Supplementary Figures for "SDEvelo: a deep generative approach for transcriptional dynamics with cell-specific latent time and multivariate stochastic modeling"

Xu Liao[1†], Lican Kang[2†], Yihao Peng[1], Xiaoran Chai[3], Chengqi Lin[4],

Hongkai Ji[5], Yuling Jiao[6*], and Jin Liu[1*]

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

[2]Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, Singapore,

[3]Cancer and Stem Cell Biology Program, Duke-NUS Medical School, Singapore,

[4]Key Laboratory of Developmental Genes and Human Disease, School of Life Science and Technology, Southeast University, Nanjing, China,

[5]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA,

[6]School of Mathematics and Statistics, Wuhan University, Wuhan, China.

[†]Equal contributions.

[*]Corresponding author. Email: yulingjiaomath@whu.edu.cn, liujinlab@cuhk.edu.cn.

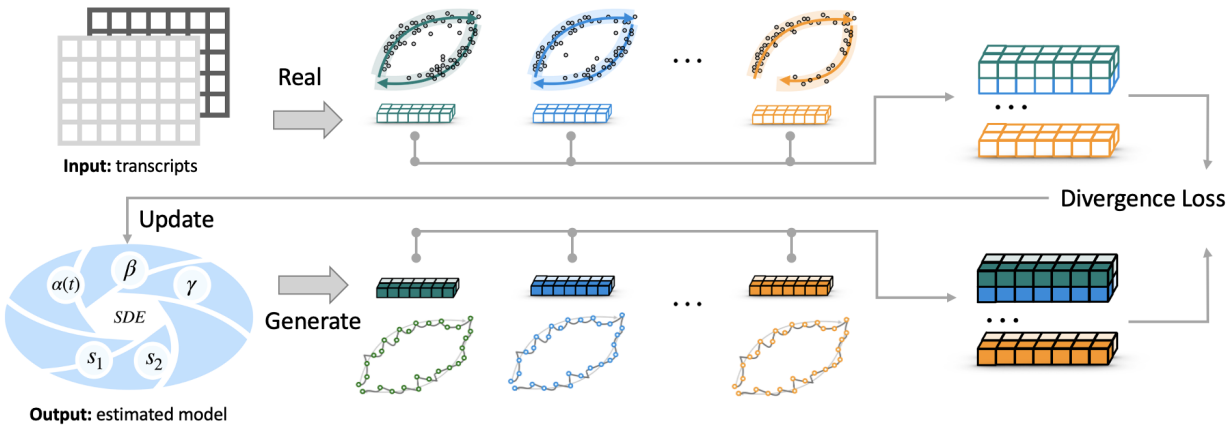

Figure S1: **Overview of the novel algorithm proposed for efficiently and effectively estimating parameters in SDEvelo.** The core idea is based on adversarial learning, where the generative SDE model is trained to match the distribution of real data. The algorithm starts with the input of unspliced and spliced matrices (representing real data), while simultaneously obtaining generated trajectories from the generative SDE model (representing generated data). In this model, parameters controlling the dynamics, kinetics, and noise will be updated. Iterative minimization of the distributional divergence between real data and SDE-generated data is performed to update the parameters of SDEvelo.

