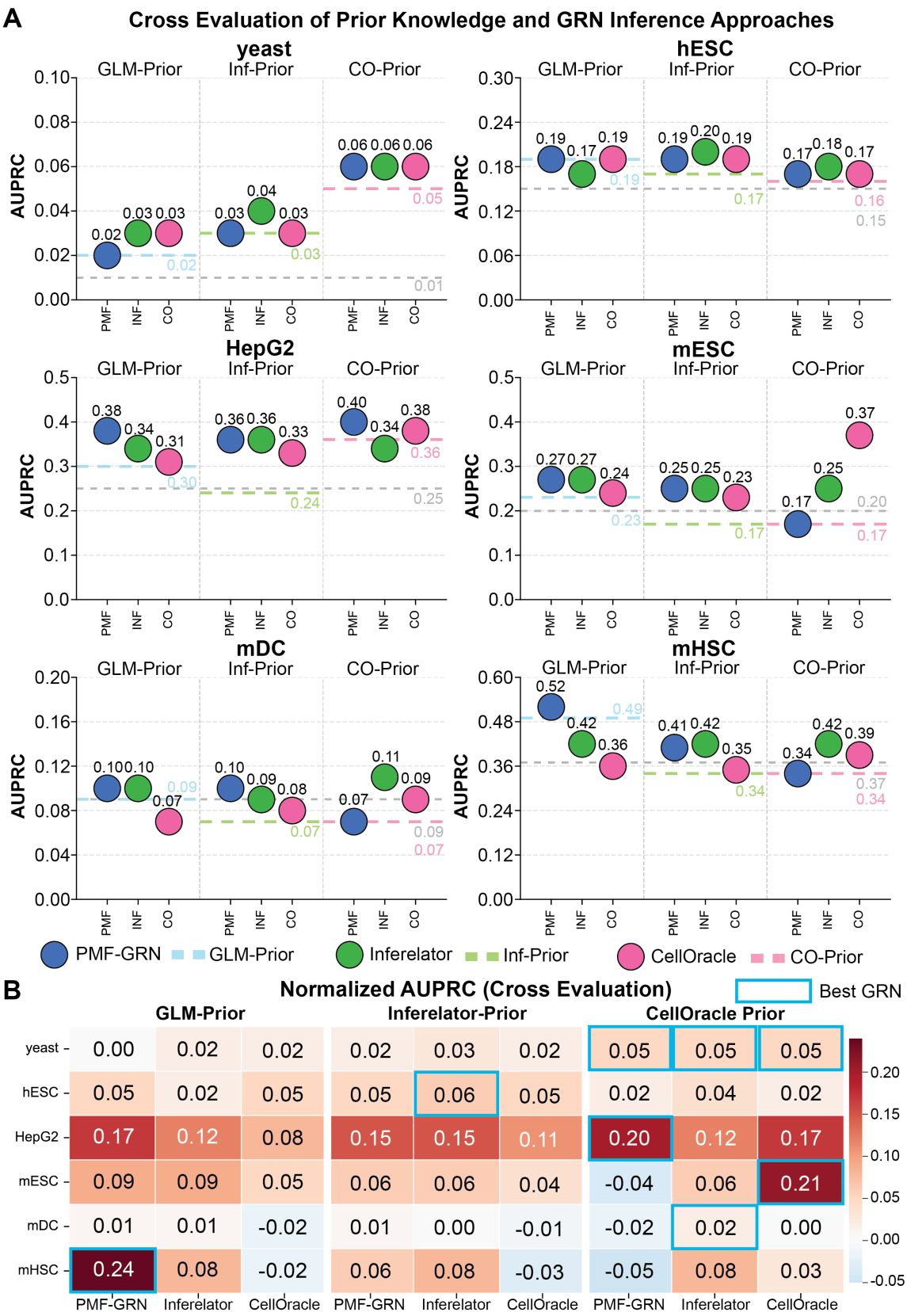

Figure 7: Cross-method comparison disentangles prior and GRN inference contributions. **(A)** AUPRC for all nine combinations of prior and GRN inference methods across six cell lines. Gray dashed lines indicate chance performance, while light blue, light green, and light pink dashed lines represent baseline performance of GLM-Prior, Inferelator-Prior, and CellOracle prior, respectively, across each cell line. **(B)** Normalized AUPRC of all nine combinations of prior and GRN for directly comparing across method and cell line. Blue boxes indicate highest GRN performance per cell line.

of information encoded in a prior using expression data, rather than the primary driver of performance with a single universally superior algorithm.

The cross-method comparison further reveals characteristic interactions between each GRN inference method and the priors. PMF-GRN appears highly prior sensitive, where switching priors can produce large swings in performance within a given species. In mHSC, for instance, normalized performance for PMF-GRN ranges from -0.05 with CellOracle-Prior to 0.24 with GLM-Prior. In mESC, performance ranges from -0.04 with CellOracle Prior to 0.09 with GLM-Prior. This pattern indicates that PMF-GRN is very efficient at exploiting high-quality priors, but struggles to rescue very weak ones. The Inferelator, by contrast, is comparatively robust to prior choice. Within each species, the variation in normalized AUPRC across priors is typically modest (often within the range of 0.02 to 0.04). In mHSC, all three priors yield identical normalized performance (0.08) with the Inferelator. This suggests that the Inferelator leans more heavily on expression-based regularized regression, making it less sensitive to exact prior structure, but also preventing it from reaching the highest performance in most cell lines. CellOracle exhibits strong synergy with its own prior in specific settings. For example, in mESC, CellOracle's prior + Cell Oracle perform well (0.21), significantly outperforming all other combinations, indicating that the regression-based pruning used by CellOracle is particularly well matched when the prior necessitates the removal of false positive edges.

