

# Supplementary Figures for “Comprehensive benchmarking of RNA velocity methods across single-cell datasets”

Yida Wu<sup>1†</sup>, Chuihan Kong<sup>1†</sup>, Xu Liao<sup>1</sup>, Zhixiang Lin<sup>2</sup>, Xiaobo Sun<sup>3\*</sup>,

and Jin Liu<sup>1\*</sup>

<sup>1</sup>School of Data Science, The Chinese University of Hong Kong, Shenzhen, China.

<sup>2</sup>Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.

<sup>3</sup>Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, USA.

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding author. Email: xiaobo.sun@emory.edu,  
liujinlab@cuhk.edu.cn

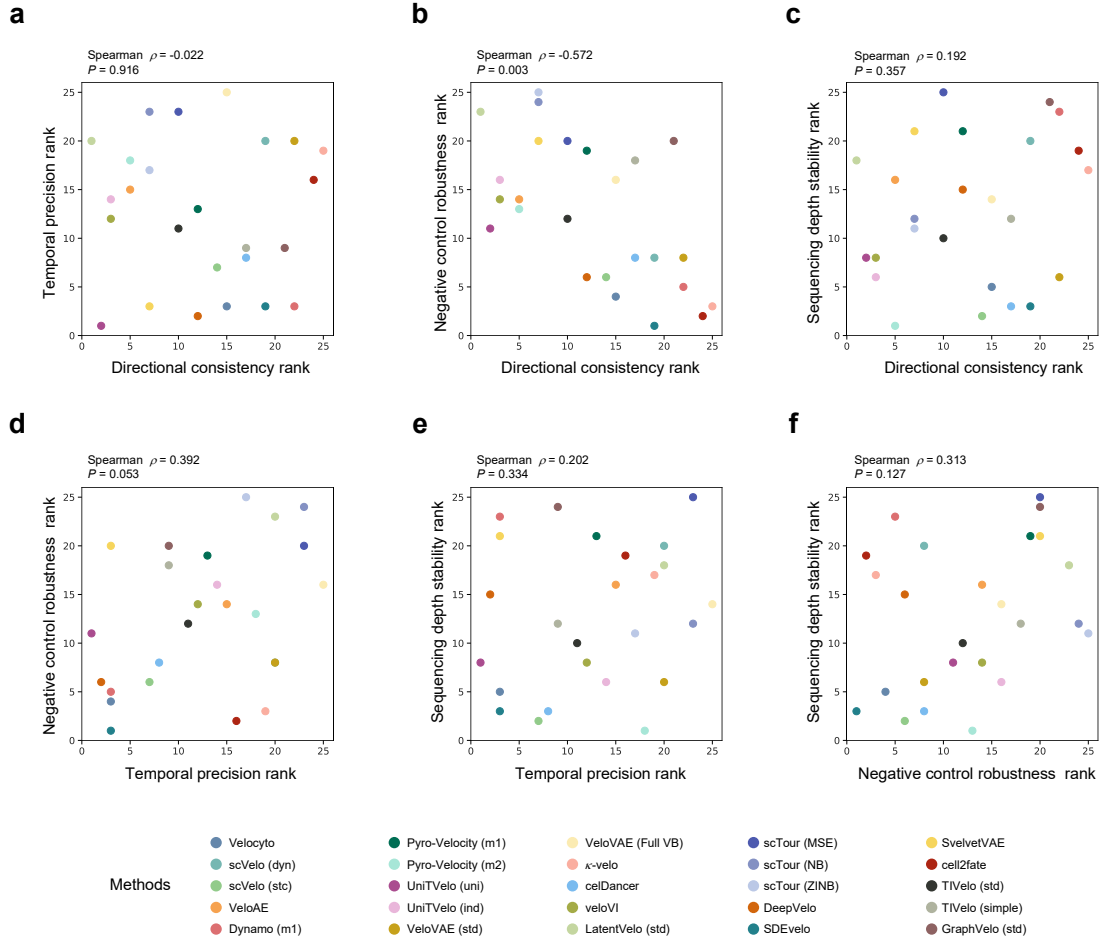


**Fig. S1 Pairwise correlation analysis of four core tasks.** Scatter plots showing the relationships between method rankings across the four benchmarking tasks: directional consistency, temporal precision, negative control robustness, and sequencing depth stability. Each point represents one RNA velocity method listed in the legend. Spearman's  $\rho$  and the associated  $P$ -value for the two rankings are shown above each panel. **a** Temporal precision versus directional consistency rankings. **b** Negative control robustness versus directional consistency rankings. **c** Sequencing depth stability versus directional consistency rankings. **d** Negative control robustness versus temporal precision rankings. **e** Sequencing depth stability versus temporal precision rankings. **f** Sequencing depth stability versus negative control robustness rankings.

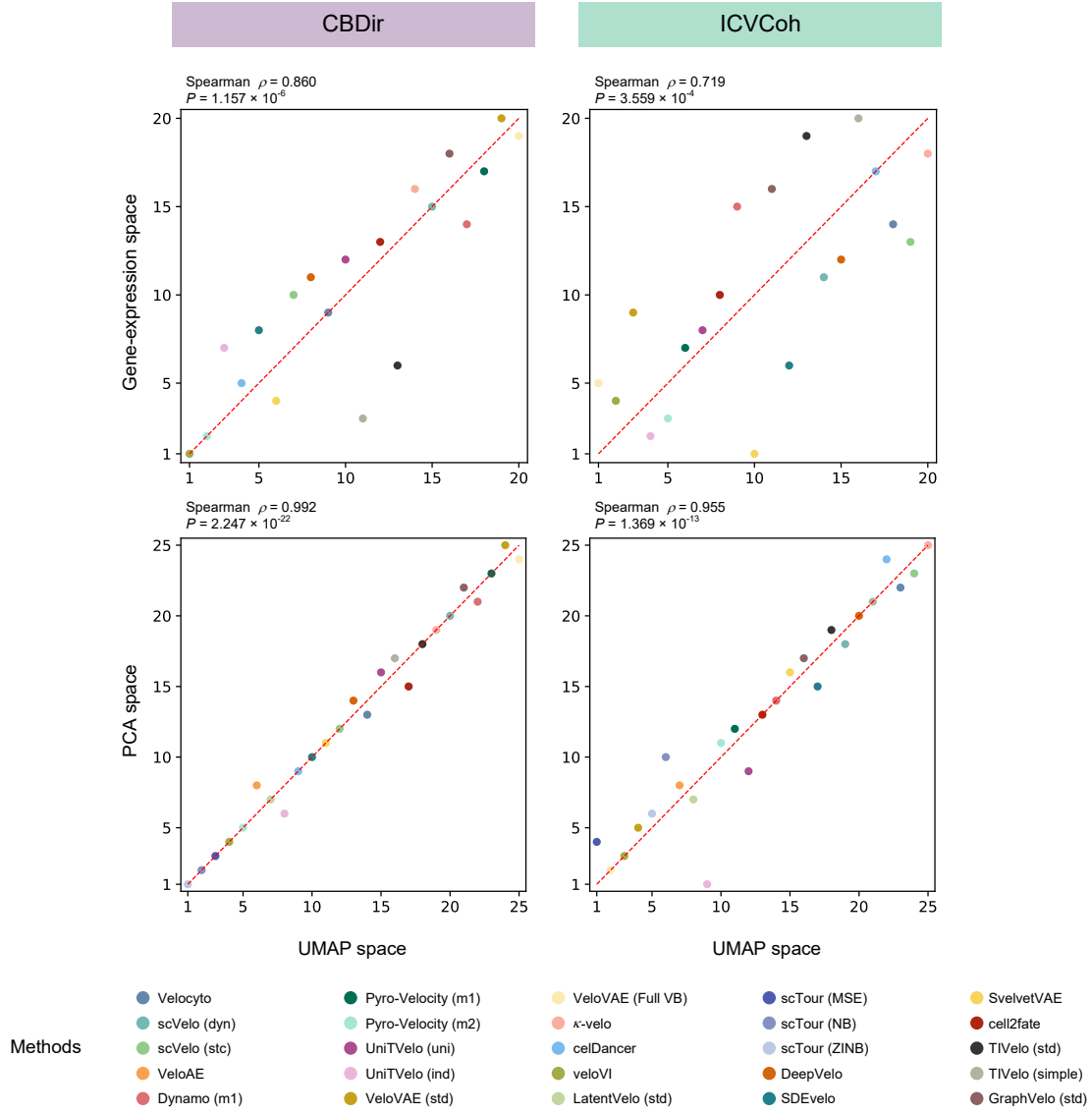


**Fig. S2 Sensitivity of directional consistency metrics to the choice of evaluation space.** Comparison of cross-boundary directional correctness (CBBDir) and in-cluster velocity coherence (ICVCoh) computed in different representation spaces using Data 3, with CBBDir shown in the left panels and ICVCoh shown in the right panels. The top panels compare results obtained in gene-expression space and UMAP space for 20 RNA velocity methods that output gene-space velocities. The bottom panels compare results obtained in PCA space and UMAP space across all evaluated RNA velocity methods. Each point represents one RNA velocity method listed in the legend. Spearman's  $\rho$  and the associated  $P$ -value are shown in each panel.