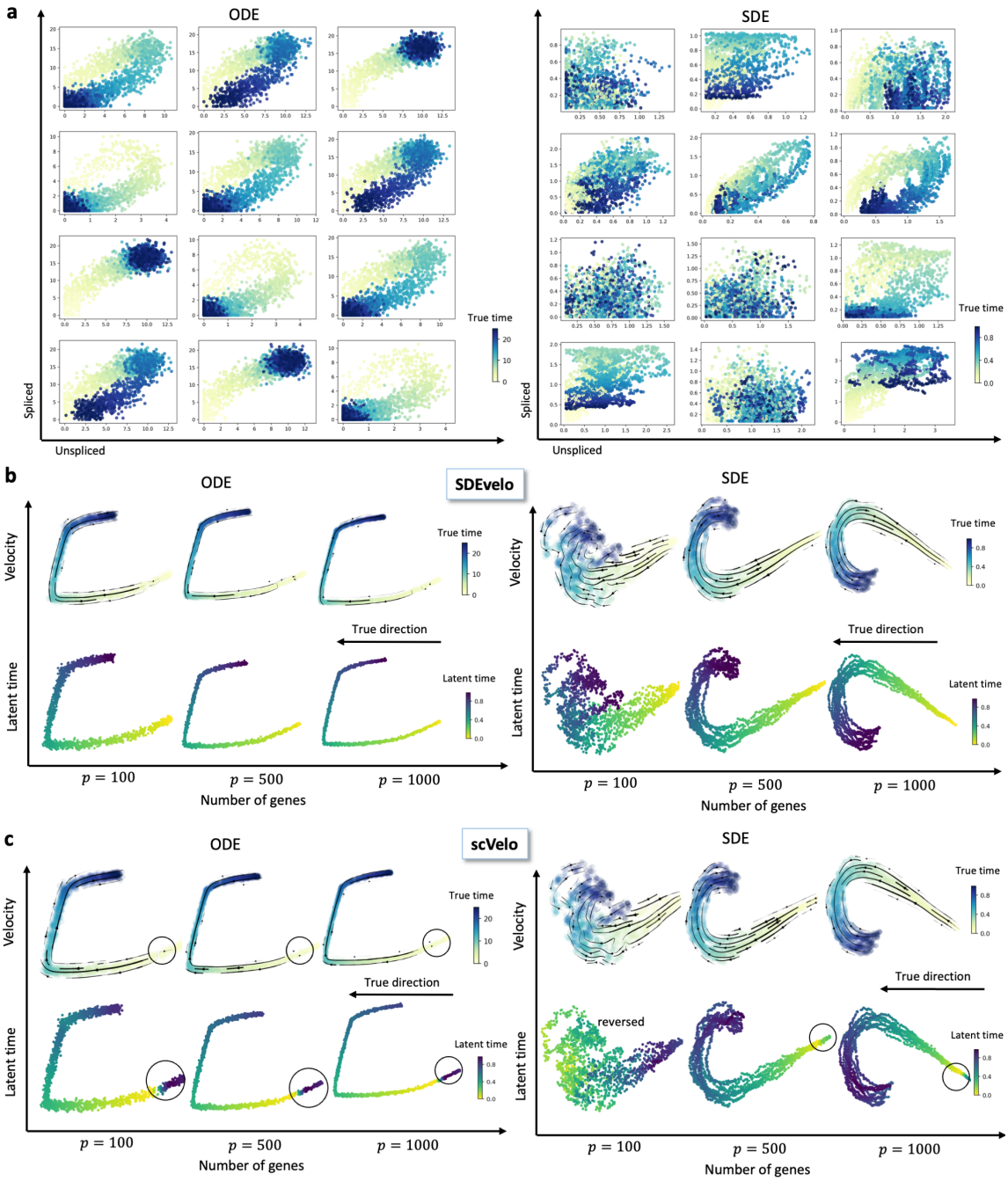

Figure S2: **Comparative analysis of ODE and SDE simulated settings across varying gene numbers** (100 **to** 1,000)**. (a)** Scatter plots of examples from ODE and SDE simulated settings when the number of genes was 100. Left panel: ODE scatter plots exhibit a more regular pattern, influenced solely by external Gaussian noises. Right panel: In contrast, scatter plots from the SDE setting exhibit inherent randomness during the generation process. **(b)** Streamline plots and latent time heatmaps estimated by SDEvelo, with the number of genes varying from 100 to 1,000. Across all settings, SDEvelo accurately identified the correct velocities in streamline plots, and the latent times aligned with the true ones. The true direction is indicated by the black arrow. **(c)** Streamline plots and latent time heatmaps estimated by scVelo, with the number of genes varying from 100 to 1,000. scVelo identified some inconsistent directions, and parts of erroneous latent times are highlighted with circles.

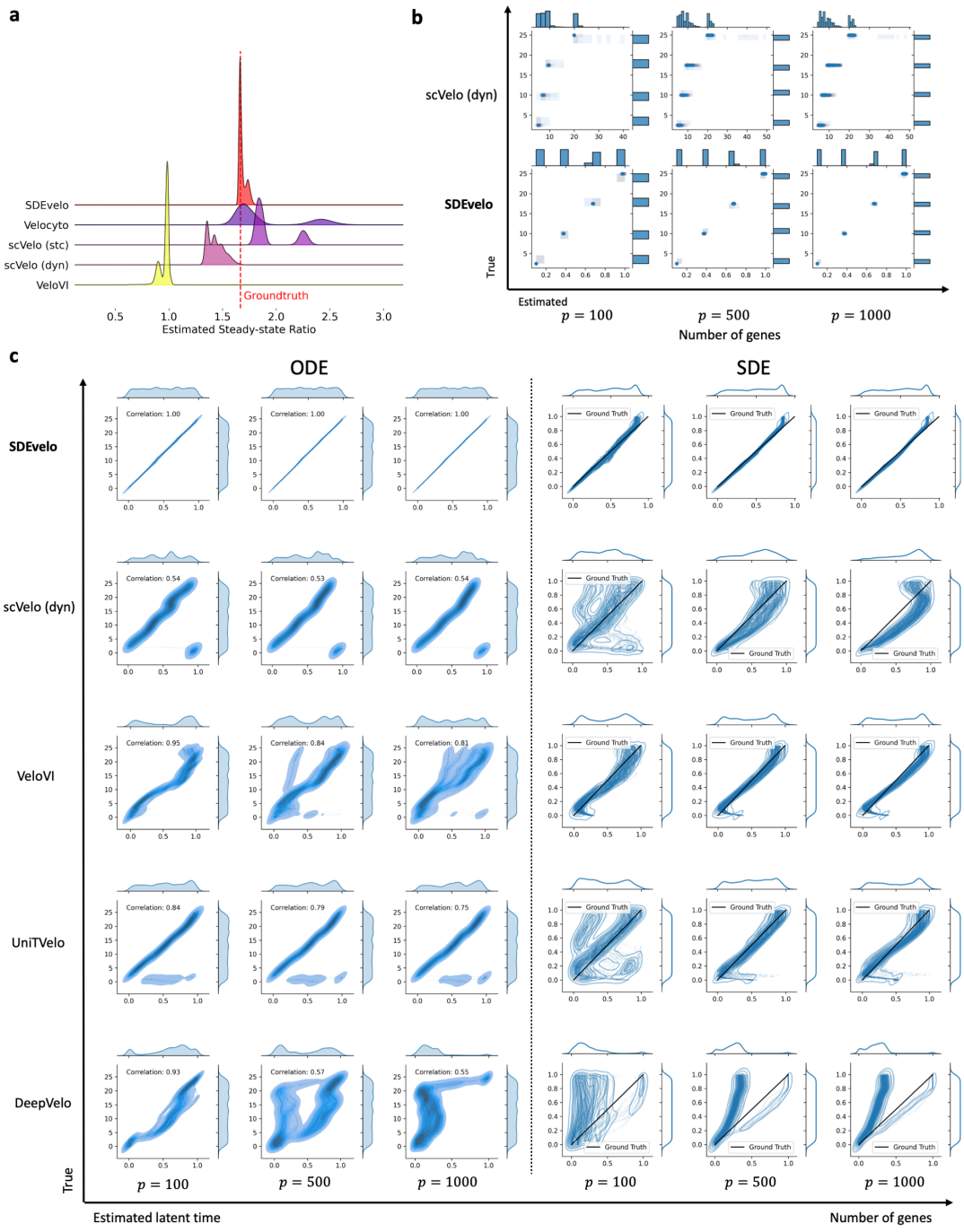

Figure S3: **Benchmarks of switch time points and latent time estimations on simulated settings. (a)** Ridgeline plot comparing the distributions of estimated ratios across all methods. Other methods exhibited biased estimations; notably, Velocyto and scVelo (stc) showed obvious bimodal distributions. However, the distribution from SDEvelo closely aligned with the ground truth. The red dotted line indicates the true constant of the steady-state ratio. **(b)** KDE plots comparing the true and estimated switch time points between SDEvelo and scVelo (dyn), with the number of genes varying from 100 to 1,000. **(c)** KDE plots comparing the true and estimated latent times across different methods, with the number of genes varying from 100 to 1,000. In the ODE setting, we demonstrate the correlation between the true and estimated latent times. In the SDE setting, the ground truth line, ranging from 0 to 1 and highlighted in a dark color, since the true latent time is on the same scale as the estimated one.

