

Supporting Information Appendix

1 Supplementary Methods

1.1 Topic modeling and differential expression analysis

In the probabilistic topic model for scRNA-seq data, for cell i , $x_i = (x_{i1}, \dots, x_{iM})$ is assumed to be drawn from a multinomial distribution

$$x_{i1}, \dots, x_{iM} \mid t_i \sim \text{Multinom}(t_i; \pi_{i,1}, \dots, \pi_{i,M}) \forall 1 \leq i \leq C \quad (1)$$

where C is the number of cells, M is the number of genes, x_{im} is number of mRNA transcripts for gene m in cell i , and $t_i = \sum_{m=1}^M x_{im}$ is the total number of transcripts in cell i . The multinomial probabilities are

$$\pi_{i,m} = \sum_{k=1}^K L_{ik} F_{mk} \quad (2)$$

where K is the user-specified number of topics; $L \in \mathbb{R}_+^{C \times K}$ is the cell topic weight matrix, and L_{ik} is the probability of topic k in cell i ; $F \in \mathbb{R}_+^{M \times K}$ is the gene topic weight matrix, and F_{mk} is the weight of gene m in topic k .

For a given K , FastTopics exploits the equivalence of the maximum likelihood estimates for Poisson non-negative matrix factorization (NMF) and the multinomial topic model [1]. The negative of the log-likelihood of the topic model for cell i and gene m is:

$$-\log p(x_{im} \mid L, F) = -\log \left(\frac{(L_i^T F_m)^{x_{im}} e^{-L_i^T F_m}}{x_{im}!} \right) \quad (3)$$

where L_i and F_m are column vectors (of size K) containing the i th row of L and the m th row of F , respectively.

After discarding the terms that are not related to L and F and summing over all cells and genes, we arrive at a suitable loss function [1]:

$$\begin{aligned} \text{minimize } l(L, F) &= \sum_{i=1}^C \sum_{m=1}^M L_i^T F_m - x_{im} \log(L_i^T F_m) \\ \text{subject to } L &\geq 0, F \geq 0, \sum_k L_{ik} = 1 \forall i, \sum_m F_{mk} = 1 \forall m \end{aligned} \quad (4)$$

In other words, the optimal L and F are fitted such that, accounting for the heterogeneity in cells over the topics and the contributions of individual genes to each topic, the input count matrix should be recovered on expectation.

Since each transcript count is generated by a Poisson model, the differentially expressed genes can be identified by computing the log fold change (LFC) of each gene in topic k [1, 2], defined as

$$LFC(k) = \log_2 \left(\frac{F_{mk}}{F_{mk'}} \right). \quad (5)$$

$$k' = \arg \min_{k' \neq k} \left| \frac{F_{mk}}{F_{mk'}} - 1 \right|$$

The posterior distribution of the LFC and local false sign rate (lfsr) [3] are then estimated with MCMC and stabilized with adaptive shrinkage. While the majority of differentially expressed (DE) genes within a topic tend to be positively associated with the topic, DE analysis in FastTopics produce genes with both positive and negative LFC. We believe the genes with negative LFC are also important for inferring dynamics, and they were kept for downstream analysis.

1.2 Topic modeling evaluation metrics

The first metric we considered uses average distance among topics to measure stability [4]:

$$\text{correlation}(k, k') = \frac{\sum_{m=1}^M F_{mk} F_{mk'}}{\sqrt{\sum_{m=1}^M (F_{mk})^2} \sqrt{\sum_{m=1}^M (F_{mk'})^2}} \quad (6)$$

$$\text{ave_cosine_dis} = \frac{\sum_{k=1}^K \sum_{k'=k+1}^K \text{correlation}(k, k')}{K(K-1)/2}$$

where $\text{correlation}(k, k')$ is the standard cosine distance between topics k and k' . A smaller ave_cosine_dis indicates more stability.

The second metric we considered is the information divergence between all pairs of topics [5]:

$$D(k||k') = \frac{1}{2} \sum_{m=1}^M F_{mk} \log\left(\frac{F_{mk}}{F_{mk'}}\right) + F_{mk'} \log\left(\frac{F_{mk'}}{F_{mk}}\right) \quad (7)$$

$$\text{ave_info_dis} = \frac{\sum_{k,k'} D(k||k')}{K(K-1)}$$

where $D(k||k')$ is the Jensen-Shannon distance between two topics. A bigger ave_info_dis indicates more independence and more information in the topic model.

We also tested a few coherence measures, which are based on the point-wise mutual information (PMI) of the top-weighted or highest ranked (by log-fold change) topic-specific genes [6]:

$$\text{PMI}(g_m, g_{m'}) = \log \frac{P(g_m, g_{m'}) + \epsilon}{P(g_m) \cdot P(g_{m'})} \quad (8)$$

where $P(g_m, g_{m'})$ is the joint probability of observing genes g_m and $g_{m'}$ in a cell, and $P(g_m)$ and $P(g_{m'})$ are the marginal probabilities of observing gene g_m and $g_{m'}$ in a cell, respectively; ϵ is a small number (e.g. 10^{-12}) to prevent the PMI from reaching 0. For the top N genes (either topic-specific or highest weighted), the *UCI coherence* is calculated as:

$$\text{Coh}_{UCI} = \frac{2}{N(N-1)} \sum_{m=1}^{N-1} \sum_{m'=i+1}^N \text{PMI}(g_m, g_{m'}) \quad (9)$$

Small values of $|C_{UCI}|$ indicate higher topic coherence and higher probability that the top genes are co-expressed.

To prevent overfitting, we also considered the Akaike information criterion (AIC) and the Bayesian information criterion (BIC):

$$AIC = 2 \cdot M \cdot (K - 1) - 2 \cdot \mathcal{L}_K \quad (10)$$

$$BIC = (K - 1) \cdot M \cdot \log(C) - 2 \cdot \mathcal{L}_K \quad (11)$$

where \mathcal{L}_K is the log-likelihood of the model for K topics.

1.3 RNA velocity parameter estimation via the one-state model

The one-state transcription model is governed by the following master equation:

$$\begin{aligned} \frac{\partial p(u, s, t)}{\partial t} = & \alpha \left[p(u-1, s, t) - p(u, s, t) \right] \\ & + \beta \left[(u+1)p(u+1, s-1, t) - up(u, s, t) \right] \\ & + \gamma \left[(s+1)p(u, s+1, t) - sp(u, s, t) \right] \end{aligned} \quad (12)$$

in which α is the rate of transcription, β is the splicing rate, and γ is the degradation rate. The steady-state distribution when $\beta \neq \gamma$ is the product of two independent Poisson distributions for u and s respectively [7]:

$$p(u, s) = \frac{\left(\frac{\alpha}{\beta}\right)^u \left(\frac{\alpha}{\gamma}\right)^s}{u!s!} \exp\left(-\frac{\alpha}{\beta} - \frac{\alpha}{\gamma}\right) \quad (13)$$

Then the log likelihood for observing C cells at steady state with unspliced and spliced counts $\{u_i, s_i\}_{i=1}^C$ conditioned on a set of kinetic parameters is:

$$\begin{aligned} \mathcal{L}(\{u_i, s_i\}_{i=1}^C | \alpha, \beta, \gamma) &= \ln \prod_{i=1}^C p(u_i, c_i) \\ &= -C \left(\frac{\alpha}{\beta}\right) + \ln\left(\frac{\alpha}{\beta}\right) \sum_{i=1}^C u_i - \sum_{i=1}^C \ln(u_i!) \\ &\quad - C \left(\frac{\alpha}{\gamma}\right) + \ln\left(\frac{\alpha}{\gamma}\right) \sum_{i=1}^C s_i - \sum_{i=1}^C \ln(s_i!) \end{aligned} \quad (14)$$

The maximum likelihood estimate of γ/β is:

$$\begin{aligned}
0 &= \frac{\partial}{\partial(\alpha/\gamma)} \mathcal{L}(\{u_i, s_i\}_{i=1}^C | \alpha, \beta, \gamma) \\
\frac{\alpha}{\gamma} &= \frac{1}{C} \sum_{i=1}^C s_i = \langle s \rangle \\
\text{similarly, } \frac{\alpha}{\beta} &= \langle u \rangle \\
\rightarrow \frac{\gamma}{\beta} &= \gamma' = \frac{\langle u \rangle}{\langle s \rangle}
\end{aligned} \tag{15}$$

where $\langle \cdot \rangle$ denotes expectation, and $\langle s \rangle$ and $\langle u \rangle$ are the average abundance of u and s over all cells in steady-state.

1.4 RNA velocity parameter estimation via the geometric burst model

To verify the correctness of this estimation approach, we compared the simulated joint distribution for parameters $k_{\text{on}} = 0.5$, $b = 5$, $\gamma = 3$ with the joint distribution simulated from the inferred parameters; the two distributions are nearly identical (Supp. Fig. S1a). To visualize the path of the optimization, we plotted it on the KL divergence landscape of k_{on} versus b for γ fixed at 0.3, and the KL divergence landscape of γ versus b with for k_{on} fixed at 0.5; we observed that the optimizer ended very close to the ground truth (Supp. Fig. S1b, c).

We aimed to find choices for the number of simulation steps (or reactions) and the maximum number of iterations for the Nelder-Mead optimization that perform well in this inference scheme. We considered a total of 27 parameter combinations across different dynamical regimes in which average mRNA abundances vary over the span of several orders of magnitudes (Supp. Fig. S1d). First, we fixed the number of simulation steps at 5×10^5 and analyzed how different choices for the maximum number of optimization iterations affect the performance. For each choice, we simulated 10 replicates of the 27 combinations and computed the average KL divergence within each replicate (Supp. Fig. S1e). We observed that the performance stopped improving when more than 50 iterations were used. Similarly, we fixed the maximum number of iterations at 50 and examined how different choices for the number of simulation steps affected inference. We observed that the performance improvement became negligible when more than 5×10^5 steps were used (Supp. Fig. S1f). Therefore, we chose 5×10^5 as the number of reactions and 50 as the maximum number iterations for parameter inference. In both settings, the bimodality of the average KL divergence that we observed may be due to the optimizers getting stuck in local minima. To ameliorate this effect in the analysis of real datasets, we perform 5 independent optimizations for each gene and, for downstream analysis, use the set of parameters that corresponds to the lowest KL-divergence.

The joint distribution of spliced and unspliced counts was typically computed from the size-normalized data (not on the log scale), rounded to the nearest integer. Under the steady-state assumption, a time-invariant splicing rate $\beta = 1$ was assumed for all genes; k_{on} , b , and γ were estimated for each gene. To illustrate the parameter inference scheme on real data, we used the observed distributions of *Grin2b* from the granule mature cells in the dentate gyrus dataset [8],

which serve as a proxy for steady state since these cells are terminally differentiated. By capturing the diffusiveness and low expression regimes of the distribution more accurately, the geometric burst model recovers a joint distribution that is closer than the one inferred via the one-state model to the observed distribution (Supp. Fig. S1g). The inferred parameters from the burst model for *Grin2b* are located within a regime of low KL divergence as shown on the KL divergence landscapes (Supp. Fig. S1h). We performed the analogous analysis for the gene *Btbd9* and observed similar results (Supp. Fig. S1i,j).

1.5 Determination of topic-associated cells

For each topic within a given dataset, topic-associated cells were defined as cells above a certain topic weight. Kinetic parameters for topic-specific genes were then inferred from topic-associated cells, which in this model represent a topic-specific steady state. In the *scVelo* implementation, the up-regulation and down-regulation steady states for a given gene are modeled as the top right corner and the bottom left corner of the phase plot, respectively. The exact determination is dependent on arbitrary expression thresholds; the default setting uses the 5th and 95th percentiles. Instead of assuming each gene has its own set of steady-state cells, *TopicVelo* uses topic association as a criterion for choosing topic-specific steady state cells, which tends to be more robust and biologically meaningful because the genes in a given topic have correlated expression patterns.

1.6 RNA velocity evaluation metrics

To provide some context for the evaluation measures in this manuscript, we summarize and compare the various quantitative measures that can be used to assess the quality of RNA velocity estimates and the accuracy of RNA-velocity-enabled trajectory inference.

- *Velocity Coherence.* The *scVelo* paper uses *velocity coherence*, which captures the similarity of velocities within cell neighborhoods, defined as:

$$\text{Coh}_i^v = \frac{1}{|N(i)|} \sum_{j \in N(i)} \cos(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j) \quad (16)$$

where $N(i)$ is the neighborhood of cell i computed from the global PCA. Velocity coherence has been used as a confidence metric to quantify the quality of velocity estimation.

We adapted velocity coherence in the framework of *TopicVelo* by introducing a topic-informed velocity coherence measure (distinct from a topic coherence measure used in text mining literature), to assess *TopicVelo* results. Briefly, we compute topic-specific coherence for local neighborhoods among topic-associated cells, and then compute a topic-weighted sum of topic-specific coherence scores for each cell i :

$$\text{Coh}_i^{tv} = \sum_{k=1}^K \tilde{L}_{ik} \frac{1}{|N_k(i)|} \sum_{j \in N_k(i)} \cos(\tilde{\mathbf{v}}_{i,k}, \tilde{\mathbf{v}}_{j,k}) \quad (17)$$

While velocity coherence measures capture consistency, they do not reflect information about trajectory correctness. Hence, RNA velocity estimates may have high velocity coherence and still lead to biologically incorrect transitions, while low velocity coherence may still correspond to biologically correct transitions. Moreover, in a multifurcating progenitor population, it is unclear how much velocity coherence should be expected.

- *Stationary Distribution.* Without loss of generality, suppose the integrated transition matrix T is irreducible. By construction, T is positive-recurrent and aperiodic. The stationary distribution π is the solution to the eigenvalue problem $\pi^T = \pi^T T$. The stationary distribution of the transition matrix is helpful for assessing the overall directionality of the transition matrix and for identifying terminal states. Therefore, it is particularly useful when strong differentiation signals are expected. However, the stationary distribution does not reveal dynamical information about the trajectories. It has been used in *CellRank* [9] to assess the transition matrix.
- *Pseudotime.* *CellRank* and *scVelo* both use *pseudotime* to evaluate results. They construct pseudotime by first computing the stationary distribution of the transition matrix to identify terminal states (those with high values), and then inferring trajectories from terminal states to ancestral states using a biased diffusion process with reversed transitions. The authors have also used a similar, biased-diffusion approach to pseudotime inference, without velocities [10]. Pseudotime analysis is restricted to one-dimensional view of the dynamics, which may be too limited in some contexts. In contrast, mean first-passage time (defined next) offers a richer summary of the dynamics.
- *Mean First-Passage Time (MFPT).*

For an ergodic Markov chain characterized by a transition matrix T , the *first passage time* from a source state i to a target state j is defined as

$$\{\min t > 0 : W(t) = j | W(0) = i\}, \quad (18)$$

where $W(t)$ is the state at time t . Because there may be many paths from i to j of varying travel times and probabilities, the mean value of the first passage time (MFPT) from i to j , $\mathcal{M}_{i,j}$, is a useful statistic to assess the expected timescale for this transition. For scRNA-seq data, MFPT can nicely capture the timescale of cell-state transitions. For example, a common problem of interest is to identify the likely progenitors of a set of terminally differentiated cells, which may be identified via the stationary distribution π or from prior knowledge. By computing the MFPTs to this set of terminally differentiated cells from all other subpopulations, one may identify subpopulations with lower MFPTs as reasonable candidates for progenitors, which may then be validated experimentally. MFPT is used in this case as an alternative to traditional pseudotime inference.

The framework of MFPT is more general than pseudotime and applicable in settings in which a traditional notion of pseudotime does not apply. One such example is transdifferentiation. The pseudotimes for two distinct, terminally differentiated cell populations from a common progenitor reflect the differentiation process from the progenitor but not information about

possible transitions between the two cell states. However, MFPTs to the two cell states can reveal potential transdifferentiation processes. Furthermore, the asymmetry of MFPT (i.e., $\mathcal{M}_{i,j} \neq \mathcal{M}_{j,i}$ in general) makes it useful for studying asymmetric bidirectional transitions, such as the ILC2-quiescent transition in the skin ILCs data and in cellular reprogramming.

While our paper introduces the use of MFPT to analyze RNA velocity, the conceptual framework has been applied previously in the context of trajectory inference. Weinreb, et al. [11] use MFPT to predict a temporal ordering based on a population balance analysis, with known starting and end points and without velocity inference. Qiu, Zhang, et al. [12] compute a *least action path* (LAP) between two subsets of cells, which, the authors note, is associated with an action that contributes the most to the MFPT and hence should be similar to computing the MFPT when there is negligible noise around the LAP. The LAP framework would need substantial adaptation to be implemented for *TopicVelo*, where velocities are computed on overlapping, topic-specific sub-graphs for topic-specific gene sets. We instead compute the MFPT matrix \mathcal{M} directly from the integrated transition matrix by solving the following matrix equation:

$$(I - T)\mathcal{M} = J - T(I \odot (\pi \mathbf{1}^T))^{-1} \quad (19)$$

where I is the identity matrix, J is a matrix of all ones, and \odot is the Hadamard (entrywise) matrix product. For a set A of target cells, the vector $\mathcal{M}_A = \{\mathcal{M}_{i,A}\}_{i=1}^C$ of MFPTs to A , is defined as:

$$\mathcal{M}_{i,A} = \begin{cases} 0 & \forall i \in A \\ \frac{1}{|A|} \sum_{j \in A} \mathcal{M}_{i,j} & \forall i \notin A \end{cases} \quad (20)$$

In analyses that compare MFPTs to a target set A across velocity methods (e.g., Fig. 2h), we rescale each vector \mathcal{M}_A by the median of nonzero elements in \mathcal{M}_A to obtain the *rescaled mean first-passage time* (rMFPT).

Note that if extremely distinct populations are contained within a given dataset, T may be reducible (i.e., T contains multiple disconnected components), in which case the stationary distribution and MFPTs must be analyzed within individual irreducible components. This was not an issue in the datasets analyzed here.

