

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

Yida Wu^{1†}, Chuihan Kong^{1†}, Xu Liao¹, Zhixiang Lin², Xiaobo Sun^{3*},

and Jin Liu^{1*}

¹School of Data Science, The Chinese University of Hong Kong-Shenzhen, Shenzhen, China.

²Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.

³Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, USA.

[†]These authors contributed equally to this work.

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

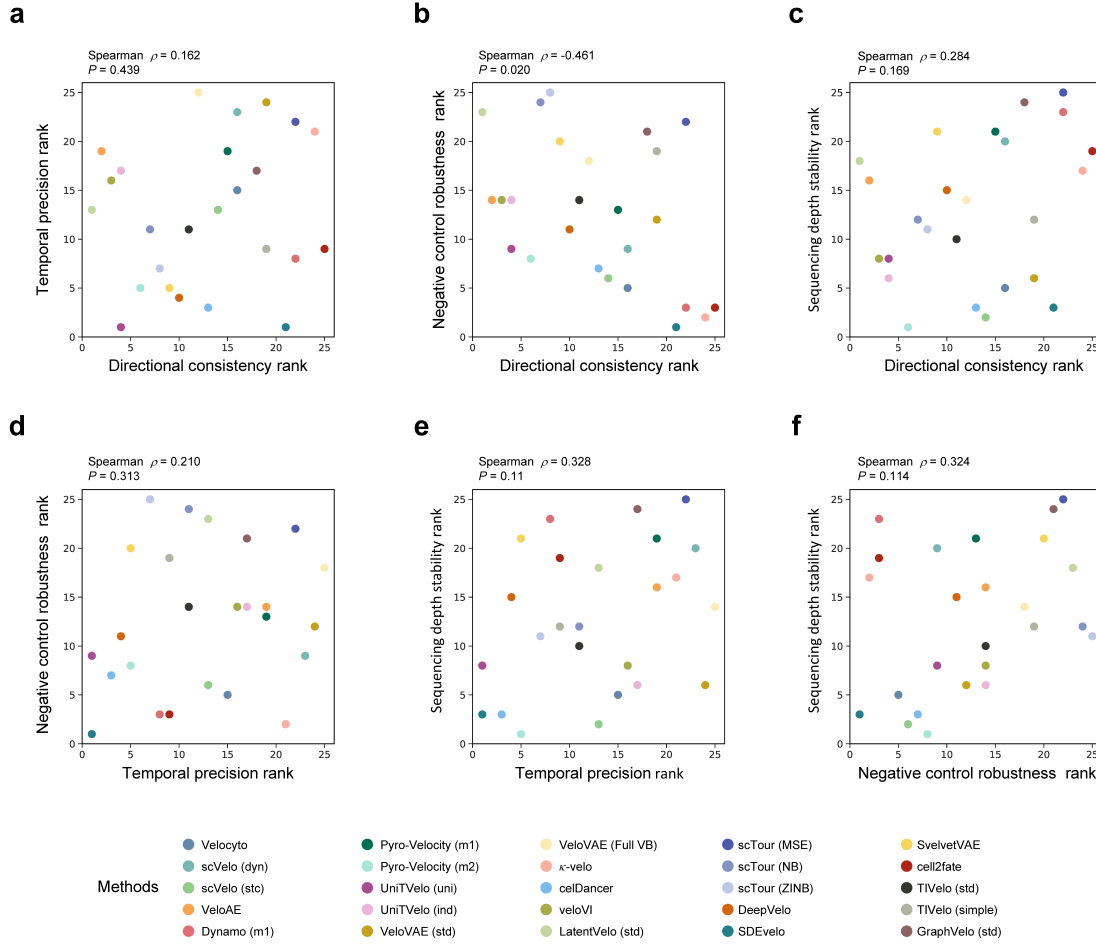


Figure S1: Pairwise correlation analysis of four core tasks. Scatter plots showing the relationships between method rankings across the four benchmarking tasks: directional consistency, temporal precision, negative control robustness, and sequencing depth stability. Each point represents one RNA velocity method listed in the legend. Spearman's ρ 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.

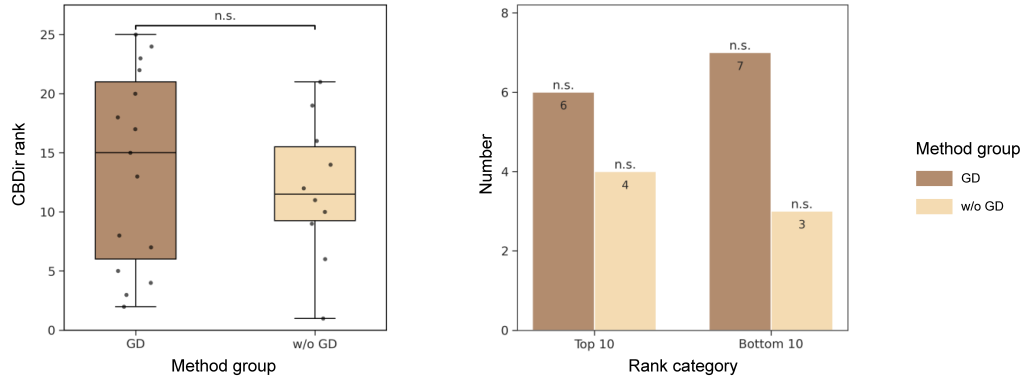
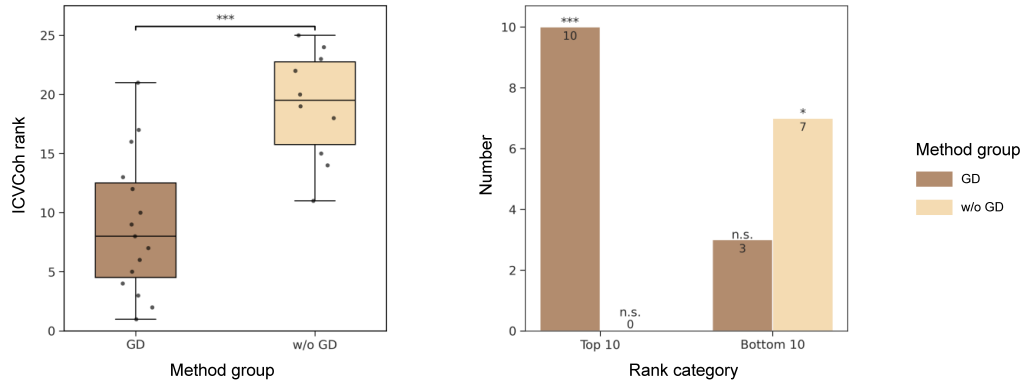
a**b**

