PLOS Computational Biology

RESEARCH ARTICLE

Prior-guided factorization for reliable imputation of scRNA-seq data

You Wu[1,2], **Li Xu**[1,2*], **Ye Win Aung**[3], **Alex Michel Daoud**[4]

1 College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang, China, **2** National Engineering Laboratory for Modeling and Emulation in E-Government, Beijing, China, **3** Defense Services Medical Research Centre, Nay Pyi Taw, Myanmar, **4** Division of Neurosurgery, Department of Surgery, Irmandade da Santa Casa de Misericórdia de São Paulo, São Paulo, SP, Brazil

* xuli@hrbeu.edu.cn

OPEN ACCESS

Citation: Wu Y, Xu L, Aung YW, Daoud AM (2026) Prior-guided factorization for reliable imputation of scRNA-seq data. PLoS Comput Biol 22(3): e1014051. https://doi.org/10.1371/journal.pcbi.1014051

Editor: Lihua Zhang, Wuhan University, CHINA

Received: November 9, 2025

Accepted: February 23, 2026

Published: March 20, 2026

Data availability statement: Data for this paper is available for download from: https://doi.org/10.5281/zenodo.17542487. Sourcecode for this paper is available for download from the public GitHub repository: https://github.com/A-uo/scZN.

Funding: This work was supported by the National Key Research and Development Program of China (LX) [Grant No.

Abstract

Single-cell RNA sequencing (scRNA-seq) provides an important means to reveal the heterogeneity and dynamic processes of tissues, organisms, and complex diseases, but technical capture loss (dropout) often obscures true biological expression, and existing imputation methods have difficulty distinguishing biological zeros (silent expression) from technical noise. To address this, we propose the imputation framework scZN. scZN assumes that the observed scRNA-seq data arise from a combination of RNA's two-state transcription process and dropout, and formulates imputation as nonnegative factorization: decomposing the raw count matrix into two interpretable nonnegative factors, performing learning and optimization under constraints from prior knowledge and multiple regularizations, thereby reconstructing the cellular expression landscape. Experiments show that scZN can capture the true distributional characteristics at both the gene and cell levels and significantly suppress spurious activation of genes that should not be expressed. Across multiple real datasets, it outperforms dozens of state-of-the-art methods. Especially in complex experimental design scenarios, scZN markedly improves trajectory inference for embryonic stem cells and mouse dentate gyrus data. In Alzheimer's disease data, scZN can also effectively recover pathways related to neuroinflammation, improving downstream scRNA-seq analysis. Overall, scZN provides a unified framework for missing-value imputation and expression reconstruction that combines accuracy and interpretability.

Author summary

We aim to better understand true gene expression states in single-cell RNA sequencing data, where technical limitations introduce many zero values arising from both genuine gene silencing and missing signals due to insufficient capture efficiency. Distinguishing these two types of zeros is essential for revealing cellular heterogeneity but remains challenging for existing methods. Here, we present

SQ2025YFE0204912], and the National Natural Science Foundation of China (LX) [Grant No. 62172122]. The funders had a role in decision to publish. The funders had a role in study design. The funders had no role in data collection and analysis, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

scZN, a single-cell data imputation framework based on a stochastic two-state model of RNA transcription that explicitly accounts for dropout. By formulating imputation as a biologically constrained nonnegative factorization problem, scZN recovers gene expression while maintaining interpretability. Across multiple real datasets, scZN more effectively suppresses spurious gene activation and improves downstream analyses such as developmental trajectory inference and pathway analysis, particularly in complex experimental designs and disease-related data, providing a unified and biologically meaningful solution for handling missing values in single-cell RNA sequencing.

Introduction

Single-cell RNA sequencing (scRNA-seq) technology offers unprecedented resolution for studying individual cells [1], enabling comprehensive characterization of cellular heterogeneity [2], identification of novel cell types [3], reconstruction of complex differentiation and developmental trajectories [4], and deeper insights into human diseases [5]. However, despite continual improvements in sequencing methodologies, inherent technical limitations—such as amplification bias [6], cell cycle effects [7], and low RNA capture efficiency [8]—inevitably introduce significant noise into scRNA-seq experiments, resulting in count matrices replete with zeros. These zeros comprise both true zeros, which indicate genuine absence of gene expression reflective of cell-specific transcriptomes, and false zeros, which arise from technical noise (i.e., dropout events) [9]. The prevalence of false zeros undermines the integrity of the biological signal, thereby impeding downstream analyses [10].

To address this challenge, early approaches primarily comprised smoothing-based imputation strategies and statistical modeling methods, which leverage intercellular similarity and the underlying data structure [11–17]. Although smoothing-based methods perform well in reconstructing trajectories from time-series scRNA-seq data [18], most count matrices encountered in practice lack intrinsic temporal structure. These approaches can therefore induce substantial changes to expression values, potentially distorting the original gene-expression landscape. In contrast, methods that purely model data structure may preserve global patterns but often overlook biological context, imputing all zeros indiscriminately, which can overestimate performance and diminish biological interpretability.

In recent years, deep learning methods have increasingly been applied to scRNA-seq data imputation. DeepImpute [19] uses neural networks to impute missing values, but its approach of allocating 95% of the dataset for training may lead to overfitting. AutoImpute [20] combines autoencoders with the inherent data distribution for imputation, yet its tendency to over-impute in pursuit of optimal results often produces a substantial number of invalid (negative) values. DCA [21] enhances scRNA-seq imputation by integrating the negative binomial distribution, the zero-inflated negative binomial distribution, and autoencoders. scVI [22,23], a hierarchical Bayesian model, projects the data into a latent space for imputation. However, it struggles

to handle cases where cell counts exceed gene counts. DISC [24] treats unexpressed genes as trainable parameters to improve the model's resistance to overfitting and enhance imputation reliability, although it imposes significant computational demands on large datasets. sciGANs [25] redefines the imputation task as an image restoration problem by converting cellular gene expression into a square matrix using GAN-based methods, which introduces considerable noise. In our previous work [26], we further improved scIGANs using masking and attention mechanisms, developing the scMASKGAN algorithm that achieved higher imputation performance. scGNN [27] leverages graph learning to model the topology of scRNA-seq data, but the resulting topology can greatly affect the expression of marker genes. Moreover, although many other deep learning models have been developed [28–30], they often pay limited attention to the biological significance of imputation, and relying solely on clustering metrics prevents their application to downstream analyses.

To address these challenges, we propose scZN. The framework starts from the physical processes of RNA transcription and capture dropout in single-cell sequencing and explicitly models gene-level overdispersion and zero inflation. scZN assumes that each cell's transcriptome is a nonnegative additive mixture of regulatory modules. Leveraging nonnegative matrix factorization (NMF) [31], we decompose the global count matrix into a soft assignment matrix mapping cells to cell types and a cell-type gene expression matrix. Building on this, we inject prior biological knowledge into the factor matrices via a simple linear shrinkage scheme, thereby improving interpretability and suppressing biologically implausible activations, which in turn reduces spurious signals during imputation. The entire model is trained end-to-end with multiple regularizers that jointly reconstruct scRNA-seq data, yielding efficient and accurate gene-expression estimation and dropout imputation.

Results

Overview of scZN

We propose a single-cell RNA sequencing (scRNA-seq) imputation framework, scZN, which supports prior-guided supervised imputation as well as fully unsupervised imputation. Specifically, scZN assumes that transcriptional bursting follows a Gamma–Poisson (negative binomial) process and that technical dropout introduces excessive zeros, and it explicitly uses a zero-inflated negative binomial (ZINB) to model these two sources of sparsity. To mitigate the non-convexity of factorization, scZN linearly injects biological structure—such as lineage features and cell types—as priors into the decomposition process. scZN is implemented end-to-end in PyTorch and optimized using the Adam optimizer together with multiple regularization strategies. (see Fig 1).

Performance benchmarking

To improve the reliability and accuracy of downstream analyses and biological discovery, the core objective of scRNA-seq imputation is to correct technical noise. Accordingly, we performed a systematic benchmark across multiple real datasets (see Methods) with cell-type labels. In the raw data, pervasive dropout often manifests as cluster intermixing in dimensionality-reduced embeddings such as UMAP, an effective imputation method should attenuate this technical artifact, enhance the resolution of cell-population heterogeneity, and yield consistent improvements in clustering metrics. Using these criteria, we quantitatively evaluated the imputation performance of scZN and compared it objectively against baseline methods.

Using the D1 dataset (see Methods), we systematically compared 14 baseline imputation algorithms. We first performed UMAP visualization to examine whether each method could recover cellular heterogeneity. Fig 2a shows the UMAP embeddings of the ERCC spike-in dataset [32] processed by different methods. Compared with the raw data, scZN and scZN_priorNMF exhibit clearer cluster separation. In contrast, several methods (e.g., DrImpute, SAVER, VIPER, and SCRABBLE) generate an imputed number of cells exceeding the size of the ground-truth label set, introducing spurious structures. As a result, UMAP embeddings and external consistency metrics cannot be computed in these cases.

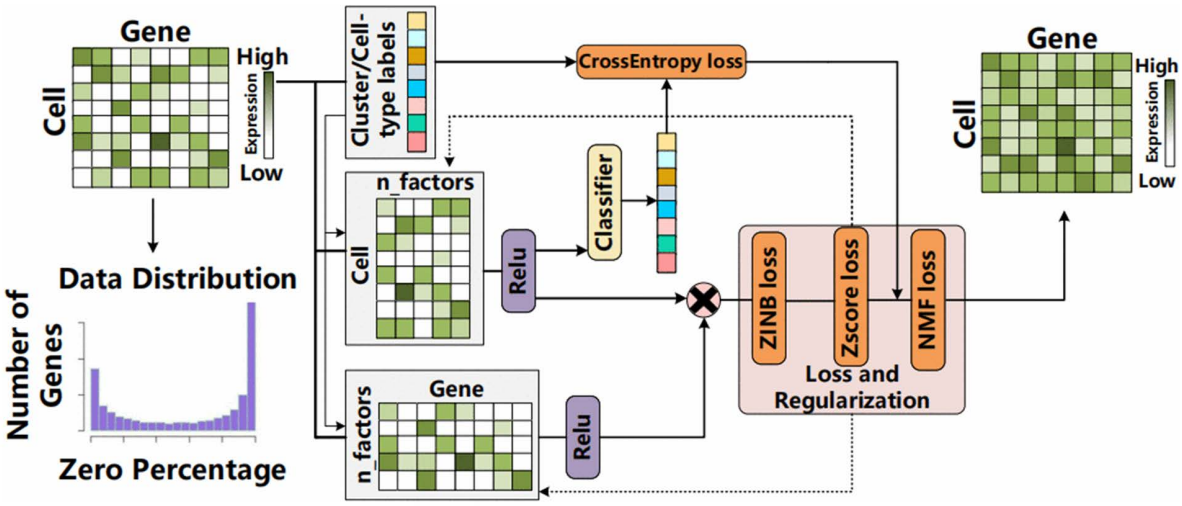

Fig 1. The framework of scZN.

https://doi.org/10.1371/journal.pcbi.1014051.g001

However, we observed that the circular structure corresponding to cell-cycle stages appeared to be partially disrupted in the UMAP embeddings. We therefore further evaluated the accuracy of cell-cycle phase classification on the ERCC spike-in dataset. Specifically, we treated cell-cycle phase prediction as a supervised multi-class classification task, using the imputed expression profiles produced by each method as input features and the provided phase labels as ground truth. For each imputation method, we trained a multinomial logistic regression classifier using standardized features. Performance was evaluated using 5-fold stratified cross-validation, in which cells were randomly split into five folds with preserved phase proportions. In each fold, the classifier was trained on 80% of the cells and tested on the remaining 20%. The results indicate that, among all compared methods, scZN_priorNMF consistently achieves the highest classification accuracy and macro-F1 score (Fig 2b). This demonstrates that, although the circular pattern is less apparent in low-dimensional embeddings, scZN_priorNMF enhances the discriminative information of cell-cycle phases in a quantitative sense.

