Probabilistic Matrix Factorization for Gene Regulatory Network Inference

Omar Mahmood*
Center for Data Science
New York University

Claudia Skok Gibbs*
Center for Data Science
New York University

Richard Bonneau
Genomics & Systems Biology,
New York University
and
Prescient Design, Genentech

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

Abstract

Gene regulatory network inference methods are often heuristically designed for specific datasets and biological problems, making it challenging to include extra information and compare against other algorithms. Here, we propose the use of probabilistic matrix factorization for gene regulatory network (PMF-GRN) inference from single-cell gene expression datasets of any size, incorporating experimental evidence into prior distributions over latent factors. We use variational inference to infer GRNs, enabling hyperparameter search and direct comparison to other generative models. We evaluate our method using *Saccharomyces cerevisiae* and *Bacillus subtilis*, benchmarking against database-derived gold standard GRNs. We find that PMF-GRN infers GRNs more accurately than the current state-of-the-art method, while reducing the need for heuristic model selection. We demonstrate the necessity of incorporating prior information into any matrix factorization approach to GRN inference. Finally, we find that PMF-GRN's uncertainty estimates are well-calibrated.

Introduction

It is an essential problem in systems biology to extract information from biological sequencing data in order to unravel the complexity of gene regulation and understand the mechanisms controlling cellular processes and heterogeneity [1]. Gene regulatory networks (GRN) describing regulatory relationships between transcription factors (TF) and their target genes [2] have proven to be useful models for describing diverse cell functions [3, 4, 5, 6]. These functions include, but are not limited to, development [7], response to environmental perturbations [8], and dysregulation in the context of disease [9, 10, 11].

GRNs, however, cannot be directly measured with current sequencing technology. Instead, methods must be developed to piece together snapshots of transcriptional processes in order to reconstruct the regulatory landscape [12]. Typically, GRNs are inferred using transcript information obtained from single-cell RNA-seq [13], bulk RNA-seq [14, 15, 13], or previously microarray data [16, 17, 18, 19, 20, 21]. However, inferring an informative GRN from observed sequencing data presents a myriad

*Equal contribution

of challenges [22] despite the increasing availability of relevant data. On one hand, instrumental limitations inherently introduce noise, artifacts, dropouts, and issues related to sequencing depth [23]. On the other hand, there are modeling challenges when attempting to estimate the latent variables of interest from such biological data. It is therefore essential to implement a GRN inference method capable of modeling constraints introduced by high-throughput technologies, in addition to providing a normative approach for inferring regulatory relationships.

In this work, we propose to adapt and use probabilistic matrix factorization [24] for inferring gene regulatory networks, to which we refer as PMF-GRN. We evaluate PMF-GRN using the model organisms *Saccharomyces cerevisiae* and *Bacillus subtilis* to infer informative GRNs and benchmark our predictions using database derived gold standards. Through PMF-GRN, we are able to address current GRN inference challenges by formulating our problem to directly model the data generating distributions of TF-target gene regulatory relationships and transcription factor activity (TFA). We implement matrix factorization to decompose observed single-cell gene expression into latent factors representing these two variables of interest. We infer posterior distributions over these latent factors while incorporating known TF-target gene relationships, which can be obtained from genomics databases or ATAC-seq used in combination with TF motif databases, into our prior distribution. Imposing prior distributions further introduces modeling flexibility as we can incorporate new high-throughput datasets or assumptions as desired. We carry out inference with our model on a GPU using stochastic gradient descent (SGD), which allows us to easily scale to a large number of cells in the observed gene expression dataset. To address the issues caused by instrumental limitations, we include data generating distributions for both sequencing depth and noise into our estimated posterior distributions. Incorporating such information provides a more accurate depiction of the data generating process, and better clarifies the underlying GRN.

Through our normative approach to GRN inference, we are able to decouple the data generating process from the inference procedure, and provide a principled approach to model selection. This approach is unlike existing heuristic approaches, which lack systematic objective functions and require new inference procedures. Heuristic approaches also lack a unified model selection procedure to determine which GRN model is optimal in a given setting. We demonstrate competitive performance to one such existing method, the Inferelator [13]. We focus on direct comparisons between GRNs inferred by our method against similarly inferred GRNs by the Inferelator. We choose the Inferelator for comparison as it is well benchmarked against other available single-cell GRN inference methods [25, 26], and further computes the same latent parameters, TFA and TF-target gene interactions, as part of its pipeline. We show that our method achieves better recovery of the underlying GRN in terms of the Area Under the Precision Recall Curve (AUPRC). We also demonstrate the necessity of using prior information for obtaining meaningful results from any matrix factorization approach that tackles GRN inference. Finally, our model, unlike previous methods, provides uncertainty estimates over the TF-target gene interactions that it infers. We evaluate and find that these uncertainty estimates are well-calibrated, as the accuracy of inferred TF-target gene interactions increases when the associated uncertainty decreases.