As each GRN inference method is applied to every prior, we can also ask whether cross-pairing ever allows a "strong" GRN inference method to elevate a weaker prior above combinations that use a stronger prior with a different inference method. Such reversals are rare. In mHSC, for example, no method applied to Inferelator-Prior or CellOracle prior achieves performance comparable to GLM-Prior + PMF-GRN. In HepG2, there is a modest example of beneficial cross-pairing, where CellOracle prior + PMF-GRN achieves an AUPRC of 0.20, outperforming the original CellOracle prior + CellOracle pairing (0.17). The performance gains here are likely due to PMF-GRN providing additional edges to the CellOracle's prior using expression data. In mDC, CellOracle prior + Inferelator (0.02) is better than CellOracle prior + CellOracle (0) or CellOracle prior + PMF-GRN (0.02), although all values remain near chance. These exceptions are incremental, not complete reversals of the prior-driven ranking, and reinforce the view that GRN inference can fine-tune performance but typically cannot turn a weak prior into one that rivals the best priors.

Overall, the cross-method comparison in Figure 7 supports three main conclusions. First, prior quality remains the primary determinant of GRN reconstruction performance, even when each GRN method is given access to each prior. Second, there is no universally better GRN inference algorithm: PMF-GRN, the Inferelator, and CellOracle perform best in different prior and species regimes, reflecting differences in how they use expression data to refine or re-weight the prior. Third, GRN inference methods modulate, rather than overturn, the structure imposed by the prior. Changing the inference algorithm can yield meaningful gains in some contexts, particularly when paired with an appropriate prior, but rarely elevates a weak prior to the level of a strong one. These findings reinforce our claim that prior construction is the main driver of GRN inference performance, with expression-based GRN inference primarily serving a prior-dependent refining role.

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, and integrate it with PMF-GRN, a probabilistic matrix factorization framework for expression-based GRN inference. This dual-stage design decouples prior construction from downstream GRN inference, allowing us to evaluate how different priors and GRN inference algorithms jointly shape GRN reconstruction across yeast, mouse, and human cell lines.

Our results first demonstrate that GLM-Prior's performance is tightly linked to the composition of the training data. Across six cell lines spanning yeast, human, and mouse, predictive accuracy scales with the abundance and diversity of positive labels and TF coverage. Human and mouse cell lines with rich, well-annotated regulatory datasets support substantially higher AUPRCs than yeast, where extreme class imbalance and sparse labels limit generalization despite good within-training metrics. These trends highlight

that even with powerful sequence models, generalization is fundamentally constrained by the available supervision. Here, we find that when only a small subset of true interactions are labeled, the model can capture some regulatory grammar but cannot fully resolves novel gene-TF combinations in held-out contexts.

By systematically varying training paradigms, we further show that GLM-Prior generalizes in a biologically interpretable manner across species. Transfer learning between evolutionarily related species, such as human and mouse, maintains or modestly improves performance relative to single-species models, indicating that GLM-Prior captures conserved sequence-level features that can be reused across lineages. In contrast, models trained in mammals transfer poorly to yeast, where performance remains near chance, consistent with deeper divergence in regulatory grammar, TF binding preferences, and genome organization. Multi-species training on human, mouse, and yeast preserves accuracy comparable to the best single-species or transfer models in most cell lines, with only mild trade-offs in some human contexts, and yields a domain-stable prior that can be applied to new systems without species-specific retraining. Together, these findings establish GLM-Prior as a flexible framework for constructing priors across diverse data regimes, with strong performance in well-annotated human and mouse cell lines and more limited gains in sparsely labeled settings such as yeast.

Furthermore, benchmarking GLM-Prior against widely used accessibility-based priors, including Inferelator-Prior and CellOracle's baseGRN, across six cell lines demonstrates that GLM-Prior consistently performs at or above chance and provides the strongest prior in four of six contexts (hESC, mESC, mDC, and mHSC). At the same time, accessibility-based priors remain competitive and even superior in settings dominated by proximal regulatory logic, such as yeast, where promoter-focused chromatin accessibility and motif enrichment capture a large fraction of regulatory edges. In more complex mammalian systems, accessibility-based priors often fall below chance while GLM-Prior remains predictive, suggesting that sequence-based models may better capture distal enhancer regulation and long-range TF-gene coupling that are not easily recovered from proximity-based peak-gene assignment. These complementary performance profiles indicate that GLM-Prior provides the most robust and transferable prior knowledge across diverse mammalian contexts, while accessibility-based priors remain particularly effective in promoter-dominated scenarios such as yeast, together capturing distinct and complementary facets of regulatory architecture.