- *Cross-Boundary Correctness.* A *cross-boundary correctness* measure is used by Qiao, et al. [13], though not by *scVelo*, *CellRank*, or *velocity*, to assess whether flow across a cell-type boundary is biased in the expected direction. The implementation computes cosine similarity between velocity vectors and target cells in the two-dimensional embedding used for visualization. This approach is similar to that used for streamline visualization, which is known to be problematic [14, 15]. The measure also assumes that flows across cell-type boundaries should be strongly biased, which may not hold in some biological settings. Hence, we defined *relative flux* as an alternate version of this measure.

- *Relative Flux.* Inspired by the notion of cross-boundary correctness, relative flux has the advantage of being independent of the visualization embedding. A high value in this measure indicates that flows across cell-type boundaries are strongly biased. Briefly, for two neighborhoods A and B of cells that share at least one edge in the k_G -NN graph, we define the *flux* from A to B as the sum of all transition probabilities from A to B :

$$\text{flux}(A, B) = \sum_{\{c \in A, c' \in B\}} T_{cc'} \quad (21)$$

and the relative flux from A to B as:

$$\text{RF}_{A \rightarrow B} = \frac{\text{flux}(A, B) - \text{flux}(B, A)}{\text{flux}(A, B) + \text{flux}(B, A)} \quad (22)$$

If the populations of A and B are very unbalanced due to non-biological reasons, it may be more appropriate to define a normalized flux by normalizing the flux by the size of the population at the boundary. Analogously, a normalized relative flux can be defined.

1.7 scRNA-seq data analysis

Human hematopoiesis scNT-seq data analysis. This dataset contains count matrices with or without metabolic labels. We focused our analysis on the latter. We used default settings for the *scVelo* stochastic and dynamical models to infer velocities and obtained streamline embeddings. The dynamical model gave streamline embeddings more consistent with biological expectation. We then applied topic modeling with 8 topics and identified topic-specific genes. A gene was selected if its lfsr is less than or equal to 0.001 and the LFC is at least 0.5 in absolute value. We removed topic-specific genes from topics 5 and 6 for downstream analysis because they are strongly associated with minichromosomal and ribosomal genes and are ubiquitously expressed. We selected topic-associated cells as those with weights above the 35th-percentile for each topic to infer the kinetic parameters and global transition matrix. We compared topic-specific velocities to the global velocity inferred by the *scVelo* dynamical model.

For the normalization study, we only changed the matrix used for velocity parameter inference. For studying the performance of *TopicVelo* with different number of topics, we use the same selection criterion for topic genes (i.e. same lfsr and LFC thresholds, and the removal of minichromosomal and ribosomal gene programs) and topic cells (cells above the 35th-percentile for each topic). For the burst model only ablative approach, we inferred kinetic parameters on all 2,000 highly variable genes and then kept genes whose KL-divergence was less than 0.005 (313 genes) for downstream analysis. For the combination of *scVelo* and topic modeling ablative approach, the only change was the use of *scVelo* for parameter inference of topic genes within topic-associated cells.

Mouse gastrulation data analysis. After standard preprocessing, we applied the *scVelo* stochastic model, rather than the dynamical model, which is more prone to wrongly inferring transcriptional boosting as down-regulation [16]. We performed topic modeling with 2 topics, which resulted in

a large number (1,961) of topic-specific genes. A gene was selected if its *lfsr* is less than or equal to 0.001 and the LFC is at least 0.5 in absolute value. To focus the analysis on the genes with the best signal given the low unspliced/spliced ratio in this dataset, we removed genes for which the ratio of the maximum of spliced counts to the maximum of unspliced counts is greater than 10 or less than 0.01, as the ratios of spliced over unspliced counts in this dataset tend to be very high (Supp. Fig. S21a). Furthermore, we observed that this dataset contains many genes that are highly expressed in specific cell subsets, which standard size-normalization steps do not handle well, potentially leading to improper assessment of the dynamics of lowly expressed genes. To avoid this possibility, we used the raw counts (rather than the size-normalized counts) to infer kinetic parameters. We used cells with topic weights above the 40th percentile for each topic to infer the kinetic parameters and global transition matrix. We compared topic-specific velocities to the global velocity inferred by the *scVelo* dynamical model.

Human bone marrow data analysis. After standard preprocessing, we applied the *scVelo* stochastic model. Like the mouse gastrulation data, this data contain genes with transcriptional bursting patterns that are not handled well by the dynamical model [17]. We performed topic modeling with 10 topics. For each topic, a gene was selected if its *lfsr* is less than or equal to 0.001 and the LFC is at least 0.5 in absolute value. We used cells with topic weights above the 65th-percentile for each topic to infer the kinetic parameters and to construct the global transition matrix. For the burst model only ablative approach, we inferred kinetic parameters on all 2,000 highly variable genes and then kept genes whose KL-divergence was less than 0.005 (353 genes) for downstream analysis. For the combination of *scVelo* and topic modeling ablative approach, the only change was the use of *scVelo* for parameter inference of topic genes within topic-associated cells.

Mouse ILCs data analysis. This dataset contains data from five different days. We only used the data collected on day 3 for the main figure. To maintain comparisons with the previous analysis, we focused on the highly variable genes as determined previously by a variance stabilizing transformation [18]. Doublet detection was done with *Scrublet* [19] where the number of UMIs is the sum of spliced and unspliced transcripts. We still performed gene filtering, and then performed topic modeling with 10 topics. For each topic, a gene was selected if its *lfsr* is less than or equal to 0.01 and the LFC is at least 0.5 in absolute value. For the global analysis, we used the 88nd-percentile for each topic to infer the kinetic parameters and global transition matrix. For analyzing velocities, we focused on comparing the *TopicVelo* topic-specific velocities and the global velocity from the *scVelo* dynamical model; in this dataset, the stochastic and dynamical *scVelo* models give very similar results. For the mean-first passage time analysis, the target cells were selected as the cells above 95th-percentile from topics 4, 6, and 9 respectively.

We used 9 topics for day 0 and 10 topics for days 1, 2, and 4. The criterion for the selection of topic genes, topic cells, and target groups for the MFPT analyses were the same for these days as for day 3. For the identification of paths of the bidirectional quiescent-ILC2 transitions, for each day of days 0 and 3, we computed positive edge weights for an adjacency matrix on the cells from that day by taking the element-wise negative log likelihood of the transition probabilities in the integrated transition matrix. Then we ran Dijkstra's shortest-path algorithm [20] for the pairs (i, j) of cells such that i has an ILC2-like topic weight in the top 0.2% and j has a quiescent-like topic

weight in the top 0.2%. There are 11 in each group from day 0 and 14 cells in each group from day 3, and hence 121 paths for each direction on day 0 and 196 paths for each direction on day 3.

Mouse dentate gyrus data analysis. After standard preprocessing, we applied the *scVelo* dynamical model. Though we allowed more EM updates than the default parameters for recovering the dynamics and computed the corresponding transition matrix multiple times, we were unable to reproduce the oligodendrocyte lineages [21]. We performed topic modeling with 10 topics. For each topic, a gene was selected if its lfsr is less than or equal to 0.001 and the LFC is at least 0.5 in absolute value. Topic-specific genes for topic 1 had a high level of expression across nearly all cells and were therefore removed for downstream analysis. We used cells with topic weights above the 80th, 80th, 90th, 70th, 80th, 60th, 60th, 90th, and 70th-percentile for topic 0 and topics 2–9, respectively, to infer the kinetic parameters and to integrate the transition matrices into the global transition matrix.

Mouse pancreas data analysis. After standard preprocessing, we applied the *scVelo* dynamical model. We performed topic modeling with 6 topics. For each topic, a gene was selected if its lfsr is less than or equal to 0.001 and the LFC is at least 0.5 in absolute value. Topic-specific genes for topics 0 and 1 are primarily related to cell cycle and were excluded. We used cells with topic weights above the 65th-percentile for topics 0 and topics 3–5, respectively, to infer the kinetic parameters and integrated transition matrix.

2 Run-time and memory complexity analysis

For topic modeling, we use the FastTopics software package [1], which has complexity $O(K(N_0 + C + M))$ for sparse matrices and $O(KCM)$ otherwise, where K is the number of topics, C is the number of cells, M is the number of genes, and N_0 is the number of non-zeros in the matrix.

For the velocity parameter inference, the runtime complexity is $O(C + Mn_{\text{maxiters}}n_{\text{simsteps}})$, where n_{maxiters} is the maximum number of stochastic simulations and n_{simsteps} is the number of samples from each trajectory. As shown in Supp. Fig. 1, we found empirically that the performance gain (reduction in KL-divergence) plateaued for $n_{\text{maxiters}} > 50$. We also observed that $n_{\text{simsteps}} > (5 \times 10^5)$ did not result in noteworthy reductions in KL-divergence. Therefore, in practice, the upper bound of the run time is effectively $O(M)$.

For each topic, computing the topic-specific transition matrix requires computing the cosine correlation between the velocities of a cell and the transcriptional profiles of its k_G neighbors in the k_G -nearest-neighbor graph. (We use the notation k_G here to indicate the number of nearest neighbors, to distinguish it from K , the number of topics.) Hence, the run-time complexity is $O(k_G CM)$ per topic, and the run-time complexity for constructing the integrated matrix is $O(Kk_G CM)$.

The memory complexity for the stochastic simulation and inference is $O(u_{\text{max}}s_{\text{max}})$, where u_{max} and s_{max} are the maximum number of unspliced and spliced transcripts, respectively, over all cells and genes. The memory complexity of computing the topic-specific velocity matrices is $O(KCM)$, while the integrated transition matrix has memory complexity $O(Kk_G C)$. Both are on par with existing methods.

To empirically test run times, for each dataset, we computed the topic models and differential expression using FastTopics, as well as the RNA velocity estimates and integrated transition matrix

using *TopicVelo*, on a basic laptop with a good amount of RAM (specifically, a 2018 MacBook Pro with a 2.9 GHz 6-Core Intel Core i9 processor and 32GB 2400MHz DDR4 Memory). The total computation time needed for *TopicVelo* ranged from about 6 minutes to nearly 1 hour (Table 1). We note that, as is the case for any per-gene or per-topic inference, the *TopicVelo* computations can easily be made much more efficient by parallelizing them.

dataset	C	M	K	topic modeling	<i>TopicVelo</i>
scNT-seq	1,947	383	8	2m35s	6m12s
mouse gastrulation	9,815	1,961	2	15m25s	57m28s
human bone marrow	5,780	612	10	10m41s	25m31s
mouse ILC day 3	6,525	790	10	13m19s	34m52s
dentate gyrus	2,930	445	10	4m27s	10m46s
pancreas	3,696	1,379	6	3m52s	19m24s

Table 1: Run times on datasets of varying numbers of cells C , genes M , and topics K . Topic modeling was performed with FastTopics.

3 *TopicVelo* identifies key lineages during dentate gyrus development

In the dentate gyrus data [8], we observed that *scVelo* recovers the main trajectory involving the astrocytes and granule lineages, though it fails to infer an established transition from oligodendrocyte precursor cells (OPCs) to oligodendrocytes (OLs) (Supp. Fig. S7a). With 10 topics, *TopicVelo* infers a similar main trajectory and also captures the transition in the smaller oligodendrocyte lineage clearly (Supp. Fig. S7b).

Topic modeling results reveal a gene program (topic 2) associated with OPC and OL cells, and other programs corresponding to different stages of differentiation and terminal cell types (Supp. Fig. S7c, Supp. Fig. S8, Supp. Table 1). Topic-2 specific genes with high log-fold changes include *Ppp1rl14a*, which was previously observed to be up-regulated during OL differentiation [22], and *Enpp2*, which has been shown to regulate oligodendrocyte differentiation *in vivo* in the hindbrain of developing zebrafish [23] (Supp. Fig. S7c, d, Supp. Table 1). The observed expression in mouse also suggests an up-regulation of *Enpp2* along the OPC-OL differentiation trajectory, which is more consistent with the positive velocities in OPCs inferred by *TopicVelo* than with the negative velocities inferred by *scVelo* in this population (Supp. Fig. S7d).

We considered multiple quantitative measures of the results. The velocity coherence suggests greater uniformity in *scVelo* results, though it is not clear what coherence should be biologically expected (Supp. Fig. S7e). The stationary distributions for both methods assign approximately half of the probability to terminally differentiated cell types, though they vary in which types, and to later time points (Supp. Fig. S7f). Relative flux differs across the results from each method, with *TopicVelo* generally inferring less extreme values of relative flux than *scVelo* for several expected transitions, including less negative flux (i.e., reverse flow) for the immature to mature granule transition (Supp. Fig. S7g), consistent with the greater proportion of mature granule cells in the stationary distribution (Supp. Fig. S7f).

4 *TopicVelo* recovers multifurcation of murine pancreatic endocrinogenesis

In the pancreatic endocrinogenesis data [24], we observed that *scVelo* reveals cell cycling in ductal progenitors and predicts their eventual commitments into α , β , and ϵ , but not δ , cells (Supp.

Fig. S9a). Applying *TopicVelo*, we obtained largely consistent qualitative results and also captured the differentiation of pre-endocrine cells into δ cells at the streamline level (Supp. Fig. S9b).

The six topics in the model correspond to previously identified populations, including β , ϵ , *Ngn3*-expressing progenitor, and ductal cells (Supp. Fig. S10a–d). Topic 4 is strongly associated with both α and δ cells, and is characterized by genes, such as *Gcg* and *Sst*, which are well-known markers of α and δ cells, respectively [25] (Supp. Fig. S9c), as well as genes, such as *Ppy* and (unspliced) *Pcdh15*, that are expressed in both populations (Supp. Fig. S10e). Another topic-4 specific gene is *Spock3* (Supp. Table 1), whose specific expression in human δ cells has been observed [26]. Though *scVelo* suggests very little up-regulation of *Spock3* in the δ lineage, *TopicVelo* predicts specific, positive velocities in those cells (Supp. Fig. S9d). Pre-endocrine-associated topic 5 also features specific expression of *Chgb* and *Fev* (Supp. Fig. S9e), genes previously reported to be enriched at the branch point of α and β commitments [27].

A few quantitative measures were computed to assess the quality of RNA velocity inference and trajectory inference. Velocity coherence looks reasonable for *TopicVelo* though its distribution is wider than that of *scVelo* (Supp. Fig. S9f). Whereas the stationary distribution from *scVelo* is largely dominated by β cells, *TopicVelo* assigns the majority of probability to α cells; both are terminally differentiated cell types (Supp. Fig. S9g). *TopicVelo* uniquely assigns a sizable probability to the terminally differentiated ϵ cells (Supp. Fig. S9g). As in the dentate gyrus, *TopicVelo* generally infers less extreme values of relative flux than *scVelo*, which may reflect the actual biology, including possible *beta*-cell self-renewal [28], or technical aspects, such as less smoothing or difficulty distinguishing pre-endocrine cells and islet cells at the pre-defined discrete transition boundaries (Supp. Fig. S9h).

5 Doublet detection in ILCs data

Doublet detection is challenging, particularly in cells undergoing transcriptional changes. We performed a few analyses to try to assess whether the thin bridge between ILC2-like and ILC3-like cells likely represents doublets or some of the ILC2-ILC3 transitioning cells (Supp. Fig. S11a). First, we observed that cells in the thin bridge are similar to many ILC3-like cells in having relatively high numbers of UMIs detected, perhaps due to an inflammatory response (Supp. Fig. S11b). It has been suggested that the number of genes detected may be relatively high in heterotypic doublets [29]. However, we observed that, in contrast to cells expressing a proliferation program (those in the thin branch on the right side of the embedding), the ILC3-associated cells and cells in the bridge do not have particularly high numbers of genes detected (Supp. Fig. S11c). In fact, the bridge cells have somewhat low numbers of genes detected, relative to the number of UMIs (Supp. Fig. S11d). Finally, we performed doublet detection with *Scrublet* [19], which did not identify cells in the thin bridge as particularly likely doublets (Supp. Fig. S11e). We also note that doublet detection algorithms risk erroneously identifying transitioning cells as doublets, due to the heterogeneous transcriptional profiles of these cells [29]. Taken together, this evidence suggests that the cells in the thin bridge cells are unlikely to be doublets. The thin bridge may be an artifact of the visualization or represent cells that are transitioning differently from other cells, i.e., perhaps more rapidly.

6 Potential bi-directional ILC2-quiescent transition in skin ILCs

We investigated whether the potential transitions in each direction likely proceed along the same path. Specifically, for each of day 0 and day 3, we computed a non-negative adjacency matrix for a graph on the cells, with edge weight from cell i to j equal to the element-wise negative log likelihood of the transition probability of i to j in the integrated transition matrix. Then we ran Dijkstra's shortest-path algorithm [20] on all pairs (i, j) of cells such that i has an ILC2-like topic weight in the top 0.2% and j has a quiescent-like topic weight in the top 0.2%. (There are 11 and 14 cells in each group, and hence 121 and 196 paths for each direction, for days 0 and 3, respectively.) For each intermediate cell in a shortest path, we calculated the overlap of its neighborhood in the original k_G -NN graph with other intermediate cells from paths going in each direction. While intermediate cells in each direction do have overlapping neighborhoods, we found a statistically significant difference in the overlap with cells from paths going the same direction versus those going the opposite direction, suggesting that the transitions in each direction associate with different regions of transcriptomic space (Supp. Fig. S15).

7 Comparison of ILCs dynamics across days

The analysis in the original study [18] uses cells from all days and finds convergent responses to ILC3-like cells from other subpopulations. While *TopicVelo* applied to all cells gave results consistent with those in the original study, we were interested in whether *TopicVelo* could identify key transitions from just a single day, as a demonstration of the power of *TopicVelo*-inferred RNA velocity. Hence, our analysis focuses on data just from day 3, which is characterized by the emergence of relatively large numbers of ILC3-like cells (as calculated below).

