ARTICLES
https://doi.org/10.1038/s41587-020-0591-3
nature biotechnology

Generalizing RNA velocity to transient cell states through dynamical modeling

Volker Bergen[1,2], Marius Lange[1,2], Stefan Peidli[2], F. Alexander Wolf[1]✉ and Fabian J. Theis[1,2]✉

RNA velocity has opened up new ways of studying cellular differentiation in single-cell RNA-sequencing data. It describes the rate of gene expression change for an individual gene at a given time point based on the ratio of its spliced and unspliced messenger RNA (mRNA). However, errors in velocity estimates arise if the central assumptions of a common splicing rate and the observation of the full splicing dynamics with steady-state mRNA levels are violated. Here we present scVelo, a method that overcomes these limitations by solving the full transcriptional dynamics of splicing kinetics using a likelihood-based dynamical model. This generalizes RNA velocity to systems with transient cell states, which are common in development and in response to perturbations. We apply scVelo to disentangling subpopulation kinetics in neurogenesis and pancreatic endocrinogenesis. We infer gene-specific rates of transcription, splicing and degradation, recover each cell's position in the underlying differentiation processes and detect putative driver genes. scVelo will facilitate the study of lineage decisions and gene regulation.

Single-cell transcriptomics has enabled the unbiased study of biological processes such as cellular differentiation and lineage choice at single-cell resolution[1,2]. The resulting computational problem is known as trajectory inference. Starting from a population of cells at different stages of a developmental process, trajectory inference algorithms aim to reconstruct the developmental sequence of transcriptional changes leading to potential cell fates. A multitude of such methods have been developed, commonly modeling the dynamics as the progression of cells along an idealized, potentially branching trajectory[3–8]. A central challenge in trajectory inference is the destructive nature of single-cell RNA sequencing (scRNA-seq), which reveals only static snapshots of cellular states. To move from descriptive toward predictive trajectory models, additional information is required to constrain the space of possible dynamics that could give rise to the same trajectory[9,10]. As such, lineage-tracing assays can add information via genetic modification to enable the reconstruction of lineage relationships[11–17]. However, these assays are not straightforward to set up and are technically limited in many systems, such as human tissues.

The concept of RNA velocity has enabled the recovery of directed dynamic information by leveraging the fact that newly transcribed, unspliced pre-mRNAs and mature, spliced mRNAs can be distinguished in common scRNA-seq protocols, the former detectable by the presence of introns[18]. Assuming a simple per-gene reaction model that relates abundance of unspliced and spliced mRNA, the change in mRNA abundance, termed RNA velocity, can be inferred. Positive RNA velocity indicates that a gene is upregulated, which occurs for cells that show higher abundance of unspliced mRNA for that gene than expected in steady state. Conversely, negative velocity indicates that a gene is downregulated. The combination of velocities across genes can then be used to estimate the future state of an individual cell. The original model[18] estimates velocities under the assumption that the transcriptional phases of induction and repression of gene expression last sufficiently long to reach both an actively transcribing and an inactive silenced steady-state equilibrium. After inferring the ratio of unspliced to spliced mRNA abundance that is in a constant transcriptional steady state, velocities are determined as the deviation of the observed ratio from its steady-state ratio. Inferring the steady-state ratio makes two fundamental assumptions, namely that (1), on the gene level, the full splicing dynamics with transcriptional induction, repression and steady-state mRNA levels are captured; and (2), on the cellular level, all genes share a common splicing rate. These assumptions are often violated, in particular when a population comprises multiple heterogeneous subpopulations with different kinetics. We refer to this modeling approach as the 'steady-state model'.

To resolve the above restrictions, we developed scVelo, a likelihood-based dynamical model that solves the full gene-wise transcriptional dynamics. It thereby generalizes RNA velocity estimation to transient systems and systems with heterogeneous subpopulation kinetics. We infer the gene-specific reaction rates of transcription, splicing and degradation and an underlying gene-shared latent time in an efficient expectation-maximization (EM) framework. The inferred latent time represents the cell's internal clock, which accurately describes the cell's position in the underlying biological process. In contrast to existing similarity-based pseudo-time methods, this latent time is grounded only on transcriptional dynamics and accounts for speed and direction of motion.

We demonstrate the capabilities of the dynamical model on various cell lineages in hippocampal dentate gyrus neurogenesis[19] and pancreatic endocrinogenesis[20]. The dynamical model generally yields more consistent velocity estimates across neighboring cells and accurately identifies transcriptional states as opposed to the steady-state model. It provides fine-grained insights into the cell states of cycling pancreatic endocrine precursor cells, including their lineage commitment, cell cycle exit and, finally, endocrine cell differentiation. Here our inferred latent time is able to reconstruct the temporal sequence of transcriptomic events and cellular fates. Moreover, scVelo identifies regimes of regulatory changes such as transition states and stages of cell fate commitment. Here scVelo identifies putative driver genes of these transcriptional changes. Driver genes display pronounced dynamic behavior and are systematically detected via their characterization by high likelihoods

[1]Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany. [2]Department of Mathematics, Technical University of Munich, Munich, Germany. ✉e-mail: alex.wolf@helmholtz-muenchen.de; fabian.theis@helmholtz-muenchen.de

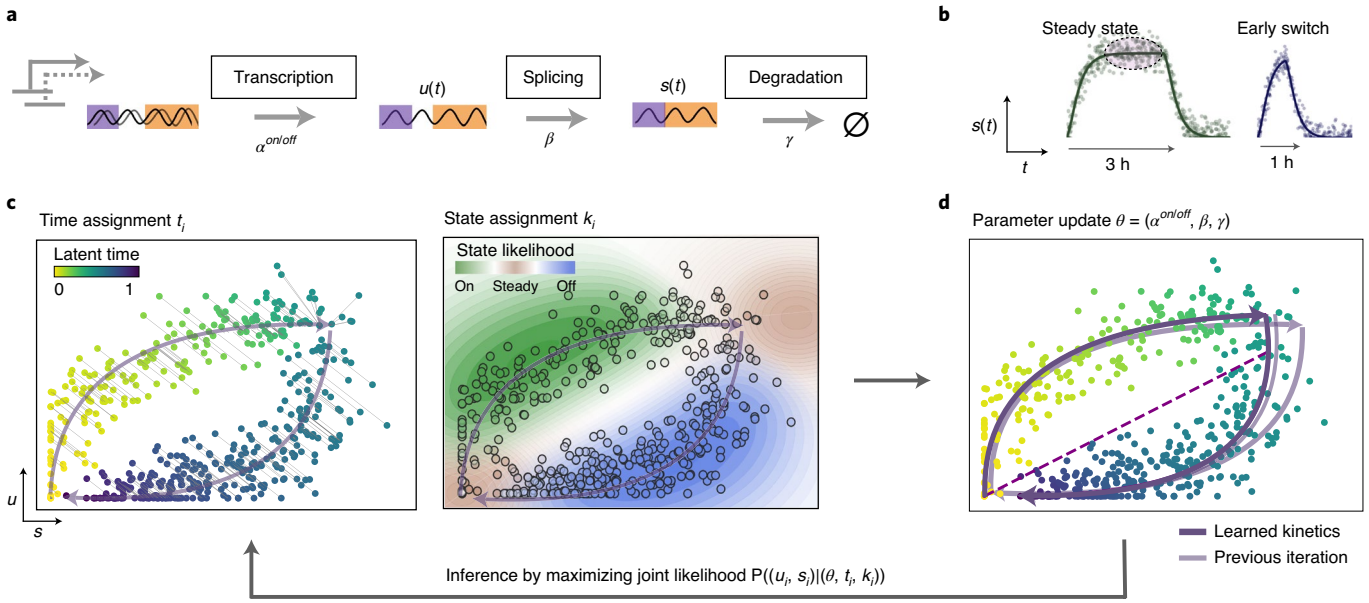

Fig. 1 | Solving the full splicing kinetics generalizes RNA velocity to transient populations. a, Modeling transcriptional dynamics captures transcriptional induction and repression ('on' and 'off' phase) of unspliced pre-mRNAs, their conversion into mature, spliced mRNAs and their eventual degradation. **b**, An actively transcribed and an inactive silenced steady state is reached when the transcriptional phases of induction and repression last sufficiently long, respectively. In particular in transient cell populations, however, steady states are often not reached as, for example, induction might terminate before mRNA-level saturation, displaying an 'early switching' behavior. **c**, We propose scVelo, a likelihood-based model that solves the full gene-wise transcriptional dynamics of splicing kinetics, which is governed by two sets of parameters: (1) reaction rates of transcription, splicing and degradation, and (2) cell-specific latent variables of transcriptional state and time. The parameters are inferred iteratively via EM. For a given estimate of reaction rate parameters, time points are assigned to each cell by minimizing its distance to the current phase trajectory. The transcriptional states are assigned by associating a likelihood to respective segments on the trajectory—that is, induction, repression and active and inactive steady state. **d**, The overall likelihood is then optimized by updating the model parameters of reaction rates. The dashed purple line links the inferred (unobserved) inactive with the active steady state.

in the dynamic model. This procedure presents a dynamics-based alternative to the standard differential expression paradigm.

Finally, we propose to further account for stochasticity in gene expression, obtained by treating transcription, splicing and degradation as probabilistic events. We show how this can be achieved for the steady-state model and demonstrate its capability of capturing the directionality inferred from the full dynamical model to a large extent. We illustrate its considerable improvement over the steady-state model while being as efficient in computation time. The dynamical, the stochastic as well as the steady-state model are available within scVelo as a robust and scalable implementation (https://scvelo.org). For the latter two, scVelo achieves a tenfold speedup over the original implementation (velocyto)[18].

