A Variational Inference Approach to Single-Cell Gene Regulatory Network Inference using Probabilistic Matrix Factorization

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

Abstract

Inferring gene regulatory networks (GRNs) from single-cell gene expression datasets is challenging, as current methods are often designed heuristically for specific datasets and lack the flexibility to incorporate additional information or compare against other algorithms. To overcome these challenges, we introduce Probabilistic Matrix Factorization for Gene Regulatory Network Inference (PMF-GRN), to learn scalable GRNs from single-cell expression data. PMF-GRN allows the introduction of additional experimental evidence into prior distributions and uses variational inference to facilitate hyperparameter search for principled model selection and direct comparison to other generative models. We evaluate our method against state-of-the-art GRN inference algorithms using the model organism *S. cerevisiae*, benchmarking against database-derived gold standard interactions. On average, PMF-GRN infers GRNs more accurately than current state-of-the-art single-cell GRN inference methods and offers well-calibrated uncertainty estimates, as it performs GRN inference in a probabilistic setting.

1. Introduction

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 (Hecker et al., 2009). Gene regulatory networks (GRNs) that annotate regulatory relationships between transcription factors (TFs) and their target genes (Chai et al., 2014) have proven to be useful models for stratifying functional differences between cells (Nachman et al., 2004; Karlebach & Shamir, 2008; Äijö & Lähdesmäki, 2009; Burdziak et al., 2019) that can arise during normal development (Allaway et al., 2021), responses to environmental signals (Jackson et al., 2020) and dysregulation in the context of disease (Ciofani et al., 2012; Ji et al., 2019; Yosef et al., 2013).

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 (Mercatelli et al., 2020). Modern GRN inference methods exploit single-cell RNA-seq, a technique which has enabled the characterization of gene expression profiles within heterogeneous populations (Saliba et al., 2014), vastly increasing the potential for GRN inference algorithms (Lähnemann et al., 2020; Akers & Murali, 2021). 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 (Hu et al., 2020). So far, these integrative approaches to GRN inference have been successfully implemented using regularized regression (Skok Gibbs et al., 2022), self-organizing maps (Jansen et al., 2019), tree-based regression (Van de Sande et al., 2020), and Bayesian Ridge regression (Kamimoto et al., 2023).

Although regression-based methods for inferring GRNs from single-cell data are available, they still suffer from significant limitations (Äijö & Bonneau, 2016). Firstly, these methods heavily rely on the input data used to learn the GRN, 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. 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 predic-

[*]Equal contribution [1]Center for Data Science, New York University [2]Genomics and Systems Biology, New York University [3]Prescient Design, Genentech. Correspondence to: Kyunghyun Cho <kyunghyun.cho@nyu.edu>.

The 2023 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2023.

tions 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 large amounts of computational time.

In this study, we introduce PMF-GRN, a novel approach that uses probabilistic matrix factorization (Mnih & Salakhutdinov, 2007) 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 use a principled hyperparameter selection process with the Evidence Lower Bound (ELBO) objective function, 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, which can be useful in the situation where there are limited validated interactions or a gold standard is incomplete. We perform inference on a GPU, by using stochastic gradient descent (SGD), 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 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. In comparison to three state-of-the-art regression-based methods for inferring single-cell GRNs, namely the Inferelator (Skok Gibbs et al., 2022), Scenic (Van de Sande et al., 2020), and Cell Oracle (Kamimoto et al., 2023), our method demonstrates an overall improved performance in recovering the true underlying GRN.

2. Methods

The goal of our PMF 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 $U \in \mathbb{R}^{N\times K}_{>0}$, a TF-target gene interaction matrix $V \in \mathbb{R}^{M\times K}$, observation noise $\sigma_{obs} \in (0,\infty)$ and sequencing depth $d \in (0,1)^N$, with N the number of cells, M the number of genes and K the number of TFs. We further decompose 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 Figure 1.

We assume that 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 motif databases, and direct measurement of TF-binding along the chromosome.