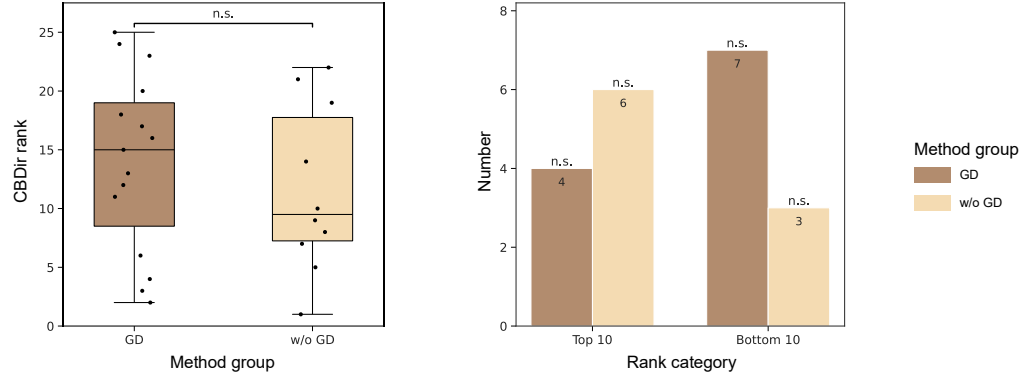
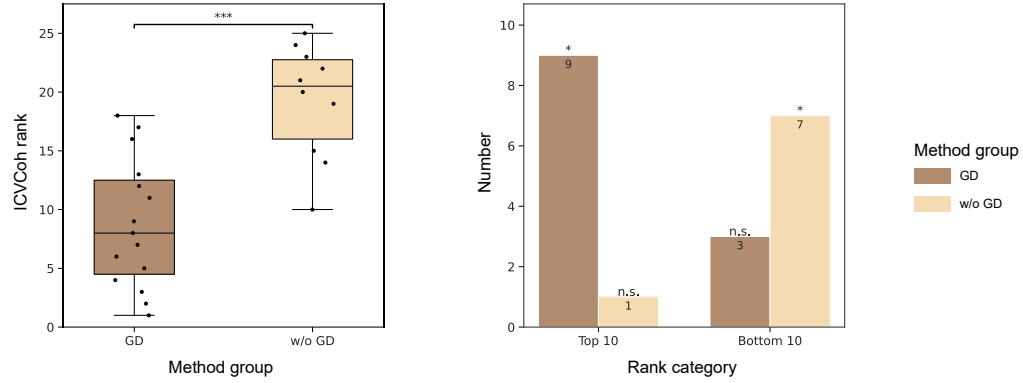
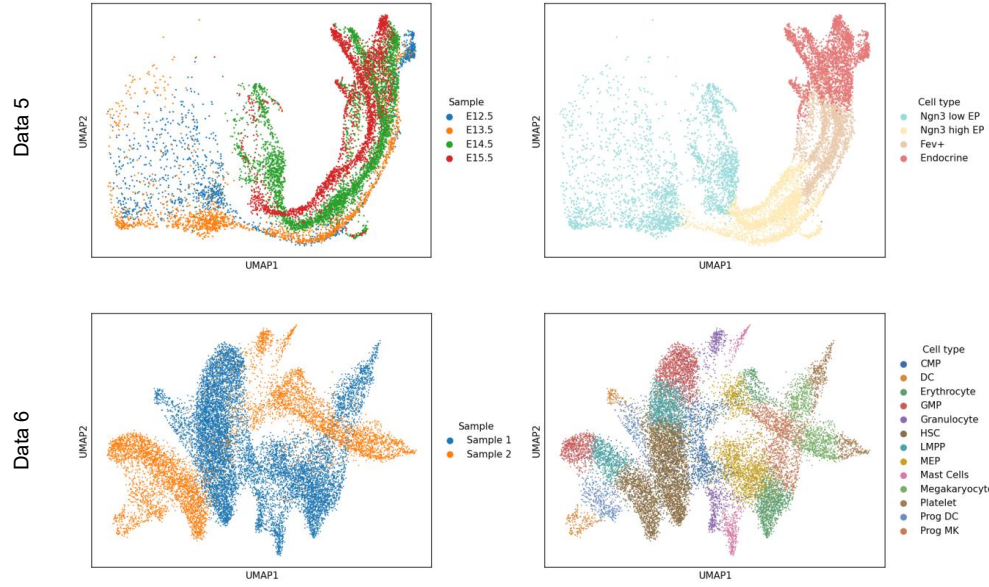
**a****b**

**Fig. S3 Impact of gene dependence on performance rankings.** **a** Left: Comparison of CBBir rankings shows no significant difference (n.s.) between gene dependence (GD) methods and those without gene dependence (w/o GD) (two-sided Wilcoxon rank-sum test). Right: Neither category is significantly enriched in the top 10 or bottom 10. **b** Left: Comparison of ICVCoh rankings indicates that methods utilizing gene dependence achieve significantly superior rankings (two-sided Wilcoxon rank-sum test, \*\*\* :  $P < 0.001$ ). Right: Methods with GD are enriched in the top 10 (\*\*\* :  $P < 0.001$ ), whereas methods without GD are enriched in the bottom 10 (\* :  $P < 0.05$ ).

**a**



**b**

|        | Method           | Bio conservation |            |            |                  |       | Batch correction |       |      |                               |      | Aggregate score  |                  |       |
|--------|------------------|------------------|------------|------------|------------------|-------|------------------|-------|------|-------------------------------|------|------------------|------------------|-------|
|        |                  | Isolated labels  | KMeans NMI | KMeans ARI | Silhouette label | cLISI | BRAS             | iLISI | KBET | Graph connectivity comparison | PCR  | Batch correction | Bio conservation | Total |
| Data 5 | Harmony          | 0.63             | 0.72       | 0.72       | 0.64             | 1.00  | 0.94             | 0.45  | 0.63 | 0.97                          | 0.33 | 0.66             | 0.74             | 0.71  |
|        | scTour (NB)      | 0.71             | 0.75       | 0.78       | 0.72             | 1.00  | 0.84             | 0.46  | 0.52 | 0.98                          | 0.00 | 0.56             | 0.79             | 0.70  |
|        | scTour (ZINB)    | 0.71             | 0.74       | 0.76       | 0.72             | 1.00  | 0.84             | 0.43  | 0.47 | 0.98                          | 0.00 | 0.54             | 0.79             | 0.69  |
|        | scTour (MSE)     | 0.71             | 0.71       | 0.69       | 0.72             | 1.00  | 0.81             | 0.38  | 0.34 | 0.98                          | 0.00 | 0.50             | 0.77             | 0.66  |
|        | LatentVelo (std) | 0.70             | 0.76       | 0.77       | 0.70             | 1.00  | 0.73             | 0.05  | 0.05 | 0.99                          | 0.00 | 0.36             | 0.79             | 0.62  |
|        | PCA              | 0.63             | 0.73       | 0.72       | 0.63             | 1.00  | 0.82             | 0.07  | 0.03 | 0.98                          | 0.00 | 0.38             | 0.74             | 0.60  |
| Data 6 | Harmony          | 0.54             | 0.51       | 0.30       | 0.53             | 0.98  | 0.54             | 0.00  | 0.15 | 0.91                          | 0.40 | 0.40             | 0.57             | 0.50  |
|        | LatentVelo (std) | 0.49             | 0.26       | 0.12       | 0.49             | 0.85  | 0.79             | 0.31  | 0.20 | 0.64                          | 0.86 | 0.56             | 0.44             | 0.49  |
|        | scTour (ZINB)    | 0.50             | 0.46       | 0.26       | 0.49             | 0.94  | 0.32             | 0.00  | 0.00 | 0.78                          | 0.22 | 0.27             | 0.53             | 0.43  |
|        | PCA              | 0.53             | 0.49       | 0.28       | 0.52             | 0.98  | 0.36             | 0.00  | 0.00 | 0.62                          | 0.00 | 0.20             | 0.56             | 0.41  |
|        | scTour (NB)      | 0.47             | 0.46       | 0.28       | 0.46             | 0.93  | 0.14             | 0.00  | 0.00 | 0.62                          | 0.00 | 0.15             | 0.52             | 0.37  |
|        | scTour (MSE)     | 0.46             | 0.42       | 0.23       | 0.44             | 0.94  | 0.14             | 0.00  | 0.00 | 0.61                          | 0.00 | 0.15             | 0.50             | 0.36  |

**Fig. S4 Visualization and evaluation of batch-confounded datasets.** **a** UMAP embeddings of the batch-confounded datasets, Data 5 and Data 6, colored by sample and cell type. **b** Evaluation of biological conservation and batch-correction performance for Harmony, LatentVelo (std), scTour (MSE/NB/ZINB), and PCA using **scib-metrics** on Data 5 (top panel) and Data 6 (bottom panel).

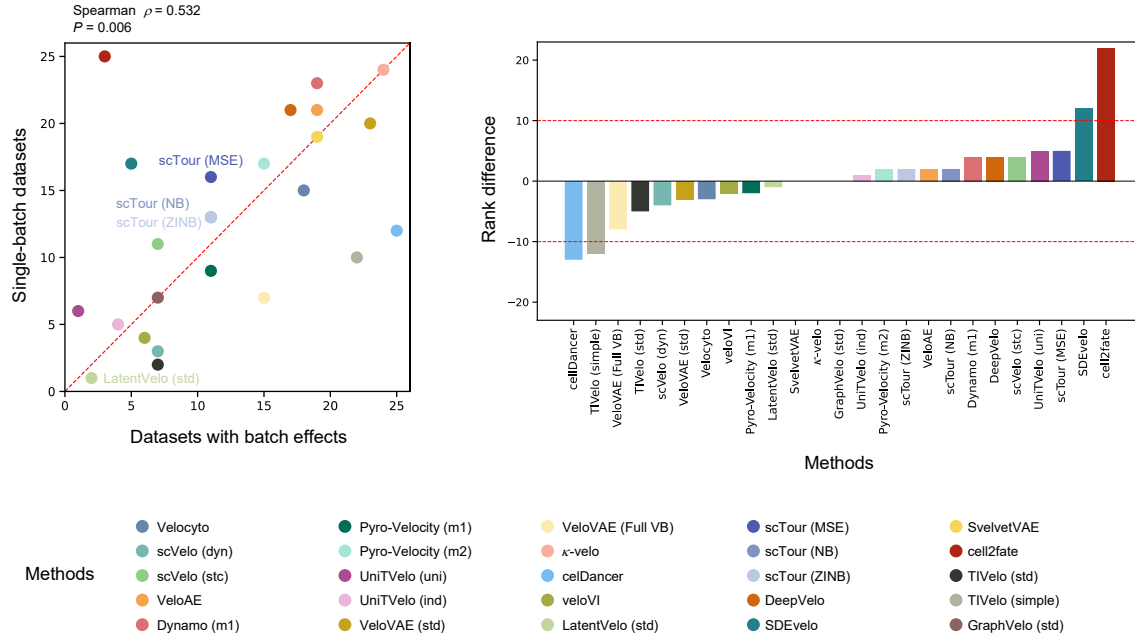


**Fig. S5 Comparison of method rankings between the single-batch subsets and the original batch-confounded datasets, together with rank differences across the two settings.** Left: Scatter plot showing the relationship between method rankings obtained on the single-batch subsets and those obtained on the original batch-confounded dataset. Each point represents one RNA velocity method listed in the legend. Spearman’s  $\rho$  and the corresponding  $P$ -value comparing the two rankings are shown above the panel. Right: Rank differences for each method between the single-batch subsets and the original batch-confounded dataset. The horizontal dashed lines at  $y = 10$  and  $y = -10$  indicate that a method ranks 10 positions higher and 10 positions lower, respectively, on the original batch-confounded datasets than on the corresponding single-batch subsets.

