PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

# Dissection and integration of bursty transcriptional dynamics for complex systems

Cheng Frank Gao[a], Suriyanarayanan Vaikuntanathan[a,b,1], and Samantha J. Riesenfeld[b,c,d,e,1,2]

Edited by Bonnie Berger, Massachusetts Institute of Technology, Cambridge, MA; received April 26, 2023; accepted March 6, 2024

**RNA velocity estimation is a potentially powerful tool to reveal the directionality of transcriptional changes in single-cell RNA-sequencing data, but it lacks accuracy, absent advanced metabolic labeling techniques. We developed an approach, *TopicVelo*, that disentangles simultaneous, yet distinct, dynamics by using a probabilistic topic model, a highly interpretable form of latent space factorization, to infer cells and genes associated with individual processes, thereby capturing cellular pluripotency or multifaceted functionality. Focusing on process-associated cells and genes enables accurate estimation of process-specific velocities via a master equation for a transcriptional burst model accounting for intrinsic stochasticity. The method obtains a global transition matrix by leveraging cell topic weights to integrate process-specific signals. In challenging systems, this method accurately recovers complex transitions and terminal states, while our use of first-passage time analysis provides insights into transient transitions. These results expand the limits of RNA velocity, empowering future studies of cell fate and functional responses.**

single-cell RNA-seq | RNA velocity | trajectory inference | probabilistic topic models | systems immunology

One of the key challenges in single-cell data science, trajectory inference (TI) leverages genome-wide transcriptional profiles to estimate the position of each cell in an underlying, ordered biological process (1–3). TI is used to analyze a variety of dynamic processes, most commonly, embryonic development and cellular differentiation, but also immune responses and tumorigenesis (4, 5). The destructive nature of single-cell RNA-sequencing (scRNA-seq) technologies limits the input data to static snapshots, rather than temporal records. Computational innovations glean true dynamic information by exploiting inadvertently captured reads from unspliced pre-mRNA, as well as targeted reads from mature, spliced mRNA, to model the transcriptional kinetics of genes and thereby estimate a time derivative of the transcriptional state, known as RNA velocity (6, 7).

Unlike similarity-based "pseudotime" TI methods (reviewed in ref. 3), RNA velocity reveals the directions and patterns of complex flows, and hence precursor and terminal cell populations, even in a single time point. Its unique capabilities and possible extensions make it a potentially powerful tool in the study of diverse biological systems, particularly where there is limited prior knowledge. Yet, despite advances, the effective use of RNA velocity has been impeded by a lack of robustness and accuracy, driven by multiple factors (8–12). Recent approaches use a variety of techniques to improve it (13–22) but do not generally account for pluripotency or distinct processes, beyond lineages, occurring simultaneously. Moreover, as most methods rely on ordinary differential equations, they do not model intrinsic transcriptional stochasticity. The persistent gap between the promise and reality of RNA velocity has largely restricted its application.

To create a more broadly effective RNA velocity tool for investigating complex systems, including immune responses, we created *TopicVelo* (Fig. 1), an approach that disentangles potentially simultaneous processes using a probabilistic topic model (23, 24), also known as a grade-of-membership model (25, 26). This highly interpretable, Bayesian nonnegative matrix factorization allows *TopicVelo* to focus on the specific cells and genes involved in distinct processes to better capture their dynamics. To infer kinetic parameters for process-specific genes, *TopicVelo* fits integer transcript counts to a physically meaningful transcriptional burst model (27). Using the topic weights for each cell, *TopicVelo* integrates the process-specific dynamics to infer a global model of cell transitions.

In addition to using standard visualizations of streamlines, we assessed RNA velocity results with Markovian techniques, including mean first-passage time analyses that identify transient transitions not observed via traditional approaches. In diverse datasets, *TopicVelo* offers distinctive insights and performs significantly better than the

## Significance

The study of dynamic biological phenomena, such as differentiation, immune responses, and cancer—which involve multiple, simultaneously occurring biological processes—using destructively measured transcriptomic profiles remains a central challenge in single-cell data science. State-of-the-art methods show promise but fail in many settings. Here, we present a method that incorporates a probabilistic topic model to dissect and then integrate simultaneous, yet distinct, bursty transcriptional dynamics. We demonstrate its effectiveness for inferring biologically informative velocity for key genes, identifying complex cell-state transitions and providing insights on transient transitions and terminal state distributions in several challenging biological systems.

Author contributions: C.F.G., S.V., and S.J.R. designed research; C.F.G. developed, implemented, and tested the algorithm; C.F.G. and S.J.R. analyzed data; C.F.G., S.V., and S.J.R. interpreted the results; S.V. and S.J.R. supervised research; and C.F.G., S.V., and S.J.R. wrote the paper.

Competing interest statement: The authors disclose the following patent filing: "Computational Models to Analyze RNA Velocity"; Case number: UCHI 22-T-208-001; Date filed: 17 January 2023.

This article is a PNAS Direct Submission.

[1] S.V. and S.J.R. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: sriesenfeld@uchicago.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2306901121/-/DCSupplemental.

Published April 26, 2024.

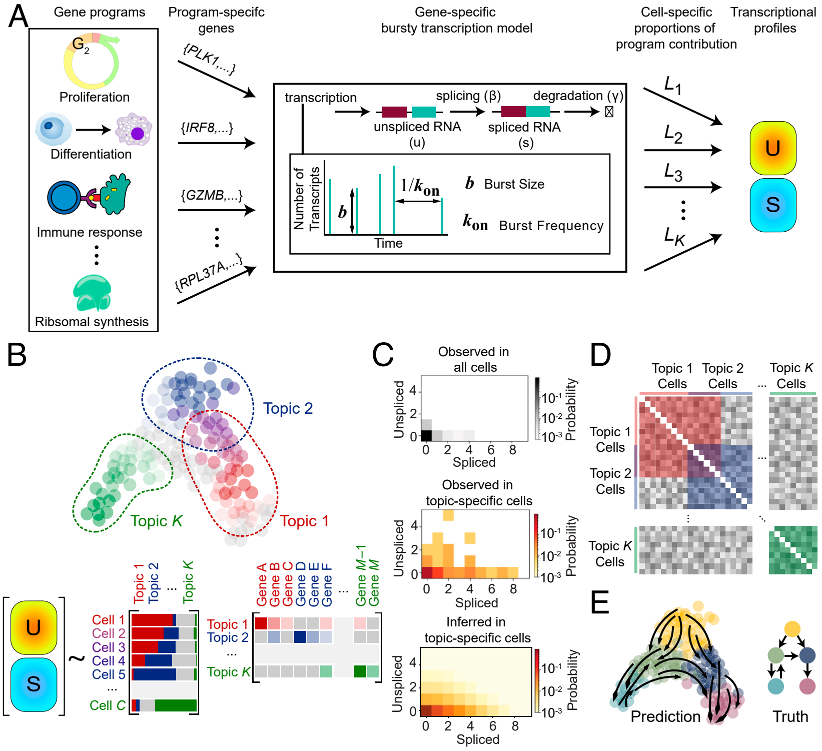

**Fig. 1.** *TopicVelo* combines topic modeling and a burst model for accurate, robust RNA velocity inference. (*A*) The generative model motivating *TopicVelo* accounts for distinct stochastic dynamics of transcriptional processes for different gene programs (*Left*). Program- and gene-specific transcription follows a bursty transcriptional model governed by several parameters: the typical burst frequency $k_{\mathrm{on}}$, the burst size $b$, which has a geometric distribution, the splicing rate parameter $\beta$, and the degradation rate $\gamma$ (*Middle*). By accounting for the varying activity levels ($L_k$) of each program $k$ across cells, the transcriptional profiles can be generated and characterized by the matrices $U$ and $S$, specifying the number of unspliced and spliced transcripts, respectively, of all genes in all cells (*Right*). (*B*) A probabilistic topic model gives a Bayesian low-rank nonnegative matrix factorization of a multinomial probability matrix that generates the combined $U$ and $S$ matrix for a heterogeneous population of cells, which reveals distinct, possibly overlapping, cells and genes associated with underlying, individual programs (topics), thereby capturing cellular pluripotency or multifaceted functionality. (*C*) For many genes, the joint distribution over all cells of spliced and unspliced transcripts is concentrated at (0,0), as the gene is not involved in most cell states (*Top*). Zooming in, the joint distribution of a topic-specific gene in topic-associated cells reveals detailed, process-specific dynamics (*Middle*). To infer those dynamics, we fit the burst model of transcription by minimizing the KL divergence between inferred and experimentally observed joint distributions of spliced and unspliced transcripts (*Bottom*). (*D*) Cell-specific topic weights are leveraged to integrate process-specific transition signals into a global transition matrix. (*E*) Results enable robust, accurate trajectory inference, as assessed by transition streamline visualizations, as well as by new mean first-passage time, terminal states, and relative flux analyses.

state-of-the-art approach *scVelo* (7), without the aid of metabolic labeling or multiple time points, by recovering velocities, transition flows, and terminal states that are more consistent with known biology.

In the rest of the paper, we give an overview of *TopicVelo* and highlight its performance in a human hematopoiesis dataset, for which the correct dynamics were previously inferred only with the aid of metabolic labeling (22). We also illustrate the capability of *TopicVelo* to handle complex developmental systems with stage-dependent dynamics (8, 28–30). Last, we show *TopicVelo* infers validated, complex, convergent trajectories underlying the inflammatory responses of skin lymphocytes, using only a single time point (31).