To more rigorously assess imputation quality, we applied *k*-NN clustering to the imputed matrices and evaluated external consistency metrics across all datasets (see S1 Table). The results show that only scMASKGAN, scImpute, scZN, and scZN_priorNMF achieve overall improvements compared with the raw data, whereas the average performance of the other methods decreases (Fig 2c), indicating that most imputation strategies do not consistently improve data quality across datasets. Among them, scZN_priorNMF achieves the largest improvement across all datasets and all metrics. We also provide box plots of the evaluation metrics across all datasets in S1 Fig. These box plots summarize the performance distributions of all methods across datasets and clearly demonstrate the robustness and stability of scZN_priorNMF, which consistently achieves strong and stable performance across multiple evaluation metrics.

Notably, the imputation methods compared in this study include both unsupervised and supervised approaches. The unsupervised methods include AGImpute [33], MAGIC, DCA, scGAIN, VIPER, AutoImpute, DeepImpute, and scZN, whereas the supervised methods include SCRABBL, scMASKGAN, scIGANs, scFP, scGNN, SAVER, and scZN_priorNMF. The results show that, among unsupervised and supervised methods, scZN and scZN_priorNMF achieve the best overall performance across the evaluated metrics, respectively (Fig 2d).

Finally, in a hybrid computing environment with two vGPUs (32 GB) and a 13th-generation i9 CPU, we systematically quantified the runtime and memory usage of 15 algorithms across single-cell matrices of varying sizes. The results

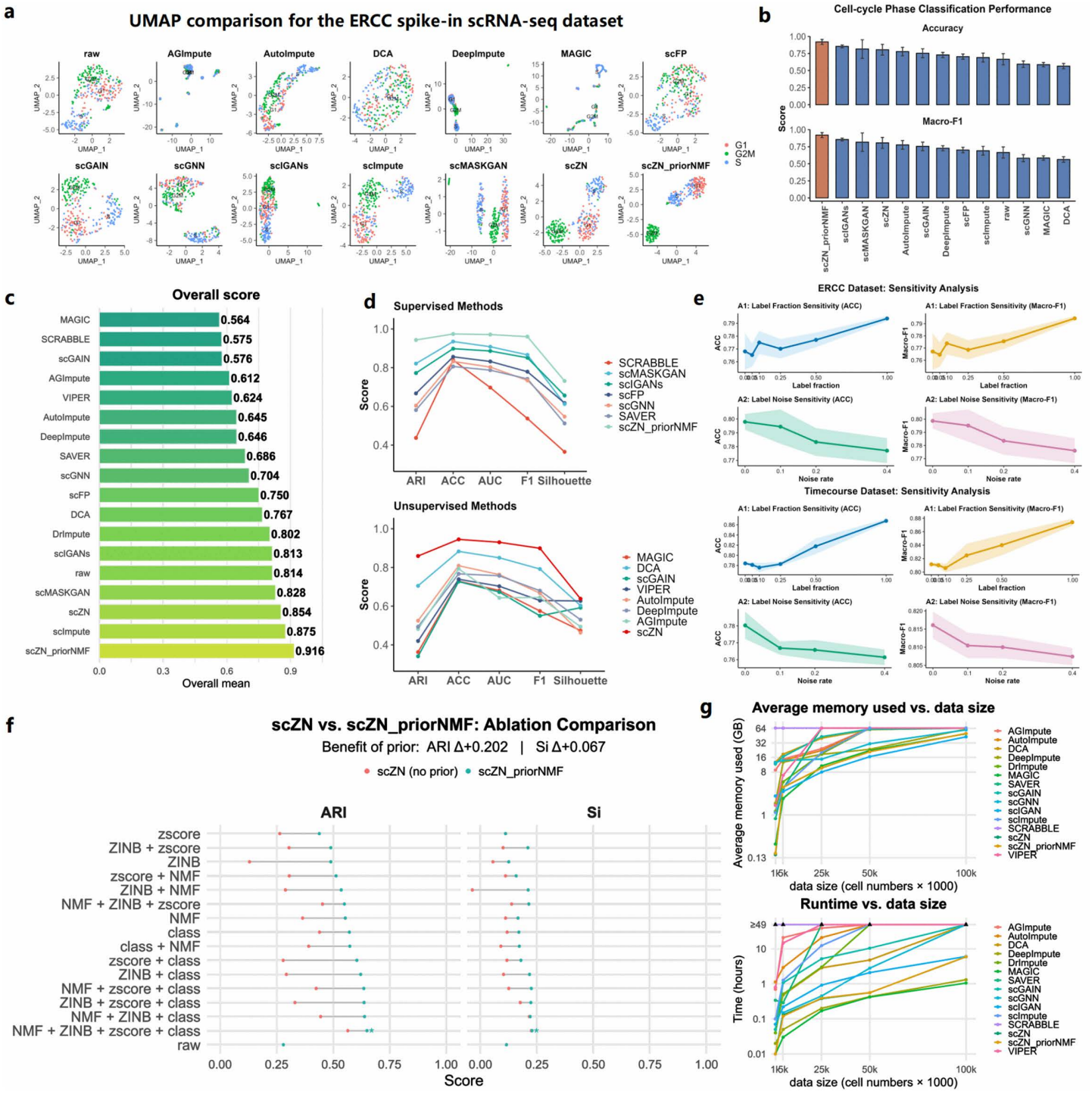

Fig 2. Benchmarking of imputation methods. (a) UMAP projections of the ERCC spike-in scRNA-seq dataset comparing the raw data with imputed results from 13 methods, including AGImpute, AutoImpute, DCA, DeepImpute, MAGIC, scFP, scGAIN, scGNN, scIGANs, scImpute, scMASKGAN, scZN, and scZN_priorNMF. **(b)** Cell-cycle phase classification performance on the ERCC spike-in dataset, evaluated using Accuracy (ACC) and Macro-F1 score. **(c)** Bar plot showing the overall performance score, computed as the average across all evaluation metrics. **(d)** External consistency evaluation on the D1 datasets, where all baseline methods are grouped into supervised and unsupervised categories. **(e)** Comparison of memory consumption and runtime between scZN and all baseline methods across datasets of different sizes. **(f)** Ablation study evaluating different combinations of regularization hyperparameters in scZN (red points) and scZN_priorNMF (blue points), illustrating the effect of prior injection on ARI and silhouette coefficient. The

blue star marks the hyperparameter configuration achieving the best performance. **(g)** Robustness analysis of scZN_priorNMF with respect to different qualities of prior labels. **(h)** Label sensitivity analysis during the imputation process on Time-course datasets, shaded regions around the curves indicate variability.

https://doi.org/10.1371/journal.pcbi.1014051.g002

(Fig 2e) show that scZN and scZN_priorNMF both rank within the overall top three, exhibiting excellent computational efficiency and memory economy while maintaining leading imputation accuracy, making them suitable for routine and high-throughput processing of large-scale single-cell datasets. The dataset used for this evaluation was randomly generated.

Hyperparameter ablation experiment

scZN incorporates multiple regularization techniques to ensure that the generated data are biologically meaningful, specifically including nonnegative matrix factorization(NMF) reconstruction, zero-inflated negative binomial (ZINB) [34] negative log-likelihood, z-score [35] regularization, and cell-type classification (see section Methods). Each regularization term has its own hyperparameter. The specific hyperparameters are described as follows: **Loss weight coefficients** λ_{NMF}, λ_{ZINB}, λ_{zscore}, and λ_{class} are binary switches in {0, 1}:

$$\lambda = \begin{cases} 1, & \text{enable the corresponding loss term,} \\ 0, & \text{disable the corresponding loss term.} \end{cases}$$

We evaluated all 16 subsets of $\{\lambda_{\text{NMF}}, \lambda_{\text{ZINB}}, \lambda_{\text{zscore}}, \lambda_{\text{class}}\}$. For each configuration, the model was trained for a fixed number of epochs on the ERCC spike-in scRNA-seq data, and performance on the held-out data was assessed using the ARI and average silhouette coefficient (Si). As shown in Fig 2f, activating a single loss term resulted in moderate performance improvements—with $\mathcal{L}_{\text{class}}$ alone achieving the highest ARI of 0.572—while certain pairs and triplets demonstrated clear synergistic effects (e.g., $\mathcal{L}_{\text{ZINB}} + \mathcal{L}_{\text{zscore}} + \mathcal{L}_{\text{class}}$ reached an ARI of 0.637). Notably, the full four-term combination ($\lambda_{\text{NMF}} = \lambda_{\text{ZINB}} = \lambda_{\text{zscore}} = \lambda_{\text{class}} = 1$) stood out, achieving the highest ARI of 0.650 and the highest silhouette coefficient of 0.228—more than doubling the clustering accuracy relative to the unregularized baseline. These findings indicate that low-rank reconstruction, zero-inflation modeling, variance alignment, and supervised classification each play unique and complementary roles, and that their joint optimization yields the most biologically consistent imputation outcomes. Therefore, we adopted this four-term configuration in all subsequent experiments to ensure optimal performance and interpretability. However, when the raw data have already been normalized, $\mathcal{L}_{\text{zscore}}$ often provides limited benefit and may become redundant, and is therefore not recommended. We also compared method performance across different hyperparameter combinations before and after incorporating priors, and assessed stability with versus without priors. The results show that introducing prior knowledge can effectively improve model performance.

However, we observe that under the configurations of $\mathcal{L}_{\text{NMF}}$, $\mathcal{L}_{\text{ZINB}}$, and $\mathcal{L}_{\text{class}}$, scZN appears to outperform scZN_PRIORNMF. Therefore, we conducted multiple rounds of experiments on ERCC Spike-in, Time-course, and Dentate Gyrus scRNA-seq datasets under the joint regularization of the NMF, ZINB, and classification losses. The results demonstrate that scZN_PRIORNMF consistently achieves superior and more stable performance. In contrast, the seemingly better results obtained by scZN in some cases can be largely attributed to fluctuations caused by random initialization. When considering both the best-case performance and the average performance across multiple runs, scZN_PRIORNMF still exhibits overall better performance (see S2 Table) and effectively improve stability in the matrix optimization results.

Robustness analysis of prior labels

In addition, we systematically examined the robustness of scZN_priorNMF to the quality of prior labels. We conducted additional experiments on a human brain dataset using three types of labels: (i) randomly generated incorrect labels, (ii)

clustering-based labels obtained using the Leiden algorithm, and (iii) high-quality manual annotations. Under the guidance of manually annotated labels, scZN_priorNMF achieves the best performance across multiple evaluation metrics. Even when using clustering-based labels, scZN_priorNMF still demonstrates strong imputation performance (Fig 2g).

To systematically assess how model performance depends on the availability and reliability of cell-type labels, we conducted two complementary sensitivity analyses (A1 and A2) on all datasets in D1 (Results for all datasets are presented in S2 Fig).

(A1) Label fraction sensitivity. We varied the fraction of available cell-type labels from 0%, 5%, 10%, 25%, 50%, to 100%, and evaluated downstream classification performance using Accuracy and Macro-F1. Fig 2h shows the results on the Time-course scRNA-seq dataset. Performance improves smoothly as more labels are provided and gradually saturates when sufficient labels become available. Importantly, no abrupt performance jump is observed in the high-label regime, indicating that the method does not degenerate into a purely supervised model. Even with only partial labels, the generative structure learned from expression data remains effective.