Integrating these priors with multiple GRN inference algorithms reveals a consistent pattern: the quality and structure of the prior largely determines the achievable performance, and GRN inference algorithms mostly modulate performance within a range set by the prior. First, when we use GLM-Prior as input to PMF-GRN, AUPRC is equal to or higher than the prior alone in all six cell lines, with modest gains when the prior is already strong and larger improvements when the prior is weaker. Edge-flip analyses in this setting show that PMF-GRN typically makes relatively small, targeted changes, treating the sequence-derived prior as a stable scaffold and adding or pruning a limited subset of edges rather than reconstructing the network from scratch. We then extend this analysis to the full cross-method setting, in which all three priors (GLM-Prior, Inferelator-Prior, and CellOracle's baseGRN) are combined with all three GRN inference methods (PMF-GRN, Inferelator, and CellOracle). In this fully crossed design, the best-performing GRNs in each cell line almost always arise from combinations that use the strongest prior for that context, regardless of which inference algorithm is applied. At the same time, no GRN inference method is uniformly superior across priors or species. Each GRN algorithm performs best in different prior–species regimes and interacts with priors in characteristic ways, from highly prior-sensitive behavior to more robust, expression-driven refinement. Together, these analyses indicate that GRN performance is primarily limited by prior construction, with inference algorithms providing prior-dependent gains rather than defining the overall performance ceiling.

Our results recast GRN inference as a problem in which the central bottleneck is the construction of high-quality prior knowledge, rather than the specific choice of the downstream expression-based inference algorithm. GLM-Prior addresses this bottleneck by leveraging a transformer-based genomic language model to learn regulatory grammar directly from DNA, enabling priors that generalize across cell types, species, and training paradigms. Critically, because GLM-Prior operates directly on genome sequence, it removes the dependence on cell-type- and assay-specific measurements such as ATAC-seq, and thus makes it possible, in principle, to construct regulatory priors in understudied or experimentally inaccessible species where chromatin accessibility or expression data are sparse or unavailable. While accessibility-based priors remain valuable in contexts dominated by relatively simple, proximal promoter–gene regulation, they struggle to

capture regulatory architecture in more complex mammalian cell lines and cannot be readily transferred to unseen species or cell types, in contrast to sequence-based priors such as GLM-Prior.

Several limitations of our study point to promising directions for future work. First, GRN benchmarks are constrained by incomplete and noisy interaction databases, in which many negative labels likely represent unknown or untested edges rather than true absences of regulation. This incompleteness affects both prior construction and evaluation, and motivates the development of benchmarks that better distinguish between confirmed negatives and unlabeled interactions. Further, as GLM-Prior downsamples negative examples from the input training data, we likely discard false negative edges during training and thus never learn these regulatory interactions using our sequence-based approach. Future work could include more sampling techniques to avoid losing access to interactions previously defined as negative example that could be in fact classified as positive examples during training. Second, our models are trained on a limited set of cell lines and regulatory labels. Extending GLM-Prior to broader compendia of cell types, developmental stages, and species will be important for assessing how far sequence-based priors can be pushed as a general regulatory scaffold that other methods can refine. Third, while we consider multiple GRN inference algorithms, they share a common edge-centric view of network reconstruction. Exploring integration strategies that move beyond edge-wise scoring, such as jointly modeling regulatory modules, trajectories, or causally inferred dynamics (39), may reveal additional ways that expression data can complement strong sequence-derived priors.

In summary, this work introduces GLM-Prior as a scalable, sequence-based framework for constructing regulatory priors and systematically characterizes how these priors interact with GRN inference methods across yeast, mouse, and human. By demonstrating that prior quality is the dominant determinant of GRN reconstruction performance, our study shifts the emphasis in GRN modeling toward learning robust, generalizable priors from sequence, and provides a foundation for future methods that integrate these priors with expression and other modalities in more flexible and biologically grounded ways.

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 (26) 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 (27; 28; 29; 30) and TRRUST (31; 32), 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 (27; 28; 29; 30) and TRRUST (31; 32)), 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 (40) 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,

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

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:

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

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) (35) and GSE144820 (Y2) (36); Human embryonic stem cells: GSE75748 (41); Human HepG2 cells: GSE81252 (42; 43); Mouse embryonic stem cells: GSE98664 (44); Mouse dendritic Cells: GSE48968 (45); and Mouse hematopoietic stem cells: GSE81682 (46). The YEASTRACT prior knowledge was derived from the YEASTRACT database (11).

The prior knowledge labels used for training GLM-Prior in yeast were obtained from the YEASTRACT database (11). Prior knowledge labels for human and mouse were obtained from the STRING (27; 28; 29; 30) and TRRUST (31; 32) databases, which were provided by the BEELINE benchmark (34) and can be found at https://zenodo.org/records/7682713. Prior knowledge for Inferelator-Prior and CellOracle's baseGRN were generated using ATAC-seq peaks from the following accessions: human embryonic stem cells: 4DNFIPGM38K4, human HepG2 cells: ENCFF913MQB, mouse embryonic stem cells: 4DNFIAEQI3RP, mouse dendritic cells: ENCFF109TUH, and mouse hematopoietic stem cells: ENCFF931CIR.