## Results

**Overview of the *TopicVelo* Method.** One scRNA-seq snapshot may capture multiple biological processes, even within one cell type, including ubiquitous processes, such as proliferation and ribosomal synthesis, as well as system-specific processes, such as differentiation and immune responses (Fig. 1*A*). Each process involves a set of genes, or gene program, for which the process- and gene-specific kinetics are typically governed by a bursty transcription model (32). The resulting transcriptional profiles of cells in the system also reflect the varying degrees to which different processes have been active in each cell up to the time of capture. These considerations are absent in existing RNA velocity approaches but must be accounted for in an accurate model of the generative processes of scRNA-seq data. The need to capture these key biological features motivated our approach to *TopicVelo*. Because the joint inference of all parameters in such a generative model may be computationally intractable, *TopicVelo* separates the inference of program-specific genes and cell-specific activity levels from the inference of kinetic parameters. Specifically, *TopicVelo* operates in these three stages:

***Process-specific inference.*** Inspired by previous works that effectively use probabilistic topic models to distinguish biologically relevant signals in scRNA-seq data (e.g., refs. 31 and 33–36), we apply topic modeling (35) to the combined unspliced and spliced transcript matrix (Fig. 1*B*) (*Materials and Methods* and *SI Appendix*, section 1). The result is a representation of each cell as a probability distribution over topics (gene programs, in the context of scRNA-seq), while each topic is a probability distribution over individual genes (Fig. 1*B*). Process-associated cells, i.e., cells with relatively high weights in a topic, and process-specific genes, determined using previous strategies (31, 36), serve as the input for inferring process-specific kinetic parameters. Within process-associated cells, process-specific genes can reveal important dynamic information that is hidden at the global scale and hence missed by existing methods (Fig. 1*C*).

The number of topics is a user-selected parameter, which, like clustering resolution, often has multiple, biologically meaningful settings. We explored several topic-quality metrics developed in natural language processing (e.g., refs. 37–39) and also used the biological literature to assess interpretability of topic-specific gene programs (*Materials and Methods* and *SI Appendix*, section 1).

***Bursty transcription model.*** In contrast to the ODE-based one-state model underlying *scVelo*, *TopicVelo* efficiently fits a more faithful physical model that accounts for transcriptional bursting (*Materials and Methods*), adapting a previous model for studying mRNA transport (27) (Fig. 1 *A* and *C*). The chemical master equation of the model for a given gene is:

$$\begin{aligned}\frac{\partial p(u,s,t)}{\partial t} &= k_{\mathrm{on}}\left[\sum_{z=0}^{u} p_z p(u-z,s,t) - p(u,s,t)\right]\\ &+ \beta\left[(u+1)p(u+1,s-1,t) - up(u,s,t)\right]\\ &+ \gamma\left[(s+1)p(u,s+1,t) - sp(u,s,t)\right], \qquad [1]\end{aligned}$$

where $p(u,s,t)$ is the probability of observing a cell with $u$ unspliced pre-mRNA transcripts and $s$ spliced mature mRNA transcripts at time $t$; $k_{\mathrm{on}}$ is the rate of the Poisson process governing the burst event; $p_z$, the probability of producing $z$ unspliced pre-mRNA transcripts during a single burst event, is governed by a geometric distribution; $\beta$ is the splicing rate; and $\gamma$ is the rate of degradation of spliced mRNA. Parameters are initialized with the method of moments or another heuristic. Cells with weights that are relatively high for a given topic are assumed to be in steady state for topic-associated genes. For a topic-associated gene and parameter setting, we use the Gillespie algorithm (40) to estimate the joint distribution of unspliced and spliced transcript counts in steady-state cells for the transcriptional burst model. The maximum likelihood values of parameters are then estimated using an implementation of the Nelder–Mead algorithm (41) (*SI Appendix*, Fig. S1). Hence, each set of topic-associated genes shares a common splicing rate, but genes from different topics may operate on different time scales.

***Integration of process-specific dynamics.*** A key feature of *TopicVelo* is the capability to integrate process-specific transition matrices into a global transition matrix (Fig. 1*D*) (*Materials and Methods*). First, from the inferred process-specific kinetic parameters, *TopicVelo* constructs process-specific transition matrices, based on a previous approach (7), namely by applying an exponential kernel to the cosine similarities between velocities and differences in spliced expression among nearest neighbors. Each transition matrix characterizes the probabilistic flow of process-specific transcriptional changes across process-associated cells. Then, a global transition matrix is constructed by linearly combining process-specific transition matrices, using the topic weights of cells. This strategy enables locally important dynamics to be accurately recovered and then woven into larger-scale, complex trajectories. The user-selected topic-weight threshold, which determines topic-associated cells, balances an inherent trade-off between the benefit of separating dynamic processes and the risk of losing dynamic range and/or information in overlaps among topic-associated cells (*Materials and Methods* and *SI Appendix*, section 1).

***Analysis of the integrated transition matrix.*** To reveal cell-state transitions and assess the accuracy of RNA-velocity–based inference, we use both customary streamline visualizations and several quantitative approaches (*Materials and Methods* and *SI Appendix*, section 1). We compute the stationary distribution of the integrated transition matrix to identify terminal cell populations. We also introduce the use of mean first-passage time (MFPT), which captures the expected time needed for a cell-state transition to occur, to gain insights into transient transitions invisible at the global scale with traditional approaches. Furthermore, we introduce another measure, relative flux, to quantify relative transitions across cell type boundaries.

Our analysis of the run-time and memory complexity of software (*SI Appendix*, section 2), as well as its performance in diverse applications detailed below, revealed its capacity to efficiently offer interpretable biological insights (Fig. 1*E*).

## *TopicVelo* Enables Challenging Trajectory Inference in Human Hematopoiesis without Metabolic Labeling.

RNA velocity inference without metabolic labeling is often inaccurate (22), but incorporating metabolic labeling into scRNA-seq remains an experimental challenge (43). To test the effectiveness of *TopicVelo*, we applied it to human hematopoiesis data from a recent study in which RNA velocity was extended to leverage single-cell metabolic labeling techniques that distinguish newly synthesized versus preexisting transcripts (22). The published analysis reconstructed a complex, multifurcating trajectory of transitions, which *scVelo* fails to capture. Using *TopicVelo* on the data without the metabolic labels, we inferred the correct transitions, including streamlines that accurately delineate the trajectories of monocytes, basophils, erythrocytes, and megakaryocytes (Fig. 2*A*).

To obtain a global transition matrix, we first performed topic modeling (35, 36), resulting in an 8-topic model that identifies gene programs associated with known cell types (topics 1 and 3) and heterogeneous cell states during differentiation (*SI Appendix*, Fig. S2 and Table 1). For example, megakaryocyte-associated topic 3 appropriately features the gene *F13A1*, a subunit of plasma factor XIII known to be produced by megakaryocytes (42) (*SI Appendix*, Figs. S2*D* and S3*A*). Though a global phase plot of *F13A1* indicates little transcriptional activity, focusing on cells with highest weight in topic 3 brings the dynamical features of *F13A1* into relief (Fig. 2*B*).

Based on the burst model, *TopicVelo* then inferred topic-specific kinetic parameters for topic-specific genes. By assuming a steady state can be approximated by the joint distributions of spliced and unspliced counts of topic-specific genes in topic-associated cells, *TopicVelo* substantially improved upon the parameter estimates inferred from the one-state model underlying *scVelo*. For example, it more accurately recovered the experimental joint distribution of *F13A1* over topic-3 high cells (Fig. 2*C*). Indeed, while velocities of topic-3 specific genes *F13A1*, *PLEK*, and *ZYX* were inferred to be negative by *scVelo*, *TopicVelo* inferred them to be positive, consistent with experimental evidence that these genes are up-regulated during megakaryocytic differentiation (44, 45) (*SI Appendix*, Fig. S3 *A–C*). Similarly, whereas *scVelo* inferred downregulation of the basophil-associated, topic-1 specific genes *GATA2* and *HPGD*, *TopicVelo* predicted their upregulation in the basophil lineage, consistent with previous experiments showing that *GATA2* is critical for basophil development (46) and *HPGD* is enriched in basophils (47) (*SI Appendix*, Fig. S3 *D* and *E*). Using the inferred topic-specific signals, *TopicVelo* then created topic-specific transition matrices, whose corresponding streamlines were consistent with those inferred for the same regions using metabolic labeling data (Fig. 2*D*).

