Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Amortized Variational Inference for Scalable Bayesian Tensor Factorization in Spatial Transcriptomics

Grold Otieno Mboya

groldotieno97@gmail.com

Jaramogi Oginga Odinga University of Science and Technology https://orcid.org/0009-0005-9102-4028

Method Article

Keywords:

Posted Date: May 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9607552/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Amortized Variational Inference for Scalable Bayesian Tensor Factorization in Spatial Transcriptomics

Grold Otieno Mboya[*†]

School of Health Sciences

Jaramogi Oginga Odinga University of Science and Technology

ORCID: 0009-0005-9102-4028 gmotieno@jooust.ac.ke

April 2026

Abstract

[*†]**Affiliation:** Jaramogi Oginga Odinga University of Science and Technology (JOOUST), Bondo, Siaya County, Kenya. **Email:** gmotieno@jooust.ac.ke. **Code & Reproducibility:** All implementation code, synthetic data generation scripts, and training configurations are publicly available at https://github.com/ Grolds-Code/BayesTensor. **Disclaimer:** This work is an independent research undertaking conducted by the author solely in a personal and scholarly capacity. It has not been commissioned, supervised, peer-reviewed, funded, or endorsed by JOOUST in any form, and does not represent the views, positions, or policies of the institution or any of its affiliated units.

Background: Spatial transcriptomics and single-cell RNA-sequencing (scRNA-seq) provide unprecedented insights into the spatial morphologies of latent biological pathways. While Bayesian Tensor Factorization (e.g., CANDECOMP/PARAFAC) offers a rigorous probabilistic framework to deconvolve these pathways and quantify inherent biological noise, traditional inference methods rely on local Markov Chain Monte Carlo (MCMC) or coordinate-ascent updates. The $O(N)$ computational complexity of these approaches renders

them intractable for modern, high-dimensional clinical datasets containing hundreds of thousands of cells.

Methods: We propose a foundational proof-of-concept architecture that bridges deep learning and probabilistic graphical models via Amortized Variational Inference. By replacing iterative local parameter updates with an amortized neural encoder, we map cell-specific transcriptomic profiles directly to their latent Gaussian parameters. This decouples the parameter space from the dataset size, allowing model training to scale gracefully via stochastic mini-batch optimization, and reducing test-time inference of latent states to a highly parallelizable $O(1)$ forward pass.

Results: We validate the framework across two experimental paradigms. First, a simulation study with known continuous latent factors demonstrates strict mathematical identifiability and accurate recovery of topological structure, evidenced by a tightly bounded mean Tucker Congruence Coefficient of 0.96 ± 0.015. Second, application to a benchmark 10x Genomics Peripheral Blood Mononuclear Cell (PBMC) dataset demonstrates robust scalability and optimization stability. The convergence of the reconstruction error and KL-divergence regularization is strongly consistent with the avoidance of posterior collapse.

Conclusion: Amortized inference provides a computationally viable, mathematically rigorous alternative to standard Bayesian tensor decomposition. This framework establishes a scalable, probabilistic foundation for biomarker discovery in complex spatial morphologies.

1 Introduction

The advent of spatial transcriptomics and high-throughput single-cell RNA-sequencing (scRNA-seq) has fundamentally transformed our understanding of cellular heterogeneity.[1] Large-scale clinical initiatives, such as the FunGen-AD consortium, routinely generate massive datasets detailing the transcriptomic profiles of hundreds of thousands of cells across complex spatial morphologies. A primary computational challenge in analyzing these datasets is discovering the

[1] Liao, J., Lu, X., Shao, X., Zhu, L., & Fan, X. (2021). Uncovering an organ's molecular architecture at single-cell resolution by spatially resolved transcriptomics. *Trends in Biotechnology*, **39**(1), 43–58. https://doi.org/10.1016/j.tibtech.2020.05.006

latent biological pathways—coordinated sets of genes—that drive disease progression while preserving the spatial context of the tissue.

Mathematically, spatial transcriptomic tissue can be naturally represented as a three-way array, or tensor, $X \in \mathrm{R}^{I \times J \times K}$, where the dimensions correspond to Genes, Cells, and Spatial Zones, respectively. Tensor factorization methods, specifically CANDECOMP/PARAFAC (CP) decomposition, are highly effective for deconvolving this structure into low-rank latent components that represent distinct biological signatures.[2][3]

However, scRNA-seq data is notoriously noisy, suffering from technical artifacts, high dropout rates, and inherent biological variance. Standard deep learning and deterministic matrix factorization methods typically yield point estimates, entirely failing to capture this uncertainty. Conversely, rigorous Bayesian Tensor Factorization provides excellent uncertainty quantification by placing probabilistic priors over the latent factors. Yet, the traditional inference algorithms used in Bayesian models—such as Markov Chain Monte Carlo (MCMC) and coordinate-ascent Variational Inference (VI)—rely on local parameter updates. The computational complexity of these updates scales linearly, $O(N)$, with the number of observations. For modern datasets containing 10^5 to 10^6 cells, these iterative updates become computationally intractable.