(A2) Label perturbation analysis. To further verify that performance gains are not caused by label leakage, we performed a controlled perturbation experiment. Specifically, we progressively injected noise into the label files used to construct H_{Prior} and W^{+} at noise levels of 10%, 20%, 30%, and 40%, while still evaluating against the original ground-truth labels. As shown in Fig 2h, both ARI and SI decrease monotonically as noise increases, confirming that the model's sensitivity to label quality is interpretable and controllable, and that no label information leakage occurred.

scZN does not introduce spurious biological signals

Most existing imputation methods rarely systematically assess whether they introduce exogenous noise or spurious biological signals, leaving the reliability of imputed data in doubt. To address this, we performed a multi-faceted evaluation on the Human Brain dataset. First, we examined post-imputation cell distributions and global structure using UMAP (Fig 3a). scZN_priorNMF improved both the overall structure and the sharpness of cluster boundaries relative to the raw data and baseline methods.

To further evaluate the impact on biological signal, we generated heatmaps showing two marker genes per cell type for each method (Fig 3b; S3 Fig). We found that AutoImpute, DCA, MAGIC, scGNN, scIGANs, and SCRABBLE substantially perturbed marker-gene expression patterns: genes expected to be highly expressed were suppressed to near silence, and aberrant upregulation appeared in unrelated cell types. scFP performs imputation only on the top 2,000 highly variable genes selected by Scanpy [36]. Because HVG sets differ across tools, key markers were omitted, preventing complete heatmap rendering. Its imputed values effectively resembled noise injection, limiting utility for downstream analyses. By contrast, the remaining methods largely preserved the original biological signal across cell types.

Having identified methods that preserve native expression patterns, we next asked whether gene-level imputation systematically alters expression. Using volcano plots, we observed that most methods elevate a subset of genes—as expected when mitigating dropout—whereas scMASKGAN and SAVER showed virtually no up- or down-regulated genes (Fig 3c; S4 Fig). This behavior likely reflects an overly conservative design that restricts imputation to the local manifold around each cell type. While such conservatism minimizes artifactual changes, it also fails to recover bona fide biological signal.

To assess marker specificity, we compared differential significance after imputation. Using GAD1 as a sentinel marker—expected to be significant in neurons but not in OPCs—only MAGIC, DeepImpute, scZN, and scZN_priorNMF reproduced the anticipated pattern, preserving neuronal significance without spurious elevation in OPCs. By contrast, SCRABBLE and scIGANs disrupted this behavior: they abolished the expected neuronal significance ($P > 0.05$) and/or produced spurious changes, undermining biological plausibility (Fig 3d; S5 Fig). Moreover, we extended our analysis to multiple canonical Neuron and OPC marker genes, and validated the results on the human brain dataset. Specifically, we selected the Neuron markers SLC17A7, SLC6A1, and GRIN1, as well as the OPC markers CLDN11, CSPG4, and

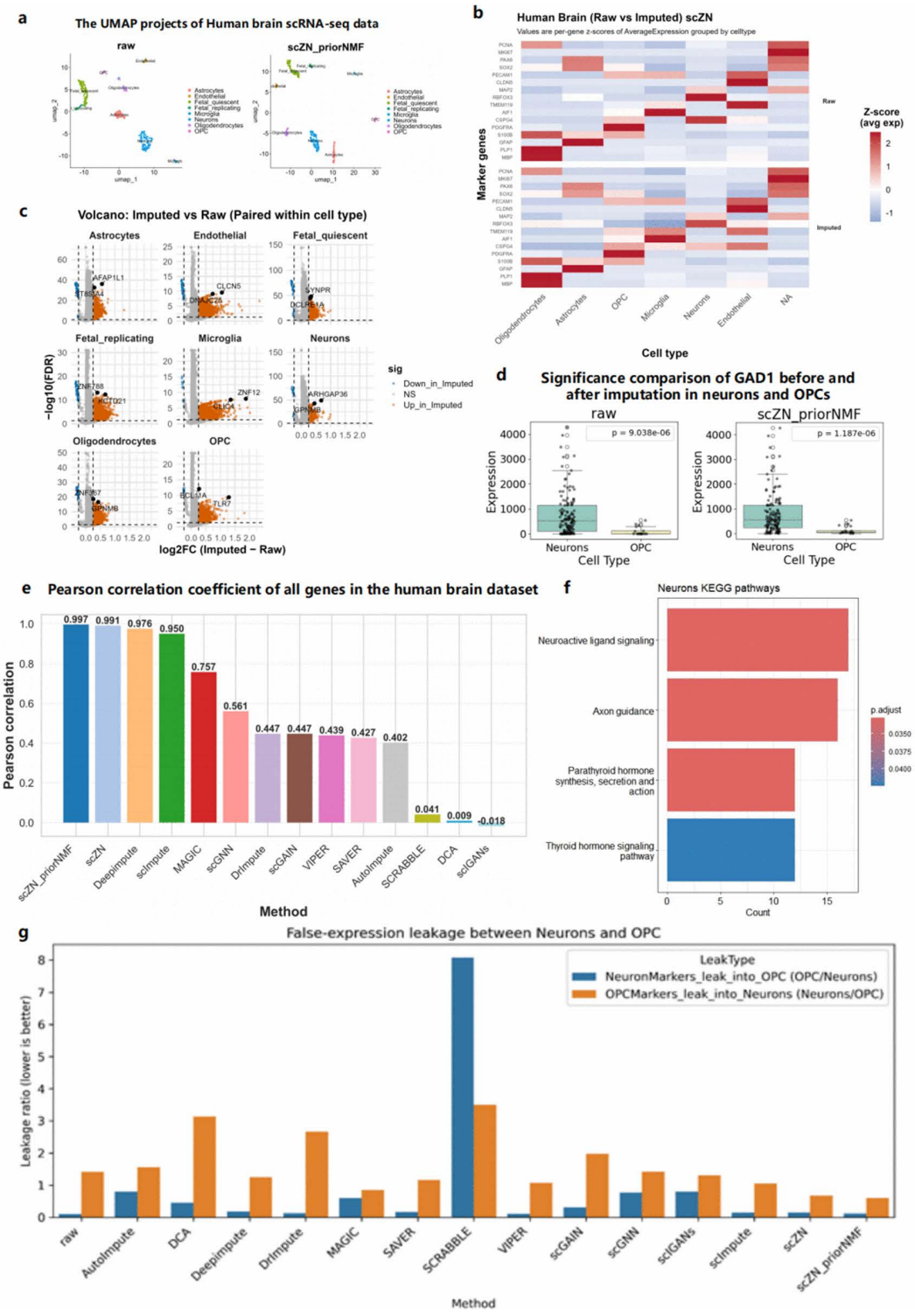

Fig 3. (a) UMAP comparison between the raw data and the scZN-imputed results. (b) Heatmap of two sets of marker genes per cell in the human brain dataset after scZN imputation. (c) Volcano plot comparison of all cell imputation results in the human brain dataset. (d) Significant changes in the imputed data relative to the raw data. (e) Gene–gene correlation before and after imputation, quantified by Pearson correlation coefficients. (f) KEGG pathways enriched among significantly upregulated genes in human brain Neurons following scZN imputation. (g) False-expression leakage between Neurons and OPC.

https://doi.org/10.1371/journal.pcbi.1014051.g003

SLC44A1. We compared the expression distributions between neurons and OPCs. For genes that already show clear cell-type separation in the raw data (e.g., SLC17A7 and CLDN11), scZN_priorNMF further strengthens the statistical significance between neurons and OPCs. More importantly, for SLC6A1, GRIN1, CSPG4, and SLC44A1, which fail to show significant cell-type differences in the raw data due to severe dropout, scZN_priorNMF successfully recovers their expected expression patterns in neurons and OPCs and restores statistically significant cell-type differences(S6 Fig).

Finally, we compared Pearson correlation coefficients before and after imputation across methods. scZN and scZN_priorNMF ranked first and second, respectively, achieving the highest Pearson correlation (Fig 3e). These results indicate that both methods denoise while preserving the native correlation structure, whereas several baselines substantially deviated from the original data—consistent with their reduced marker specificity.

Building on this, we performed KEGG enrichment on upregulated genes for each cell type using scZN_priorNMF. In Neurons (Fig 3f), significantly enriched pathways included Neuroactive ligand–receptor interaction and Axon guidance—two canonical neuronal programs—as well as endocrine modules (Thyroid hormone signaling and Parathyroid hormone synthesis, secretion, and action) that modulate neuronal differentiation and excitability. Comparable, cell type–appropriate enrichments were observed across other lineages (S7 Fig), with no inflation of irrelevant pathways. Collectively, these analyses show that scZN delivers top-tier denoising performance without introducing spurious biological signals.

To quantitatively assess whether imputation introduces spurious biological signals, we performed a false-positive expression leakage analysis using 20 groups of differentially expressed marker genes for Neurons and OPCs, respectively. This analysis measures two complementary error modes: (i) leakage of neuronal marker gene expression into OPCs, and (ii) leakage of OPC marker gene expression into neurons. These leakage rates directly quantify the extent to which an imputation method blurs true cell-type-specific expression patterns. The leakage ratio is defined as the ratio of the average expression of marker genes in the incorrect cell type to that in the corresponding correct cell type, with lower values indicating less false-positive expression leakage. As shown in Fig 3g, scZN_priorNMF achieves the lowest leakage rates between Neuron and OPC, indicating that it introduces the fewest false-positive expression signals across cell types, followed by scZN. This suggests that scZN_priorNMF and scZN produce the most reliable imputation results.

Robust reconstruction of gene–gene correlations

We further evaluated how imputation affects the internal gene–gene correlation structure of the expression matrix. Because bulk RNA-seq can estimate gene co-expression patterns more reliably than sparse single-cell data, it serves as a useful reference for assessing correlation structure. We therefore used a Human ESC scRNA-seq dataset comprising H1 human embryonic stem cells (hESCs) and differentiated endoderm/mesoderm-like cells (DEC), and compared gene–gene correlations derived from raw scRNA-seq data, scZN_priorNMF-imputed scRNA-seq data, and bulk RNA-seq data. This analysis consists of three parts. First, we visualized gene–gene correlation matrices using heatmaps (Fig 4a). The results show that correlation patterns derived from scZN_priorNMF-imputed data exhibit substantially higher concordance with bulk RNA-seq than those derived from raw scRNA-seq data.

Second, to evaluate the robustness of gene–gene correlation structures after imputation, we performed a stability analysis. Specifically, we treated the bulk RNA-seq correlation matrix as a fixed reference and randomly subsampled 20% of cells from both the raw and imputed scRNA-seq expression matrices. Agreement with the bulk reference was quantified using two complementary metrics: (i) the Pearson correlation between the upper-triangular elements of the correlation matrices (similarity; higher is better), and (ii) the Frobenius norm of their difference (distance; lower is better). This procedure was repeated for 50 iterations. The distributions of these metrics were visualized using box plots (Fig 4b). The results indicate that scZN_priorNMF consistently increases similarity to bulk RNA-seq correlations while reducing distance, demonstrating that the recovered correlation structure is both effective and robust.