To unveil day-specific dynamics, we used *TopicVelo* on the cells from each day separately. We performed topic modeling with 9 topics for day 0 and with 10 topics for days 1, 2, and 4. For day 0, we identified a topic that is strongly associated with quiescent cells and two topics that contain genes associated in the original paper with ILC2-like cells (Supp. Fig. S16a). For days 1, 2, and 4, we identified ILC3-like, quiescent-like, and ILC2-like programs consistent with the analogous topics observed in day 3 (Supp. Fig. S16b–d). Using the largest topic weight as a proxy for cell type, we calculated the percentage of ILC3-like cells in each day: day 3 had the largest percentage at 9.3%, compared with 0 on day 0, 2.1% on day 1, 3.4% on day 2, and 7.0% on day 4. This variation suggests a possible crescendo of the ILC3-associated inflammatory response, culminating on day 3 and potentially beginning to decrease on day 4.

Analysis of the dynamics for each day shows expected consistency across days 0, 1, and 3 (Supp. Fig. S17). Although the day 0 streamlines do not reveal much about transitions between the quiescent and ILC2 cells, the MFPT analysis shows that quiescent-like cells may indeed transition to ILC2-like cells, consistent with previous *in vitro* experiments (Supp. Fig. S17a,b). Streamlines and MFPT analysis from days 1 and day 3 reveal the quiescent-to-ILC3, ILC2-to-ILC3, quiescent-to-ILC2, and ILC2-to-quiescent transitions, consistent with the findings in the original study (Supp. Fig. S17c–f).

Analysis of data from days 2 and 4 reveals overall less dynamic information. This could reflect biological reality, but could also result from a combination of technical artifacts and erroneous inference. (Notably, days 2 and 4 also have the smallest numbers of cells after quality control: 4,058 and 4,953 cells, respectively, compared with 5,290, 6,051, and 6,525 cells for days 0, 1, and 3,

respectively, which could possibly indicate quality issues specific to days 2 and 4.) While day 2 and day 4 streamlines suggest the ILC2-to-ILC3 and quiescent-to-ILC3 transitions, the MFPT analyses only support the quiescent-to-ILC3 transition in day 2 and the ILC2-to-ILC3 transition in day 4 (Supp. Fig. S18a–d). As for all days, the ILC3-like cells at day 4 appear to be a strong sink, exemplified by long MFPTs to other states; however, in contrast to other days, the quiescent population shows a lower tendency to transition at day 4 (Supp. Fig. S18d). These results could suggest that by day 4, the quiescent cells are no longer involved in cell-state transitions.

8 Analysis of the impact of topic number choice

In a thorough analysis of the scNT-seq data, we tested procedures to identify the regime for an appropriate choice of the number of topics. We then carried out both qualitative and quantitative analysis to confirm that key results from *TopicVelo* are not sensitive to the exact choice of topic number, provided the chosen number is within the reasonable regime.

First, for each value of K , $2 \leq K \leq 20$, we fit a topic model and computed topic coherence (Methods, SI Appendix 1). For each value of K , we repeated this process 50 times, to account for randomness in the topic model fitting, and computed the mean and standard deviation (Supp. Fig. S19a). We observed that an “elbow” in the plot of topic coherence as a function of topic number occurs at $K = 8$, suggesting that key granular processes should be captured for $K \geq 8$. We carried out the robustness study for values of K , $6 \leq K \leq 10$. Though average topic coherence values continue to rise as K increases beyond that, the number of topic-associated genes becomes too small to capture detailed topic-associated dynamics (Supp. Fig. S19b). Velocity coherence is highest for $K = 8$ and $K = 10$ (Supp. Fig. S19c); $K = 8$ is the value used in the main analysis (Fig. 2).

In the streamline visualizations, all topic number choices ($6 \leq K \leq 10$) recover the differentiation trajectories to terminally differentiated cell types. However, the results for $K = 6, 7, 9$ do not display the transitions from HSCs to the MEP-like and GMP-like intermediate cell states, which are seen for $K = 8, 10$ (Supp. Fig. S19d). In fact, none of the topic number choices results in a clear directionality for those transitions based on relative fluxes (Supp. Fig. S19e). The low relative flux values may be due to the fact that these three subpopulations are very transcriptionally similar, as demonstrated by their shared high weights in topics 2, 5, and 6 for $K = 8$. Though the model for $K = 6$ assigns high stationary probability to progenitor states, similar to *scVelo*, the models for $K = 7, 8, 9, 10$ all assign the majority of the stationary probability to terminal states (Supp. Fig. S19f).

For megakaryocytes as the target group, we would expect the MFPT values from HSCs and MEP-like cells to be lower than those from other populations. For the models with $K = 6$ and $K = 7$, we did not find this pattern, but for $K = 8$ and $K = 9$, we did find statistically significantly lower MFPT values for HSCs and MEP-like cells (Supp. Fig. S19g). For $K = 10$, only the HSC values were statistically significantly lower, though quite a few MEP-like cells (in the bottom of the violin plot) show short transition times to megakaryocytes. The higher median MFPT value for MEP-like cells for $K = 10$ may be due to *TopicVelo* predicting that MEP-like cells are more likely to transition to other terminal states, such as erythroid cells.

9 Algorithmic ablation analysis of *TopicVelo*

We performed an ablation study to understand the individual contributions of the following ablative approaches:

1. Bursty transcriptional model only (“Burst”)
2. Topic modeling + kinetic model used in *scVelo* (“*scVelo*+TM”).

Our analysis focused on the scNT-seq and human bone marrow datasets. The results suggest that each component improves trajectory inference, albeit in different ways.

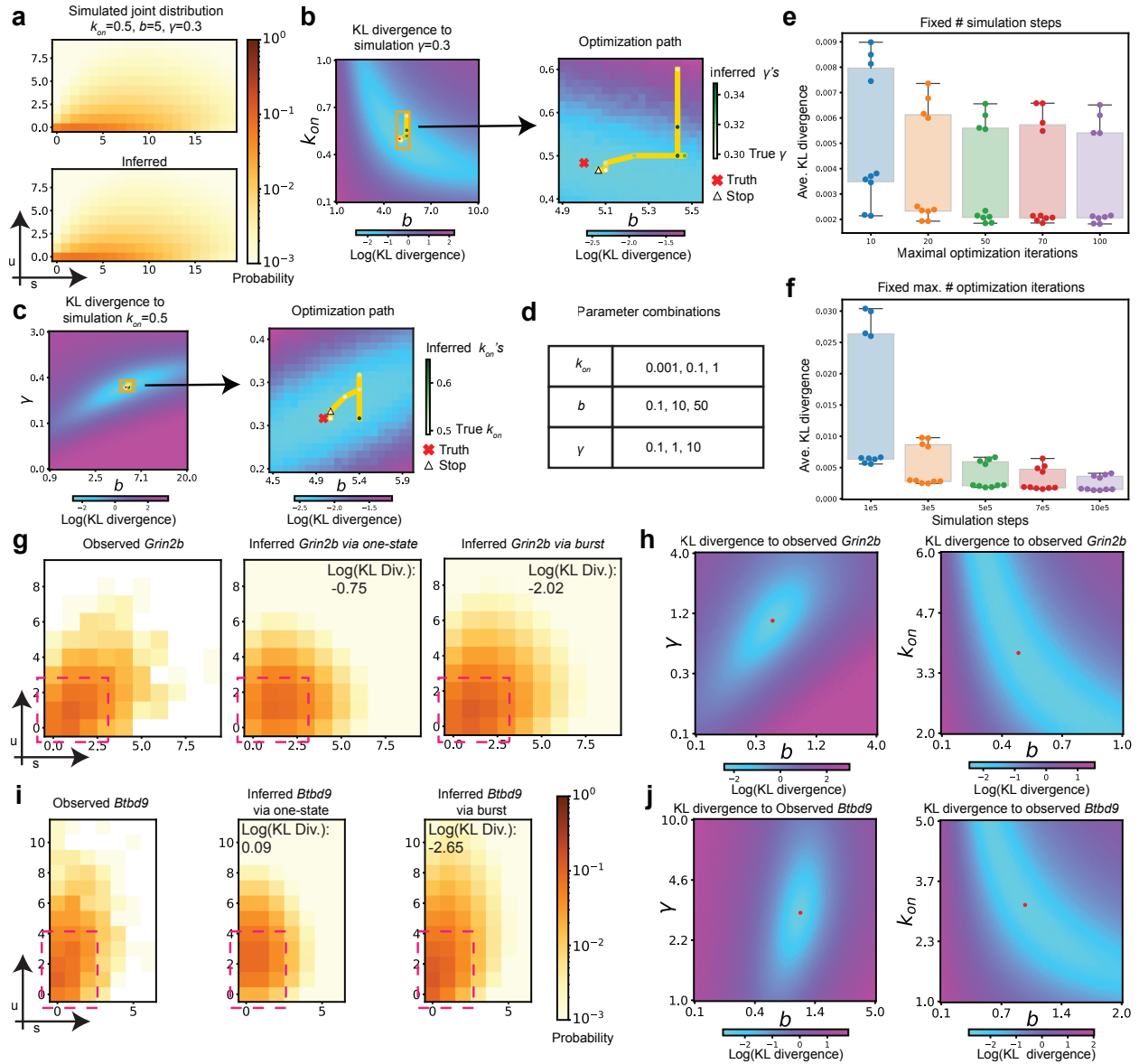
For the scNT-seq data, streamline visualizations suggest that, compared with using *scVelo* only, Burst additionally identifies the differentiation of early erythroid cells to mature erythroid cells, while *scVelo*+TM additionally identifies the differentiation of early megakaryocytes to mature megakaryocytes and that of early basophils to mature basophils (Supp. Fig. S20a). For both approaches, the stationary distributions consist mostly of progenitor states (Supp. Fig. S20b). Moreover, unlike *TopicVelo*, the two ablative approaches fail to improve relative flux (Supp. Fig. S20c). We note, however, the relative flux we computed only reflects transitions between identified cell types and cannot account for other transitions, such as the maturation of erythroid cells. Velocity coherence is also not improved (Supp. Fig. S20d). Finally, in contrast to *TopicVelo*, neither ablative approach results in mean first-passage times to megakaryocytes that are statistically significantly lower for the progenitor populations than for the others (Supp. Fig. S20e). These results suggest that in some cases, the algorithmic components synergistically improve trajectory inference.

For the human bone marrow data, the streamline visualizations show that Burst improves upon *scVelo* results by identifying the differentiation trajectory of early HSCs to precursors and the erythroid lineage. *scVelo*+TM approach appears different but not obviously better than using *scVelo* alone (Supp. Fig. S20f). The stationary distribution reflects these observations; like *TopicVelo*, Burst assigns high weights to mature erythroid cells, while *scVelo*+TM assigns sizeable probability to the precursors, monocytes, DCs, early erythroid cells, and megakaryocytes (Supp. Fig. S20g). For relative flux, *scVelo* has notably negative (biologically erroneous) values for nearly all transitions, while *TopicVelo* and the ablative approaches have positive values for at least half of the eight known transitions (Supp. Fig. S20h). *TopicVelo* has positive values for the most transitions (six out of eight). The ablative approaches do not lead to higher velocity coherence (Supp. Fig. S20i).

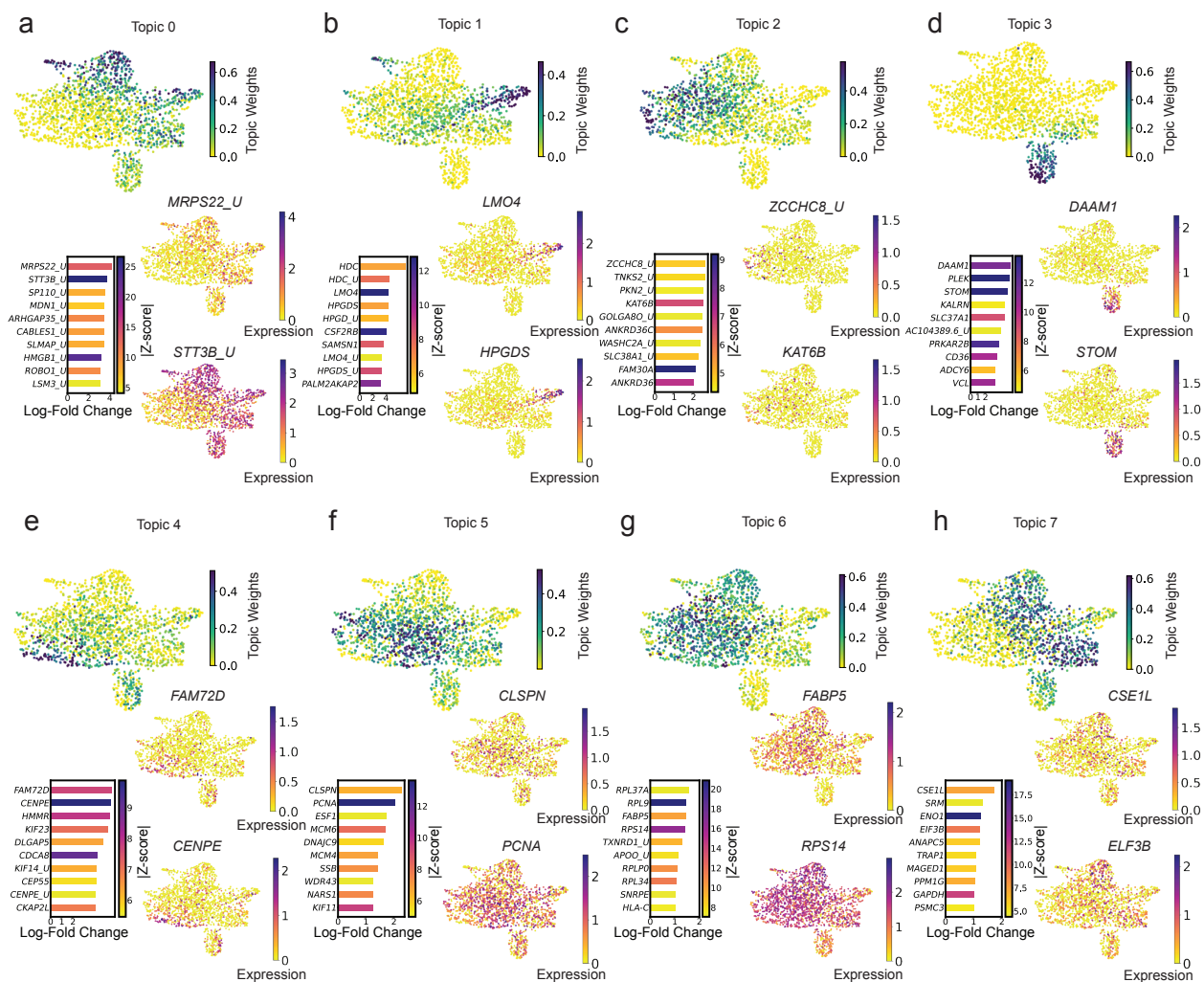
Supporting Information References

1. Carbonetto, P., Sarkar, A., Wang, Z. & Stephens, M. Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv* **2105.13440**. arXiv: 2105.13440. <https://arxiv.org/abs/2105.13440> (2021).
2. Carbonetto, P. *et al.* Interpreting structure in sequence count data with differential expression analysis allowing for grades of membership. *bioRxiv*. eprint: <https://www.biorxiv.org/content/early/2023/03/06/2023.03.03.531029.full.pdf>. <https://www.biorxiv.org/content/early/2023/03/06/2023.03.03.531029> (2023).
3. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18**, 275–294. issn: 1465-4644. eprint: <https://academic.oup.com/biostatistics/article-pdf/18/2/275/11057424/kxw041.pdf>. <https://doi.org/10.1093/biostatistics/kxw041> (Oct. 2016).
4. Cao, J., Xia, T., Li, J., Zhang, Y. & Tang, S. A density-based method for adaptive LDA model selection. *Neurocomputing* **72**. Advances in Machine Learning and Computational Intelligence, 1775–1781. issn: 0925-2312. <https://www.sciencedirect.com/science/article/pii/S092523120800372X> (2009).
5. Deveaud, R., SanJuan, E. & Bellot, P. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* **17**, 61–84 (2014).
6. Röder, M., Both, A. & Hinneburg, A. *Exploring the Space of Topic Coherence Measures in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (Association for Computing Machinery, Shanghai, China, 2015), 399–408. isbn: 9781450333177. <https://doi.org/10.1145/2684822.2685324>.
7. Li, T., Shi, J., Wu, Y. & Zhou, P. On the Mathematics of RNA Velocity I: Theoretical Analysis. *CSIAM Transactions on Applied Mathematics* **2**, 1–55. issn: 2708-0579. http://global-sci.org/intro/article_detail/csiam-am/18653.html (2021).
8. Hochgerner, H., Zeisel, A., Lönnerberg, P. & Linnarsson, S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nature Neuroscience* **21**, 290–299. issn: 1546-1726. <https://doi.org/10.1038/s41593-017-0056-2> (Feb. 2018).
9. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nature Methods* **19**, 159–170. <https://doi.org/10.1038/s41592-021-01346-6> (2022).
10. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science (New York, N.Y.)* **360**. 29700225[pmid], eaar3131. issn: 1095-9203. <https://doi.org/10.1126/science.aar3131> (June 2018).
11. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences* **115**, E2467–E2476. issn: 0027-8424. eprint: <https://www.pnas.org/content/115/10/E2467.full.pdf>. <https://www.pnas.org/content/115/10/E2467> (2018).
12. Qiu, X. *et al.* Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690–711.e45. issn: 0092-8674. <https://www.sciencedirect.com/science/article/pii/S0092867421015774> (2022).
13. Qiao, C. & Huang, Y. Representation learning of RNA velocity reveals robust cell transitions. *Proceedings of the National Academy of Sciences* **118**, e2105859118. eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.2105859118>. <https://www.pnas.org/doi/abs/10.1073/pnas.2105859118> (2021).
14. Gorin, G., Fang, M., Chari, T. & Pachter, L. RNA velocity unraveled. *PLOS Computational Biology* **18**, 1–55. <https://doi.org/10.1371/journal.pcbi.1010492> (Sept. 2022).
15. Zheng, S. C., Stein-O'Brien, G., Boukas, L., Goff, L. A. & Hansen, K. D. Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biology* **24** (2023).
16. Barile, M. *et al.* Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biology* **22**. issn: 1474-760X (July 2021).