Figure S2: **Impact of gene dependence on performance rankings.** **a** Left: Comparison of cross-boundary directional correctness (CBDir) rankings shows no significant difference (n.s.) between gene dependence (GD) methods and methods 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 in-cluster velocity coherence (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 (* : $P < 0.05$).

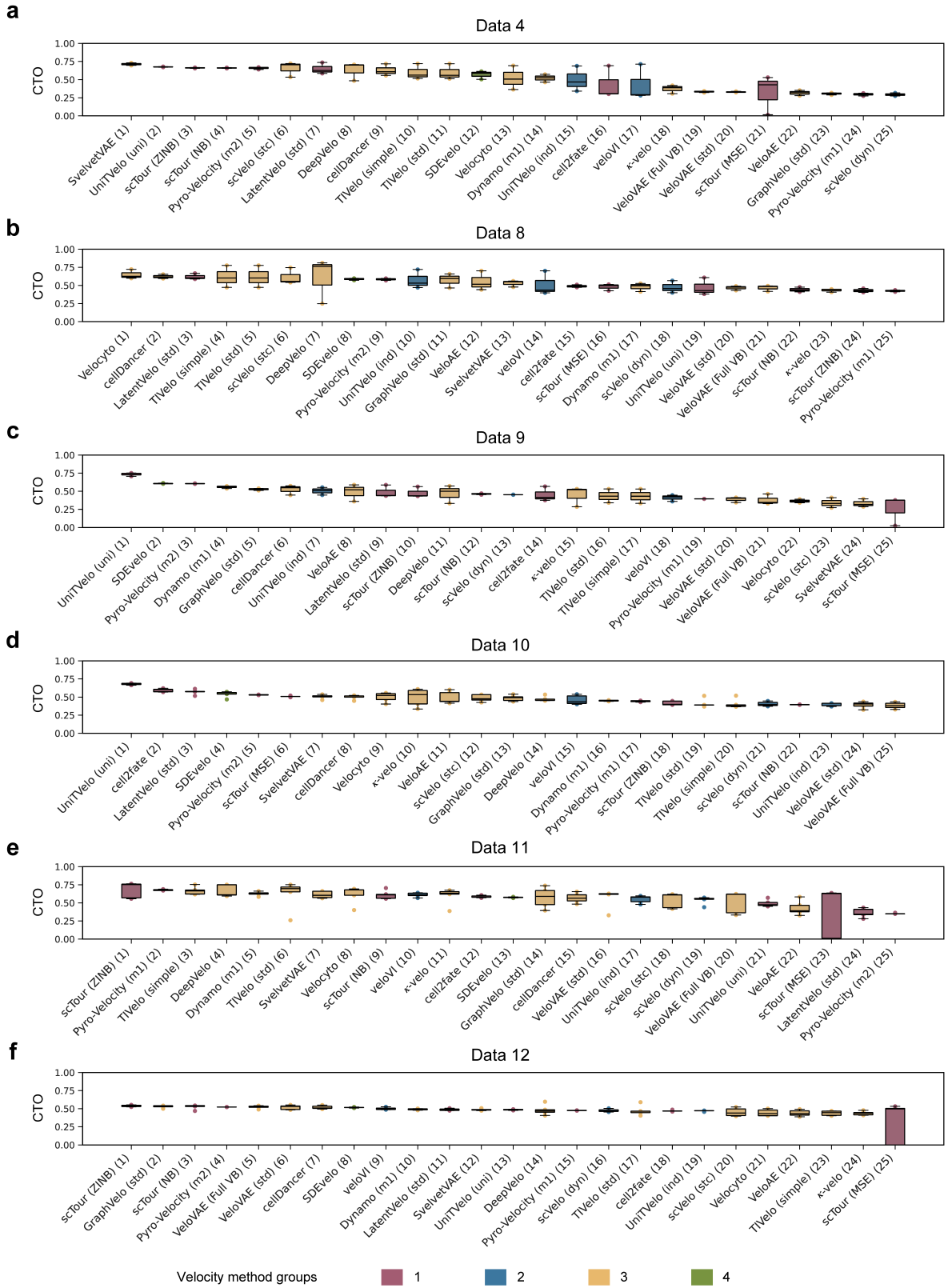


Figure S3: **Benchmarking RNA velocity methods based on cluster temporal ordering accuracy.** **a–f** Box plots showing cluster temporal ordering accuracy (CTO) scores for RNA velocity methods across six distinct datasets: Data 4 (**a**), Data 8 (**b**), Data 9 (**c**), Data 10 (**d**), Data 11 (**e**), and Data 12 (**f**). 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.