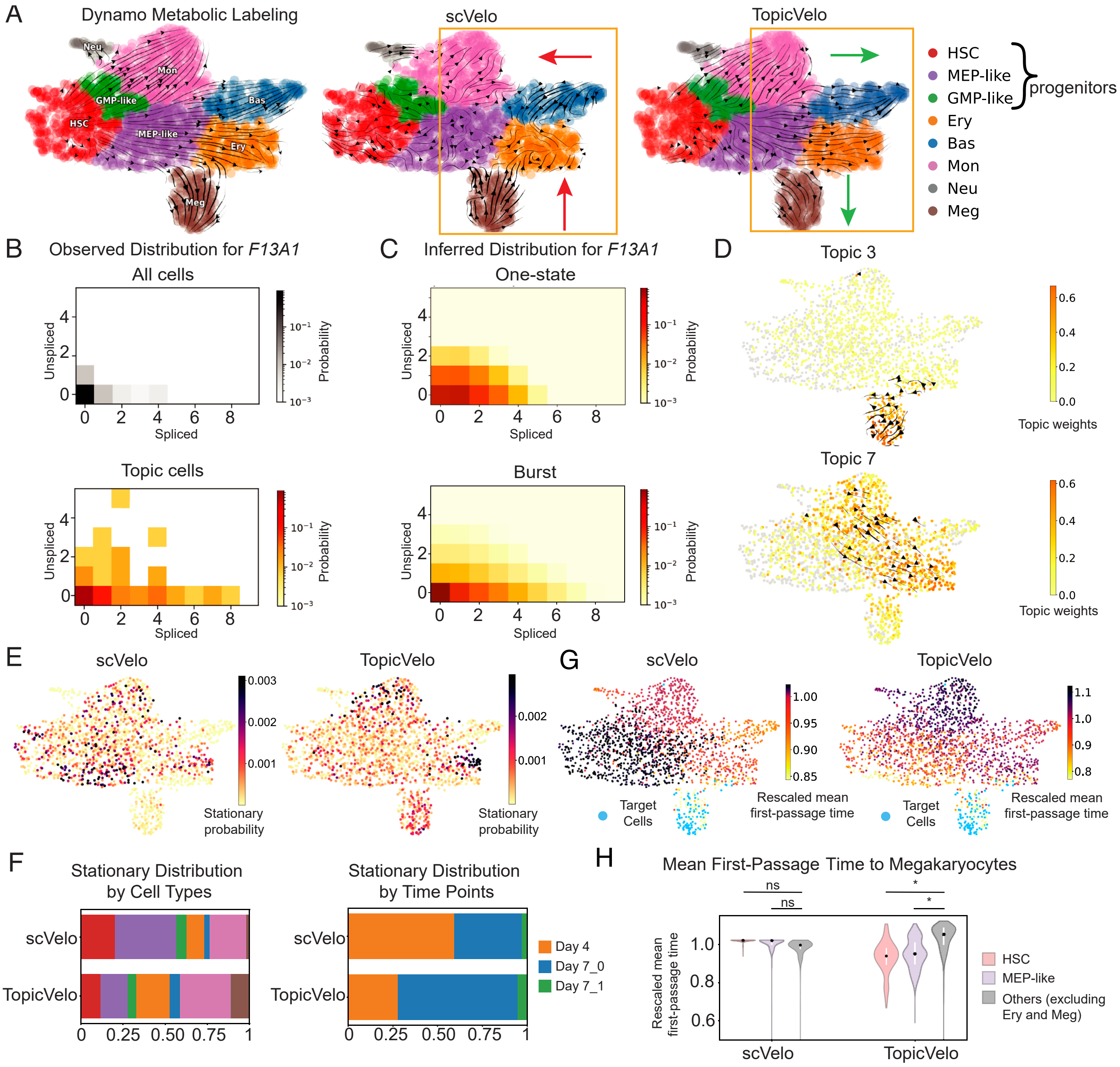

**Fig. 2.** *TopicVelo* inferred multifurcating trajectories of human hematopoiesis whose recovery previously required metabolic labeling. (*A*) Previously published (22) UMAP embedding of hematopoiesis data shows cells colored by annotated progenitor (HSC, hematopoietic stem cell; MEP-like, megakaryocyte and erythrocyte progenitor; GMP-like, granulocyte and monocyte progenitor) and terminal (Ery, erythrocyte; Bas, basophil; Mon, monocyte; Neu, neutrophil; Meg, megakaryocyte) cell types. Streamlines (arrows) were inferred either with metabolic labeling, by *Dynamo* (*Left*), or without it, by the *scVelo* dynamical model (*Middle*), and by *TopicVelo* with an 8-topic model (*Right*); *TopicVelo* but not *scVelo* captures key cell-type differentiation (green versus red arrows). (*B*) Plots show the experimental joint distribution of spliced and unspliced mRNA counts in all cells, or cells with highest weight in topic 3, of the topic-3 specific gene *F13A1*, which is known to be expressed in megakaryocytes (42). (*C*) Plots show the joint distribution of *F13A1* in topic-3 high cells, inferred using the one-state model, or maximum likelihood estimates for the burst model; the latter better captures both the diffuseness of the joint distribution and the empirical concentration at (0,0). (*D*) Topic-specific streamlines obtained from topic-specific transition matrices for topics 3 and 7, respectively. The color bar indicates the topic weights for cells used in the parameter inference. The topic-3 plot demonstrates transitions into mature megakaryocytes, and the topic-7 plot suggests transitions into erythroid. (*E* and *F*) *TopicVelo* identified terminal states missed by *scVelo*. UMAPs (*E*) show stationary probabilities for *scVelo* (*Left*) and *TopicVelo* (*Right*) transition matrices, which are summarized in bar charts (*F*) by cell type (*Left*, colored as in panel *A*) and by time point (*Right*) that highlight relatively high probabilities from *TopicVelo* for terminal cell types, such as megakaryoctyes, versus progenitor cell types, and for late versus early time points. (*G* and *H*) *TopicVelo* estimated shorter transition times for true differentiation pathways. UMAPs (*G*) show mean first-passage times to megakaryocytes (Target, blue), rescaled by median, based on *scVelo* (*Left*) and *TopicVelo* (*Right*); summary violin plots (*H*) highlight shorter transition times from progenitors versus others estimated by *TopicVelo*, but not scVelo. (Black dot: median, white vertical lines: 25th to 75th percentile.) $^*P < 0.0001$ by one-sided permutation test; ns, $P \geq 0.1$.

Finally, these topic-specific transition matrices were integrated to obtain the global transition matrix and corresponding streamlines (Fig. 2*A*). To quantitatively evaluate the quality of inference by *TopicVelo*, we computed the stationary distribution as a proxy for identifying terminal states. While both *scVelo* and *TopicVelo* assigned relatively high stationary probabilities to erythroid cells and monocytes, *TopicVelo* additionally recognized megakaryocytes as terminal states (Fig. 2*E*). Furthermore, aggregation of the stationary probabilities by cell types illustrated that, compared to *scVelo*, *TopicVelo* suggested higher stationary probability for terminal cell types and lower probability for progenitors, consistent with the expected cell-fate transitions (Fig. 2*F*).

To investigate the differentiation dynamics and trajectories, we used MFPT analysis to gauge the identities of ancestral subpopulations and assess the likelihood of subpopulations transitioning into terminal states. For instance, we computed MFPTs to megakaryocyte-like cells and observed that the MFPTs derived from *scVelo* versus *TopicVelo* displayed very different trends (Fig. 2 *G* and *H*). In particular, *TopicVelo* estimated lower MFPTs for progenitors than for other, nonmegakaryocyte terminal cell types, whereas *scVelo* estimated the opposite. The inference from *TopicVelo* agrees better with established biological understanding that megakaryocytes originate directly from progenitors, rather than from other terminally differentiated populations.

Collectively, these results demonstrate the capacity of *TopicVelo* to identify biologically meaningful dynamic genes, infer more biologically accurate RNA velocity, and provide more meaningful insights into the terminal states and trajectories of differentiation.

***TopicVelo* Recovers Complex Developmental Trajectories in Mouse and Human.** Several studies have observed that certain genes exhibit developmental-stage–dependent transcription rates, termed "multiple rate kinetics (MURK)" (7–10, 28). *scVelo* does not account for this stage dependency and erroneously produced reversed streamlines for mouse erythropoiesis when MURK genes were included in the data (8). In contrast, *TopicVelo* produced the correct trajectories in this setting (Fig. 3*A*). A stationary distribution analysis further confirmed the streamline visualization; whereas *scVelo* falsely identified intermediate erythroid stages as terminal states, *TopicVelo* results suggested that essentially all of the stationary probability is in the erythroid-3 cell state (Fig. 3*B*). To investigate the relative proportion of cell–cell transitions entering versus leaving a terminal cell type, we computed the relative flux between cells at different developmental stages (*Materials and Methods* and *SI Appendix*, section 1). *TopicVelo* predicted overall positive flow toward more mature erythroid cells, while *scVelo* predicted negative flow in later stages (Fig. 3*C*).

Biologically informative results were achieved by *TopicVelo* using a model with two topics, which accurately captured expression patterns during the maturation of blood progenitors to erythroid cells (*SI Appendix*, Fig. S4 and Table S1). Topic 0 has weights increasing across the developmental process and features the archetypal red blood cell genes *Hba-x* and *Hbb-y* (8), and their unspliced counterparts, as well as *Smim1*, which influences red blood cell traits (48) (Fig. 3*D* and *SI Appendix*, Fig. S4*A*). Inversely, topic-1 weights decrease across the developmental process, as does the expression of topic-1 specific genes, such as *Gata2*, *Fn1* and *Fscn1* (Fig. 3*E* and *SI Appendix*, Fig. S4*B*). These results corroborate previous observations that *Gata2* is highly expressed in progenitors, with expression declining after erythroid commitment (49), and that *Ccnd2* expression is anticorrelated with erythroid progression (50).

We then turned to a challenging setting of human hematopoietic stem cell (HSC) differentiation (28). *TopicVelo* used a 10-topic model to recover the expected trajectories and identify key genes involved in cell-fate commitments, without the prior knowledge of the starting state required by pseudotime inference (*SI Appendix*, Fig. S5). Unlike *scVelo*, *TopicVelo* did not infer erroneous reversals in directionality (Fig. 3*F*). The stationary distribution analysis confirmed that *scVelo* incorrectly identified early stage HSCs as terminal states, whereas the stationary probability derived from *TopicVelo* was predominantly associated with true terminal states (Fig. 3 *G* and *H*).

The inferred topics characterized different stages of development and identified key, lineage-specific genes, leading to velocity predictions that are more consistent with known biology. For example, topic six is relatively high in erythroid cells and includes the gene *KLF1*, previously shown to be associated with erythroid commitment (28). In contrast to *scVelo* predictions that early erythroid cells (Ery 1) down-regulate *KLF1*, *TopicVelo* accurately predicted that they up-regulate *KLF1* (Fig. 3*I*). *TopicVelo* also highlighted several other patterns previously observed in the literature, including upregulation of *MPO* during monocyte commitment (28) (Fig. 3*J*), upregulation of *CA1* in the peripheral blood erythroid cells (51), association of *IRF8* with monocyte development and dendritic cell function (52), expression of *SELP* during megakaryocyte development (53), downregulation of *CRHBP* in HSCs during differentiation (54), and expression of the chemotactic gene *AZU1* in monocytes (55) (*SI Appendix*, Fig. S6 and Table S1).

To investigate the performance of *TopicVelo* on additional examples, we applied it to human dentate gyrus (29) and murine pancreatic endocrinogenesis (30) datasets on which *scVelo* was tested (7) (*SI Appendix*, Figs. S7–S10 and sections 3 and 4). In these data, *TopicVelo* infers the appropriate transitions as well as offers insights into key genes in rare cell types.