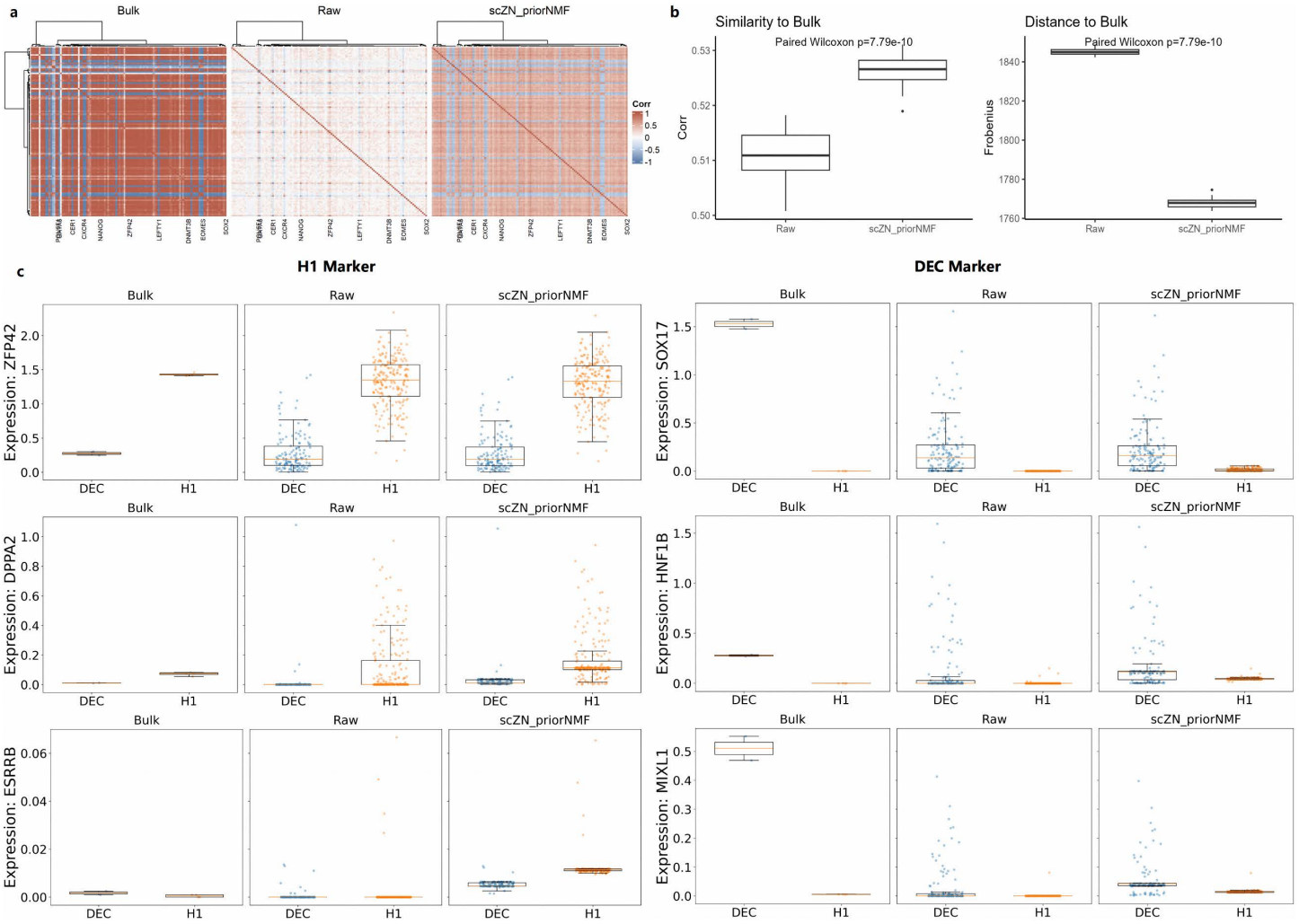

Fig 4. Consistency with bulk RNA-seq and recovery of cell-type–specific gene expression. (a) Gene–gene correlation heatmaps computed from bulk RNA-seq data, raw single-cell RNA-seq data, and scZN_priorNMF-imputed data. Hierarchical clustering is applied consistently across panels. **(b)** Quantitative comparison of similarity to bulk RNA-seq. Left: distribution of correlation similarity scores between single-cell data and bulk RNA-seq. Right: Frobenius distance between gene–gene correlation matrices derived from single-cell and bulk RNA-seq data. Statistical significance is assessed using paired Wilcoxon signed-rank tests. **(c)** Expression distributions of representative marker genes for the H1 and DEC cell types, shown for bulk RNA-seq, raw single-cell data, and scZN_priorNMF-imputed data. Box plots and overlaid points illustrate gene expression levels across cell types.

https://doi.org/10.1371/journal.pcbi.1014051.g004

Third, we visualized the expression patterns of H1 marker genes (ZFP42, DPPA2, and ESRRB) and DEC marker genes (SOX17, HNF1B, and MIXL1) (Fig 4c). The results show that, relative to bulk RNA-seq, scZN_priorNMF-imputed data preserve the gene–gene correlation structure observed in the raw data for genes that already display differential expression, while introducing minimal additional noise. Notably, for certain genes that exhibit differential expression in bulk RNA-seq but appear non-differential in raw scRNA-seq data due to dropout, scZN_priorNMF is able to recover these biologically meaningful differences. In particular, ESRRB is expected to be highly expressed in H1 cells. However, this pattern is not clearly observed in the raw scRNA-seq data. After scZN_priorNMF imputation, the expected cell-type-specific expression of ESRRB is successfully restored. These results indicate that scZN_priorNMF effectively recovers biologically meaningful gene expression patterns and preserves coherent gene–gene correlation networks.

scZN strengthens cellular trajectory reconstruction

Single-cell RNA-seq (scRNA-seq) not only improves cell-type resolution, but also enables ordering cells along temporal and developmental axes via trajectory inference. Standard trajectory algorithms typically reconstruct a latent path that cells are assumed to traverse and then place cells onto that path. However, these methods generally do not account for zero inflation/dropout, which can distort both manifold geometry and ordering.

To evaluate whether imputation can improve trajectory reconstruction, we analyzed the D2 time-course scRNA-seq dataset of H1 embryonic stem cells differentiating toward definitive endoderm cells (DEC) using Monocle 3 [37]. We first generated a 2-D UMAP embedding from the raw data, annotated cells by induction time, and connected timepoints sequentially to depict the empirical chronology (Fig 5a, left). We then computed pseudotime with Monocle 3 and overlaid the learned principal graph (Fig 5a, right).

The results indicated two key issues. First, consistent with dropout artifacts, cells at 72 h and 96 h were poorly resolved and appeared intermixed in the raw time-course embedding. Second, the pseudotime ordering was largely discordant: aside from progenitor cells being placed plausibly early, most subsequent timepoints were misassigned. Marker dynamics reinforced these concerns. The pluripotency genes DNMT3B and POU5F1 (OCT4), and the DEC marker HNF1B, showed

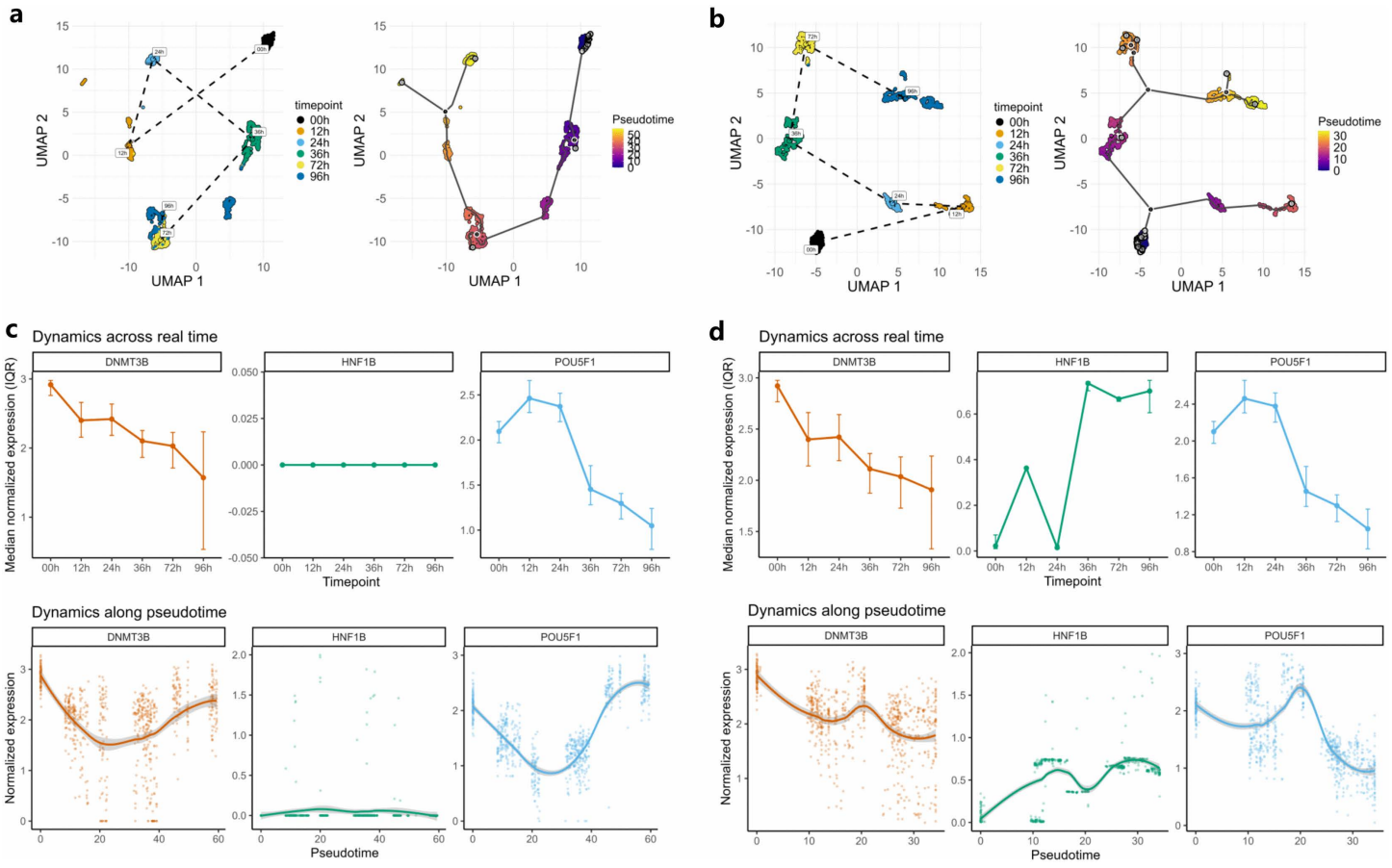

Fig 5. scZN enhances pseudotime analysis with Monocle 3. (a) Pseudotime analysis of the raw embryonic stem cell differentiation data; the left subpanel shows the correct differentiation trajectory derived from temporal labels, and the right subpanel shows Monocle 3's pseudotime result. (b) Analysis results on the imputed data. (c) and (d): Before/after imputation analyses for DNMT3B, POU5F1, and the DEC marker HNF1B; the top row shows the true temporal progression, and the bottom row shows the pseudotime inference.

https://doi.org/10.1371/journal.pcbi.1014051.g005

trajectories that were globally inconsistent with the raw data; in particular, HNF1B exhibited near-absent expression in the raw counts even though it is expected to increase during endoderm specification. Taken together, these observations indicate that pseudotime estimation on the raw dataset is unreliable and likely confounded by dropout, motivating a direct comparison with imputed data to assess whether imputation restores biologically coherent trajectories.