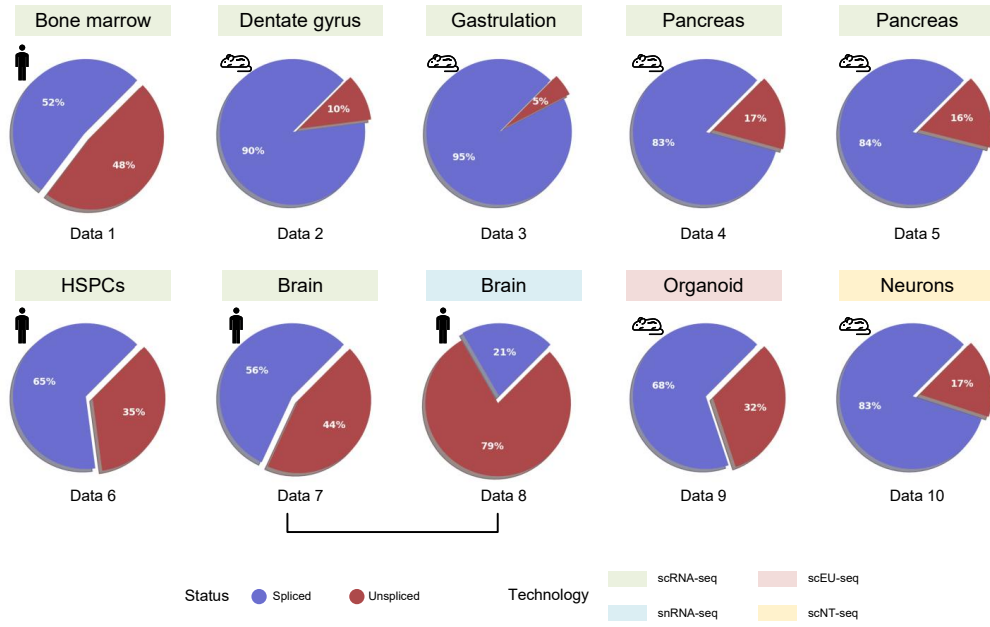


**Fig. S6 Comparison of spliced and unspliced transcript proportions across datasets used in Task I.** The proportions of spliced and unspliced transcripts are shown for the ten datasets used in Task I, with annotations indicating the tissue, species (human or mouse), and sequencing technology (scRNA-seq, snRNA-seq, scEU-seq, and scNT-seq) of each dataset. Data 7 and Data 8 further enable a direct comparison of transcript composition between scRNA-seq and snRNA-seq datasets from the same species and tissue type.

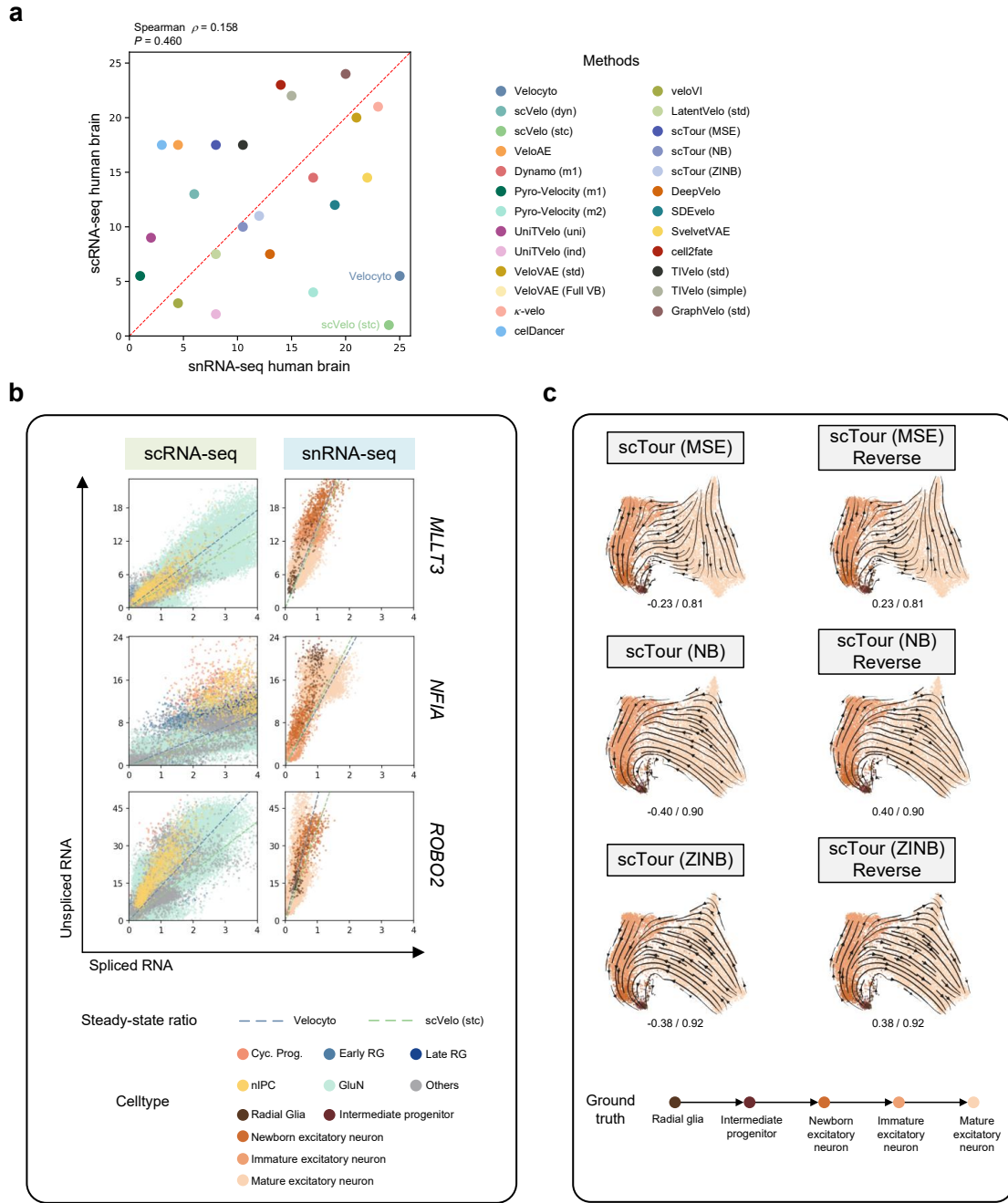


**Fig. S7 Evaluation of RNA velocity methods on snRNA-seq data.** **a** Directional consistency ranking comparison between matched scRNA-seq and snRNA-seq human brain datasets. Each point represents one RNA velocity method listed in the legend. Spearman's  $\rho$  and the corresponding  $P$ -value comparing the two rankings are shown above the panel. **b** The spliced-unspliced relationships of *MLLT3*, *NFIA*, and *ROBO2* in the scRNA-seq and snRNA-seq human brain datasets. For each gene, the  $x$ - and  $y$ -axis tick marks are kept consistent between the two datasets to facilitate direct comparison. The dashed lines indicate the steady-state ratios estimated by Velocyto and scVelo (stc), respectively. **c** Streamline visualizations of the three scTour variants on Data 8 (snRNA-seq), using both original and reversed velocity vectors, with the reference cell-state transitions shown below. CDBir (left) and ICVCoh (right) values computed using the full dataset are shown below each streamline plot.

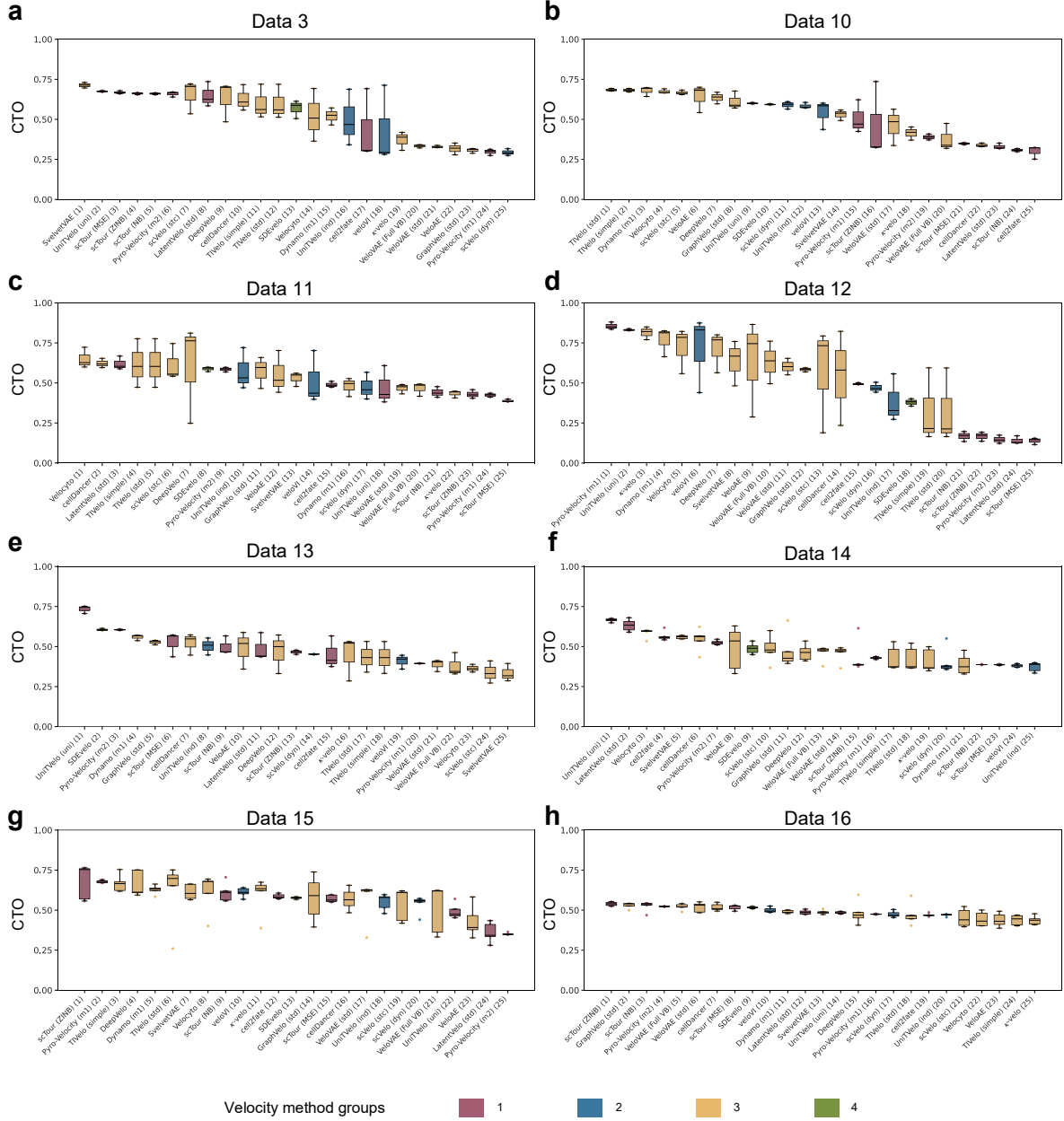


**Fig. S8 Benchmarking RNA velocity methods based on cluster temporal ordering accuracy.** **a–h** Box plots showing cluster temporal ordering accuracy (CTO) scores for RNA velocity methods across eight distinct datasets: Data 3 (**a**), Data 10 (**b**), Data 11 (**c**), Data 12 (**d**), Data 13 (**e**), Data 14 (**f**), Data 15 (**g**), and Data 16 (**h**). Methods are arranged along the  $x$ -axis according to their rankings (indicated in parentheses), and boxes are color-coded by velocity method group as shown in the legend. Individual points represent scores for specific cross-validation folds in each dataset.