Results

PMF-GRN Model Overview

The goal of our probabilistic matrix factorization approach is to decompose observed gene expression into latent factors, representing TF activity (TFA) and TF target gene regulatory interactions, as shown in Figure 1 (A). 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.

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 by analyzing other data types, such as the measurement of chromosomal accessibility, TF-binding motif databases, and direct measurement of TF-binding along the chromosome (see 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-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 [27, 28], where we approximate the true posterior distributions with tractable, approximate (variational) posterior distributions.

We minimize 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 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.

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 estimate of the GRN [13, 26, 25]. 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/comparison, as evident in [13]. Heuristic model selection in available GRN inference methods presents the challenge of determining and selecting an 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 above, across a diverse set of generative models. Inference can easily be performed in the same way for each such 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 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 type of procedure.

This framework allows us to encode any prior knowledge via the prior distributions of latent variables. For instance, we incorporate prior domain knowledge about TF-gene interactions as hyperparameters

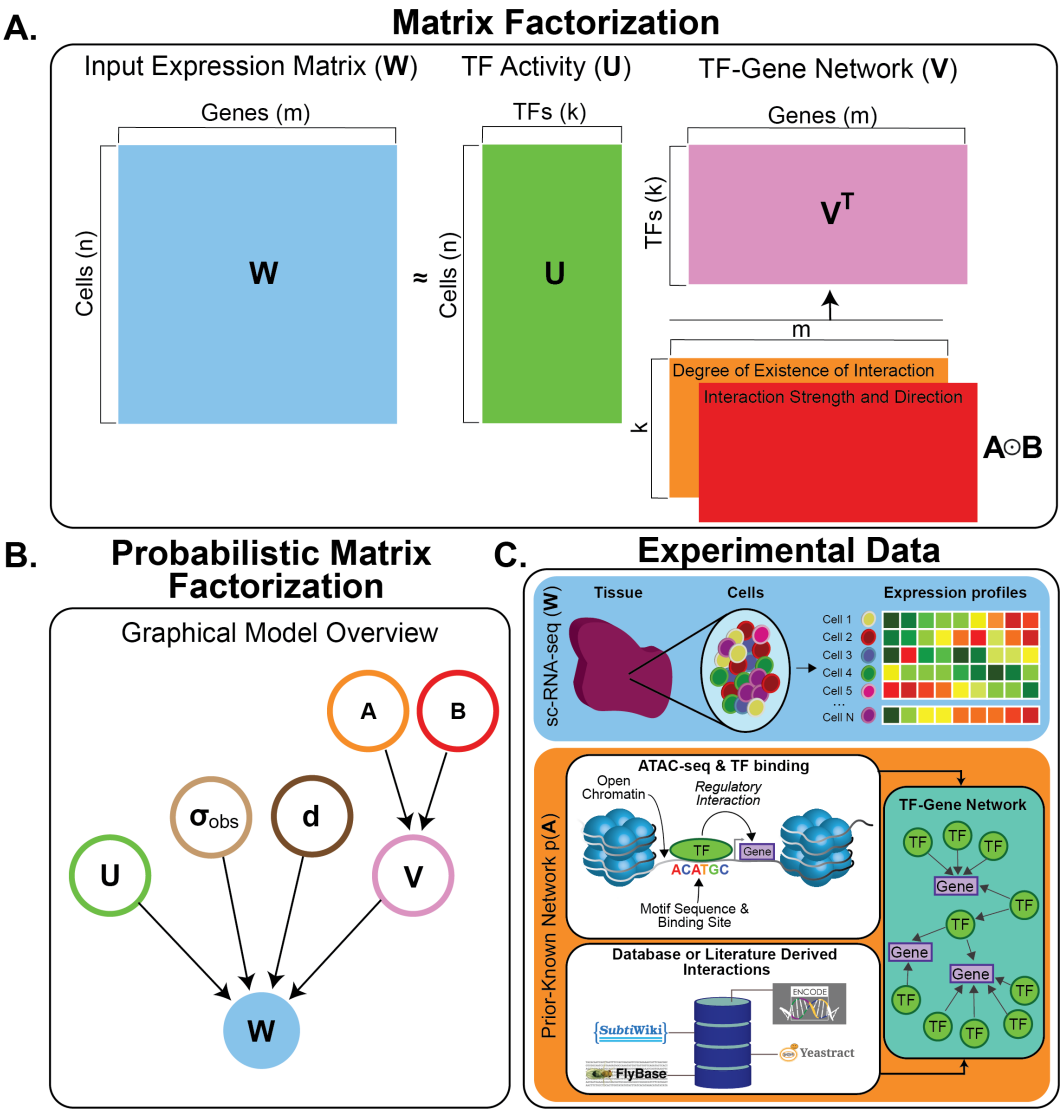

Figure 1: PMF-GRN schematic overview. (**A**) The matrix factorization set up takes a gene expression matrix **W** as input, and decomposes it into two latent factor matrices **U** and **V**, representing TF activity and TF-gene interactions respectively. **V** is further decomposed into **A** and **B**, representing degree of existence of interaction, and strength and direction of interaction, respectively. (**B**) Graphical model representation of PMF where **W** can be further modeled by observation noise σ_{obs} and sequencing depth d. (**C**) Experimental data inputs for PMF-GRN includes sc-RNA-seq data for **W**. ATAC-seq paired with peak calling and motif enrichment can be used to create a known-prior network for p(**A**). If experimental information for p(**A**) is unavailable, databases can be used to construct a known-prior network.

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 demonstrate PMF-GRN's ability to infer GRNs, we use scRNA-seq datasets obtained from two *S.cerevisiae* experiments, for which we have a reliable gold standard. *S.cerevisiae* is a relatively simple and well studied eukaryote, which allows us to evaluate our model's performance. We infer two GRNs individually in this experiment and average the posterior means of A to build a final combined network. In addition to the effectiveness of PMF-GRN, this experiment also demonstrates how we can use PMF-GRN on multiple observation matrices to learn a consensus GRN by simple averaging.