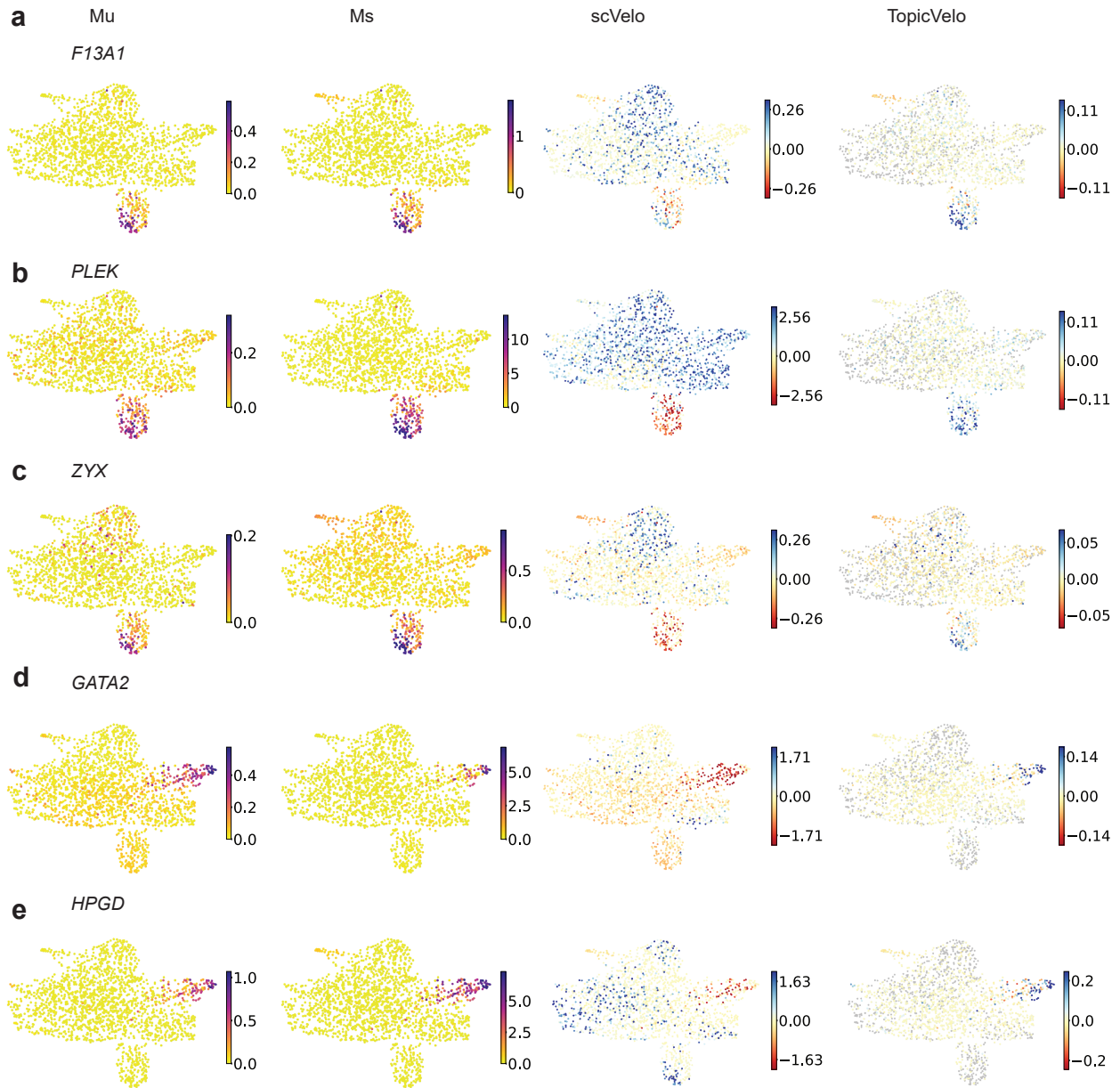
17. Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity-current challenges and future perspectives. *Molecular Systems Biology* **17**, e10282. ISSN: 1744-4292. <https://doi.org/10.15252/msb.202110282> (Aug. 2021).
18. Bielecki, P. *et al.* Skin-resident innate lymphoid cells converge on a pathogenic effector state. *Nature* **592**, 128–132. ISSN: 1476-4687. <https://doi.org/10.1038/s41586-021-03188-w> (Apr. 2021).
19. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019).
20. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik* **1**, 269–271 (1959).
21. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* **38**, 1546–1696 (2020).
22. Cahoy, J. D. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci* **28**, 264–278 (2008).
23. Yuelling, L. W., Waggener, C. T., Afshari, F. S., Lister, J. A. & Fuss, B. Autotaxin/ENPP2 regulates oligodendrocyte differentiation in vivo in the developing zebrafish hindbrain. *Glia* **60**, 1605–1618. <https://doi.org/10.1002/glia.22381> (July 2012).
24. Bastidas-Ponce, A. *et al.* Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* **146**. dev173849. ISSN: 0950-1991. eprint: <https://journals.biologists.com/dev/article-pdf/146/12/dev173849/1857007/dev173849.pdf>. <https://doi.org/10.1242/dev.173849> (June 2019).
25. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385–394.e3 (2016).
26. Xin, Y. *et al.* RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metabolism* **24**, 608–615. ISSN: 1550-4131. <https://www.sciencedirect.com/science/article/pii/S155041311630434X> (2016).
27. Byrnes, L. E. *et al.* Lineage dynamics of murine pancreatic development at single-cell resolution. *Nature Communications* **9**, 3922. ISSN: 2041-1723. <https://doi.org/10.1038/s41467-018-06176-3> (Sept. 2018).
28. Teta, M., Rankin, M. M., Long, S. Y., Stein, G. M. & Kushner, J. A. Growth and Regeneration of Adult β Cells Does Not Involve Specialized Progenitors. *Developmental Cell* **12**, 817–826. ISSN: 1534-5807. <https://www.sciencedirect.com/science/article/pii/S153458070700158X> (2007).
29. Xi, N. M. & Li, J. J. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Syst.* **12**, 176–194.e6 (2021).



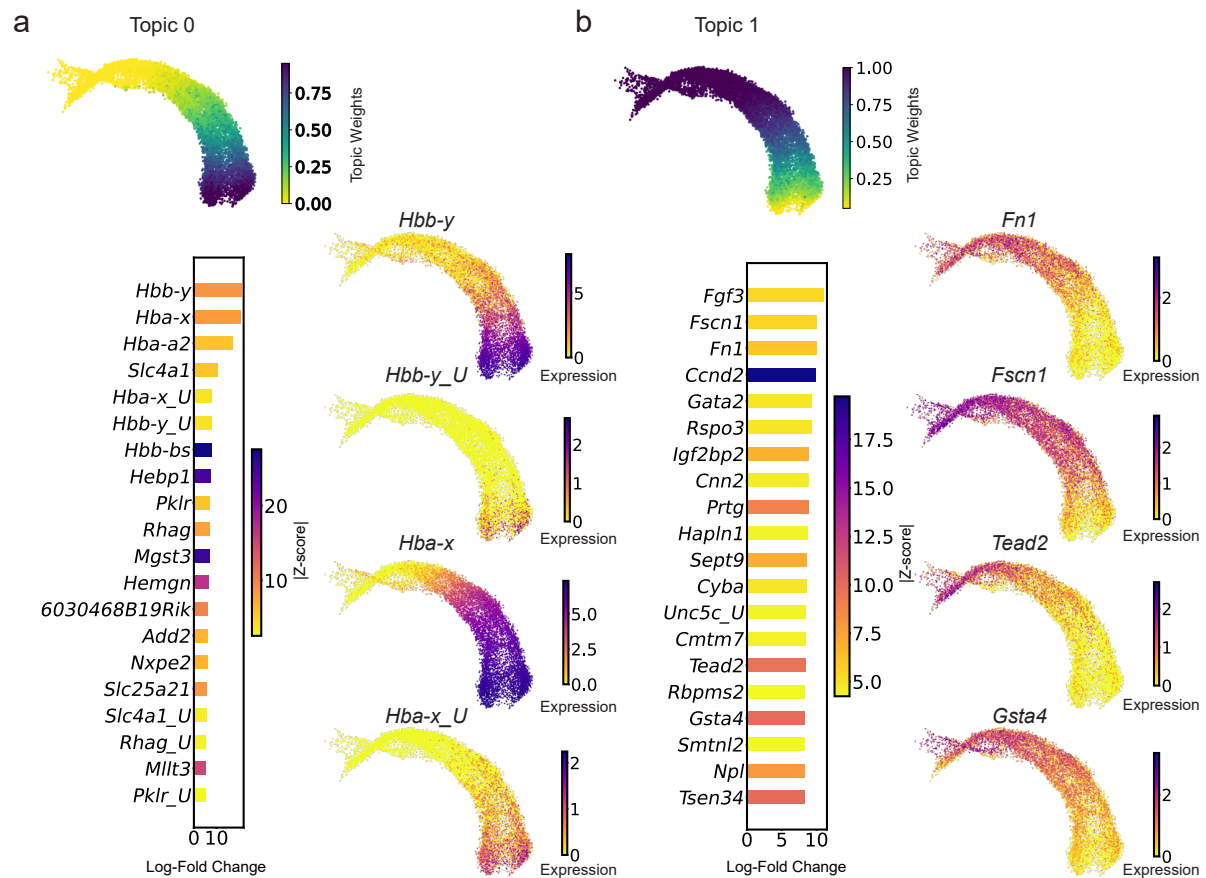
Supporting Figure S1: The geometric burst model more accurately recovers experimental distributions than the one-state model. **a**, An example joint distribution of spliced (s) and unspliced (u) transcript counts, as simulated by the Gillespie algorithm for the geometric burst model with fixed parameters $k_{on} = 0.5$, $b = 5$, and $\gamma = 0.3$ (top), and with maximum-likelihood estimates (MLEs) inferred from the simulated data (bottom). **b**, Log KL divergence (color) landscape for $\gamma = 0.3$ over a range of values of b and k_{on} , with close-up (right) of restricted range (orange box, left). True parameter values marked by red cross; optimization path (yellow) shown across iterations (points, colored by inferred γ value) to end point (triangle). **c**, Analogous to **b**, for $k_{on} = 0.5$ and varying b and γ . **d**, Table of 27 parameter combinations used to assess effects of the number of simulation steps and maximum number of optimization iterations. **e**, For a fixed number ($5 \cdot 10^5$) of simulation steps and varying maximum number of iterations (x axis, color), bar plots show the average KL divergence across the 27 parameter combinations in **d**, for 10 replicates (points). **f**, Analogous to **e**, for a fixed maximum number (50) of optimization iterations. **g**, The joint distribution of the gene *Grin2b* in the granule mature cells in the dentate gyrus dataset [8], as observed (left), computed from MLEs for the one-state model (middle), and simulated from the MLEs for the geometric burst model (right), annotated by log KL divergence from the observed. The burst model better matches the observed values in the region where probability mass is concentrated (pink dashed box). **h**, For the burst model, plots show the log KL divergence to the observed distributions of *Grin2b* for b versus γ with k_{on} fixed to the MLE (left), and b versus k_{on} with γ fixed to the MLE (right), with optimizer end points indicated (red dots). **i, j**, Analogous to **g, h**, but for *Btbd9*. The log of KL divergence is base 10.



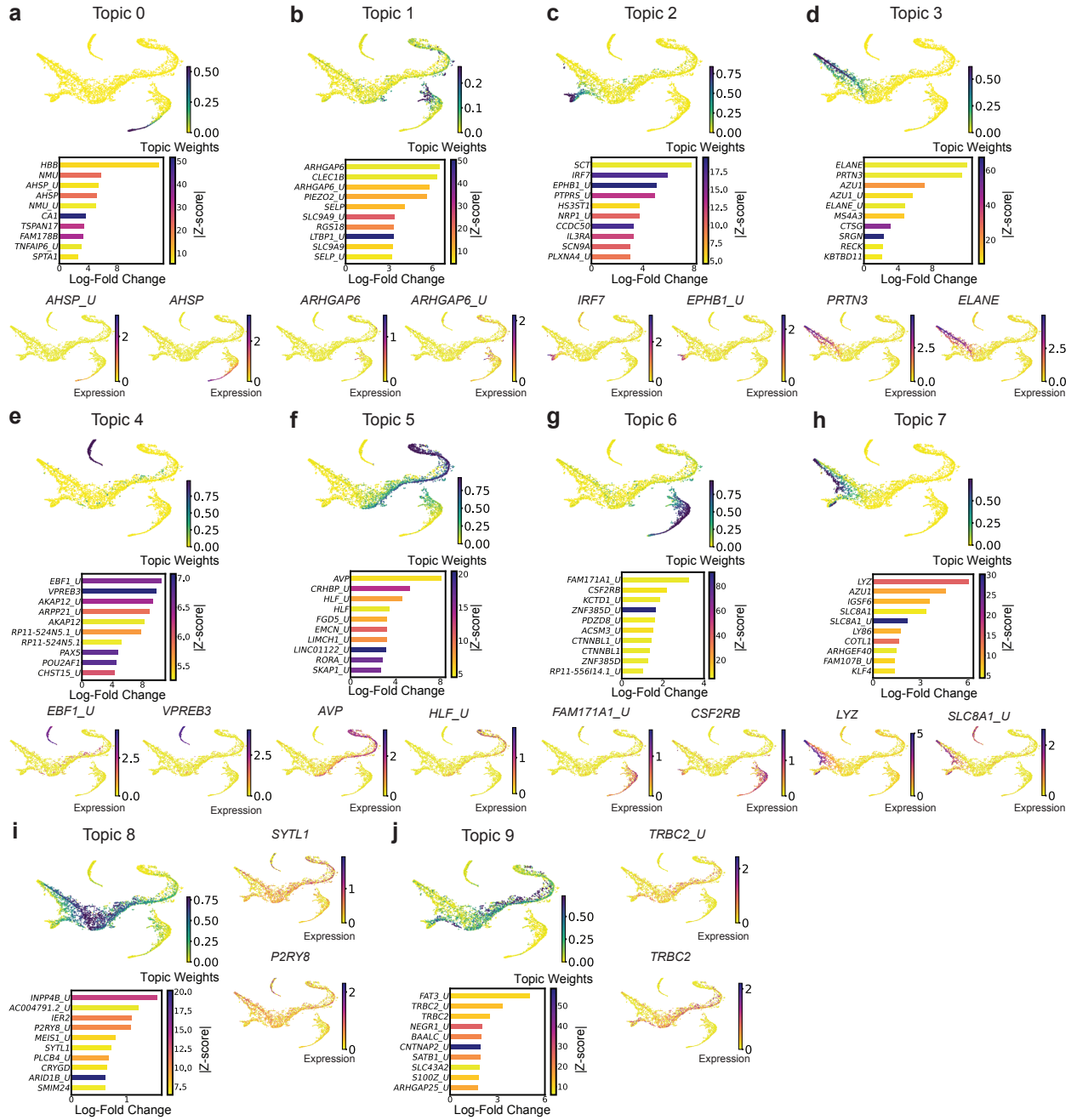
Supporting Figure S2: Topic modeling analysis of scNT-seq data from human hematopoiesis. **a**, For topic 0, UMAP plots shows cells colored by topic weights (top) and by log-normalized expression of topic-specific genes (bottom right); bar plot (bottom left) shows top 10 topic-specific genes ranked by log-fold change (x axis) and colored by absolute value of z-score; ‘_U’ indicates unspliced transcripts. **b–h**, Analogous to **a**, for topics 1–7, respectively.



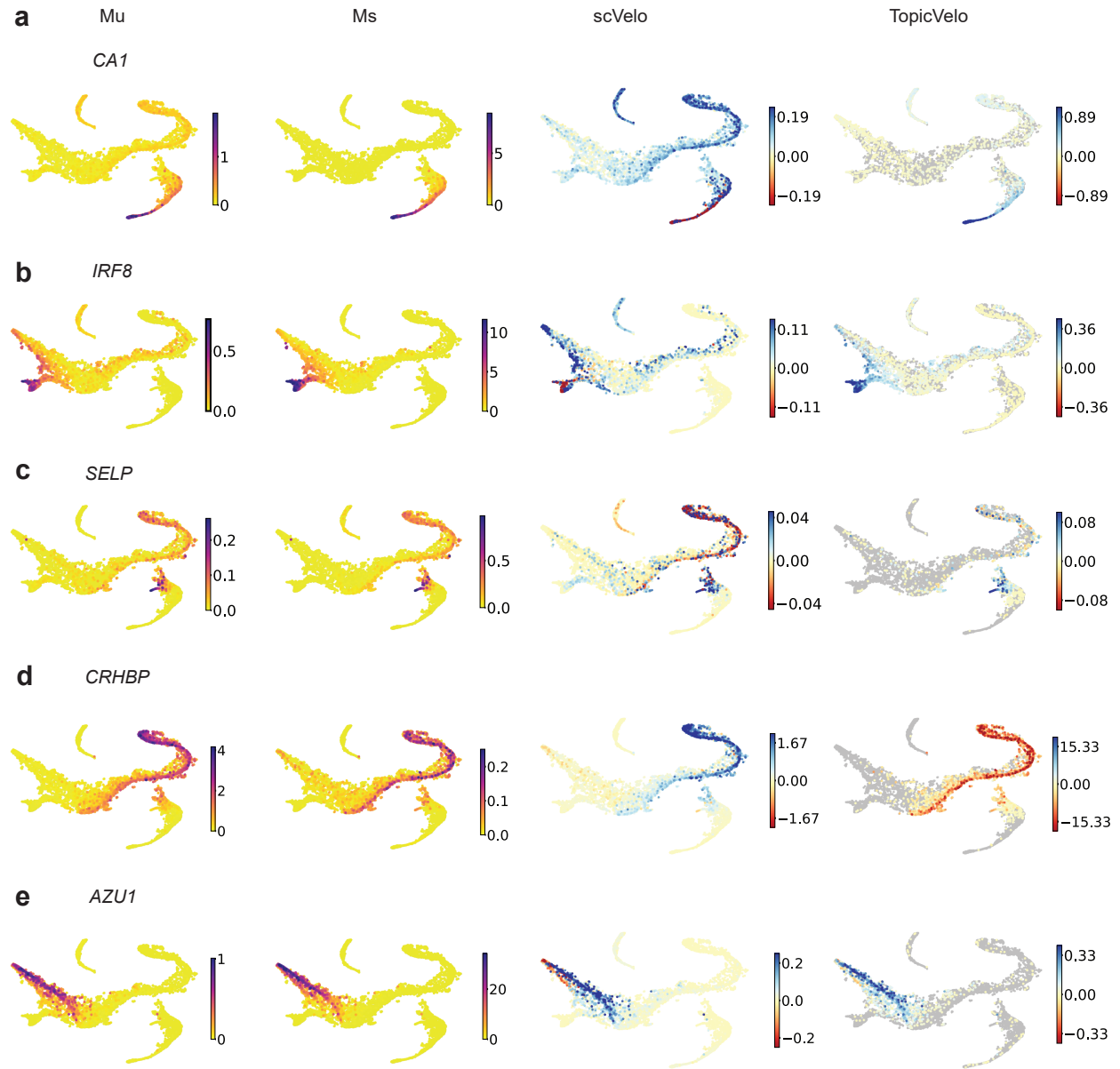
Supporting Figure S3: *TopicVelo* recovers more biologically plausible velocity estimates than those of *scVelo* for the scNT-seq data. **a–c,** Analysis of topic-3 specific genes. UMAP plots colored by smoothed size-normalized counts of unspliced (Mu) (far left) and spliced (Ms) (middle left) transcripts, and by velocities inferred by *scVelo* (middle right) and *TopicVelo* (far right), for the genes *F13A1* (**a**), *PLEK* (**b**), and *ZYX* (**c**). **d, e,** Analysis of topic-1 specific genes *GATA2* and *HPGD*, analogous to a–c.