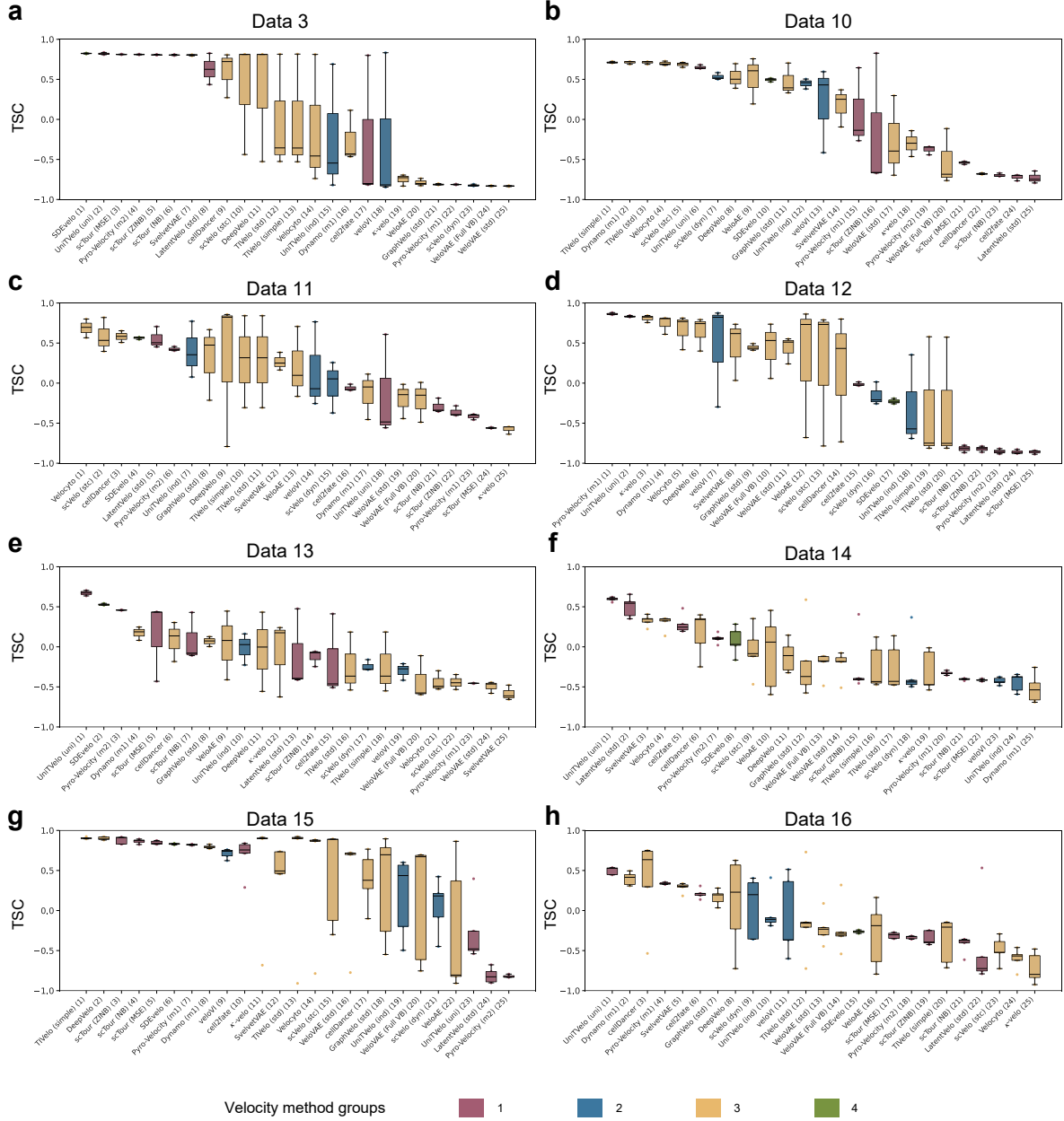


**Fig. S9 Benchmarking RNA velocity methods based on temporal Spearman correlation.** a–h Box plots showing the temporal Spearman correlation (TSC) scores for various RNA velocity methods across eight distinct datasets: Data 3 (a), Data 10 (b), Data 11 (c), Data 12 (d), Data 13 (e), Data 14 (f), Data 15 (g), and Data 16 (h). Methods are arranged along the  $x$ -axis according to their rankings (indicated in parentheses), and boxes are color-coded by velocity method group as shown in the legend. Individual points represent scores for specific cross-validation folds in each dataset.

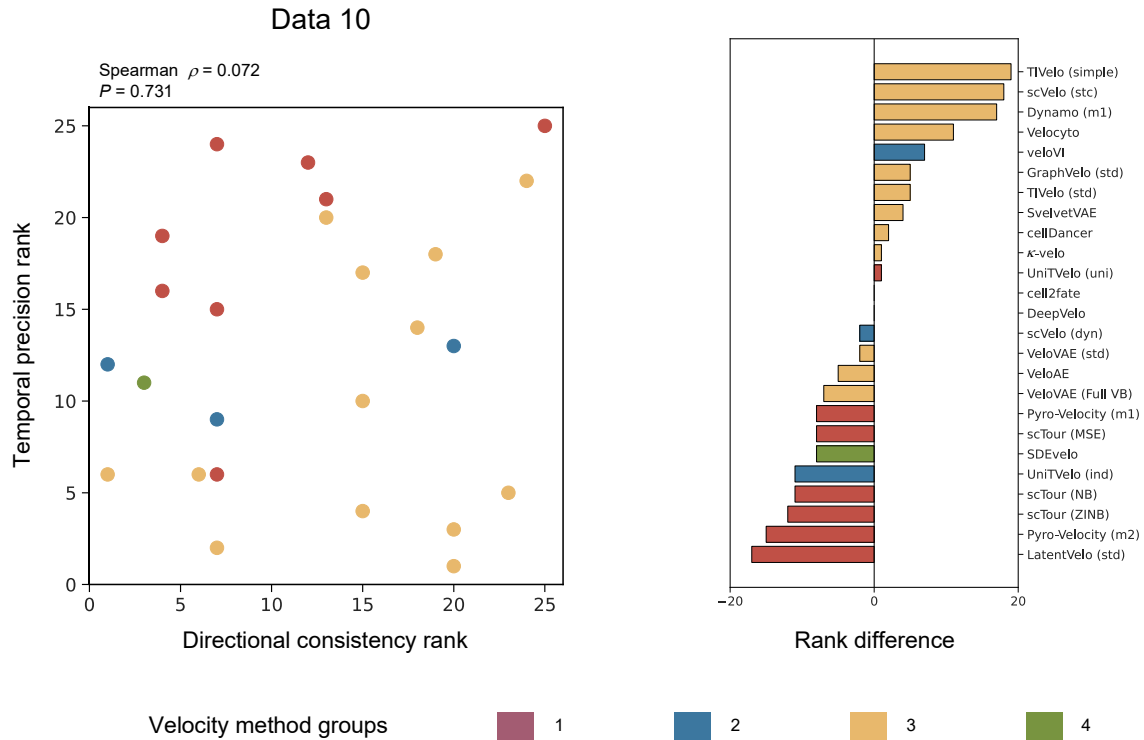


**Fig. S10 Comparison between temporal precision and directional consistency on Data 10.** Left: Scatter plot comparing directional consistency rank and temporal precision rank on Data 10. Each point represents one RNA velocity method and is colored according to its velocity method group. Spearman's  $\rho$  and the associated  $P$ -value between the two rankings are shown above the panel. Right: Rank differences between directional consistency and temporal precision for each RNA velocity method. Positive rank differences indicate methods with better temporal precision rankings than directional consistency rankings, whereas negative values indicate the opposite pattern.

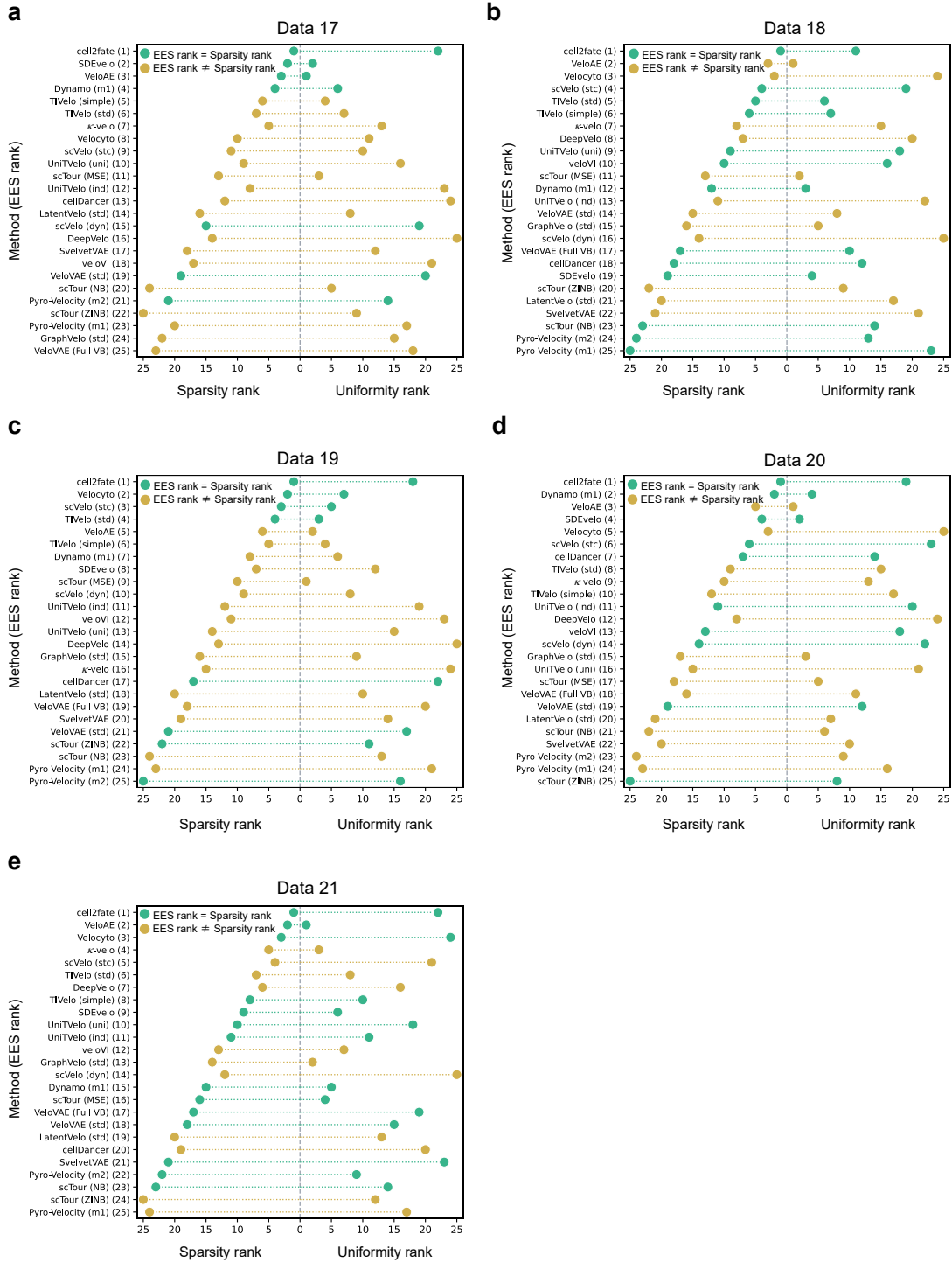


**Fig. S11 Relationship between EES rankings and transition sparsity or uniformity rankings.** Scatter plots comparing EES-based method rankings with transition sparsity and uniformity rankings across five PBMC negative-control datasets. The  $y$ -axis shows RNA velocity methods ordered by ascending EES rank, with the EES rank indicated in parentheses; the  $x$ -axis shows the rank of the mean effective number of retained transition probabilities,  $m_c^{(0.95)}$ , on the left and the rank of  $\text{EES}/m_c^{(0.95)}$  on the right. Panels **a–e** correspond to Data 17–21, respectively.