Results

Solving the full gene-wise transcription dynamics at single-cell resolution. As in the original framework[18], we model transcriptional dynamics (Fig. 1a) using the basic reaction kinetics described by

$$\begin{aligned}\frac{du(t)}{dt} &= \alpha^{(k)}(t) - \beta u(t),\\ \frac{ds(t)}{dt} &= \beta u(t) - \gamma s(t),\end{aligned}$$

for each gene, independent of all other genes. As opposed to the original framework, to account for non-observed steady states (Fig. 1b), we solve these equations explicitly and infer the splicing kinetics that is governed by two sets of parameters: (1) the reaction rates of transcription $\alpha_k(t)$, splicing β and degradation γ; and (2) cell-specific latent variables— that is, a discrete transcriptional state k_i and a continuous time t_i, where i represents a single observed cell. The parameters of the reaction rates can be obtained if the latent variables are given and vice versa. Hence, we infer the parameters by EM, iteratively estimating the reaction rates and latent variables via maximum likelihood. In the expectation step, for a given model estimate of the unspliced/spliced phase trajectory, $\mathcal{X} = (\hat{u}(t), \hat{s}(t))_t$, a latent time t_i is assigned to an observed mRNA value $x_i = (u_i, s_i)$ by minimizing its distance to the phase trajectory χ (Fig. 1c). The transcriptional states k_i are then assigned by associating a likelihood to respective segments on the phase trajectory χ—that is, $k_i \in \{\text{on}, \text{off}, \text{ss}_{\text{on}}, \text{ss}_{\text{off}}\}$ labeling induction, repression and active and inactive steady states. In the maximization step, the overall likelihood is then optimized by updating the parameters of reaction rates (Fig. 1d and Methods). Convergence to an optimal parameter set is achieved for genes that display evident kinetics (Supplementary Fig. 5). Note that, for efficiency reasons, we use an approximation to the optimal time assignment, which essentially yields the same results at a 30-fold speedup (Supplementary Fig. 7 and Methods).

The resulting gene-specific trajectory χ, parametrized by interpretable parameters of reaction rates and transcriptional states, explicitly describes how mRNA levels evolve over latent time. Whereas the steady-state model uses linear regression to fit assumed steady states and fails if these are not observed, the dynamical model resolves the full dynamics of unspliced and spliced mRNA abundances and thus enables unobserved steady states to also be faithfully captured (Supplementary Fig. 1). RNA velocity is then explicitly given by the derivative of spliced mRNA abundance, parametrized by the inferred variables.

To make the inferred parameters of reaction rates relatable across genes, the gene-wise latent times are coupled to a universal, gene-shared latent time that proxies a cell's internal clock (Supplementary Fig. 2 and Methods). This universal time allows us

to resolve the cell's relative position in a biological process with support from the splicing dynamics of all genes. Also, transcriptional states can be identified more confidently by sharing information between genes. On simulated splicing kinetics, latent time is able to reconstruct the underlying real time at near-perfect correlation and correct scale, clearly outperforming diffusion pseudo-time. In contrast to pseudo-time methods[3,21], our latent time is grounded on transcriptional dynamics and internally accounts for speed and direction of motion. Hence, scVelo's latent time yields faithful gene expression time courses to delineate dynamical processes and to extract gene cascades.

Further, the coupling to a universal latent time allows us to identify the kinetic rates up to a global gene-shared scale parameter. Employing the overall time scale of the developmental process as prior information, the absolute values of kinetic rates can eventually be identified (Supplementary Fig. 3).

Identifying reaction rates in transient cell populations. To validate the sensitivity of both models with respect to varying parameters in simulated splicing kinetics, we randomly sampled 2,000 log-normally distributed parameters for each reaction rate and time events following the Poisson law. The total time spent in a transcriptional state is varied between 2 and 10 h.

The ratio inferred by the steady-state model yields a systematic error as the time of transcriptional induction decreases such that mRNA levels are less likely to reach steady-state equilibrium levels (Supplementary Fig. 3a). By contrast, the dynamical model yields a consistently smaller error and is completely insensitive with respect to variability in induction duration. Furthermore, the Pearson correlation between the true and inferred steady-state ratio increases from 0.71 to 0.97 when using the dynamical model. Imposing the overall time scale of the splicing dynamics of 20 h as prior information, the dynamical model reliably recovers the true parameters of the simulated splicing kinetics, achieving correlations of 0.85 and higher (Supplementary Fig. 3b).

Resolving the heterogeneous population kinetics in dentate gyrus development. To test whether scVelo's velocity estimates allow identification of more complex population kinetics, we considered a scRNA-seq experiment from the developing mouse dentate gyrus[19] comprising two time points (P12 and P35) measured using droplet-based scRNA-seq (10× Genomics Chromium Single Cell Kit V1; Methods). The original publication aimed to elucidate the relationship between developmental and adult dentate gyrus neurogenesis. Although they linked transient intermediate states to neuroblast stages and mature granule cells, the commitment of radial glia-like cells could not be conclusively determined.

After basic pre-processing, we apply both the steady-state and the dynamical model and display the vector fields using streamline plots[22] in a uniform manifold approximation and projection (UMAP)-based embedding[23] of the data (Fig. 2a). The dominating structure is the granule cell lineage, in which neuroblasts develop into granule cells. Simultaneously, the remaining population forms distinct cell types that are fully differentiated (for example, Cajal–Retzius (CR) cells) or cell types that form a sublineage (for example, GABA cells). Whether a cell type is still in transition or already terminal is indicated by two experimental time points (Supplementary Fig. 6a) and experimental analysis[19], both supporting the overall velocity-inferred directionality. Notably, the velocities derived from both models settle previously ambiguous evidence of the fate choice of radial glia-like cells in favor of astrocytes over neurogenic intermediate progenitor cells.

Whereas the main lineage toward mature granule cells is captured by both models, the single-cell velocities illustrate pronounced differences in sublineages and subclusters. As such, only scVelo correctly identifies the oligodendrocyte precursor cells (OPCs) differentiating into myelinating oligodendrocytes (OLs) and CR cells as terminal. The steady-state model erroneously assigns high velocities to CR cells, which can be traced back to gene-resolved velocities. With *Fam155a*, the incongruous CR velocities from the steady-state model become evident. The splicing dynamics, particularly well illustrated by *Fam155a*, clearly suggests that the CR population is terminal. Also, expression patterns show no evidence for any further maturation within the CR population. Yet, as the steady-state model determines velocities as deviations from steady state that are computed for the whole population, the model is biased to assign high velocities to outlier cells, such as the CR population (Fig. 2b). The dynamical model assigns CR cells to steady state with high likelihood, as it cannot be confidently linked to any transient state.

Tmsb10 is the major contributor to the inferred dynamics and illustrates another fundamental difference. Velocities derived from the dynamical model are more consistent across velocities of neighboring cells than those derived from the steady-state model, which results in a higher overall coherence of the velocity vector field (Fig. 2a, top right, and Supplementary Fig. 9).

Both the steady-state and the dynamical model yield additional dynamic flow within the mature compartment of granule cells, which was expected to be terminal and might be worthwhile to follow up experimentally. It is further noteworthy that, even though mossy cells are positioned next to neuroblast cells, velocity-inferred cell-to-cell transition probabilities do not show any likely transitions between the two populations, thus suggesting that mossy cells form their own lineage (Supplementary Fig. 6b).

Determining dynamical genes beyond differential expression testing. scVelo computes a likelihood for each gene and cell for a model-optimal latent time and transcriptional state, explaining how well a cell is described by the learned spliced/unspliced phase trajectory. Aggregating over cells to obtain an overall gene likelihood, we rank genes according to their goodness of fit. This enables us to identify genes that display pronounced dynamic behavior, which makes them candidates for important drivers of the main process in the population (Fig. 2c and Supplementary Fig. 4). The top likelihood-ranked genes show clear indication of splicing dynamics, whereas the expression of low-ranked genes is governed by noise or nonexisting transient states. Moreover, partial gene likelihoods—that is, likelihoods computed for a subset of cells—enable us to identify potential drivers for particular transition phases, branching regions, specific cell types or cycling subpopulations. Many of the top-ranked genes have been reported to play a crucial role in neurogenesis (for example, *Grin2b*, *Map1b* and *Dlg2*)[24,25], whereas some of these genes were connected to the CA1 region in the hippocampal circuit (for example, *Tmsb10* and *Hn1*)[26]. *Ppp3ca*, the gene with the highest likelihood, which mostly contributes to the velocity vector field, is elevated toward granule cells. Its vital role has been demonstrated by associating a reduction of *Ppp3ca* activity with tauopathy in Alzheimer's disease[27]. By showing that excluding the top likelihood-ranked genes results in non-reconstructability of the dynamics, we show computationally that the inferred directionality is mainly governed by these driver genes (Supplementary Fig. 8).