Applying the same workflow to the scZN_priorNMF–imputed dataset yielded markedly improved structure. In the UMAP embedding, clusters were cleanly separated and the temporal path exhibited minimal overlap (Fig 5b, left). The Monocle 3 pseudotime reconstruction correctly ordered most time points, with only a localized misassignment between 12 h and 24 h (Fig 5b, right). Marker dynamics were likewise improved. In particular, HNF1B expression was restored in a manner consistent with definitive endoderm specification.

A broader comparison across imputation methods (S8 Fig) supported these findings. Among alternatives, only scIGANs and scImpute effectively mitigated UMAP-level mixing. For temporal ordering, scIGANs ranked second to scZN_priorNMF. All other methods failed to improve trajectory inference. Collectively, these results indicate that appropriate imputation—especially scZN_priorNMF—can materially enhance trajectory analysis in time-course scRNA-seq.

scZN enhances RNA velocity analysis

RNA velocity is a powerful approach that leverages the ratio of spliced to unspliced mRNA to estimate instantaneous changes in gene expression, thereby inferring cell-state transition trajectories. Compared with pseudotime-based analyses, it is more sensitive to subtle cellular changes. However, existing scRNA-seq velocity methods have not accounted for the impact of dropout events. We hypothesize that appropriate imputation can improve RNA-velocity analysis.

Using the mouse dentate gyrus dataset with known ground-truth lineages—Radial glia-like → Astrocytes and nIPC → Neuroblast → Granule immature → Granule mature—we ran two RNA-velocity inference models (scVelo and VeloVI) before and after imputation. We first evaluated cross-boundary directional consistency to assess trajectory accuracy. As shown in Fig 6a, after imputation both methods improved by more than 25%. We also present latent time (Fig 6b), UMAP velocity streamlines, velocity-derived pseudotime, and gene heatmaps along the inferred timeline. After imputation, these temporal trends become clearer and more accurate. In addition, fitting the spliced/unspliced dynamics for the genes in the heatmap yielded trajectories and temporal ordering that were more consistent with expectations (S9 Fig). Together, these results demonstrate that scZN mitigates dropout effects and enhances RNA velocity analysis.

Alzheimer's disease scRNA-seq analysis

To evaluate the applicability of scZN in real biological contexts and on extremely sparse scRNA-seq data, we performed imputation and downstream analyses on the GSE138852 dataset (10,850 genes and 13,124 cells, dropout rate 93.53%, including Alzheimer's disease (AD) and control (ct)). In the original data, the UMAP representation showed pronounced cluster adhesion and weak separation of cellular heterogeneity (Fig 7a). After scZN imputation, cellular heterogeneity increased and clustering metrics improved (Fig 7b). Pairwise differential analyses with volcano plots (Fig 7c) and KEGG enrichment (Fig 7d) further indicated that among upregulated pathways the AD group was significantly enriched for IgSF CAM signaling, Spinocerebellar ataxia, and EGFR tyrosine kinase inhibitor resistance, which suggests enhanced adhesion and immune activity and stress responses downstream of receptor tyrosine kinases. The ct group showed upregulation of the MAPK signaling pathway, Non-alcoholic fatty liver disease, and Carbon metabolism, indicating more intact canonical signaling and mitochondrial carbon-metabolic homeostasis. This pattern aligns with the canonical molecular landscape of AD, characterized by heightened neuroinflammation, blood–brain barrier and glial responses along with impaired energy metabolism and synaptic homeostasis [38]. Therefore, scZN can effectively enhance cell-type resolution in highly sparse single-cell transcriptomic data while preserving biologically meaningful differential signals.

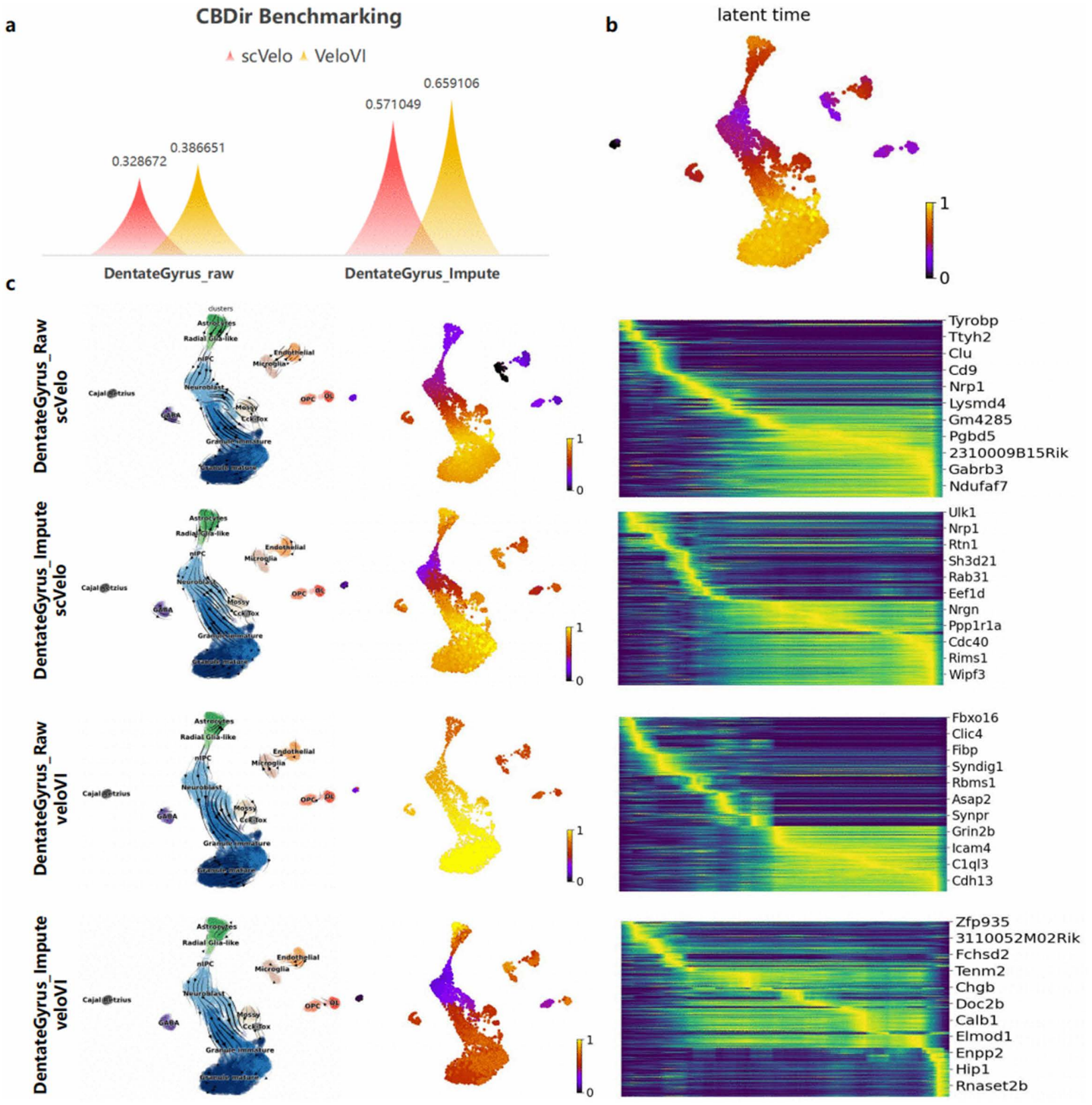

Fig 6. RNA velocity analysis with scVelo and VeloVI on datasets before and after scZN imputation. (a) Evaluation of velocity flow accuracy along cell boundaries before and after imputation using the CBDir metric. **(b)** Expression dynamics across latent time in mouse dentate gyrus. **(c)** RNA velocity analyses by scVelo and VeloVI on raw and imputed data, showing (from left to right) the velocity field on the UMAP embedding, velocity-inferred pseudo-time, and gene-expression heatmaps.

https://doi.org/10.1371/journal.pcbi.1014051.g006

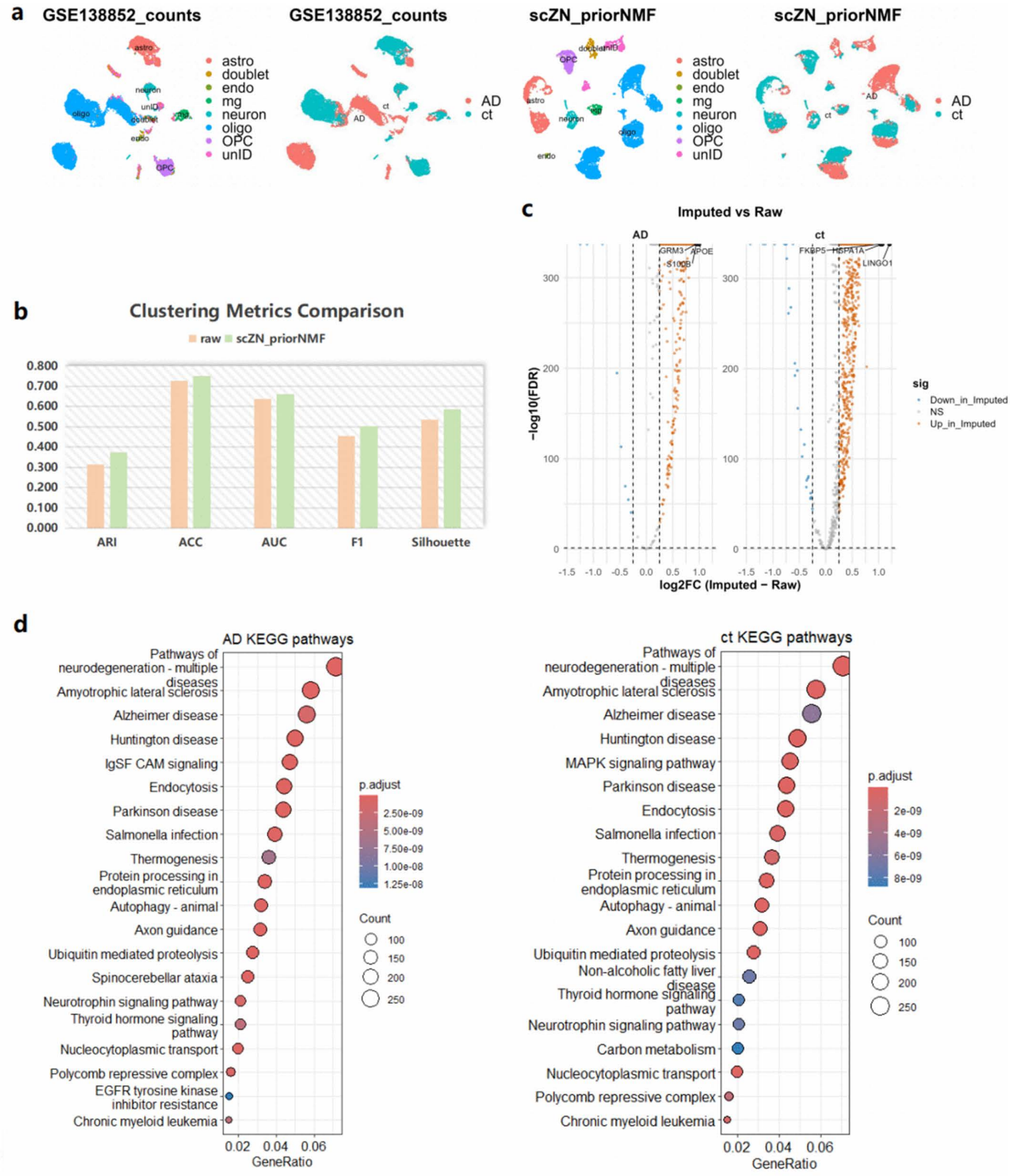

Fig 7. scRNA-seq analysis of Alzheimer's disease (AD) data with scZN imputation. (a): UMAP cell-type classifications and AD vs. control comparisons before and after imputation. (b): External clustering metrics compared before and after imputation. (c): Genes significantly upregulated in the imputed data relative to the original AD and control datasets. (d): KEGG pathway analysis of the two upregulated gene sets.

https://doi.org/10.1371/journal.pcbi.1014051.g007

Discussion

We present scZN, a single-cell RNA-seq (scRNA-seq) imputation framework that directly addresses two pervasive shortcomings of current deep learning methods: (i) treating all zeros as missing values and applying global smoothing, and (ii) restricting imputation to highly variable genes (HVGs) selected by Scanpy, which forces downstream analyses to depend on a consistent HVG selection tool. The core innovation of scZN is not to propose another smoother, but to redefine imputation as an interpretable factorization process jointly constrained by a count-based observation model and biological priors.