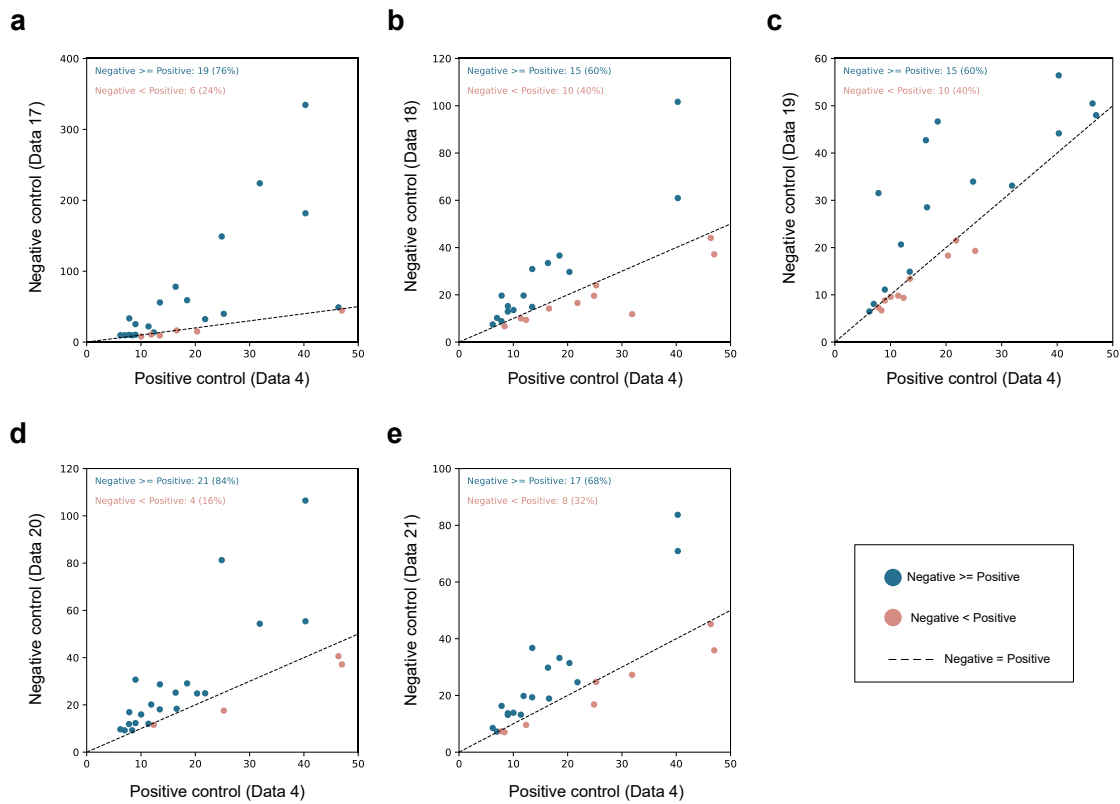


**Fig. S12 Comparison of EES scores between negative-control and positive-control datasets.** EES scores are compared between each PBMC negative-control dataset and the positive-control dataset Data 4 across RNA velocity methods. Panels **a–e** correspond to the negative-control datasets Data 17–21, respectively. Points above the diagonal indicate methods with higher EES scores on negative-control data than on the positive-control dataset.

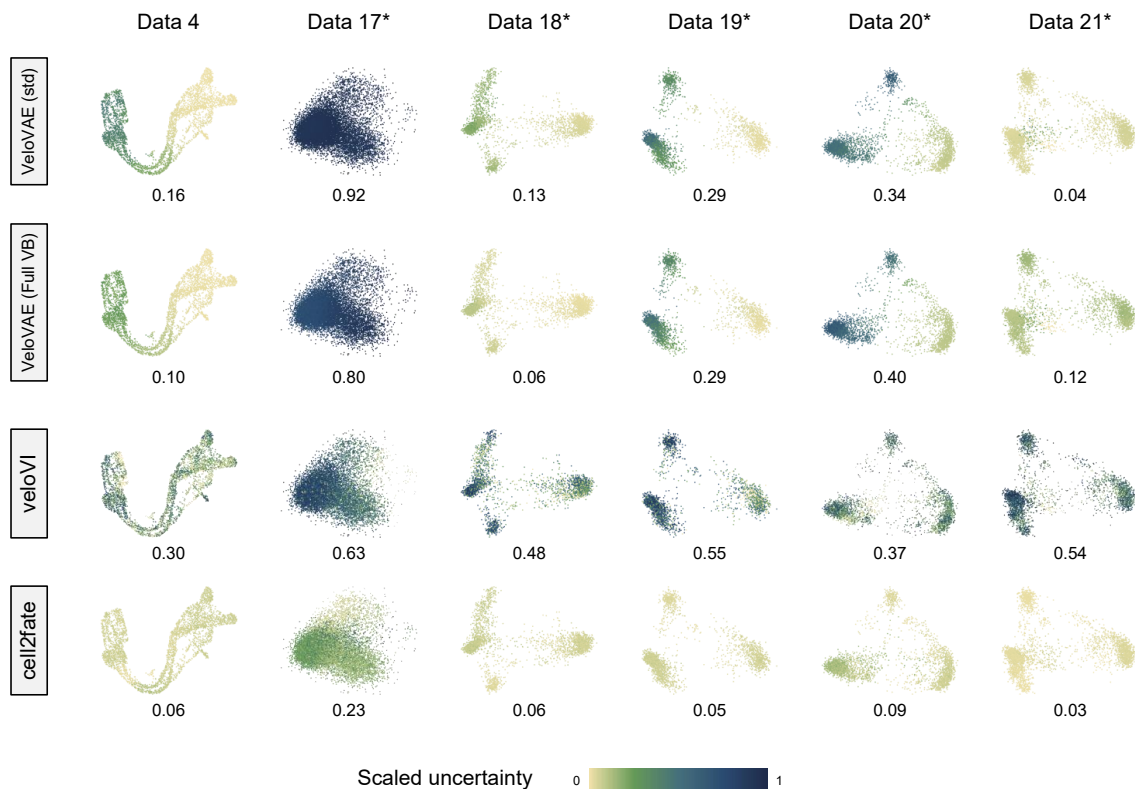


**Fig. S13 Visualization of uncertainty scores across RNA velocity methods.** Low-dimensional embeddings colored by uncertainty scores for VeloVAE (std), VeloVAE (Full VB), veloVI, and cell2fate. Each row corresponds to a specific method. Uncertainty scores are min-max scaled independently for each method across six datasets (Data 4 and Data 17–21). The first column shows Data 4 (dynamic reference), while the remaining columns represent peripheral blood mononuclear cell (PBMC) datasets. The superscript asterisk (\*) indicates that only one cross-validation fold was used for the dataset.

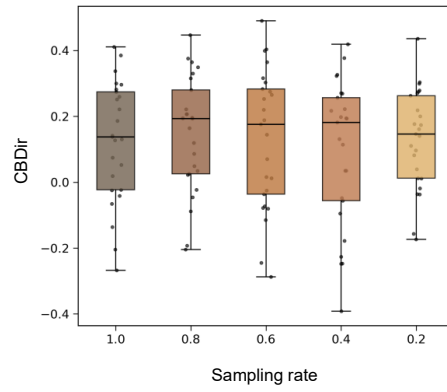
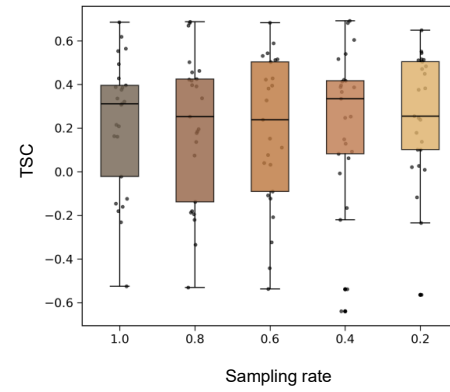
**a****b**

**Fig. S14 Robustness to sequencing depth variations.** **a** Box plots showing CDBir scores across sampling rates ranging from 1.0 to 0.2 on Data 2, 3, and 9. Each point represents the average CDBir value across the three datasets for a specific method. **b** Box plots showing TSC scores across sampling rates ranging from 1.0 to 0.2 on Data 3, 11, and 13. Each point represents the average TSC value across the three datasets for a specific method.