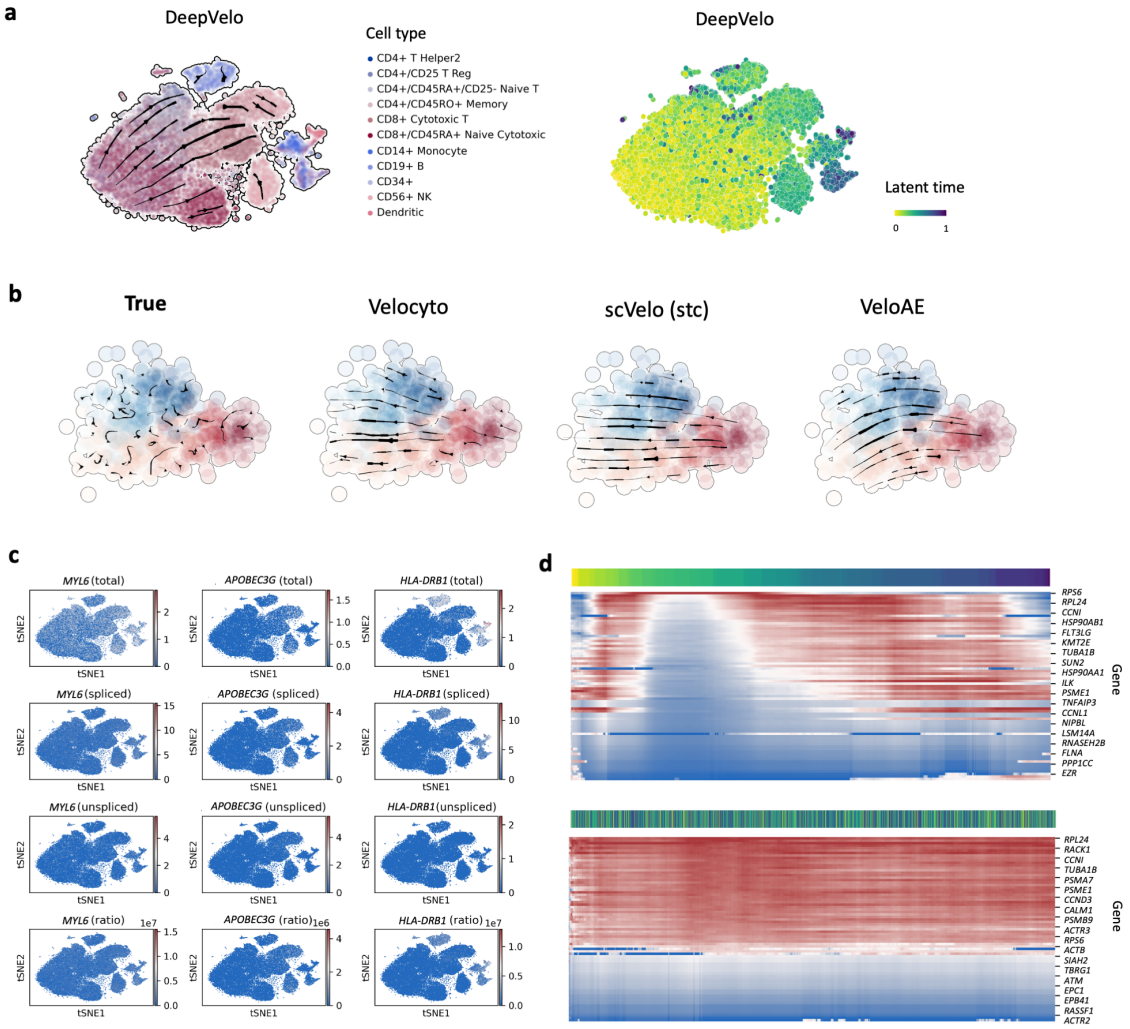

Figure S4: **Assessment of negative controls in mature-state cell populations. (a)** Streamline plot and estimated latent time heatmap of DeepVelo. **(b)** Streamline plots of true velocity and velocity estimated by other methods. Compared with the random pattern of the true velocity, the other methods displayed strong erroneous directions. **(c)** Heatmap of gene expression (total), unspliced, spliced, and ratio between unspliced and spliced data (ratio) for *MYL6*, *APOBEC3G*, and *HLA-DRB1*. **(d)** Gene expression heatmap for cell-cycle genes in CD8+ Cytotoxic T cells mapped against latent time as estimated by SDEvelo (upper) versus permutated time (lower).

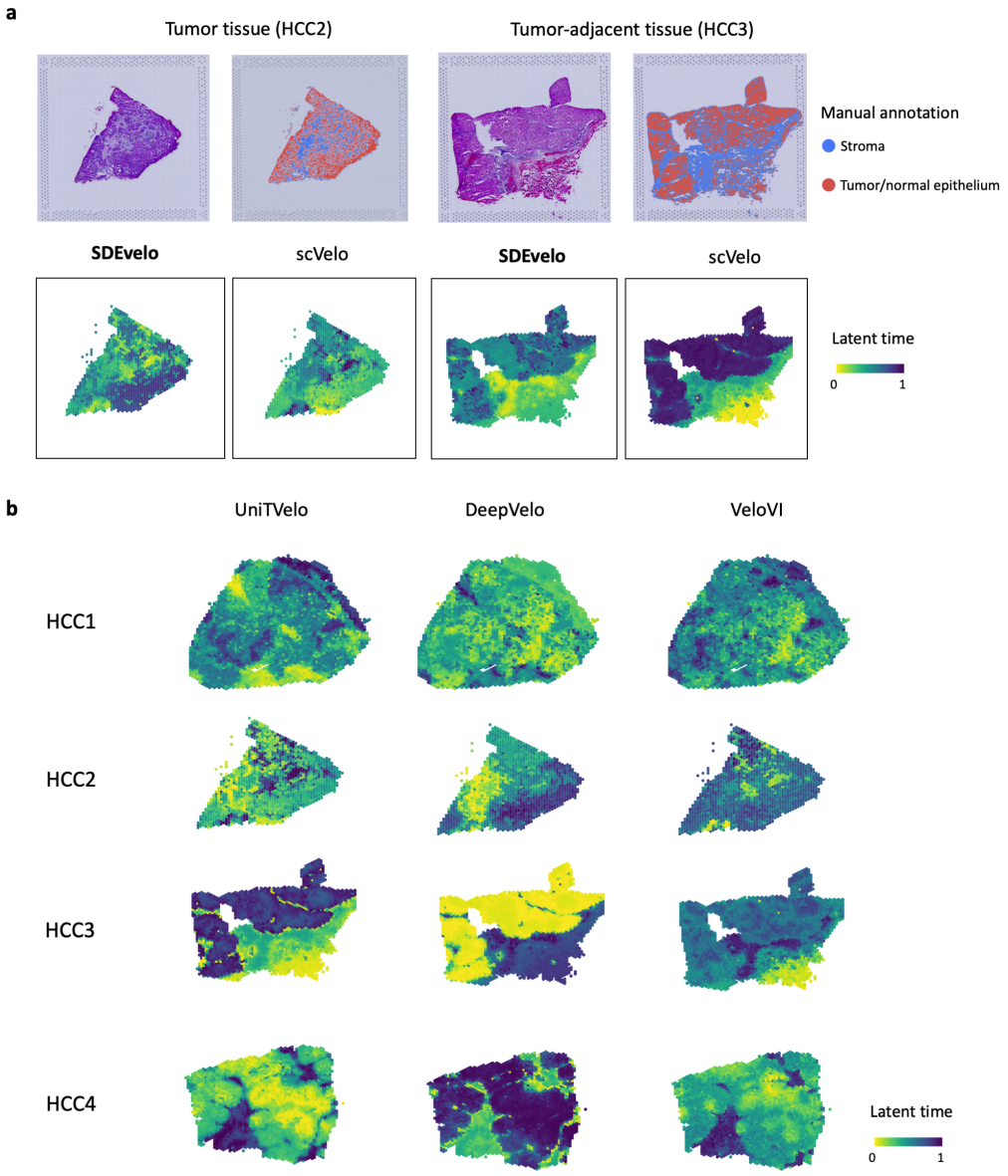

Figure S5: **Comparison results of HCC data in spatial transcriptomics. (a)** Top panel: H&E images and manual annotations by a pathologist from tumor (HCC2) and tumor-adjacent (HCC3) tissue sections. Bottom panel: Estimated latent time heatmap for SDEvelo and scVelo (dyn). Compared with scVelo (dyn), SDEvelo accurately identified the correct dynamics of carcinogenesis. **(b)** Latent time heatmaps projected onto the spatial space, estimated by UniTVelo, DeepVelo, and VeloVI across four tissue sections. These methods either identified reversed dynamics or produced unclear patterns that are hard to distinguish between stroma and TNE.