Specifically, scZN performs optimization over the entire gene set within an NMF space: we use per-cell library-size factors for observation-level normalization to decouple technical scale from biological signal and inject cell-type priors via linear shrinkage and anchor gene–module factors to prototype expression means, thereby alleviating the non-convexity of nonnegative matrix factorization (NMF) and the instability of random initialization. We impute only zero values to preserve reliable nonzero counts, and employ composite regularization to further refine the estimates. In comprehensive benchmarking across multiple real datasets, scZN consistently outperforms more than a dozen competing methods while maintaining biological consistency and improving pseudotime and RNA velocity analyses. In sum, the core contribution of scZN is to place imputation within an interpretable NMF space that simultaneously accounts for count statistics and biological priors, thereby addressing common problems of deep learning models at their source. Moreover, during the AD analysis, we observed upregulated neuroinflammation-related pathways, which both align with previous research and validate the effectiveness and real-world applicability of scZN.

Despite scZN's strong performance across multiple benchmarks, it still has limitations: it requires pre-specifying the factorization rank k, which is currently chosen via clustering, and a poor choice can degrade performance. After introducing deep-learning–based optimization, the computational cost and training time become higher than with purely statistical models. Moreover, under the unsupervised setting, scZN achieves only limited improvement over the raw data and ranks third among the evaluated methods (Fig 2c). This result highlights a fundamental limitation: without external supervision or biological priors, recovering complex gene expression structure from sparse scRNA-seq data remains intrinsically challenging. In the absence of label or module-level constraints, the factorization problem is highly underdetermined, particularly under severe dropout, which limits the achievable performance gains of unsupervised imputation. Therefore, in future work we will improve the framework to enhance its imputation performance in unsupervised settings. In addition, we will extend interpretable factorization—constrained by count statistics and biological priors—to multi-omics and spatial data (e.g., scATAC-seq, multiome, and spatial transcriptomics). For reference mapping and data integration, we will use spatial coordinates or histological images as priors to build spatially constrained, interpretable factorizations that enable cross-modal "spatial remapping" and unified imputation. We also plan to incorporate spliced/unspliced counts into the modeling to improve the joint analysis of pseudotime and RNA velocity.

Methods

Ethics statement

No human or animal subjects were involved in this study. All data analyzed were obtained from published literature or public databases, and thus ethical approval was not required.

Data preparation

We compiled a comprehensive collection of scRNA-seq datasets, which can be categorized into four distinct groups. The first group (D1), sourced from Xu et al. [25], comprises seven datasets representing a broad spectrum of data types and experimental platforms. These datasets include Human brain scRNA-seq data, which provides high-quality data from human brain tissues for assessing imputation in complex biological systems. (dropout 81%) https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE67835;

ERCC spike-in RNA scRNA-seq dataset, featuring spike-in RNA molecules and an atypical setting with more cells than genes, is used to benchmark imputation performance on extremely small gene expression matrices. (dropout 33%) https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-2512;

A mouse embryonic stem cell (ESC) scRNA-seq dataset, encompassing embryonic stem cell differentiation and large-scale measurements, is used to evaluate imputation performance in developmental contexts. (dropout 70%) https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE65525;

Time-course scRNA-seq data are used to evaluate the ability of imputation methods to preserve temporal gene expression dynamics under a high dropout rate. (dropout 55%) https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE75748;

sc_Drop-seq data generated by droplet-based sequencing are used to evaluate cross-platform robustness of imputation methods. (dropout 62%) https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE118706;

sc_CEL-seq2 data generated by plate-based sequencing are employed to benchmark the adaptability of imputation methods to high-precision platforms. (dropout 74%) https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE117617;

sc_10X data generated by the 10X Genomics platform are used to evaluate the broad applicability of the method. (dropout 45%) https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE111108.

The second group (D2), derived from Aivazidis et al. [39], comprises dentate gyrus scRNA-seq data, a challenging setting where most RNA velocity inference methods perform poorly. This dataset is therefore used to evaluate whether imputation can enhance inference accuracy.

The third group (D3) consists of an Alzheimer's disease (AD) scRNA-seq dataset that provides gene expression profiles for both AD and control samples and exhibits an extremely high dropout rate (93.53%). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138852. We therefore used it to assess the practical utility of scZN.

The fourth group (D4) consists of a human embryonic stem cell (ESC) scRNA-seq dataset for differential expression analysis. This dataset includes six bulk RNA-seq samples, which are used as a reference to compare gene–gene relationships before and after imputation. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75748

In addition, prior to all result analyses, we performed Seurat-based quality control on the scRNA-seq data [40], specifically excluding mitochondrial (MT) genes and selecting 2,000 highly variable genes for analysis.

The framework of scZN

scZN generative framework. In scZN, we assume that whether a zero entry for a gene in a given cell should be imputed is not determined by the zero value itself, but rather by the position of this entry in a latent similarity space. Specifically, if a gene exhibits consistent and stable nonzero expression among cells that are similar to the target cell, then an observed zero is more likely to arise from technical dropout; conversely, if the gene is intrinsically lowly expressed or highly variable among similar cells, the zero more likely reflects true biological absence and should not be corrected. This principle motivates a generative formulation in which imputation is driven by a latent biological expression generator defined over a similarity space, rather than by heuristic rules applied directly to observed zeros.

Let the scRNA-seq count matrix be denoted by $\mathbf{X} \in \mathbb{N}_0^{n\times m}$, where n is the number of cells and m is the number of genes. We first estimate a cell-specific size factor that captures library-size and capture-rate variation exclusively at the observation layer:

$$s_i = \frac{\sum_{j=1}^{m} X_{ij}}{\text{median}_r \sum_{j=1}^{m} X_{rj}}, \qquad i = 1, \ldots, n. \tag{1}$$

This choice is justified by Poisson thinning: if the biological molecule counts satisfy $Y_{ij}^{\text{bio}} \sim \text{Poisson}(\lambda_{ij})$ and each molecule is captured independently with a cell-specific probability p_i, then the observed counts follow $Y_{ij}^{\text{obs}} \sim \text{Poisson}(p_i\lambda_{ij})$, implying

$s_i \propto p_i$. Thus, technical scaling is handled by s_i at the observation layer, without entangling sequencing depth with the biological mean.

Our purpose is to generate a matrix $\boldsymbol{\mu}$ that represents the imputed results inferred by the model. Using PyTorch [41], we first initialize two trainable factor matrices: $\mathbf{W} \in \mathbb{R}^{n\times k}$ and $\mathbf{H} \in \mathbb{R}^{k\times m}$, where k denotes the number of cell types, cell clusters, or latent biological modules. The generated matrix $\boldsymbol{\mu}$ is defined as the matrix product of these two factors:

$$\boldsymbol{\mu} = \mathbf{WH} \tag{2}$$

The i-th row $\mathbf{w}_i$ of matrix $\mathbf{W}$ represents the nonnegative module membership probabilities of cell i, which can be interpreted as a probability distribution over k modules; the l-th row $\mathbf{h}_l$ of matrix $\mathbf{H}$ describes the gene expression profile associated with module l. The key interpretation of this generative matrix is as follows: for a given cell–gene pair (i, j), if cell i is close, in the induced similarity space, to a group of cells that stably and highly express gene j (i.e., $\mathbf{w}_i$ is close to the representations of those cells), then a relatively reasonable value $\boldsymbol{\mu}_{i,j}$ will be produced via the inner product as the imputed result. This supports the interpretation that the observed zero value is more likely due to technical dropout.

Conversely, if the gene exhibits low expression or high variability among similar cells, the generator will naturally produce a very small value of $\boldsymbol{\mu}_{i,j}$, thereby avoiding over-correction of zero entries.

When cell type labels or other module annotations are available, scZN incorporates this prior information to link latent cell abundance factors with known biological characteristics, resulting in scZN_priorNMF.

Let $\mathbf{P} \in \{0, 1\}^{n\times k}$ denote a one-hot (or soft) encoding matrix of cell type or module assignments. To ensure nonnegativity, we define $\widehat{\mathbf{W}} = \mathrm{ReLU}(\mathbf{W})$ and $\widehat{\mathbf{H}} = \mathrm{ReLU}(\mathbf{H})$, and obtain a prior-aligned cell factor via an equally weighted row-wise softmax operation:

$$\mathbf{W}^{+} = \tfrac{1}{2}\left(\mathbf{P} + \mathrm{softmax}(\widehat{\mathbf{W}})\right) \tag{3}$$

To guide $\boldsymbol{\mu}$ toward canonical expression patterns and reduce sensitivity to random initialization, we generate a simple prototype expression matrix $\mathbf{H}_{\mathrm{prior}} \in \mathbb{R}^{k\times m}$ based on cell-type-wise averages:

$$\mathbf{H}_{\mathrm{prior}}(i, j) = \frac{1}{|\mathcal{C}_i|}\sum_{c\in\mathcal{C}_i} X_{c,j}, \qquad i = 1, \ldots, k,\ j = 1, \ldots, m, \tag{4}$$

where $\mathcal{C}_i$ denotes the index set of cells assigned to the i-th cell type, and $|\mathcal{C}_i|$ is the number of cells in the index set.

$$\mathbf{H}^{+} = \tfrac{1}{2}\left(\mathbf{H}_{\mathrm{prior}} + \widehat{\mathbf{H}}\right), \qquad \widehat{\mathbf{H}} = \mathrm{ReLU}(\mathbf{H}). \tag{5}$$

Finally, we obtain the imputed expression matrix through the prior-aligned factors:

$$\boldsymbol{\mu} = \mathbf{W}^{+}\mathbf{H}^{+} \tag{6}$$

which represents the inferred biological-scale mean expression for each cell–gene pair. To compare the inferred means with the observed scRNA-seq counts, we project $\boldsymbol{\mu}$ back to the observation layer by accounting for cell-specific sequencing depth. Let $\mathbf{s} \in \mathbb{R}^{n}_{>0}$ denote the size-factor vector; the reconstructed count matrix is given by

$$\widehat{\mathbf{X}}^{(\mathrm{recon})} = \mathrm{diag}(\mathbf{s})\,\boldsymbol{\mu} + \varepsilon \tag{7}$$

where ε denotes stochastic observation noise.

When gene-specific batch effects are present, we further incorporate a multiplicative correction term. Let $\kappa_{j,b(i)}$ denote the batch effect for gene j in the batch $b(i)$ of cell i. The reconstruction then becomes

$$\widehat{X}_{ij}^{(\text{recon})} = s_i \, \kappa_{j,b(i)} \, \mu_{ij} + \varepsilon \tag{8}$$

where the identifiability constraint ensures that batch effects are centered on the log scale and do not confound the biological mean expression.