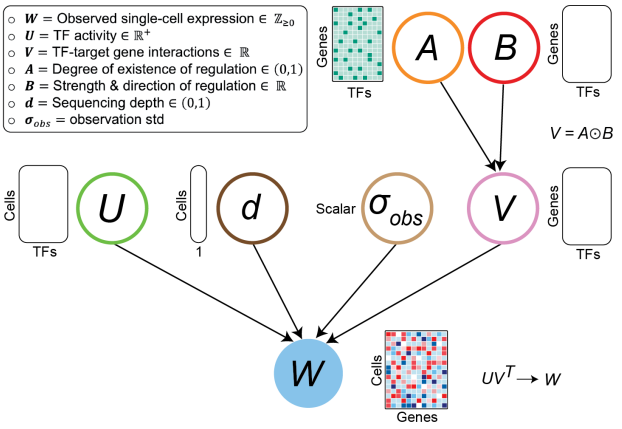

Figure 1. PMF-GRN graphical model overview. Input single-cell gene expression W is decomposed into latent factors U and V, representing TF activity and TF-gene interactions respectively. V is further decomposed into A and B, representing the degree of existence of interaction, and the strength and direction of an interaction, respectively. Information obtained from chromatin accessibility data or genomics databases is incorporated into the prior distribution for A. Additional latent variables are included to model observation noise σ_{obs} and sequencing depth d, in order to better model our observed single-cell gene expression input data.

The observations, W, result from a matrix product $UV^\top$. We assume noisy observations by defining a likelihood

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}. We then 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 (Ranganath et al., 2014; Blei et al., 2017), 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, 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. For more details about the PMF-GRN model, see Appendix A.

3. Results

3.1. PMF-GRN Inference in Eukaryotes

To demonstrate PMF-GRNs ability to infer informative and robust GRNs, we use two single-cell RNA-seq datasets from the model organism *S. cerevisiae*. We perform three experiments using two independently collected single-cell RNA-seq datasets in *S. cerevisiae* (Jackson et al., 2020; Jariani et al., 2020) to test PMF-GRN and compare our performance against three state-of-the-art GRN inference methods, the Inferelator (AMuSR, BBSR, StARS) (Skok Gibbs et al., 2022), Scenic (Van de Sande et al., 2020), and CellOracle (Kamimoto et al., 2023). In the first experiment, we infer a GRN for each of the two single-cell datasets and average the posterior means of A to simulate a "multi-task" GRN inference approach for building the final combined network. Using AUPRC, we show that PMF-GRN outperforms AMuSR, StARS, and Scenic, while performing competitively with BBSR and CellOracle (Figure 2). To provide a baseline for each method in the scenario where data cannot be cleanly separated into tasks, we combine the two expression datasets into one observation before inferring a GRN. This baseline demonstrates a large performance decrease for BBSR, indicating that the method may be limited to gene expression data that is organized into tasks. This could present challenges when attempting to infer GRNs in more complicated organisms where cell-types or conditions are less easily defined.

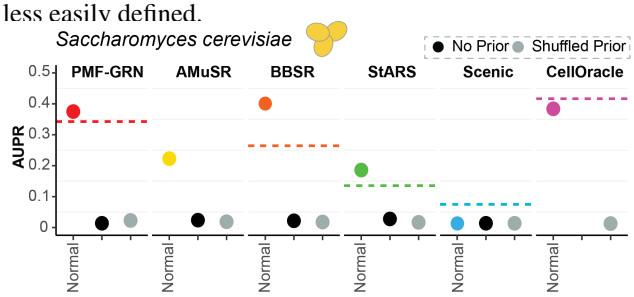

Figure 2. GRN inference in *S. cerevisiae*. Consensus network AUPR using gold standard network. Performance of each algorithm denoted by colored dot. The baseline for each method (dashed line), demonstrates the performance when the expression data is combined into one task. Two negative controls, no prior information (black) and shuffled prior information (gray), are inferred to ensure reliable results.

In the second experiment, we implement a 5 fold cross-validation approach to establish a baseline for each model. In the context of GRN inference, cross-validation is particularly important because it helps us assess the performance of PMF-GRN in predicting TF-target gene interactions based on limited data, which is often the case in experimental settings. We first combine the two *S. cerevisiae* single-cell RNA-seq datasets into one observation matrix for simplicity. To perform cross-validation, the gold standard is divided into an $80\% - 20\%$ split, where a network is inferred using 80% of the gold standard as prior-known information, and evaluated using the remaining 20%. We repeat this cross-validation process five times using different random splits of the gold standard to obtain meaningful results. We observe that PMF-GRN outperforms Scenic and CellOracle, while achieving competitive performance to BBSR and StARS (Fig. 3A in A). 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).