Together, the results show that *TopicVelo* outperforms the state-of-the-art in complex settings, recovering biologically accurate trajectories and highlighting informative genes.

***TopicVelo* Predicts Bidirectional and Convergent Immune Responses of Innate Lymphoid Cells.** An important motivation for developing *TopicVelo* was to meet the challenge of analyzing complex immune responses, including those involving unconventional trajectories, such as convergence on one cell state from multiple origins and functional plasticity between cell types (31, 56, 57). With different gene programs involved in conversions in opposite directions between cell types, traditional approaches to RNA velocity and trajectory inference do not reveal such intricacies. In our previously published study of skin-resident innate lymphoid cells (ILCs) from a mouse model of psoriasis (31), the ILC transcriptional states and their trajectories during the immune response are modeled by leveraging scRNA-seq data collected from mice killed at five time points (days 0 to 4) during induction of inflammation (Fig. 4*A*). The study's detailed analysis, which combines topic modeling with density-based pseudotime inference, and extensive experimental validations, demonstrates multiple possible transitions to a pathogenic ILC3-like state. These include an ILC2-ILC3 transition, confirmed using a transgenic fate-mapped mouse, which may occur via two routes, as well as a quiescent-ILC3 transition and possibly bidirectional quiescent-ILC2 transition (Fig. 4*B*).

To assess the capability of *TopicVelo* to predict these complex immune response trajectories, without information from multiple time points or specification of root and terminal states, we used data from day 3 only. First, we verified that the cells in a thin bridge connecting ILC2 and ILC3 cells are unlikely to represent doublets (*SI Appendix*, Fig. S11 and section 5). Next, we performed topic modeling to obtain a 10-topic model that was consistent with the published analysis (Fig. 4 *C*–*E* and

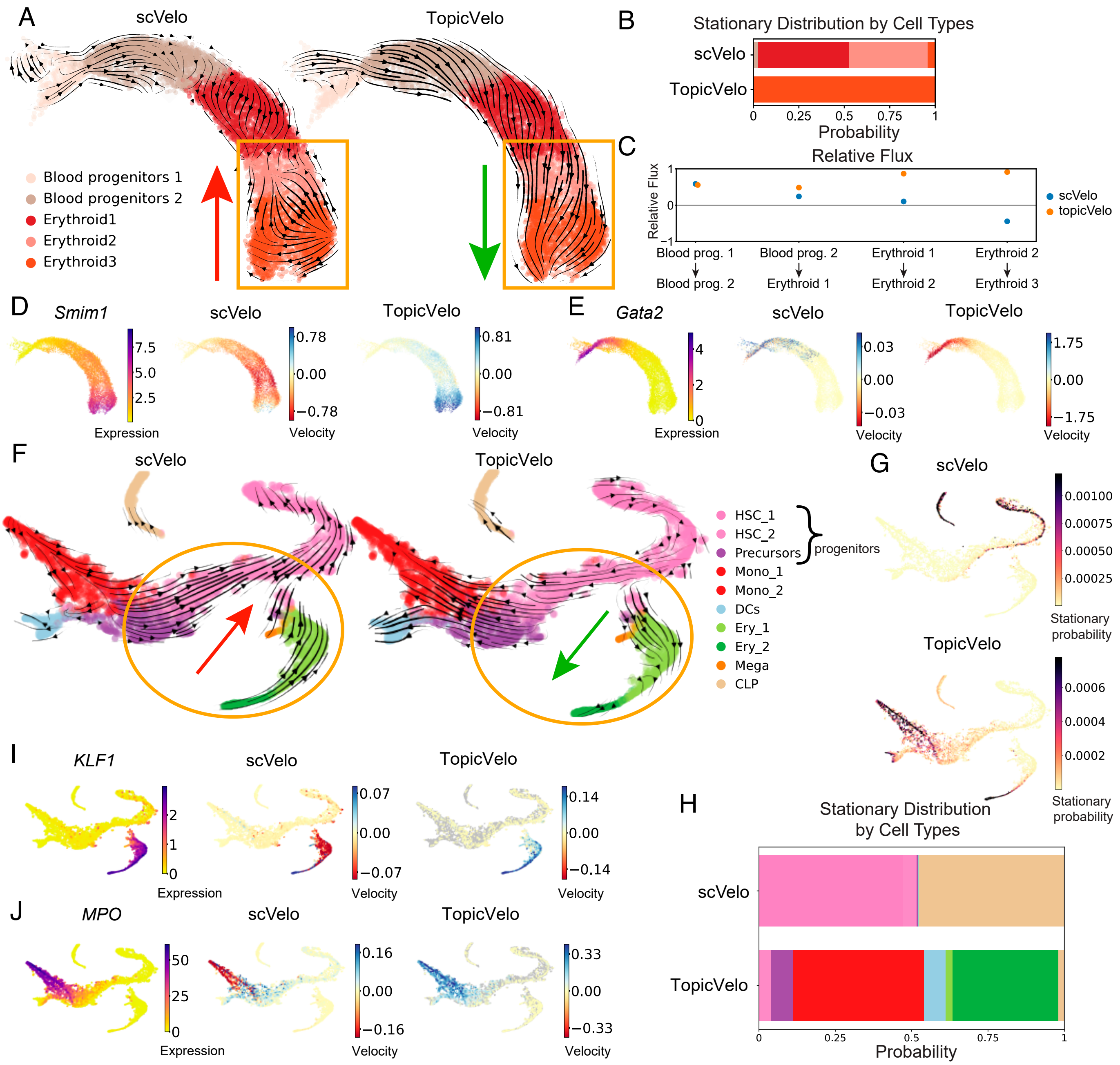

**Fig. 3.** *TopicVelo* correctly captured mouse erythropoiesis and human bone marrow development trajectories. (*A* and *B*) *TopicVelo* accurately identified erythroid 3 as a terminal state. Previously published (8, 10) UMAP embedding of cells in erythropoiesis (*A*), colored by cell-type annotation, shows streamlines (arrows) inferred by the *scVelo* stochastic model (*Left*), which erroneously suggests differentiation of erythroid 3 into erythroid 2 cells (red arrow), or by *TopicVelo* (*Right*), which recovers the expected differentiation trajectory (green arrow). Bar charts (*B*) show the stationary probability distributions for each method (row), aggregated and colored by cell type. (*C*) *TopicVelo* predicted positive flux toward more mature erythroid cells, whereas *scVelo* predicted negative flux. For each method (color), the plot (*C*) shows the relative flux (*y* axis) between pairs of cell subpopulations in the direction of the arrow (*x* axis). (*D* and *E*) UMAP plots for topic-specific genes *Smim1* (*D*) and *Gata2* (*E*), with cells colored by smoothed gene expression (*Left*) and by velocities (negative, red; positive, blue), as inferred by *scVelo* (*Middle*) or *TopicVelo* (*Right*). (*F*–*H*) *TopicVelo* correctly discovered terminal cell types in human bone marrow development. Previously published (28) *t*-SNE plot of cells from human bone marrow, colored by annotated cell type (*F*), shows streamlines inferred by *scVelo* stochastic model (*Left*), which incorrectly predicted that precursors, megakaryocytes (Mega), and erythrocytes (Ery) differentiate into hematopoietic stem cells (HSC) (red arrow), or by *TopicVelo* (*Right*), using 10 topics, which recovered the expected trajectories for all major lineages (green arrow). (Mono: monocyte, DC: dendritic cell, CLP: common lymphoid progenitor.) The *t*-SNE plots show cells colored by stationary probability (*G*) as inferred by *scVelo* (*Top*) or *TopicVelo* (*Bottom*). Bar charts (*H*) show the stationary probability distributions for each method (row), aggregated by cell type (color, as in panel *F*). (*I* and *J*) *TopicVelo* gave markedly different velocity results from those of *scVelo* for topic-specific genes. For the erythroid-associated gene *KLF1* (*I*) and monocyte-associated gene *MPO* (*J*), *t*-SNE plots show cells colored by smoothed gene expression (*Left*) and by velocities, as inferred by *scVelo* (*Middle*) or *TopicVelo* (*Right*).

*SI Appendix*, Fig. S12 and Table S1). In particular, topic 4 is strongly associated with the ILC3-like cells and characterized by proinflammatory, ILC3- and $T_H$17-associated genes, such as *Il17a*, *Il23r*, *Gzmb*, and *Il1r1* (58) (Fig. 4*C*). Topic 6 features a gene program previously identified as "quiescent-like" (31), including *Klf2*, a transcription factor associated with T cell quiescence (59) (Fig. 4*D*). Topic 9 features ILC2- and $T_H$2–associated genes, such as *Il1rl1* (ST2, the receptor for IL-33) (58),