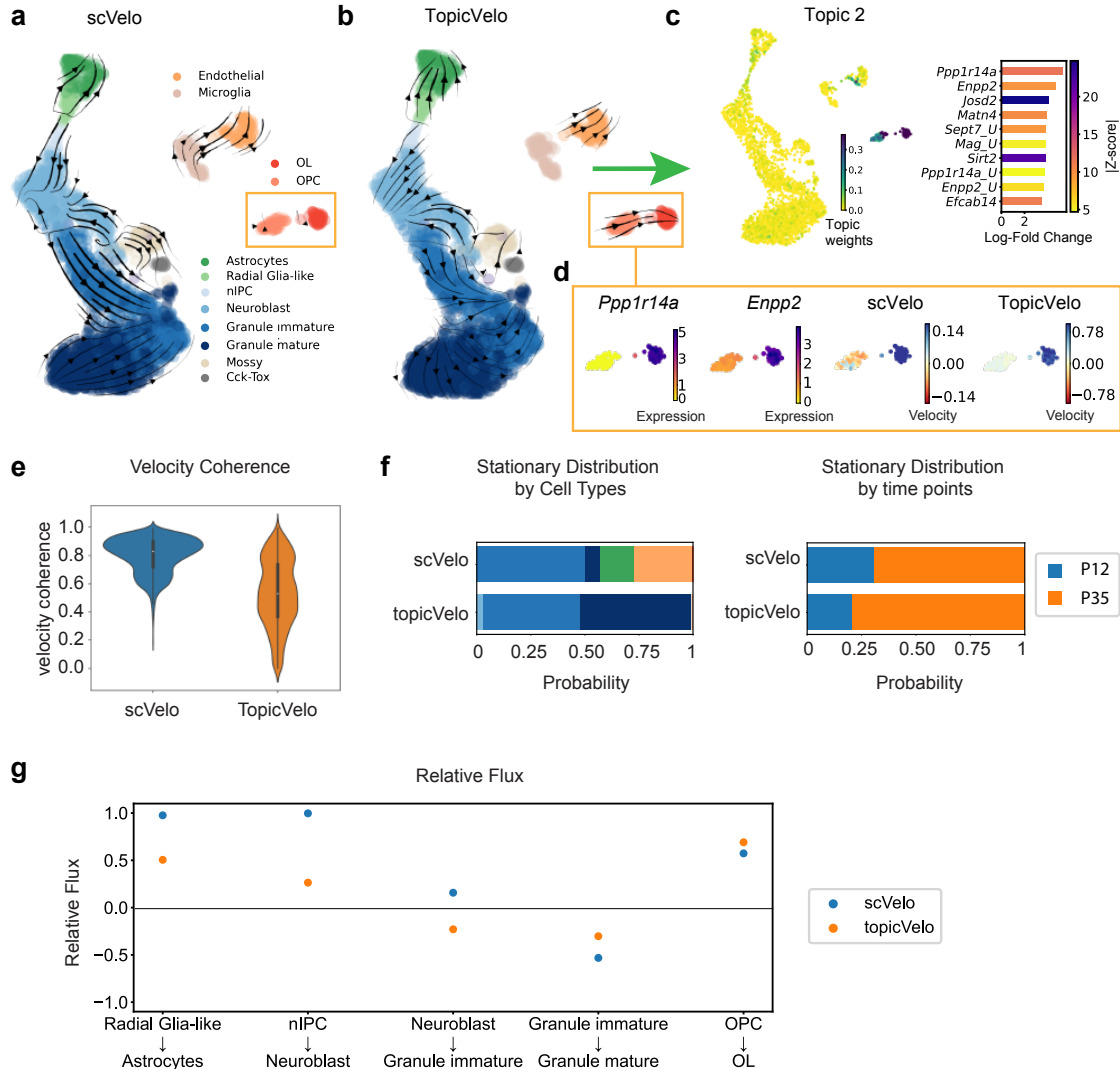
Supporting Figure S4: Topic modeling of the gastrulation data reveals key genes underlying the differentiation of blood progenitors to erythroid. **a**, For topic 0, UMAP shows cells colored by topic weights (top) and by log-normalized expression of topic-specific genes (bottom right); bar plot (bottom left) shows top 20 topic-specific genes ranked by log-fold change (x axis) and colored by absolute value of z-score; '_U' indicates unspliced transcripts. **b**, Analogous to a, for topic 1.



Supporting Figure S5: Topic modeling of the human bone marrow data provides insights into the different stages of differentiation along all lineages. **a**, For topic 0, *t*-SNE plots shows cells colored by topic weights (top) and by log-normalized expression of topic-specific genes (bottom); bar plot (middle) shows top 10 topic-specific genes ranked by log-fold change (x axis) and colored by absolute value of z-score; ‘_U’ indicates unspliced transcripts. **b–j**, Analogous to **a**, for topics 1–9, respectively. Topics are generally associated with annotated stages of development: topic 0 (mature erythroid), 1 (megakaryocyte), 2 (dendritic cells), 3 (one lineage of monocytes), 4 (common lymphoid progenitors), 5 (hematopoietic stem cells), 6 (erythroid), 7 (another lineage of monocytes), 8 (precursors to monocytes), 9 (hematopoietic stem cells and precursors to dendritic cells).

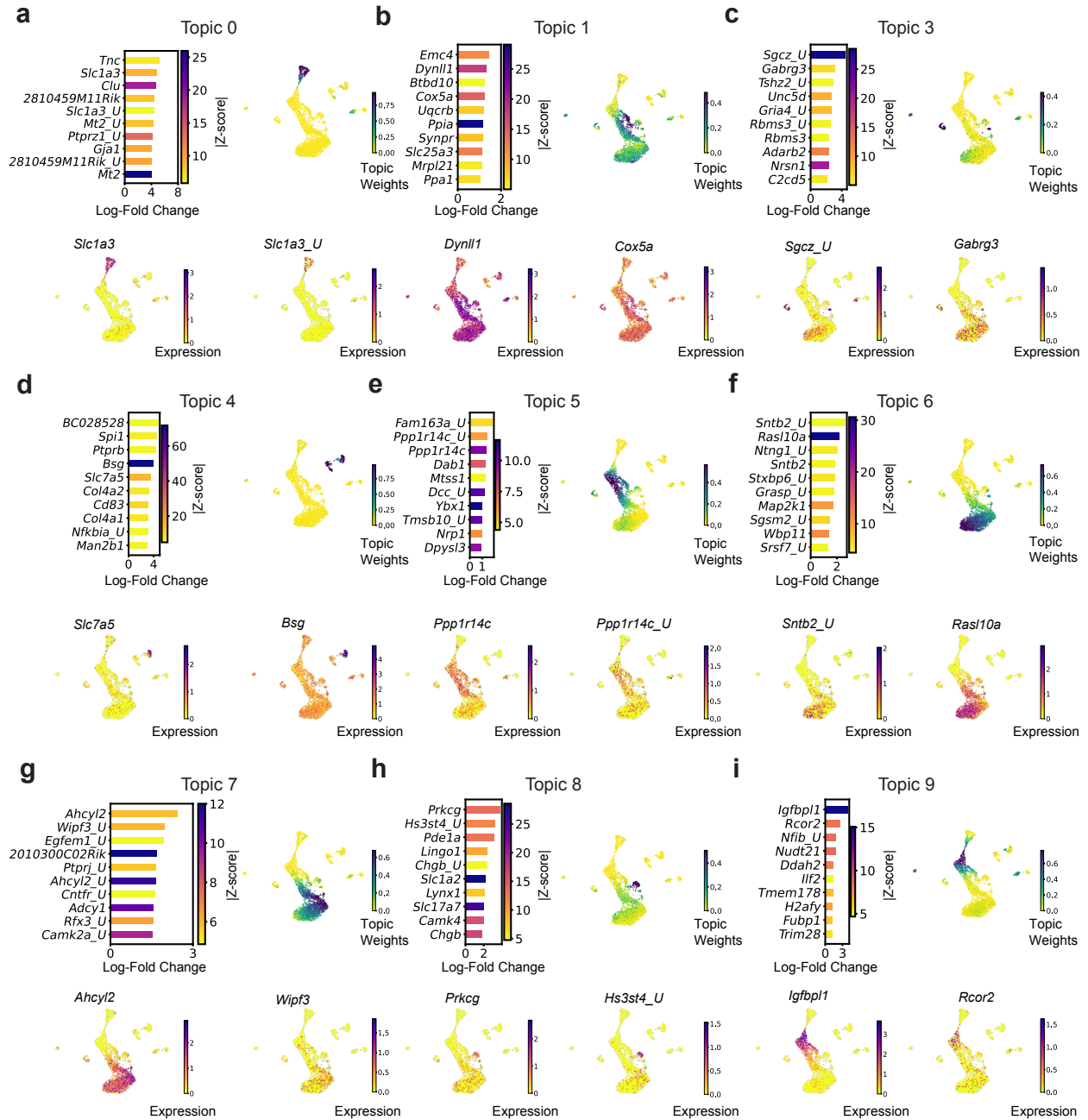


Supporting Figure S6: *TopicVelo* recovers better biologically supported velocities than *scVelo* for the human bone marrow data. **a**, For topic-0 specific gene *CA1*, UMAP plots are colored by smoothed size-normalized counts of unspliced (Mu) (far left) and spliced (Ms) (middle left) transcripts, and by velocities inferred by *scVelo* (middle right) and *TopicVelo* (far right). **b–e**, Analogous to **a**, for topic-2 specific gene *IRF8* (Supp. Table 1), topic-1 specific gene *SELP*, topic-5 specific gene *CRHBP*, and topic-7 specific gene *AZU1*, respectively.

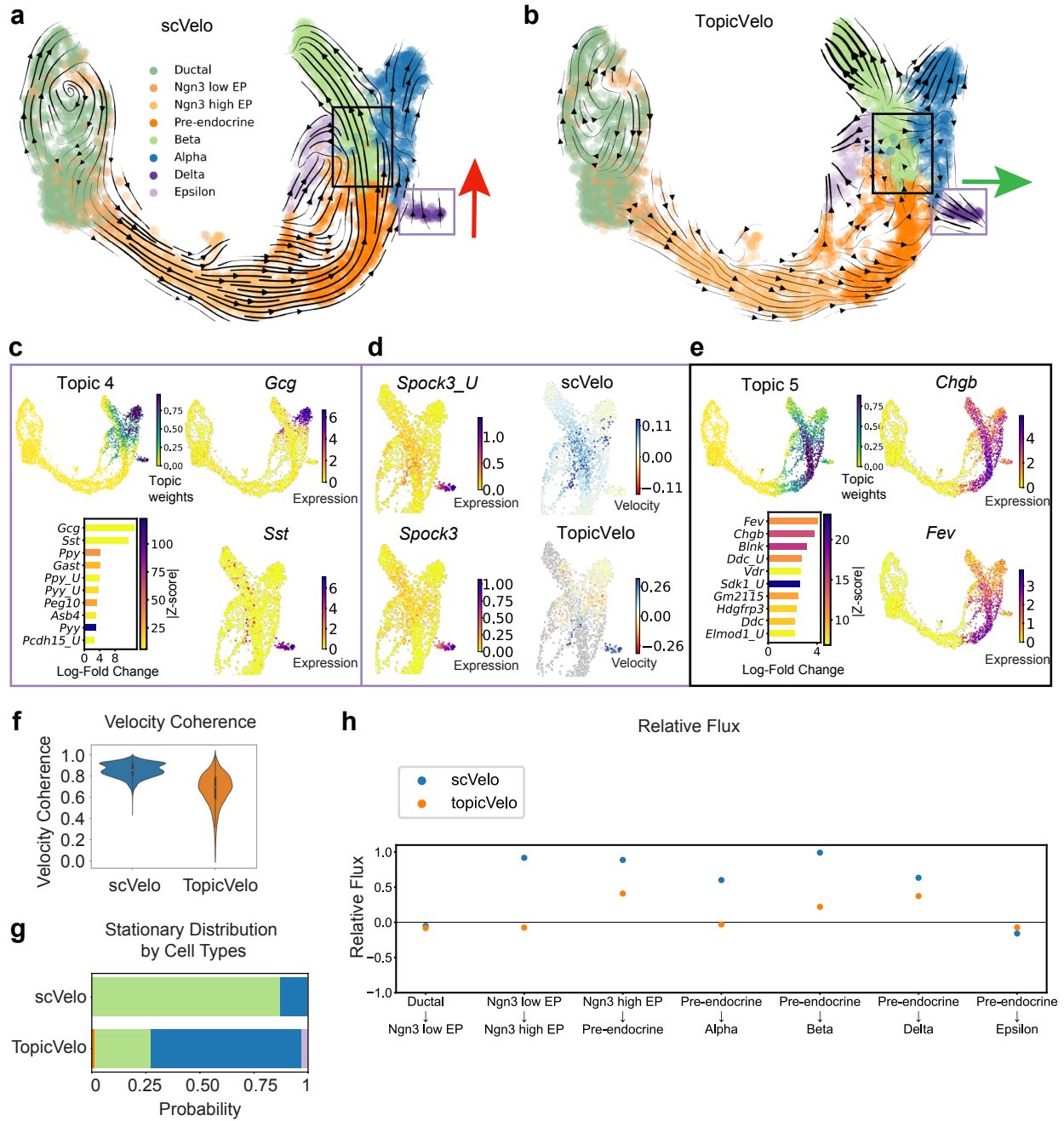


Supporting Figure S7: TopicVelo correctly infers transitions, including between small cell populations, during dentate gyrus development.

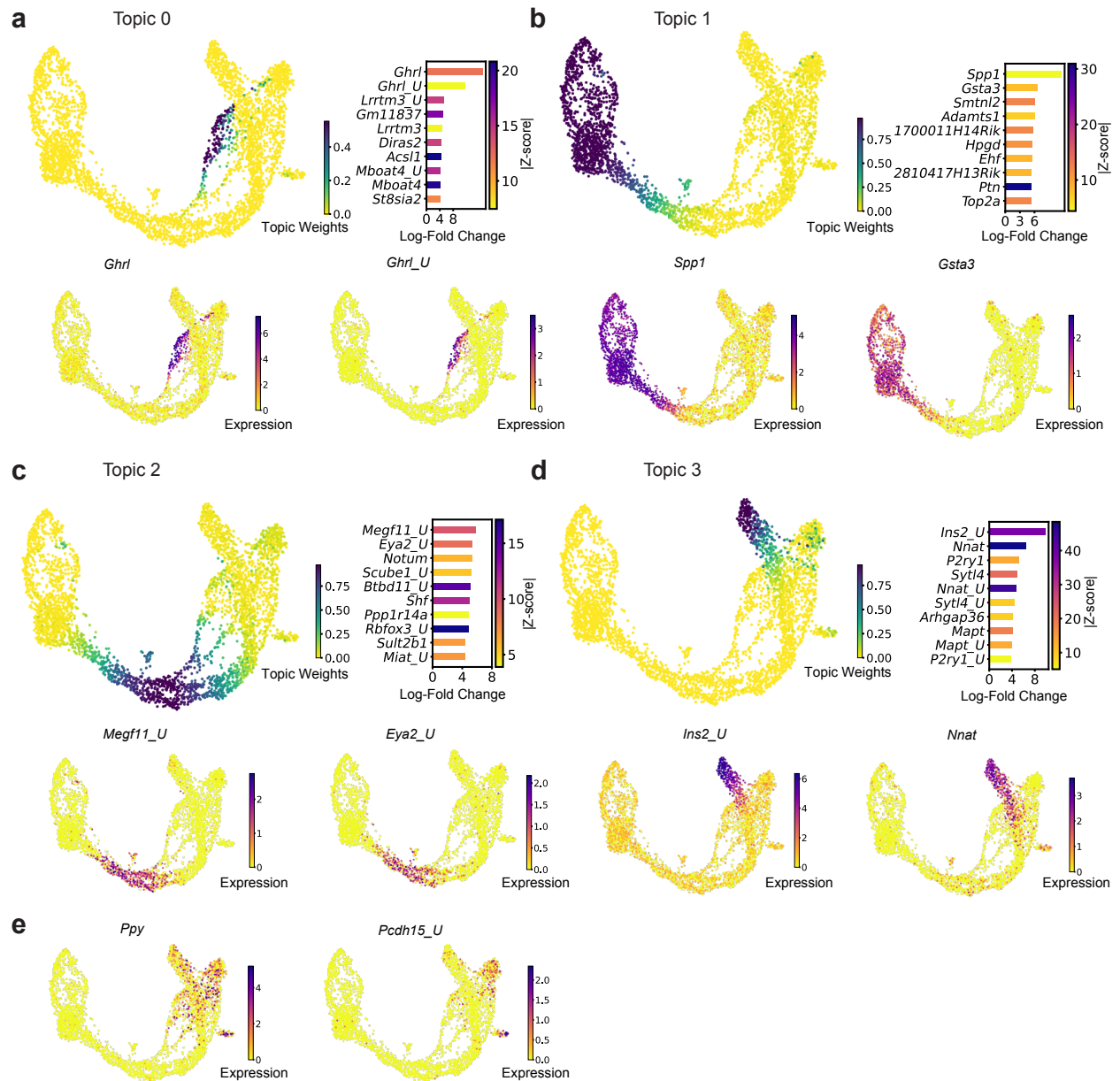
a, Previously published UMAP of dentate gyrus data [8, 21], colored by cell-type annotation, shows streamlines (black arrows) from the *scVelo* dynamical model, which reveal major trajectories but not the transition from oligodendrocyte precursors (OPCs) to oligodendrocytes (OLs) (orange box). **b**, UMAP shows streamlines from *TopicVelo* with 10 topics, which the OPC-OL transition (green arrow), as well as the other major trajectories. **c**, For topic 2 of the topic model (as in Supp. Fig. S8), UMAP (left) is colored by topic weights, and the bar chart (right) shows the top 10 topic-specific genes by largest log-fold change (x axis), colored by z-score. **d**, Close-up of UMAP highlighting OL lineage, is colored by smoothed, size-normalized counts of topic-2 specific genes (left two panels), and by the velocities for *Enpp2* inferred by *scVelo* and *TopicVelo* (right two panels). **e**, Violin plots show the velocity coherence for *scVelo* and *TopicVelo*. (White dot: median, black vertical line: 25th–75th percentile.) **f**, Bar charts show the stationary distributions from *scVelo* and *TopicVelo*, aggregated and colored by cell type (left) and by time point (right). **g**, Scatter plot shows the relative flux (y axis) from *scVelo* and *TopicVelo* (color) for known transitions between pairs of subpopulations, in the direction of the arrow (x axis).



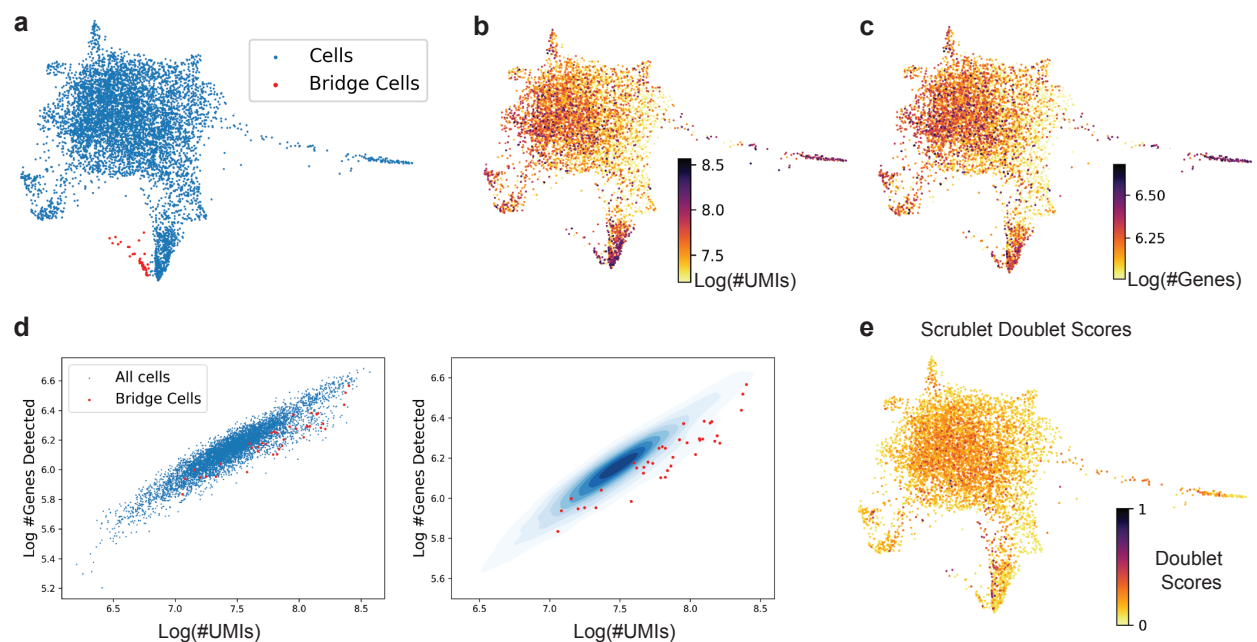
Supporting Figure S8: Topic modeling analysis of dentate gyrus data reveals lineage stages and key genes of rare cell types. **a**, For topic 0, (top, left) bar plot shows top 10 topic-specific genes ranked by log-fold change (x axis) and colored by absolute value of z-score; (top, right) UMAP shows cells colored by topic weights; (bottom) UMAP plots colored by log-normalized expression of topic-specific genes. '_U' indicates unsplined transcripts. **b–i**, Analogous to **a**, for topics 1 and 3–9, respectively. Topic 2 shown in Supp. Fig. S7.



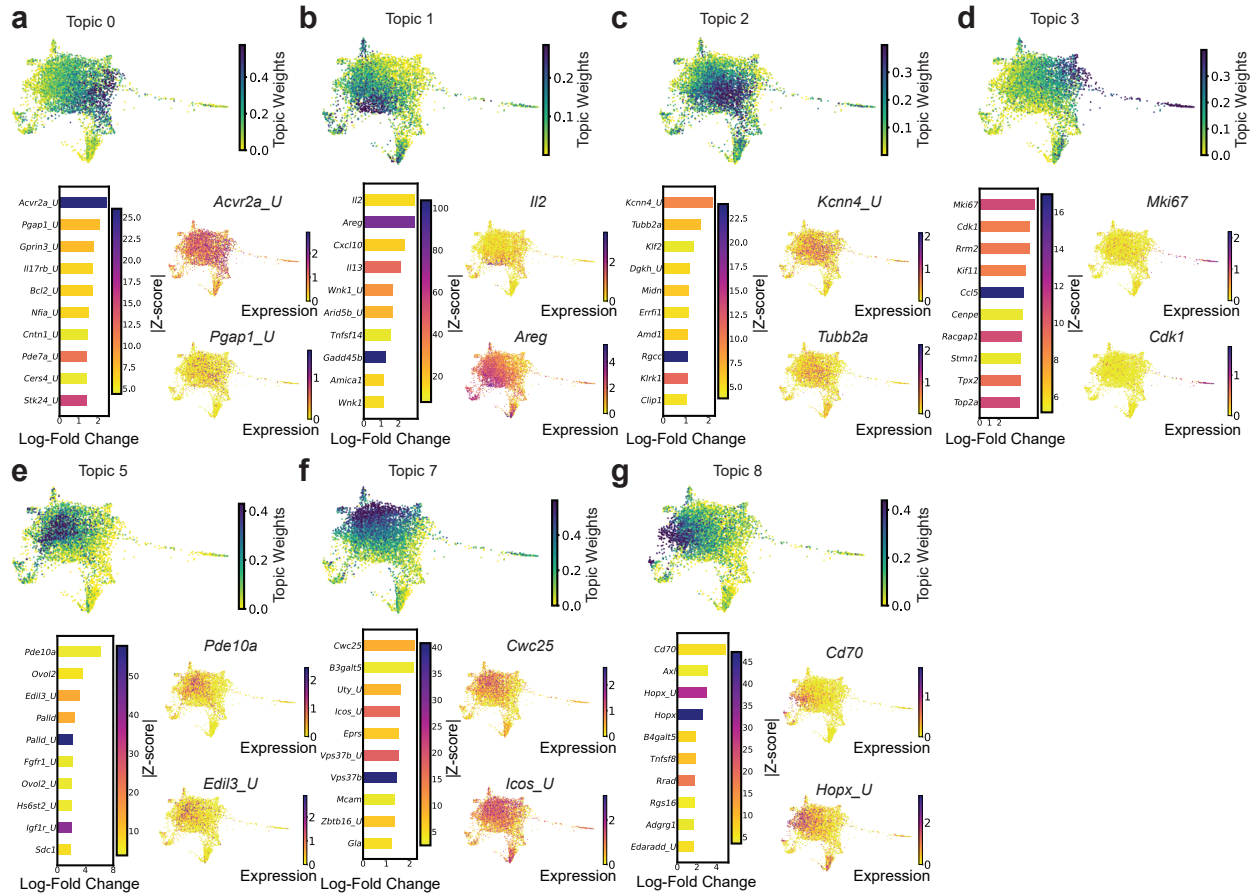
Supporting Figure S9: *TopicVelo* correctly infers multifurcating trajectories during pancreatic endocrinogenesis. **a**, **b**, Previously published UMAP embedding of pancreatic endocrinogenesis data [21, 24], colored by cell type, shows streamlines from the *scVelo* dynamical model (**a**) and *TopicVelo* (**b**). *scVelo* streamlines show the major trajectories but miss the pre-endocrine- δ transition (purple box, red arrow) and suggest a dominant pre-endocrine- β transition (black box). *TopicVelo* with 6 topics also recovers the major transitions, as well as the δ -cell commitment (purple box, green arrow), and suggest greater uncertainty in the α - β branch point (black box). **c**, For topic 4 of the topic model (as in Supp. Fig. S10), UMAP plots are colored by topic weights (upper left) and log-normalized expression of topic-associated genes *Gcg* and *Sst* (right); the bar chart (bottom left), shows the top 10 topic-specific genes by largest log-fold change (x axis), colored by z-score. **d**, For topic-4 specific gene *Spock3* (Supp. Table 1), close-up of UMAP highlighting the δ lineage is colored by smoothed, size-normalized counts of (left, top) unspliced and (left, bottom) spliced transcripts, and by velocities inferred by (right, top) *scVelo* and (right, bottom) *TopicVelo*. **e**, Analogous to **c**, for topic 5. **f**, Violin plots show the velocity coherence for *scVelo* and *TopicVelo*. (White dot: median, black vertical line: 25th–75th percentile.) **g**, Bar charts show the stationary distributions from *scVelo* and *TopicVelo*, aggregated and colored by cell type (left) and by time point (right). **h**, Scatter plot shows the relative flux (y axis) from *scVelo* and *TopicVelo* (color) for known transitions between pairs of subpopulations, in the direction of the arrow (x axis).



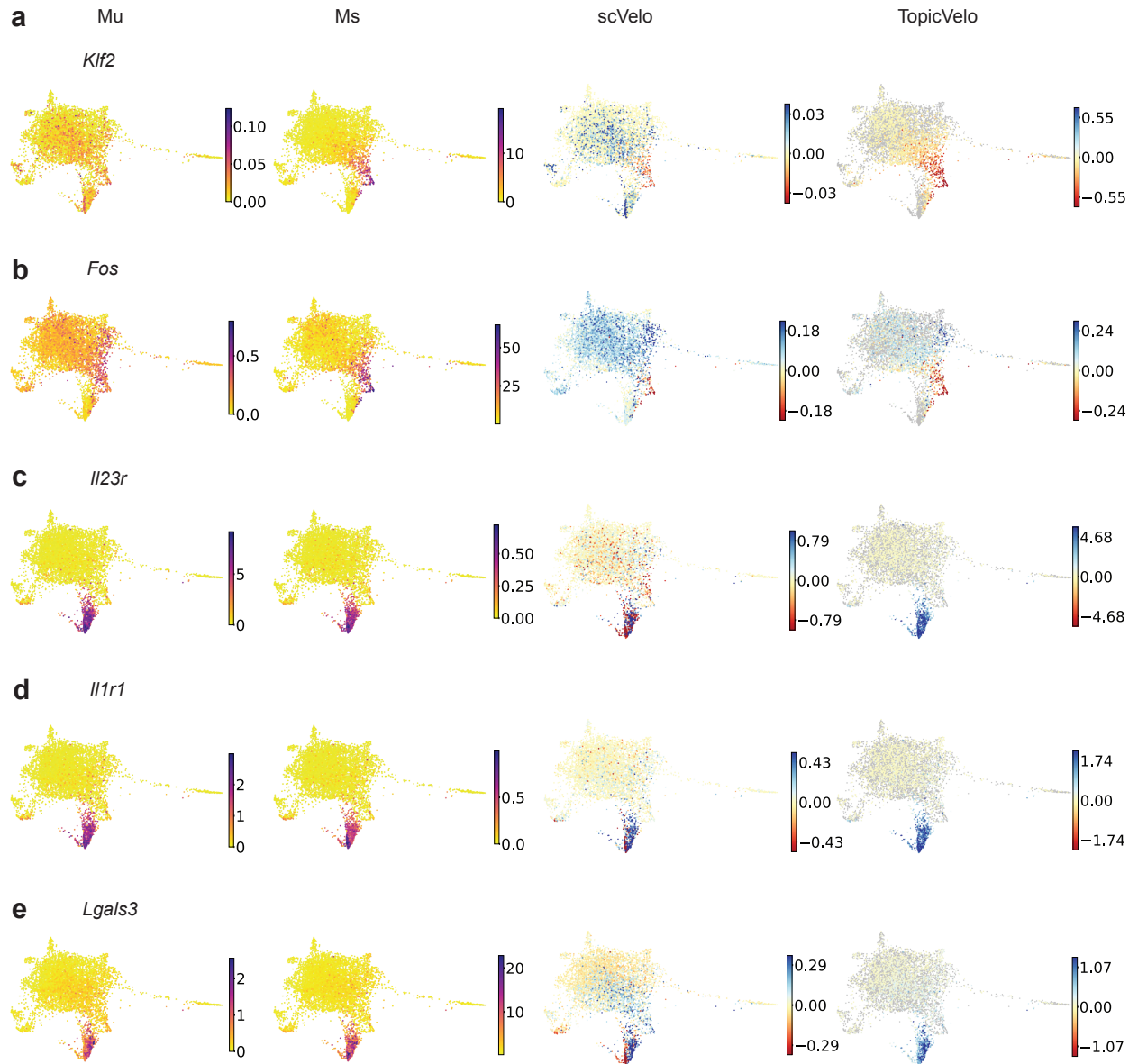
Supporting Figure S10: Topic modeling analysis of the pancreas dataset captures the cycling ductal cells, transient *Ng3n*-expressing endocrine progenitors, and maturation of ϵ and β cells. **a**, For topic 0, (left, top) UMAP shows cells colored by topic weights; (right, top) bar plot shows top 10 topic-specific genes ranked by log-fold change (x axis) and colored by absolute value of z-score; (bottom) UMAP plots colored by log-normalized expression of topic-specific genes. ‘_U’ indicates unsplined transcripts. **b–d**, Analogous to a, for topics 1–3, respectively. **e**, For topic 4 (as in Supp. Fig. S9), UMAP plots are colored by log-normalized expression of topic-4 specific genes *Ppy* and *Pcdh15_U*. ‘_U’ indicates unsplined transcripts.



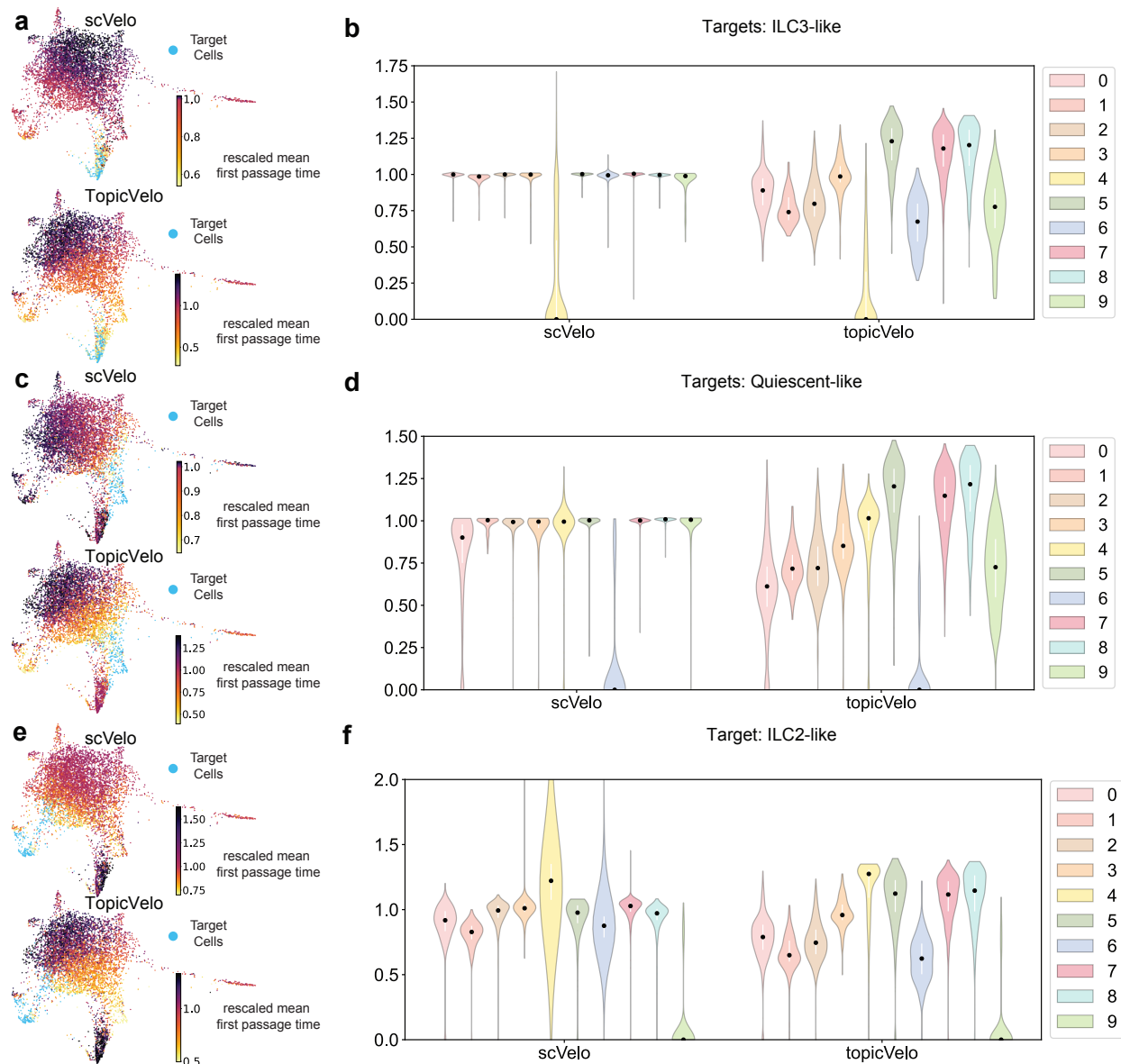
Supporting Figure S11: The bridge cells in the ILCs data are unlikely to be doublets. **a**, The force-directed layout (FDL) embedding (as in Fig. 4), with cells in the “bridge” colored red and others in blue. **b c**, The FDL embedding colored by the natural logarithm of the number of UMIs (**b**) and number of genes detected (**c**) in each cell. **d**, Scatter plot (left) and density plot (right) of the # UMIs versus # genes detected. **e**, The FDL embedding colored by the doublet scores computed by *Scrublet* [19].



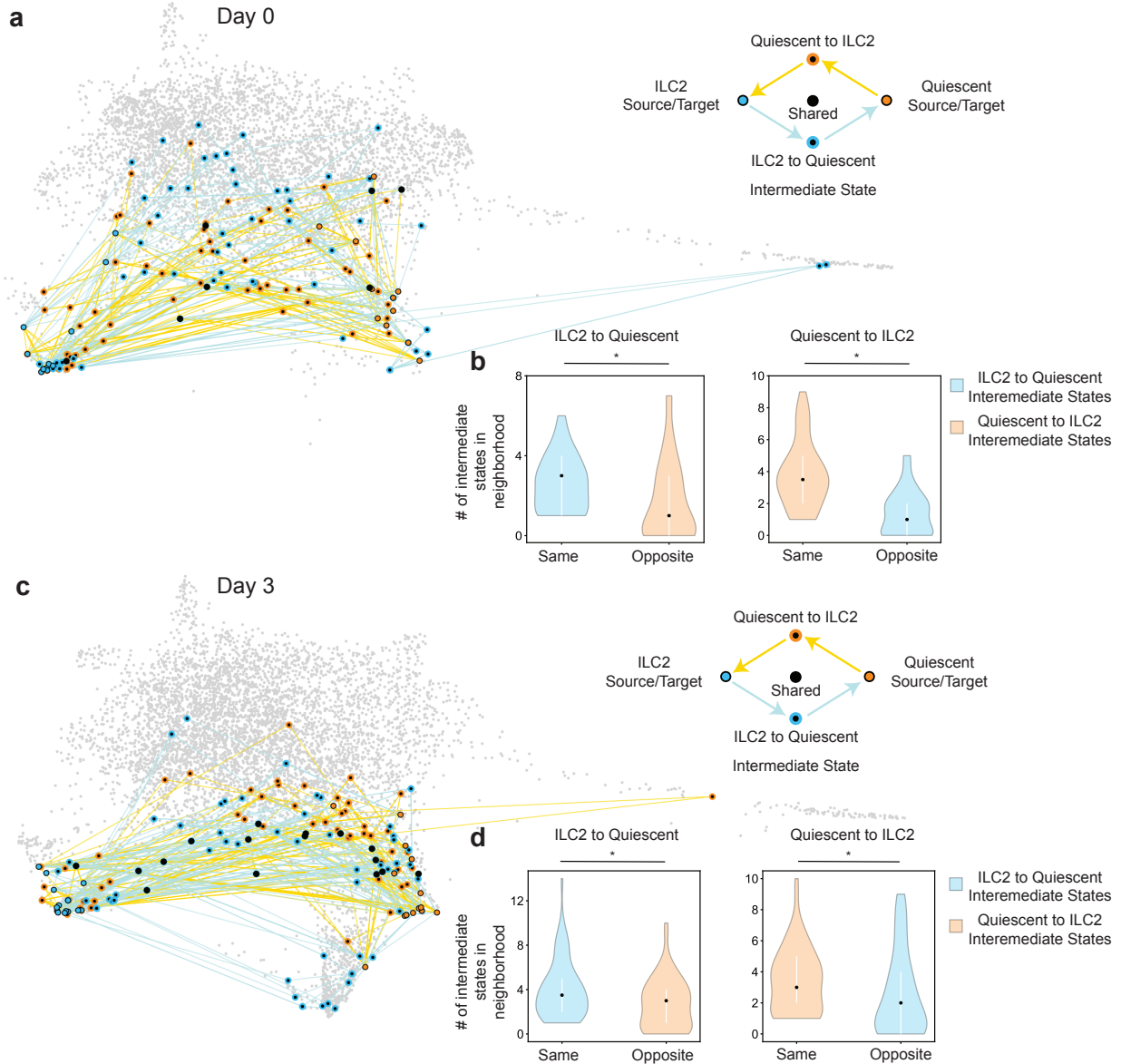
Supporting Figure S12: Topic modeling analysis of the ILCs data from only day 3. **a**, For topic 0, force-directed layout (FDL) embeddings shows cells colored by topic weights (top) and by log-normalized expression of topic-specific genes (bottom right); bar plot (bottom left) shows top 10 topic-specific genes ranked by log-fold change (x axis) and colored by absolute value of z-score; '.U' indicates unsplined transcripts. **b–g**, Analogous to a, for topics 1–8, respectively.



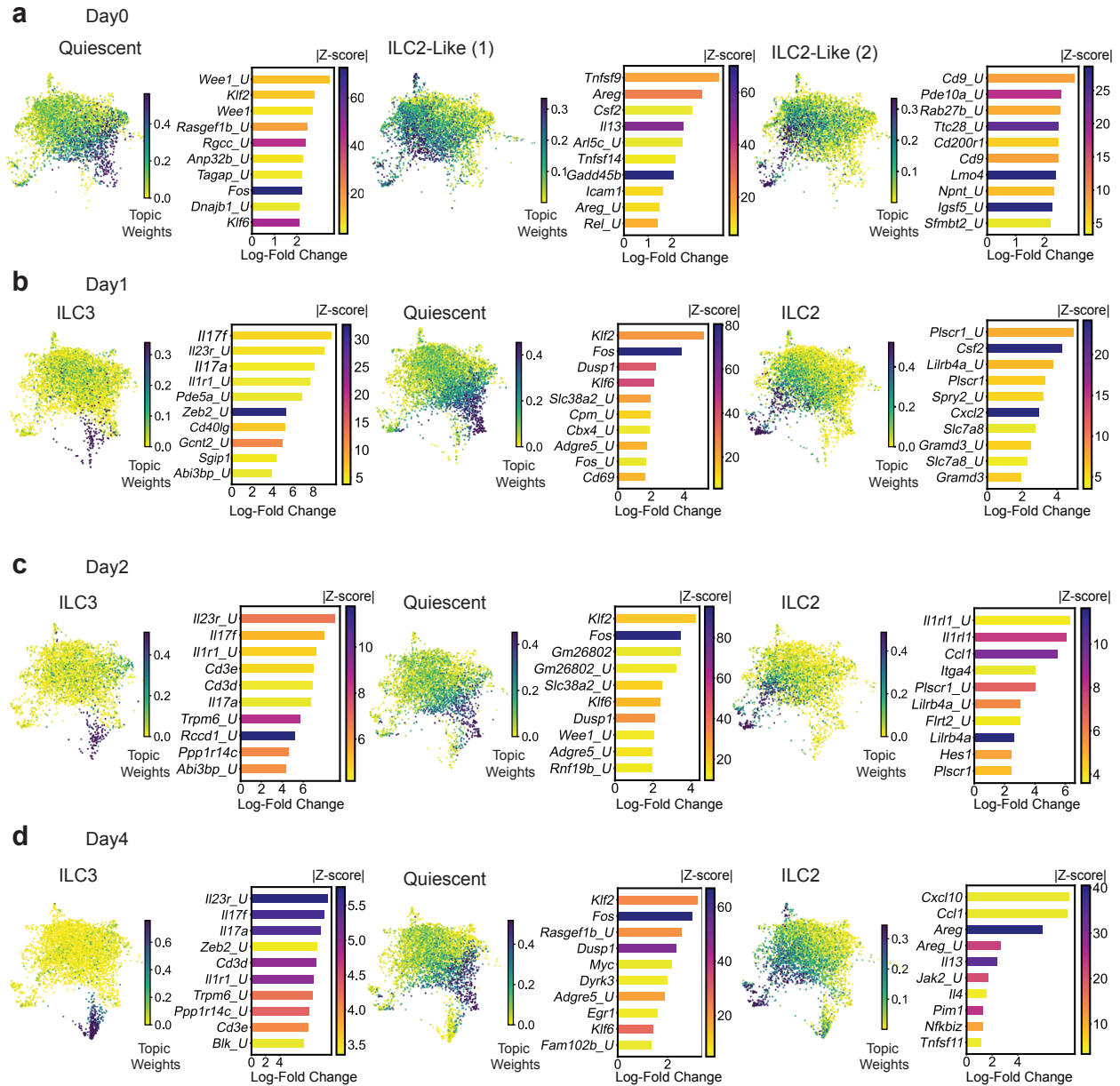
Supporting Figure S13: *TopicVelo* recovers more biologically plausible velocities than *scVelo* for the skin ILCs data. **a, b**, Analysis of topic-6 specific genes. FDL plots colored by smoothed size-normalized counts of unspliced (Mu) (far left) and spliced (Ms) (middle left) transcripts, and by velocities inferred by *scVelo* (middle right) and *TopicVelo* (far right), for the genes *Klf2* (**a**) and *Fos* (**b**). **c–e** Analysis of topic-4 specific genes *Il23r*, *Il1r1*, and *Lgals3*, analogous to a, b.



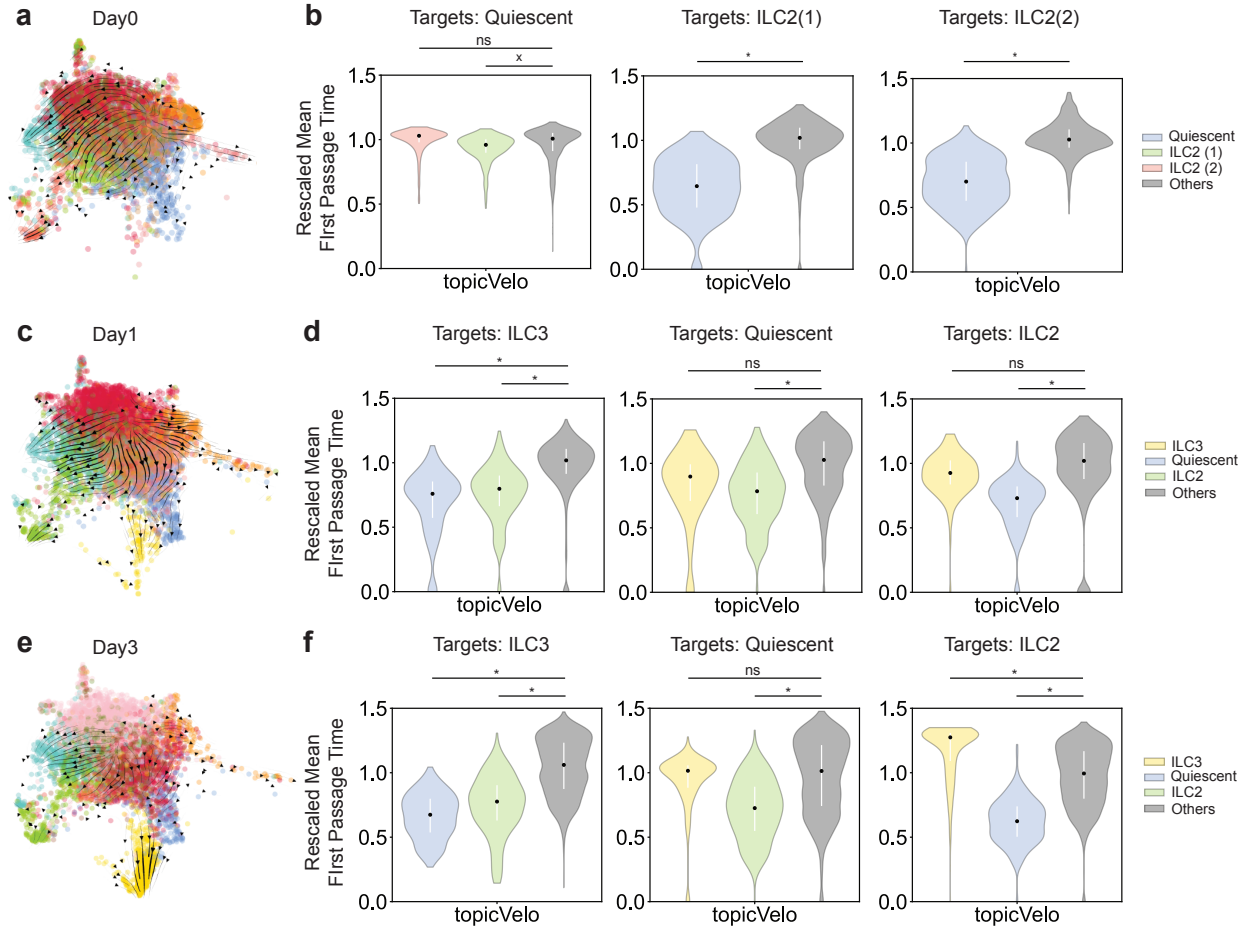
Supporting Figure S14: Mean first-passage time analysis of the skin ILCs data. **a, b**, Median-rescaled mean first-passage times (rMFPTs) to target group ILC3-like cells. FDL plots (**a**) show target cells in blue and other cells colored by rMFPT, as estimated by *scVelo* (top) and *TopicVelo* (bottom). Violin plots (**b**), show the distributions of rMFPTs to ILC3-like cells, aggregated and colored by the highest-weight topic in each cell, as computed by *scVelo* (left) and *TopicVelo* (right). (Black dot: median, white vertical line: 25th–75th percentile). **c, d**, Analogous to **a, b**, with target group set to quiescent-like cells. **e, f**, Analogous to **a, b**, with target group set to ILC2-like cells.



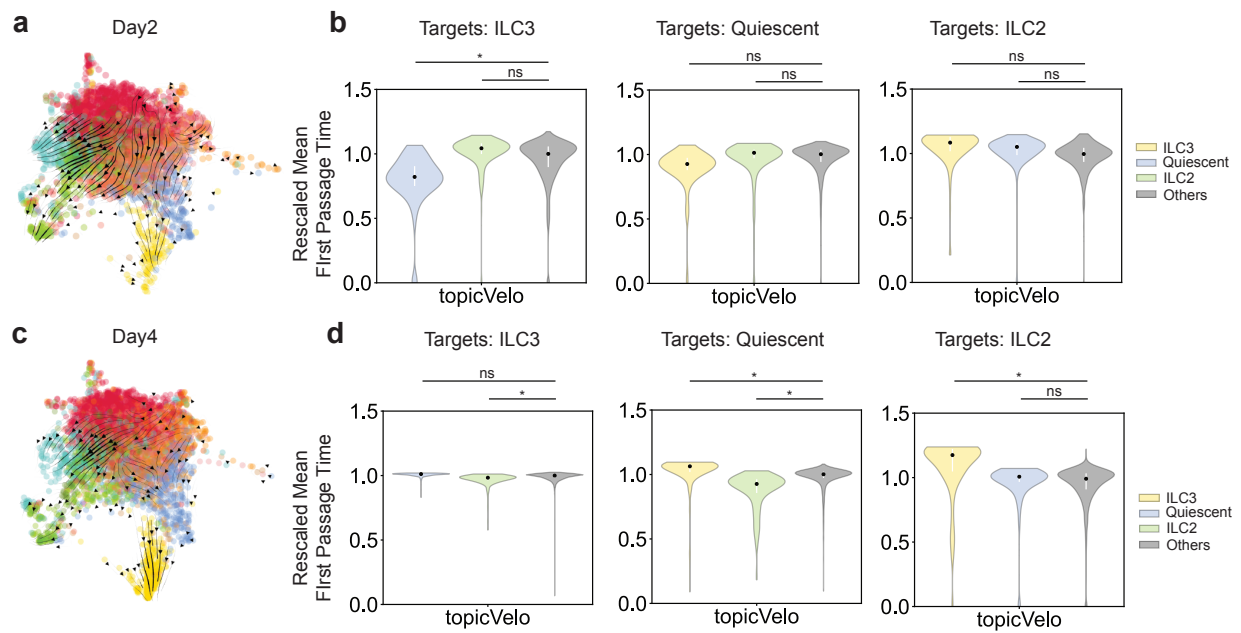
Supporting Figure S15: *TopicalVelo* predicts that the most probable transition paths between the highest weighted ILC2-like and quiescent-like cells traverse different but overlapping regions of transcriptomic space in each direction. **a**, For cells from day 0 only, the FDL embedding shows the top 0.2% quiescent-like cells in orange with black outlines and the top 0.2% ILC2-like cells in blue with black outlines; for quiescent-to-ILC2 paths (orange lines), intermediate cells are shown in black with orange outlines; for ILC2-to-quiescent paths (blue lines), intermediate cells are shown in black with blue outlines; intermediate states that occur in paths in both directions are represented as solid black dots. Other cells from day 0 are in gray. (Note: the aspect ratio of the FDL embedding has been stretched for better visibility.) **b**, Violin plots show the distribution, over intermediate states i in ILC2-to-quiescent paths (left panel) and quiescent-to-ILC2 paths (right panel), of the number of intermediate states j from paths in the same and opposite directions (x axis, color) such that j is in the neighborhood of i in the underlying k_G -NN graph. (Black dot: median, white vertical line: 25th–75th percentile). * indicates $P < 0.01$ by one-sided permutation tests. **c, d**, Analogous to panels a and b, respectively, for day 3.



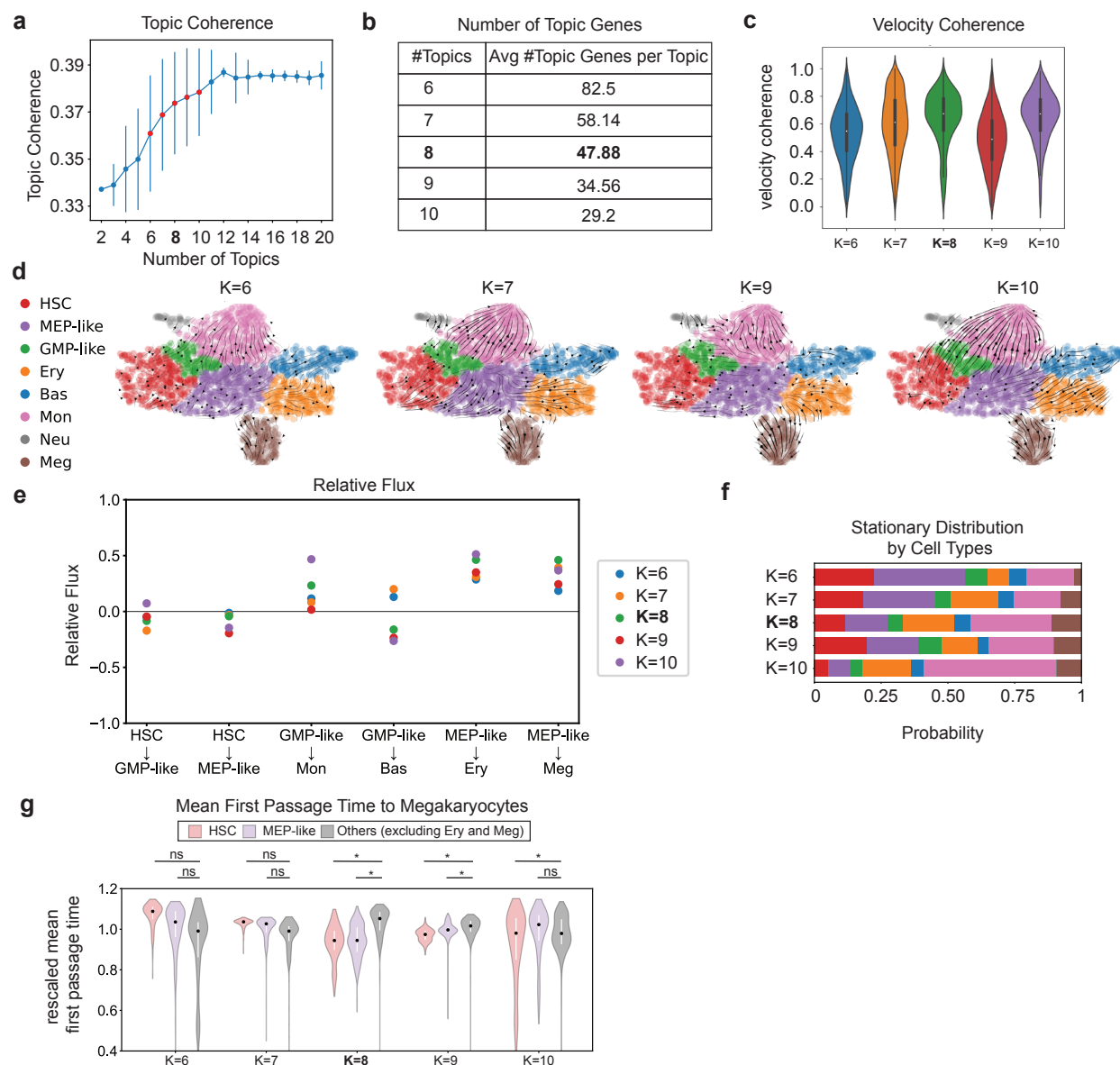
Supporting Figure S16: Topic modeling reveals consistent gene programs for ILCs sampled on different days. **a**, For day 0 (before induction), the FDL plots show cells colored by weights for topics associated with quiescence (left) and ILC2s (middle, right); corresponding bar plots show the top 10 topic-specific genes, ranked by log-fold change (x axis) and colored by absolute value of z-score. '_U' indicates unspliced transcripts. **a**, For day 1 (before induction), the FDL plots show cells colored by weights for topics associated with ILC3s (left), quiescence (middle), and ILC2s (right); corresponding bar plots show the top 10 topic-specific genes, ranked by log-fold change (x axis) and colored by absolute value of z-score. **c,d**, Analogous to panel b, for days 2 and 7, respectively.



Supporting Figure S17: Days 0 and 1 exhibit ILC dynamics consistent with those on day 3. **a**, FDL as in Fig. 4, with cells colored by weight of dominant topic, shows the streamlines from *TopicVelo* for cells from day 0. **b**, Violin plots show the distributions of median-rescaled mean first-passage times, estimated using *TopicVelo* on cells from day 0, from different groups of non-target cells (colors) to different target subpopulations, i.e., quiescent-like (left), and the two ILC2-like (middle and right) subpopulations. (Black dot: median, white vertical line: 25th–75th percentile). *, $P < 0.0001$ by one-sided permutation test; x, $0.001 \leq P < 0.1$; ns, $P \geq 0.1$. **c**, Analogous to panel a, for day 1. **d**, Analogous to panel b for day 1, with target groups as ILC3-like (left), quiescent-like (middle), and ILC2-like (right) subpopulations. **e**, Analogous to panel a, for day 3. **f**, Analogous to d for day 3.

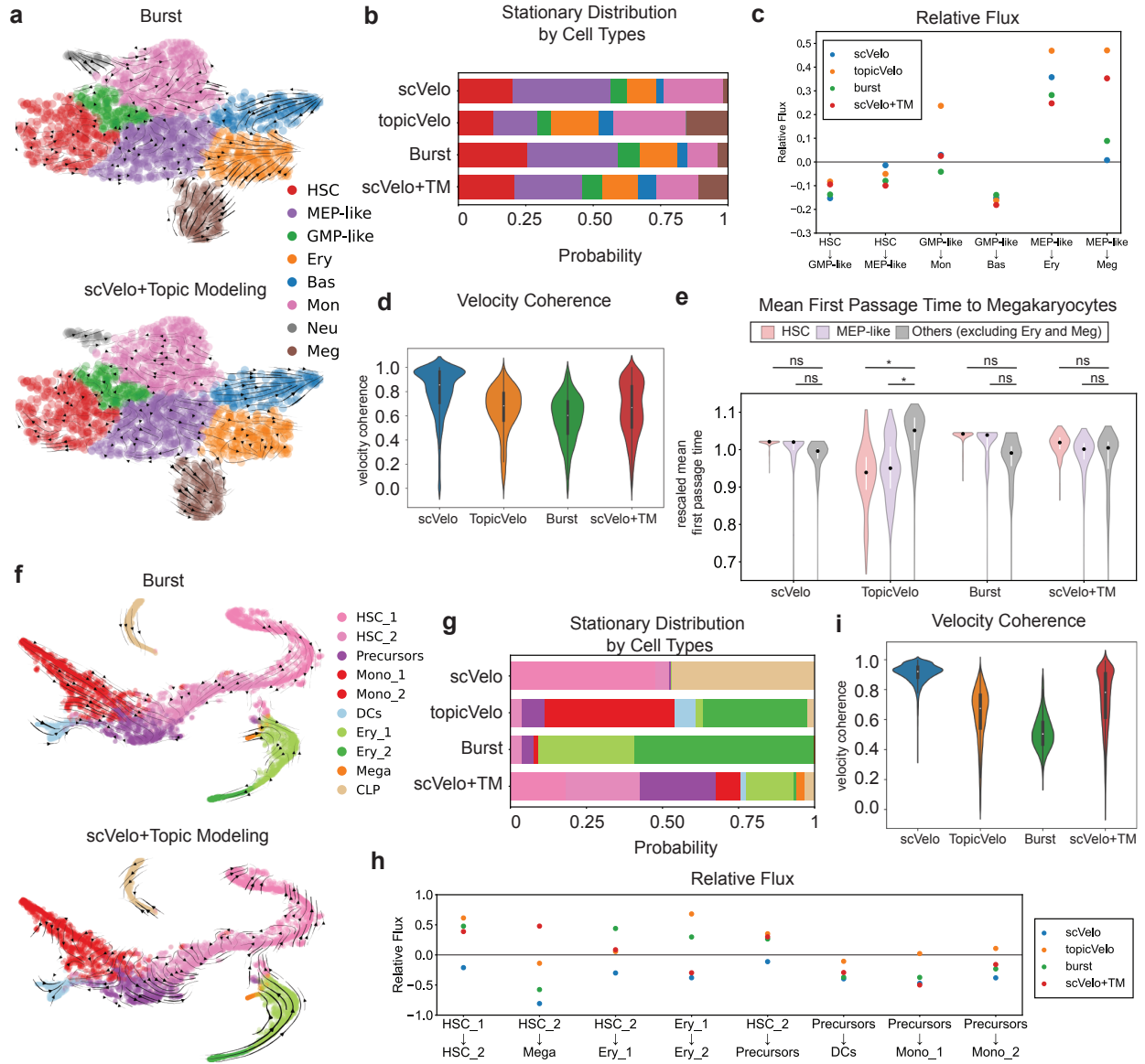


Supporting Figure S18: Fewer clear signals of ILC state transitions were observed on days 2 or 4. **a**, FDL as in Fig. 4, with cells colored by weight of dominant topic, shows the streamlines from *TopicVelo* for cells from day 2. **b**, Violin plots show the distributions of median-rescaled mean first-passage times, estimated using *TopicVelo* on cells from day 2, from different groups of non-target cells (colors) to different target subpopulations, i.e., ILC3-like (left), quiescent-like (middle), and ILC2-like (right) cells. (Black dot: median, white vertical line: 25th–75th percentile). *, $P < 0.0001$ by one-sided permutation test; ns, $P \geq 0.1$. **c**, Analogous to panel a, for day 4. **d**, Analogous to panel b for day 4.

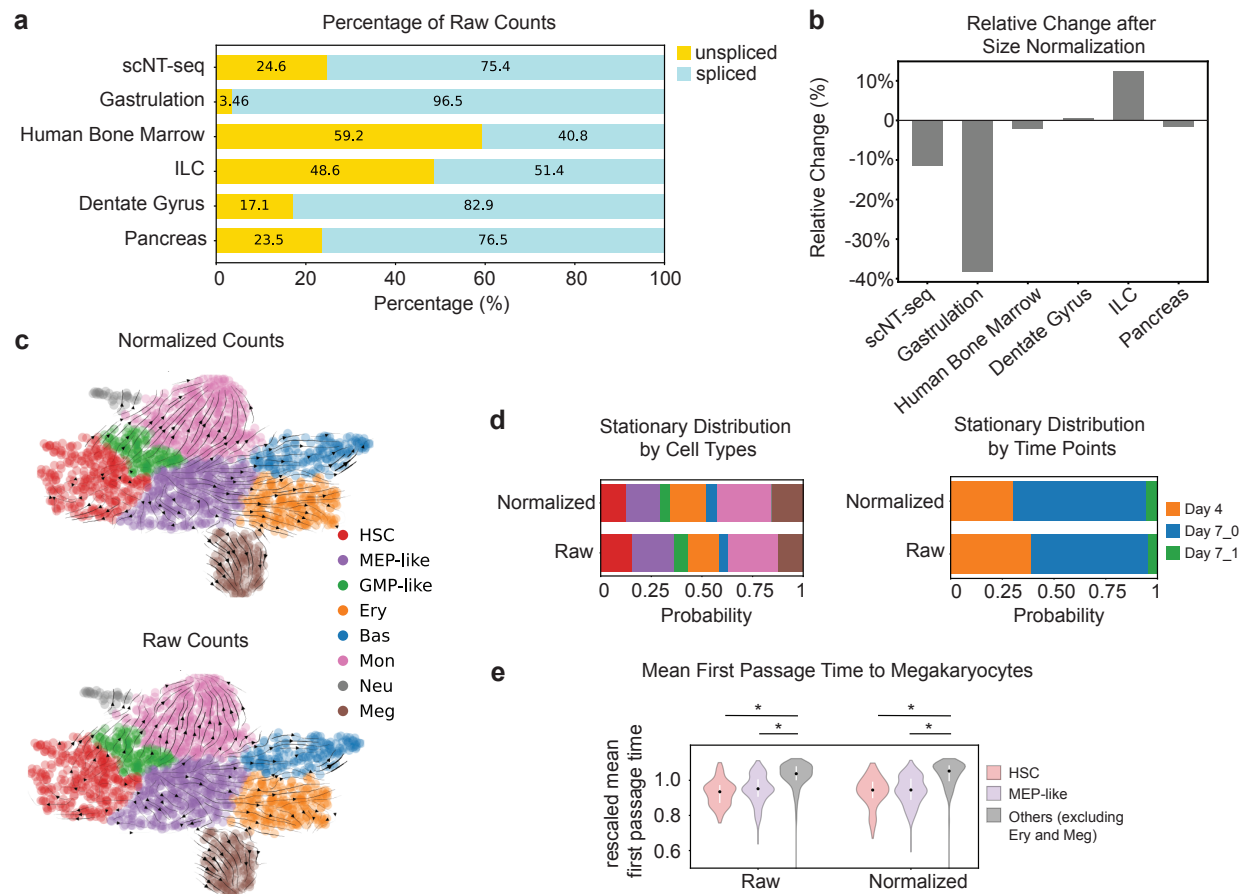


Supporting Figure S19: Overall results from *TopicVelo* are insensitive to the exact choice of topic number within an appropriate regime.

All results shown are on the scNT-seq hematopoiesis data (as in Fig. 2). ***K* = 8** (in bold in multiple panels) is the value used in the main analysis (e.g., Fig. 2). **a**, Topic coherence (y axis) is shown as a function of the number of topics (x axis), with bars corresponding to the standard deviation over 50 repetitions of model fitting. Red dots correspond to the topic numbers chosen for the demonstration. **b**, For the topic models used in the demonstration, the table gives the average number of topic genes per topic. **c**, Violin plots show velocity coherence (y axis) for models with different topic numbers (x axis). (White dot: median, black vertical lines: 25th-75th percentile). **d**, Each UMAP plot (as in Fig. 2) is colored by cell type and shows streamlines inferred by *TopicVelo* using a model with a different topic number. **e**, Plot shows the relative flux (y axis) for known transitions between pairs of cell subpopulations in the direction of the arrow (x axis) for *TopicVelo* results using models with different topic numbers (color). **f**, Bar plot shows the stationary distributions, aggregated and colored by cell type, for different topic numbers (rows). **g**, Violin plots show median-rescaled mean first-passage time to megakaryocytes from two progenitor subpopulations (in color, HSC and MEP-like) and other cells (in gray, excluding erythroid cells) for *TopicVelo* results using different topic numbers (x axis). (Black dot: median, white vertical lines: 25th-75th percentile.). *, $P < 0.0001$ by one-sided permutation test; ns, $P \geq 0.1$.



Supporting Figure S20: Algorithmic ablation studies indicate that topic modeling and the burst model extract distinct, biologically meaningful signals. **a**, UMAP (as in Fig. 2) for the scNT-seq data, colored by cell type, shows streamlines created using the burst transcriptional model only (top) and using *scVelo* combined with topic modeling (bottom). **b**, Bar charts show the stationary distributions, aggregated and colored by cell type, for the results using *scVelo*, *TopicVelo*, and the two ablative approaches (rows). **c**, Plot shows the relative flux (y axis) for known transitions between pairs of cell subpopulations in the direction of the arrow (x axis) for the different methods (color). **d**, Violin plots show velocity coherence (y axis) for the results from different methods (x axis, color). (White dot: median, black vertical lines: 25th-75th percentile). **e**, Violin plots show median-rescaled mean first-passage time to megakaryocytes from two progenitor subpopulations (in color, HSC and MEP-like) and other cells (in gray, excluding erythroid cells) for the different methods (x axis). (Black dot: median, white vertical lines: 25th-75th percentile.). *, $P < 0.0001$ by one-sided permutation test; ns, $P \geq 0.1$. **f**, A *t*-SNE plot (as in Fig. 3) for the bone marrow data, colored by cell type, shows streamlines created using the burst transcriptional model only (top) and using *scVelo* combined with topic modeling (bottom). **g-i**, Panels are analogous to panels b-d, for the bone marrow data.



Supporting Figure S21: *TopicVelo* is robust with respect to size normalization. **a**, Bar charts show the percentages (x axis, annotations in white) of the total raw gene counts that correspond to unspliced and spliced (color) transcripts in the 6 data sets (rows) considered in our study. **b**, The bar chart show the relative change (y axis), after size normalization and rounding, in the ratio of the total number of unspliced transcripts to the total number of spliced transcripts for each of the datasets (x axis). **c–e**, Size normalization does not strongly impact *TopicVelo* results in the scNT-seq human hematopoiesis data. UMAP plots (as in Fig. 2) (**c**), colored by cell type, show the streamlines from the results of *TopicVelo* using either size-normalized (top) or raw (bottom) counts. Bar plots (**d**) show stationary distributions, aggregated and colored by cell type (left) and time point (right), from the results of *TopicVelo* using size-normalized or raw counts (rows). Violin plots (**e**), show the median-rescaled mean first-passage time to megakaryocytes from different cell subpopulations (colors) inferred from *TopicVelo* using raw counts or size-normalized counts (x axis). (Black dot: median, white vertical line: 25th-75th percentiles.) *, $P < 0.0001$ by one-sided permutation test.