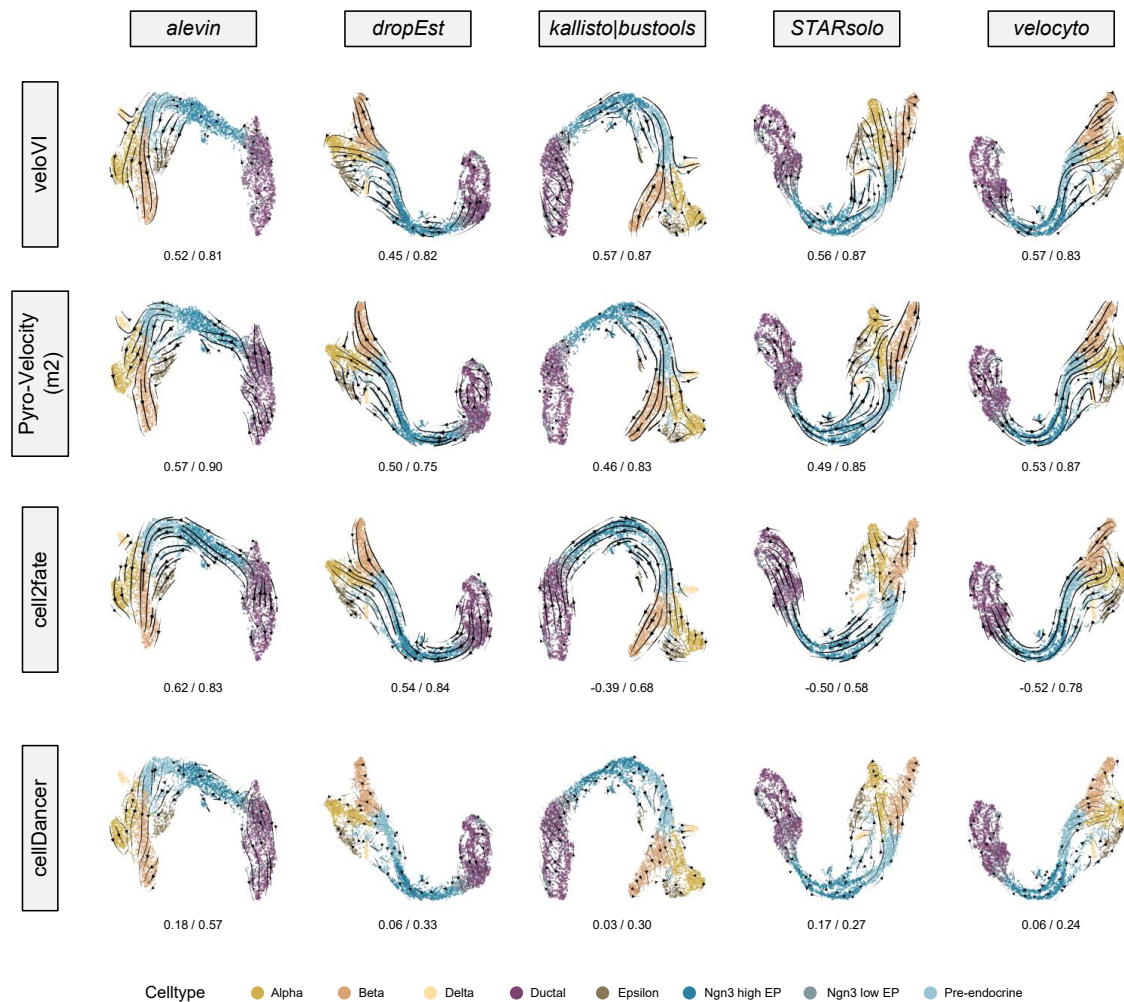
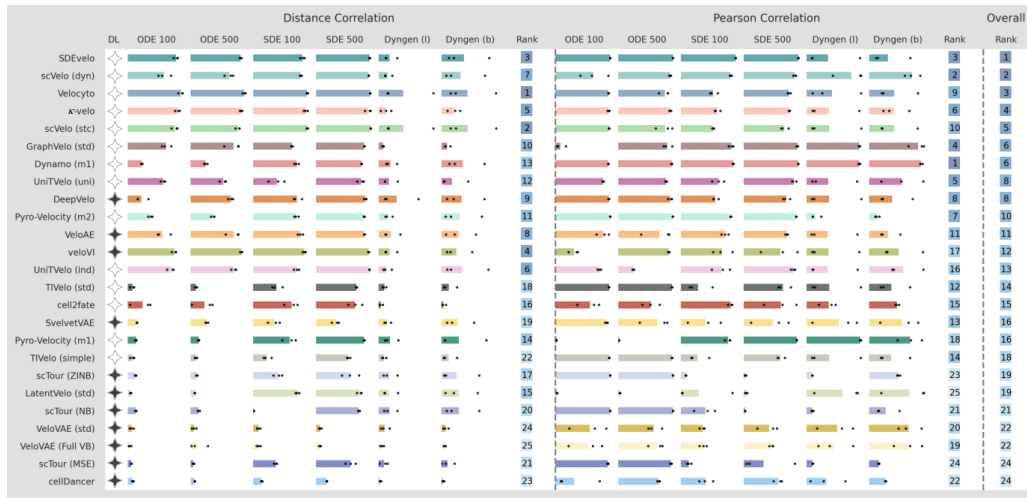
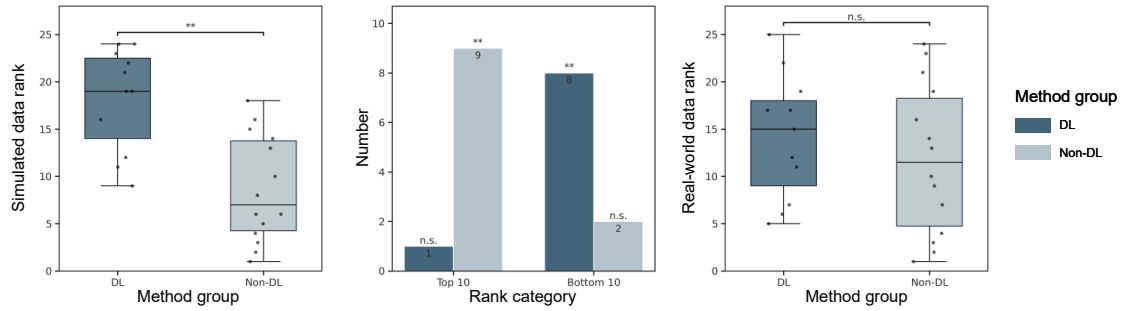


**Fig. S15 Robustness to quantification methods.** Low-dimensional embeddings colored by cell type for veloVI, Pyro-Velocity (m2), cell2fate, and cellDancer. Each row represents a specific RNA velocity method, and each column corresponds to a quantification algorithm. Values shown below each embedding indicate CDBir (left) and ICVCoh (right) scores.

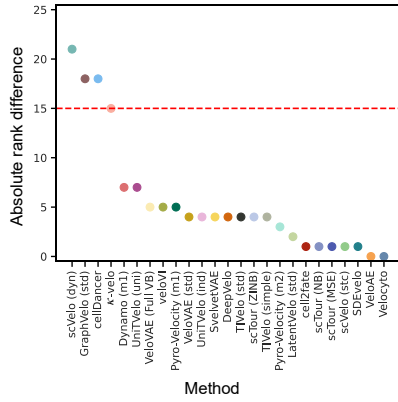
**a**



**b**



**c**



**d**

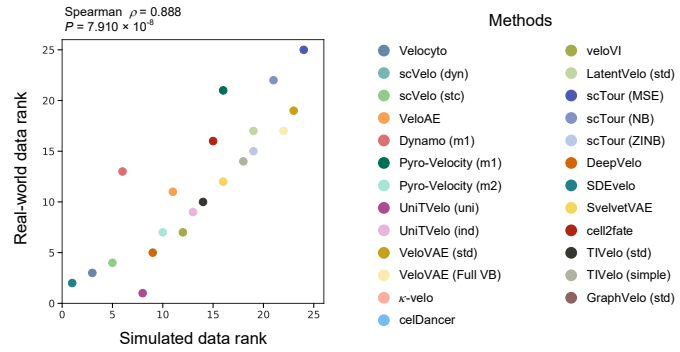
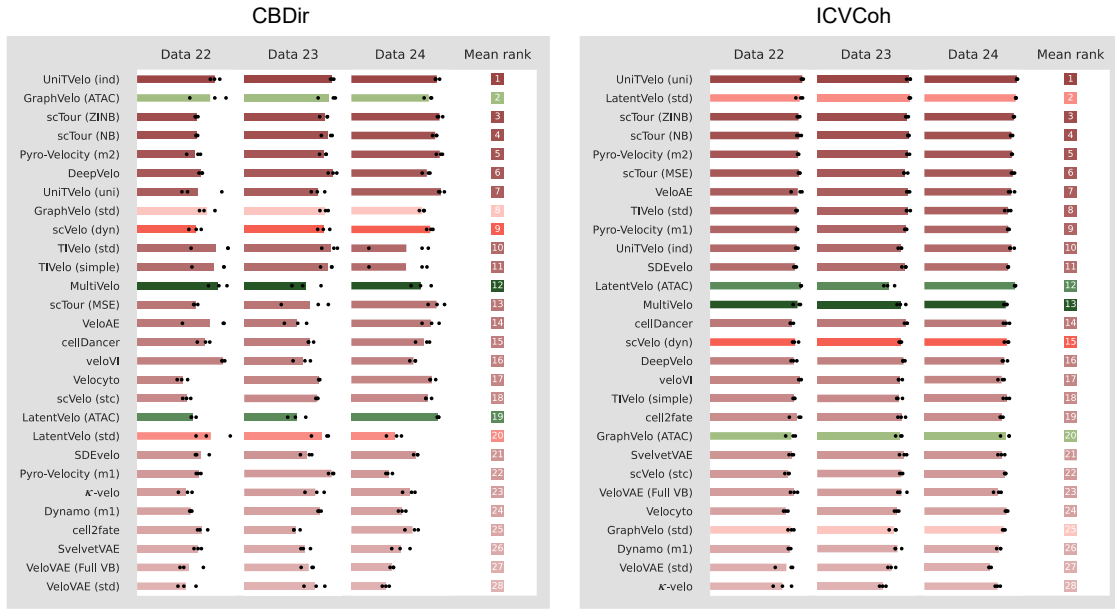


**Fig. S16 Performance evaluation on simulated datasets.** **a** Ranking summary of 25 RNA velocity methods based on distance correlation and Pearson correlation across six simulated datasets (two ODE-based, two SDE-based, and two dyngen-generated). Methods are ordered by mean metric values so that stronger performers appear in the top rows. The overall rank is calculated as the average of the distance correlation and Pearson correlation ranks. Individual points represent scores for specific cross-validation folds in each dataset. **b** Comparison of method categories. Left: Non-deep learning (DL) frameworks achieve significantly better rankings than DL methods on simulated data (two-sided Wilcoxon rank-sum test,  $** : P < 0.01$ ). Middle: Non-DL methods are significantly enriched in the top 10 ( $** : P < 0.01$ ), whereas DL methods are enriched in the bottom 10 ( $** : P < 0.01$ ). Right: Comparison of rankings on real-world data shows no significant difference (n.s.) between the groups. **c** Absolute rank differences between simulated and real-world data for each method. The red dotted line indicates a rank difference of 15. **d** Scatter plot comparing rankings on simulated data versus real-world data. Spearman’s  $\rho$  and the associated  $P$ -value for the two rankings are shown above the panel.