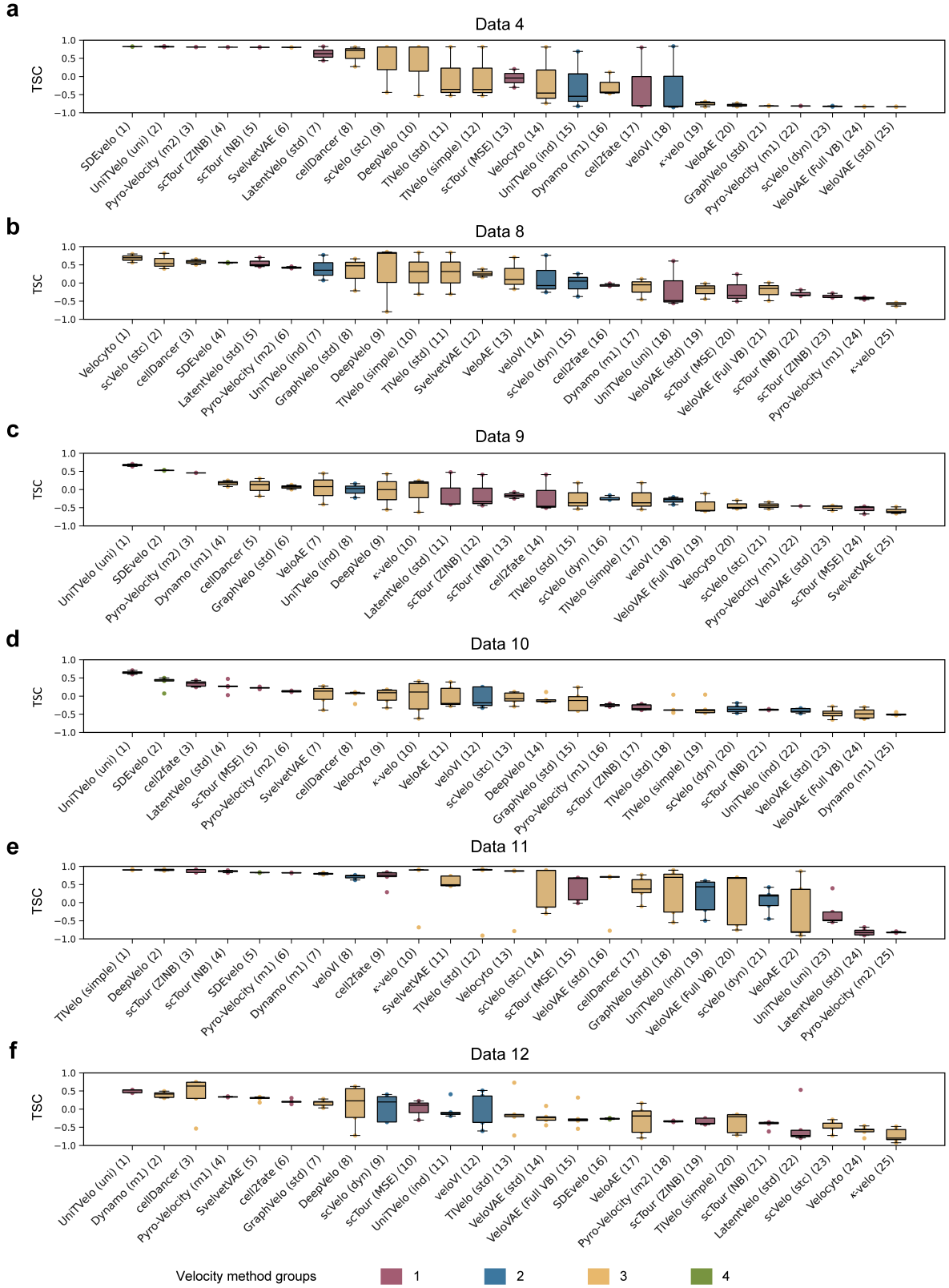


Figure S4: **Benchmarking RNA velocity methods based on temporal Spearman correlation.** **a–f** Box plots showing the temporal Spearman correlation (TSC) scores for various RNA velocity methods across six distinct datasets: Data 4 (**a**), Data 8 (**b**), Data 9 (**c**), Data 10 (**d**), Data 11 (**e**), and Data 12 (**f**). 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.