Delineating cycling progenitors, commitment and fate transitions in endocrinogenesis. Next, we demonstrate scVelo's capabilities to delineate transient lineages in endocrine development in the mouse pancreas, with transcriptome profiles sampled from E15.5 (ref. [20]). Endocrine cells are derived from endocrine progenitors located in the pancreatic epithelium, marked by transient expression of the transcription factor *Ngn3*. Endocrine commitment terminates in four major fates: glucagon-producing α-cells, insulin-producing β-cells, somatostatin-producing δ-cells and ghrelin-producing $\in$-cells[28]. Although in previous work RNA velocity illuminated the directional flow in the endocrine lineage,

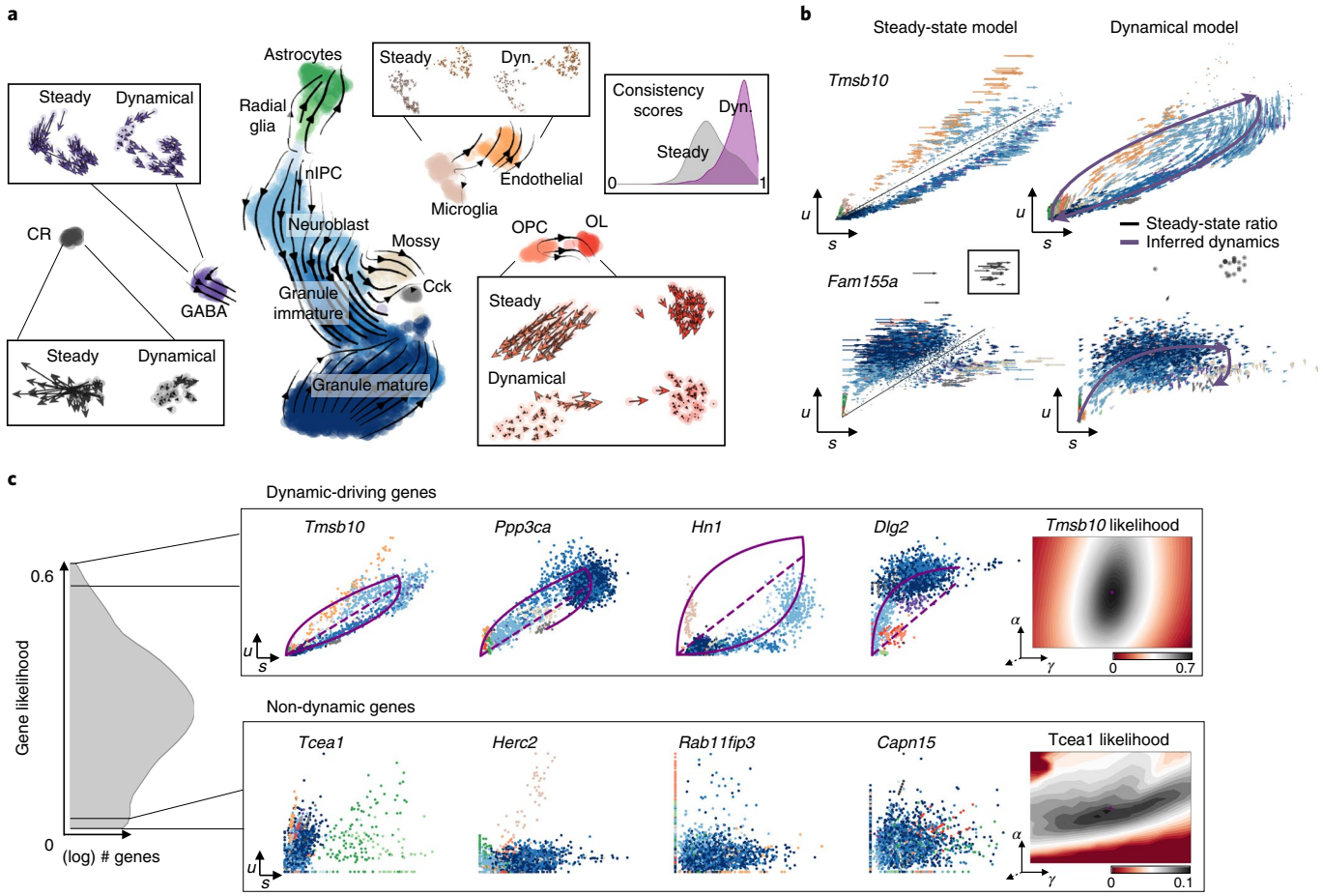

Fig. 2 | Resolving subpopulation kinetics and identifying dynamical genes in neurogenesis. a, Velocities derived from the dynamical model for dentate gyrus neurogenesis[19] are projected into a UMAP-based embedding. The main gene-averaged flow visualized by velocity streamlines corresponds to the granule lineage, in which neuroblasts develop into granule cells. The remaining populations form distinct cell types that are either differentiated, for example CR cells, or cell types that form sublineages, for example the GABA and oligodendrocyte lineages (OPC to OL). When zooming into the cell types to examine single-cell velocities, fundamental differences between the velocities derived from the steady-state and dynamical model become apparent. Only the dynamical model identifies CR cells to be terminal by assigning no velocity and indicates that OPCs indeed differentiate into OLs. By contrast, the steady-state model displays a high velocity in CR cells and points OPCs away from OLs. Overall, the dynamical model yields a more coherent velocity vector field as illustrated by the consistency scores (in the top-right corner, defined for each cell as the correlation of its velocity with the velocities of neighboring cells). **b**, Gene-resolved velocities allow further interpreting the inferred directionality on the cellular level. For instance, *Tmsb10* is the major contributor to the gene-averaged flow that describes neuroblasts as differentiating into granule cells. With *Fam155a*, the incongruous CR velocities from the steady-state model become evident. By reducing velocity estimation to steady-state deviations, this model is biased to assign high velocities to outlier cells, such as the CR population. In contrast, the dynamical model assigns CR cells to a steady state with high likelihoods, as they are not well explained by the overall kinetics and cannot be confidently linked to the transient induction state. **c**, The dynamical model allows to systematically identify putative driver genes as genes characterized by high likelihoods. Whereas genes selected by high likelihoods (upper row) display pronounced dynamic behavior, expression of low-likelihood genes (lower row) is governed by noise or nonexisting transient states. nIPC, neurogenic intermediate progenitor cell.

the endocrine fates could not be clearly delineated and incongruous subpopulation flows emerged[20].

We demonstrate the additional fine-grained insights into the developmental processes that we gain from the dynamical model when compared to the steady-state model. First, scVelo accurately delineates the cycling population of ductal cells and endocrine progenitors (Fig. 3a), biologically affirmed by cell cycle scores (standardized scores of mean expression levels of phase marker genes[29]) and previous analysis[30]. Further, scVelo illuminates cell states of lineage commitment, cell cycle exit and endocrine cell differentiation. By contrast, the steady-state model does not capture the cell cycle and yields incongruous backflows in later endocrine stages (Fig. 3b). For instance, α-cells that erroneously seem to be de-differentiating can be traced back to false state identifications—for example, in *Cpe*-assigning α-cells in parts to both induction and repression phases (Fig. 3c). The inferred dynamics by scVelo are reported in several recent studies that have shed light on the time-resolved programs along the lineage stages[28,31,32]. For instance, lineage-tracing analyses revealed endocrine cells to be derived from $Ngn3^+$ precursors via intermediate stages of Fev^+ endocrine cells[28].

Relating cell fates and disentangling dynamical regimes through latent time. We infer a universal gene-shared latent time that represents the cell's internal clock. This latent time is a more faithful reconstruction of real time than similarity-based diffusion pseudo-time (Supplementary Fig. 2 and Methods). We compared pseudo-time and latent time in the chronology of endocrine cell fates. In real time, α-cells are produced earlier (before E12.5) than β-cells (E12.5–E15.5)[20]. This ordering is captured by latent time but not by pseudo-time (Fig. 3d). Furthermore, the inferred