The gold standard for the yeast datasets was obtained from (33). Reference networks used as gold standards in human embryonic stem cells, human HepG2 cells, mouse embryonic stem cells, mouse dendritic cells, and mouse hematopoietic stem cells were obtained from the BEELINE benchmarks (34).

Models for experiments from yeast, human and mouse can be found at https://huggingface.co/cskokgibbs as well as their corresponding tokenized datasets. Priors and GRNs created with CellOracle and Inferelator software can also be found here.

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] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Zhaonan Zou, Tazro Ohta, and Shinya Oki. Chip-atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Research*, 52(W1):W45–W53, 2024.

[38] Huilei Xu, Caroline Baroukh, Ruth Dannenfelser, Edward Y Chen, Christopher M Tan, Yan Kou, Yujin E Kim, Ihor R Lemischka, and Avi Ma'ayan. Escape: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database*, 2013:bat045, 2013.

[39] Jang-Hyun Kim, Claudia Skok Gibbs, Sangdoo Yun, Hyun Oh Song, and Kyunghyun Cho. Large-scale targeted cause discovery via learning from simulated data. *Transactions on Machine Learning Research*, 2025.

[40] 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.

[41] 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.

[42] Felipe Mora-Bermúdez, Farhath Badsha, Sabina Kanton, J Gray Camp, Benjamin Vernot, Kathrin Köhler, Birger Voigt, Keisuke Okita, Tomislav Maricic, Zhisong He, et al. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *Elife*, 5:e18683, 2016.

[43] J Gray Camp, Keisuke Sekine, Tobias Gerber, Henry Loeffler-Wirth, Hans Binder, Malgorzata Gac, Sabina Kanton, Jorge Kageyama, Georg Damm, Daniel Seehofer, et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature*, 546(7659):533–538, 2017.

[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] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, et al. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014.

[46] Sonia Nestorowa, Fiona K Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, The Journal of the American Society of Hematology*, 128(8):e20–e31, 2016.

[47] Abdul Muntakim Rafi, Brett Kiyota, Nozomu Yachie, and Carl de Boer. Detecting and avoiding homology-based data leakage in genome-trained sequence models. *bioRxiv*, pages 2025–01, 2025.

[48] Thomas Madden. The blast sequence analysis tool. *The NCBI handbook*, 2(5):425–436, 2013.

Appendix

A Single-Species Experiments

A.1 Yeast

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 (26) 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). Training for yeast was conducted across 5,893 genes and 123 TFs, with 660 positive labels and 724,179 negative labels for these interactions derived from YEASTRACT.

Validation Metrics for Yeast Single-Species GLM-Prior

Metric	Score
Best F1 Score	0.76
ROC AUC	1.00
Best Classification Threshold	1.00
Positive Class Precision	0.89
Positive Class Recall	0.67
Negative Class Precision	1.00
Negative Class Recall	1.00
AUPRC (vs. held-out set (33))	0.02

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).

A hyperparameter sweep over 1 epoch of training using different class-weights and downsampling rates for the negative class revealed $[0.5, 1.0]$ 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 yeast single-species GLM-Prior model on YEASTRACT database labels, we ran inference on all genes and TFs, including those in a held-out set not seen during training. These held-out genes and TFs corresponded to those present in the gold standard from (33). Our held-out set comprised 933 genes and 98 TFs, with 986 positive labels, and 95,911 negative labels. After inference, the model predictions were evaluated using AUPRC with the corresponding labels from the gold standard, with a chance baseline calculated by taking the positive rate of the held-out set.

A.1.1 Sequence homology experiments in yeast

Sequence homology refers to shared sequence similarity between loci, often introduced by duplication and other genome rearrangements. Because nucleotide language models are optimized to exploit recurring sequence patterns, high homology across train-test splits can, in principle, make held-out prediction easier by allowing the model to match familiar fragments rather than rely on fully generalizable features. For GLM-Prior, which is trained on labeled TF-sequence pairs, this motivates a simple check: whether genes or TFs are unusually similar to sequences seen during training, and whether controlling for that similarity changes performance at inference time. We therefore use hashFrag (47), which leverages BLAST (48)

to quantify cross-split sequence similarity and generate homology-aware partitions (or prune highly similar training sequences), and then re-evaluate GLM-Prior on these homology-controlled splits to test whether sequence similarity contributes to our yeast performance.

In Figure A1A we depict the homology-aware partitioning scheme. We then quantify leakage among genes by computing, for each test gene, its maximum BLAST score to any training gene (Figure A1B-C). Using an operational cutoff $t = 150$, and holding the test set fixed ($n = 978$), pruning 346 homologous training genes reduces the fraction of "leaking" test genes from 24.8% to 7.8%.