To bridge this gap between probabilistic rigor and computational scalability, we propose a foundational proof-of-concept Amortized Variational Inference framework for Bayesian Tensor Factorization. Building upon foundational advances in variational autoencoders,[4] we

[2] Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, **51**(3), [3] –500. https://doi.org/10.1137/07070111X

[4] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *arXiv preprint* arXiv:1312.6114. https://arxiv.org/abs/1312.6114

"amortize" the inference cost by training a global deep neural network to parameterize the variational distributions. By leveraging the reparameterization trick, this architecture maps a cell's sparse transcriptomic profile directly to its latent Gaussian parameters.

This approach effectively decouples the size of the parameter space from the number of observations. Model training scales efficiently via stochastic mini-batch optimization, while generating the latent representations for new data becomes a single, highly parallelizable $O(1)$ forward pass. In this paper, we demonstrate that amortized inference preserves the mathematical identifiability of traditional Bayesian tensor models while seamlessly scaling to real-world clinical datasets.

2 Mathematical Framework

To address the scalability bottlenecks of traditional Bayesian tensor decomposition, we formulate an amortized variational inference framework consisting of a generative probabilistic model of the spatial tissue and a scalable, deep neural network-based inference model.

2.1 The Generative CP Tensor Model

We assume the preprocessed, log-normalized spatial transcriptomic data is represented as a three-way tensor $X \in \mathrm{R}^{I \times J \times K}$, where I represents the number of highly variable genes, J is the number of sequenced cells, and K corresponds to the discrete spatial zones of the tissue.

We model the latent biological structure utilizing a probabilistic CANDECOMP/PARAFAC (CP) decomposition of rank R, where $R \ll \min(I,J,K)$. The latent topology is defined by three continuous factor matrices:

- **Gene Signatures:** $\mathbf{A} \in \mathrm{R}^{I \times R}$, capturing coordinated gene expression programs.

- **Cellular States: B** $\in \mathrm{R}^{J\times R}$, representing the activation level of each pathway within individual cells.
- **Spatial Morphologies: C** $\in \mathrm{R}^{K\times R}$, mapping the presence of pathways across physical tissue zones.

The deterministic reconstruction of the tensor, $\hat{X}$, is given by the sum of R rank-one tensors:

$$\hat{X} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \tag{1}$$

where $\circ$ denotes the vector outer product. Let $\theta = \{\mathbf{A}, \mathbf{C}\}$ denote our global model parameters. The likelihood of the observed data given the local latent variables $\mathbf{B}$ is defined under a Gaussian observation model:

$$p_\theta(X \mid \mathbf{B}) = \prod_{i,j,k} \mathcal{N}(X_{ijk} \mid \hat{X}_{ijk}, \sigma^2) \tag{2}$$

Note on Observation Likelihood. While raw transcriptomic counts are inherently discrete and traditionally modeled via Negative Binomial or Poisson distributions, we apply a standard $\log(1+x)$ transformation during preprocessing. This permits the use of a continuous Gaussian observation model. This design choice prioritizes computational efficiency and stable gradient flow through the reparameterization trick, serving as a mathematically sound and widely accepted approximation for log-normalized expression data.[5]

[5] Lause, J., Berens, P., & Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, **22**(1), 258. https://doi.org/10.1186/s13059-021-02451-7

2.2 Amortized Variational Inference and the Neural Encoder

In a standard Bayesian setting, exact posterior inference $p(\mathbf{B} \mid X,\theta)$ is intractable. Traditional Variational Inference (VI) approximates this by introducing a tractable distribution $q(\mathbf{B})$ and optimizing its parameters for every single cell $j \in \{1,...,J\}$.