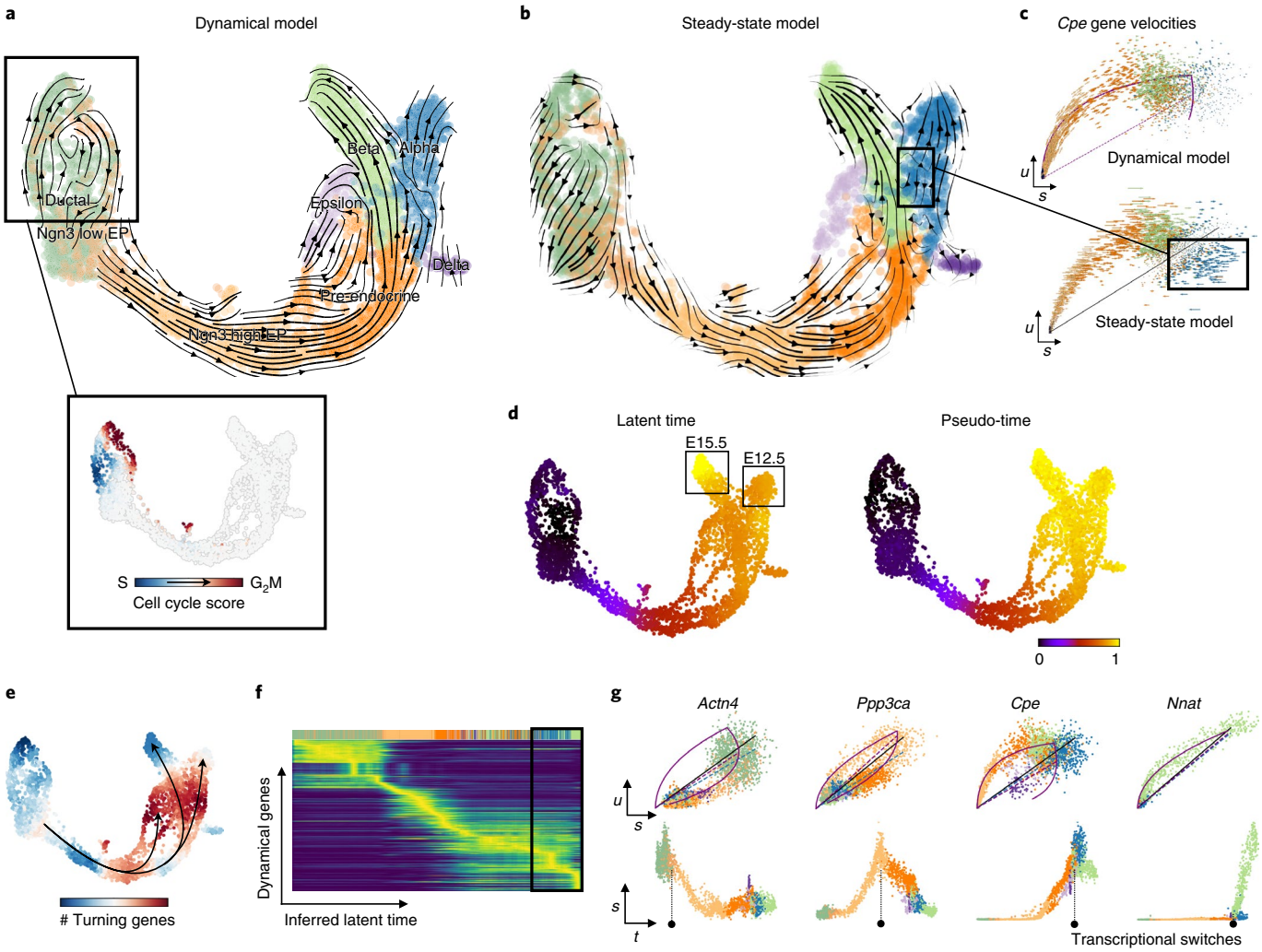

Fig. 3 | Delineating cycling progenitors and lineage commitment and disentangling cell fates and regimes of transcriptional activity through latent time in pancreatic endocrinogenesis. a, Velocities derived from the dynamical model for pancreatic endocrinogenesis[20] are visualized as streamlines in a UMAP-based embedding. The dynamical model accurately delineates the cycling population of endocrine progenitors, their lineage commitment, cell cycle exit and endocrine differentiation. Inferred S and G_2M phases based on cell cycle scores affirms the cell cycle identified by the dynamical model. **b**, The steady-state model does not capture the cycle and yields incongruous backflows directed against the lineage in later endocrine stages. **c**, Single-gene velocities illustrate the limitations of the steady-state model. Incongruous backflows in α-cells can be traced back to false state identifications—for example, in *Cpe* it assigns α-cells in parts to both induction and repression phases. **d**, scVelo's latent time is based only on its transcriptional dynamics and represents the cell's internal clock. It captures aspects of the actual time better than similarity-based diffusion pseudo-time, as observed in the chronology of endocrine cell fates: α-cells are produced earlier in actual time (before E12.5), whereas β-cells are produced later (E12.5–E15.5). Whereas latent time enables the temporal relation of the two fates, diffusion pseudo-time does not distinguish their temporal position. **e**, By using latent time to infer and count switching points between transcriptional states (for example, from induction to homeostasis), lineage commitment and branching points become apparent. **f**, Gene expression dynamics resolved along latent time shows a clear cascade of transcription in the top 300 likelihood-ranked genes. **g**, Putative driver genes are identified by high likelihoods. Phase portraits (top) and expression dynamics along latent time (bottom) for these driver genes characterize their activity. Whereas *Actn4* switches at cycle exit and endocrine commitment, the three other genes switch or start to express at the branching points.

velocities in α-cells are lower than the strong directional flow in β-cells, which, again, suggests that α-cells have already been produced at an earlier stage. Moreover, the inferred gene-specific switching time points indicate regions of transcriptional changes. The number of identified genes turning from one transcriptional state to another—for example, from induction to repression—give rise to regions of lineage commitment, transition states and branching points (Fig. 3e). Within these regions, putative driver genes can be identified by their likelihoods, among which the top-ranked genes have been associated with hormone processing (for example, *Cpe* and *Pcsk2*) and secretion (*Abcc8*)[33–35]. Their transcriptional activity is shown by gene expression dynamics resolved along latent time (Fig. 3f,g).

Extending the model to account for stochasticity in gene expression. The partial stochasticity of gene expression[36] has been addressed by a variety of modeling approaches in systems biology[37]. The flexibility of scVelo's likelihood-based approach allows us to extend the deterministic ordinary differential equation (ODE) model to account for stochasticity by treating transcription, splicing and degradation as probabilistic events. For simplicity, we demonstrate how this can be achieved for the steady-state model (Methods). The resulting Markov jump process is commonly approximated by moment equations[38], which can be solved in closed form in the linear ODE system under consideration. By including second-order moments, we exploit not only the balance of unspliced to spliced mRNA levels but also their covariation. The stochastic steady-state

model is capable of capturing the results of the full dynamical model to a greater extent than the deterministic steady-state model (Supplementary Fig. 9a), which indicates that stochasticity adds valuable information. For instance, the stochastic model resolves the sublineages in the dentate gyrus of granule, astrocyte and GABA maturation. In pancreatic endocrinogenesis, it is capable of resolving the cycling progenitors and endocrine lineage commitment to a great extent yet also yields backflows in the α-cells like the deterministic model. Overall, the stochastic model displays higher consistency than the deterministic model (Supplementary Fig. 9b), while remaining as efficient in computation time (Supplementary Fig. 13). Investigations of a stochastic dynamical model are left for future work.

Accounting for different kinetic regimes and insufficiently observed kinetics. One important concern is dealing with systems that represent multiple lineages and processes, where genes are likely to show different kinetic regimes across subpopulations. Distinct cell states and lineages are typically governed by different variations in the gene regulatory networks and might, hence, exhibit different splicing kinetics. This gives rise to genes that display multiple trajectories in phase space. To address this, we perform a likelihood-ratio test for differential kinetics to detect clusters that display kinetic behavior that cannot be well explained by a single model of the overall dynamics (Supplementary Fig. 11 and Methods). Clustering cell types into their different kinetic regimes then allows us to fit each regime separately.

Another difficulty concerns insufficiently observed splicing kinetics. For instance, we might detect only a small portion of the overall dynamics at the very ends of the process (Supplementary Fig. 12a). This manifests as a straight line in the unspliced to spliced phase portraits rather than a curve (Supplementary Fig. 12b). Observing partial dynamics in this way leads the steady-state and stochastic models to incorrectly fit this line and erroneously assign positive and negative velocities. The lack of observed curvature also challenges the dynamical model when determining whether an upregulation or downregulation should be fit. This ambiguity can be observed in two application scenarios where only a small fraction of the kinetics is disclosed: (1) a gene is active only in a small window of the observed process, or (2) the observed time frame in the data covers only a small part of the timeframe of the underlying dynamical process. The former case occurs when a gene is upregulated only at the very end or downregulated at the very beginning of a developmental process. The latter case can occur when a dynamical process occurs in a fast or synchronous fashion such that the snapshot captured in an scRNA-seq data set recovers little of the full dynamics. Here, the overall developmental time scale of the sampled population might be far shorter than the potential duration of the kinetics. We address this issue by extending the dynamical model with a 'root prior'. This prior can either be internally obtained from genes that are sufficiently informative to uncover the root of the process, or it can be obtained from prior knowledge, such as the first experimental time point or a known progenitor population (Supplementary Fig. 12c and Methods).

To that end, we advise the user not to limit biological conclusions to the projected velocities but to examine individual gene dynamics via phase portraits to understand how inferred directions are supported by particular genes. Thereby, finding the most relevant genes is greatly facilitated by the dynamical model. We also encourage the user to challenge the underlying assumptions and, in particular, test for differential kinetics, insufficiently observed kinetics and time scale mismatches.

Tenfold speedup for the steady-state model and large-scale applicability. The dynamical, the stochastic as well as the steady-state model are available within scVelo as a robust and scalable implementation (https://scvelo.org). Illustratively, on pancreas development with 25,919 transcriptome profiles, scVelo runs the full pipeline for the steady-state as well as stochastic model from pre-processing the data to velocity estimation to projecting the data in any embedding in less than 1 min (Supplementary Fig. 13). That is obtained by memory-efficient, scalable and parallelized pipelines via integration with scanpy[21], by leveraging efficient nearest-neighbor search[39], analytical closed-form solutions, sparse implementation and vectorization. The scVelo pipeline thereby achieves more than a tenfold speedup over the original implementation (velocyto). The full splicing dynamics, including kinetic rate parameters, latent time and velocities, are inferred in a longer but practicable runtime of 20 min for 1,000 genes across 35,000 profiles. As it scales in near-linear time with the number of cells and genes, its runtime is exceeded by velocyto's quadratic runtime on large cell numbers of 35,000 and higher. For large cell numbers, also memory efficiency becomes a critical aspect. On an Intel Core i7 CPU with 3.7 GHz and 64 GB of RAM, velocyto cannot process more than 40,000 cells, whereas scVelo scales to more than 300,000 cells. Notably, the stochastic steady-state model is solved in closed form and remains computationally efficient. It serves as a tradeoff between efficiency and accuracy and is advised to be used if runtime is of particular importance.

Discussion

scVelo enables velocity estimation without assuming either the presence of steady states or a common splicing rate across genes. It maintains the weaker assumptions of constant gene-specific splicing and degradation rates and two transcription rates each for induction and repression. These assumptions might be violated in practice and can be addressed by extending scVelo toward more complex regulations. On the gene level, full-length scRNA-seq protocols, such as Smart-seq2 (ref. [40]), allow accounting for gene structure, alternative splicing and state-dependent degradation rates. These can be incorporated into scVelo's likelihood-based inference by adapting the ODE model. In particular, spatial single-cell RNA profiling at transcriptome scale[41,42] might provide additional information on relative cell positions necessary to resolve spatial dependencies in gene regulation. Spatial coordinates as well as experimental time might also be leveraged as additional constraints to extend the concept of latent time—for instance, to capture the progression around a cell cycle. Stochastic variability may be leveraged beyond steady state, which has been dubbed as 'listening to the noise' and shown to improve parameter identifiability[43]. Extending the kinetic model to protein translation has been proposed within the steady-state formulation[44] and can be likewise included into the dynamical model. Metabolic labeling, for example using single-cell SLAM-seq[45,46], enables the quantification of total RNA levels together with newly transcribed RNA. This additional readout can be easily included into the dynamical model, incorporating varying labeling lengths as additional prior. A further extension would be to couple the single-gene dynamical models to formulate regulatory motifs, which may be inferred by leveraging recent parameter inference techniques for scalable estimation and model selection[47]. Downstream of scVelo, existing trajectory inference methods may be extended toward informing directionality by robustly integrating velocities to better model cell fate decisions. As such, PAGA[7] has made a first suggestion for inferring directed abstracted representations of trajectories through RNA velocity. Further, scVelo's latent time and velocities could be used together with expression profiles to jointly learn better latent space representations.