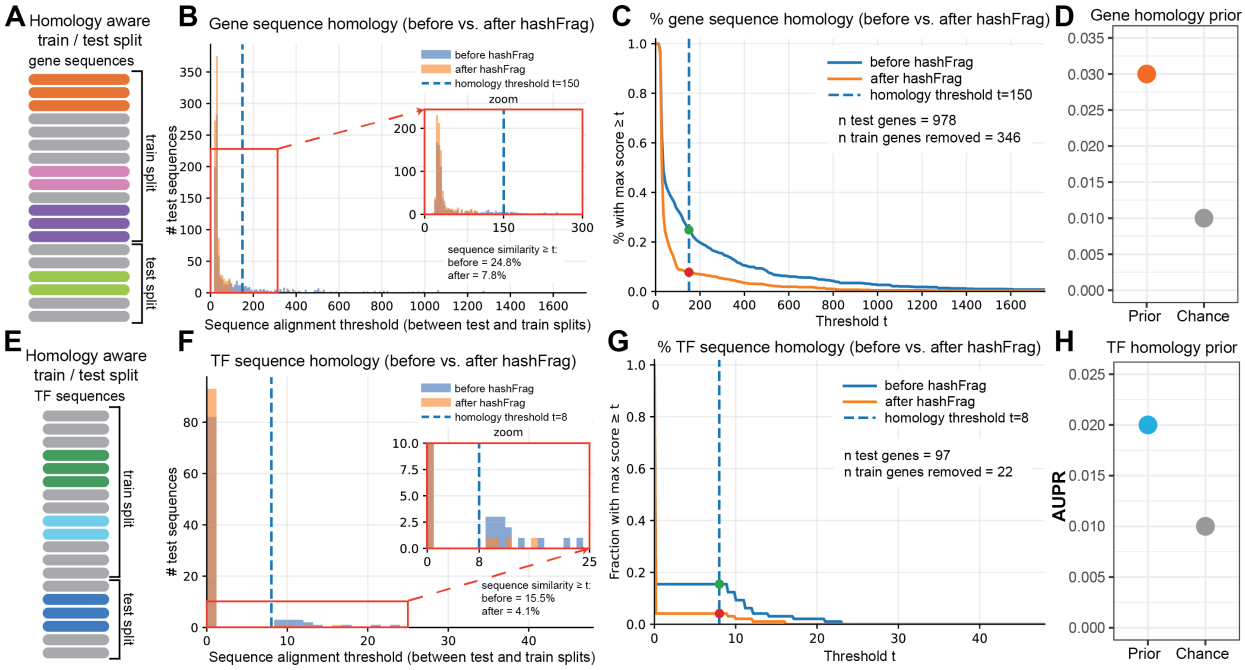

Figure A1: **Sequence-homology analysis of yeast gene and TF splits using hashFrag. (A)** Schematic of homology-aware partitioning, illustrating how highly similar sequences are identified and separated to reduce cross-split homology between training and test sets. **(B)** Histogram of per-test-gene maximum BLAST alignment score to any training gene, shown before (blue) and after (orange) hashFrag pruning of the training gene set. Blue dashed line indicates the similarity threshold used to define leakage. **(C)** Leakage curve for genes showing, for each similarity threshold t, the fraction of test genes whose maximum BLAST score to any training gene is $\geq t$. Blue dashed line marks the leakage threshold. **(D)** AUPRC evaluation of GLM-Prior trained and tested on the hashFrag-defined gene partitions, with the gray dot indicating chance performance. **(E)** Schematic of homology-aware partitioning for TF sequences, analogous to panel A. **(F)** Histogram of per-test-TF maximum BLAST alignment score to any training TF, shown before (blue) and after (orange) hashFrag pruning of the training TF set. Blue dashed line indicates the similarity threshold. **(G)** Leakage curve for TFs showing, for each similarity threshold t, the fraction of test TFs whose maximum BLAST score to any training TF is $\geq t$. Blue dashed line marks the leakage threshold. **(H)** AUPRC evaluation of GLM-Prior trained and tested on the hashFrag-defined TF partitions, with the gray dot indicating chance performance.

We next ask whether removing this homology affects model accuracy. Training on the hashFrag-defined gene partitions yields a test AUPRC of 0.03, compared to 0.01 chance performance (Figure A1D). This magnitude is comparable to the yeast results (AUPRC = 0.02) reported in Section 1, indicating that sequence leakage is not responsible for our observed performance in yeast.

We repeat the same analysis for TF sequences (Figure A1E-H). As with genes, hashFrag pruning decreases cross-split similarity, and model performance on the homology-clean TF splits remains similar to the main yeast model, again motivating against leakage-driven gains.

Overall, these experiments show that while homology-aware partitioning is methodologically important and successfully reduces cross-split similarity, it does not materially affect our model performance, which is primarily dependent on access to sufficient positive training examples. Because aggressive pruning removes substantial numbers of sequences, and thus labels, from both training and evaluation, it can reduce statistical power without measurable benefit here. For this reason, we include this analysis in the appendix to demonstrate its value, and motivate why we avoid homology pruning in the primary experiments to preserve sufficient data for training and evaluation.

A.2 Human cell lines (hESC & HepG2)

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 (26), under *H. sapiens*.

Validation Metrics for Human Single-Species GLM-Prior

Metric	**hESC**	**HepG2**
Best F1 Score	0.29	0.35
ROC AUC	0.85	0.83
Best Classification Threshold	0.94	0.97
Positive Class Precision	0.39	0.53
Positive Class Recall	0.22	0.27
Negative Class Precision	0.99	0.99
Negative Class Recall	1.00	1.00
AUPRC (vs. reference network)	0.19	0.30

Table 2: Validation performance of the GLM-Prior model trained on human, evaluated in hESC and HepG2. Evaluation is conducted on held-out sets using validation metrics that consider the contribution of the positive and negative classes and AUPRC against ChIP-seq-derived reference networks after inference of the prior-knowledge matrix.