Finally, in the third experiment, we demonstrate the robustness of each GRN inference method against noisy prior information. To do so, we infer GRNs where increasing amounts of noise have been added to the input prior-known information. Here, we show that as noise increases, PMF-GRN's AUPRC decreases similarly to CellOracle, while on average, performing better than BBSR, StARS and CellOracle, demonstrating that it is one of the most robust approaches to inferring accurate GRNs from noisy priors (Fig. 3B in A).

3.2. 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 (Skok Gibbs et al., 2022; Kamimoto et al., 2023; Van de Sande et al., 2020). This *ad hoc* nature of model construction 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, presenting 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, enabling 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 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 SGD in variational inference facilitates significant computational advantage when performing GRN inference on very large single-cell datasets without any constraint on the number of observations. This approach is further sped up by using modern hardware, such as GPUs.

Under this probabilistic framework, we perform 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 domain 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.

As 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 PMF for GRN inference is scalable, generalizable and aware of uncertainty, which makes its use much more advantageous compared to most existing methods.

4. Discussion

In this paper, we present a framework for probabilistic matrix factorization, optimized using automatic variational inference, for inferring GRNs from single-cell expression data. In contrast with previous methods, our framework decouples the model that defines the data generation process from the inference procedure. This flexibility facilitates incorporating various sequencing datasets and modeling assumptions into the model without defining a new inference prodecure, which has previously been the case. Furthermore, PMF-GRN provides a principled approach to model selection and hyperparameter configuration by using the same objective function and inference procedure across all models.

We demonstrate sucessful GRN inference in *S. cerevisiae* and compare our results to GRNs inferred by the Inferelator, Scenic and CellOracle, with respect to a gold standard. We find that PMF-GRN recovers consistent and competitive GRNs when learning a single-task or multi-task network, performing cross-validation, and inferring GRNs from noisy priors. In contrast to existing GRN inference methods, our model provides well-defined uncertainty estimation in addition to point estimation 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, indicating 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. Finally, 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 single-cell RNA-seq datasets of any size.

Ultimately, the study of GRN inference is far from complete and has required new computational models to keep up with relevant sequencing technologies. It is thus essential to develop an adaptable and scalable approach to GRN inference that can be easily modified as new biological datasets become available. We have thus proposed PMF-GRN as a modular, principled and probabilistic approach that can be easily adapted to both new biological data without redesigning a new GRN inference method.

Software and Data

The datasets used in this work are publicly available and are referenced in Section 3 and are available at `https://github.com/nyu-dl/pmf-grn`. Code, inferred GRNs, 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 Experiments, 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. We would like to thank the reviewers for their insightful comments, questions and feedback on this work. CSG would also like to thank Yanis Bahroun for his thoughtful feedback on this manuscript. 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

Äijö, T. and Bonneau, R. Biophysically motivated regulatory network inference: progress and prospects. *Human heredity*, 81(2):62–77, 2016.

Äijö, T. and Lähdesmäki, H. Learning gene regulatory networks from gene expression measurements using nonparametric molecular kinetics. *Bioinformatics*, 25(22): 2937–2944, 2009.

Akers, K. and Murali, T. Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*, 26:87–97, 2021.

Allaway, K. C., Gabitto, M. I., Wapinski, O., Saldi, G., Wang, C.-Y., Bandler, R. C., Wu, S. J., Bonneau, R., and Fishell, G. Genetic and epigenetic coordination of cortical interneuron development. *Nature*, 597(7878): 693–697, 2021.

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

Burdziak, C., Azizi, E., Prabhakaran, S., and Pe'er, D. A nonparametric multi-view model for estimating cell type-specific gene regulatory networks. *arXiv preprint arXiv:1902.08138*, 2019.

Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., and Zakaria, Z. A review on the computational approaches for gene regulatory network construction. *Computers in biology and medicine*, 48:55–65, 2014.

Ciofani, M., Madar, A., Galan, C., Sellars, M., Mace, K., Pauli, F., Agarwal, A., Huang, W., Parkurst, C. N., Muratet, M., et al. A validated regulatory network for th17 cell specification. *Cell*, 151(2):289–303, 2012.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., and Guthke, R. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, 96 (1):86–103, 2009.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Hu, X., Hu, Y., Wu, F., Leung, R. W. T., and Qin, J. Integration of single-cell multi-omics for gene regulatory network inference. *Computational and Structural Biotechnology Journal*, 18:1925–1938, 2020.