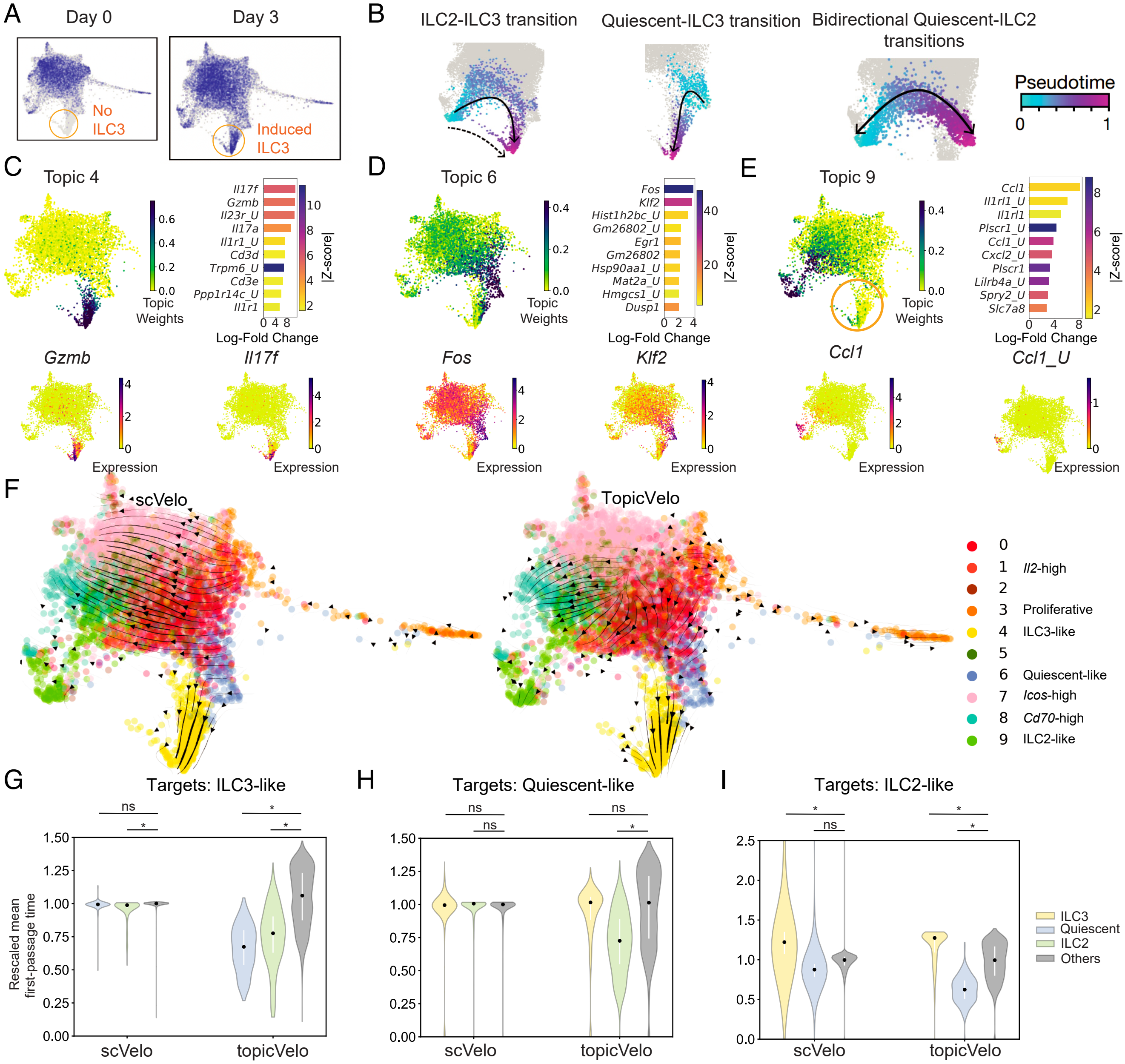

**Fig. 4.** Using data from only one of five time points, *TopicVelo* revealed complex transitions underlying the inflammatory responses of skin ILCs. (*A* and *B*) Previously published force-directed layout (FDL) embedding of scRNA-seq profiles of skin ILCs from a mouse model of psoriasis, colored by day of collection (*A*) and by pseudotime (*B*), as previously inferred via diffusion-based trajectories (panels, *B*), with directionality (arrows) imposed by the presence of ILC3-like cells (orange circle, *A*) on day 3 but not day 0 (31). (*C–E*) Highlights of three topics from a 10-topic model of both spliced and unspliced mRNA transcripts from only day-3 cells. For ILC3-like topic 4 (*C*), quiescent-like topic 6 (*D*), and ILC2-like topic 9 (*E*), the FDL plots (as in panel *A*) show only day-3 cells, colored by topic weight (*Top Left*) and by log-normalized expression of topic-specific genes (*Bottom Left*, *Right*), and the bar chart (*Top Right*) shows the top 10 topic-specific genes by largest log-fold change, colored by z-score ('_*U*' appended to gene symbol indicates unspliced transcript). A subset of induced cells have relatively high topic weights for both topics 4 and 9 (orange circle, *E*). (*F–I*) *TopicVelo* disentangled simultaneous but distinct dynamics of ILC responses. FDL plots of day-3 cells, colored by most strongly associated topic (*F*), show streamlines (arrows) from the *scVelo* dynamical model (*Left*) or *TopicVelo* (*Right*), using the topic model shown. Focusing on transitions to ILC3-like cells (yellow, high in topic 4), streamlines suggest that both methods predicted the transition from quiescent-like cells (blue, high in topic 6), but only *TopicVelo* correctly predicted the experimentally validated transition from ILC2-like cells (green, high in topic 9). Violin plots show the distributions of median-rescaled mean first-passage times, estimated using *scVelo* and *TopicVelo* (*x*-axis), from different groups of nontarget cells (colors) to different target populations: ILC3-like (*G*), quiescent-like (*H*), and ILC2-like (*I*) cells. Smaller values indicate faster inferred transition times, suggesting better support for that biological transition. (Black dot: median; white vertical line: 25–75th percentile.) *$P < 0.0001$ by one-sided permutation tests; ns, $P \geq 0.1$.

as well as chemokines, such as *Ccl1* and *Cxcl2*, and their unspliced counterparts (Fig. 4*E*).

Though the RNA velocity analyses of these data by both *TopicVelo* and *scVelo* suggested a quiescent-ILC3 transition (Fig. 4*F*) and predicted the observed downregulation of *Klf2* and *Fos* (31) during the transition (*SI Appendix*, Fig. S13 *A* and *B*), only *TopicVelo* revealed the transition path of the biologically important ILC2-ILC3 trajectory or suggested a possible bidirectional quiescent-ILC2 transition (Fig. 4*F*). We found that discrepancies between *TopicVelo* and *scVelo* results were at least partly due to differences in velocity estimates. For example, the observed upregulation of *Il23r*, *Il1r1*, and *Lgals3* during ILC3 response (31) was more faithfully captured by *TopicVelo* than *scVelo* (*SI Appendix*, Fig. S13 *C–E*).

To quantitatively confirm these intertwined transitions, we computed rescaled mean first-passage times (rMFPT) to different target cell populations. First, we used cells very strongly associated with the ILC3-like gene programs as target cells. The rMFPTs derived from *scVelo* show little variation across cells, whereas results from *TopicVelo* showed a clear distinction suggesting that, relative to transitions from other populations, the quiescent-ILC3 and ILC2-ILC3 transitions may both occur at relatively fast timescales (Fig. 4*G* and *SI Appendix*, Fig. S14 *A* and *B*). For quiescent-like cells as the target group, both methods agreed that a reverse ILC3-quiescent transition was unlikely. However, *TopicVelo* suggested a possible ILC2-quiescent conversion (Fig. 4*H* and *SI Appendix*, Fig. S14 *C* and *D*). Finally, for ILC2-like cells as targets, both methods again agreed that a reverse ILC3-ILC2 transition is unlikely. *TopicVelo* also specifically identified a transition from quiescent-like cells to ILC2s as significantly more likely than transitions to ILC2s from other populations (Fig. 4*I* and *SI Appendix*, Fig. S14 *E* and *F*). Our analysis of the potential bidirectional quiescent-ILC2 transitions suggests that the most likely trajectories in each direction occur through different but overlapping parts of transcriptional space (*SI Appendix*, Fig. S15 and section 6).

While the analysis of the day 3 data demonstrates that *TopicVelo* can infer immune response dynamics without requiring a time course, we also investigated the dynamics inferred for other days (*SI Appendix*, Figs. S16–S18 and section 7). We found good consistency, particularly between days 3 and 1, with greater differences between those days and day 2 or 4 (*SI Appendix*, Figs. S17 and S18), which could be caused by batch effects or interday fluctuations in immune response dynamics.

Taken together, our results demonstrate the effectiveness of *TopicVelo* in the analysis of immune responses, where cells may exhibit functional plasticity or reflect varying contributions of simultaneous, very distinct, dynamic processes.

## Discussion

RNA velocity inference has recently been improved via different machine learning techniques (16–22, 60, 61), but challenges remain. We present *TopicVelo*, a method and framework for RNA velocity that improves on the state of the art and conceptually complements other approaches. Existing methods typically include genes based on their fit to a velocity model (7, 19–21), making strong assumptions about a globally determined steady state and potentially excluding genes that are informative for locally dynamic processes. In contrast, by using topic modeling to discover biologically relevant gene programs or processes ("topics") and the cells in which their activity levels are relatively high, *TopicVelo* hones in on genes that are informative for the kinetic parameters for different processes, while preventing cells that are not associated with a process from distorting its parameter estimates. To provide a global view of cell-state transitions, *TopicVelo* leverages the probabilistic topic weights to integrate process-specific transition matrices into a unified transition matrix. The number of topics, which ideally reveals granular processes without compromising statistical power, can be selected using a combination of measures to assess the quality of topic models and biological interpretability (*Materials and Methods* and *SI Appendix*, sections 1 and 8). Our detailed analysis shows that results from *TopicVelo* are robust to the exact choice, provided the number of topics lies in an appropriate regime (*SI Appendix*, Fig. S19 and section 8). Future work may incorporate the use of hierarchical Dirichlet processes (62), which infer the number of topics from data in an unsupervised fashion.

*TopicVelo* infers gene-specific parameters of a transcriptional burst model by efficiently estimating the full joint distribution of unspliced and spliced gene counts given by a chemical master equation, thus explicitly accounting for higher-order moments. In contrast, the leading method *scVelo* (7) and others (18, 19, 21, 22, 60), which infer kinetic parameters based on ordinary differential equations (ODEs) from counts smoothed across cell neighborhoods in the $k_G$-nearest neighbors ($k_G$-NN) graph, can distort second- or higher-order moments (11). A recent study also incorporated a global burst model, fit via numerical gradient descent, rather than the simplex-based optimization in *TopicVelo*, but focused on analyzing the effects of gene-length–dependent capture rates of unspliced RNA (15). To assess how the burst model and topic modeling each contribute to *TopicVelo* performance, we performed an algorithmic ablation study (*SI Appendix*, section 9). We found that the ablative approaches, i.e., the burst model without topic modeling and topic modeling (combined with *scVelo*) without the burst model, offer different improvements, though none as remarkable as those achieved by their combination (*SI Appendix*, Fig. S20). The flexible conceptual framework of *TopicVelo* allows future incorporation of more sophisticated topic models (63) and transcriptional models (e.g., ref. 64).