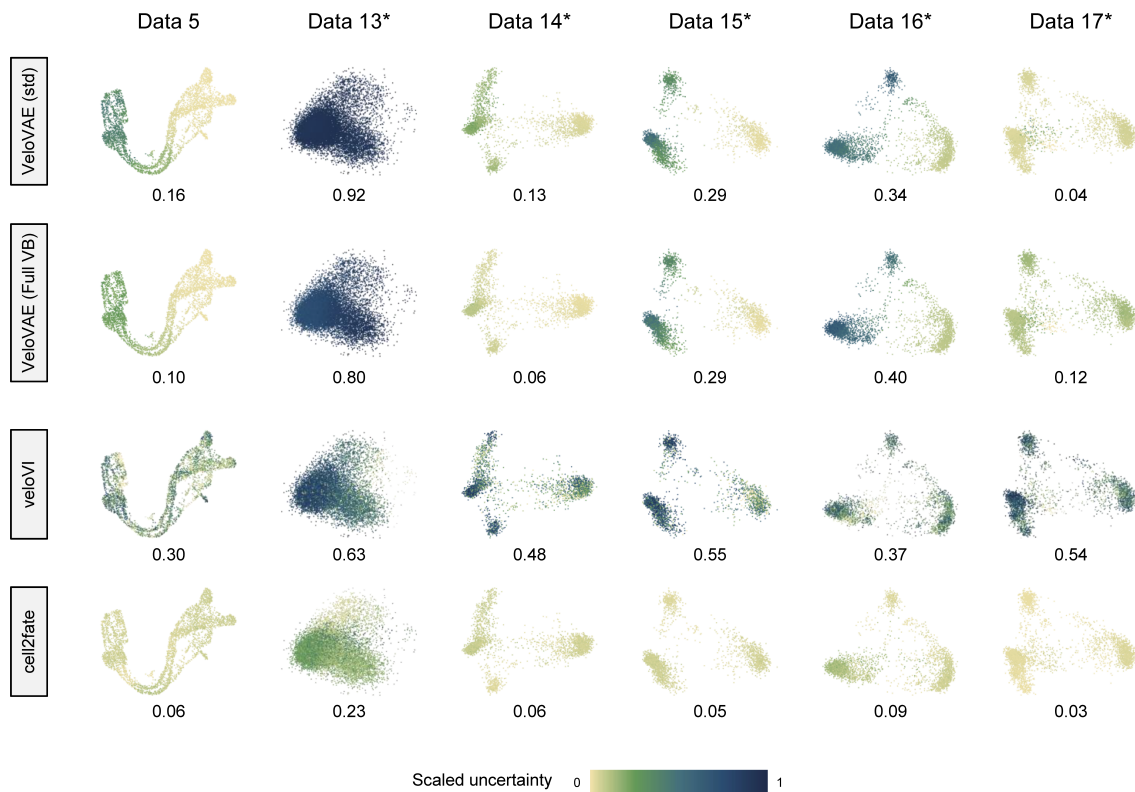


Figure S5: **Visualization of uncertainty scores across RNA velocity methods.** Low-dimensional embeddings colored by uncertainty scores for VeloVAE (std), VeloVAE (Full VB), veloVI, and cell2fate. Each row corresponds to a specific method. Uncertainty scores are min-max scaled independently for each method across six datasets (Data 5 and Data 13–17). The first column shows Data 5 (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.

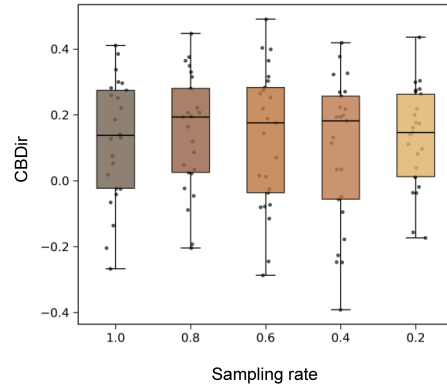