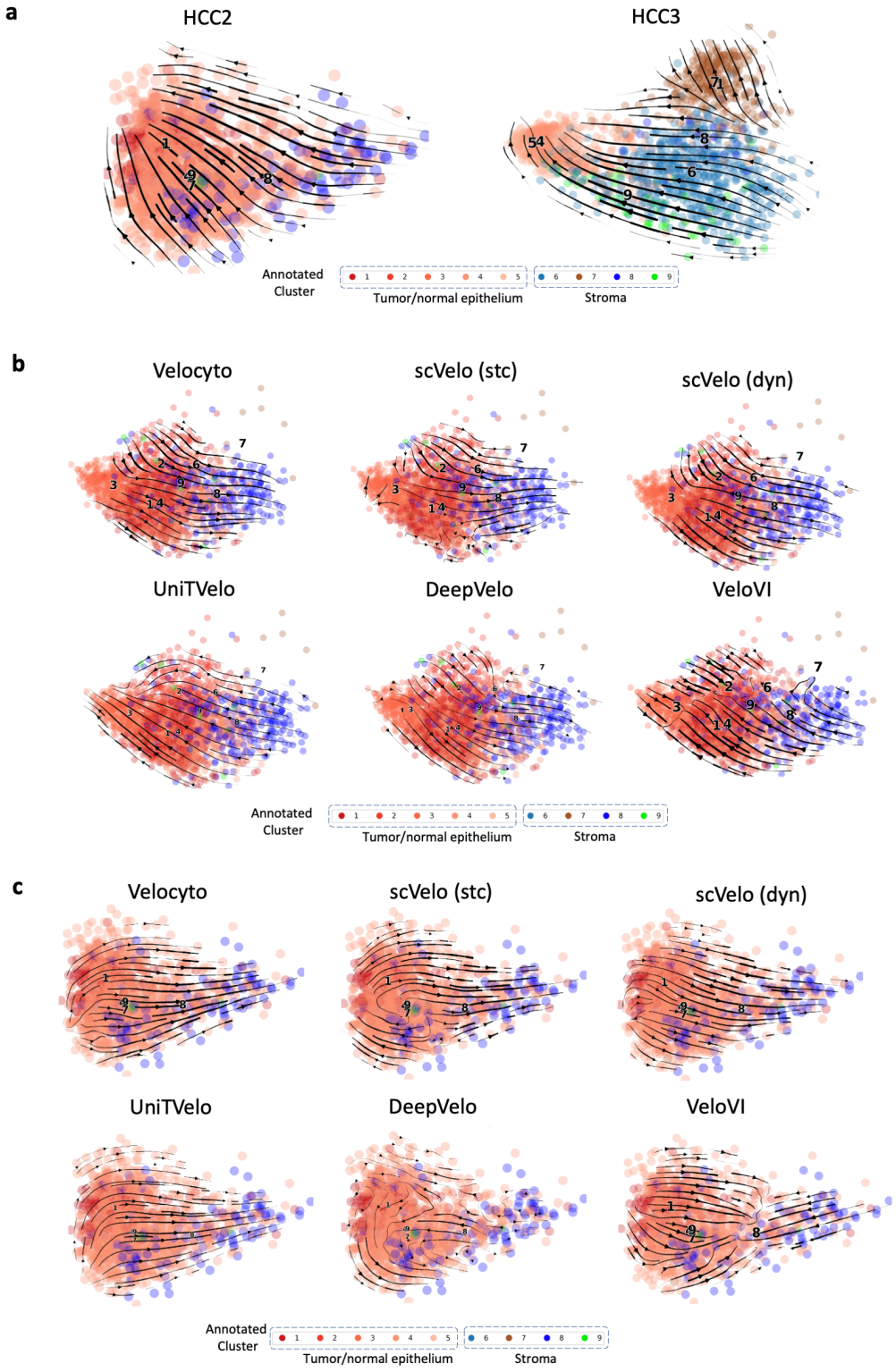

Figure S6: **Comparative velocity estimation results across all methods in HCC data. (a)** Streamline plots on PCA with the velocity estimated by SDEvelo in HCC2-3. **(b)** Streamline plots projected onto PCA with the velocity estimated by other methods in HCC1. **(c)** Streamline plots on PCA representing velocities as estimated by the other methods for HCC2.