Beyond the identification of trajectories and the dynamics of single genes, the dynamic activation of pathways is of central importance. By combining scVelo with enrichment techniques, activated pathways can be inferred in a systematic way, without relying on clustering and differential expression analysis, in analogy to how

we demonstrated the inference of dynamically regulated genes. The identification of dynamic pathways and transcription factors immediately lead to testable hypotheses for contributions to cell state transitions. scVelo's suitability for characterizing transient populations makes it a promising candidate for studying cellular responses to perturbation, which often display drastic switching behaviors. In particular, scVelo could help to mechanistically understand recent machine learning approaches to modeling such response[48] and point to ways to extend them to incorporate splicing dynamics.

In the meantime, scVelo is continuously advanced by the community, bringing efficiency enhancements to the RNA velocity workflow[49]. It has, for instance, contributed to the detailed study of dynamic processes in human lung regeneration[50] and is expected to facilitate the study of lineage decisions and gene regulation, particularly in humans.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0591-3.

Received: 28 October 2019; Accepted: 5 June 2020;
Published online: 3 August 2020

References

1. Griffiths, J. A. et al. Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* **14**, e8046 (2018).
2. Kulkarni, A. et al. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr. Opin. Biotechnol.* **58**, 129–136 (2019).
3. Haghverdi, L. et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
4. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
5. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotem- poral ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
6. Cannoodt, R. et al. Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* **46**, 2496–2506 (2016).
7. Wolf, F. A. et al. PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
8. Saelens, W. et al. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547 (2019).
9. Weinreb, C. et al. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl Acad. Sci. USA* **115**, E2467–E2476 (2018).
10. Tritschler, S. et al. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146**, dev170506 (2019).
11. Junker, J. P. et al. Massively parallel clonal analysis using CRISPR/Cas9 induced genetic scars. Preprint at https://www.biorxiv.org/content/10.1101/056499v2 (2017).
12. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
13. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9- induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
14. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
15. Alemany, A. et al. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
16. Kester, L. & van Oudenaarden, A. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* **23**, 166–179 (2018).
17. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e22 (2019).
18. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494 (2018).
19. Hochgerner, H. et al. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.* **21**, 290–299 (2018).
20. Bastidas-Ponce, A. et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* **146**, dev173849 (2019).
21. Wolf, F. A. et al. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
22. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
23. McInnes, L. & Healy, J. UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
24. Duric, V. et al. Altered expression of synapse and glutamate related genes in post-mortem hippocampus of depressed subjects. *Int. J. Neuropsychopharmacol.* **16**, 69–82 (2013).
25. Ryley Parrish, R. et al. Status epilepticus triggers early and late alterations in brain-derived neurotrophic factor and NMDA glutamate receptor Grin2b DNA methylation levels in the hippocampus. *Neuroscience* **248**, 602–619 (2013).
26. Artegiani, B. et al. A single-cell RNA sequencing study reveals cellular and molecular dynamics of the hippocampal neurogenic niche. *Cell Rep.* **21**, 3271–3284 (2017).
27. Seo, J.-S. et al. Transcriptome analyses of chronic traumatic encephalopathy show alterations in protein phosphatase expression associated with tauopathy. *Exp. Mol. Med.* **49**, e333–e333 (2017).
28. Byrnes, L. E. et al. Lineage dynamics of murine pancreatic development at single-cell resolution. *Nat. Commun.* **9**, 1–17 (2018).
29. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
30. Bechard, M. E. et al. Precommitment low-level Neurog3 expression defines a long-lived mitotic endocrine-biased progenitor pool that drives production of endocrine-committed cells. *Genes Dev.* **30**, 1852–1865 (2016).
31. Krentz, N. A. J. et al. Single-cell transcriptome profiling of mouse and hESC-derived pancreatic progenitors. *Stem Cell Rep.* **11**, 1551–1564 (2018).
32. Ramond, C. et al. Understanding human fetal pancreas development using subpopulation sorting, RNA sequencing and single-cell profiling. *Development* **145**, dev165480 (2018).
33. Yan, F.-F. et al. Congenital hyperinsulinism-associated ABCC8 mutations that cause defective trafficking of ATP-sensitive K+ channels. *Diabetes* **56**, 2339–2348 (2007).
34. Liew, C. W. et al. Insulin regulates carboxypeptidase E by modulating translation initiation scaffolding protein eIF4G1 in pancreatic β cells. *Proc. Natl Acad. Sci. USA* **111**, E2319–E2328 (2014).
35. Wasserfall, C. et al. Persistence of pancreatic insulin mRNA expression and proinsulin protein in type 1 diabetes pancreata. *Cell Metab.* **26**, 568–575.e3 (2017).
36. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
37. Wilkinson, D. J. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.* **10**, 122–133 (2009).
38. Fröhlich, F. et al. Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput. Biol.* **12**, e1005030 (2016).
39. Malkov, Y. A. & Yashunin, D. A. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 824–836 (2018).
40. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
41. Moor, A. E. & Itzkovitz, S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* **46**, 126–133 (2017).
42. Xia, C. et al. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl Acad. Sci. USA* **116**, 19490–19499 (2019).
43. Munsky, B. et al. Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.* **5**, 318 (2009).
44. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 39 (2020).
45. Erhard, F. et al. scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* **571**, 419–423 (2019).
46. Qiu, X. et al. Mapping vector field of single cells. Preprint at https://www.biorxiv.org/content/10.1101/696724v1 (2019).
47. Fröhlich, F. et al. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput. Biol.* **13**, e1005331 (2017).
48. Lotfollahi, M. et al. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
49. Melsted, P. et al. Modular and efficient pre-processing of single-cell RNA-seq. Preprint at https://www.biorxiv.org/content/10.1101/673285v2 (2019).
50. Strunz, M. et al. Longitudinal single cell transcriptomics reveals Krt8+ alveolar epithelial progenitors in lung regeneration. Preprint at https://www.biorxiv.org/content/10.1101/705244v2 (2019).

Methods

Preparing the scRNA-seq data for velocity estimation. The raw data set of hippocampal dentate gyrus neurogenesis[19] is available in the National Center for Biotechnology Information's Gene Expression Omnibus repository under accession number GSE95753. We included samples from two experimental time points: P12 and P35.

The raw data set of pancreatic endocrinogenesis[20] has been deposited under accession number GSE132188. We included samples from the last experimental time point: E15.5.

Annotations of unspliced/spliced reads were obtained using velocyto CLI[18]. Alternatively, reads can be pseudo-aligned with kallisto[49].

The data sets are directly accessible in our Python implementation (https://scvelo.org).

```
import scvelo as scv
adata = scv.datasets.dentategyrus()
adata = scv.datasets.pancreatic_endocrinogenesis()
```

All analyses and results are obtained using default parameters and default data preparation procedures. The count matrices are size normalized to the median of total molecules across cells. The top 2,000 highly variable genes are selected out of those that pass a minimum threshold of 20 expressed counts commonly for spliced and unspliced mRNA. A nearest-neighbor graph (with 30 neighbors) was calculated based on Euclidean distances in principal component analysis space (with 30 principal components) on logarithmized spliced counts. For velocity estimation, first- and second- order moments (means and uncentered variances) are computed for each cell across its 30 nearest neighbors. These are the default procedures in scVelo.

```
scv.pp.filter_and_normalize(adata, min_shared_
counts=20, n_top_genes=2000)
scv.pp.moments(adata, n_neighbors=30, n_pcs=30)
```

After velocity estimation, the gene space can be further restricted to genes that pass a minimum threshold for the coefficient of determination (R^2, derived from the steady-state model) or gene likelihood ($\mathbb{P}((u,s)|(\theta,\eta))$, derived from the dynamical model). For downstream analysis, the UMAP-based embedding, clustering by Louvain community detection[51–53] and diffusion pseudo-time[3] for comparison against latent time are obtained via scanpy[21].

Modeling transcriptional dynamics. On the basis of the dynamical model of transcription shown in Fig. 1, we developed a computational framework for robust and scalable inference of RNA velocity. In the following, we first briefly outline the problem of modeling splicing kinetics, explain the steady-state model and, thereafter, describe the novel dynamical model.

The model of transcriptional dynamics captures transcriptional induction ('on' phase) and repression ('off' phase) of unspliced precursor mRNAs $u(t)$ with state-dependent rates $\alpha^{(k)}$, its splicing into mature mRNAs $s(t)$ with rate β (that is, removing introns from pre-mRNAs and joining adjacent exons to produce spliced mRNAs) and eventual degradation with rate γ, that is,

$$\emptyset \xrightarrow{\alpha^{(k)}} u(t) \xrightarrow{\beta} s(t) \xrightarrow{\gamma} \emptyset$$

Assuming splicing and degradation rates to be constant (time independent), we obtain the gene-specific rate equations

$$\begin{aligned} \frac{du(t)}{dt} &= \alpha^{(k)}(t) - \beta u(t), \\ \frac{ds(t)}{dt} &= \beta u(t) - \gamma s(t), \end{aligned} \tag{1}$$

which describe how the mRNA abundances evolve over time. The time derivative of mature spliced mRNA, termed RNA velocity, is denoted as $\nu(t) = \frac{ds(t)}{dt}$.

The quantities $u(t)$ and $s(t)$ are size-normalized abundances of unspliced and spliced mRNA, respectively, for a cell measured at time point t. In general, the sampled population is not time resolved, and t is a latent variable. Likewise, the cell's transcriptional state k is a latent variable that is not known, and the rates $\alpha^{(k)}$, β and γ are usually not experimentally measured.