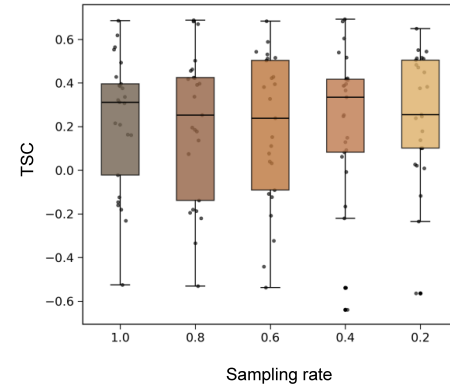
a**b**

Figure S6: **Robustness to sequencing depth variations.** **a** Box plots showing CDBDir scores across sampling rates ranging from 1.0 to 0.2 on Data 3, 4, and 6. Each point represents the average CDBDir 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 4, 8, and 9. Each point represents the average TSC value across the three datasets for a specific method.

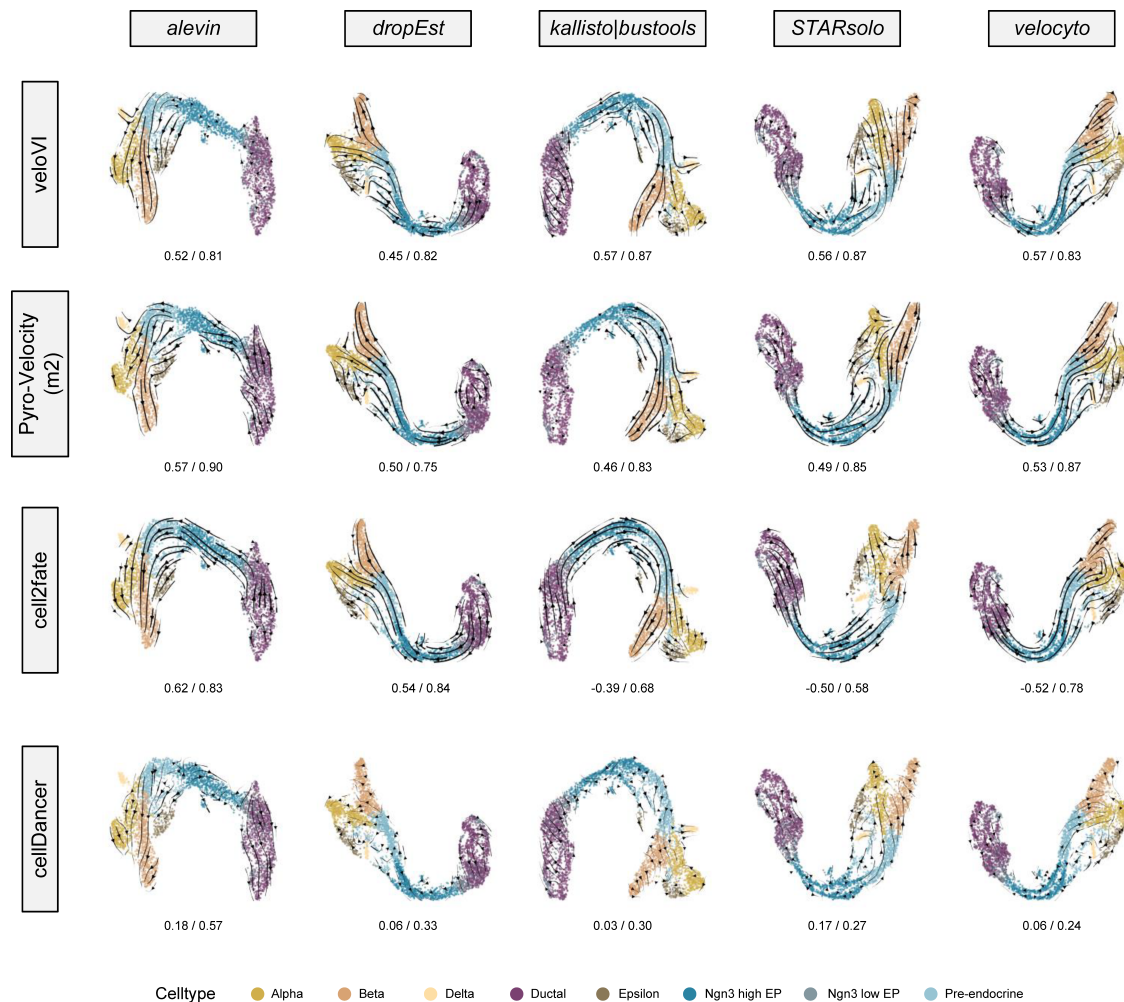
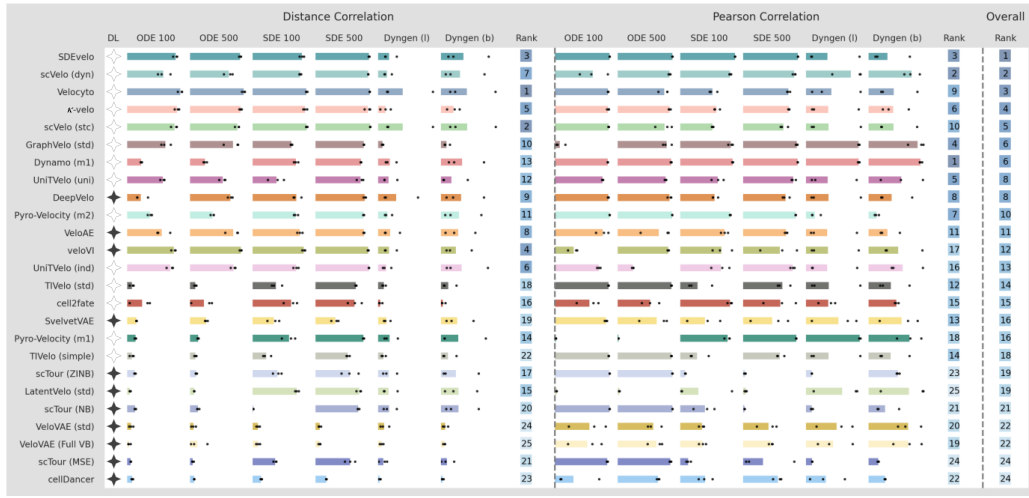
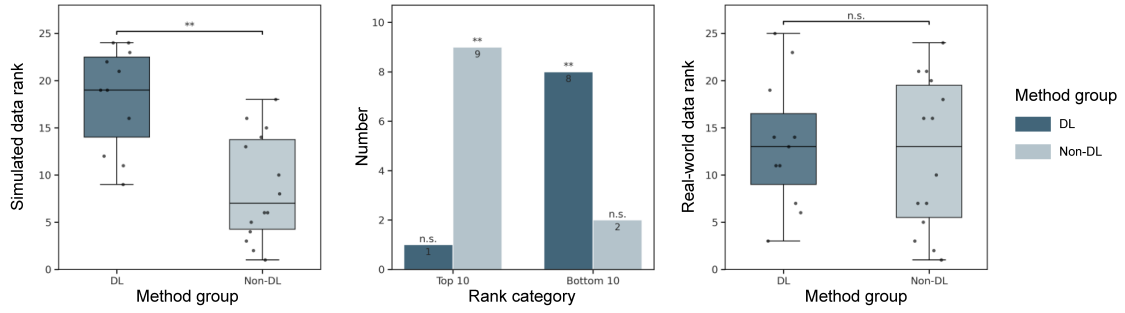