**a**



**b**

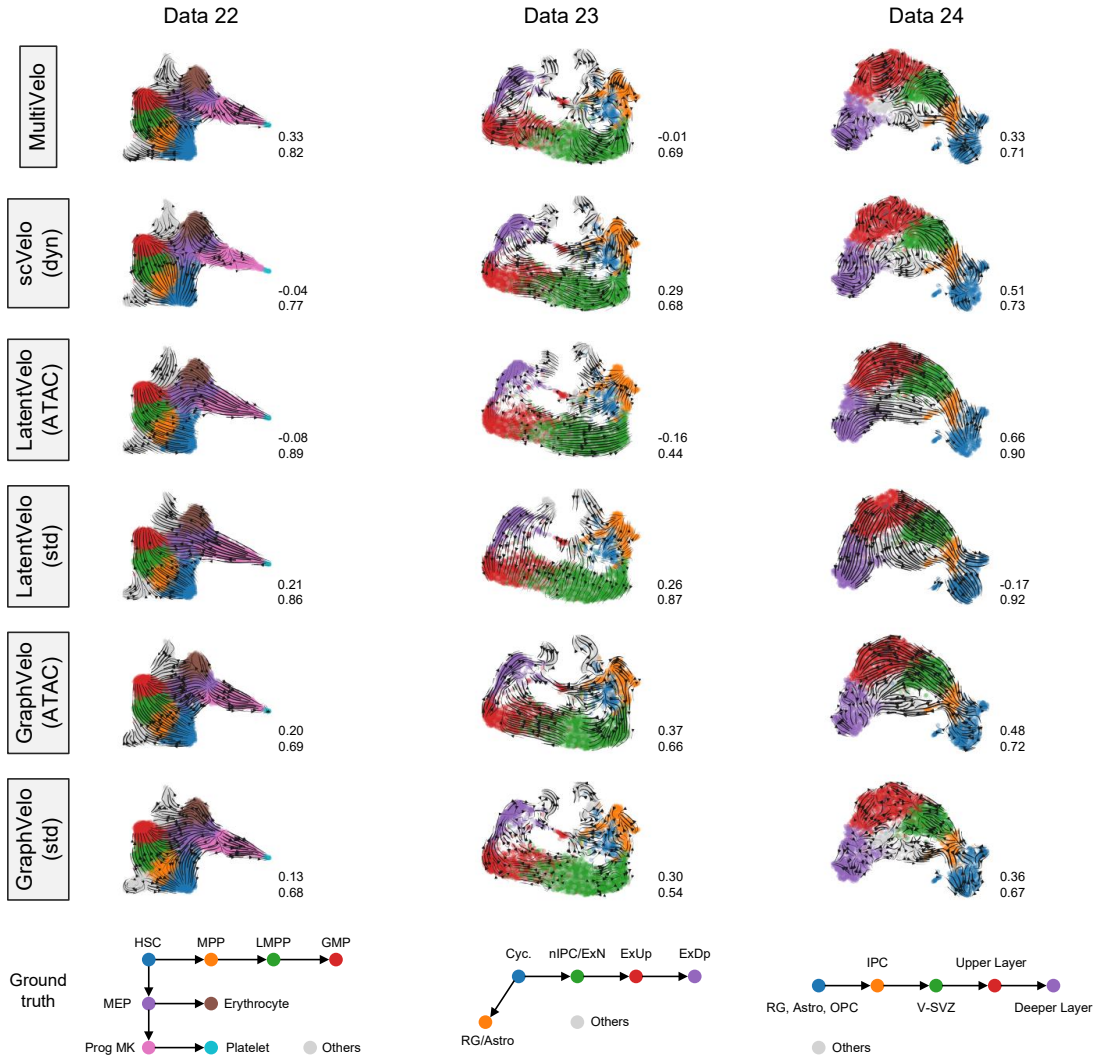
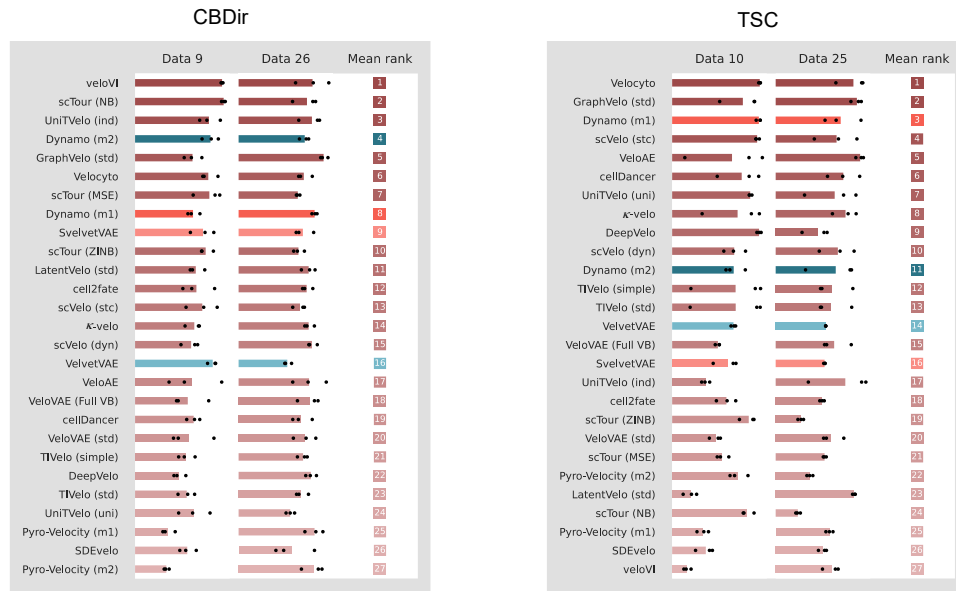
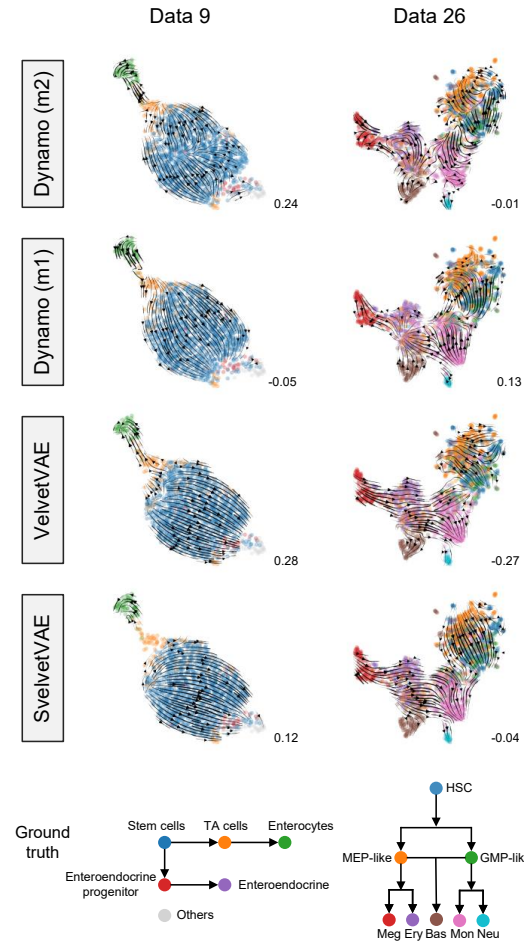


**Fig. S17 Benchmarking chromatin accessibility-enhanced RNA velocity methods against 25 RNA-only methods. a** Ranking summary of 28 RNA velocity methods based on CBBDir (left) and ICVCoh (right). Methods are ordered by their average rank across three datasets (Data 22, 23, 24), and longer bars indicate better ranking performance. Individual points represent scores for specific cross-validation folds in each dataset. The three chromatin accessibility-enhanced methods are highlighted with green bars. **b** Velocity streamlines projected onto UMAP embeddings for three chromatin accessibility-enhanced methods and their corresponding RNA-only baselines using Data 22, 23, and 24. Comparisons include MultiVelo versus scVelo (dyn), LatentVelo (ATAC) versus LatentVelo (std), and GraphVelo (ATAC) versus GraphVelo (std), with the reference cell-state transitions shown in the legend. Quantitative performance scores appear in the bottom-right corner of each plot, with the first score representing CBBDir and the second representing ICVCoh.

**a**



**b**



**c**

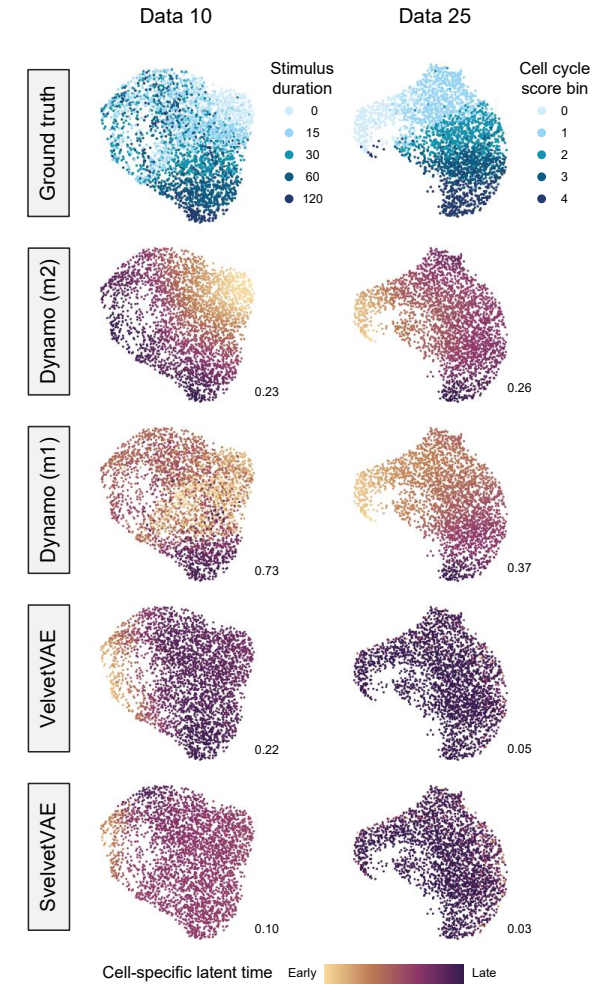
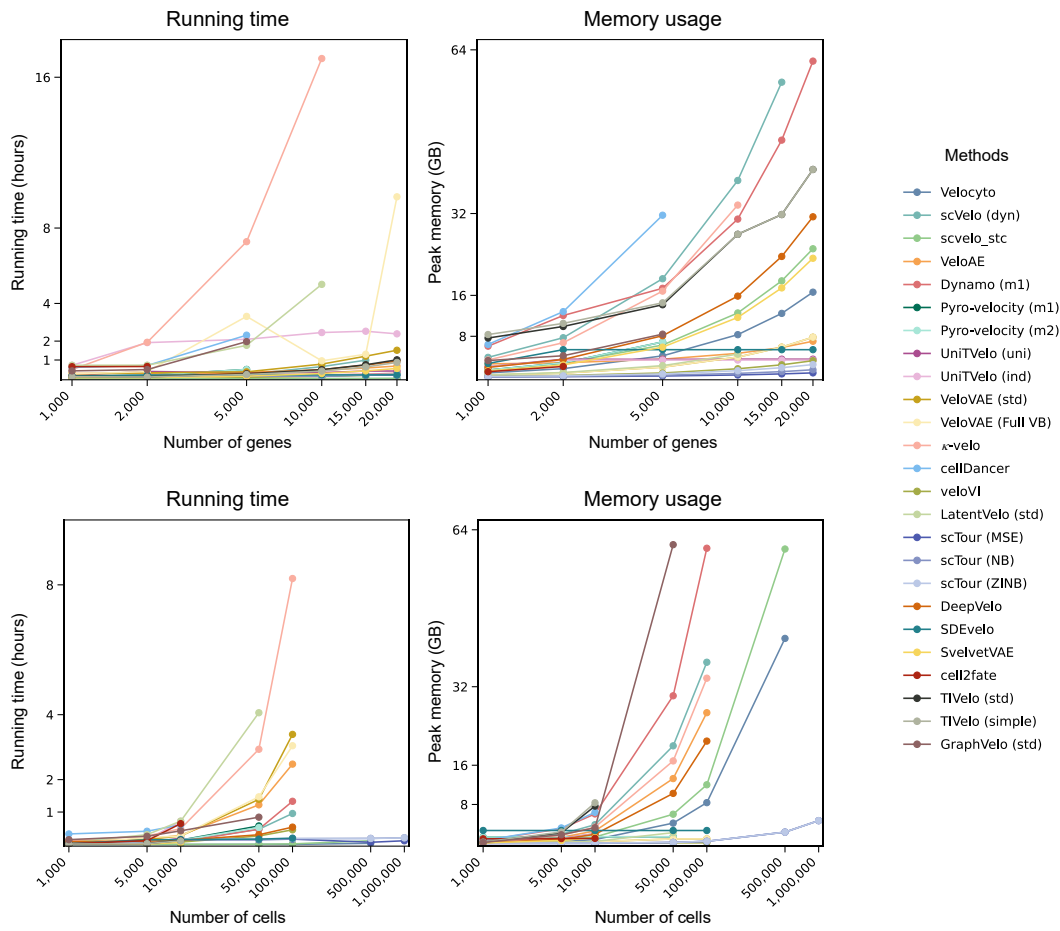


**Fig. S18 Benchmarking metabolic labeling–enhanced RNA velocity methods against 25 RNA-only methods.** **a** Ranking summary of 27 RNA velocity methods across selected datasets. Left: CDBir rankings for Data 9 and Data 26. Right: TSC rankings for Data 10 and Data 25. Methods are sorted by their average rank within each metric group, with longer bars indicating better ranking performance. Individual points represent scores for specific cross-validation folds in each dataset. The two metabolic labeling–enhanced methods are highlighted with blue bars. **b** Velocity streamlines projected onto UMAP embeddings for two metabolic labeling–enhanced methods and their corresponding RNA-only baselines using Data 9 and Data 26. Comparisons include Dynamo (m2) versus Dynamo (m1) and VelvetVAE versus SvelvetVAE, with the reference cell-state transitions shown in the legend. CDBir scores appear in the bottom-right corner of each plot. **c** The first row displays scatter plots in the projected UMAP space, with cells colored by labeling time for Data 10 and by cell cycle score bin for Data 25. Rows 2–5 show cell-specific latent time heatmaps for selected methods. Comparisons include Dynamo (m2) versus Dynamo (m1) and VelvetVAE versus SvelvetVAE. TSC scores appear in the bottom-right corner of each plot.

**a**



**b**

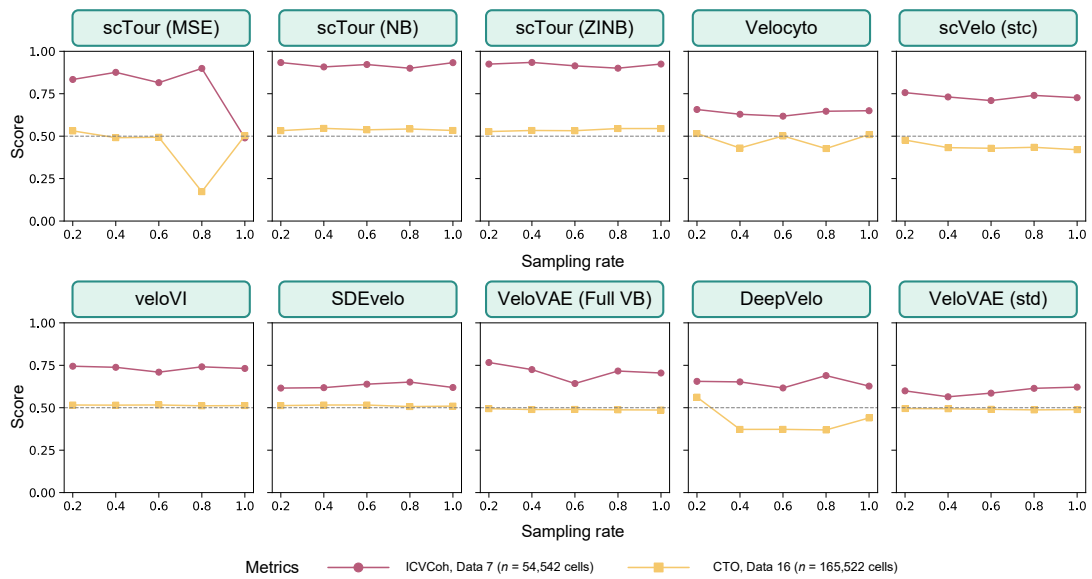


**Fig. S19 Computational scalability analysis of 25 RNA-only methods.** **a** Line plots showing the running time (hours, left) and peak memory usage (GB, right) for each method. Methods are color-coded as indicated in the legend. Top: Performance scaling with increasing number of genes, ranging from 1,000 to 20,000 genes with a fixed number of cells ( $n = 10,000$ ). Bottom: Performance scaling with increasing number of cells, ranging from 1,000 to 1,000,000 cells with a fixed number of genes ( $p = 1,000$ ). **b** Line plot of accuracy versus downsampling rates for the top 10 methods in terms of  $\text{Scalability}_{\text{cell}}$  using two real-world datasets. The red line indicates ICVCoh using Data 7, while the yellow line indicates TSC using Data 16. Sampling rates range from 0.2 to 1.0 (the original data). The dashed line at  $y = 0.5$  serves as a baseline.

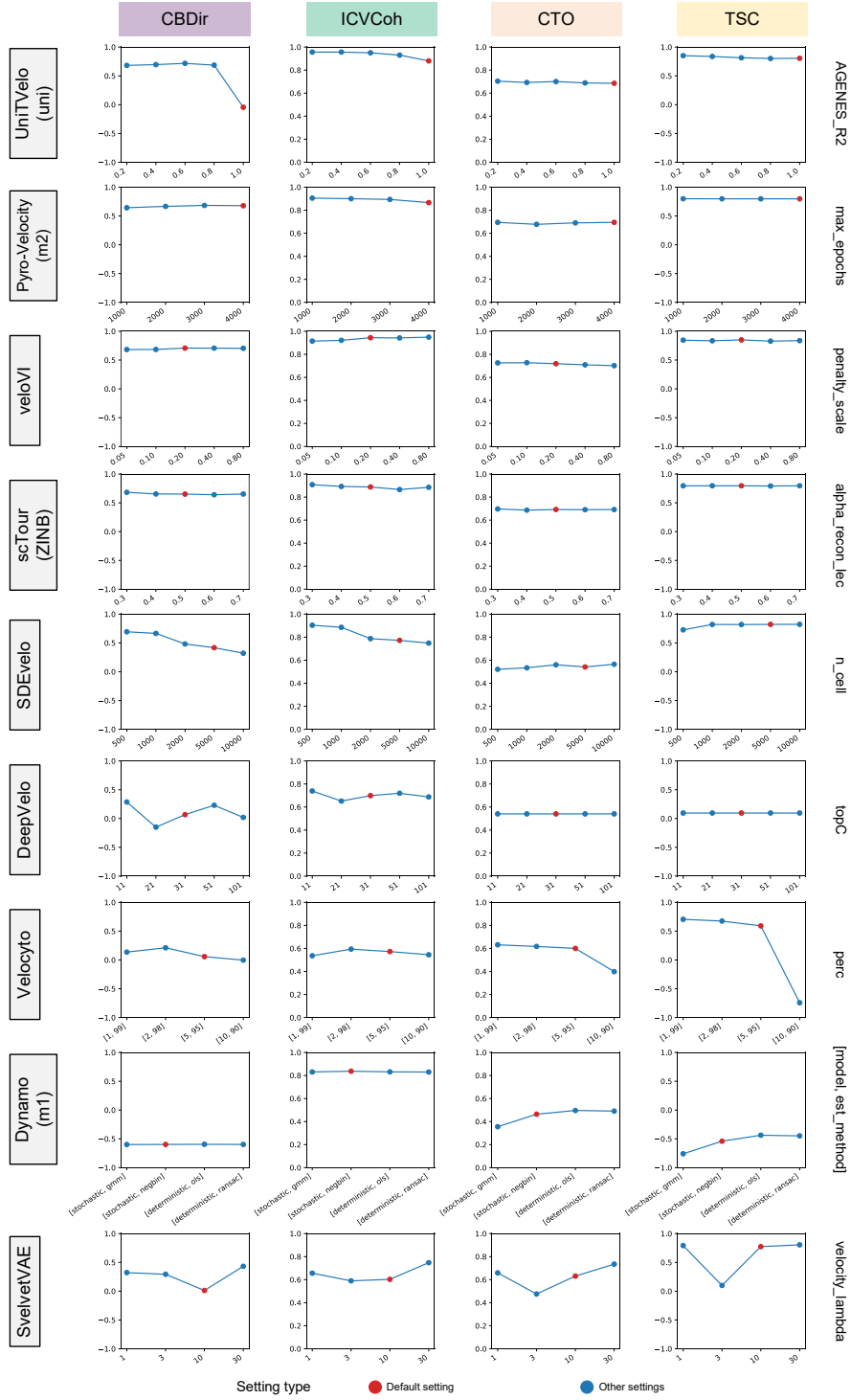


**Fig. S20 Hyperparameter sensitivity analysis for guideline-recommended RNA velocity methods across Task I and Task II.** The left labels show the RNA velocity methods included in the analysis, and the right labels show their corresponding hyperparameter settings. Each column represents one evaluation metric, ordered from left to right as CBDir, ICVCoh, CTO, and TSC. Red points indicate the default parameter settings, whereas blue points indicate alternative parameter settings.