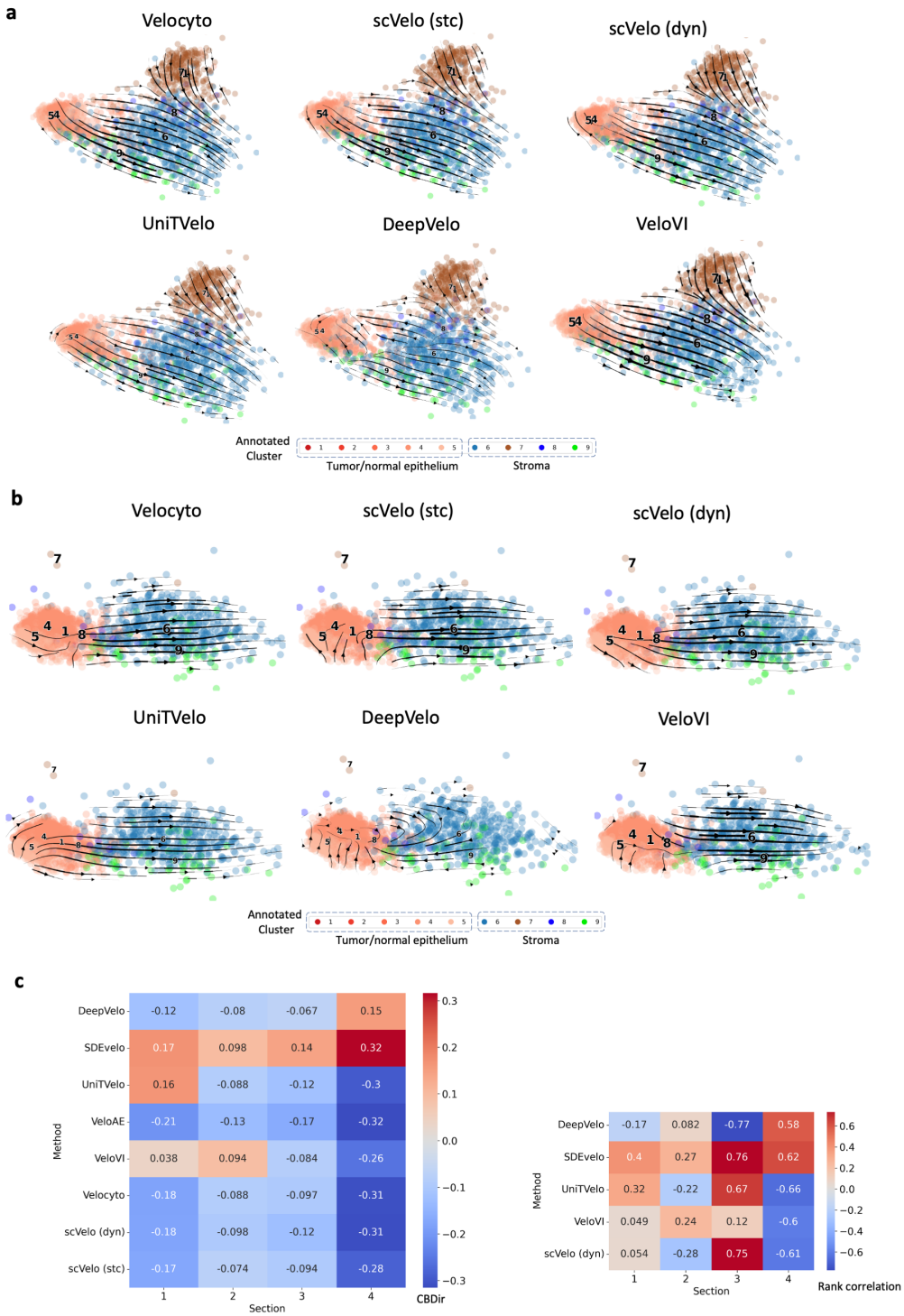

Figure S7: **Comparative velocity estimation results of others methods in HCC data.** **(a)** **(a)** Streamline plots projected onto PCA with the velocity estimated by other methods in HCC3. **(b)** Streamline plots on PCA representing velocities as estimated by the other methods for HCC4. **(c)** Heatmaps of cross-boundary direction correctness (CBDir) (right) and rank correlation (left) across four HCC sections to benchmark the performance of all methods. CBDir measures the consistency between estimated velocity and biological transitions, where a higher value indicates better performance. A higher Spearman rank correlation suggests that the estimated latent time is more consistent with the true transitions.

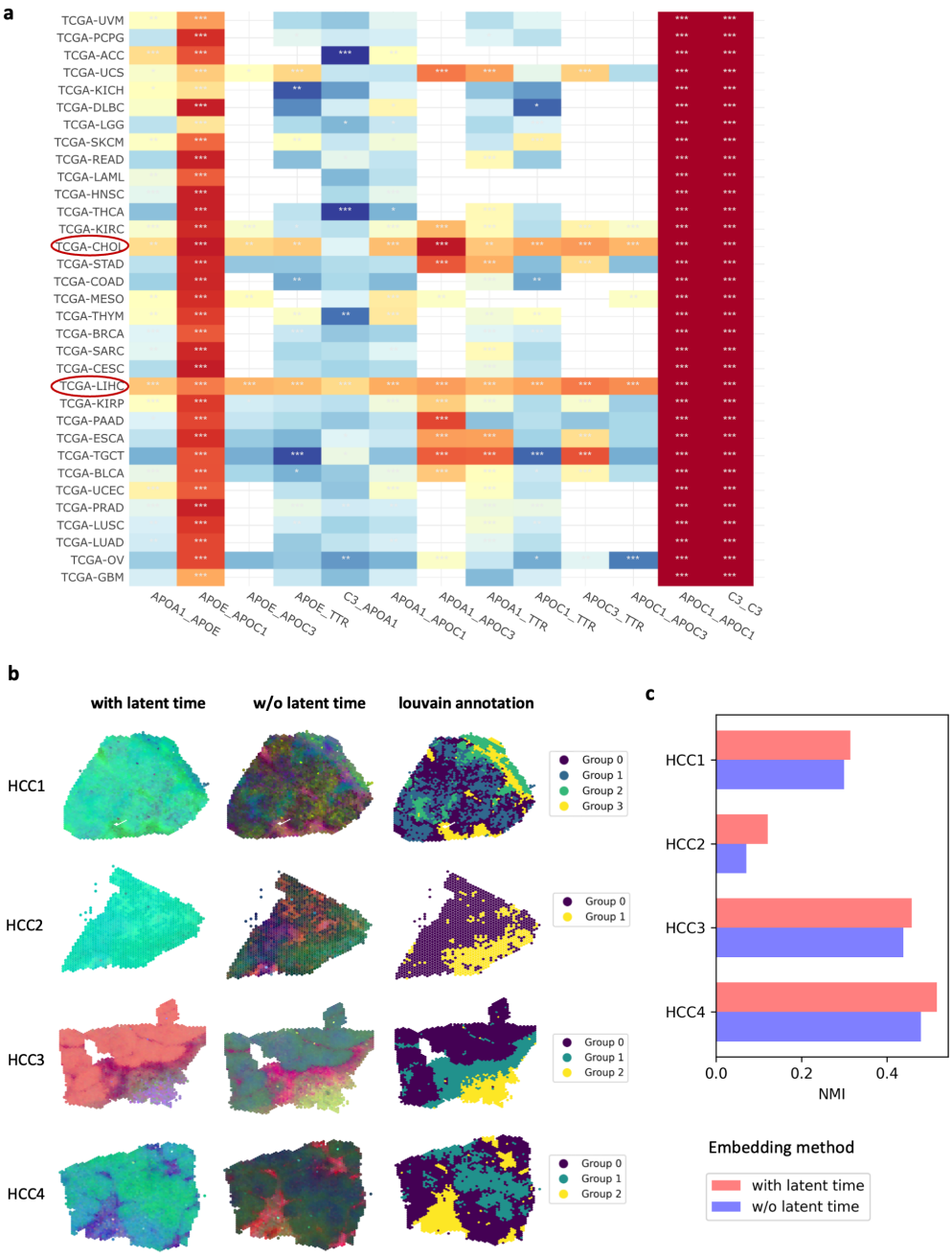

Figure S8: **SDEvelo facilitated downstream analyses.** **(a)** Heatmap of mRNA expression Spearman correlation coefficients for interacting protein pairs in the PPI network across cancer types on TCGA, emphasizing The Cancer Genome Atlas Cholangiocarcinoma (TCGA-CHOL) and The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) for their significant high values. **(b)** Left panel: UMAP RGB plots of HCC1-4, comparing scenarios with and without latent time information in spatial space. Right panel: Louvain annotation with embeddings, integrated with latent time information in spatial space. **(c)** Bar plot illustrating the comparison of gene expression clustering performance in terms of Normalized Mutual Information (NMI) values, both with and without integrating the estimated latent time from SDEvelo. A higher NMI indicates better clustering accuracy for the alignment with annotations from PRECAST.