Training for hESC was conducted across $1,801$ genes and 504 TFs, with $5,305$ positive labels and $902,399$ negative labels for these interactions derived from STRING and TRRUST. Training for hESC was conducted across $2,121$ genes and 527 TFs, with $8,087$ positive labels and $1,109,680$ negative labels for these interactions derived from STRING and TRRUST.

A hyperparameter sweep over 1 epoch of training using different class-weights and downsampling rates for the negative class revealed $[0.8, 1.0]$ to be the optimal class-weights and 0.3 to be the optimal negative class downsampling rate in hESC, and $[1.0, 1.0]$ to be the optimal class-weights and 0.4 to be the optimal negative class downsampling rate in HepG2. These weights were used during final training and achieved the validation metrics found in Table 2.

Following training of the human single-species GLM-Prior model on STRING and TRRUST database labels, we ran inference on all genes and TFs, including those in a held-out set not seen during training. These held-out genes and TFs corresponded to those present in the BEELINE hESC and HepG2 reference ChIP-seq networks, respectively. For hESC, our held-out set comprised $4,773$ genes and 79 TFs, with $58,337$ positive labels, and $318,730$ negative labels. For HepG2, our held-out set comprised $4,338$ genes and 53 TFs, with $56,567$ positive labels, and $173,347$ negative labels. After inference, the model predictions were evaluated using AUPRC with the corresponding labels from the hESC and HepG2 reference ChIP-seq networks, with a chance baseline calculated by taking the positive rate of the held-out set.

A.3 Mouse cell lines (mESC, mDC, & mHSC)

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 (26), under *M. musculus*.

Validation Metrics for Mouse Single-Species GLM-Prior

Metric	mESC	mDC	mHSC
Best F1 Score	0.19	0.20	0.19
ROC AUC	0.70	0.80	0.74
Best Classification Threshold	0.97	0.99	0.98
Positive Class Precision	0.23	0.17	0.50
Positive Class Recall	0.16	0.19	0.12
Negative Class Precision	0.99	0.99	0.99
Negative Class Recall	0.99	0.99	1.00
AUPRC (vs. reference network)	0.23	0.09	0.49

Table 3: Validation performance of the GLM-Prior model trained on mouse, evaluated in mESC, mDC, and mHSC. Evaluation is conducted on held-out sets using validation metrics that consider the contribution of the positive and negative classes and AUPRC against ChIP-seq-derived reference networks after inference of the prior-knowledge matrix.

Training for mESC was conducted across $1,097$ genes and 437 TFs, with $3,129$ positive labels and $476,260$ negative labels for these interactions derived from STRING and TRRUST. Training for mDC was conducted across $2,837$ genes and 462 TFs, with $11,551$ positive labels and $1,299,134$ negative labels for these interactions derived from STRING and TRRUST. Training for mHSC was conducted across $1,129$ genes and 419 TFs, with $2,808$ positive labels and $470,243$ negative labels.

A hyperparameter sweep over 1 epoch of training using different class-weights and downsampling rates for the negative class revealed $[0.2, 1.0]$ to be the optimal class-weights and 0.5 to be the optimal negative class downsampling rate in mESC, and $[0.1, 1.0]$ to be the optimal class-weights and 0.4 to be the optimal negative class downsampling rate in mDC, and $[0.1, 1.0]$ to be the optimal class-weights and 0.5 to be the optimal negative class downsampling rate in mHSC. These weights were used during final training, achieving the validation metrics shown in Table 3.

Following training of the mouse single-species GLM-Prior model on STRING and TRRUST database labels, we ran inference on all genes and TFs, including those in a held-out set not seen during training. These held-out genes and TFs corresponded to those present in the BEELINE mESC, mDC and mHSC reference ChIP-seq networks, respectively. For mESC, our held-out set comprised $5,711$ genes and 59 TFs, with $65,933$ positive labels, and $271,016$ negative labels. For mDC, our held-out set comprised $3,373$ genes and 29 TFs, with $8,708$ positive labels, and $89,109$ negative labels. For mHSC, our held-out set comprised $6,777$ genes and 72 TFs, with $182,005$ positive labels, and $305,939$ negative labels. After inference, the model predictions were evaluated using AUPRC with the corresponding labels from the mESC, mDC and mHSC reference ChIP-seq networks, with a chance baseline calculated by taking the positive rate of the held-out set.

B Transfer-Learning Experiments

Transfer learning experiments were conducted by first ensuring that no overlap existed between the training set (human, mouse, or yeast), and the corresponding inference set with held-out evaluation labels. For the

human and mouse model that transferring knowledge to yeast, no information was overlapping between training and inference *a priori*.

AUPRC Results from Transfer Learning

Training Dataset	Inference Dataset	AUPRC
Human and Mouse	Yeast	0.02
Mouse	hESC	0.19
Mouse	HepG2	0.31
Human	mESC	0.23
Human	mDC	0.09
Human	mHSC	0.46

Table 4: 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.