Steady-state model. Under the assumption that we observe both transcriptional phases of induction and repression, and that these phases last sufficiently long to reach a transcribing (active) and a silenced (inactive) steady-state equilibrium, velocity estimation can be simplified as follows. In steady states, we obtain, on average, a constant transcriptional state where $\frac{ds}{dt} = 0$ which, by solving Equation 1, yields $\gamma\prime = \frac{\gamma}{\beta}$ as the steady-state ratio of unspliced to spliced mRNA. It indicates where mRNA synthesis and degradation are in balance. Steady states are expected at the lower and upper quantiles in phase space, that is, where mRNA levels reach minimum and maximum expression, respectively. Hence, the ratio can be approximated by a linear regression on these extreme quantiles.

It can be solved analytically via a least square fit and is given by

$$\gamma\prime = \frac{u^T s(t)}{\|s\|^2}, \tag{2}$$

where $u = (u_1, \ldots, u_n)$ and $s = (s_1, \ldots, s_n)$ are vectors of size-normalized unspliced and spliced counts for a particular gene that lie in the lower or upper extreme quantile—that is, n is only a fraction of the total number of cells. A positive offset can be included into the least square fit to account for basal transcription. The steady-state ratio is then given by $\gamma\prime = \frac{\mathrm{Cov}(u,s)}{\mathrm{Var}(s)}$, and the offset is given by $o = \bar{u} - \bar{s}\gamma\prime$, where $\bar{u}$ and $\bar{s}$ are the means of u and s, respectively.

Then, velocities can be computed as deviations from this steady-state ratio—that is,

$$\nu_i = u_i - \gamma\prime s_i. \tag{3}$$

Whereas a constant transcriptional state is reflected by zero velocity, the direction and relative speed during a dynamic process are given by the sign and magnitude of non-zero velocity.

Taken together, under this simplified model, velocities are estimated along two simple equations as steady-state deviations. With this notion, the cumbersome problem of estimating latent time is circumvented. Further, velocities depend on only one ratio instead of absolute values of kinetic rates, which technically corresponds to measuring all entities in units of splicing rate, thus effectively assuming one common splicing rate $\beta = 1$ across all genes.

scVelo hosts an efficient estimation procedure of velocities derived from the steady-state model.

```
scv.tl.velocity(adata, mode='steady_state')
```

Dynamical model. *Model description.* In recognition that steady states are not always captured and that splicing rates differ between genes, we establish a framework that does not rely on these restrictions. The analytical solution to the gene-specific rate equations in Equation 1 is found by integration, which yields

$$\begin{aligned} u(t) &= u_0 e^{-\beta\tau} + \tfrac{\alpha^{(k)}}{\beta}\left(1 - e^{-\beta\tau}\right), \\ s(t) &= s_0 e^{-\gamma\tau} + \tfrac{\alpha^{(k)}}{\gamma}(1 - e^{-\gamma\tau}) + \tfrac{\alpha^{(k)} - \beta u_0}{\gamma - \beta}\left(e^{-\gamma\tau} - e^{-\beta\tau}\right), \tau = t - t_0^{(k)}, \end{aligned} \tag{4}$$

with parameters of reaction rates $\theta = \left(\alpha^{(k)}, \beta, \gamma\right)$, cell-specific time points $t \in (t_1, \ldots, t_N)$ and initial conditions $u_0 = u(t_0)$, $s_0 = s(t_0)$.

Gene activity is orchestrated by transcriptional regulation, implying that gene upregulation and downregulation are inscribed by alterations in the state-dependent transcription rate $\alpha^{(k)}$. That is, $\alpha^{(k)}$ can have multiple configurations each encoding one transcriptional state. For the model, this requires an additional parameter set, assigning a transcriptional state k to each cell. Consequently, not only $\alpha^{(k)}$ but also the initial conditions $u_0^{(k)}$, $s_0^{(k)}$ are state dependent, as well as the time point of switching states $t_0^{(k)}$. In the following, we consider four phases, induction ($k = 1$) and repression ($k = 0$), each with an associated potential steady state ($k = ss_1$ and $k = ss_0$). Consider a transition from one state k to a subsequent state $k\prime$—for example, from induction to repression. Then, the initial conditions of the next state are given by evaluating the trajectory of the current state at its respective switching time point

$$\begin{aligned} u_0^{(k\prime)} &= u\left(t_0^{(k\prime)} | \theta^{(k)}\right), \\ s_0^{(k\prime)} &= s\left(t_0^{(k\prime)} | \theta^{(k)}\right), \end{aligned} \tag{5}$$

where $t_0^{(k\prime)}$ is learned jointly with the parameters of reaction rates, as will be described later.

Being at state k, abundances can potentially reach their steady state in the limit

$$\left(u_\infty^{(k)}, s_\infty^{(k)}\right) = \left(\tfrac{\alpha^{(k)}}{\beta}, \tfrac{\alpha^{(k)}}{\gamma}\right). \tag{6}$$

The number of potential steady states equals the number of transcriptional states.

Parameter inference. Recovering the splicing kinetics entails inferring the model parameters—that is, reaction rates $\theta^{(k)}$—at time point t_i for each cell that couples the measurement to the system of differential equations by assignment onto the phase trajectory, where state $k_i \in \{1, 0, ss_1, ss_0\}$ to which each cell is assigned, and at switching time point $t_0^{(k)}$ of transitioning to another state.

Let the observations u_i^{obs} and s_i^{obs} be size-normalized unspliced and spliced counts for a particular gene, where u_i^{obs} is rescaled to have the same variance as s_i^{obs}. We further consider only counts that are non-zero in both unspliced and spliced mRNA. Now, let the model estimate be $\hat{x}(t) = (\hat{u}(t), \hat{s}(t))$. We aim to find a phase trajectory specified by $\hat{x}(t)$ that best describes the observations. We define the residuals of the observations to the phase trajectory as signed Euclidean distances by $e_i = \mathrm{sign}(s_i^{obs} - \hat{s}_{t_i}) \cdot \left\| x_i^{obs} - x_{t_i}(\theta) \right\|$, also referred to as Deming

residuals. Under the assumption that the residuals are normally distributed with $e_i \sim N(0, \sigma^2)$ with a gene-specific σ constant across cells within one transcriptional state and that the observations are independent and identically distributed, the likelihood-based framework is derived in the following.

The likelihood for a particular gene writes

$$\mathcal{L}(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2n} \sum_{i}^{n} \frac{\left\| x_i^{obs} - x_{t_i}(\theta) \right\|^2}{\sigma^2} \right). \tag{7}$$

Accordingly, the negative log-likelihood to be minimized is given by

$$l(\theta) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\frac{1}{n}\sum_{i}^{n}\left\| x_i^{obs} - x_{t_i}(\theta) \right\|^2, \tag{8}$$

where $\theta = \left(\alpha^{(k)}, \beta, \gamma\right)$.

Alternatively, it can also be modelled with Laplacian residuals, thereby changing the least squares to least absolute residuals, which fundamentally portrays the same results.

The cell-specific latent time points are required for coupling an observation to the system of differential equations to obtain a mapping of x_i^{obs} to $\hat{x}(t|\theta)$. Hence, solving for kinetic rates relies on also estimating latent time, which illustrates the problem complexity. This is solved by EM iterating between finding the optimal parameters of kinetic rates and the latent variables of time and state, initialized with parameters derived from the steady-state model—that is, $\beta = 1$ and $\gamma\prime = \frac{\mathbf{u}^T\mathbf{s}}{\|\mathbf{s}\|}$, where $\mathbf{u}$ and $\mathbf{s}$ are size-normalized unspliced and spliced counts from extreme quantile cells. The cell-specific state is initialized to be induction or repression depending on whether the sample lies above or below the steady-state ratio, respectively—that is,

$$k_i = \begin{cases} 1, & \text{if } u_i - \gamma\prime s_i > 0 \\ 0, & \text{else} \end{cases}$$

The transcription rates are initialized with $\alpha^{(k_i=1)} = \max(s)$ and $\alpha^{(k_i=0)} = 0$. Note that only two initial states are necessary to initialize the transcription rates, whereas, during the optimization, we also explicitly model potential steady states to accurately model the accumulating noise therein.

After initializing the system with meaningful parameters, the EM algorithm iteratively applies the following two steps:

- E-step: Given $\hat{x}(t|\theta)$ parametrized by the current estimate of θ, we assign a latent time t_i to the observed value x_i^{obs} by minimizing the distance to the phase trajectory $(\hat{x}(t|\theta))_t$ in each transcriptional state. State likelihoods are then assigned to each cell, which yield an expected value of the log-likelihood (by integrating over all possible outcomes for transcriptional states).
- M-step: The parameters θ and t_0 are updated to maximize the log-likelihood.

We explicitly model both transient states of induction and repression as well as (potentially unobserved) active and inactive steady states. The state likelihoods are determined by the distance of the observations to the four segments of the phase trajectory, parametrized by kinetic rates and latent time.

For latent time, we adopt an explicit formula that approximates the optimal time assignment for each cell. This is applied throughout the EM framework mainly for computation efficiency reasons, whereas (exact) optimal time assignment is applied in the last iterations. The approximation of optimal latent time is obtained as follows.

The equation for spliced mRNA levels can be rewritten as a function of unspliced mRNA levels:

$$s(t) = \tilde{\beta}u(t) + \frac{\alpha}{\gamma} - \tilde{\beta}\frac{\alpha}{\beta} + \left(s_0 - \frac{\alpha}{\gamma} - \tilde{\beta}\left(u_0 - \frac{\alpha}{\beta} \right) \right) e^{-\gamma\tau}, \tag{9}$$

where $\tilde{\beta} := \frac{\beta}{\gamma - \beta}$ constitutes the linear dependence of unspliced on spliced mRNA.