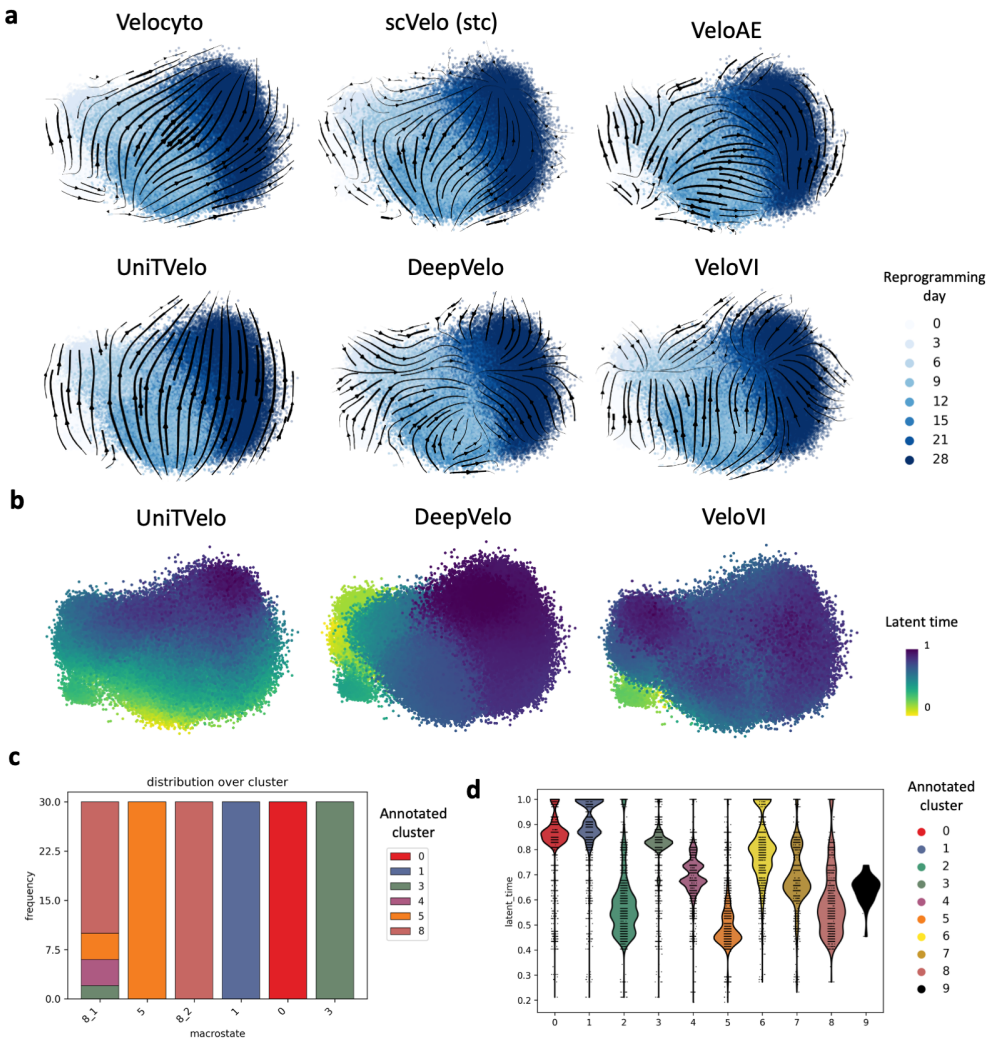

Figure S9: **Integrative analysis of mouse embryonic reprogramming dynamics: streamline visualization, latent time heatmaps, macrostate distribution, and latent time estimation across clusters.** **(a)** Streamline plots from other methods applied to mouse embryonic reprogramming data, with each cell colored according to reprogramming days, ranging from day 0 to 28. **(b)** Latent time heatmaps estimated using the other methods. **(c)** Stacked frequency histograms of macrostates over categorical cluster annotations, illustrating the relationship between the macrostates estimated by CellRank and the original cluster annotations. **(d)** Violin plot of latent time estimated by SDEvelo across various annotated clusters.