Figure S7: **Robustness to quantification methods.** Low-dimensional embeddings colored by cell type for veloVI, Pyro-Velocity (m2), cell2fate, and cellDancer. Each row represents a specific RNA velocity method, and each column corresponds to a quantification algorithm. Values shown below each embedding indicate CDBir (left) and ICVCoh (right) scores.

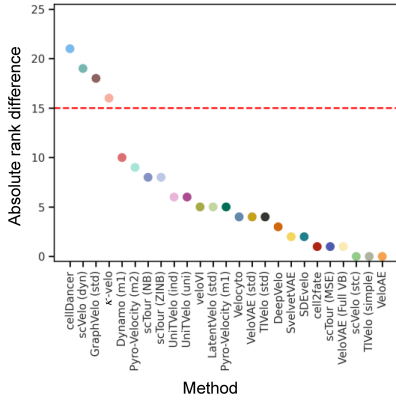
a



b



c



d

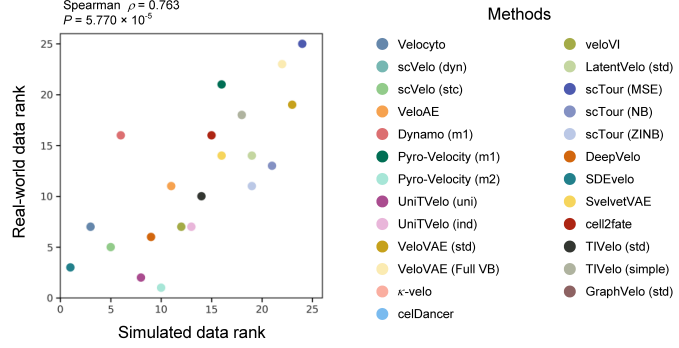
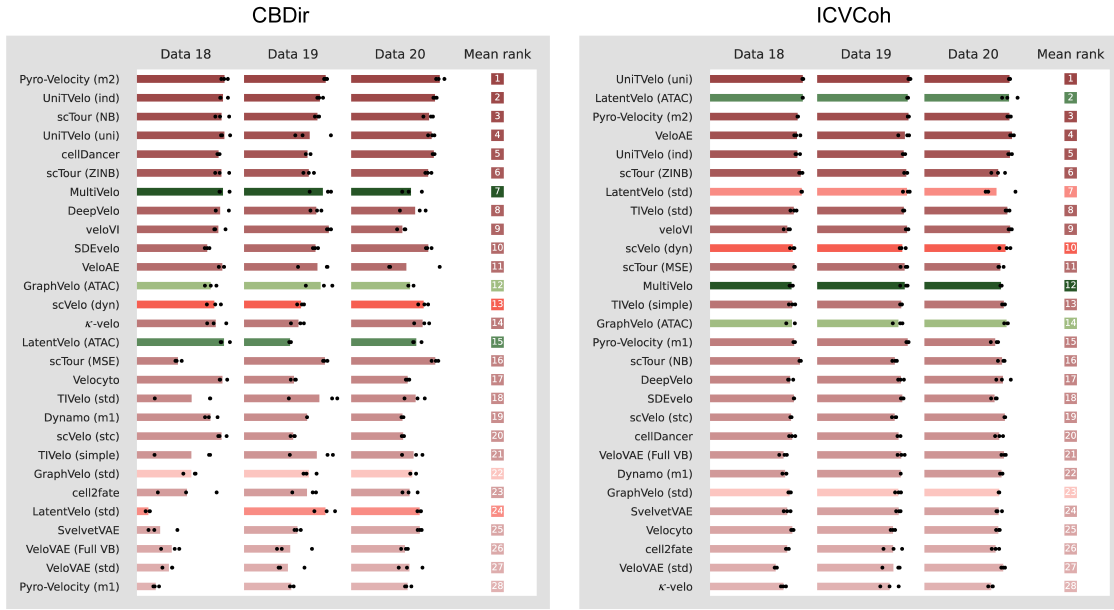