For the mouse trained model which transferred knowledge to hESC and HepG2, two separate models were trained, one in which overlaps between the mouse genes, TFs and labels with hESC were removed, and the second in which the overlaps between mouse genes, TFs and labels with HepG2 were removed. For the human trained model which transferred knowledge to mESC, mDC, and mHSC, three separate models were trained. In the first model, overlaps between human genes, TFs, and labels with mESC were removed. In the second model, overlaps between human genes, TFs, and labels with mDC were removed. In the third model, overlaps between human genes, TFs, and labels with mHSC were removed.

Training six separate models allowed us to ensure an optimal number of gene and TF sequences and their corresponding labels were seen during training, while maintaining complete independence from the inference set and corresponding evaluation performance. Results for the transfer learning experiments can be found in Table 4.

C Multi-Species Experiments

To train a multi-species model, we consecutively trained GLM-Prior on human, mouse, and yeast. To ensure no sequence overlap between the genes, TFs, and labels used during training and inference, we removed all overlapping inference sets from each individual organisms training set. Here, we ensured there was no inter-species leakage, as well as intra-species leakage, between training and inference sets.

Following the consecutive training of the multi-species model, we ran inference in yeast, hESC, HepG2, mESC, mDC, and mHSC held-out sets, respectively. Results for this multi-species model inference can be found in Table 5.

D Inferelator-Prior and CellOracle baseGRN

We generated prior knowledge for Inferelator-Prior (6) and CellOracle's baseGRN (8), using each respective methods published Python software. To generate prior knowledge with Inferelator-Prior and CellOracle baseGRN for each of the six cell lines, we obtained ATAC-seq datasets from the following accessions: Human embryonic stem cells: 4DNFIPGM38K4, Human HepG2 cells: ENCFF913MQB, Mouse embryonic stem cells: 4DNFIAEQI3RP, Mouse dendritic cells: ENCFF109TUH, and Mouse hematopoietic stem cells: ENCFF931CIR. In yeast, prior knowledge datasets 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.

Results for the Inferelator-Prior and CellOracle baseGRN prior knowledge for each six cell lines using the same reference sets described in Section A are provided in Table 6.

AUPRC Results from Multi-Species Training		
Training Dataset	**Inference Dataset**	**AUPRC**
Human + Mouse + Yeast	Yeast	0.02
	hESC	0.18
	HepG2	0.29
	mESC	0.23
	mDC	0.09
	mHSC	0.49

Table 5: 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 cell-line specific reference networks.

	AUPRC		
Cell Line	**GLM-Prior**	**Inferelator-Prior**	**CellOracle baseGRN**
Yeast	0.02	0.03	0.05
hESC	0.19	0.17	0.16
HepG2	0.30	0.24	0.36
mESC	0.23	0.17	0.17
mDC	0.09	0.07	0.07
mHSC	0.49	0.34	0.34

Table 6: AUPRC of GLM-Prior, Inferelator-Prior and CellOracle baseGRN across six cell lines.

E GRN Inference with PMF-GRN, the Inferelator, and CellOracle

GRN inference for each of the six cell lines was performed using PMF-GRN, the Inferelator 3.0, and CellOracle Python software, respectively.

Single cell gene expression datasets were obtained from GSE125162 ($38,225$ cells by $6,763$ genes) (35) and GSE144820 ($6,118$ cells by $6,763$ genes) (36) without modification. The dataset was then combined by concatenating GSE125162 and GSE144820 on the cells axis ($44,343$ cells by $6,763$ genes). The remaining hESC (758 cells by $17,735$ genes), HepG2 (425 cells by $11,515$ genes), mESC (421 cells by $18,385$ genes), mDC (383 cells by $7,371$ genes), and mHSC ($2,807$ cells by $4,762$ genes) single-cell expression datasets were obtained from BEELINE (34).

	AUPRC		
Cell Line	**PMF-GRN**	**Inferelator 3.0**	**CellOracle**
Yeast	0.02	0.04	0.06
hESC	0.19	0.20	0.17
HepG2	0.38	0.36	0.38
mESC	0.27	0.25	0.37
mDC	0.10	0.09	0.09
mHSC	0.52	0.42	0.39

Table 7: AUPRC of three GRN inference methods (PMF-GRN, Inferelator 3.0, and CellOracle) across six cell lines.

Prior knowledge for each method is required to produce a GRN from single-cell expression data. For the experiments in this section, we use the specific cell line prior knowledge generated by the corresponding

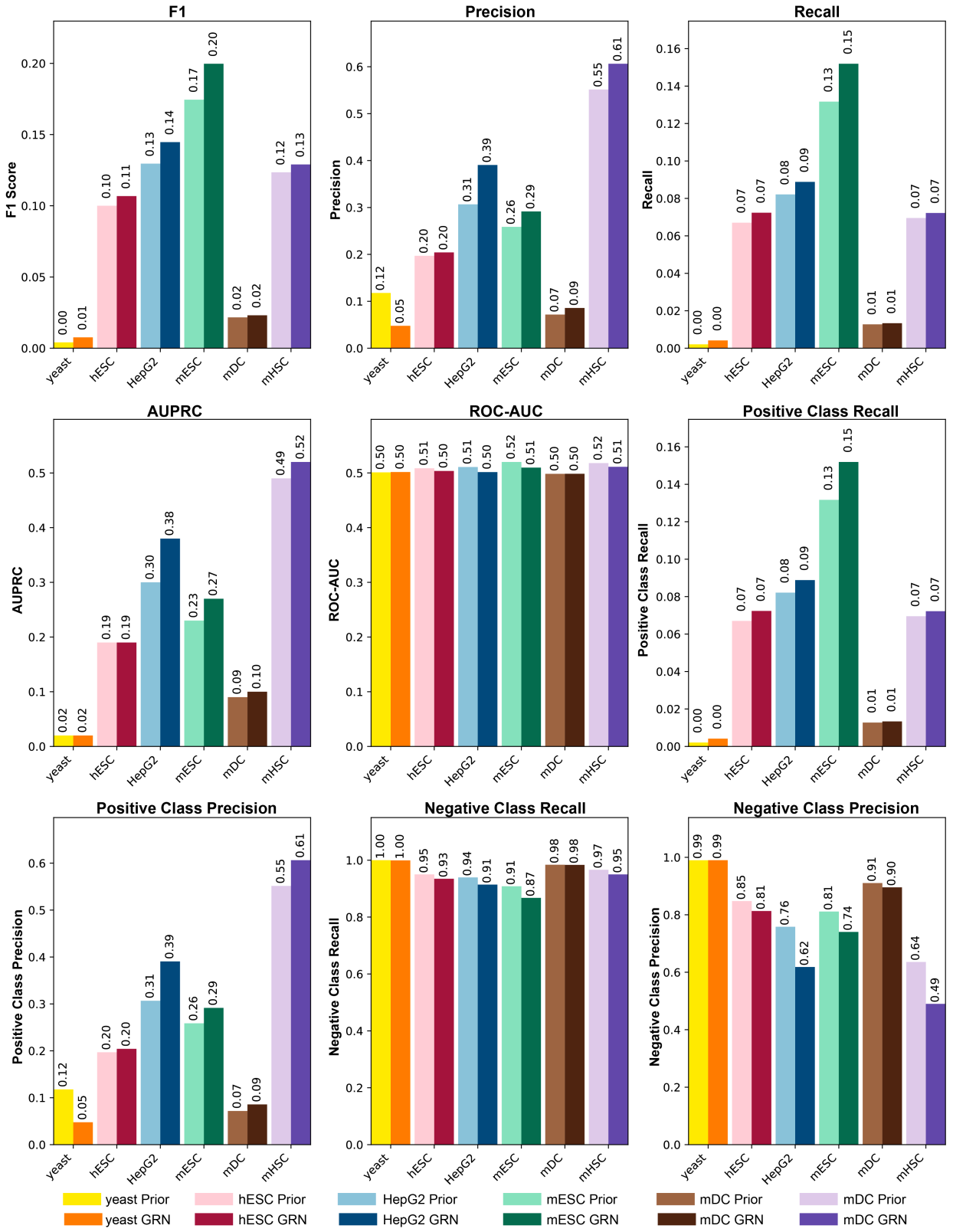

Figure A2: **Performance Metrics**. Additional performance metrics for GLM-Prior and PMF-GRN across six cell lines.

algorithm (described in Section A for GLM-Prior and Section 6 for Inferelator-Prior and CellOracle baseGRN. Results for these GRN inference experiments can be found in Table 7).

Additional metrics for GLM-Prior and PMF-GRN, including F1 score, precision, recall, ROC-AUC, positive class recall, positive class precision, negative class recall, and negative class precision, are provided in A2.

F Cross-Method Comparison

In order to disentangle the performance contributions provided by each prior knowledge method (GLM-Prior, Inferelator-Prior, and CellOracle baseGRN) from their corresponding GRN inference method (PMF-GRN, Inferelator 3.0, and CellOracle), we provide a cross-method comparison in which each GRN inference algorithm is applied to each prior. No new datasets are introduced in this section. The results from the cross-method comparison can be found in Table 8.

	CellOracle Prior			**Inferelator-Prior**			**GLM-Prior**		
Cell Line	**PMF**	**Inf**	**CO**	**PMF**	**Inf**	**CO**	**PMF**	**Inf**	**CO**
Yeast	0.06	0.06	0.06	0.03	0.04	0.03	0.02	0.03	0.03
hESC	0.17	0.18	0.17	0.19	0.20	0.19	0.19	0.17	0.19
HepG2	0.40	0.34	0.38	0.36	0.36	0.33	0.38	0.34	0.31
mESC	0.17	0.25	0.37	0.25	0.25	0.23	0.27	0.27	0.24
mDC	0.07	0.11	0.09	0.10	0.09	0.08	0.10	0.10	0.07
mHSC	0.34	0.42	0.39	0.41	0.42	0.35	0.52	0.42	0.36

Table 8: Cross-method AUPRC comparison for six cell lines, combining three prior-knowledge constructions (CellOracle prior, Inferelator-Prior, GLM-Prior) with three GRN inference methods (PMF-GRN, Inferelator, CellOracle).