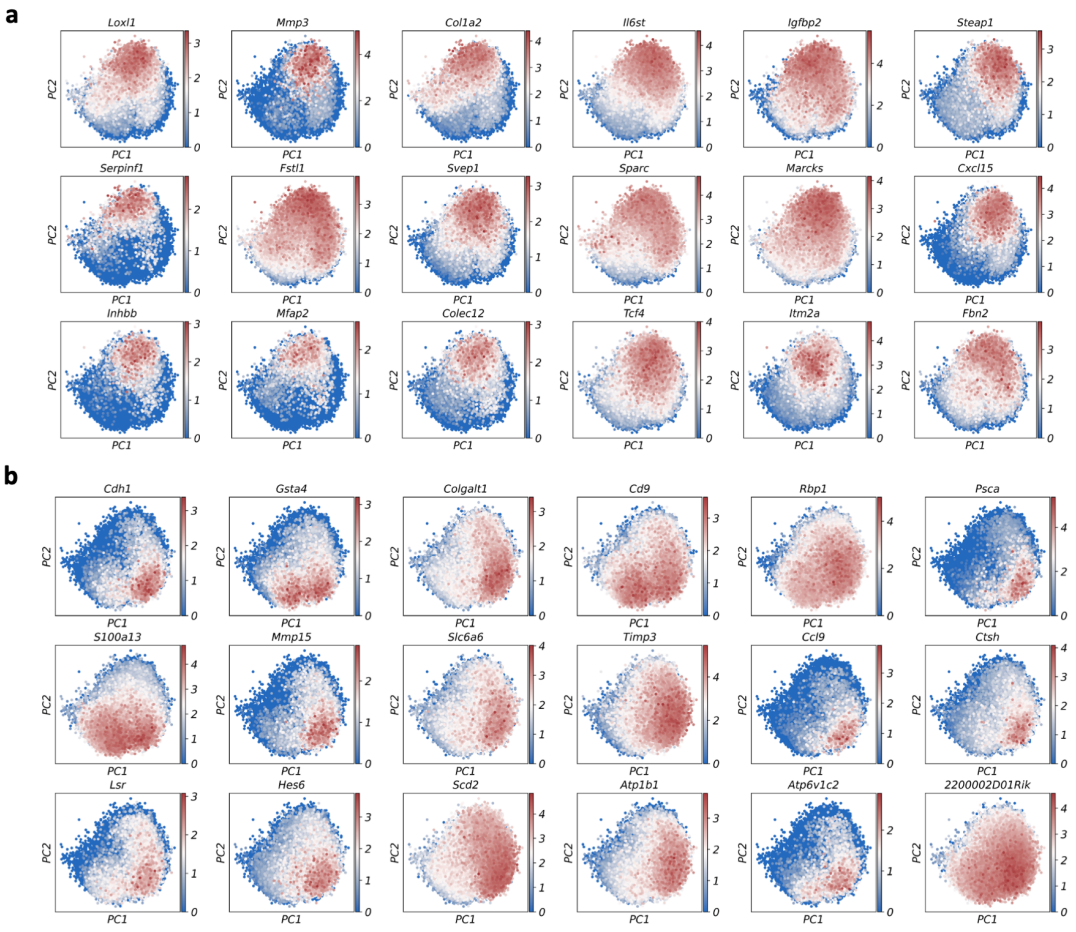

Figure S10: **Gene expression heatmaps in mouse embryonic reprogramming data.** **(a)** Expression heatmaps for genes associated with dead-end fates. **(b)** Expression heatmaps for genes associated with reprogramming fates.

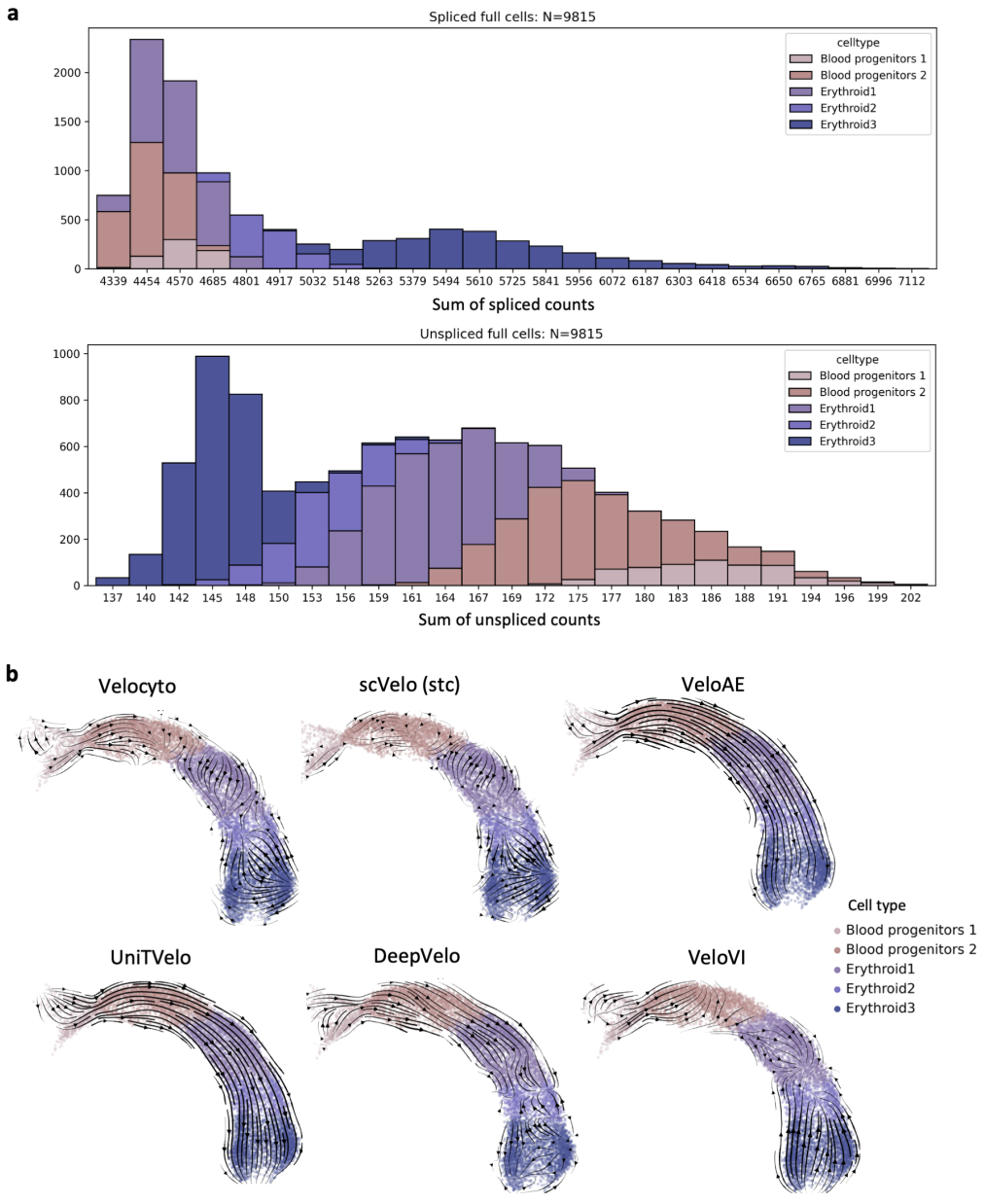

Figure S11: **Results for mouse erythroid data.** **(a)** Stacked histograms displaying transcript abundance for spliced data (upper panel) and unspliced data (lower panel) across different cell types. **(b)** Streamline plots from other methods, colored by cell types.