Jackson, C. A., Castro, D. M., Saldi, G.-A., Bonneau, R., and Gresham, D. Gene regulatory network reconstruction using single-cell rna sequencing of barcoded genotypes in diverse environments. *Elife*, 9:e51254, 2020.

Jansen, C., Ramirez, R. N., El-Ali, N. C., Gomez-Cabrero, D., Tegner, J., Merkenschlager, M., Conesa, A., and Mortazavi, A. Building gene regulatory networks from scatac-seq and scrna-seq using linked self organizing maps. *PLoS computational biology*, 15(11):e1006555, 2019.

Jariani, A., Vermeersch, L., Cerulus, B., Perez-Samper, G., Voordeckers, K., Van Brussel, T., Thienpont, B., Lambrechts, D., and Verstrepen, K. J. A new protocol for single-cell rna-seq reveals stochastic gene expression during lag phase in budding yeast. *elife*, 9:e55320, 2020.

Ji, Z., He, L., Regev, A., and Struhl, K. 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.

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, pp. 1–10, 2023.

Karlebach, G. and Shamir, R. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., and Giorgi, F. M. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6): 194430, 2020.

Mnih, A. and Salakhutdinov, R. R. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20, 2007.

Monteiro, P. T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., Galocha, M., Godinho, C. P., Martins, L. C., Bourbon, N., et al. Yeastract+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic acids research*, 48(D1):D642–D649, 2020.

Nachman, I., Regev, A., and Friedman, N. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20(suppl_1):i248–i256, 2004.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Artificial intelligence and statistics*, pp. 814–822. PMLR, 2014.

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

Skok Gibbs, C., Jackson, C. A., Saldi, G.-A., Tjärnberg, A., Shah, A., Watters, A., De Veaux, N., Tchourine, K., Yi, R., Hamamsy, T., et al. High-performance single-cell gene regulatory network inference at scale: the inferelator 3.0. *Bioinformatics*, 2022.

Tchourine, K., Vogel, C., and Bonneau, R. Condition-specific modeling of biophysical parameters advances inference of regulatory networks. *Cell reports*, 23(2): 376–388, 2018.

Teixeira, M. C., Monteiro, P. T., Palma, M., Costa, C., Godinho, C. P., Pais, P., Cavalheiro, M., Antunes, M., Lemos, A., Pedreira, T., et al. Yeastract: an upgraded database for the analysis of transcription regulatory networks in saccharomyces cerevisiae. *Nucleic acids research*, 46 (D1):D348–D353, 2018.

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., et al. A scalable scenic workflow for single-cell gene regulatory network analysis. *Nature Protocols*, 15(7):2247–2276, 2020.

Yosef, N., Shalek, A. K., Gaublomme, J. T., Jin, H., Lee, Y., Awasthi, A., Wu, C., Karwacz, K., Xiao, S., Jorgolli, M., et al. Dynamic regulatory network controlling th17 cell differentiation. *Nature*, 496(7446):461–468, 2013.

A. Appendix

A.1. 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:

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

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.

A.2. Identifiability

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 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-target gene interactions into the prior distribution over A is therefore essential to give the inference algorithm information about which column corresponds to which TF.

With this identifiability issue in mind, we design an inference procedure that can be used on 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-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.

A.3. 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 $\beta \geq 1$ (Higgins et al., 2017):

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

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.

A.4. 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.