A critique (12) of *scVelo* notes that excessive smoothing can lead to a potentially problematic dependence of parameters, especially in the dynamical model, on the global $k_G$-NN graph structure and the visualization embedding. *TopicVelo* circumvents this issue at the gene level by inferring kinetic parameters from unsmoothed counts. Furthermore, by computing a $k_G$-NN for each topic, *TopicVelo* loosens the coupling between the transition matrix and UMAP embedding. Like *scVelo*, *TopicVelo* uses the inferred velocity matrix and matrix of differences of smoothed spliced counts to compute transition probabilities, but the *TopicVelo* framework also naturally permits the computation of (noisier) transition probabilities from differences of unsmoothed counts.

Using its dissection-then-integration approach, *TopicVelo* inferred robust, accurate dynamics in complex systems, including functionally plastic immune responses and multifurcating differentiation, without requiring multiple time points or the support of metabolic labeling. To use information from ordered time points, a future extension could add weak penalties to transitions in the integrated transition matrix between cells from later time points to cells from earlier time points, similar in spirit to a biased diffusion approach (e.g., ref. 65). A more intricate potential extension could capture topic evolution using a dynamic topic model (66).

The combination of topic modeling with a steady-state transcriptional model may allow *TopicVelo* to implicitly handle some non-steady-state contexts. Relaxing the steady-state assumption in the chemical master equation framework to fully account for transient states presents considerable challenges. Recent work (11) considers the applicability of general master equations to RNA velocity inference, but not their efficient implementation. Future research may focus on efficiently sampling from the distributions of transient states and inferring time-dependent stochastic dynamics.

Challenges for the future also include developing methods that merge the advantages of *TopicVelo* with other recent, complementary advances, such as batch correction, improving and removing biases in transcript quantification (15, 67), a Bayesian deep generative framework for quantifying statistical uncertainty, which was developed for ODE velocity models (18, 60, 61), greater robustness from postprocessing noisy velocity vectors using representation learning (16, 68), and multiomic data and models (13, 14).

The interpretation of RNA velocity data represents another set of challenges. Traditional approaches heavily rely on streamline visualizations and pseudotime, which may be inadequate or misleading. Our detailed discussion of various quantitative measures (*Materials and Methods* and *SI Appendix*, section 1) may help practitioners more confidently interpret RNA velocity. In the vein of our use of fundamental Markovian techniques to quantitatively assess transition matrices, future work may borrow ideas from nonequilibrium statistical mechanics and relevant sampling frameworks (69), potentially leading to more reliable tools to provide mechanistic insights into cell-state transitions. More broadly, *TopicVelo* provides a potential framework for developing more sophisticated RNA velocity methods, while serving as a valuable biological tool to accurately infer the dynamics of interpretable gene programs and cell-state transitions in diverse systems.

## Materials and Methods

**Topic Modeling and Differential Expression Analysis.** We use the tomotopy Python package (70) to efficiently infer topic models for a range of values of $K$, the number of topics. After evaluating those results to select a final value for $K$, we use the FastTopics R package (35) to infer the final model and compute topic-specific differentially expressed genes (36) (*SI Appendix*, section 1). For optimized $K$, the above procedures were performed as follows:

```
  topic_model_fit <- fit_topic_model(count_matrix, k=K)
  de_results <- de_analysis(topic_model_fit,
      count_matrix)
```

where the input count matrix is constructed by stacking the raw spliced count matrix and the raw unspliced count matrix for top 2,000 highly variable genes.

**Topic Modeling Evaluation Metrics.** To estimate the optimal number $K$ of topics, we computed established metrics, including average cosine distance (37), information divergence (39), and topic coherence (38) (*SI Appendix*, section 1) on topic models inferred using tomotopy (70) for a range of values of $K$. For each dataset, at least one of these metrics plateaued as a function of increasing values of $K$, and we selected the smallest value of $K$ in the intersection of those regimes across metrics. To prevent overfitting, we also considered the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (*SI Appendix*, section 1).

Another criterion was interpretability, i.e., a reasonable number of potentially biologically meaningful differentially expressed genes. For most datasets, "topic-specific genes" were selected from the differentially expressed genes for downstream analysis (e.g., RNA velocity) if, for either the spliced or unspliced form, the local false sign rate (lfsr) was at most 0.001 and the log fold change (LFC) was at least 0.5 in absolute value. This criterion is a very conservative estimate of differential expression and, in practice, produces 50 to 250 topic genes for each topic.

**RNA Velocity Parameter Estimation Via the One-State Model.** The one-state transcription model is governed by this master equation:

$$\begin{aligned}\frac{\partial p(u,s,t)}{\partial t} =&\alpha\Big[p(u-1,s,t)-p(u,s,t)\Big]\\ &+\beta\Big[(u+1)p(u+1,s-1,t)-up(u,s,t)\Big]\\ &+\gamma\Big[(s+1)p(u,s+1,t)-sp(u,s,t),\Big]\end{aligned} \qquad [2]$$

where $\alpha$ is the rate of transcription, $\beta$ is the splicing rate, and $\gamma$ is the degradation rate. Previous work showed that the steady-state distribution when $\beta \neq \gamma$ is the product of two independent Poisson distributions for $u$ and $s$ respectively (71). By identifying the maximum likelihood estimates for observing the transcriptional profiles of cells at steady state, we obtained $\frac{\gamma}{\beta} = \frac{\langle u\rangle}{\langle s\rangle}$ 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 (*SI Appendix*, section 1).

**RNA Velocity Parameter Estimation Via the Geometric Burst Model.** To estimate the steady-state joint distributions, we implemented a Gillespie algorithm (40) to simulate the master equation (Eq. **1**) in Python, accelerated via Numba (72). For a trajectory with burn-in period $t_{\text{burn-in}}$ (i.e., before the system converges to a steady state) and total simulation time $t_{\text{total}}$, the probability $p(u,s)$ of observing a cell with $u$ unspliced and $s$ spliced transcripts for a given gene in the steady state is

$$p(u,s) = \frac{1}{t_{\text{total}} - t_{\text{burn-in}}} \int_{t_{\text{burn-in}}}^{t_{\text{total}}} \delta(u,s,t)\, dt, \qquad [3]$$

where $\delta(u,s,t) = 1$ if the cell has $u$ unspliced counts and $s$ spliced counts at time $t$, and $\delta(u,s,t) = 0$ otherwise.

We initialize the kinetic parameters with the method of moments, which was previously derived (15, 27):

$$\hat{b} = \frac{\langle u^2\rangle}{\langle u\rangle} - 1, \quad \hat{k}_{\text{on}} = \frac{\langle u\rangle}{\hat{b}}, \quad \hat{\gamma} = \frac{\langle u\rangle}{\langle s\rangle}, \qquad [4]$$

where the moments are estimated from the observed distribution. Then, to find the optimal kinetic parameters, the KL divergence is minimized using the Nelder–Mead algorithm implemented in SciPy (41). In some cases, the method of moments estimate is a local minimum that is close to the global minimum, and the optimizer can get stuck. In this case, we used $3\hat{b}$, $\hat{k}_{\text{on}}/3$, and $\hat{\gamma}$ to restart the search for the global minimum. The convergence criterion was chosen to be a relative change in KL divergence between two subsequent iterations smaller than 1/1,000 or reaching a maximum number of iterations. To verify the performance and robustness of this inference scheme, we performed detailed analysis on both simulated data and real biological datasets. Our results indicated that the approach is efficient, recovers the ground truth on simulated data, and outperforms the one-state model for real data (*SI Appendix*, section 1 and Fig. S1).

**Determination of Topic-Associated Cells.** *TopicVelo* uses the cells associated with a topic to analyze the steady state for that topic (*SI Appendix*, section 1). While one approach for choosing a topic weight threshold is to associate each cell with the topic for which it has the highest weight, which discretely clusters the cells, this has several drawbacks: 1) A cell may have relevant information about a topic for which it does not have the highest weight; 2) the cells assigned to a topic may not capture the full dynamic range of the topic-associated process; and 3) there is no potential for transitions between cells assigned to different topics.

In general, we instead used the following procedure to identify a reasonable range for the choice of topic weight threshold as a topic weight percentile. For a given topic $k$, we denote by $A_{n,k}^{+}$ the set of cells with topic-$k$ weights above the $n$th-percentile, and by $A_{n,k}^{-}$ the set of cells with topic-$k$ weights at most the $n$th-percentile. Note that $A_{n,k}^{+} \bigcup A_{n,k}^{-} = A$ where $A$ is the set of all cells. For integers $n$ from 1 to 99, we compute what we call an *average rescaled KL divergence*, denoted by $D_{n,k}^{+}$ as follows: For each topic-specific gene, we compute the KL divergence of the joint $u$-$s$ distribution of $A_{n,k}^{+}$ to that of $A$, and rescale the divergence to $[0,1]$; then, we average the rescaled KL divergences over the genes. We perform an analogous procedure to compute $D_{n,k}^{-}$, the average rescaled KL divergence for the distribution from $A_{n,k}^{-}$ to that of $A$. $D_{n,k}^{+}$ approaches 0 as $n$ approaches 0. We observed a sharp decline in $D_{n,k}^{+}$ at a relatively large value of $n$, which we denote by $n_k^{+}$. If the topic weight threshold is chosen in the regime $n > n_k^{+}$, the full dynamic range of topic-associated process is not properly accounted for.