Matrix factorization based GRN is only identifiable up to the permutation of the columns of U and V, as identical permutation of the columns of both U and V does not change the resulting observation matrix W. 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 as the order in which the inference algorithm implicitly assigns TFs to these columns is $\frac{1}{K!}$, which is essentially 0 for any reasonable value of K. Incorporating prior knowledge of TF-Gene interactions into the prior distribution over A is therefore essential to give the inference algorithm information about which column corresponds to which TF.

Keeping this identifiability issue in mind, we design an inference procedure that can be used with any dataset. 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 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.

In order to benchmark our results, we compare our predictions against those of a comparable GRN inference algorithm, the Inferelator [13]. We choose the Inferelator for network comparison, because; 1. it is a widely used inference pipeline that scales well to very large single-cell datasets, 2. it has been extensively compared to other methods and tested in real biological settings, and 3. it can incorporate multiple data-types to enable a direct comparison. The Inferelator comes with several inference algorithms, including BBSR, StARS-LASSO and AMuSR. It is, however, not possible to choose the best inference algorithm among these, unlike the proposed PMF-GRN. Therefore we test and report the results using all three algorithms.

The results of our experiments on the *S. cerevisiae* datasets are in Figure 2. (The numbers used to create this figure are given in Supplementary Table 1.) The first main observation is that on average the proposed PMF-GRN performs better than the Inferelator in recovering the GRN, regardless of whether we pick the mean or median Inferelator algorithm in terms of AUPRC. Specifically, we see that PMF-GRN performs markedly better than 2 out of the 3 Inferelator algorithms (AMuSR and StARS) and similarly to the remaining algorithm (BBSR).

The second main observation is that our approach eliminates the high variance associated with choosing between different inference algorithms. Implementing the Inferelator on the *S. cerevisiae* datasets yields AUPRCs approximately in the range 0.2 to 0.4, without any a priori information on which of these algorithms to use. The resulting inferred GRN could be arbitrarily accurate or inaccurate depending on which algorithm happened to be chosen. In contrast, our method is reliable as it provides one set of results, chosen using a principled objective function; furthermore, this set of

results performs reliably well, on par with the best-performing Inferelator algorithm. We emphasize that this objective function does not require any gold standard network.

In order to highlight the identifiability issue and ensure that prior information is well-incorporated into the prior distributions, we show in Figure 2 with black dots the AUPRC corresponding to results where we do not use prior information (e.g. all prior logistic normal means of A are set to the same low number). We also show with grey dots the AUPRC obtained when we randomly shuffle the prior TF-target gene interaction hyperparameters before using them to build the prior distribution for A. As anticipated, the resulting AUPRC scores are close to 0, implying that our approach, as well as the Inferelator, are capable of taking into account such prior information well and that the prior information we provided is useful and reliable (see Methods section for details).