Figure S8: **Performance evaluation on simulated datasets.** **a** Ranking summary of 25 RNA velocity methods based on distance correlation and Pearson correlation across six simulated datasets (two ODE-based, two SDE-based, and two dyngen-generated). Methods are ordered by mean metric values so that 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 superior rankings compared with 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 ρ and the associated P -value for the two rankings are shown above the panel.

a



b

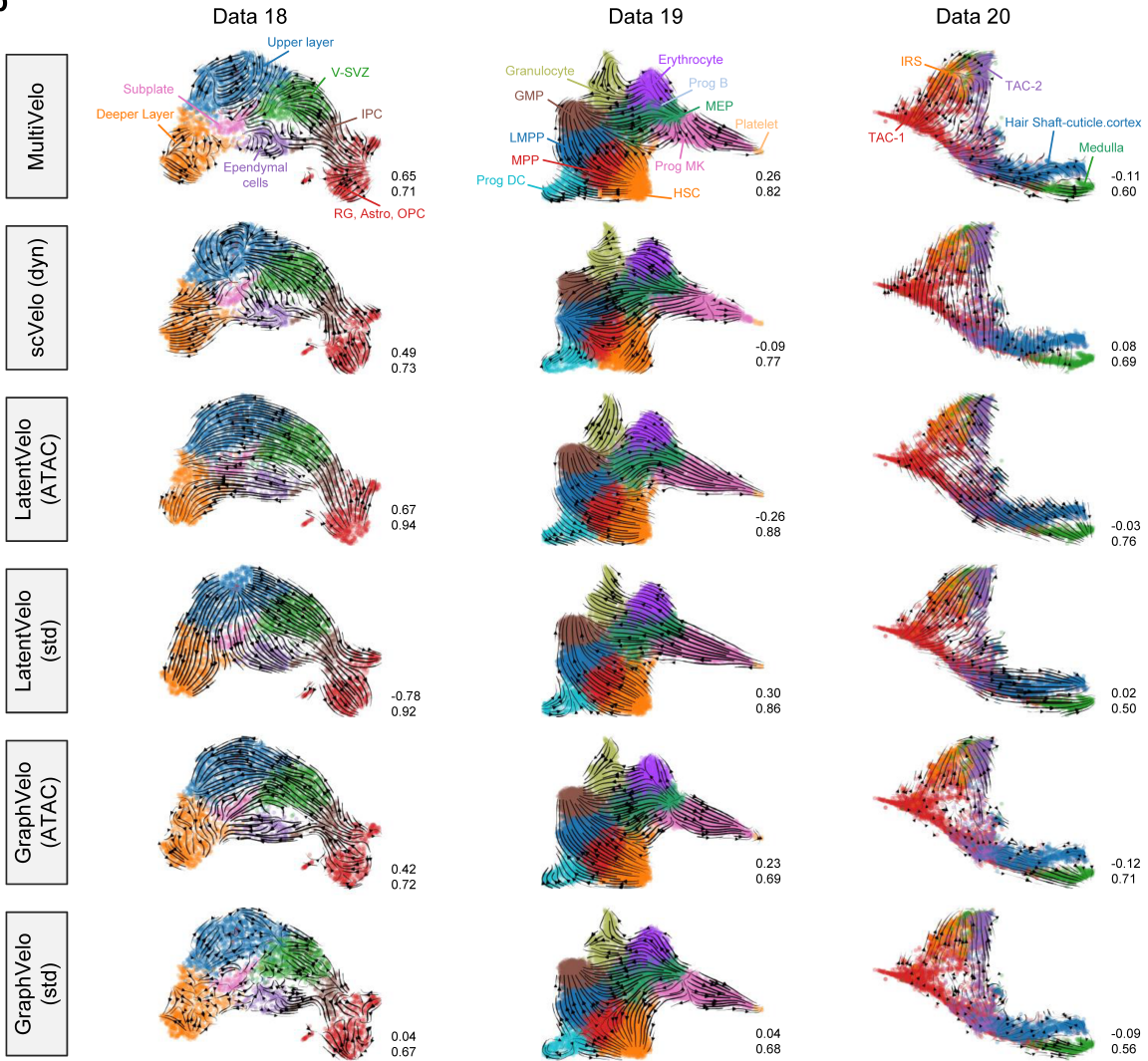
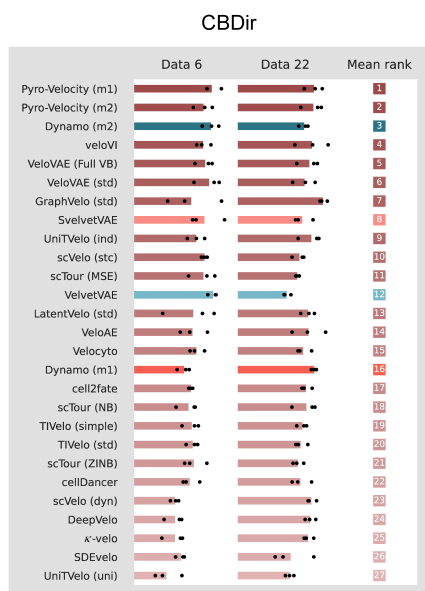
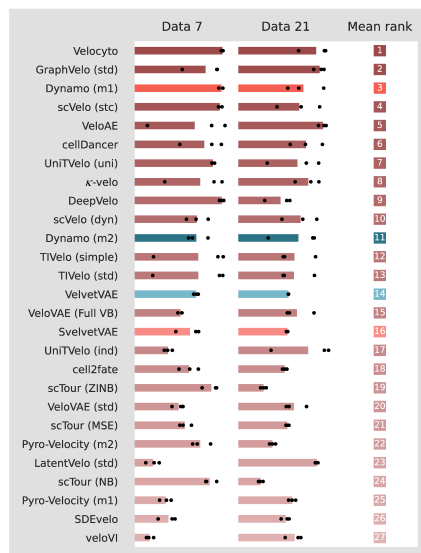