Similarly, $D^{-}_{n,k}$ approaches 0 as $n$ approaches 100, and a sharp decline in $D^{-}_{n,k}$ is observed for a relatively small value of $n$ denoted by $n^{-}_{k}$. Topic weight thresholds in the regime $n \leq n^{-}_{k}$ risk including cells not meaningfully associated with the topic-associated process. The interval $[n^{-}_{k}, n^{+}_{k}]$ is a natural and simple heuristic for the range of suitable thresholds for topic $k$. For the majority of topics and datasets, we observed $[n^{-}_{k}, n^{+}_{k}] = [30, 70]$ to be a range in which both $D^{-}_{n,k}$ and $D^{+}_{n,k}$ were relatively flat, though in other cases, corresponding to a rare cell type or very distinct process, the range was around $[75, 95]$.

**Construction of Topic-Specific Transition Matrices.** While we use unsmoothed counts for kinetic parameter inference, we compute the transition flows on smoothed counts to remove noise in the visualization. We did not observe significant distortions in the overall trends using smoothed versus unsmoothed counts. For cell $i$ and gene $m$, the first moments $\tilde{u}_{im}$ and $\tilde{s}_{im}$ represent the smoothed counts, computed as the number of unspliced and spliced transcripts, respectively, averaged over the cells in the neighborhood of $i$ in the $k_G$-NN graph with $k_G = 30$, computed from the top 30 principal components (PCs) of the global PCA of the log-normalized spliced expression matrix.

The velocity vector for cell $i$ associated to topic $k$ is $\tilde{\mathbf{v}}_{i,k} = (\tilde{v}_{i1,k}, \tilde{v}_{i2,k} \ldots, \tilde{v}_{iM_k,k})$, for topic-specific velocity vector $\tilde{v}_{im,k}$ defined as $\tilde{v}_{im,k} = \tilde{u}_{im} - \gamma'_{m,k}\tilde{s}_{im}$ for gene $m$, where $M_k$ is the number of topic-specific genes, and $\gamma'_{m,k}$ is the topic-specific degradation rate for gene $m$. Across small neighborhoods in the $k_G$-NN graph, the first moments of the smoothed data are not as distorted as higher-order moments, and the velocity $\tilde{\mathbf{v}}_{i,k}$ is a reasonable smoothed approximation.

Then, a cosine similarity between the velocity vectors and the differences in spliced expression can be computed, as previously (7):

$$\tilde{p}_{ij,k} = \cos(\tilde{\mathbf{s}}_{j,k} - \tilde{\mathbf{s}}_{i,k}, \tilde{\mathbf{v}}_{i,k}), \tag{5}$$

where $\tilde{\mathbf{s}}_{i,k}$ is the vector of smoothed spliced counts in cell $i$ for topic-$k$ specific genes.

For each topic $k$, a topic-specific $k_G$-NN graph is constructed on the topic-associated cells using the top 30 PCs of the global PCA. The topic-specific transition probability $p_{ij,k}$ from cell $i$ to $j$ for topic $k$ is obtained by applying an exponential kernel to the cosine similarities over the set $N_k(i)$ of cells in the topic-specific neighborhood of cell $i$:

$$p_{ij,k} = \frac{1}{z_{ik}} \exp\left(\frac{\tilde{p}_{ij,k}}{\sigma^2}\right), \tag{6}$$

where $\sigma$ is kernel width, and the normalization factor $z_{ik}$ is $z_{ik} = \sum_{j \in N_k(i)} \exp\left(\frac{\tilde{p}_{ij,k}}{\sigma^2}\right)$.

**Integration of Process-Specific Dynamics.** Because the topic-associated cells and global set of cells may have different indices, we switch to using $c$ to denote the identity of a cell. To compute the global transition matrix, we first renormalize the topic weights $\tilde{L}_{ck}$ over just the topics that cell $c$ is associated to:

$$\tilde{L}_{ck} = \frac{L_{ck}}{\sum_{k' \in \{k_c\}} L_{ck'}} \quad \text{if} \quad k \in \{k_c\},\ 0 \text{ otherwise}, \tag{7}$$

where $\{k_c\}$ is the set of topics associated to cell $c$.

We compute the probability of a transition from cell $c$ to $c'$ as:

$$T_{cc'} = \sum_{k=1}^{K} \tilde{L}_{ck} p'_{cc',k}, \tag{8}$$

where $p'_{cc',k} = p_{cc',k}$ if $k \in \{k_c\} \bigcap \{k'_c\}$ and $p'_{cc',k} = 0$ otherwise. In the rare case that a cell is not included in any topic-associated process, the transition probability of the cell is assigned such that it can transition to any of its neighbors with uniform probability.

**RNA Velocity Evaluation Metrics.** We apply and compare several quantitative measures (*SI Appendix*, section 1) that go beyond pseudotime-based evaluations (7, 17) to assess the quality of RNA velocity estimates and accuracy of downstream inferred trajectories. Briefly, these measures include: 1) velocity coherence, also used by *scVelo* (7), to quantify the consistency (but not correctness) of velocity estimates; 2) the stationary distribution of the transition matrix, also used by *CellRank* (17), to identify terminal states; 3) mean first-passage time (MFPT), used before in velocity-independent TI (73) and related to a least action path (LAP) approach (22), to capture the expected timescales of transitions between subpopulations of cells; and 4) relative flux (*SI Appendix*, section 1), which we defined as a visualization-embedding–independent version of cross-boundary correctness (16), to capture the relative transition probability in each direction between two subpopulations.

**Permutation Tests.** To calculate the statistical significance of comparisons of the MFPT distributions between two groups of cells, we use a permutation test with 99,999 permutations. Specifically, for two subpopulations $A$ and $B$ of cells, each permutation consists of randomly permuting the MFPT values for $A \cup B$ and then splitting the values again into two groups of sizes $|A|$ and $|B|$ to compute the means. We perform a similar permutation test to calculate statistical significance for the difference in means in the number of intermediate states in a neighborhood (*SI Appendix*, Fig. S15).

**Preprocessing of scRNA-seq Data.** For each dataset, a gene was removed if there were not at least 20 cells with both spliced and unspliced transcripts for it. Following previous studies, we account for cell sizes by using size-normalized counts, which is also consistent with assumption in the classical derivation of the chemical master equation that reactions occur in a container of constant volume (74). Because the master equations inference requires integer counts, we round the size-normalized counts. In general, the normalization does not cause severe distortions to the abundances or proportions of the unspliced and spliced counts. However, distortions became problematic for the very small ratios of unspliced to spliced counts in the gastrulation data (*SI Appendix*, Fig. S21 *A* and *B*); hence, there we used raw counts. We verified that *TopicVelo* is robust with respect to the choice to size normalize by also using the raw counts to infer kinetic parameters for topic-associated processes, with other procedures remaining the same, on the scNT-seq dataset. The results from using normalized or raw counts were qualitatively consistent for the streamlines, stationary distributions, and mean first-passage times (*SI Appendix*, Fig. S21 *C*–*E*).

A principal components analysis was performed on the log-normalized spliced counts matrix using the top 2,000 highly variable genes. From the top 30 principal components, a $k_G$-nearest-neighbor ($k_G$-NN) graph was constructed (using the default of $k_G = 30$). (The notation $k_G$ is used to distinguish the parameter for the number of nearest neighbors from the completely independent parameter $k$ in the topic model.) Then the first and second moments of each cell were estimated over the $k_G$-NN graph. The above procedures were performed via *scVelo* (7):

```
scVelo.pp.filter_and_normalize(adata,
    min_shared_counts=20)
scVelo.pp.moments(adata, n_pcs=30, n_neighbors=30)
```

***TopicVelo* Analysis of scRNA-seq Data.** For all datasets, we identified the number of topics, selected topic-associated genes and cells, constructed integrated transitions, and computed RNA velocity evaluation metrics, as described (*SI Appendix*, section 1).

**Data, Materials, and Software Availability.** The source code, Jupyter notebooks, and R markdown files for reproducing figures and results in this paper are available at https://doi.org/10.5281/zenodo.10826412 (75). *TopicVelo* is available as an open-source Python package for public use at https://github.com/RiesenfeldGroup/TopicVelo (76). The gastrulation (10), bone marrow (28), dentate gyrus (29), and pancreas (30) data are available in the *scVelo* package (7). The human hematopoiesis scNT-seq (22) and ILCs data (31) are available in the NCBI Gene Expression Omnibus (GEO) under accession numbers GSE193517 and GSE149622, respectively.

**ACKNOWLEDGMENTS.** We thank Hope Anderson for discussions on interpreting the topic modeling results for the immune datasets, Hanna Hieromnimon for help in the design and creation of the overview figure, and Luis Barreiro for insightful comments on a manuscript draft. S.V. and C.F.G. were supported by the NIH NIGMS Award R35GM147400.

