Genome Biology

METHOD

Open Access

PMF-GRN: a variational inference approach to single-cell gene regulatory network inference using probabilistic matrix factorization

Claudia Skok Gibbs[1†], Omar Mahmood[1†], Richard Bonneau[1,2,3] and Kyunghyun Cho[1,2*]

†Claudia Skok Gibbs and Omar Mahmood contributed equally to this work.

*Correspondence: kyunghyun.cho@nyu.edu

[1] Center for Data Science, New York University, New York, NY 10011, USA
[2] Prescient Design, Genentech, New York, NY 10010, USA
[3] Center for Genomics and Systems Biology, New York University, New York, NY 10003, USA

Abstract

Inferring gene regulatory networks (GRNs) from single-cell data is challenging due to heuristic limitations. Existing methods also lack estimates of uncertainty. Here we present Probabilistic Matrix Factorization for Gene Regulatory Network Inference (PMF-GRN). Using single-cell expression data, PMF-GRN infers latent factors capturing transcription factor activity and regulatory relationships. Using variational inference allows hyperparameter search for principled model selection and direct comparison to other generative models. We extensively test and benchmark our method using real single-cell datasets and synthetic data. We show that PMF-GRN infers GRNs more accurately than current state-of-the-art single-cell GRN inference methods, offering well-calibrated uncertainty estimates.

Keywords: Probabilistic matrix factorization, Variational inference, Gene regulatory network inference, Single cell, Gene expression

Background

An essential problem in systems biology is to extract information from genome wide sequencing data to unravel the mechanisms controlling cellular processes within heterogeneous populations [1]. Gene regulatory networks (GRNs) that annotate regulatory relationships between transcription factors (TFs) and their target genes [2] have proven to be useful models for stratifying functional differences between cells [3–6] that can arise during normal development [7], responses to environmental signals [8], and dysregulation in the context of disease [9–11].

GRNs cannot be directly measured with current sequencing technology. Instead, methods must be developed to piece together snapshots of transcriptional processes in order to reconstruct a cell's regulatory landscape [12]. Initial approaches to GRN inference relied on microarray technology [13–15], a hybridization-based method to measure the expression of thousands of genes simultaneously [16]. This technology was biased as it was limited to only those genes that were annotated at the time, which in turn

presented challenges for inferring the complete regulatory landscape [1]. Subsequently, the high-throughput sequencing method RNA-seq provided a genome wide readout of transcriptional output, allowing for the detection of novel transcripts [17] and thus improving GRN inference potential. More recently, single-cell RNA-seq technology has enabled the characterization of gene expression profiles within heterogeneous populations [18], vastly increasing the potential for GRN inference algorithms [19, 20]. In contrast to bulk RNA experiments (microarray and RNA-seq) that average measurements of gene expression across heterogenous cell populations, GRNs inferred from single-cell data have the advantage of unmasking biological signal in individual cells [21].

Several matrix factorization approaches have been proposed to overcome the limitations of reconstructing GRNs from microarray data [22]. These include use of statistical techniques such as singular value decomposition and principal component analysis [23], Bayesian decomposition [24], and non-negative matrix factorization [25–27]. More recently, matrix factorization approaches have been applied to integrative analysis of DNA methylation and miRNA expression data [28] as well as single-cell RNA-seq and single-cell ATAC-seq data [29]. However, to the best of our knowledge, these matrix factorization approaches have not yet been used to infer GRNs from single-cell gene expression data. Meanwhile, several regression-based methods have been proposed to learn GRNs from single-cell RNA-seq and single-cell ATAC-seq to capture regulatory relationships at single-cell resolution [30]. So far, these integrative approaches to GRN inference have been successfully implemented using regularized regression [31], self-organizing maps [32], tree-based regression [33], and Bayesian Ridge regression [34].

Although regression-based methods for inferring GRNs from single-cell data are available, they still suffer from significant limitations [35]. Firstly, these methods are designed for specific input datasets, such as bulk or single-cell RNA-seq, causing issues when new data becomes available or new assumptions are required in the model. This can result in inaccurate predictions if the new data or assumptions are not well integrated into the existing model, leading to the need for a complete re-design of the algorithm, which can be costly and time-consuming. Additionally, these methods typically focus on inferring a single GRN that explains the available data, without performing hyperparameter search to determine the optimal model. This can lead to heuristic model selection, with no justification for the approach taken or evidence that the best possible model has been selected. Conversely, hyperparameter search ensures the accuracy of the GRN inference algorithm by finding the optimal model that fits the data well while avoiding overfitting. Regression-based GRN inference algorithms that do not perform hyperparameter search may miss important data features or overemphasize irrelevant ones, leading to inaccurate or incomplete models. Moreover, these methods do not provide an indication of their uncertainty about the predictions that they make. Finally, several regression-based GRN inference algorithms struggle to scale optimally to the size of typical single-cell datasets, limiting inference to small subsets of data or requiring enormous amounts of computational time.

In this study, we introduce PMF-GRN, a novel approach that uses probabilistic matrix factorization [36] to infer gene regulatory networks from single-cell gene expression and chromatin accessibility information. This approach extends previous methods that applied matrix factorization for GRN inference with microarray data, to address the

current limitations in regression-based single-cell GRN inference. We implement our approach in a probabilistic setting with variational inference, which provides a flexible framework to incorporate new assumptions or biological data as required, without changing the way the GRN is inferred. We also use a principled hyperparameter selection process, which optimizes the parameters of our probabilistic model for automatic model selection. In this way, we replace heuristic model selection by comparing a variety of generative models and hyperparameter configurations before selecting the optimal parameters with which to infer a final GRN. Our probabilistic approach provides uncertainty estimates for each predicted regulatory interaction, serving as a proxy for the model confidence in each predicted interaction. Uncertainty estimates can be useful in the situation where there are limited validated interactions or a gold standard is incomplete. By using stochastic gradient descent (SGD), we perform GRN inference on a GPU, allowing us to easily scale to a large number of observations in a typical single-cell gene expression dataset. Unlike many existing methods, PMF-GRN is not limited by pre-defined organism restrictions, making it widely applicable for GRN inference.

To demonstrate the novelty and advantages of PMF-GRN, we apply our method to datasets from *Sacchromyces cerevisiae*, human peripheral blood mononuclear cells (PBMCs), and BEELINE. In our first experiment, we apply our method to two single-cell gene expression datasets for the model organism *S. cerevisiae*. We evaluate our model's performance in a normal inference setting as well as with cross-validation and noisy data. To assess the accuracy of predicted regulatory interactions, we evaluate all regulatory predictions using area under the precision recall curve (AUPRC) against database derived gold standards. Our findings show that the uncertainty estimates are well-calibrated for inferred TF-target gene interactions, as the accuracy of predictions increases when the associated uncertainty decreases. Here, in comparison to three state-of-the-art regression-based methods for inferring single-cell GRNs, namely the Inferelator [31], Scenic [33], and Cell Oracle [34], our method demonstrates an overall improved performance in recovering the true underlying GRN. Additionally, we apply our method to a PBMC dataset and explore the inferred TFA profiles in the context of annotated cell types and specific immune TFs. We investigate regulatory edges in our inferred GRN and find compelling support for our predictions. Lastly, we benchmark our method using six synthetic datasets generated from BEELINE [37] and demonstrate consistent outperformance of PMF-GRN compared to the baseline.

Results

The PMF-GRN model

The goal of our probabilistic matrix factorization approach is to decompose observed gene expression into latent factors, representing TF activity (TFA) and regulatory interactions between TFs and their target genes. These latent factors, which represent the underlying GRN, cannot be measured experimentally, unlike gene expression. We model an observed gene expression matrix $W \in \mathbb{R}^{N \times M}$ using a TFA matrix $U \in \mathbb{R}_{>0}^{N \times K}$, a TF-target gene interaction matrix $V \in \mathbb{R}^{M \times K}$, observation noise $\sigma_{obs} \in (0, \infty)$, and sequencing depth $d \in (0, 1)^N$, where N is the number of cells, M is the number of genes, and K is the number of TFs. We rewrite V as the product of a matrix $A \in (0, 1)^{M \times K}$, representing

the degree of existence of an interaction, and a matrix $B \in \mathbb{R}^{M \times K}$ representing the interaction strength and its direction:

$$V = A \odot B,$$

where $\odot$ denotes element-wise multiplication. An overview of the graphical model is shown in Fig. 1A.

These latent variables are mutually independent a priori, i.e., $p(U, A, B, \sigma_{obs}, d) = p(U)p(A)p(B)p(\sigma_{obs})p(d)$. For the matrix A, prior hyperparameters represent an initial guess of the interaction between each TF and target gene which need to be provided by a user. These can be derived from genomic databases or obtained

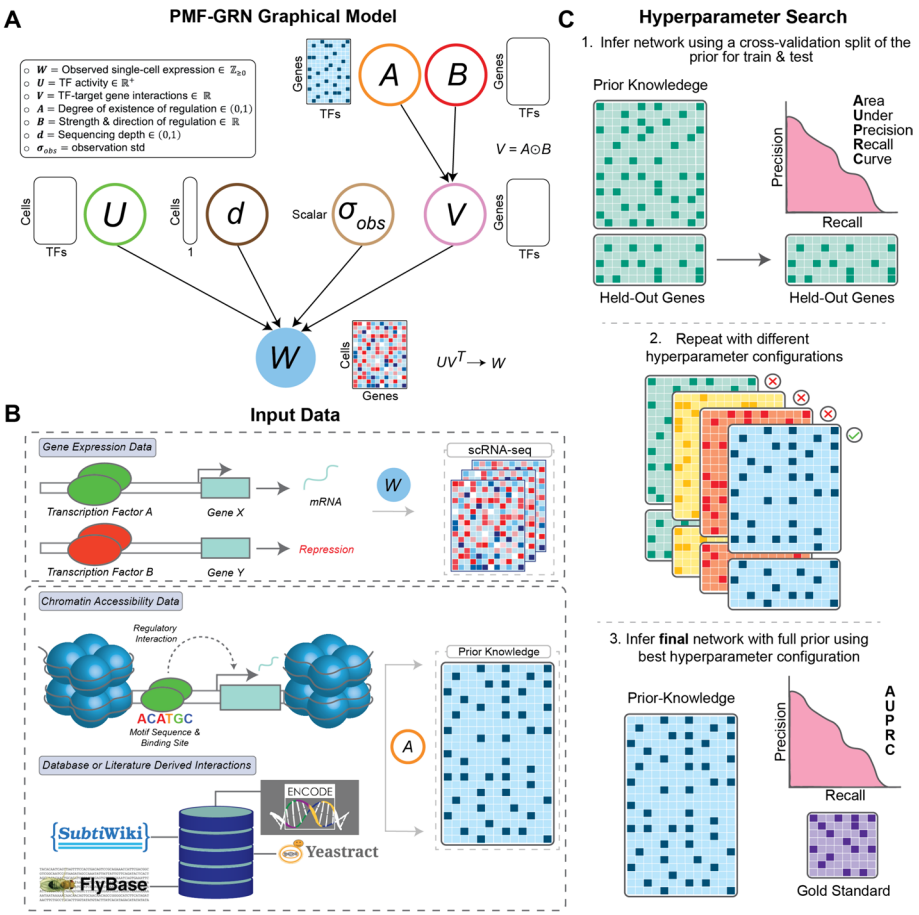

Fig. 1 **A** PMF-GRN graphical model overview. Input single-cell gene expression *W* is decomposed into several latent factors. Information obtained from chromatin accessibility data or genomics databases is incorporated into the prior distribution for *A*. **B** Input experimental data for PMF-GRN includes single-cell RNA-seq gene expression data. Prior-known TF-target gene interactions can be obtained using chromatin accessibility in parallel with known TF motifs or through databases or literature derived interactions. **C** Hyperparameter selection process is performed for optimal model selection. The provided prior-known network is split into a train and validation dataset. 80% of the prior-known information is used to infer a GRN, while the remaining 20% is used for validation by computing AUPRC. This process is repeated multiple times, using different hyperparameter configurations in order to determine the optimal hyperparameters for the GRN inference task at hand. Finally, using the optimal hyperparameters, a final network is inferred using the full prior and evaluated using an independent gold standard

by analyzing other data types, such as the measurement of chromosomal accessibility, TF motif databases, and direct measurement of TF-binding along the chromosome, as shown in Fig. 1B (see the "Methods" section for details).