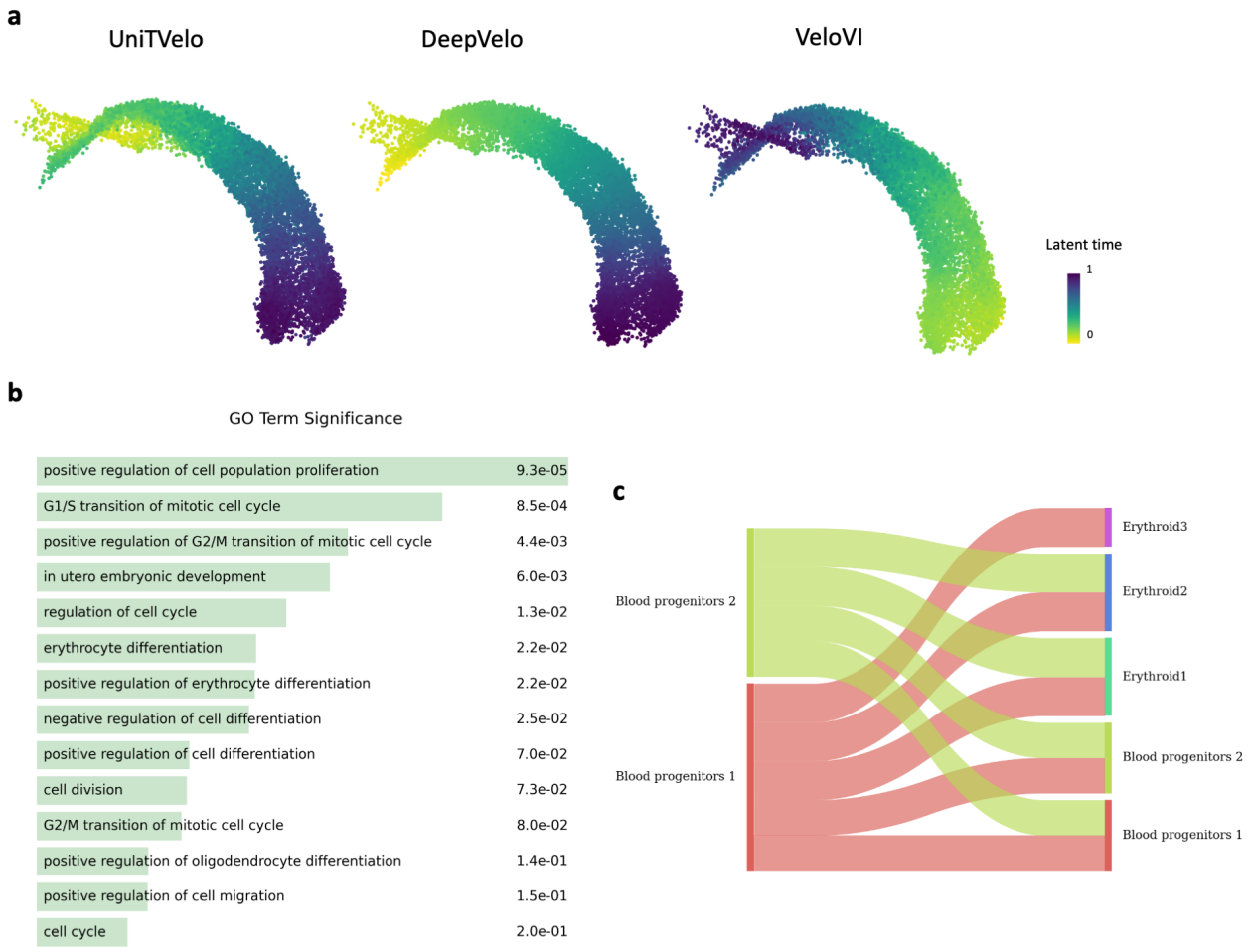

Figure S12: **Results for mouse erythroid data.** **(a)** Latent time heatmaps of other methods. **(b)** GO enrichment terms and their corresponding p-values for genes selected by SDEvelo. **(c)** Cell-cell communication plot for mouse erythroid data.

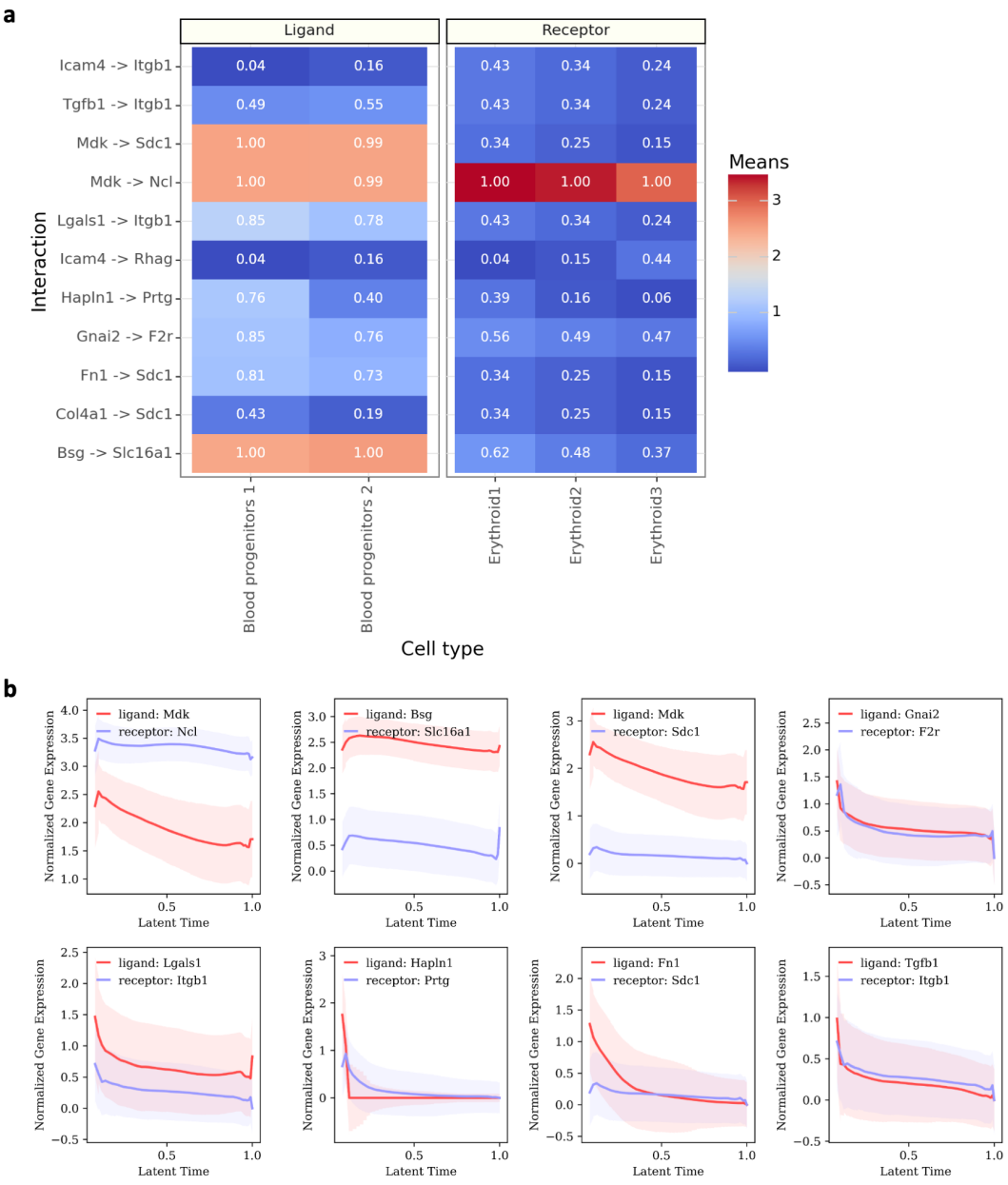

Figure S13: **Cell-cell interaction results for mouse erythroid data.** **(a)** Heatmap showing the intensity of ligand-receptor interactions, with blood progenitors as the ligand source and erythroids as the receptor target. **(b)** Plots of normalized gene expression trends for ligand-receptor pairs along the latent time estimated by SDEvelo.