Author affiliations: [a]Department of Chemistry, University of Chicago, Chicago, IL 60637; [b]Institute for Biophysical Dynamics, University of Chicago, Chicago, IL 60637; [c]Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL 60637; [d]Department of Medicine, University of Chicago, Chicago, IL 60637; and [e]Committee on Immunology, Biological Sciences Division, University of Chicago, Chicago, IL 60637

1. P. V. Kharchenko, The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* **18**, 723–732 (2021).
2. D. Lähnemann *et al.*, Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
3. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
4. F. Ginhoux, A. Yalin, C. A. Dutertre, I. Amit, Single-cell immunology: Past, present, and future. *Immunity* **55**, 393–404 (2022).
5. J. Fan, K. Slowikowski, F. Zhang, Single-cell transcriptomics in cancer: Computational challenges and opportunities. *Exp. Mol. Med.* **52**, 1452–1465 (2020).
6. G. La Manno *et al.*, RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
7. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1546–1696 (2020).
8. M. Barile *et al.*, Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biol.* **22**, 197 (2021).
9. V. Bergen, R. A. Soldatov, P. V. Kharchenko, F. J. Theis, RNA velocity-current challenges and future perspectives. *Mol. Syst. Biol.* **17**, e10282 (2021).
10. B. Pijuan-Sala *et al.*, A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
11. G. Gorin, M. Fang, T. Chari, L. Pachter, RNA velocity unraveled. *PLoS Comput. Biol.* **18**, 1–55 (2022).
12. S. C. Zheng, G. Stein-O'Brien, L. Boukas, L. A. Goff, K. D. Hansen, Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biol.* **24**, 246 (2023).
13. G. Gorin, V. Svensson, L. Pachter, Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 39 (2020).
14. C. Li, M. Virgilio, K. L. Collins , J. D. Welch , Single-cell multi-omic velocity infers dynamic and decoupled gene regulation. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.12.13.472472 (Accessed 5 October 2022).
15. G. Gorin, L. Pachter, Length biases in single-cell RNA sequencing of pre-mRNA. *Biophys. Rep.* **3**, 100097 (2022).
16. C. Qiao, Y. Huang, Representation learning of RNA velocity reveals robust cell transitions. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2105859118 (2021).
17. M. Lange *et al.*, Cell Rank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
18. A. Gayoso *et al.*, Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.08.12.503709 (Accessed 7 November 2022).
19. M. Gao, C. Qiao, Y. Huang, UniTVelo: Temporally unified RNA velocity reinforces single-cell trajectory inference. *Nat. Commun.* **13**, 6586 (2022).
20. S. Farrell, M. Mani, S. Goyal, Inferring single-cell dynamics with structured dynamical representations of RNA velocity. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.08.22.504858 (Accessed 7 November 2022).
21. H. Cui, H. Maan, M. D. Taylor, B. Wang, DeepVelo: Deep learning extends RNA velocity to multi-lineage systems with cell-specific kinetics. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.04.03.486877 (Accessed 7 November 2022).
22. X. Qiu *et al.*, Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690–711.e45 (2022).
23. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
24. D. M. Blei, Probabilistic topic models. *Science* **55**, 77–84 (2012).
25. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
26. E. A. Erosheva *et al.*, "Bayesian Statistics 7" in *Bayesian Estimation of the Grade of Membership Model*, J. M. Bernardo, Ed. (Oxford University Press, Oxford, UK, 2003), pp. 501–510.
27. A. Singh, P. Bokes, Consequences of mRNA transport on stochastic variability in protein levels. *Biophys. J.* **103**, 1087–1096 (2012).
28. M. Setty *et al.*, Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
29. H. Hochgerner, A. Zeisel, P. Lönnerberg, S. Linnarsson, Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.* **21**, 290–299 (2018).
30. A. Bastidas-Ponce *et al.*, Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* **146**, dev173849 (2019).
31. P. Bielecki *et al.*, Skin-resident innate lymphoid cells converge on a pathogenic effector state. *Nature* **592**, 128–132 (2021).
32. D. Levens, D. R. Larson, A new twist on transcriptional bursting. *Cell* **158**, 241–242 (2014).
33. K. K. Dey, C. J. Hsiao, M. Stephens, Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet.* **13**, e1006599 (2017).
34. Y. Zhao, H. Cai, Z. Zhang, J. Tang, Y. Li, Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat. Commun.* **12**, 5261 (2021).
35. P. Carbonetto, A. Sarkar, Z. Wang, M. Stephens, Non-negative matrix factorization algorithms greatly improve topic model fits. arXiv [Preprint] (2021). https://doi.org/10.48550/arXiv.2105.13440 (Accessed 19 May 2022).
36. P. Carbonetto *et al.*, Interpreting structure in sequence count data with differential expression analysis allowing for grades of membership. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.03.03.531029 (Accessed 5 December 2022).
37. J. Cao, T. Xia, J. Li, Y. Zhang, S. Tang, A density-based method for adaptive LDA model selection. *Neurocomputing* **72**, 1775–1781 (2009).
38. M. Röder, A. Both, A. Hinneburg, "Exploring the space of topic coherence measures" in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15* (Association for Computing Machinery, New York, NY, 2015), pp. 399–408.
39. R. Deveaud, E. SanJuan, P. Bellot, Accurate and effective latent concept modeling for ad hoc information retrieval. *Docu. Numér.* **17**, 61–84 (2014).
40. D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
41. P. Virtanen *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
42. N. Songdej *et al.*, Transcription factor RUNX1 regulates factor FXIIIA subunit (F13A1) expression in megakaryocytic cells and platelet F13A1 expression is downregulated in RUNX1 haplodeficiency. *Blood* **136**, 25–26 (2020).
43. F. Erhard *et al.*, Time-resolved single-cell RNA-seq using metabolic RNA labelling. *Nat. Rev. Methods Primers* **2**, 77 (2022).
44. B. Psaila *et al.*, Single-cell analyses reveal megakaryocyte-biased hematopoiesis in myelofibrosis and identify mutant clone-specific targets. *Mol. Cell* **78**, 477–492.e8 (2020).
45. M. H. Shim, A. Hoover, N. Blake, J. G. Drachman, J. A. Reems, Gene expression profile of primary human CD34+CD38lo cells differentiating along the megakaryocyte lineage *Exp. Hematol.* **32**, 638–648 (2004).
46. Y. Li, X. Qi, B. Liu, H. Huang, The STAT5-GATA2 pathway is critical in basophil and mast cell differentiation and maintenance. *J. Immunol.* **194**, 4328–4338 (2015).
47. M. Karlsson *et al.*, A single-cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, eabh2169 (2021).
48. A. Cvejic *et al.*, SMIM1 underlies the vel blood group and influences red blood cell traits. *Nat. Genet.* **45**, 542–545 (2013).
49. M. Suzuki *et al.*, GATA factor switching from GATA2 to GATA1 contributes to erythroid differentiation. *Genes Cells* **18**, 921–933 (2013).
50. B. K. Tusi *et al.*, Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
51. H. Yan *et al.*, Developmental differences between neonatal and adult human erythropoiesis. *Am. J. Hematol.* **93**, 494–503 (2018).
52. D. Sichien *et al.*, IRF8 transcription factor controls survival and function of terminally differentiated conventional and plasmacytoid dendritic cells, respectively. *Immunity* **45**, 626–640 (2016).
53. H. Wang *et al.*, Decoding human megakaryocyte development. *Cell Stem Cell* **28**, 535–549.e8 (2021).
54. D. Pellin *et al.*, A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, 2395 (2019).
55. O. Chertov *et al.*, Identification of human neutrophil-derived cathepsin G and azurocidin/CAP37 as chemoattractants for mononuclear cells and neutrophils. *J. Exp. Med.* **186**, 739–747 (1997).
56. M. Colonna, Innate lymphoid cells: Diversity, plasticity, and unique functions in immunity. *Immunity* **48**, 1104–1117 (2018).
57. J. J. O'Shea, W. E. Paul, Mechanisms underlying lineage commitment and plasticity of helper CD4+ T cells *Science* **327**, 1098–1102 (2010).
58. E. Vivier *et al.*, Innate lymphoid cells: 10 years on. *Cell* **174**, 1054–1066 (2018).
59. Z. Cao, X. Sun, B. Icli, A. K. Wara, M. W. Feinberg, Role of Kruppel-like factors in leukocyte development, function, and disease. *Blood* **116**, 4404–4414 (2010).
60. Y. Gu, D. Blaauw, J. D. Welch, Bayesian inference of RNA velocity from multi-lineage single-cell data. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.07.08.499381 (Accessed 7 November 2022).
61. Q. Qin, E. Bingham, G. L. Manno, D. M. Langenau , L. Pinello , Pyro-Velocity: Probabilistic RNA velocity inference from single-cell data. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.09.12.507691 (Accessed 7 November 2022).
62. T. Griffiths, M. Jordan, J. Tenenbaum, D. Blei, "Hierarchical topic models and the nested Chinese restaurant process" in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, B. Schölkopf, Eds. (MIT Press, 2003), vol. 16.
63. I. Vayansky, S. A. Kumar, A review of topic modeling methods. *Inf. Syst.* **94**, 101582 (2020).
64. G. Gorin, J. J. Vastola, M. Fang, L. Pachter, Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.09.06.459173 (Accessed 24 July 2022).
65. J. A. Farrell *et al.*, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science (New York, N.Y.)* **360**, eaar3131 (2018).
66. D. M. Blei, J. D. Lafferty, "Dynamic topic models" in *Proceedings of the 23rd International Conference on Machine Learning, ICML '06* (Association for Computing Machinery, New York, NY, 2006), pp. 113–120.
67. C. Soneson, A. Srivastava, R. Patro, M. B. Stadler, Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLoS Comput. Biol.* **17**, 1–26 (2021).

68. Z. Chen, W. C. King, A. Hwang, M. Gerstein, J. Zhang, DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. *Sci. Adv.* **8**, eabq3745 (2022).
69. R. Zakine, E. Vanden-Eijnden, Minimum-action method for nonequilibrium phase transitions. *Phys. Rev. X* **13**, 041044 (2023).
70. M. Lee, bab2min/tomotopy: 0.12.3 (version v0.12.3, Zenodo, 2022).
71. T. Li, J. Shi, Y. Wu, P. Zhou, On the mathematics of RNA velocity. I: Theoretical Analysis. *CSIAM Trans. Appl. Math.* **2**, 1–55 (2021).
72. S. K. Lam, A. Pitrou, S. Seibert, "Numba: A LLVM-based python JIT compiler" in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (2015), pp. 1–6.
73. C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, A. M. Klein, Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E2467–E2476 (2018).
74. D. Gillispie, A rigorous derivation of the chemical master equation. *Physica A* **188**, 404–425 (1992).
75. F. C. Gao, S. Vaikuntanathan, S. J. Riesenfeld, Code for Reproducing Results in "Dissection and Integration of Bursty Transcriptional Dynamics for Complex Systems." Zenodo. https://doi.org/10.5281/zenodo.10826412. Deposited 10 April 2024.
76. F. C. Gao, S. Vaikuntanathan, S. J. Riesenfeld, TopicVelo. Github. https://github.com/RiesenfeldGroup/TopicVelo. Deposited 30 September 2023.