The observations W result from a matrix product $UV^\top$. We assume noisy observations by defining a distribution over the observations with the level of noise σ_{obs}, i.e., $p(W|U, V = A \odot B, \sigma_{obs}, d)$.

Given this generative model, we perform posterior inference over all the unobserved latent variables——U, A, B, d, and σ_{obs}——and use the posterior over A to investigate TF-target gene interactions. Exact posterior inference with an arbitrary choice of prior and observation probability distributions is, however, intractable. We address this issue by using variational inference [38, 39], where we approximate the true posterior distributions with tractable, approximate (variational) posterior distributions.

We minimise the KL-divergence $D_{KL}(q\|p)$ between the two distributions with respect to the parameters of the variational distribution q, where p is the true posterior distribution. This allows us to find an approximate posterior distribution q that closely resembles p. This is equivalent to maximizing the evidence lower bound (ELBO), i.e., a lower bound to the marginal log likelihood of the observations W:

$$\begin{aligned}\log p(W) \geq \mathbb{E}_{U,A,B,\sigma_{obs},d\sim q(U,A,B,\sigma_{obs},d)}[&\log p(W|U, V = A \odot B, \sigma_{obs}, d)\\ &+ \log p(U, A, B, \sigma_{obs}, d)\\ &- \log q(U, A, B, \sigma_{obs}, d)]\end{aligned}$$

The mean and variance of the approximate posterior over each entry of A, obtained from maximizing the ELBO, are then used as the degree of existence of an interaction between a TF and a target gene and its uncertainty, respectively.

It is important to note that matrix factorization based GRN inference is only identifiable up to a latent factor (column) permutation. In the absence of prior information, the probability that the user assigns TF names to the columns of U and V in the same order that the inference algorithm implicitly assigns TFs to these columns is $\frac{1}{K!}$, is essentially 0 for any reasonable value of K. Incorporating prior-knowledge of TF-target gene interactions into the prior distribution over A is therefore essential in order to provide the inference algorithm with the information of which column corresponds to which TF.

With this identifiability issue in mind, we design an inference procedure that can be used on any prior-knowledge edge matrices, described in Fig. 1C. The first step is to randomly hold out prior information for some percentage of the genes in $p(A)$ (we choose 20%) by leaving the rows corresponding to these genes in A but setting the prior logistic normal means for all entries in these rows to be the same low number.

The second step is to carry out a hyperparameter search using this modified prior-knowledge matrix. The early stopping and model selection criteria are both the 'validation' AUPRC of the posterior point estimates of A, corresponding to the held out genes, against the entries for these genes in the full prior hyperparameter matrix. This step is motivated by the idea that inference using the selected hyperparameter configuration should yield a GRN whose columns correspond to the TF names that the user has assigned to these columns.

The third step is to choose the hyperparameter configuration corresponding to the highest validation AUPRC and perform inference using this configuration with the full

prior. An importance weighted estimate of the marginal log likelihood is used as the early stopping criterion for this step. The resulting approximate posterior provides the final posterior estimate of A.

Advantages of PMF-GRN

Existing methods almost always couple the description of the data generating process with the inference procedure used to obtain the final estimated GRN [31, 33, 34]. Designing a new model thus requires designing a new inference procedure specifically for that model, which makes it difficult to compare results across different models due to the discrepancies in their associated inference algorithms. Furthermore, this ad hoc nature of model building and inference algorithm design often leads to the lack of a coherent objective function that can be used for proper hyperparameter search as well as model selection and comparison, as evident in [31]. Heuristic model selection in available GRN inference methods presents the challenge of determining and selecting the optimal model in a given setting.

The proposed PMF-GRN framework decouples the generative model from the inference procedure. Instead of requiring a new inference procedure for each generative model, it enables a single inference procedure through (stochastic) gradient descent with the ELBO objective function, across a diverse set of generative models. Inference can easily be performed in the same way for each model. Through this framework, it is possible to define the prior and likelihood distributions as desired with the following mild restrictions: we must be able to evaluate the joint distribution of the observations and the latent variables, the variational distribution and the gradient of the log of the variational distribution.

The use of stochastic gradient descent in variational inference comes with a significant computational advantage. As each step of inference can be done with a small subset of observations, we can run GRN inference on a very large dataset without any constraint on the number of observations. This procedure is further sped up by using modern hardware, such as GPUs.

Under this probabilistic framework, we carry out model selection, such as choosing distributions and their corresponding hyperparameters, in a principled and unified way. Hyperparameters can be tuned with regard to a predefined objective, such as the marginal likelihood of the data or the posterior predictive probability of held out parts of the observations. We can further compare and choose the best generative model using the same procedure.

This framework allows us to encode any prior knowledge via the prior distributions of latent variables. For instance, we incorporate prior knowledge about TF-gene interactions as hyperparameters that govern the prior distribution over the matrix A. If prior knowledge about TFA is available, this can be similarly incorporated into the model via the hyperparameters of the prior distribution over U.

Because our approach is probabilistic by construction, inference also estimates uncertainty without any separate external mechanism. These uncertainty estimates can be used to assess the reliability of the predictions, i.e., more trust can be placed in interactions that are associated with less uncertainty. We verify this correlation between the degree of uncertainty and the accuracy of interactions in the experiments.

Overall, the proposed approach of probabilistic matrix factorization for GRN inference is scalable, generalizable and aware of uncertainty, which makes its use much more advantageous compared to most existing methods.

PMF-GRN recovers true interactions in simple eukaryotes

To evaluate PMF-GRN's ability to infer informative and robust GRNs, we leverage two single-cell RNA-seq datasets from the model organism *Saccharomyces cerevisiae* [8, 40]. This eukaryote, being relatively simple and extensively studied, provides a reliable gold standard [41] for assessing the performance of different GRN inference methods. We conduct three experiments to compare the performance of three state-of-the-art GRN inference methods, the Inferelator (AMuSR, BBSR, and StARS) [31], SCENIC [33], and CellOracle [34]. Throughout these experiments, each method is provided with the exact same single-cell RNA-seq datasets (GSE125162 [8]: N cells = 38, 225, GSE144820 [40]: N cells = 6118, combined: N cells = 44, 343 by M genes = 6763), prior-knowledge (M genes = 6885 by K TFs = 220), and gold standard (M genes = 993 by K TFs = 98).

In the first experiment, we infer GRNs for each of the two yeast datasets and average the posterior means of A to simulate a "multi-task" GRN inference approach. Using AUPRC, we demonstrate that PMF-GRN outperforms AMuSR, StARS, and SCENIC, while performing competitively with BBSR and CellOracle (Fig. 2A). We next combine the two expression datasets into one observation to test whether each method can discern the overall GRN accurately when data is not cleanly organized into tasks. This experiment reveals a substantial performance decrease for BBSR, indicating its dependence on organized gene expression tasks. This finding suggests potential challenges for BBSR in more complex organisms with less well-defined cell types or conditions. For benchmarking purposes, we provide two negative controls for each method, a GRN inferred without prior information (no prior), and a GRN inferred using shuffled prior information (shuffled prior). For all methods, these negative controls achieve an expected low AUPRC. It is essential to note that for CellOracle, an experiment with no prior information could not be performed. This is due to the fact that by design, CellOracle cannot learn regulatory edges that are not included in the prior information.

In our comparitive GRN inference analysis, we assess the number of edges predicted in common by each algorithm, on the individual *S. cerevisiae* datasets. We do so by computing the Intersection over Union (IoU) score, filtering each GRN to the top 25% of interactions to remove noisy predictions. Notably, PMF-GRN obtains an IoU score of 15.69%, outperforming alternative algorithms such as SCENIC (3.17%), AMuSR (12.46%), BBSR (14.56%), and StARS (11.78%). The superior performance of PMF-GRN can be attributed to an ability to discern meaningful regulatory interactions, thereby enriching the consensus among predictions. Importantly, our findings underscore a limitation of CellOracle, which achieves an IoU score of 30.28%. This algorithm, while proficient, can only ascertain edges present in the prior-knowledge matrix. Consequently, the two yeast GRNs inferred display high similarity, reflecting an inherent constraint. This characteristic imparts a degree of predictability to CellOracle, limiting its capacity to discover novel interactions beyond the established prior-knowledge. In contrast, PMF-GRNs IoU score is indicative of a more diverse and comprehensive set of common

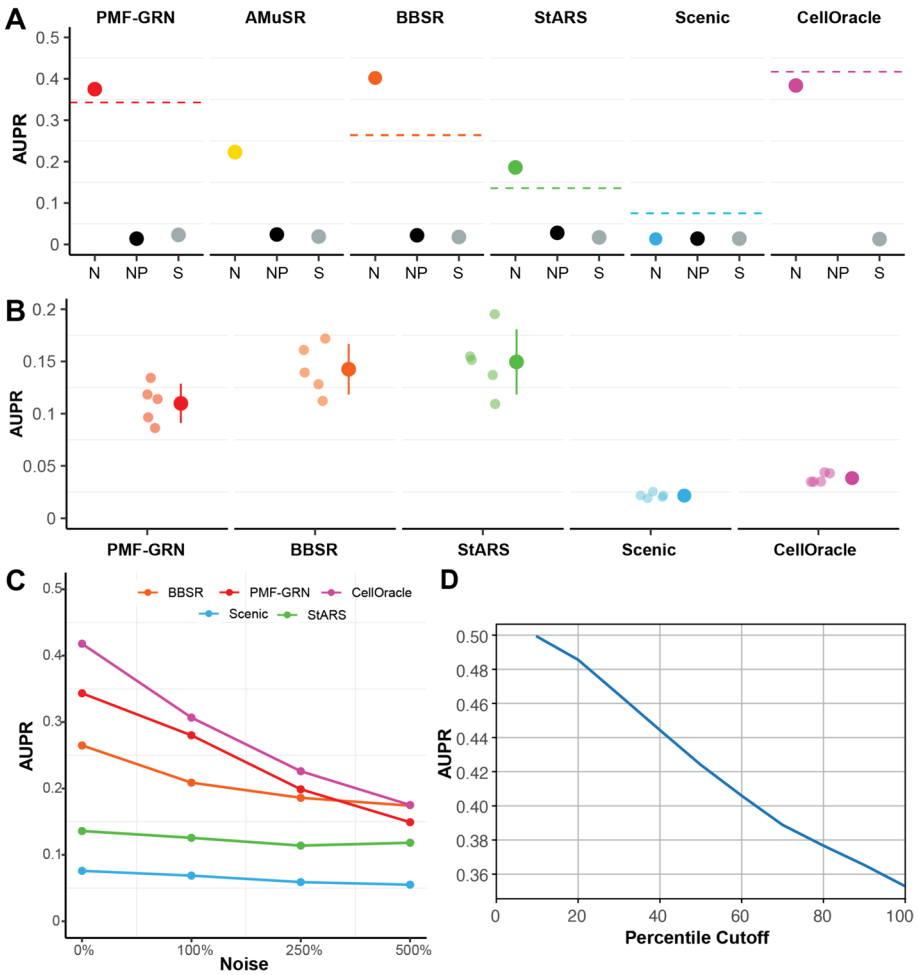

Fig. 2 GRN inference in *S. cerevisiae*. **A** Consensus network AUPR with a normal prior-knowledge matrix (N): PMF-GRN (red) performance compared to Inferelator algorithms (AMuSR in yellow, BBSR in orange, StARS in green), SCENIC (blue), and CellOracle (purple). Dashed line represents the baseline if expression data is combined. Negative controls: no prior information (NP——black) and shuffled prior information (S——gray). **B** 5-fold cross-validation baseline: each dot with low opacity represents one of the five experiments. Colored dots and lines depict the mean AUPR ± standard deviation for each GRN inference method. **C** GRNs inferred with increasing amounts of noise added to the prior. **D** Calibration results on *S.cerevisiae* (GSE144820 [8] only) dataset. Posterior means are cumulatively placed in bins based on their posterior variances. AUPRC for each of these bins is computed against the gold standard (see the "Methods" section for details)

edges. This highlights PMF-GRNs capability to capture nuanced regulatory relationships as a robust and versitle tool for GRN inference.

In a second experiment, we implement a 5-fold cross-validation approach to establish a baseline for each model. Cross-validation is crucial for evaluating the generalization ability of machine learning models like PMF-GRN, particularly in predicting TF-target gene interactions with limited data, a common scenario in experimental settings. To streamline the analysis, we combine the two *S. cerevisiae* single-cell RNA-seq datasets into a single observation matrix. The cross-validation process involves an 80–20% split of the gold standard, where a network is inferred using 80% as "prior-known information" and evaluated using the remaining 20%. This process is iterated five times with different random splits to yield meaningful results. We observe that PMF-GRN

outperforms SCENIC and CellOracle, while achieving similar performance to BBSR and StARS (Fig. 2B). We note that for this experiment, we are unable to implement the AMuSR algorithm as it is a multi-task inference approach that requires more than one task (dataset).

In a third experiment, we evaluate the robustness of each GRN inference method in the presence of noisy prior information. We conduct GRN inference with increasing levels of noise introduced into the prior knowledge. Specifically, the prior information begins with 1% non-zero edges, and we systematically introduce noise to observe the performance of each method. The noise levels are varied from zero noise (original prior, 1% non-zero edges), to 100% noise (resulting in 2% non-zero edges), 250% noise (3.5% non-zero edges), and 500% noise (6% non-zero edges). Our findings, illustrated in Fig. 2C, reveal that as the noise in the prior information increases, PMF-GRNs AUPRC experiences a slow decline, mirroring the behavior observed in CellOracle. Notably, PMF-GRN consistently outperforms BBSR, StARS, and SCENIC under these noise conditions, showcasing its robustness in accurately inferring GRNs from noisy priors. These results underscore PMF-GRN as one of the most robust approaches in the face of noisy prior information, thereby emphasizing its utility in practical applications.

To further emphasize PMF-GRN's robustness in a diverse number of settings, we perform the following two experiments. In the first experiment, we examine the performance of PMF-GRN using different sizes of downsampled yeast expression (Fig. 3A). The downsampling procedure involved reducing the expression data to sizes of 80%, 60%, 40%, and 20%, with each size undergoing random sampling five times to generate five distinct datasets per sample size. Remarkably, the AUPRC performance exhibits noteworthy stability across the downsampling variations. Despite the reduction in dataset size, PMF-GRN consistently demonstrates an ability to learn accurate GRNs as evidenced by the sustained AUPRC performance. These findings underscore the robustness of PMF-GRN, suggesting its reliability even under conditions of diminished dataset sizes, a critical consideration for practical applications where data availability may be limited.

In a subsequent experiment, we explore the impact of different cross-validation split sizes on hyperparameter tuning for PMF-GRN using the *S. cerevisiae* prior-knowledge (Fig. 3B). Four distinct cross-validation splits, ranging from 80% training and 20% validation to 20% training and 80% validation, were employed. For each split, we conducted a hyperparameter search across five samples, selecting the optimal hyperparameters based on the highest validation AUPRC. We then selected the best overall hyperparameters from each split to learn a GRN on the full dataset, in order to demonstrate the downstream effect of cross validation split choice on GRN inference. Surprisingly, our results revealed that the choice of cross-validation split size had a marginal impact on the overall performance of the inferred GRN. Specifically, the AUPRC values for the full GRN remained nearly unchanged regardless of whether an 80% train and 20% validation or 60% train and 40% validation split where employed. Even with more disparate splits, such as 40% train and 60% validation, or 20% train and 80% validation, the decrease in AUPRC was only minor. This implies that PMF-GRN exhibits robustness in hyperparameter selection, with the algorithm consistently converging to optimal settings across varying cross-validation scenarios.

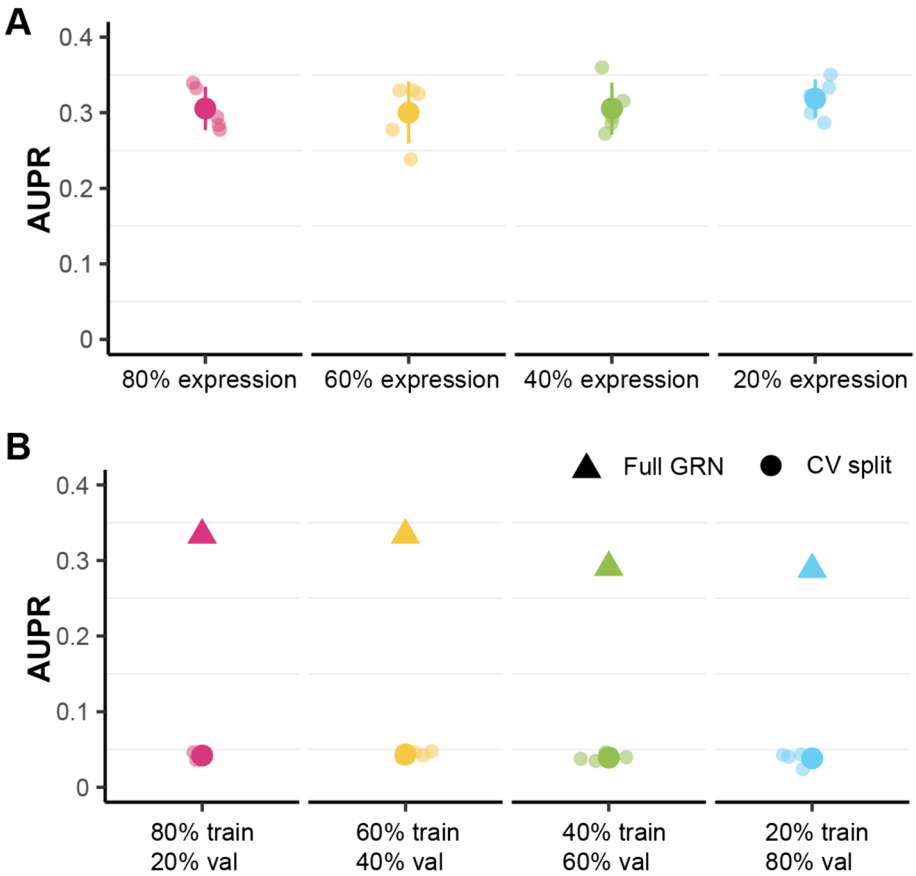

Fig. 3 **A** GRNs inferred by downsampling *S. cerevisiae* expression data. **B** Hyperparameter search performed on 4 different ratios of cross-validation. Dots represent validation AUPRC from hyperparameter search during cross-validation, triangle represents AUPRC from a GRN learned using the most optimal hyperparameters for each ratio

From our experiments on *S. cerevisiae* data, several key observations emerge. First, PMF-GRN consistently outperforms the Inferelator in recovering true GRNs, surpassing two Inferelator algorithms (AMuSR and StARS) and performing similarly to BBSR. Notably, when expression data is not separated into tasks, PMF-GRN outperforms BBSR. In comparison to CellOracle, PMF-GRN demonstrates competitive performance during normal inference and significantly outperforms CellOracle in cross-validation. However, PMF-GRN, in contrast to CellOracle, is not constrained to predicting edges solely within the confines of the prior-knowledge matrix. Furthermore, PMF-GRN consistently outperforms SCENIC across all experiments.

A second key observation is that our approach addresses the high variance associated with heuristic model selection among different inference algorithms. When implementing the Inferelator on *S. cerevisiae* datasets under normal conditions, AUPRCs fall within the range of 0.2 to 0.4, showcasing significant variability without a priori information to guide algorithm selection. This diversity among Inferelator algorithms constitutes heuristic model selection, as one cannot predict a priori which algorithm will perform better or discern the reasons behind their divergent performances. In contrast, our method offers reliable results grounded in a principled objective function, delivering competitive

performance akin to the best-performing Inferelator algorithm (BBSR) and CellOracle. This underscores the importance of a consistent and robust approach in the face of uncertainty associated with heuristic model selection among disparate algorithms.

To underscore the identifiability issue and affirm the utility of prior-known information, we showcase PMF-GRN's performance when prior information is unused (e.g., all prior logistic normal means of A set to the same low number). This process is replicated for other GRN inference algorithms by providing an empty prior. Additionally, we assess PMF-GRN's performance when prior-known TF-target gene interaction hyperparameters are randomly shuffled before building the prior distribution for A. The results, along with those for the Inferelator and CellOracle, indicate the capability of these approaches to accommodate such prior information effectively.

PMF-GRN provides well-calibrated uncertainty estimates

Through our inference procedure, we obtain a posterior variance for each element of A, in addition to the posterior mean. We interpret each variance as a proxy for the uncertainty associated with the corresponding posterior point estimate of the relationship between a TF and a gene. Due to our use of variational inference as the inference procedure, our uncertainty estimates are likely to be underestimates. However, these uncertainty estimates still provide useful information as to the confidence the model places in its point estimate for each interaction. We expect posterior estimates associated with lower variances (uncertainties) to be more reliable than those with higher variances.

In order to determine whether this holds for our posterior estimates, we cumulatively bin the posterior means of A according to their variances, from low to high. We then calculate the AUPRC for each bin as shown for the GSE125162 [8] *S.cerevisiae* dataset in Fig. 2D. We observe that the AUPRC decreases as the posterior variance increases. In other words, inferred interactions associated with lower uncertainty are more likely to be accurate than those associated with higher uncertainty. This is in line with our expectations as the more certain the model is about the degree of existence of a regulatory interaction, the more accurate it is likely to be, indicating that our model is well-calibrated.

PMF-GRN integrates single-cell multi-omic data for GRN and TFA inference in human PBMCs

We next evaluate PMF-GRN's ability to learn informative GRNs in a human cell line by focusing on peripheral blood mononuclear cells (PBMCs). PBMCs represent an essential component of the human immune system and consist of lymphocytes (CD4 and CD8 T cells, B cells, and natural killer cells), monocytes, and dendritic cells. Unraveling the distinct regulatory landscape of PBMCs is an essential task to provide insight into how these immune cells interact as well as coordinate to maintain homeostasis and respond effectively to infections.

To infer an informative and comprehensive PBMC GRN, we harness information from a large, paired single-cell RNA and ATAC-seq multi-omic dataset [42]. We adopt a prior-knowledge matrix of TF-target gene interactions (M genes = 18,557 by K TFs = 860) as previously constructed by [43] for GRN inference with this multi-omic dataset. In this work, the ATAC-seq data was used as a regulatory mask for ENCODE-derived TF ChIP-seq peaks. Regulatory associations were established

through the Inferelator-Prior package based on the proximity of TFs to their potential target genes within 50 kb upstream and 2 kb downstream of the gene transcription start site. We integrate this prior knowledge with the raw expression profiles of 11,909 PBMCs from a healthy donor to infer a global PBMC GRN and analyze the TFA profiles of eight annotated cell types and several families of immune TFs within this cell line.

We first investigate whether our predicted TFA clusters into distinct cell-type groups, as annotated by [42]. Using UMAP dimensionality reduction, we are able to determine a near clear distinction between each cell type within PBMCs (Fig. 4A). Interestingly, the TFA profiles for each of the T cell sub-types (CD4 T, CD8 T, and other T cells) are closely grouped together, suggesting that these cell types may have a similar lineage or TFA patterns, and may share common transcriptional programs or regulatory networks.

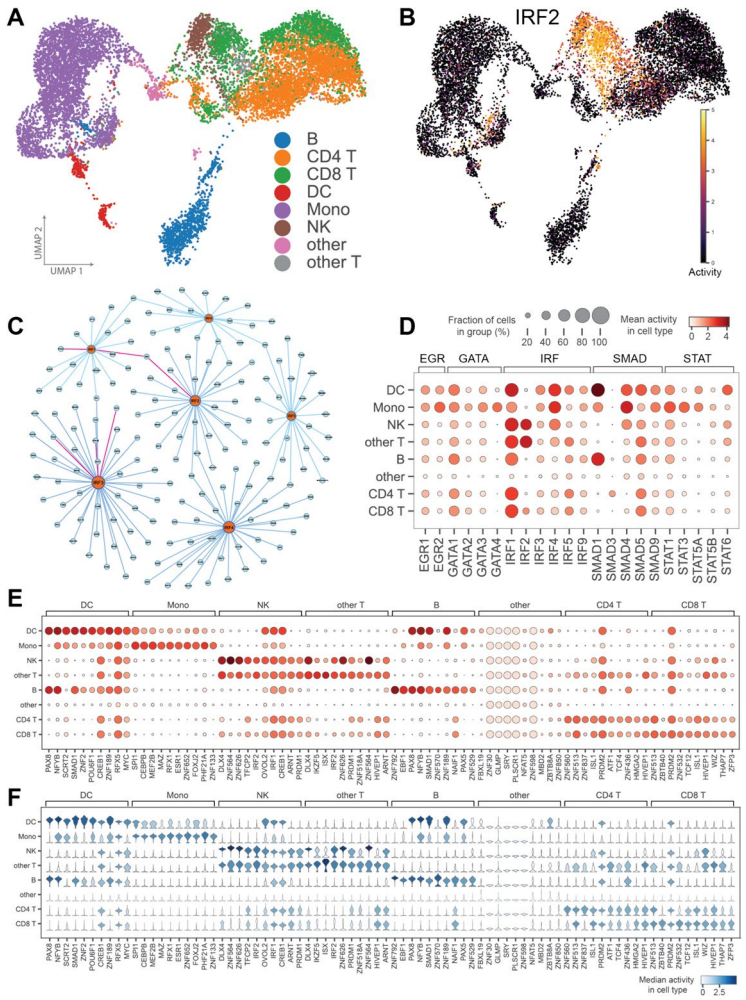

Fig. 4 GRN and TFA inference in PBMC. **A** UMAP projection of predicted TFA for each annotated PBMC cell type. **B** Predicted IRF2 TFA demonstrates high activity in NK and CD8 T cells. **C** GRN between IRF TFs and their targets. Pink edges indicate literature support for interaction. **D** Heat-map dot-plot depicting TFA of selected immune TFs across annotated PBMC cell types. **E** Heat-map dot-plot indicates ten most highly active TFs for each PBMC cell type. **F** Violin plot demonstrates corresponding distribution of TFA profiles for ten most highly active TFs

We next explore the activity profiles of specific immune TF families, starting with the family of TFs belonging to IRF. In PBMCs, IRF contributes to the activation of immune cells that modulate antiviral immunity. Notably, the UMAP projection for IRF2 indicates a high activity pattern within natural killer cells and CD8 T cells (Fig. 4B). Indeed, IRF2 is essential for the development and maturation of natural killer cells [44] and acts as a CD8 T cell nexus to translate signals from inflammatory tumor microenvironments [45].

In order to support our predicted TFA for the family of IRF TFs, we additionally investigate the regulatory interactions inferred by PMF-GRN (Fig. 4C). To do this within a reasonable scale, we first threshold our predicted GRN interactions (described in detail in the "Methods" section). Within our thresholded GRN, we predict regulatory edges between IRF1 and the target genes B2M and BTN3A1. IF1 has been documented as a transactivator of B2M [46], while BTN3A1, a defense-related gene, has been found to be upregulated via the IRF1 pathway [47]. Furthermore, we predict that IRF2 also regulates B2M. Supporting evidence demonstrates that IRF2 has been shown to directly bind to genes linked to the interferon response and MHC class I antigen presentation, including B2M [48]. Finally, we predict regulatory edges between IRF3 and GPR108, RNF5, and TRAF2. GPR108 has been shown to be a regulator of type I interferon responses by targeting IRF3 [49]. Evidence supporting the interaction between IRF3 and RNF5 indicates that RNF5 has an inhibitory effect on the activation of IRF3 [50]. Lastly, TRAF3 has been shown to be a critical component in the activation of IRF3 during the innate immune response to viral infections [51].

In addition to the IRF TFs, several other families of TFs, such as SMAD, STAT, GATA, and EGR, collectively play pivotal roles in PBMCs. These roles contribute to a wide spectrum of functions, including antiviral responses (IRF), fine-tuning immune responses (SMAD), immune cell development (GATA), immediate early responses to signals (EGR), and central regulation of T cells, B cells, and natural killer cells (STAT). Their coordinated activities orchestrate the complex interplay of immune cells, enabling PBMCs to effectively respond to diverse stimuli and maintain immune homeostasis.

Similarly to IRF, we also explore edges in our thresholded PBMC GRN for these immune TFs to identify regulatory edges supported by literature. Of the five families of immune TFs that we investigate, we find supporting literature for 60 regulatory edges predicted by PMF-GRN. We provide these literature supported edges, along with their supporting references in Additional file 1: Table S10. Additionally, we provide a graph representation of each immune TF GRN in Additional file 2: Fig. S2.

We next explore the TFA profiles of each of these immune TFs within the eight PBMC cell types. In Fig. 4D, a heat-map dot-plot provides a visual representation of TFA for each immune TF family across the different PBMC cell types. In particular, we observe that within the IRF family, IRF1 is highly active in CD4 T cells. Previous studies have confirmed the pivotal role of IRF1 in CD4+ T cells, where it is essential for promoting the development of TH1 cells through the activation of the Il12rb1 gene [52]. Additionally, SMAD5 is predicted as highly active in B cells. SMAD5 is a key component of the TGF-β signaling pathway, and has been shown to play a crucial role in maintaining immune homeostasis in B cells [53]. We provide a UMAP of the TFA profiles for each of these immune TFs in Additional file 2: Fig. S3.

We further explore our predicted TFA profiles from our global PBMC GRN and calculate the ten most active TFs across the eight distinct cell types. For this experiment, we provide a heat-map dot-plot demonstrating the mean TFA value for each of the top TFs as well as a corresponding violin plot depicting the distributions of these TFA profiles (Fig. 4E and F). Visualizing these distinct activity profiles provides a concise and informative snapshot of the predominant TFs contributing significant transcriptional activity within each cell population. For example, within B cells we observe high activity for the TF PAX5. PAX5 is known to play a crucial role in B cell development by guiding the commitment of lymphoid progenitors to the B lymphocyte lineage while simultaneously repressing inappropriate genes and activating B lineage-specific genes [54].

For each annotated cell-type in the PBMC dataset, a set of marker genes were provided. From our ten most active TFs per cell-type analysis combined with their edges to target genes from our thresholded GRN, we find that several of these TFs are predicted to regulate marker genes. For example, within dendritic cells, the marker gene HLA-DQA1 is predicted to be regulated by the TFs SMAD1 and RFX5; the marker gene HLA-DPA1 is predicted to be regulated by ZNF2 and RFX5; and the marker gene HLA-DRB1 is predicted to be regulated by RFX5. Within CD4 T cells, the marker gene LTB is predicted to be regulated by the TF ZNF436. Within natural killer cells, the marker gene PRF1 is predicted to be regulated by ZNF626. Finally, within B cells, the marker gene BANK1 is predicted to be regulated by the TFs ZNF792, EBF1, PAX8, and PAX5; and the marker gene HLA-DQA1 is predicted to be regulated by the TF SMAD1.

From the predicted edges between a snapshot of highly active TFs and annotated marker genes, we find the following supporting evidence. The regulatory relationship between RFX5 and HLA-DQA1 involves the inability of RFX5 to bind to the proximal promoter region of HLA-DQA1, potentially due to DNA methylation, hindering the assembly of active regulatory regions [55]. Additionally, EBF1 orchestrates direct transcriptional regulation of BANK1, leading to the observed downregulation of BANK1 expression [56].

Pairing the intensity (dot-plot) with the distribution (violin plot) of TFA offers a comprehensive view of the key TFs guiding our regulatory networks. This approach illuminates the variability in their activity levels across diverse immune cell populations, providing a nuanced understanding of the transcriptional dynamics in PBMCs. This information can be used to guide insights into the functional specialization and diversity of immune cells within PBMCs. Furthermore, this comparison provides a sound starting point for exploring the commonalities and differences in the transcriptional regulation of various immune cell populations.

Evaluating PMF-GRN with BEELINE synthetic data

We next evaluated PMF-GRN using synthetic datasets curated from the BEELINE benchmark [37]. This benchmark provides six synthetic networks, linear (LI), linear long (LL), cycle (CY), bifurcating (BF), trifurcating (TF), and bifurcating converging (BFC). In repetitions of ten, expression datasets of increasing cell sizes (e.g., $n = 100, 200, 500, 2000$, and 5000) were generated by sampling. Using these generated expression datasets, as well as the provided reference GRNs, we inferred 300 GRNs using

PMF-GRN (Fig. 5A). For each of the six synthetic datasets, PMF-GRN outperforms the BEELINE baseline, represented in Fig. 5A with a black dashed line.

To further evaluate PMF-GRN, we calculate the AUPRC ratio of PMF-GRN over the baseline random predictor to compare to the similarly computed ratios in the original BEELINE paper (Fig. 5B). We observe that for the linear, cycle, and bifurcating converging, PMF-GRN achieves competitive AUPRC ratios in comparison to the original methods used in the BEELINE benchmark. Interestingly, PMF-GRN does not perform competitively on long linear. This could be due to a number of factors, such as the larger number of intermediate genes introducing additional complexity which PMF-GRN struggles to capture. Alternatively, the extended trajectory introduces a higher-dimensional space, which could present a challenge for our matrix factorization based approach to effectively decompose the data into meaningful latent factors. This presents an interesting avenue of consideration when developing future probabilistic matrix factorization approaches for GRN inference.

Discussion

In this paper, we introduce a robust framework for probabilistic matrix factorization, optimized through automatic variational inference, to infer GRNs from single-cell gene expression data. A distinctive feature of our approach is the decoupling of the data generation model from the inference procedure, providing unprecedented flexibility. This decoupling allows for modifications to the latent variables and their distributions, without altering the inference process. Such flexibility facilitates the seamless integration of diverse sequencing datasets and modeling assumptions. Unlike previous methods, our framework eliminates the need to define a new inference procedure for each specific dataset or biological context when building new models.

PMF-GRN not only offers a flexible and unified approach to GRN inference but also provides a principled methodology for model selection and hyperparameter configuration. The use of a consistent objective function and inference procedure across all generative models streamlines the process of hyperparameter search, reducing ambiguity

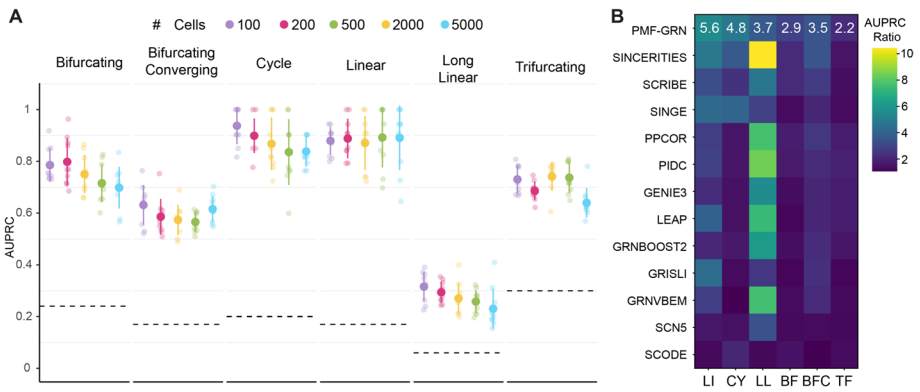

Fig. 5 PMF-GRN performance on BEELINE synthetic GRN data **A** PMF-GRN inference performance with half of the ground truth provided as prior network information and the remaining half provided as a gold standard for evaluation. Dashed lines are the expected baseline of a random predictor. **B** AUPRC ratio over the baseline random predictor for PMF-GRN in comparison to each of the GRN inference methods used in the original BEELINE benchmark

present in methods like the Inferelator. By conducting hyperparameter search across different generative models, we identify configurations corresponding to optimal values of our objective function, minimizing the reliance on heuristic model selection.

To validate the effectiveness of our approach, we applied PMF-GRN to infer GRNs from single-cell *S. cerevisiae* gene expression, comparing results with state-of-the-art single-cell GRN inference methods such as the Inferelator, SCENIC, and CellOracle. Our method demonstrates competitive, if not superior, performance in terms of AUPRC, in each experiment performed. Here, PMF-GRN provides a stable and reliable inferred GRN without the need for heuristic model selection or data separation into tasks.

Cross-validation experiments further support the robustness of PMF-GRN, BBSR, and StARS, indicating their ability to generalize well to new data without overfitting. In contrast, SCENIC and CellOracle exhibited poor performance during cross-validation, suggesting potential issues with generalizability. Notably, we assessed the robustness of each algorithm against increasing noise in the prior-knowledge, identifying PMF-GRN and CellOracle as the most resilient to noisy priors. This resilience ensures the reliability of inferred GRNs even in the presence of uncertain prior knowledge.

Our model uniquely provides well-calibrated uncertainty estimates alongside point estimates for each interaction in the final GRN. The evaluation of uncertainty estimates demonstrated that as the posterior variance decreases, the AUPRC increases, indicating that the model is well-calibrated. Biologists can leverage these uncertainty estimates for downstream experimental validation, placing more trust in estimates with lower posterior variance. Finally, the linear scalability of our models computational cost with the number of cells enables its application to single-cell RNA-seq datasets of any size.

Our investigation into PMF-GRN's application to human PBMCs provides insightful findings into the regulatory landscape of these essential immune cells. Leveraging a comprehensive multi-omic dataset, we demonstrate that our approach integrates single-cell RNA and well-curated prior knowledge derived from ATAC-seq data. The resulting global PBMC GRN unveils distinct TFA profiles for eight annotated cell types and various immune TF families. Through UMAP dimensionality reduction, we observe clear clustering of TFA profiles. Focusing on the IRF family, we identify specific TF-target gene interactions supported by literature, shedding light on regulatory relationships critical for immune responses. Extension to other immune TF families reveals their orchestrated activities within PBMCs, contributing to antiviral responses, immune cell development, and the regulation of T cells, B cells, and natural killer cells. By exploring predicted edges between active TFs and marker genes, we establish connections between regulatory networks and cellular functions. The combined dot-plot and violin plot visualization strategy provides a nuanced understanding of TF activities, offering a valuable resource for deciphering the intricate transcriptional dynamics in PBMCs. This detailed exploration sets the stage for further investigations into the functional specialization and diversity of immune cells within the PBMC population, with implications for advancing our understanding of immune responses and disease mechanisms.

In the context of synthetic datasets curated from the BEELINE benchmark, PMF-GRN demonstrates robust performance across various network structures. Outperforming the BEELINE baseline across different synthetic networks, PMF-GRN consistently achieves competitive AUPRC ratios compared to the original methods used in the BEELINE

benchmark. Notably, PMF-GRN's competitive performance is observed in linear, cycle, and bifurcating converging structures. However, challenges arise in the long linear structured synthetic data, suggesting potential limitations in capturing the complex dynamics of extended trajectories. Factors such as the increased number of intermediate genes and a higher-dimensional space may contribute to this limitation. This observation opens avenues for future development of probabilistic matrix factorization approaches, encouraging exploration of methods better suited for intricate network structures. The overall success of PMF-GRN in diverse synthetic network scenarios underscores its versatility and effectiveness in inferring GRNs, promising broad applicability in deciphering complex biological systems and regulatory interactions.

Conclusion

In conclusion, the PMF-GRN framework provides a flexible and principled approach for inferring GRNs from single-cell gene expression data. By decoupling the model and inference procedure, PMF-GRN enables easy integration of new and various sequencing datasets as well as modeling assumptions without the need for defining a new inference procedure. Additionally, PMF-GRN provides a principled approach for model selection through hyperparameter search, reducing the need for heuristic model selection. Overall, PMF-GRN consistently yields high-performing competitive results compared to other state-of-the-art single-cell GRN inference methods with a reliable gold standard and is robust to cross validation, noisy priors, and downsampling. Furthermore, PMF-GRN provides well-calibrated uncertainty estimation, enabling a reliable set of results for downstream experimental validation.

We envision many possible directions for future work to design a better algorithm for inferring GRNs under our framework. This framework could be extended to explicitly model multiple expression matrices and their batch effects. We could probabilistically model prior information for A obtained from ATAC-seq and TF motif databases and include this as part of the probabilistic model over which we carry out inference. Evaluating the posterior estimates of the direction of transcriptional regulation, provided by the matrix B, could provide a useful benchmark for the computational estimation of TF activation and repression. Research could also be carried out on improved self-supervised objectives for hyperparameter selection.

Future work could also focus on how to use results from our framework to guide experimental wet-lab work. For example, the uncertainty quantification provided by our model could open up new research directions in active learning for GRN inference. Highly ranked, uncertain interactions could be experimentally tested and the results fed back into the prior hyperparameter matrix for A. Inference with this updated matrix would ideally yield a better posterior GRN estimate. Posterior estimates of TFA provided by our model could be useful to wet lab scientists, as this quantity provides information about possible post-transcriptional modifications, which are currently challenging to measure experimentally.

Most importantly, the study of GRN inference is far from complete. GRN inference approaches have thus far required new computational models and assumptions in order to keep up with relevant sequencing technologies. It is thus essential to develop a model that can be easily adapted to new biological datasets as they become available, without having

to completely re-build each model. We have therefore proposed PMF-GRN as a modular, principled, probabilistic approach that can be easily adapted to both new and different biological data without having to design a new GRN inference method.

Methods

Model details

We index cells, genes and TFs using $n \in \{1, \cdots, N\}$, $m \in \{1, \cdots, M\}$, and $k \in \{1, \cdots, K\}$, respectively. We treat each cell's expression profile W_n as a random variable, with local latent variables U_n and d_n, and global latent variables (that are shared among all cells) σ_{obs} and $V = A \odot B$. We use the following likelihood for each of our observations:

$$p(W_n|U, V, \sigma_{obs}, d) = \mathcal{N}\left(d_n * U_n V^\top, \sigma_{obs}^2\right).$$

We assume that U, V, σ_{obs}, and d are independent, i.e., $p(U, V, \sigma_{obs}, d) = p(U)p(V)p(\sigma_{obs})p(d)$. In addition to our i.i.d assumption over the rows of U and d, we also assume that the entries of U_n are mutually independent, and that all entries of A and B are mutually independent. We choose a lognormal distribution for our prior over U and a logistic normal distribution for our prior over d:

$$\begin{aligned} p(\log(U_{nk})) &= \mathcal{N}\left(\mu_u, \sigma_u^2\right), \\ p(\text{logit}(d_n)) &= \mathcal{N}(0, 9) \end{aligned}$$

where $\mu_u \in \mathbb{R}$ and $\sigma_u \in \mathbb{R}^+$.

We use a logistic normal distribution for our prior over A, a normal distribution for our prior over B and a logistic normal distribution for our prior over σ_{obs}:

$$\begin{aligned} p(\text{logit}(A_{mk})) &= \mathcal{N}\left(\text{logit}\left(\text{clip}\left(\bar{A}_{mk}, a_{\max}, a_{\min}\right)\right), \sigma_a^2\right), \\ p(B_{mk}) &= \mathcal{N}\left(0, \sigma_b^2\right). \\ p(\log(\sigma_{obs})) &= \mathcal{N}(0, 1), \end{aligned}$$

where $\bar{A}_{mk} \in \{0, 1\}$, $a_{\max} \in (0, 1)$, $a_{\min} \in (0, 1)$, $\sigma_a \in \mathbb{R}_{>0}$, $\text{clip}(\bar{A}_{mk}, a_{\max}, a_{\min}) = \max(\min(\bar{A}_{mk}, a_{\max}), a_{\min})$, and $\sigma_b \in \mathbb{R}_{>0}$. $\bar{A}_{mk}$ is given by a pipeline that is used by other methods such as the Inferelator. The pipeline leverages ATAC-seq and TF binding motif data to provide binary initial guesses of gene-TF interactions. $a_{\max}$ and $a_{\min}$ are hyperparameters that determine how we clip these binary values before transforming them to the logit space.

For our approximate posterior distribution, we enforce independence as follows:

$$q(U, A, B, \sigma_{obs}, d) = q(U)q(A)q(B)q(\sigma_{obs})q(d).$$

We impose the same independence assumptions on each approximate posterior as we do for its corresponding prior. Specifically, we use the following distributions:

$$
\begin{aligned}
q(\log(U_{nk})) &= \mathcal{N}\left(\tilde{U}_{nk}, \tilde{\sigma}^2_{U_{nk}}\right) \\
q(\text{logit}(d_n)) &= \mathcal{N}\left(\tilde{d}_n, \tilde{\sigma}^2_{d_n}\right) \\
q(\text{logit}(A_{mk})) &= \mathcal{N}\left(\tilde{A}_{mk}, \tilde{\sigma}^2_{A_{mk}}\right) \\
q(B_{mk}) &= \mathcal{N}\left(\tilde{B}_{mk}, \tilde{\sigma}^2_{B_{mk}}\right) \\
q(\log(\sigma_{obs})) &= \mathcal{N}\left(\tilde{o}, \tilde{\sigma}^2_{o}\right),
\end{aligned}
$$

where the parameters on the right hand sides of the equations are called variational parameters; $\tilde{U}_{nk}$, $\tilde{d}_n$, $\tilde{A}_{mk}$, $\tilde{B}_{mk}$, $\tilde{o} \in \mathbb{R}$ and $\tilde{\sigma}_{U_{nk}}$, $\tilde{\sigma}_{d_n}$, $\tilde{\sigma}_{A_{mk}}$, $\tilde{\sigma}_{B_{mk}}$, $\tilde{\sigma}_o \in \mathbb{R}^+$. To avoid numerical issues during optimization, we place constraints on several of these variational parameters.

Inference

We perform inference on our model by optimizing the variational parameters to maximize the ELBo. In doing so, we minimise the KL-divergence between the true posterior and the variational posterior. In practice, to help with addressing the latent factor identifiability issue, we use a modified version of the ELBo where the prior and posterior terms are weighted by a constant $\beta \geq 1$ [57]:

$$
\begin{aligned}
\mathbb{E}_{U,A,B,\sigma_{obs},d \sim q(U,A,B,\sigma_{obs},d)}[\, &\log p(W|U, V = A \odot B, \sigma_{obs}, d) \\
&+ \beta(\log p(U, A, B, \sigma_{obs}, d) - \log q(U, A, B, \sigma_{obs}, d))]
\end{aligned}
$$

Inference is carried out using the Adam optimizer with learning rate 0.1 and beta values of 0.9 and 0.99. We clip gradient norms at a value of 0.0001. We set $a_{\min} = 0.005$, $a_{\max} = 0.995$, $\sigma_b^2 = 1$, and $\mu_u = 0$. We vary σ_a and σ_u as hyperparameters that control the strengths of the priors over A and U, respectively. We also vary β as a hyperparameter.

We choose a hyperparameter configuration using validation AUPRC as the objective function as well as the early stopping metric. We hold out hyperparameters for $p(A)$ for a fraction of the genes. We do this by setting $\bar{A}_{mk} = 0$ for m corresponding to these genes for all k. During inference, we regularly obtain posterior point estimates for these entries and measure the AUPRC against the original values of these entries as given in the full prior. This quantity is known as the validation AUPRC.

Once we have picked the hyperparameter configuration corresponding to the best validation AUPRC, we perform inference with this model using the full prior without holding out any information. We use an importance weighted estimate of the marginal log likelihood as our early stopping criterion:

$$
\log p(W) = \log\left(\mathbb{E}_{U,A,B,\sigma_{obs},d \sim q(U,A,B,\sigma_{obs},d)}\left[\frac{p(W|U, A, B, \sigma_{obs}, d)p(U, A, B, \sigma_{obs}, d)}{q(U, A, B, \sigma_{obs}, d)}\right]\right),
$$

where the expectation is computed using simple Monte Carlo and the log-$\sum$-exp trick is used to avoid numerical issues.

Computing summary statistics for the posterior

After training the model, we use $\tilde{A}$ and $\tilde{\sigma}_A$, the variational parameters of $q(A)$, to obtain a mean and a variance for each entry of A. Since $q(A)$ is logistic normal, it admits no closed form solution for the mean and variance. We therefore use Simple Monte Carlo, i.e., we sample each entry of A several times from its posterior distribution and then compute the sample mean and sample variance from these samples. We use each mean as a posterior point estimate of the probability of interaction between a TF and a gene, and its associated variance as a proxy for the uncertainty associated with this estimate.

Calculating AUPRC

The gold standards for the datasets used in this paper do not necessarily perfectly overlap with the genes and TFs that make up the rows and columns of A as defined by the prior hyperparameters, i.e., there may be genes and TFs in the gold standard with a recorded interaction or lack of interaction, that do not appear in our model at all because they are not present in the prior. The reverse is also true: the prior may contain genes and TFs that are not in the gold standard. For this reason, we compute the AUPRC using one of two methods: "keep all gold standard" or "overlap," which correspond to evaluating only interactions that are present in the gold standard or only interactions that are present in both the gold standard and the prior/posterior. We present results with "keep all gold standard" AUPRC as the evaluation metric when comparing our model to the Inferelator in Fig. 2. For our evaluation of uncertainty calibration (Fig. 2D), we use the overlap AUPRC so that bins containing a lower number of posterior means do not have artificially deflated AUPRCs (see the "Evaluating calibration of posterior uncertainty" section for further information).

Evaluating calibration of posterior uncertainty

We create 10 bins, corresponding to the lowest 10%, 20%, 30%, and so on of posterior variances. We place the posterior point estimates of TF-gene interactions associated with these variances into these bins and then calculate the "overlap AUPRC" for each bin using the corresponding gold standard. The AUPRC for each bin is calculated using those interactions that are in the gold standard and also in the bin. We use such a cumulative binning scheme because using a non-cumulative scheme could result in some bins having very small numbers of posterior interactions that are present in the gold standard, which would lead to noisier estimates of the AUPRC.

Inference and evaluation on multiple observations of W

The Inferelator method applies two scRNA-seq experiments separately on *S. cerevisiae*, with each resulting in a distinct model. These models are used to infer TF-gene interaction matrices, which are then sparsified. The final matrix is obtained by taking the intersection of the two matrices and retaining only the entries that are non-zero in both matrices. In our approach, we also train a separate model on each expression matrix and obtain a posterior mean matrix for A for each of them. To obtain the final posterior mean matrix for A, we average the posterior mean matrices from each model. While this approach works well, future research could focus on explicitly modeling separate expression matrices within the model, as mentioned in the "Discussion" section.

Measuring the impact of prior hyperparameters

We evaluate the utility of each of the prior hyperparameter matrices used in our experiments. In Fig. 2A and Additional file 2: Fig. S1, we present with grey dots the AUPRCs achieved when performing inference using shuffled prior hyperparameters for A. This corresponds to randomly assigning to each row (gene) of A, the prior hyperparameters that correspond to a different row of A. Shuffling the hyperparameters should lead to worse performance, as the posterior estimates should then also be shuffled, whereas the row/column labels for the posterior will remain unshuffled. For the "no prior" setting, shown with black dots in the figures, we set $\bar{A}_{mk} = 0 \, \forall \, m, k$. The difference in AUPRC achieved using the unshuffled vs shuffled or no hyperparameters measures the usefulness of the provided hyperparameters for the inference task on the dataset in question.

Cross-validation

For *S. cerevisiae*, we perform a five-fold cross validation experiment (Fig. 2B). Cross-validation is performed by partitioning the gold standard into an 80–20% split, where 80% of the data represents prior-known information to be used as a prior for $p(A)$, and the remaining 20% is treated as the gold standard for evaluation. This process is repeated five times to generate five random splits of the data in order to robustly evaluate GRN inference. It is important to note that PMF-GRN performs hyperparameter search before inferring a final GRN within each cross-validation split. For each of the five partitioned cross-validation folds, the 80%, or prior portion, is further split into 80% train and 20% test for hyperparameter search and evaluation. Once the optimal hyperparameters have been determined, the initial 80% split is treated as the training data, while the remaining 20%, which was not seen during hyperparameter selection, is used for evaluation.

Intersection over Union

Intersection over Union (IoU) scores were computed using the GRN learned by each algorithm for the two *S. cerevisiae* expression datasets. For each GRN, we calculate and retain the top 25% of predicted edges in order to obtain the best estimates for each algorithm and elimate noisier predictions. For each algorithm, we compute both the intersection and the union of the GRN interactions predicted from the two *S. cerevisiae* datasets. Dividing the Intersection by the Union allows us to obtain a score indicating how similar the two inferred GRNs are for each algorithm.

Downsampling expression

For *S. cerevisiae*, in repetitions of five, we randomly sample the *S. cerevisiae* expression matrix on the cell axis to obtain downsampled expression dataset sizes of 80%, 60%, 40%, and 20%. We perform a hyperparameter search, using an 80% training-20% validation split of the prior-knowledge matrix, on each of these five expression matrices for each sample size. Using these hyperparameters, we infer GRNs for each repetition within each split to obtain our final downsampled GRNs.

Exploring the effect of cross-validation ratios on hyperparameter selection

To effectively explore the influence of cross-validation split size on obtaining optimal hyperparameters for GRN inference, we methodically separate our *S. cerevisiae* prior-knowledge into 4 different split sizes. These splits consist of 80% training-20% validation, 60% training-40% validation, 40% training-60% validation, and 20% training-80% validation. For each split size, we obtain 5 random training and validation splits to ensure robust results. We then perform hyperparameter search across each 5 random splits for each split size. Using the best overall hyperparameters for each split size, we infer a final GRN to demonstrate the impact each particular split had on obtaining the optimal hyperparameters for the final GRN.

Datasets and preprocessing

We inferred each GRN using a single-cell RNA-seq expression matrix, a TF-target gene connectivity matrix, and a gold standard for bench-marking purposes. We modeled the single-cell expression matrices based on the raw UMI counts obtained from sequencing for the *S. cerevisiae* and PBMC datasets, which were therefore not normalized for the purpose of this work. For the two *B. subtilis* datasets used in this work, we demonstrate the effect of different normalization and scaling techniques and convert all data used to integers in order to create a single-cell-like dataset. We further obtained binary TF-gene matrices representing prior-known interactions, which served as prior hyperparameters over $\mathbf{A}$ and were derived from the YEASTRACT and SubtiWiki databases, as well as from [43] for PBMC. We acquired a gold standard for *S. cerevisiae* our datasets from independent work which is detailed below.

Saccharomyces cerevisiae

We used two raw UMI count expression matrices for the organism *S. cerevisiae* obtained from NCBI GEO (GSE125162 [8] and GSE144820 [40]). For this well studied organism, we employed the YEASTRACT [58, 59] literature derived network of TF-target gene interactions to be used as a prior over A in both *S. cerevisiae* networks. A gold standard for *S. cerevisiae* was additionally obtained from a previously defined network [41] and used for bench-marking our posterior network predictions. We note that the gold standard is roughly a reliable subset of the YEASTRACT prior. Additional interactions in the prior can still be considered to be true but have less supportive evidence than those in the gold standard.

Peripheral blood mononuclear cells

We used a paired multi-omic single-cell RNA-seq and ATAC-seq dataset for PBMC obtained from [42]. The single-cell expression matrix contained 11,909 cells. The prior-knowledge matrix was constructed using the ATAC-seq data from this multi-omic dataset, constructed and described in detail by [43]. The prior-knowledge matrix is 18,557 genes by 860 TFs and contains 0.5% non-zero edges.

Due to the complex and dynamic nature of PBMCs, a gold standard is currently unavailable for this cell line. To evaluate our inferred network, we implement a 5-fold cross-validation procedure where our chromatin accessibility-based prior is split

into 5 random sets, where 80% is used as prior knowledge and 20% is used as the gold standard for evaluation. We then took the intersection of the regulatory edges inferred across each of the 5 fold cross-validation experiments and filtered to retain the highest quality edges, obtaining a prediction probability of 90% or higher.

BEELINE synthetic datasets

We used the BEELINE synthetic expression datasets [37] without modification. Reference GRNs were transformed into cross-tab matrices in order to use this information for prior-knowledge and gold standard evaluation. We used 50% of the reference GRN as the prior and the remaining 50% as the gold standard, as was similarly done in [31].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03226-6.

Additional file 1. Additional tables [62–115].

Additional file 2. Additional figures.

Additional file 3. Additional experiments [116–120].

Additional file 4. Additional methods, including data generation and curation [121–123].

Additional file 5. The peer review history.

Acknowledgements
We thank members of the Bonneau lab for insightful discussions and feedback on this manuscript. We also thank the the staff of the NYU IT High Performance Computing and Flatiron Institute Scientific Computing Core. CSG is grateful to Yanis Bahroun, Daniel Berenberg, and Maggie Beheler-Amass for insightful discussions related to this work.

Review history
The review history is available as Additional file 5.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
CSG and KC contributed to the conceptualization of the project. OM and KC designed the probabilistic model. OM implemented the PMF-GRN software, experiments, and validation. CSG implemented the PMF-GRN experiments, validation, and Inferelator software. OM, CSG, and KC contributed to the methodology, software, validation, formal analysis, visualization, and writing and preparation of the original draft. CSG contributed to the data curation. KC and RB contributed to the supervision, project administration, and funding acquisition.

Funding
This work was supported by Samsung Advanced Institute of Technology (under the project ***Next Generation Deep Learning: From Pattern Recognition to AI);*** NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science; the National Institutes of Health (RM1HG011014, R01NS116350, R01NS118183, R01AI130945); and the Simons Foundation. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2234660. 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
The datasets used in this work are publicly available [60]. They are referenced in the "Methods" section and are available through https://github.com/nyu-dl/pmf-grn. Code, inferred GRNs [61], and inference and evaluation scripts can be found at https://github.com/nyu-dl/pmf-grn.

Declarations

Ethics approval and consent to participate
Ethics approval were not needed for the study.

Competing interests
The authors declare that they have no competing interests.

Received: 30 March 2023 Accepted: 26 March 2024
Published online: 08 April 2024

References

1. Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models-a review. Biosystems. 2009;96(1):86–103.
2. Chai LE, Loh SK, Low ST, Mohamad MS, Deris S, Zakaria Z. A review on the computational approaches for gene regulatory network construction. Comput Biol Med. 2014;48:55–65.
3. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008;9(10):770–80.
4. Äijö T, Lähdesmäki H. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. Bioinformatics. 2009;25(22):2937–44.
5. Nachman I, Regev A, Friedman N. Inferring quantitative models of regulatory networks from expression data. Bioinformatics. 2004;20(suppl_1):248–56.
6. Burdziak C, Azizi E, Prabhakaran S, Pe'er D. A nonparametric multi-view model for estimating cell type-specific gene regulatory networks. 2019. arXiv preprint arXiv:1902.08138. https://arxiv.org/abs/1902.08138.
7. Allaway KC, Gabitto MI, Wapinski O, Saldi G, Wang CY, Bandler RC, et al. Genetic and epigenetic coordination of cortical interneuron development. Nature. 2021;597(7878):693–7.
8. Jackson CA, Castro DM, Saldi GA, Bonneau R, Gresham D. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. Elife. 2020;9:e51254.
9. Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, et al. A validated regulatory network for Th17 cell specification. Cell. 2012;151(2):289–303.
10. Ji Z, He L, Regev A, Struhl K. Inflammatory regulatory network mediated by the joint action of NF-kB, STAT3, and AP-1 factors is involved in many human cancers. Proc Natl Acad Sci. 2019;116(19):9453–62.
11. Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, et al. Dynamic regulatory network controlling TH17 cell differentiation. Nature. 2013;496(7446):461–8.
12. Mercatelli D, Scalambra L, Triboli L, Ray F, Giorgi FM. Gene regulatory network inference resources: a practical overview. Biochim Biophys Acta (BBA) - Gene Regul Mech. 2020;1863(6):194430.
13. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS ONE. 2010;5(9):e12776.
14. Wang Y, Joshi T, Zhang XS, Xu D, Chen L. Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics. 2006;22(19):2413–20.
15. Chang C, Ding Z, Hung YS, Fung PCW. Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. Bioinformatics. 2008;24(11):1349–58.
16. Dufva M. Introduction to microarray technology. DNA Microarrays Biomed Res Methods Protocol. 2009;529:1–22.
17. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
18. Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014;42(14):8845–60.
19. Akers K, Murali T. Gene regulatory network inference in single-cell biology. Curr Opin Syst Biol. 2021;26:87–97.
20. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):1–35.
21. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front Genet. 2019;10:317.
22. Ochs MF, Fertig EJ. Matrix factorization for transcriptional regulatory network inference. In: 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). San Diego: IEEE; 2012. pp. 387–96. https://doi.org/10.1109/CIBCB.2012.6217256.
23. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci. 2000;97(18):10101–6.
24. Moloshok TD, Klevecz R, Grant JD, Manion FJ, Speier W IV, Ochs MF. Application of Bayesian decomposition for analysing microarray data. Bioinformatics. 2002;18(4):566–75.
25. Kim PM, Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. 2003;13(7):1706–18.
26. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci. 2004;101(12):4164–9.
27. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics. 2005;21(21):3970–5.
28. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2016;32(1):1–8.
29. Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc Natl Acad Sci. 2018;115(30):7723–8.
30. Hu X, Hu Y, Wu F, Leung RWT, Qin J. Integration of single-cell multi-omics for gene regulatory network inference. Comput Struct Biotechnol J. 2020;18:1925–38.
31. Skok Gibbs C, Jackson CA, Saldi GA, Tjärnberg A, Shah A, Watters A, et al. High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. Bioinformatics. 2022;38(9):2519–28.
32. Jansen C, Ramirez RN, El-Ali NC, Gomez-Cabrero D, Tegner J, Merkenschlager M, et al. Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. PLoS Comput Biol. 2019;15(11):e1006555.

33. Van de Sande B, Flerin C, Davie K, De Waegeneer M, Hulselmans G, Aibar S, et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat Protoc. 2020;15(7):2247–76.
34. Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA. Dissecting cell identity via network inference and in silico gene perturbation. Nature. 2023;1–10.
35. Äijö T, Bonneau R. Biophysically motivated regulatory network inference: progress and prospects. Hum Hered. 2016;81(2):62–77.
36. Mnih A, Salakhutdinov RR. Probabilistic matrix factorization. Adv Neural Inf Process Syst. 2007;20.
37. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17(2):147–54.
38. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. J Am Stat Assoc. 2017;112(518):859–77.
39. Ranganath R, Gerrish S, Blei D. Black box variational inference. arXiv preprint arXiv:1401.0118 (2013).
40. Jariani A, Vermeersch L, Cerulus B, Perez-Samper G, Voordeckers K, Van Brussel T, et al. A new protocol for single-cell RNA-seq reveals stochastic gene expression during lag phase in budding yeast. Elife. 2020;9:e55320.
41. Tchourine K, Vogel C, Bonneau R. Condition-specific modeling of biophysical parameters advances inference of regulatory networks. Cell Rep. 2018;23(2):376–88.
42. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87.
43. Tjärnberg A, Beheler-Amass M, Jackson CA, Christiaen LA, Gresham D, Bonneau R. Structure-primed embedding on the transcription factor manifold enables transparent model architectures for gene regulatory network and latent activity inference. Genome Biol. 2024;25(1):24.
44. Persyn E, Wahlen S, Kiekens L, Van Loocke W, Siwe H, Van Ammel E, et al. IRF2 is required for development and functional maturation of human NK cells. Front Immunol. 2022;13:1038821.
45. Lukhele S, Abd Rabbo D, Guo M, Shen J, Elsaesser HJ, Quevedo R, et al. The transcription factor IRF2 drives interferon-mediated CD8+ T cell exhaustion to restrict anti-tumor immunity. Immunity. 2022;55(12):2369–85.
46. Gobin SJ, Biesta P, Van den Elsen PJ. Regulation of human β2-microglobulin transactivation in hematopoietic cells. Blood J Am Soc Hematol. 2003;101(8):3058–64.
47. Pietz G, De R, Hedberg M, Sjöberg V, Sandström O, Hernell O, et al. Immunopathology of childhood celiac disease-key role of intestinal epithelial cells. PLoS ONE. 2017;12(9):e0185025.
48. Mercado N, Schutzius G, Kolter C, Estoppey D, Bergling S, Roma G, et al. IRF2 is a master regulator of human keratinocyte stem cell fate. Nat Commun. 2019;10(1):4676.
49. Zhao M, Zhang Y, Qiang L, Lu Z, Zhao Z, Fu Y, et al. A Golgi-resident GPR108 cooperates with E3 ubiquitin ligase Smurf1 to suppress antiviral innate immunity. Cell Rep. 2023;42(6):112655.
50. Zhong B, Zhang L, Lei C, Li Y, Mao AP, Yang Y, et al. The ubiquitin ligase RNF5 regulates antiviral responses by mediating degradation of the adaptor protein MITA. Immunity. 2009;30(3):397–407.
51. Yu JH, Moon EY, Kim J, Koo JH. Identification of small GTPases that phosphorylate IRF3 through TBK1 activation using an active mutant library screen. Biomol Ther. 2023;31(1):48.
52. Kano Si, Sato K, Morishita Y, Vollstedt S, Kim S, Bishop K, et al. The contribution of transcription factor IRF1 to the interferon-γ–interleukin 12 signaling axis and TH1 versus TH-17 differentiation of CD4+ T cells. Nat Immunol. 2008;9(1):34–41.
53. Malhotra N, Kang J. SMAD regulatory networks construct a balanced immune system. Immunol. 2013;139(1):1–10.
54. Cobaleda C, Schebesta A, Delogu A, Busslinger M. Pax5: the guardian of B cell identity and function. Nat Immunol. 2007;8(5):463–70.
55. Majumder P, Boss JM. DNA methylation dysregulates and silences the HLA-DQ locus by altering chromatin architecture. Genes Immun. 2011;12(4):291–9.
56. Treiber T, Mandel EM, Pott S, Györy I, Firner S, Liu ET, et al. Early B cell factor 1 regulates B cell gene networks by activation, repression, and transcription-independent poising of chromatin. Immunity. 2010;32(5):714–25.
57. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-VAE: learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations. 2017. https://openreview.net/forum?id=Sy2fzU9gl. Accessed 2022.
58. Monteiro PT, Oliveira J, Pais P, Antunes M, Palma M, Cavalheiro M, et al. YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. Nucleic Acids Res. 2020;48(D1):D642–9.
59. Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, et al. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. Nucleic Acids Res. 2018;46(D1):D348–53.
60. Skok Gibbs C, Mahmood O, Bonneau R. Cho K pmf-grn: datasets. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25451986.
61. Skok Gibbs C, Mahmood O, Bonneau R, Cho K. pmf-grn: gene regulatory networks. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25444810.
62. Abou El Hassan M, Huang K, Eswara MB, Xu Z, Yu T, Aubry A, et al. Properties of STAT1 and IRF1 enhancers and the influence of SNPs. BMC Mol Biol. 2017;18(1):1–19.
63. Au-Yeung N, Mandhana R, Horvath CM. Transcriptional regulation by STAT1 and STAT2 in the interferon JAK-STAT pathway. Jak-stat. 2013;2(3):e23931.
64. Johnstone AL, Andrade NS, Barbier E, Khomtchouk BB, Rienas CA, Lowe K, et al. Dysregulation of the histone demethylase KDM6B in alcohol dependence is associated with epigenetic regulation of inflammatory signaling pathways. Addict Biol. 2021;26(1):e12816.
65. Mayumi A, Tomii T, Kanayama T, Mikami T, Tanaka K, Yoshida H, et al. Activation of the STAT1-BCL-2/MCL-1 axis in leukemic cells carrying a SPAG9-JAK2 fusion. Blood. 2021;138:4326.
66. Kumari S, Bonnet MC, Ulvmar MH, Wolk K, Karagianni N, Witte E, et al. Tumor necrosis factor receptor signaling in keratinocytes triggers interleukin-24-dependent psoriasis-like skin inflammation in mice. Immunity. 2013;39(5):899–911.

67. Andoh A, Shioya M, Nishida A, Bamba S, Tsujikawa T, Kim-Mitsuyama S, et al. Expression of IL-24, an activator of the JAK1/STAT3/SOCS3 cascade, is enhanced in inflammatory bowel disease. J Immunol. 2009;183(1):687–95.
68. Edsbäcker E, Serviss JT, Kolosenko I, Palm-Apergi C, De Milito A, Tamm KP. STAT3 is activated in multicellular spheroids of colon carcinoma cells and mediates expression of IRF9 and interferon stimulated genes. Sci Rep. 2019;9(1):536.
69. Roy R, Dagher A, Butterfield C, Moses MA. ADAM12 is a novel regulator of tumor angiogenesis via STAT3 signaling. Mol Cancer Res. 2017;15(11):1608–22.
70. Kim JH, Hedrick S, Tsai WW, Wiater E, Le Lay J, Kaestner KH, et al. CREB coactivators CRTC2 and CRTC3 modulate bone marrow hematopoiesis. Proc Natl Acad Sci. 2017;114(44):11739–44.
71. Nguyen-Jackson H, Panopoulos AD, Zhang H, Li HS, Watowich SS. STAT3 controls the neutrophil migratory response to CXCR2 ligands by direct activation of G-CSF-induced CXCR2 expression and via modulation of CXCR2 signal transduction. Blood J Am Soc Hematol. 2010;115(16):3354–63.
72. Wu S, Fu J, Dong Y, Yi Q, Lu D, Wang W, et al. Golph3 promotes glioma progression via facilitating JAK2-STAT3 pathway activation. J Neuro-Oncol. 2018;139:269–79.
73. Wei T, Lambert PF. Role of IQGAP1 in carcinogenesis. Cancers. 2021;13(16):3940.
74. Nie XH, Qiu S, Xing Y, Xu J, Lu B, Zhao SF, et al. Paeoniflorin regulates NEDD4L/STAT3 pathway to induce ferroptosis in human glioma cells. J Oncol. 2022;2022:6093216.
75. Keuthan C, Santiago C, Ash JD. STAT3 is a potential genetic modifier of photoreceptor gene expression during stress. Investig Ophthalmol Vis Sci. 2019;60(9):466.
76. Wei X, Yu L, Li Y. PBX1 promotes the cell proliferation via JAK2/STAT3 signaling in clear cell renal carcinoma. Biochem Biophys Res Commun. 2018;500(3):650–7.
77. Liu W, Geng C, Li X, Li Y, Song S, Wang C. Downregulation of SLC9A8 promotes epithelial-mesenchymal transition and metastasis in colorectal cancer cells via the IL6-JAK1/STAT3 signaling pathway. Dig Dis Sci. 2023;68(5):1873–84.
78. Shibata M, Ooki A, Inokawa Y, Sadhukhan P, Ugurlu MT, Izumchenko E, et al. Concurrent targeting of potential cancer stem cells regulating pathways sensitizes lung adenocarcinoma to standard chemotherapy. Mol Cancer Ther. 2020;19(10):2175–85.
79. Li L, Zhang R, Liu Y, Zhang G. ANXA4 activates JAK-STAT3 signaling by interacting with ANXA1 in basal-like breast cancer. DNA Cell Biol. 2020;39(9):1649–56.
80. Nagel S, Pommerenke C, Meyer C, Kaufmann M, Drexler HG, MacLeod RA. Deregulation of polycomb repressor complex 1 modifier AUTS2 in T-cell leukemia. Oncotarget. 2016;7(29):45398.
81. Lessard S, Gatof ES, Beaudoin M, Schupp PG, Sher F, Ali A, et al. An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria susceptibility. J Clin Investig. 2017;127(8):3065–74.
82. Katsumura KR, Bresnick EH, Group GFM. The GATA factor revolution in hematology. Blood J Am Soc Hematol. 2017;129(15):2092–102.
83. Gao J, Chen YH, Peterson LC. GATA family transcriptional factors: emerging suspects in hematologic disorders. Exp Hematol Oncol. 2015;4:1–7.
84. Zhang Z, Parker MP, Graw S, Novikova LV, Fedosyuk H, Fontes JD, et al. O-GlcNAc homeostasis contributes to cell fate decisions during hematopoiesis. J Biol Chem. 2019;294(4):1363–79.
85. Johnson KD, Boyer ME, Kang JA, Wickrema A, Cantor AB, Bresnick EH. Friend of GATA-1-independent transcriptional repression: a novel mode of GATA-1 function. Blood J Am Soc Hematol. 2007;109(12):5230–3.
86. Chakrabarti S, Kabra M, Mandal AK, Senthil S, Kaur I. The transcription factors PBX1 and GATA1 are regulated by the mutation profiles of CYP1B1 in primary congenital glaucoma. Investig Ophthalmol Vis Sci. 2016;57(12):803.
87. Wu W, Xu N, Zhou X, Liu L, Tan Y, Luo J, et al. Integrative genomic analysis reveals cancer-associated gene mutations in chronic myeloid leukemia patients with resistance or intolerance to tyrosine kinase inhibitor. OncoTargets Ther. 2020;13:8581–91.
88. Kobayashi M, Funayama R, Ohnuma S, Unno M, Nakayama K. Wnt-β-catenin signaling regulates ABCC 3 (MRP 3) transporter expression in colorectal cancer. Cancer Sci. 2016;107(12):1776–84.
89. Kong X, Wang Q, Li J, Li M, Deng F, Li C. Mammaglobin, GATA-binding protein 3 (GATA3), and epithelial growth factor receptor (EGFR) expression in different breast cancer subtypes and their clinical significance. Eur J Histochem EJH. 2022;66(2):3315.
90. Blumenthal SG, Aichele G, Wirth T, Czernilofsky AP, Nordheim A, Dittmer J. Regulation of the human interleukin-5 promoter by Ets transcription factors: Ets1 and ets2, but not elf-1, cooperate with gata3 and htlv-i tax1. J Biol Chem. 1999;274(18):12910–6.
91. Liu X, Bai F, Wang Y, Wang C, Chan HL, Zheng C, et al. Loss of function of GATA3 regulates FRA1 and c-FOS to activate EMT and promote mammary tumorigenesis and metastasis. Cell Death Dis. 2023;14(6):370.
92. Li K, Wu Y, Li Y, Yu Q, Tian Z, Wei H, et al. Landscape and dynamics of the transcriptional regulatory network during natural killer cell differentiation. Genom Proteomics Bioinforma. 2020;18(5):501–15.
93. Yang X, Wang C, Lin Y, Zhang P. Identification of crucial hub genes and differential T cell infiltration in idiopathic pulmonary arterial hypertension using bioinformatics strategies. Front Mol Biosci. 2022;9:800888.
94. Fitch SR, Kapeni C, Tsitsopoulou A, Wilson NK, Göttgens B, de Bruijn MF, et al. Gata3 targets Runx1 in the embryonic haematopoietic stem cell niche. IUBMB Life. 2020;72(1):45–52.
95. Liao MH, Lin PI, Ho WP, Chan WP, Chen TL, Chen RM. Participation of GATA-3 in regulation of bone healing through transcriptional upregulation of bcl-xL expression. Exp Mol Med. 2017;49(11):e398–e398.
96. Hintze M, Prajapati RS, Tambalo M, Christophorou NA, Anwar M, Grocott T, et al. Cell interactions, signals and transcriptional hierarchy governing placode progenitor induction. Development. 2017;144(15):2810–23.
97. Arroyo N, Villamayor L, Díaz I, Carmona R, Ramos-Rodríguez M, Muñoz-Chápuli R, et al. GATA4 induces liver fibrosis regression by deactivating hepatic stellate cells. JCI insight. 2021;6:150059.
98. San Roman AK, Aronson BE, Krasinski SD, Shivdasani RA, Verzi MP. Transcription factors GATA4 and HNF4A control distinct aspects of intestinal homeostasis in conjunction with transcription factor CDX2. J Biol Chem. 2015;290(3):1850–60.

99. Yu TY, Chen XX, Liu QW, Ma FF, Huang HL, Zhou L, et al. Loss of GATA4 C-terminus by p. S335X mutation modulates coronary artery vascular smooth muscle cell phenotype. Mediators of Inflammation. 2021;2021:Article ID 3698386.
100. Khalid AB, Pence J, Suthon S, Lin J, Miranda-Carboni GA, Krum SA. GATA4 regulates mesenchymal stem cells via direct transcriptional regulation of the WNT signalosome. Bone. 2021;144:115819.
101. Liu Y, Harmelink C, Peng Y, Chen Y, Wang Q, Jiao K. CHD7 interacts with BMP R-SMADs to epigenetically regulate cardiogenesis in mice. Hum Mol Genet. 2014;23(8):2145–56.
102. Gao Y, Chen Q, Yue W. LAPTM5 protein can regulate TGF-β mediated MAPK and smad signaling pathways in ovarian cancer cell. Ann Oncol. 2019;30:v9.
103. Jung GS, Hwang YJ, Choi JH, Lee KM. Lin28a attenuates TGF-β-induced renal fibrosis. BMB Rep. 2020;53(11):594.
104. Jiang X, Tan J, Wen Y, Liu W, Wu S, Wang L, et al. MSI2-TGF-β/TGF-β R1/SMAD3 positive feedback regulation in glioblastoma. Cancer Chemother Pharmacol. 2019;84:415–25.
105. Hua F, Mu R, Liu J, Xue J, Wang Z, Lin H, et al. TRB3 interacts with SMAD3 promoting tumor cell migration and invasion. J Cell Sci. 2011;124(19):3235–46.
106. Hill CS. Transcriptional control by the SMADs. Cold Spring Harb Perspect Biol. 2016;8(10):a022079.
107. Wang X, Liao P, Fan X, Wan Y, Wang Y, Li Y, et al. CXXC5 associates with Smads to mediate TNF-α induced apoptosis. Curr Mol Med. 2013;13(8):1385–96.
108. Rochette L, Dogon G, Zeller M, Cottin Y, Vergely C. GDF15 and cardiac cells: current concepts and new insights. Int J Mol Sci. 2021;22(16):8889.
109. Chen L, Toke NH, Luo S, Vasoya RP, Fullem RL, Parthasarathy A, et al. A reinforcing HNF4-SMAD4 feed-forward module stabilizes enterocyte identity. Nat Genet. 2019;51(5):777–85.
110. Wang Y, Jiang L, Mo X, Lan Y, Yang X, Liu X, et al. Megakaryocytic Smad4 regulates platelet function through Syk and ROCK2 expression. Mol Pharmacol. 2017;92(3):285–96.
111. Trelford CB, Di Guglielmo GM. Canonical and non-canonical TGFβ signaling activate autophagy in an ULK1-dependent manner. Front Cell Dev Biol. 2021;9:712124.
112. Hsu LJ, Hong Q, Chen ST, Kuo HL, Schultz L, Heath J, et al. Hyaluronan activates Hyal-2/WWOX/Smad4 signaling and causes bubbling cell death when the signaling complex is overexpressed. Oncotarget. 2017;8(12):19137.
113. Chen L, Wang S, Zhou Y, Wu X, Entin I, Epstein J, et al. Identification of early growth response protein 1 (EGR-1) as a novel target for JUN-induced apoptosis in multiple myeloma. Blood J Am Soc Hematol. 2010;115(1):61–70.
114. Liu R, Liu L, Bian Y, Zhang S, Wang Y, Chen H, et al. The dual regulation effects of ESR1/NEDD4L on SLC7A11 in breast cancer under ionizing radiation. Front Cell Dev Biol. 2022;9:772380.
115. Wong KM, Song J, Wong YH. CTCF and EGR1 suppress breast cancer cell migration through transcriptional control of Nm23-H1. Sci Rep. 2021;11(1):491.
116. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science. 2012;335(6072):1103–6.
117. Arrieta-Ortiz ML, Hafemeister C, Bate AR, Chu T, Greenfield A, Shuster B, et al. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. Mol Syst Biol. 2015;11(11):839.
118. Michna RH, Zhu B, Mäder U, Stülke J. Subti Wiki 2.0-an integrated database for the model organism Bacillus subtilis. Nucleic Acids Res. 2016;44(D1):D654–D662.
119. Zhu B, Stülke J. Subti Wiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis. Nucleic Acids Res. 2018;46(D1):D743–8.
120. Pedreira T, Elfmann C, Stülke J. The current state of Subti Wiki, the database for the model organism Bacillus subtilis. Nucleic Acids Res. 2022;50(D1):D875–82.
121. Faria JP, Overbeek R, Taylor RC, Conrad N, Vonstein V, Goelzer A, et al. Reconstruction of the regulatory network for Bacillus subtilis and reconciliation with gene expression data. Front Microbiol. 2016;7:275.
122. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
123. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the international AAAI conference on web and social media. vol. 3. 2009. pp. 361–2.

Publisher's Note