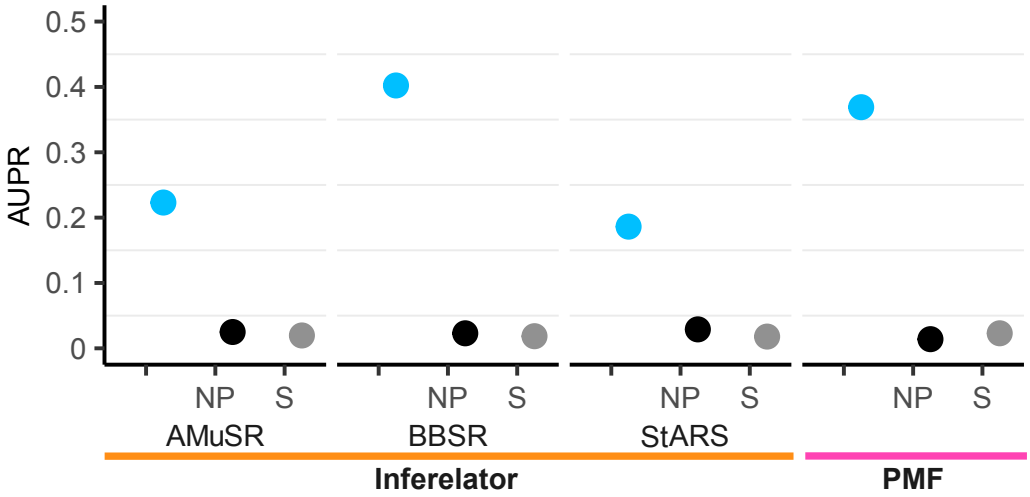

Figure 2: AUPRC computed for GRN's inferred by PMF (pink) and Inferelator (orange) from two independently collected single-cell RNA-seq datasets after averaging. We compare three Inferelator algorithms, AMuSR, BBSR, and StARS to our PMF model (blue dots). To benchmark all networks inferred, we infer two negative controls for each network. NP (No Prior) represents a network inferred without prior-known interactions (black dots), while S (Shuffled) is a network inferred using a prior in which the gene names have been shuffled to randomize the entries.

PMF-GRN Recovers True Interactions in Prokaryotes as Evaluated by Cross-Validation

We carry out experiments using two replicate experiments of a scRNA-seq *B. subtilis* dataset [29]. As we do not have prior information that is independent of the gold-standard for these datasets, we use an approach inspired by cross-validation to evaluate PMF-GRN as well as the Inferelator.

For the prior hyperparameters of A, we use the gold standard with one modification: we randomly hold out 20% of the genes in A by leaving the rows corresponding to these genes in A but setting all prior logistic normal means for these rows to be the same low number. We carry out inference and then evaluate the posterior means for this held-out part of A against the held-out set. Hence the held-out portion is used as an unseen gold-standard while the rest is used as noisy prior knowledge on the GRN. We repeat this procedure using five different randomly chosen held-out sets and calculate the mean and standard deviation of the AUPRCs achieved by the inferred GRNs.

Results from our experiments are shown in Figure 3. (The numbers used to create this figure are given in Supplementary Table 2.) We observe that PMF-GRN achieves a higher AUPRC than all the Inferelator algorithms. Furthermore, as with *S.cerevisiae*, our approach yields one set of results rather than multiple results as returned by different Inferelator algorithms. This one set of results performs reliably well, better than any of the Inferelator algorithms.

'No Prior' and 'Shuffled' results are also shown in Figure 3. For all Inferelator algorithms as well as for PMF-GRN, the corresponding AUPRCs are lower than those achieved using the Inferelator.

As with *S. cerevisiae*, this demonstrates the importance of providing prior information for matrix factorization approaches.