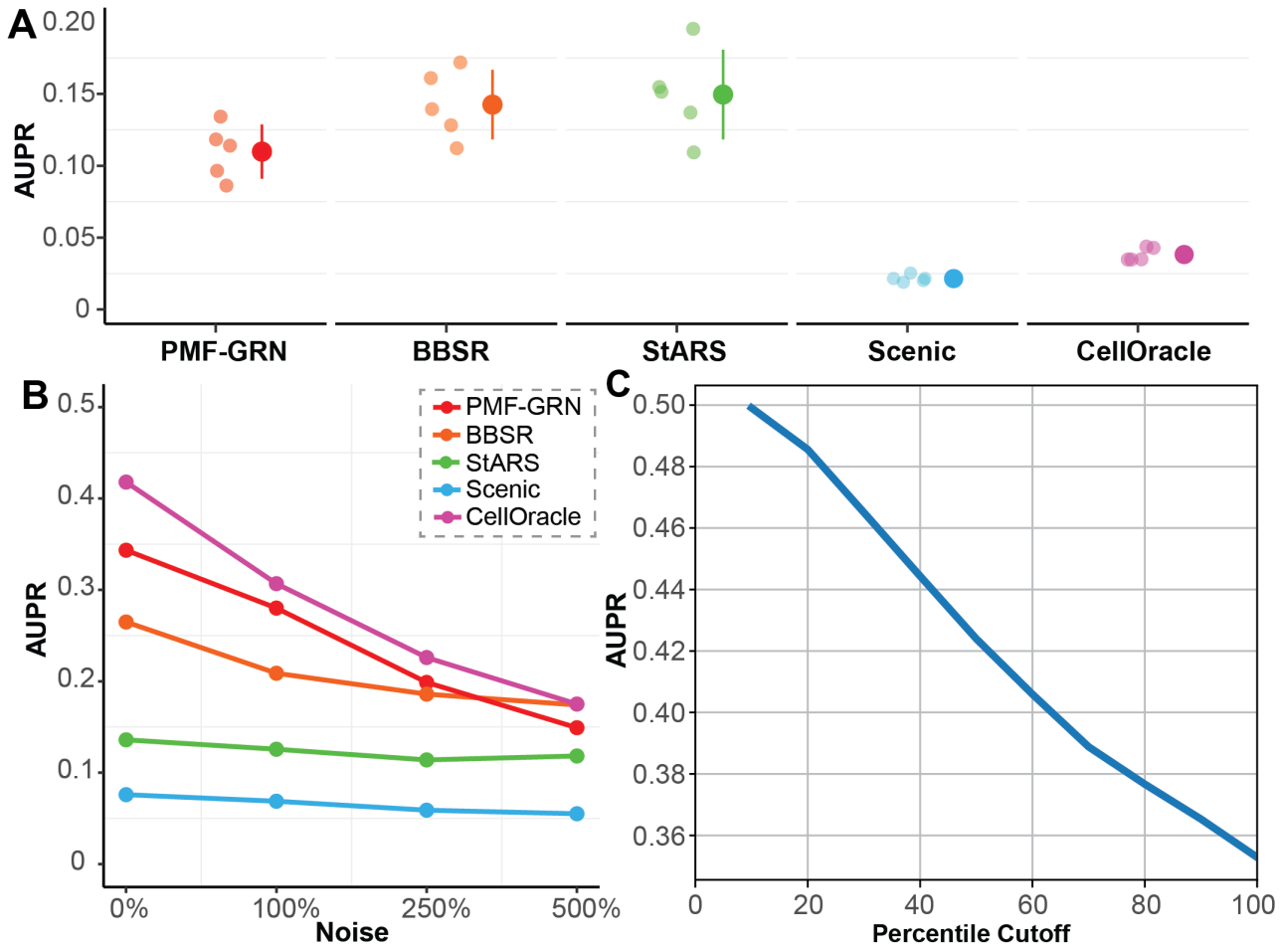

Figure 3. (**A**) 5 fold cross-validation establishes a baseline for each model. Low-opacity dots represent each of the five cross-validation experiments. The mean AUPR $\pm$ standard deviation for each GRN inference method is depicted by colored dot and line. (**B**) GRNs inferred with increasing amounts of noise added to the prior. (**C**) Calibration results on the *S.cerevisiae* (GSE144820 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 section 2 for details) calculated on these posterior means against the gold standard.

A.5. 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 3 C), we use the overlap AUPRC so that bins containing a lower number of posterior means do not have artificially deflated AUPRCs.

A.6. 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.

A.7. 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.

A.8. Measuring the Impact of Prior Hyperparameters

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

A.9. Cross-Validation

For *S. cerevisiae*, we perform a five-fold cross validation experiment. 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.

A.10. 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* datasets, 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 derived from the YEASTRACT database. We acquired a gold standard for *S. cerevisiae* our datasets from independent work which is detailed below.

S. CEREVISIAE

We used two raw UMI count expression matrices for the organism *S. cerevisiae* obtained from NCBI GEO (GSE125162 (Jackson et al., 2020) and GSE144820 (Jariani et al., 2020)). For this well studied organism, we employed the YEASTRACT (Monteiro et al., 2020; Teixeira et al., 2018) 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 (Tchourine et al., 2018) 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.