Multiple regularization design

Based on the framework described above, it is evident that the learning process in this work is essentially a non-convex optimization problem. Without sufficient structural constraints, the optimization is prone to producing numerically valid yet biologically implausible imputation results. Accordingly, we construct regularization terms from four aspects: the solution space of optimal solutions, data distribution properties, geometric relationships, and interpretability. Specifically, the overall loss is decomposed into the following components: (i) the NMF reconstruction loss ($\mathcal{L}_{\text{NMF}}$), which measures the mean Frobenius norm error between the original and reconstructed matrices; (ii) the ZINB [34] negative log-likelihood loss ($\mathcal{L}_{\text{ZINB}}$), which models the zero inflation and over-dispersion in the counts; (iii) the z-score [35] regularization loss ($\mathcal{L}_{\text{zscore}}$), which ensures consistency in the standardized gene expression profiles within each cell; and (iv) the cell-type classification loss ($\mathcal{L}_{\text{class}}$), which leverages cell-type labels to constrain the low-dimensional representation and enhance biological interpretability. The following subsections detail the formulation of each loss component.

NMF reconstruction loss. $\mathcal{L}_{\text{NMF}}$ employs the Frobenius norm, which imposes a symmetric quadratic penalty on the reconstruction error of the imputed data under an observation-aligned scale. This is equivalent to assuming a tighter Gaussian tolerance region, which stabilizes the optimization of the non-convex matrix factorization problem. Moreover, the Frobenius norm is the simplest and most stable choice widely adopted in the NMF literature to enforce alignment between the imputed values and the original count matrix. Concretely, let $\boldsymbol{\mu}$ be the biological-scale mean and let $\mathbf{s} \in \mathbb{R}^n_{>0}$ denote per-cell library-size factors, we use the depth-calibrated reconstruction $\mu_{\text{adjusted}} = \text{diag}(\mathbf{s})\,\boldsymbol{\mu}$. The loss is the mean Frobenius norm [42]:

$$\mathcal{L}_{\text{NMF}} = \frac{\left\| \mathbf{X} - \mu_{\text{adjusted}} \right\|_F}{n \times m}, \tag{9}$$

ZINB model. scRNA-seq count data typically exhibit zero inflation (arising from capture dropout and transcriptionally silent states) and overdispersion (due to bursty transcription and cell-to-cell heterogeneity). To constrain these distributional characteristics during reconstruction—without asserting the ZINB model as the true data-generating process—we adopt a ZINB-based distributional regularizer.

Let x_{ij} denote the observed count for gene j in cell i. The ZINB probability mass function $P(x_{ij})$ is defined as

$$\begin{cases} \pi_{ij} + (1-\pi_{ij}) \left(\dfrac{\theta}{\mu_{ij}+\theta} \right)^{\theta} & \text{if } x_{ij} = 0, \\ (1-\pi_{ij}) \dfrac{\Gamma(x_{ij}+\theta)}{\Gamma(\theta)\,\Gamma(x_{ij}+1)} \left(\dfrac{\mu_{ij}}{\mu_{ij}+\theta} \right)^{x_{ij}} \left(\dfrac{\theta}{\mu_{ij}+\theta} \right)^{\theta} & \text{if } x_{ij} > 0, \end{cases} \tag{10}$$

where μ_{ij} is the reconstructed mean at entry (i,j), $\pi_{ij} = \sigma(\cdot) \in [0,1]$ denotes the dropout (structural-zero) probability, and $\theta > 0$ is a shared negative binomial dispersion parameter, jointly optimized with μ_{ij} and π_{ij}.

The corresponding regularization term is given by the negative log-likelihood:

$$\mathcal{L}_{\text{ZINB}} = -\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\log P(x_{ij}). \tag{11}$$

This regularizer explicitly aligns the zero fraction and the mean–variance relationship of the reconstructed data with empirically observed scRNA-seq behavior, discouraging overly smoothed imputations that erase zero entries or underestimate expression dispersion.

Z-score regularization. Many downstream analyses in single-cell studies—such as marker gene ranking, module activity scoring, and gene co-expression analysis—depend primarily on relative expression patterns within cells rather than absolute count magnitudes. Consequently, preserving the internal geometry of gene expression profiles is often more critical than matching raw counts exactly. By computing the loss on z-score–normalized expression values, we enforce consistency between reconstructed and observed data in a scale-free space that emphasizes relative deviations. This prevents the imputation process from collapsing heterogeneous cellular expression profiles into a common structure and helps preserve intrinsic expression geometry across genes within each cell.

From a maximum a posteriori (MAP) viewpoint, minimizing the mean squared error on z-scores is equivalent to imposing a Gaussian prior on the standardized, scale-invariant relative expression geometry. Under this assumption, the MSE loss naturally arises as the negative log-likelihood associated with that prior. Therefore, for cell i,

$$\bar{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij}, \qquad \sigma_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(x_{ij}-\bar{x}_i\right)^2}, \qquad z_{ij} = \frac{x_{ij}-\bar{x}_i}{\sigma_i+\epsilon}, \tag{12}$$

and for the reconstruction,

$$\bar{\mu}_i = \frac{1}{m}\sum_{j=1}^{m} \mu_{ij}, \qquad \hat{\sigma}_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\mu_{ij}-\bar{\mu}_i\right)^2}, \qquad \hat{z}_{ij} = \frac{\mu_{ij}-\bar{\mu}_i}{\hat{\sigma}_i+\epsilon}. \tag{13}$$

We penalize standardized discrepancies:

$$\mathcal{L}_{\text{zscore}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(z_{ij}-\hat{z}_{ij}\right)^2, \tag{14}$$

so that, after removing per-cell scale/offset, the reconstruction preserves each cell's intrinsic relative expression shape.

Cell-type classification loss. The purpose of $\mathcal{L}_{\text{class}}$ is to anchor $\boldsymbol{\mu}$ to a latent semantic representation. In highly non-convex optimization settings, reconstruction and distributional constraints alone may yield solutions that are numerically valid but biologically uninterpretable. By requiring known cell identities to be linearly decodable from the latent embedding $\mathbf{W}_i^+$, the classification loss introduces an effective semantic anchor that eliminates such degenerate solutions, without enforcing a one-to-one correspondence between modules and cell types. Let $\mathbf{W}_i^+$ be the prior-aligned embedding of cell i, a linear probe produces class probabilities

$$\hat{\mathbf{y}}^{\,i} = \text{softmax}\left(\mathbf{W}_i^+\,\mathbf{w}_c + b\right), \tag{15}$$

with $\mathbf{w}_c \in \mathbb{R}^{k\times C}$, $b \in \mathbb{R}^C$, and C cell types. The cross-entropy is

$$\mathcal{L}_{\text{class}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\mathbf{1}\{y_i = c\}\,\log \hat{y}_{ic}. \tag{16}$$

Because a truly biological module coordinate should be linearly decodable, this term both tests and enforces that $\mathbf{W}^+$ carries cell-identity information, suppressing biologically implausible module mixing and stabilizing rare states.

Optimization. We minimize the total objective

$$\mathcal{L} = \lambda_{\text{NMF}}\mathcal{L}_{\text{NMF}} + \lambda_{\text{zinb}}\mathcal{L}_{\text{ZINB}} + \lambda_{\text{zscore}}\mathcal{L}_{\text{zscore}} + \lambda_{\text{class}}\mathcal{L}_{\text{class}}, \tag{17}$$

where each coefficient λ controls the contribution of its term. All trainable parameters $\Theta = \{\mathbf{W}, \mathbf{H}, \pi, \theta, \mathbf{w}_c, b\}$ are updated with Adam [43]. Let $g_t = \nabla_\Theta \mathcal{L}(\Theta)$. Adam updates are

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon_{\text{adam}}} \tag{18}$$

where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, $\epsilon_{\text{adam}} = 10^{-8}$ prevents numerical instability, and η_t is the learning rate, which may be decayed according to

$$\eta_t = \eta_0 \times \gamma^{\lfloor t/T_{\text{decay}} \rfloor} \tag{19}$$

with initial rate η_0, decay factor $\gamma < 1$, and decay interval T_{decay}.

Stopping criterion. Training stops when the objective plateaus, e.g., the relative change of the (training or validation) loss stays below a tolerance $\epsilon_{\text{loss}} = 0.001$ for W consecutive epochs:

$$\frac{\left|\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}\right|}{\mathcal{L}^{(t-1)}} < \epsilon_{\text{loss}} \quad \text{for } t \in \{t_0 - W + 1, \ldots, t_0\}. \tag{20}$$

Parameters of scZN

The main parameters of scZN are the factorization rank k and the regularization coefficients. We typically set k to the number of cell types to ensure biological interpretability, yielding a factorization into (i) a cell-by–cell-type probability matrix and (ii) a cell-type–by–gene expression matrix.

Imputation evaluation

We evaluate the accuracy of imputation using ARI, ACC, F1, AUC, and the silhouette coefficient. The computations are as follows:

ARI (Adjusted Rand Index). Pairwise agreement between the predicted partition and ground truth, corrected for chance:

$$\text{ARI} = \frac{\text{observed pairwise agreement} - \text{expected agreement at random}}{\text{maximum possible agreement} - \text{expected agreement at random}} \tag{21}$$

ACC (Clustering Accuracy). Match predicted clusters to ground-truth classes via the Hungarian algorithm, then compute the proportion of correctly assigned samples:

$$\text{ACC} = \frac{\text{correct assignments after matching}}{\text{total samples}} \tag{22}$$

F1. Compute precision and recall per class, then aggregate:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{23}$$

AUC (ROC–AUC). Using one-vs-rest continuous scores, trace the ROC curve across thresholds and take its area. For multiclass, report the macro-average:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, \text{d}(\text{FPR}), \qquad \text{AUC}_{\text{macro}} = \text{mean of classwise AUCs} \tag{24}$$

Silhouette coefficient. For each sample, compare the mean dissimilarity to its own cluster with that to the nearest other cluster:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}, \tag{25}$$

where $a(i)$ is the average distance from i to samples in its own cluster, and $b(i)$ is the minimum average distance to any other cluster. The dataset silhouette is the mean of $s(i)$ over all samples.

CBDir. We computed the CBDir score using the functions provided by the UniTVelo [44] Python package. In the UniTVelo paper, CBDir evaluates the correctness of the transition from a source cluster to a target cluster by using boundary cells defined by the ground truth. Here, the boundary of the source cluster consists of cells in that cluster that are adjacent to the target cluster, and vice versa. Boundary cells are used because they reflect short-term developmental dynamics. CBDir is defined as

$$\text{CBDir}(c) = \frac{1}{\text{Norm}} \sum_{c' \in C_A \cap N(c)} \frac{\mathbf{v}_c \cdot (\mathbf{x}_{c'} - \mathbf{x}_c)}{\|\mathbf{v}_c\|\, \|\mathbf{x}_{c'} - \mathbf{x}_c\|}, \qquad \text{Norm} = \left| \{c' \in C_A \cap N(c)\} \right| \tag{26}$$

Here, C_A denotes the set of cells in the target cluster A, $N(c)$ denotes the neighbors of cell c, and $\mathbf{v}_c$ and $\mathbf{x}_c$ are the inferred velocity and low-dimensional position vectors of cell c, respectively. Thus, $\mathbf{x}_{c'} - \mathbf{x}_c$ represents the short-term displacement in the embedding space.

Gene-level Pearson correlation. We estimate gene-level consistency before and after imputation using the Pearson correlation as follows. For each gene g, compute the Pearson correlation between its expression vector before imputation and after imputation:

$$r_g = \text{corr}\left(\text{raw}(g),\ \text{imputed}(g)\right) \tag{27}$$

Summarize the dataset-level consistency by aggregating across genes.

Supporting information

S1 Fig. The performance of all methods across D1. Box plots summarizing the performance of all methods across D1 (ERCC spike-in, Human Brain, Timecourse, mESC, scDrop-seq, scCelseq2, and sc10X). Columns correspond to evaluation metrics, and rows represent datasets.
(PDF)

S2 Fig. Label sensitivity analysis. Label sensitivity analysis during the imputation process across D1, shaded regions around the curves indicate variability.
(PDF)

S3 Fig. Human brain marker gene heatmap. This figure shows the heatmap expression of labeled genes after imputation on a human dataset using multiple methods.
(PDF)

S4 Fig. Volcano plots. Genes significantly upregulated by multiple imputation methods.
(PDF)

S5 Fig. Significance comparison of GAD1 in Neurons and OPC. Comparison of the significance of GAD1 in Neurons and OPC datasets after imputation using 14 methods.
(PDF)

S6 Fig. Significance comparison of marker genes in Neurons and OPC. Significance comparison before and after imputation of SLC17A7, CLDN11, SLC6A1, GRIN1, CSPG4, and SLC44A1.
(PDF)

S7 Fig. KEGG pathway. Changes in KEGG pathways of upregulated genes in human brain datasets after scZN imputation.
(PDF)

S8 Fig. Monocle 3–based pseudotime analysis. Focusing on comparing pseudotime results using data imputed by other methods.
(PDF)

S9 Fig. Gene analysis based on RNA velocity. Focusing on fitting the spliced/unspliced dynamics of the genes in the Fig 5c heatmap and supplementing with corresponding UMAP plots of velocity and gene expression.
(PDF)

S1 Table. External consistency evaluation. For the D1 dataset, we evaluated ARI, ACC, AUC, F1, and the silhouette coefficient on the data before and after processing by 14 imputation methods.
(XLS)

S2 Table. Robustness Analysis of scZN and scZN_priorNMF across D1. We conducted five independent runs on the ERCC Spike-in, Time-course, and Dentate Gyrus scRNA-seq datasets to assess stability with and without incorporating priors, in order to determine whether the framework's non-convexity is mitigated.
(XLS)

Author contributions

Conceptualization: You Wu.

Data curation: Ye Win Aung.

Investigation: Ye Win Aung, Alex Michel Daoud.

Methodology: You Wu.

Project administration: Li Xu.

Resources: Li Xu.

Software: You Wu.

Supervision: Li Xu.

Validation: Alex Michel Daoud.

Visualization: You Wu.

Writing – original draft: Li Xu.

Writing – review & editing: Ye Win Aung, Alex Michel Daoud.

References

1. Qu H-Q, Kao C, Hakonarson H. Single-cell RNA sequencing technology landscape in 2023. Stem Cells. 2024;42(1):1–12. https://doi.org/10.1093/stmcls/sxad077 PMID: 37934608
2. Zhu Z, Jiang L, Ding X. Advancing breast cancer heterogeneity analysis: insights from genomics, transcriptomics and proteomics at bulk and single-cell levels. Cancers (Basel). 2023;15(16):4164. https://doi.org/10.3390/cancers15164164 PMID: 37627192
3. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13(1):1246. https://doi.org/10.1038/s41467-022-28803-w PMID: 35273156
4. Serrano-Ron L, Perez-Garcia P, Sanchez-Corrionero A, Gude I, Cabrera J, Ip P-L, et al. Reconstruction of lateral root formation through single-cell RNA sequencing reveals order of tissue initiation. Mol Plant. 2021;14(8):1362–78. https://doi.org/10.1016/j.molp.2021.05.028 PMID: 34062316
5. Khan SU, Huang Y, Ali H, Ali I, Ahmad S, Khan SU, et al. Single-cell RNA sequencing (scRNA-seq): advances and challenges for cardiovascular diseases (CVDs). Current problems in cardiology. 2024;49(2):102202.
6. Kim D, Lim B, Jang M, Lim S, Kim J. RNA sequencing and data analysis: a revolutionary approach to transcriptome profiling in livestock. Bioinformatics in veterinary science: Vetinformatics. 2025. p. 23–40.
7. Weiner AC, Williams MJ, Shi H, Vázquez-García I, Salehi S, Rusk N, et al. Inferring replication timing and proliferation dynamics from single-cell DNA sequencing data. Nat Commun. 2024;15(1):8512. https://doi.org/10.1038/s41467-024-52544-7 PMID: 39353885
8. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: a brief overview. Clin Transl Med. 2022;12(3):e694. https://doi.org/10.1002/ctm2.694 PMID: 35352511
9. Jia C, Hu Y, Kelly D, Kim J, Li M, Zhang NR. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res. 2017;45(19):10978–88. https://doi.org/10.1093/nar/gkx754 PMID: 29036714
10. Brendel M, Su C, Bai Z, Zhang H, Elemento O, Wang F. Application of deep learning on single-cell RNA sequencing data analysis: a review. Genomics Proteomics Bioinformatics. 2022;20(5):814–35. https://doi.org/10.1016/j.gpb.2022.11.011 PMID: 36528240
11. Dijk D v, Nainys J, Sharma R, Kaithail P, Carr AJ, Moon KR, et al. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. BioRxiv. 2017;:111591.
12. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics. 2018;19(1):220. https://doi.org/10.1186/s12859-018-2226-y PMID: 29884114
13. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018;15(7):539–42. https://doi.org/10.1038/s41592-018-0033-z PMID: 29941873
14. Kang B, Abeysinghe E, Agarwal D, Wang Q, Pamidighantam S, Huang M, et al. Online single-cell RNA-seq data denoising with transfer learning. In: Practice and Experience in Advanced Research Computing. 2020. p. 469–72. https://doi.org/10.1145/3311790.3399617
15. Wagner F, Barkley D, Yanai I. Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis. BioRxiv. 2019:655365.
16. Peng T, Zhu Q, Yin P, Tan K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. Genome Biol. 2019;20(1):88. https://doi.org/10.1186/s13059-019-1681-8 PMID: 31060596
17. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018;9(1):997. https://doi.org/10.1038/s41467-018-03405-7 PMID: 29520097
18. Sha Y, Qiu Y, Zhou P, Nie Q. Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. Nat Mach Intell. 2024;6(1):25–39. https://doi.org/10.1038/s42256-023-00763-w PMID: 38274364
19. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biol. 2019;20(1):211. https://doi.org/10.1186/s13059-019-1837-6 PMID: 31627739
20. Talwar D, Mongia A, Sengupta D, Majumdar A. AutoImpute: autoencoder based imputation of single-cell RNA-seq data. Sci Rep. 2018;8(1):16329. https://doi.org/10.1038/s41598-018-34688-x PMID: 30397240
21. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):390. https://doi.org/10.1038/s41467-018-07931-2 PMID: 30674886

22. Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, et al. A Python library for probabilistic analysis of single-cell omics data. Nat Biotechnol. 2022;40(2):163–6. https://doi.org/10.1038/s41587-021-01206-w PMID: 35132262
23. Virshup I, Bredikhin D, Heumos L, Palla G, Sturm G, Gayoso A, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat Biotechnol. 2023;41(5):604–6. https://doi.org/10.1038/s41587-023-01733-8 PMID: 37037904
24. He Y, Yuan H, Wu C, Xie Z. DISC: a highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. Genome Biol. 2020;21(1):170. https://doi.org/10.1186/s13059-020-02083-3 PMID: 32650816
25. Xu Y, Zhang Z, You L, Liu J, Fan Z, Zhou X. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. Nucleic Acids Res. 2020;48(15):e85. https://doi.org/10.1093/nar/gkaa506 PMID: 32588900
26. Wu Y, Xu L, Cong X, Li H, Li Y. Scmaskgan: masked multi-scale CNN and attention-enhanced GAN for scRNA-seq dropout imputation. BMC Bioinformatics. 2025;26(1):130. https://doi.org/10.1186/s12859-025-06138-9 PMID: 40394489
27. Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat Commun. 2021;12(1):1882. https://doi.org/10.1038/s41467-021-22197-x PMID: 33767197
28. Yun S, Lee J, Park C. Single-cell RNA-seq data imputation using feature propagation. arXiv preprint. 2023. https://arxiv.org/abs/2307.10037
29. Lee J, Yun S, Kim Y, Chen T, Kellis M, Park C. Single-cell RNA sequencing data imputation using bi-level feature propagation. Brief Bioinform. 2024;25(3):bbae209. https://doi.org/10.1093/bib/bbae209 PMID: 38706317
30. Ahn SJ, Um D, Yeo Y, Yoon JW. Gene-gene relationship modeling based on genetic evidence for single-cell RNA-Seq data imputation. In: Advances in Neural Information Processing Systems 37, 2024. p. 18882–909. https://doi.org/10.52202/079017-0598
31. Yang X, Zhu T, Peng S, Nie F, Lin Z. Semi-supervised pivotal-aware nonnegative matrix factorization with label and pairwise constraint propagation for data clustering. Pattern Recognition. 2025;157:110933. https://doi.org/10.1016/j.patcog.2024.110933
32. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017;18(1):174. https://doi.org/10.1186/s13059-017-1305-0 PMID: 28899397
33. Zhu X, Meng S, Li G, Wang J, Peng X. AGImpute: imputation of scRNA-seq data based on a hybrid GAN with dropouts identification. Bioinformatics. 2024;40(2):btae068. https://doi.org/10.1093/bioinformatics/btae068 PMID: 38317025
34. Zhou W, Huang D, Liang Q, Huang T, Wang X, Pei H, et al. Early warning and predicting of COVID-19 using zero-inflated negative binomial regression model and negative binomial regression model. BMC Infect Dis. 2024;24(1):1006. https://doi.org/10.1186/s12879-024-09940-7 PMID: 39300391
35. Curtis A, Smith T, Ziganshin B, Elefteriades J. The mystery of the Z-score. Aorta. 2016;04(04):124–30. https://doi.org/10.12945/j.aorta.2016.16.014
36. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. https://doi.org/10.1186/s13059-017-1382-0 PMID: 29409532
37. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496–502. https://doi.org/10.1038/s41586-019-0969-x PMID: 30787437
38. Heneka MT, Van der Flier WM, Jessen F, Hoozemanns J, Thal DR, Boche D, et al. Neuroinflammation in Alzheimer disease. Nature Reviews Immunology. 2024;:1–32.
39. Aivazidis A, Memi F, Kleshchevnikov V, Er S, Clarke B, Stegle O, et al. Cell2fate infers RNA velocity modules to improve cell fate prediction. Nature Methods. 2025;:1–10.
40. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502. https://doi.org/10.1038/nbt.3192 PMID: 25867923
41. Imambi S, Prakash KB, Kanagachidambaresan G. "PyTorch", Programming with TensorFlow: solution for edge computing applications. 2021. p. 87–104.
42. Li XP, Wang Z-Y, Shi Z-L, So HC, Sidiropoulos ND. Robust tensor completion via capped frobenius norm. IEEE Trans Neural Netw Learn Syst. 2024;35(7):9700–12. https://doi.org/10.1109/TNNLS.2023.3236415 PMID: 37021988
43. Zhou P, Xie X, Lin Z, Yan S. Towards understanding convergence and generalization of AdamW. IEEE Trans Pattern Anal Mach Intell. 2024;46(9):6486–93. https://doi.org/10.1109/TPAMI.2024.3382294 PMID: 38536692
44. Gao M, Qiao C, Huang Y. UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. Nat Commun. 2022;13(1):6586. https://doi.org/10.1038/s41467-022-34188-7 PMID: 36329018