Figure S9: **Benchmarking chromatin accessibility-enhanced RNA velocity methods against 25 splicing dynamics-based methods.** **a** Ranking summary of 28 RNA velocity methods based on CBDir (left) and ICVCoh (right). Methods are ordered by their average rank across three datasets (Data 18, 19, 20), and 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 splicing dynamics baselines using Data 18, 19, and 20. Comparisons include MultiVelo versus scVelo (dyn), LatentVelo (ATAC) versus LatentVelo (std), and GraphVelo (ATAC) versus GraphVelo (std). The first row displays cell-type annotations. Quantitative performance scores appear in the bottom right corner of each plot, with the first score representing CBDir and the second representing ICVCoh.

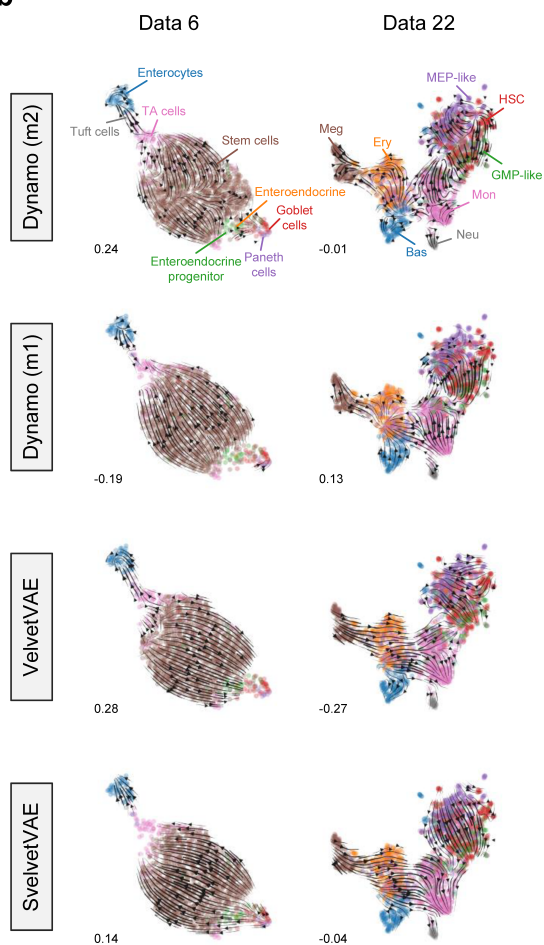
a



TSC



b



c

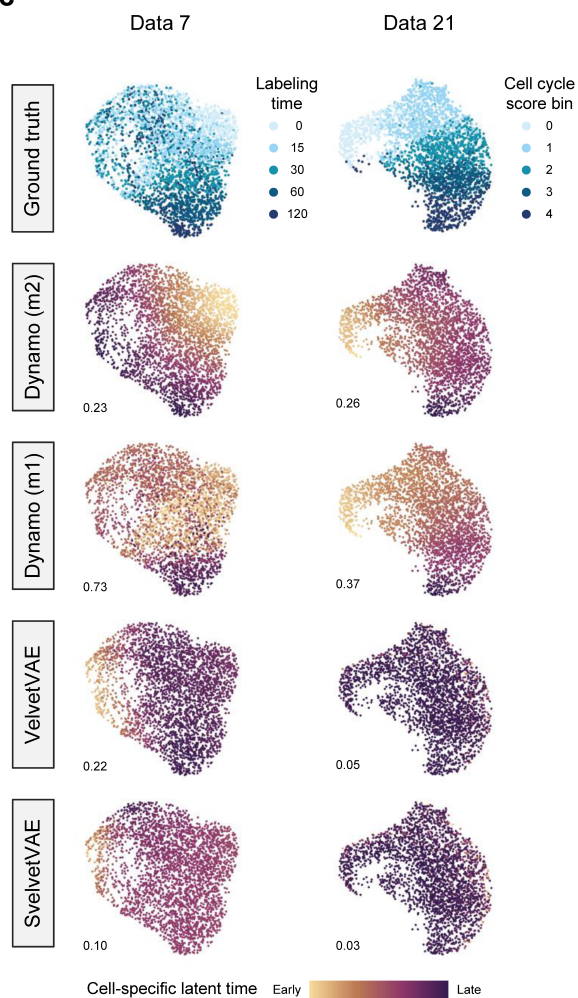
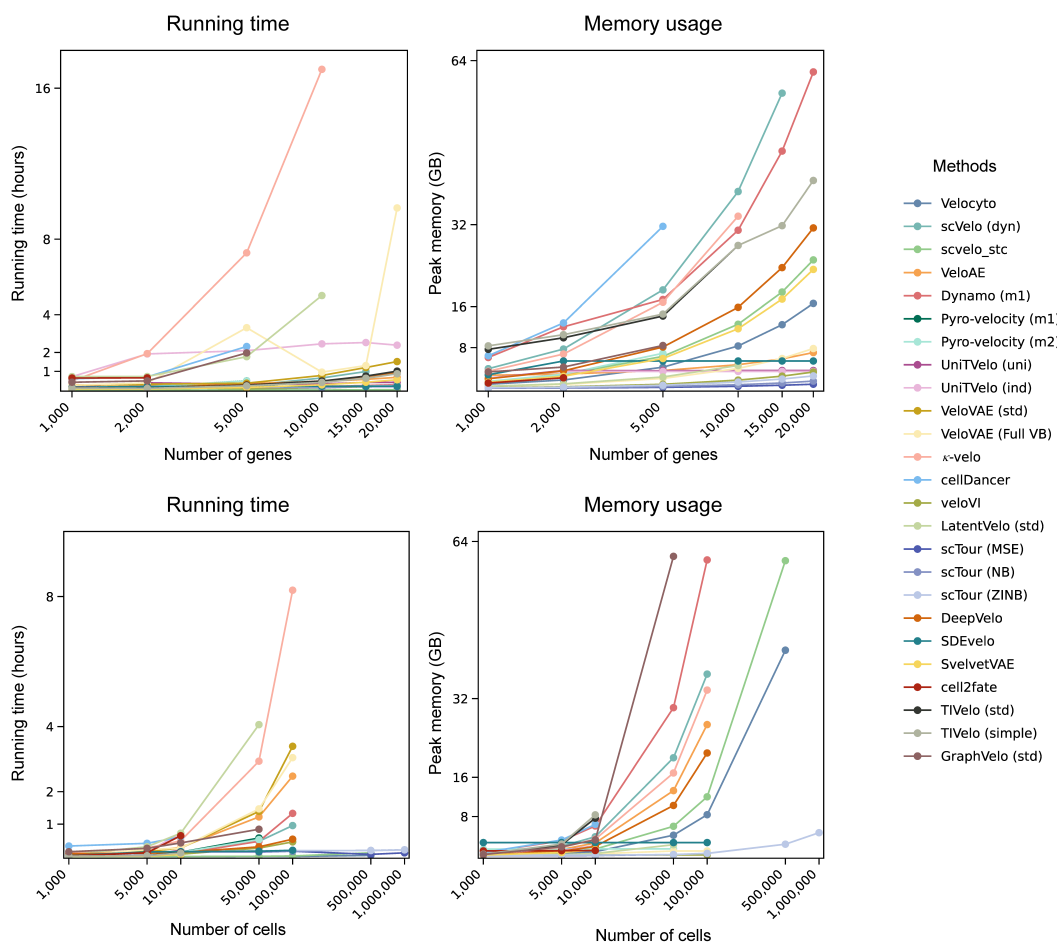


Figure S10: **Benchmarking metabolic labeling-enhanced RNA velocity methods against 25 splicing dynamics-based methods.** **a** Ranking summary of 27 RNA velocity methods across selected datasets. Left: CBDir rankings for Data 6 and Data 22. Right: TSC rankings for Data 7 and Data 21. Methods are sorted 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 splicing dynamics baselines using Data 6 and Data 22. Comparisons include Dynamo (m2) versus Dynamo (m1) and VelvetVAE versus SvelvetVAE. The first row displays cell-type annotations. CBDir scores appear in the bottom left corner of each plot. **c** The first row displays scatter plots in the projected UMAP space, with cells colored by labeling time for Data 7 and by cell cycle score bin for Data 21. 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-left corner of each plot.

a



b

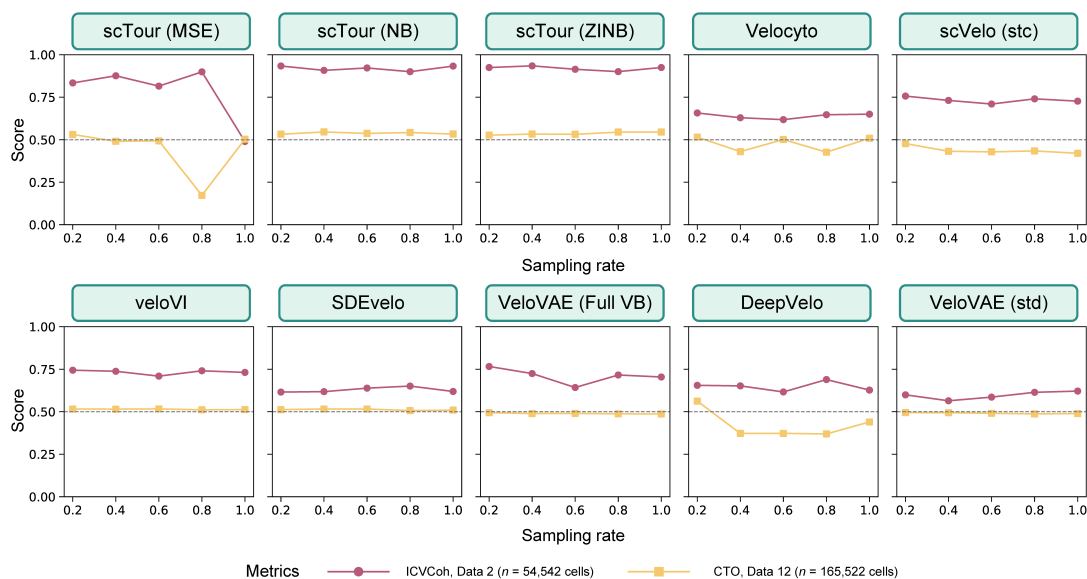


Figure S11: **Computational scalability analysis of 25 splicing dynamics–based RNA velocity methods.** **a** Line plots showing the running time (hours, left) and peak memory usage (GB, right) for each method. Methods are color-coded as indicated in the legend. Top: Performance scaling with increasing 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 2, while the yellow line indicates TSC using Data 12. Sampling rates range from 0.2 to 1.0 (the original data). The dashed line at $y=0.5$ serves as a baseline.