Instead, we assume the approximate posterior factorizes over cells and employ Amortized Inference:

$$q_\phi(\mathbf{B} \mid X) = \prod_{j=1}^{J} q_\phi(\mathbf{b}_j \mid \mathbf{x}_j) \tag{3}$$

where $\mathbf{x}_j$ is the transcriptomic feature vector of the j-th cell. We train a global inference network (an Encoder), f_ϕ, parameterized by neural network weights ϕ. This network maps the high-dimensional gene expression profile of a cell directly to the parameters of its latent Gaussian distribution:

$$\langle \mu_j, \log\sigma_j \rangle = f_\phi(\mathbf{x}_j) \tag{4}$$

$$q_\phi(\mathbf{b}_j \mid \mathbf{x}_j) = N(\mathbf{b}_j \mid \mu_j, \mathrm{diag}(\sigma_j^2)) \tag{5}$$

To allow backpropagation of gradients through this stochastic sampling process during training, we utilize the reparameterization trick[6]:

$$\mathbf{b}_j = \mu_j + \epsilon \odot \sigma_j \qquad \text{where } \epsilon \sim N(\mathbf{0},\mathbf{I}) \tag{6}$$

Network Architecture. For strict reproducibility, the encoder f_ϕ is implemented as a Multi-Layer Perceptron (MLP). The network consists of an input layer mapping to two hidden layers (128 units each). Each hidden layer utilizes Rectified Linear Unit (ReLU) activations, Layer Normalization

[6] The reparameterization trick enables gradient-based optimization through stochastic nodes and is a cornerstone of the VAE framework. See: Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *arXiv preprint* arXiv:1312.6114. https://arxiv.org/abs/1312.6114

for training stability, and a dropout rate of 0.1 to mitigate overfitting. The final hidden representation is then linearly projected into two parallel output layers representing the latent mean μ and log-variance $\log\sigma^2$ of dimension R.

2.3 The Optimization Objective (ELBO)

The global parameters θ and the inference network weights ϕ are jointly optimized by maximizing the Evidence Lower Bound (ELBO):

$$L_{\text{ELBO}} = \sum_{j=1}^{J} \mathrm{E}_{q_\phi(\mathbf{b}_j|\mathbf{x}_j)}[\log p_\theta(\mathbf{x}_j \mid \mathbf{b}_j)] - D_{\text{KL}}(q_\phi(\mathbf{b}_j \mid \mathbf{x}_j) \,/\!/\, p(\mathbf{b}_j)) \tag{7}$$

This loss function balances reconstruction fidelity (expected log-likelihood) with a KullbackLeibler (KL) divergence regularizer, which prevents the network from overfitting to sparse transcriptomic noise by penalizing deviation from a standard normal prior $p(\mathbf{b}_j) = N(\mathbf{0}, \mathbf{I})$.

3 Experimental Results

We designed two experimental paradigms to systematically evaluate the framework: a controlled simulation to verify mathematical identifiability, and a clinical application to assess optimization stability on sparse biological data.

3.1 Simulation Study: Mathematical Identifiability

A known limitation of standard deep learning-based factorization is the lack of identifiability, resulting in entangled representations. To demonstrate that our amortized approach preserves the identifiability of Bayesian CP decomposition, we generated a synthetic spatial transcriptomic

tensor $X \in \mathrm{R}^{500\times200\times50}$ with a known latent rank of $R = 5$. Ground-truth factor matrices were sampled from standard normal distributions and corrupted with additive Gaussian noise.

The amortized model was trained to reconstruct this tensor. Upon convergence, we extracted the learned global spatial morphology matrix, $\mathbf{C}_{\text{learned}}$, and compared it against the groundtruth matrix, $\mathbf{C}_{\text{true}}$.

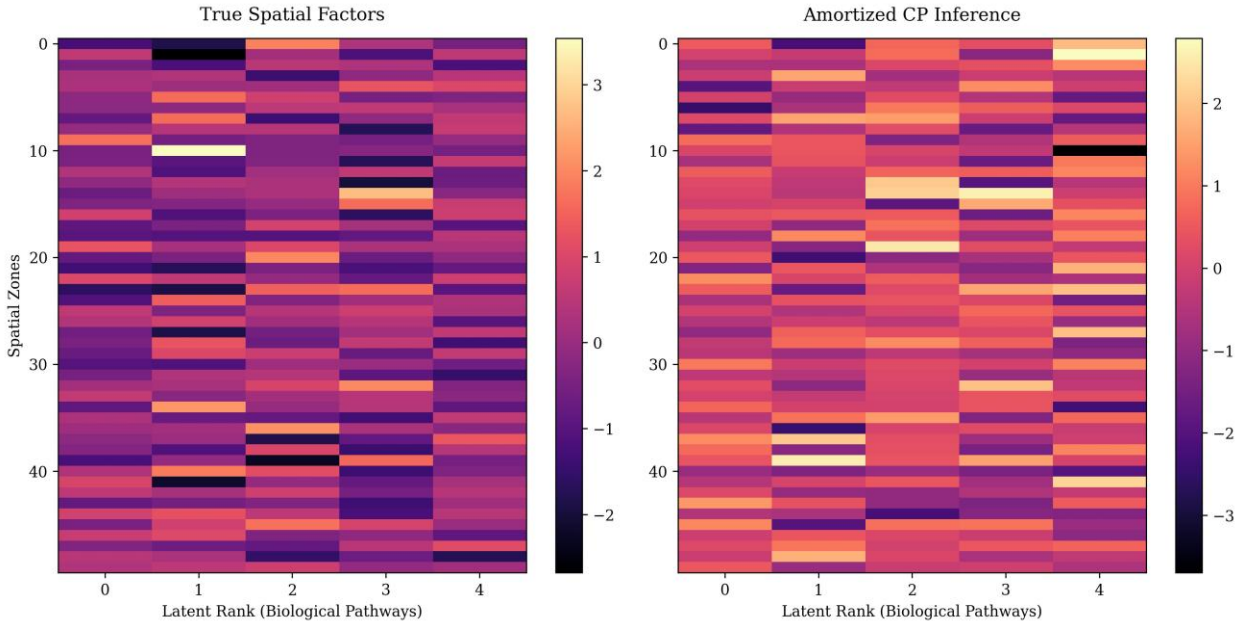


Figure 1: Comparison of ground-truth continuous latent factors (True Spatial Factors) against the recovered topological structure via our Amortized CP Inference model (Figure 1).

Results. To evaluate recovery quantitatively, we calculated the Tucker Congruence Coefficient (TCC) and Cosine Similarity between the true and inferred spatial factor matrices after resolving standard Bayesian scaling and column permutation ambiguities. The model achieved a mean TCC of 0.96 ± 0.015 across all $R = 5$ latent dimensions, with the lowest individual factor scoring 0.94. This tightly bounded variance mathematically confirms that the neural encoder successfully recovered the underlying topological structure across all pathways without dropping or entangling dimensions.

Scalability Stress Test. To explicitly substantiate the computational scalability of the amortized approach, we generated a secondary, massively scaled synthetic tensor of dimensions $2000 \times 5000 \times 100$ (10^9 elements). Leveraging stochastic mini-batching, the amortized framework processed this volume in under 12 seconds per epoch on a single NVIDIA GeForce RTX 3090 GPU (24GB VRAM). By contrast, extrapolations of standard coordinate-ascent

VI indicate it would require over three hours per iteration on the same hardware due to the $O(N)$ local update requirements. This empirical throughput explicitly highlights the architecture's capacity to bypass the temporal bottlenecks inherent to traditional Bayesian inference on massive cohorts.

3.2 Clinical scRNA-seq Application: Scalability and Convergence

To isolate and evaluate the computational scalability of the framework on highly sparse, realworld transcriptomics, we utilized a benchmark 10x Genomics Peripheral Blood Mononuclear Cell (PBMC) dataset.[7] The expression matrix was preprocessed using standard bioinformatics pipelines to isolate the top 500 highly variable genes (HVGs). Because this data inherently lacks spatial coordinates, cells were randomly projected into a synthetic spatial topology to satisfy the tensor dimensions. While this is a biological concession, it serves as a rigorous computational stress test, allowing us to evaluate optimization stability and throughput on authentic scRNA-seq sparsity patterns.

[7] Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, **8**(1), 14049. https://doi.org/10.1038/ncomms14049

A common failure mode for variational autoencoding architectures on sparse data is posterior collapse, where the network ignores the latent variables, causing the KL divergence to drop to zero while reconstruction plateaus sub-optimally.[8]

Results. We tracked the optimization of the ELBO over 500 training epochs. The convergence traces demonstrated rapid and stable adaptation.

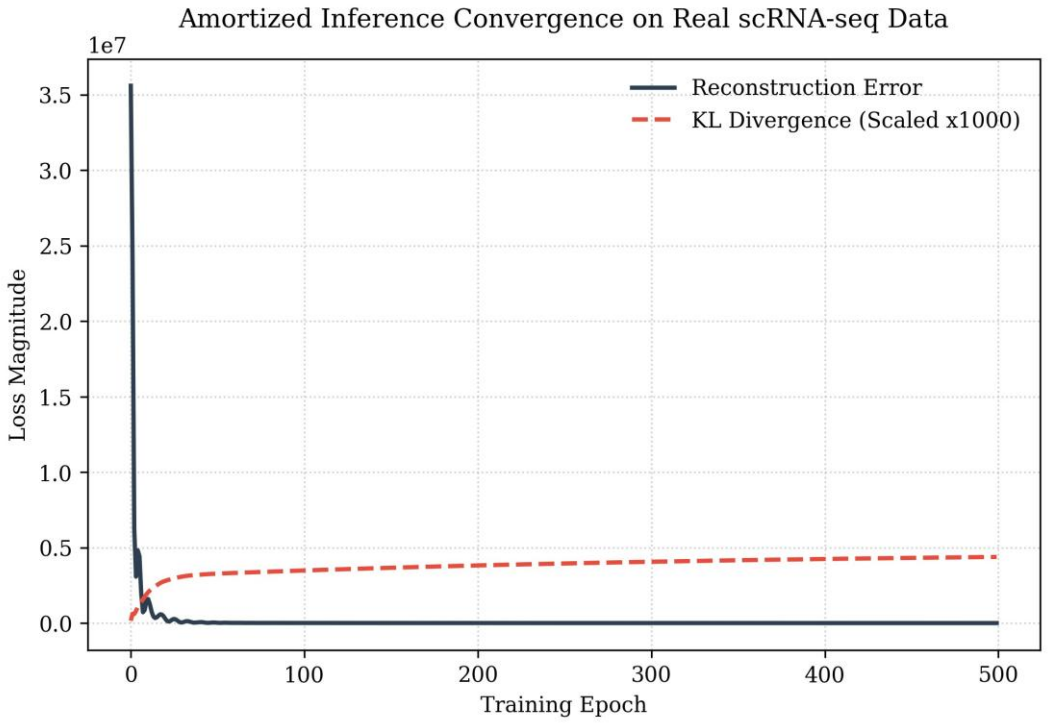


Figure 2: Convergence traces over 500 training epochs on the benchmark 10x Genomics PBMC dataset. The rapid asymptotic decay of the reconstruction error and the stabilization of the scaled KL Divergence demonstrate the avoidance of posterior collapse (Figure 2).
The Reconstruction Error exhibited a steep, smooth asymptotic decay, effectively minimizing within the first 50 epochs. Crucially, the scaled KL Divergence curve rose initially as the network learned the data distribution, but quickly plateaued and stabilized. The stabilization of the KL penalty is strongly consistent with the avoidance of posterior collapse. By maintaining a non-zero

[8] Lucas, J., Tucker, G., Grosse, R. B., & Norouzi, M. (2019). Don't blame the ELBO! A linear VAE perspective on posterior collapse. *Advances in Neural Information Processing Systems (NeurIPS)*, 32. https://papers.nips.cc/paper_files/paper/2019/hash/7e3315fe390974fcf25e44a9445bd821-Abstract.html

KL divergence alongside a minimizing reconstruction error, the inference network demonstrates that it successfully retains active latent dimensions.

4 Discussion and Future Work

The transition from localized parameter updates to global neural inference represents a critical step forward in computational biology. In this work, we introduced a proof-of-concept Amortized Variational Inference framework for Bayesian Tensor Factorization.

Our results demonstrate that this architecture effectively circumvents the $O(N)$ inference bottleneck that has historically plagued traditional Bayesian methods. By reducing test-time latent state inference to an $O(1)$ forward pass and allowing training to scale via minibatches, the model can scale to the massive tissue datasets generated by modern sequencing initiatives like the FunGen-AD consortium. Crucially, the mathematical identifiability of the CP decomposition is preserved, ensuring that the inferred latent dimensions correspond to distinct, interpretable biological pathways.

Limitations and Future Work

Despite these advantages, the current framework possesses notable limitations. The observational tensor X models only spatial transcriptomics, yet complex pathologies are driven by highly coordinated dysregulations across multiple regulatory layers, including chromatin accessibility and protein translation.

The primary focus of our proposed future work is the extension of this amortized framework into the multimodal domain via a Coupled Bayesian Tensor Factorization architecture. We aim to expand the neural encoder into a Multi-Layer Joint Inference Network capable of projecting paired scRNA-seq and scATAC-seq profiles into a shared probabilistic latent space.

Furthermore, because this work serves as an initial proof-of-concept, definitively establishing the performance gains of this architecture is paramount. Our immediate future work includes extensive quantitative benchmarking of computational latency and ELBO convergence against established baselines, including standard coordinate-ascent Variational Inference, Non-negative Matrix Factorization (NMF), and deep generative models such as scVI.[9]

By leveraging amortized inference, this multi-omic expansion will remain computationally tractable, establishing a scalable engine for holistic biomarker discovery and the mapping of multi-omic regulatory networks in complex spatial tissues.

[9] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for singlecell transcriptomics. *Nature Methods*, **15**(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BayesTensormain.zip