If we denote $\tilde{s}(t) := s(t) - \tilde{\beta}u(t)$ and $\tilde{s}_\infty := s_\infty - \tilde{\beta}u_\infty$, the equation can be rewritten as

$$\tilde{s}(t) - \tilde{s}_\infty = (\tilde{s}_0 - \tilde{s}_\infty)e^{-\gamma\tau}$$

where τ can be solved explicitly for each cell by taking the inverse

$$\tau_i = -\frac{1}{\gamma}\log\left(\frac{\tilde{s}_i - \tilde{s}_\infty}{\tilde{s}_{0,i} - \tilde{s}_\infty} \right) \tag{10}$$

If $\beta > \gamma$, the time assignment is thus obtained as inverse of a positive linear combination of unspliced and spliced mRNA dynamics. For genes with $\beta \leq \gamma$, we can instead directly take the inverse of $u(t)$, which is given by

$$\tau_i = -\frac{1}{\beta}\log\left(\frac{u_i - u_\infty}{u_{0,i} - u_\infty} \right). \tag{11}$$

This explicit time assignment is used throughout the parameter fitting, whereas, in the last iteration, latent time is solved optimally likelihood based.

Although the explicit time assignments entail a negligibly small likelihood downturn, this procedure fundamentally yields the same result at a 30-fold speedup over solving it optimally throughout the parameter fitting.

Finally, in the M-step, the parameter set of kinetic rates is updated to maximize the log-likelihood. We used the downhill simplex method, also referred to as the Nelder–Mead method, which has proven to be the most robust and efficient approach, particularly when compared against classical gradient descent algorithms. It is a derivative-free method based on a simplex—that is, a convex hull of $n + 1$ vertices, where n is the number of parameters of kinetic rates. The method performs a sequence of transformations (reflection, expansion and contraction) of the working simplex, aimed at decreasing the negative log-likelihood. Convergence is reached once the change in the parameters or the change in the resulting likelihood between subsequent iterations is smaller than 1/10,000, which takes between a few up to 100 iterations. The fitting procedure for genes that display evident kinetics usually converges within ten iterations.

The resulting likelihood in Equation 7 for a particular gene corresponds to its goodness of fit indicating whether the gene displays any evident kinetics that can be well described by the learned phase trajectory. Hence, the gene likelihood serves as a way to identify putative drivers of the underlying process.

scVelo provides a flexible and efficient module to estimate reaction rates and latent variables.

```
scv.tl.recover_dynamics(adata)
```

RNA velocity can then be estimated using the explicit description of inferred splicing kinetics.

```
scv.tl.velocity(adata, mode='dynamical')
```

Computing transition probabilities from velocities. Assuming that velocities truthfully describe the actual dynamics locally, we estimate transition probabilities of cell-to-cell transitions. Let $s_i \in \mathbb{R}^{n \times d}$ be the gene expression matrix of d genes across n cells. Further, we estimated the velocity vectors $(v_i)_{i=1,\ldots,n}$ in the previous section, of which $v_i \in \mathbb{R}^d$ predicts the change in gene expression of cell $s_i \in \mathbb{R}^d$.

Cell s_i is expected to have a high probability of transitioning toward cell s_j when the corresponding change in gene expression $\delta_{ij} = s_j - s_i$ matches the predicted change according to the velocity vector v_i. We apply cosine similarity—that is, cosine of the angle between two vectors,

$$\pi_{ij} = \cos\angle(\delta_{ij}, \mathbf{v}_i) = \frac{\delta_{ij}^T v_i}{\|\delta_{ij}\| \|v_i\|}, \tag{12}$$

where $\pi_{ii} = 0$. It solely measures similarity in directionality, not in magnitude, and ranges from -1 (opposite direction) over 0 (orthogonal, thus maximally dissimilar) to 1 (identical direction). The resulting similarity matrix π encodes a graph, which we refer to as velocity graph. Optionally, a variance-stabilizing transformation can be applied such that the cosine correlation is computed between $\text{sign}(\delta_{ij})\sqrt{|\delta_{ij}|}$ and $\text{sign}(v_i)\sqrt{|v_i|}$. Note that not every possible cell-to-cell transition is considered but only transitions within a knn graph neighborhood including direct neighbors and respective neighbors of neighbors. With the default of 30 neighbors, the recursive neighbor approach yields 100–200 potential transitions, of which only few yield relatively high probabilities. It fundamentally satisfies the same results as when computed with a fixed 100-neighbor graph yet circumvents the necessity to compute a large and computationally expensive knn graph.

An exponential kernel is applied to transform the cosine correlations into transition probabilities

$$\tilde{\pi}_{ij} = \frac{1}{z_i}\exp\left(\frac{\pi_{ij}}{\sigma_i^2} \right), \tag{13}$$

with row normalization factors $z_i = \sum_j \exp\left(\frac{\pi_{ij}}{\sigma_i^2} \right)$ and kernel width parameters σ_i optionally adjusted for each cell locally (across neighboring cells).

The transition probabilities are aggregated into a transition matrix $\tilde{\boldsymbol{\pi}}$ describing the Markov chain of the differentiation process. Throughout the literature, $\tilde{\boldsymbol{\pi}}$ is also referred to as transport map, which serves as a coupling of a developmental process. A distribution of cells $\mu = (\mu_1, \ldots, \mu_n)$ can be pushed through the transport map to obtain its descendant distribution. Reversely, a distribution μ can be pulled back through the transport map to obtain its ancestors—that is,

$$\begin{aligned} \mu^{des} &= \mu \cdot \tilde{\pi}, \\ \mu^{anc} &= \mu \cdot \tilde{\pi}^T. \end{aligned} \tag{14}$$

A descendant or ancestor distribution of a set of cells $\mathcal{S} = \{s_1, \ldots, s_n\}$ can be obtained by setting $\mu_i = 1_{s_i \in \mathcal{S}}$, where 1 denotes the indicator function.

scVelo efficiently computes the velocity graph by sparse and vectorized implementation.

```
scv.tl.velocity_graph(adata)
```

Gene-shared latent time. After inferring parameters of kinetic rates, a gene-shared latent time is computed as follows. First, gene-specific time points of well-fitted

genes (with a likelihood of at least 0.1) are normalized to a common overall time scale. The root cells of the differentiation process are obtained by computing the stationary states μ^* satisfying

$$\mu^* = \mu^* \tilde{\pi}^T \tag{15}$$

which is given by the left eigenvectors of $\tilde{\pi}^T$ corresponding to an eigenvalue of 1.

Now, for every root cell o, we compute the p-quantile Q_p^g of all respective gene-specific time increments across all genes g

$$t_{i,o} = Q_p^g\left(t_i^g - t_o^g\right). \tag{16}$$

where p is chosen such that it maximizes the correlation between the resulting gene-shared latent time course $t_{0,o}, \cdots, t_{N,o}$ and its convolution across the local neighborhood of cells. The rationale behind taking the p-quantile is the adaption to a non-uniform density of cells along the time course, as cells often tend to accumulate in later time points. We find optimal values for p, lower than the median, at around 20–30%. A high correlation with the convolution of latent time improves robustness and consistency in the estimate.

That is, for each root cell, we find the respective time increments that achieve best overall accordance with the learned dynamics and yield local coherence.

Gene-shared latent time of cell i is then obtained as mean across all root cells

$$t_i = \left\langle t_{i,o} \right\rangle_o. \tag{17}$$

Finally, for robustness, by regressing the gene-shared latent time course against its neighborhood convolution, we detect inconsistent time points and replace them with their convolutions.

```
scv.tl.recover_latent_time(adata)
```

Projection of velocities into the embedding. The projection of velocities into a lower-dimensional embedding (for example, UMAP) for a cell i is obtained on the basis of a transition matrix $\boldsymbol{\pi}$ (see previous section), which contains probabilities of cell-to-cell transitions that are in accordance with the corresponding velocity vectors,

$$\tilde{\pi}_{ij} = \frac{1}{z_i} \exp\left(\frac{\cos\angle\left(x_j - x_i, v_i\right)}{\sigma_i^2}\right),$$

with row normalization factors $z_i = \sum_j \exp\left(\frac{\pi_{ij}}{\sigma_i^2}\right)$ and kernel width parameters σ_i.

The positions of cells in an embedding, such as t-distributed stochastic neighbor embedding or UMAP, are described by a set of vectors $s_1, \ldots, s_n$. Given the normalized differences of the embedding positions $\tilde{\delta}_{ij} = \frac{\tilde{s}_j - \tilde{s}_i}{\left\|\tilde{s}_j - \tilde{s}_i\right\|}$, the embedded velocity is estimated as the expected displacements with respect to the transition matrix

$$\boldsymbol{v}_i = \mathbb{E}_{\boldsymbol{\pi}_i}\left[\tilde{\delta}_{i\cdot}\right] = \sum_{j \neq i}\left(\tilde{\pi}_{ij} - \frac{1}{n}\right)\tilde{\delta}_{ij}, \tag{18}$$

where subtracting $\frac{1}{n}$ corrects for the non-uniform density of points in the embedding.

The directional flow is visualized as single-cell velocities or streamlines in any embedding.

```
scv.pl.velocity_embedding(adata, basis='umap')
scv.pl.velocity_embedding_stream(adata, basis='umap')
```

Accounting for stochasticity through second-order moments. The model for velocity estimation can be extended with higher-order moments, obtained by treating transcription, splicing and degradation as probabilistic events. In this regard, the probabilities of all possible reactions corresponding to these events occurring within an infinitesimal time interval $(t, t + dt]$ are provided as follows:

$$\begin{aligned}
\mathbb{P}(u_{t+dt} = u_t + 1, s_{t+dt} = s_t) &= \quad \alpha\, dt,\\
\mathbb{P}(u_{t+dt} = u_t - 1, s_{t+dt} = s_t + 1) &= \quad \beta u_t\, dt,\\
\mathbb{P}(u_{t+dt} = u_t, s_{t+dt} = s_t - 1) &= \quad \gamma s_t\, dt,
\end{aligned} \tag{19}$$

where we denoted $u_t = u(t), s_t = s(t)$ to facilitate clarity.