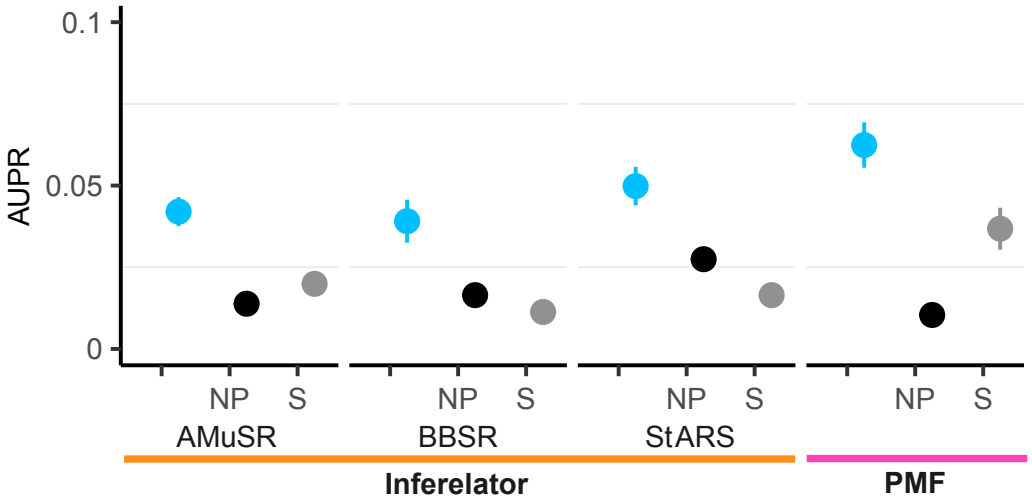

Figure 3: Following the setup of Figure 2, we infer a GRN from two replicate *B.subtilis* experiments and calculate the AUPRC after averaging the inferred network. For each algorithm, we plot the mean $\pm$ standard deviation for the 5 distinct 80%-20% prior-gold standard splits. To benchmark all networks inferred, we infer two negative controls for each network. NP (No Prior) represents the mean $\pm$ standard deviation for the same 5 network splits inferred without any prior-known information (black dots), while S (Shuffled) represents the mean $\pm$ standard deviation for networks inferred using the same 5 prior splits, where the gene names in the prior have been shuffled to randomize the entries.

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 of 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 Figure 4. It is evident from the figures that the AUPRC decreases as the posterior variance increases. Stated differently, 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. The more certain the model is about the degree of existence of a regulatory interaction, the more accurate it is likely to be, showing that our model is well-calibrated.

Discussion

In this paper we present a framework for probabilistic matrix factorization, optimized using automatic variational inference, for inferring GRNs from single cell gene expression data. In contrast with previous methods, our framework decouples the model that defines the data generation process from the inference procedure. Concretely, this means that we can modify the latent variables that constitute the model, along with their distributions, without altering the inference procedure. This flexibility will allow different sequencing data and modeling assumptions to be readily incorporated into the model.Building new models no longer requires defining a new inference procedure, which has previously been the case.

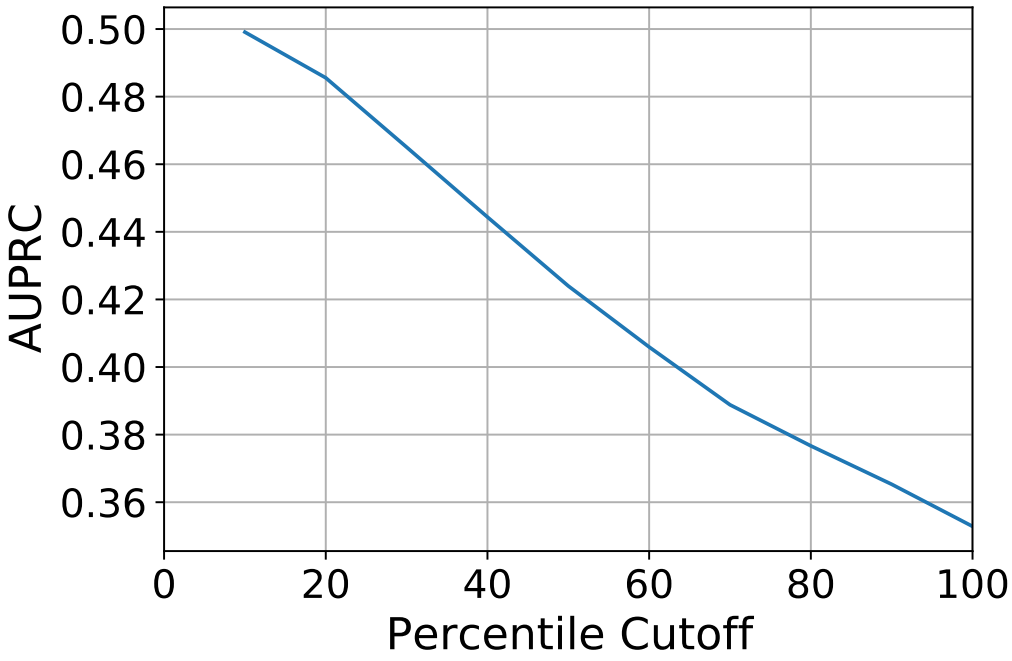

Figure 4: Calibration results on the *S.cerevisiae* (GSE144820 [8] only) dataset. Posterior means are cumulatively placed in bins based on their posterior variances. The x-coordinate x of each point in the plot represents all posterior means that correspond to the bottom $x\%$ of posterior variances. The y-coordinate is the 'overlap' AUPRC (see Methods section for details) calculated on these posterior means against the gold standard.

Additionally, PMF-GRN provides a principled way to carry out model selection and hyperparameter configuration by using the same objective function and inference procedure across all models. This feature differs from previous GRN methods, where it is often unclear which algorithm or hyperparameters to use for a given dataset. In the PMF-GRN framework, we carry out hyperparameter searches across generative models and choose the configuration that corresponds to the optimal value of the objective function. This greatly reduces the need for heuristic model selection.

In order to demonstrate successful GRN inference, we infer a consensus GRN for *S.cerevisiae* using our principled model selection method, and compare our results to GRNs inferred by the Inferelator, with respect to a reliable gold standard. Whereas the Inferelator yields a set of highly varying results across the variants, our approach results in a single inferred GRN. This GRN yields an AUPRC that is higher than the mean and median AUPRC achieved by the Inferelator's respective algorithms, and is comparable to the AUPRC achieved by performing Inferelator algorithm. Our model hence yields a reliable high-performing set of results without any need for heuristic model selection.

Using replicate experiments from a single-cell *B. subtilis* dataset, we further demonstrate that PMF-GRN is capable of learning informative GRNs. To include database information into both the prior knowledge and evaluation, we use an approach motivated by cross-validation. Here, we find that PMF-GRN outperforms all Inferelator algorithms in recovering TF-target gene interactions. Furthermore, as with *S. cerevisiae*, our method provides one set of reliable results, compared with the Inferelator, which provides multiple results that vary in terms of performance.

In order to determine the effect of incorporating our prior domain knowledge into the model, we compare results obtained using shuffled and unshuffled hyperparameters for the matrix A. We observe that for both *S. cerevisiae* and *B. subtilis*, not using prior information or shuffling the prior information results in very low AUPRCs, whereas using the prior information as intended results in significantly better AUPRCs. This result holds for PMF-GRN as well as for all Inferelator algorithms. This shows that prior information is essential for addressing the latent factor identifiability issue and obtaining interpretable results from matrix factorization approaches.

In contrast to previous methods, our model provides well-defined uncertainty estimation in addition to point estimates of GRNs. We evaluate these uncertainty estimates as provided by our model, by computing the AUPRC for inferred TF-target gene interactions corresponding to different levels of posterior uncertainty. We find that the AUPRC increases as the posterior variance decreases, demonstrating that when our model is more certain about its estimates, it produces better rankings of TF-target gene interactions compared to when it is uncertain. This indicates that our model is well-calibrated. For downstream experimental validation, biologists could therefore place more trust in model estimates that have a lower posterior variance.

We also note that the computational cost of our model scales linearly with the number of cells in the dataset. This enables application of our method to scRNA-seq datasets of any size.

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 incorporates information on post-transcriptional modifications.

Most importantly, the study of GRN inference is far from complete. So far, this has 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:

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

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 iid 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:

$$p(\log(U_{nk})) = \mathcal{N}(\mu_u, \sigma_u^2),$$
$$p(\text{logit}(d_n)) = \mathcal{N}(0, 9)$$

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}:

$$p(\text{logit}(A_{mk})) = \mathcal{N}(\text{logit}(\text{clip}(\bar{A}_{mk}, a_{\max}, a_{\min})), \sigma_a^2),$$
$$p(B_{mk}) = \mathcal{N}(0, \sigma_b^2).$$

$$p(\log(\sigma_{obs})) = \mathcal{N}(0, 1),$$

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:

$$q(\log(U_{nk})) = \mathcal{N}(\tilde{U}_{nk}, \tilde{\sigma}^2_{U_{nk}})$$

$$q(\text{logit}(d_n)) = \mathcal{N}(\tilde{d}_n, \tilde{\sigma}^2_{d_n})$$

$$q(\text{logit}(A_{mk})) = \mathcal{N}(\tilde{A}_{mk}, \tilde{\sigma}^2_{A_{mk}})$$

$$q(B_{mk}) = \mathcal{N}(\tilde{B}_{mk}, \tilde{\sigma}^2_{B_{mk}})$$

$$q(\log(\sigma_{obs})) = \mathcal{N}(\tilde{o}, \tilde{\sigma}^2_o),$$

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 $\beta \geq 1$ [30]:

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

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 Figures 2 and 3. For our evaluation of uncertainty calibration (Figure 4), 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 part of the Methods 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

For all datasets used in this paper, we have two scRNA-seq matrices corresponding to different experiments on the same organism. The Inferelator carries out inference with a separate model on each of these matrices, sparsifies each of the resulting TF-gene interaction matrices and picks as its final matrix the intersection between the two matrices, where an entry of the final matrix is non-zero if the corresponding entry in each of the two matrices is also non-zero. For our model, we also train a model on each of the matrices. We obtain a posterior mean matrix for A for each expression matrix, and we average these posterior mean matrices to get our final posterior mean matrix for A. We use this final matrix for evaluation. As mentioned in the Discussion section, explicitly modeling separate expression matrices within the model could be a direction for future work.

Measuring the Impact of Prior Hyperparameters

We evaluate the utility of each of the prior hyperparameter matrices used in our experiments. In Figures 2 and 3, 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.

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, which were therefore not normalized for the purpose of this work. We further obtained binary TF-gene matrices representing prior-known interactions, which served as prior hyperparameters over $\mathbf{A}$, and were either literature derived or estimated using the Inferelator-Prior pipeline [13]. We acquired gold standards for each of 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 [31]). For this well studied organism, we employed the YEASTRACT [32, 33] 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 [34] 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.

Bacillus subtilis

We used two replicates of raw UMI count expression matrices for the organism *B. subtilis* obtained from NCBI GEO (GSM4594095 and GSM4594096 [29]). It is important to note that *B. subtilis* is a polycistronic prokaryote, and thus the usual integer UMI count matrix is in double format. This is due to the fact that multiple genes are encoded by one transcript, and thus have their counts averaged in the UMI matrix. For this well studied organism, we use the subtiwiki database [35] to obtain a network of prior-known TF-target gene interactions to be used as a prior over A as well as a gold standard for benchmarking posterior predictions.

Data Availability

The datasets used in this work are publicly available. They are referenced in the Methods section and are available through `https://github.com/nyu-dl/pmf-grn`.

Code Availability

Code, inferred GRNs, and inference and evaluation scripts can be found at `https://github.com/nyu-dl/pmf-grn`.

Author Contributions

CSG and KC contributed to Conceptualization of the project. OM and KC designed the probabilistic model. OM implemented PMF-GRN Software, Experiments and Validation. CSG implemented PMF-GRN Validation and Inferelator Software. OM, CSG, and KC contributed to Methodology, Software, Validation, Formal Analysis, Visualization, and Writing Original Draft Preparation. CSG contributed to Data Curation. KC and RB contributed to Supervision, Project Administration and Funding Acquisition.

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. 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.

References

[1] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, 96(1):86–103, 2009.

[2] Lian En Chai, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in biology and medicine*, 48:55–65, 2014.

[3] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

[4] Tarmo Äijö and Harri Lähdesmäki. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, 25(22):2937–2944, 2009.

[5] Iftach Nachman, Aviv Regev, and Nir Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20(suppl_1):i248–i256, 2004.

[6] Cassandra Burdziak, Elham Azizi, Sandhya Prabhakaran, and Dana Pe'er. A nonparametric multi-view model for estimating cell type-specific gene regulatory networks. *arXiv preprint arXiv:1902.08138*, 2019.

[7] Kathryn C Allaway, Mariano I Gabitto, Orly Wapinski, Giuseppe Saldi, Chen-Yu Wang, Rachel C Bandler, Sherry Jingjing Wu, Richard Bonneau, and Gord Fishell. Genetic and epigenetic coordination of cortical interneuron development. *Nature*, 597(7878):693–697, 2021.

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

[9] Maria Ciofani, Aviv Madar, Carolina Galan, MacLean Sellars, Kieran Mace, Florencia Pauli, Ashish Agarwal, Wendy Huang, Christopher N Parkurst, Michael Muratet, et al. A validated regulatory network for th17 cell specification. *Cell*, 151(2):289–303, 2012.

[10] Zhe Ji, Lizhi He, Aviv Regev, and Kevin Struhl. Inflammatory regulatory network mediated by the joint action of nf-kb, stat3, and ap-1 factors is involved in many human cancers. *Proceedings of the National Academy of Sciences*, 116(19):9453–9462, 2019.

[11] Nir Yosef, Alex K Shalek, Jellert T Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, et al. Dynamic regulatory network controlling th17 cell differentiation. *Nature*, 496(7446):461–468, 2013.

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

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

[14] Daniel Marbach, Sushmita Roy, Ferhat Ay, Patrick E Meyer, Rogerio Candeias, Tamer Kahveci, Christopher A Bristow, and Manolis Kellis. Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. *Genome research*, 22(7):1334–1349, 2012.

[15] Emily R Miraldi, Maria Pokrovskii, Aaron Watters, Dayanne M Castro, Nicholas De Veaux, Jason A Hall, June-Yong Lee, Maria Ciofani, Aviv Madar, Nick Carriero, et al. Leveraging chromatin accessibility for transcriptional regulatory network inference in t helper 17 cells. *Genome research*, 29(3):449–463, 2019.

[16] Richard Bonneau, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteinn Thorsson. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology*, 7(5):1–16, 2006.

[17] Aviv Madar, Alex Greenfield, Harry Ostrer, Eric Vanden-Eijnden, and Richard Bonneau. The inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5448–5451. IEEE, 2009.

[18] Mario L Arrieta-Ortiz, Christoph Hafemeister, Ashley Rose Bate, Timothy Chu, Alex Greenfield, Bentley Shuster, Samantha N Barry, Matthew Gallitto, Brian Liu, Thadeous Kacmarczyk, et al. An experimentally supported model of the bacillus subtilis global transcriptional regulatory network. *Molecular systems biology*, 11(11):839, 2015.

[19] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166–176, 2003.

[20] Feng Gao, Barrett C Foat, and Harmen J Bussemaker. Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC bioinformatics*, 5(1):1–10, 2004.

[21] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. BioMed Central, 2006.

[22] Tarmo Äijö and Richard Bonneau. Biophysically motivated regulatory network inference: progress and prospects. *Human heredity*, 81(2):62–77, 2016.

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

[24] Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20, 2007.

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

[26] Kenji Kamimoto, Christy M Hoffmann, and Samantha A Morris. Celloracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv*, 2020.

[27] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[28] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[29] Anna Kuchina, Leandra M Brettner, Luana Paleologu, Charles M Roco, Alexander B Rosenberg, Alberto Carignano, Ryan Kibler, Matthew Hirano, R William DePaolo, and Georg Seelig. Microbial single-cell rna sequencing by split-pool barcoding. *Science*, 371(6531):eaba5257, 2021.

[30] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

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

[32] Pedro T Monteiro, Jorge Oliveira, Pedro Pais, Miguel Antunes, Margarida Palma, Mafalda Cavalheiro, Mónica Galocha, Cláudia P Godinho, Luís C Martins, Nuno Bourbon, et al. Yeastract+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic acids research*, 48(D1):D642–D649, 2020.

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

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

[35] José P Faria, Ross Overbeek, Ronald C Taylor, Neal Conrad, Veronika Vonstein, Anne Goelzer, Vincent Fromion, Miguel Rocha, Isabel Rocha, and Christopher S Henry. Reconstruction of the regulatory network for bacillus subtilis and reconciliation with gene expression data. *Frontiers in Microbiology*, 7:275, 2016.

Supplementary Information

A Supplementary Tables

Method	Prior information		
	Regular	None	Shuffled
PMF-GRN	0.375	0.014	0.023
Inferelator AmUSR	0.223	0.024	0.019
Inferelator BBSR	0.402	0.022	0.018
Inferelator StARS	0.186	0.028	0.017

Supplementary Table 1: AUPRCs achieved by PMF-GRN and Inferelator algorithms on *S. cerevisiae* datasets.

Method	Prior information		
	Regular	None	Shuffled
PMF-GRN	0.0624 ± 0.0069	0.0104 ± 0.0009	0.0368 ± 0.0064
Inferelator AmUSR	0.0420 ± 0.0044	0.0113 ± 0.0003	0.0164 ± 0.0015
Inferelator BBSR	0.0390 ± 0.0065	0.0164 ± 0.0033	0.0274 ± 0.0036
Inferelator StARS	0.0498 ± 0.0058	0.0224 ± 0.0061	0.0132 ± 0.0017

Supplementary Table 2: AUPRCs achieved by PMF-GRN and Inferelator algorithms on *B. subtilis* datasets. Results are reported as the mean AUPRC across five 'cross-validation' splits $\pm$ standard deviation

B Supplementary Methods

B.1 TF Target Gene Connectivity Matrix Generation

B.1.1 Saccharomyces cerevisiae

Datasets were obtained from [13] without further modification.

B.1.2 Bacillus subtilis

A prior-known TF-target gene interactions matrix was obtained from the Subtiwiki database [35] from "regulations" (downloaded 07/21/22). Using the columns "regulator locus" and "gene locus" a cross-tab integer matrix was created, where 1 represents the existence of an interaction and 0 represents no interaction. This matrix was randomly split 5 times in 80%-20% proportions along the gene axis to generate independent prior-known information and gold standard matrices.

B.2 Inferelator Networks

B.2.1 Saccharomyces cerevisiae

Networks were inferred using the "multitask" workflow setting of the Inferelator for the same single-cell *S.cerevisiae* datasets described in [13]. For each algorithm, BBSR, StARS, and AMuSR, the following parameters were used: gold_standard_filter_method="keep_all_gold_standard", num_bootstraps=5. Aggregated multi-task networks were used for benchmarking, while single-task networks were disregarded for the purpose of this work. To make these networks directly comparable to PMF, we did not make use of normalization, count minimum, or meta-data options available within the Inferelator workflow.

B.2.2 Bacillus subtilis

Networks were inferred using the "multitask" workflow setting on the Inferelator. Using two tasks for GSM4594095 and GSM4594096 [29] respectively, an aggregate network was learned. For each algorithm, BBSR, Stars and AMuSR, the following parameters were used: gold_standard_filter_method="keep_all_gold_standard", num_bootstraps=5. Aggregated multi-task networks were used for benchmarking, while single-task networks were disregarded for the purpose of this work. To make these networks directly comparable to PMF, we did not make use of normalization, count minimum, or meta-data options available within the Inferelator workflow.