From Equation 19, the time derivative for the uncentered moment $\left\langle u_t^l s_t^k \right\rangle = \mathbb{E}\left[u_t^l s_t^k\right]$ is derived as

$$\begin{aligned}
\tfrac{d\left\langle u_t^l s_t^k \right\rangle}{dt} = \quad & \left\langle \alpha\left((u_t+1)^l s_t^k - u_t^l s_t^k\right)\right\rangle + \left\langle \beta u_t\left((u_t-1)^l (s_t+1)^k - u_t^l s_t^k\right)\right\rangle\\
& + \left\langle \gamma s_t\left(u_t^l (s_t-1)^k - u_t^l s_t^k\right)\right\rangle.
\end{aligned}$$

Hence, the first- and second-order dynamics are given by

$$\begin{aligned}
\tfrac{d\langle u_t\rangle}{dt} &= \alpha - \beta\langle u_t\rangle,\\
\tfrac{d\langle s_t\rangle}{dt} &= \beta u_t - \gamma\langle s_t\rangle,\\
\tfrac{d\langle u_t^2\rangle}{dt} &= \alpha + 2\alpha\langle u_t\rangle + \beta\langle u_t\rangle - 2\beta\langle u_t^2\rangle,\\
\tfrac{d\langle u_t s_t\rangle}{dt} &= \alpha\langle s_t\rangle + \beta\langle u_t^2\rangle - \beta\langle u_t s_t\rangle - \gamma\langle u_t s_t\rangle,\\
\tfrac{d\langle s_t^2\rangle}{dt} &= \beta\langle u_t\rangle + 2\beta\langle u_t s_t\rangle + \gamma\langle s_t\rangle - 2\gamma\langle s_t^2\rangle,
\end{aligned} \tag{20}$$

The moments for each cell are computed among a preset number of nearest neighbors of the corresponding cell.

This extension can be easily applied to the steady-state model. Using both first- and second-order moments, the steady-state ratio is obtained from the system

$$\underbrace{\begin{pmatrix}\langle u_t\rangle\\ \langle u_t\rangle + 2\langle u_t s_t\rangle\end{pmatrix}}_{u\prime} = \gamma\prime \underbrace{\begin{pmatrix}\langle s_t\rangle\\ 2\langle s_t^2\rangle - \langle s_t\rangle\end{pmatrix}}_{s\prime} + \epsilon\prime \tag{21}$$

where $\mathbb{E}[\epsilon\prime|s\prime] = 0$ and $\mathrm{Cov}[\epsilon\prime|s\prime] = \Omega$.

The steady-state ratio can be solved explicitly by generalized least squares and is given by

$$\gamma\prime = \left(s\prime^T \Omega^{-1} s\prime\right)^{-1} s\prime^T \Omega^{-1} u\prime \tag{22}$$

The stochastic model thereby exploits not only the relationship between unspliced and spliced mRNA abundances but also their covariation.

```
scv.tl.velocity(adata, mode='stochastic')
```

Accounting for different kinetic regimes with a differential kinetic test. Distinct cell types and lineages might exhibit different kinetics regimes, as these can be governed by a different network structure. Even if cell types or lineages are related, kinetics can be differential due to alternative splicing, alternative polyadenylation and modulations in degradation.

The likelihood-based framework allows us to address this issue with a likelihood ratio test for differential kinetics to detect clusters and lineages that display kinetic behavior that cannot be sufficiently explained by a single model for the overall dynamics. Each cell type is tested whether an independent fit yields a significantly improved likelihood.

The likelihood ratio (LR) test statistic is given by

$$LR = -2\ln\frac{sup_\theta\, \mathcal{L}(\theta)}{sup_{\theta\prime}\, \mathcal{L}(\theta\prime)}, \tag{23}$$

where θ and $\theta\prime$ correspond to a one-kinetic and two-kinetic model, respectively. The alternative hypothesis is tested on the cell type that is most distant from the learned overall trajectory, which yields a new phase trajectory. The remaining cell types are clustered into the new kinetic regime if that improves the LR further. By Wilk's theorem, LR has an asymptotic χ^2 distribution, and the ratio can be tested for significance. Note that, for efficiency reasons, by default an orthogonal regression is used instead of a full phase trajectory as alternative hypothesis to detect cell types for different kinetic regimes.

Accounting for insufficiently observed kinetics with prior information. Splicing kinetics might be insufficiently observed, for instance, when, for a particular gene, only a small portion of the overall dynamics is disclosed at the very end of the process. This manifests in a straight line rather than a curve in the unspliced to spliced phase diagram, which constitutes a fundamental issue to all existing models for velocity estimation. The steady-state and stochastic model would simply fit the line and arbitrarily assign positive and negative velocities to observations that might be fully in upregulation. Also, the dynamical model is challenged to determine whether an upregulation or downregulation should be fit.

This ambiguity can be observed in two application scenarios: (1) the gene is active only in a small window—that is, it is upregulated only at the very end or downregulated at the very beginning of a developmental process, or (2) the observed time frame in the data covers only a small part of the time frame of the underlying dynamical process—for example, when the process occurs in a fast fashion.

Mathematically, these scenarios result in two local optima, thus entailing an identifiability issue, which can be solved only with additional information. We addressed these shortcomings by extending the dynamical model with a 'root prior'. Intuitively, if the model gets passed the root of the kinetics, it can conclude whether and upregulation or downregulation has to be fit, thereby resolving the ambiguity.

In scenario 1, the root prior can be internally obtained from genes that are sufficiently informative to uncover the root of the process. In scenario 2, it has to be obtained from prior knowledge, such as the first experimental time point or a known progenitor population.

Genes with partial kinetics can be easily detected by their R^2 value. We define a kinetic to be ambiguous if its R^2 value exceeds a threshold of 0.95. Given a root cell o, we penalize observations that have a latent time assigned earlier than the root, resulting in a regularization term

$$R(\theta) = \frac{1}{n}\sum_{i}^{n} 1_{\{t_i < t_j\}} \left\| x_i^{obs} - x_o^{obs} \right\|^2 \quad (24)$$

that is added to the negative log-likelihood $l(\theta) + \lambda R(t, \theta)$.

Validation metrics. To validate the coherence of the velocity vector field, we define a consistency score for each cell i as the mean correlation of its velocity v_i with velocities from neighboring cells,

$$c_i = \left\langle \mathrm{corr}\left(v_i, v_j\right) \right\rangle_j, \text{ where cell } j \text{ is neighboring cell } i.$$

To validate the contribution of a selection of genes (for example, top likelihood-ranked genes) to the overall inferred dynamics, we define a reconstructability score as follows. The velocity graph consisting of correlations between velocities and cell-to-cell transitions (see previous sections) is computed once (1) including all genes yielding $\boldsymbol{\pi}$ and once (2) including only the selection of genes yielding $\boldsymbol{\pi}\prime$. The reconstructability score is defined as the median correlation of outgoing transitions from cell i to all cells that it can potentially transition to—that is,

$$r = \mathrm{median}_i\, \mathrm{corr}\left(\pi_i, \pi_i'\right) \text{ across all cells } i.$$

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data sets analyzed in this paper are publicly available and published. The annotated data, results and Python implementation are available at https://scvelo.org. The raw data set of hippocampal dentate gyrus neurogenesis is available in the National Center for Biotechnology Information's Gene Expression Omnibus repository under accession number GSE95753. We included samples from two experimental time points: P12 and P35. The raw data set of pancreatic endocrinogenesis has been deposited under accession number GSE132188. We included samples from the last experimental time point: E15.5.

Code availability

The results reported in this paper and our Python implementation are available at https://scvelo.org.

References

51. Blondel, V. D. et al. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
52. Gu, G. et al. Global expression analysis of gene regulatory pathways during endocrine pancreatic development. *Development* **131**, 165–179 (2004).
53. de Lichtenberg, K. H. et al. Notch controls multiple pancreatic cell fate regulators through direct hes1-mediated repression. Preprint at https://www.biorxiv.org/content/10.1101/336305v1 (2018).

Acknowledgements

We thank P. Kharchenko and S. Linnarsson for stimulating discussions, M. Luecken for valuable feedback on the manuscript and S. Tritschler for valuable feedback on the biological applications. This work was supported by BMBF grants (01IS18036A and 01IS18053A); by the German Research Foundation (DFG) within the Collaborative Research Centre 1243, Subproject A17; by the Helmholtz Association (sparse2big and ZT-I-0007); and by the Chan Zuckerberg Initiative DAF (182835). M.L. further acknowledges financial support by the DFG through the Graduate School of Quantitative Biosciences Munich (GSC 1006), by the Joachim Herz Stiftung Foundation and by the Bayer Foundation.

Author contributions

V.B. designed and developed the method, implemented scVelo and analyzed the data. F.J.T. conceived the study with contributions from V.B. and F.A.W. V.B., F.A.W. and F.J.T. wrote the manuscript with contributions from the coauthors. S.P. contributed to developing scVelo. M.L. contributed to developing validation metrics. All authors read and approved the final manuscript.

Competing interests

F.A.W. is a full-time employee of Cellarity Inc.; the present work was carried out as an employee of Helmholtz Munich. F.J.T. reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0591-3.

Correspondence and requests for materials should be addressed to F.A.W. or F.J.T.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Fabian J. Theis, F. Alexander Wolf

Last updated by author(s): May 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data is available through https://github.com/theislab/scvelo_notebooks.
Data analysis	The main code is available through https://github.com/theislab/scvelo and based on open source Python packages.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data is available through https://github.com/theislab/scvelo_notebooks.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No experiments in study
Data exclusions	No experiments in study
Replication	No experiments in study
Randomization	No experiments in study
Blinding	No experiments in study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging