nature biotechnology

SUPPLEMENTARY INFORMATION

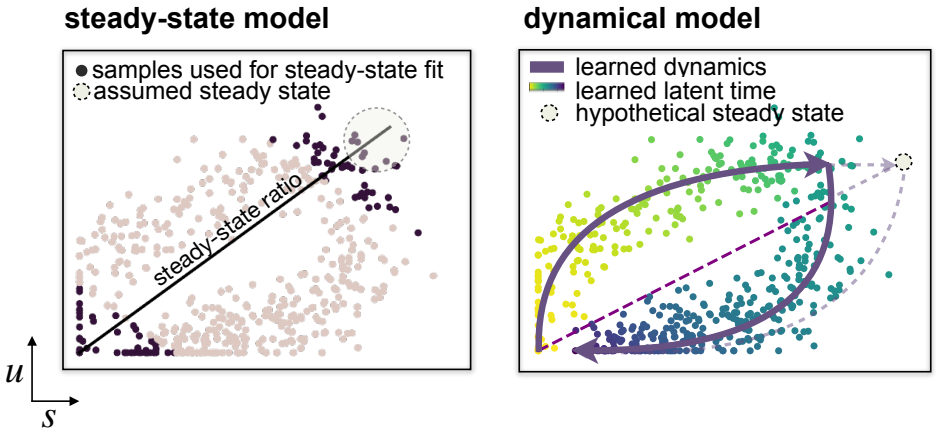
https://doi.org/10.1038/s41587-020-0591-3

In the format provided by the authors and unedited.

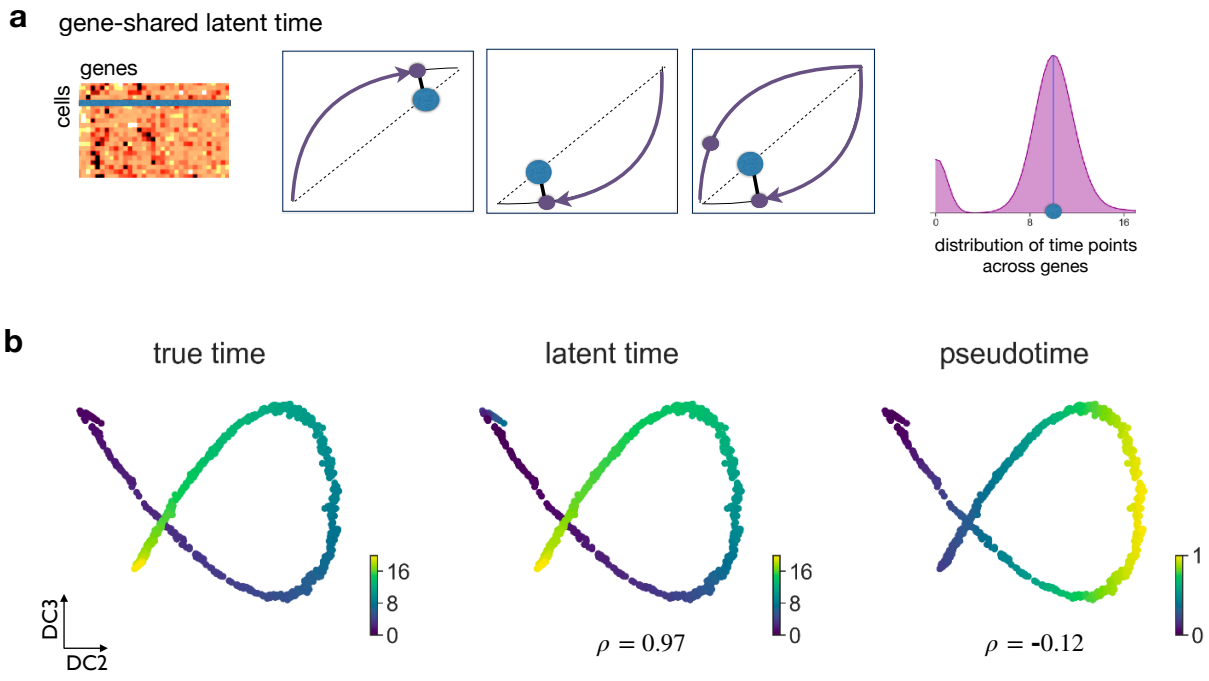
Generalizing RNA velocity to transient cell states through dynamical modeling

In the format provided by the authors and unedited

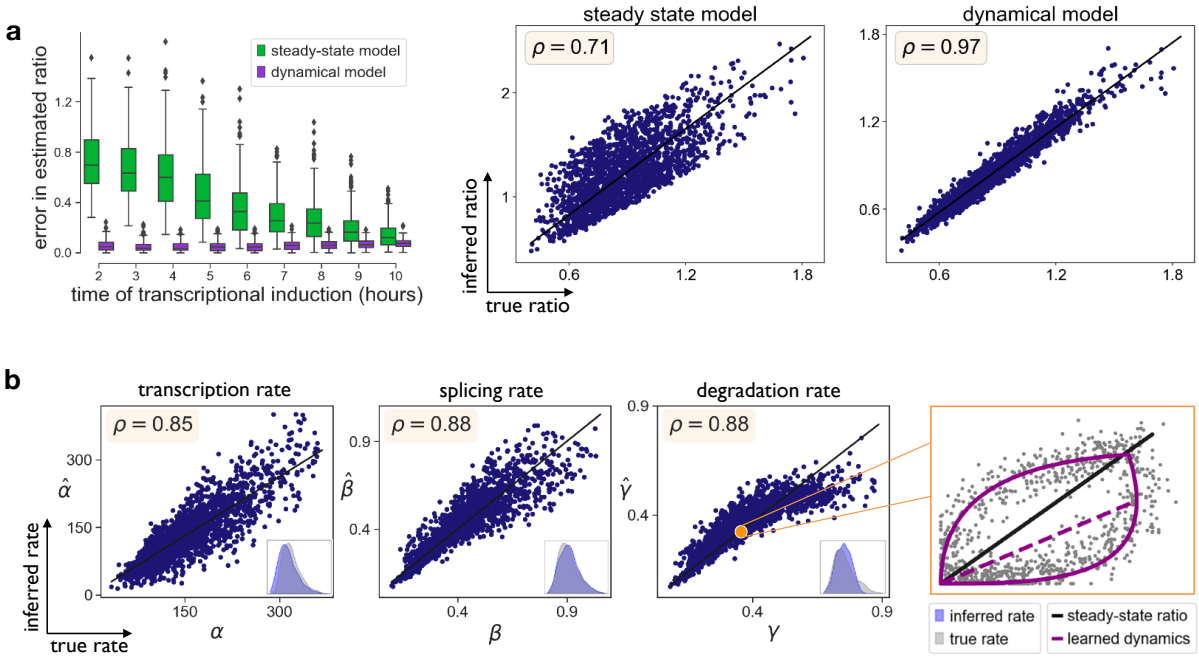
Supplementary Information



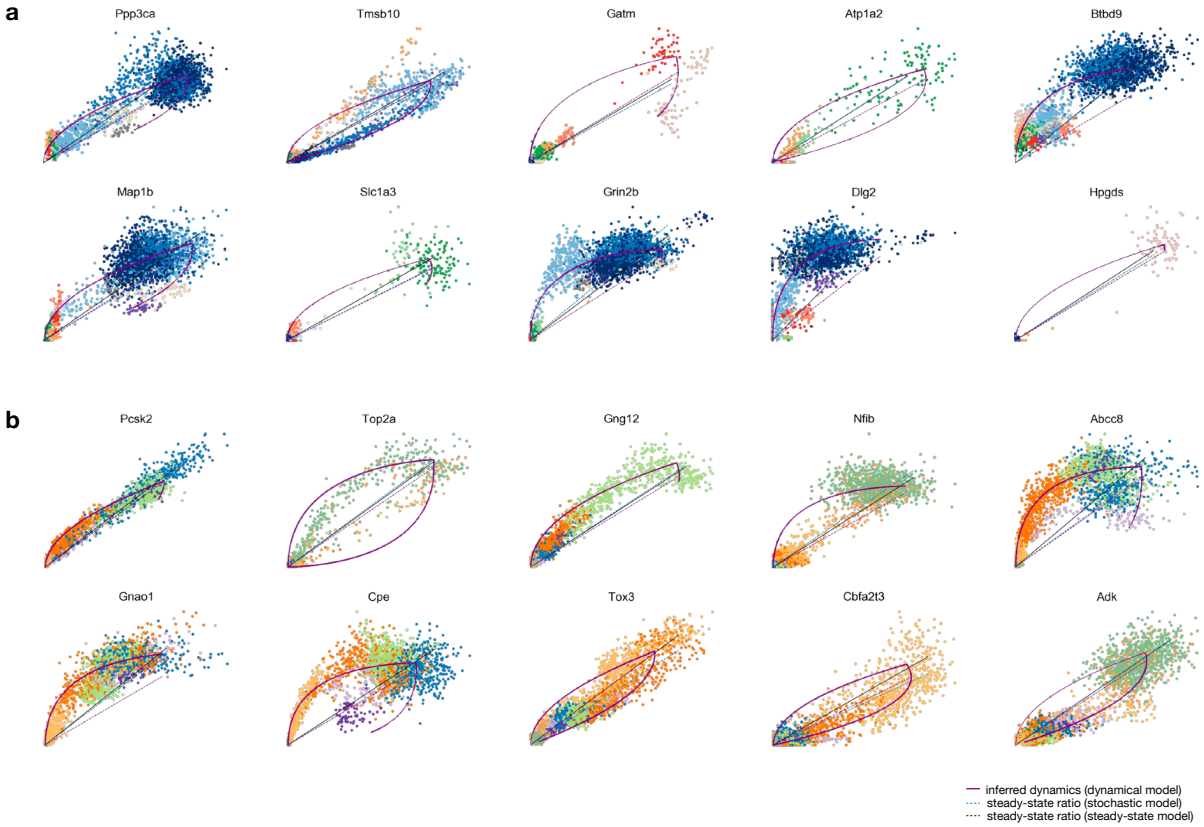
Supplementary Figure 1 | Model fit comparison.
The steady-state model determines velocities by quantifying how observations deviate from a ratio of unspliced to spliced mRNA describing an assumed steady-state equilibrium. This steady-state ratio is obtained by performing a linear regression restricting the input data to the extreme quantiles. By contrast, the dynamical model solves the full splicing kinetics and thereby captures non-observed steady states. The error made by the steady-state model for such cases becomes evident when comparing the slope of the line describing the steady-state ratio.



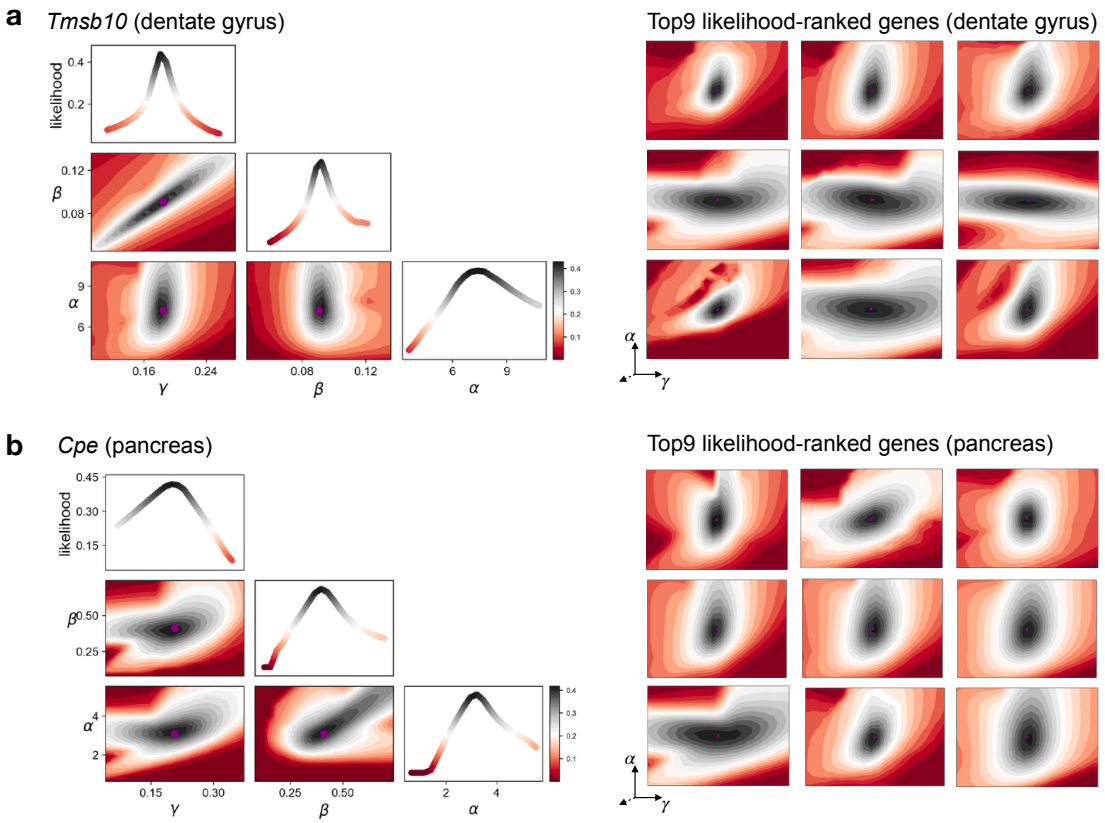
Supplementary Figure 2 | Inferring a universal gene-shared latent time.
a. A shared latent time is constructed by coupling gene-wise latent times to a unified one-dimensional time covariate across genes. For instance, if a cell is confidently assigned to a later time point in two genes (left) while being ambiguously positioned in a third gene (right), the decision for the later time point is imposed by the majority. Thereby, the cell's position can be confidently identified by sharing information across genes.
b. The latent time recovered by the dynamical model faithfully captures the underlying real time. While latent time yields a near-perfect Pearson correlation, diffusion pseudotime is unable to reconstruct real time.



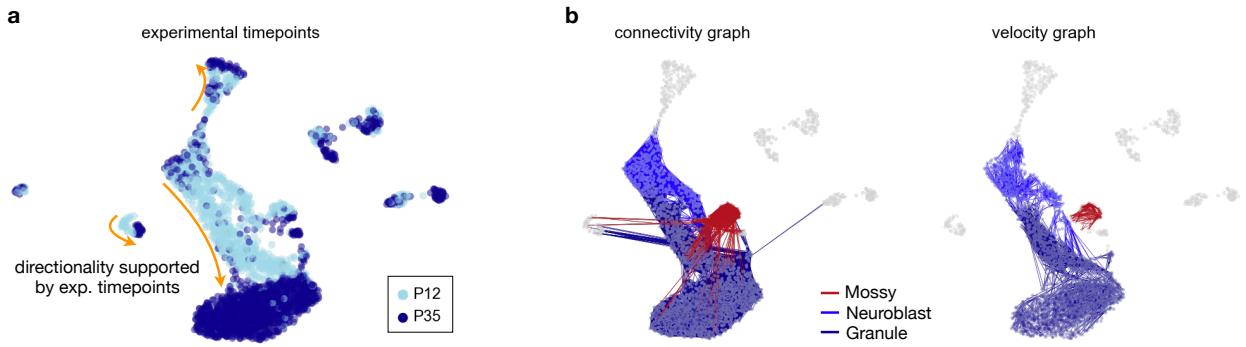
Supplementary Figure 3 | Identifying parameters of reaction rates in transient populations.
a. To validate the sensitivity of both models with respect to varying parameters in simulated splicing kinetics, we randomly sampled 2,000 log-normally distributed parameters of reaction rates $\theta=(\alpha,\beta,\gamma)$ with $log(\theta)\sim N(\mu,\Sigma)$ with $\mu=(5,0.2,0.05)$ and $\Sigma_{ii}=0.4^2$, $\Sigma_{12}=\Sigma_{12}=0.4^2\cdot 0.8$, $\Sigma_{23}=0.4^2\cdot 0.2$, and 1,000 time events following the Poisson law, i.e., the number of arrivals in the time interval τ being Poisson$(\lambda\tau)$ distributed with $\lambda=0.1$. The time point of the transcriptional switch (here: from induction to repression) was uniformly varied between 2 and 10 hours. As transcriptional induction progresses the abundances converge to an equilibrium steady-state. However, for shorter duration where no equilibrium state has yet been reached, the steady-state model systematically fails to capture the true steady-state ratios. In contrast, the likelihood-based dynamical model consistently infers the true ratio regardless of induction time. The Pearson correlation between the true and inferred steady-state ratio increases from 0.71 to 0.97 when using the dynamical model.
b. The dynamical model reliably recovers the true parameters of the splicing kinetics. For all kinetic rates (transcription, splicing and degradation), we obtain a correlation of at least 0.85 across the range of 2,000 parameter values for each rate. By imposing an overall timescale of 20 hours as prior information, the dynamical model recovers the absolute values of the reaction rates. As illustrated by the phase diagram, the dynamical model leverages the property of the curvature to estimate the parameter values, thereby also correctly identifies the relative steady-state ratio (dashed purple line) as opposed to the ratio obtained by the steady-state model (black line).



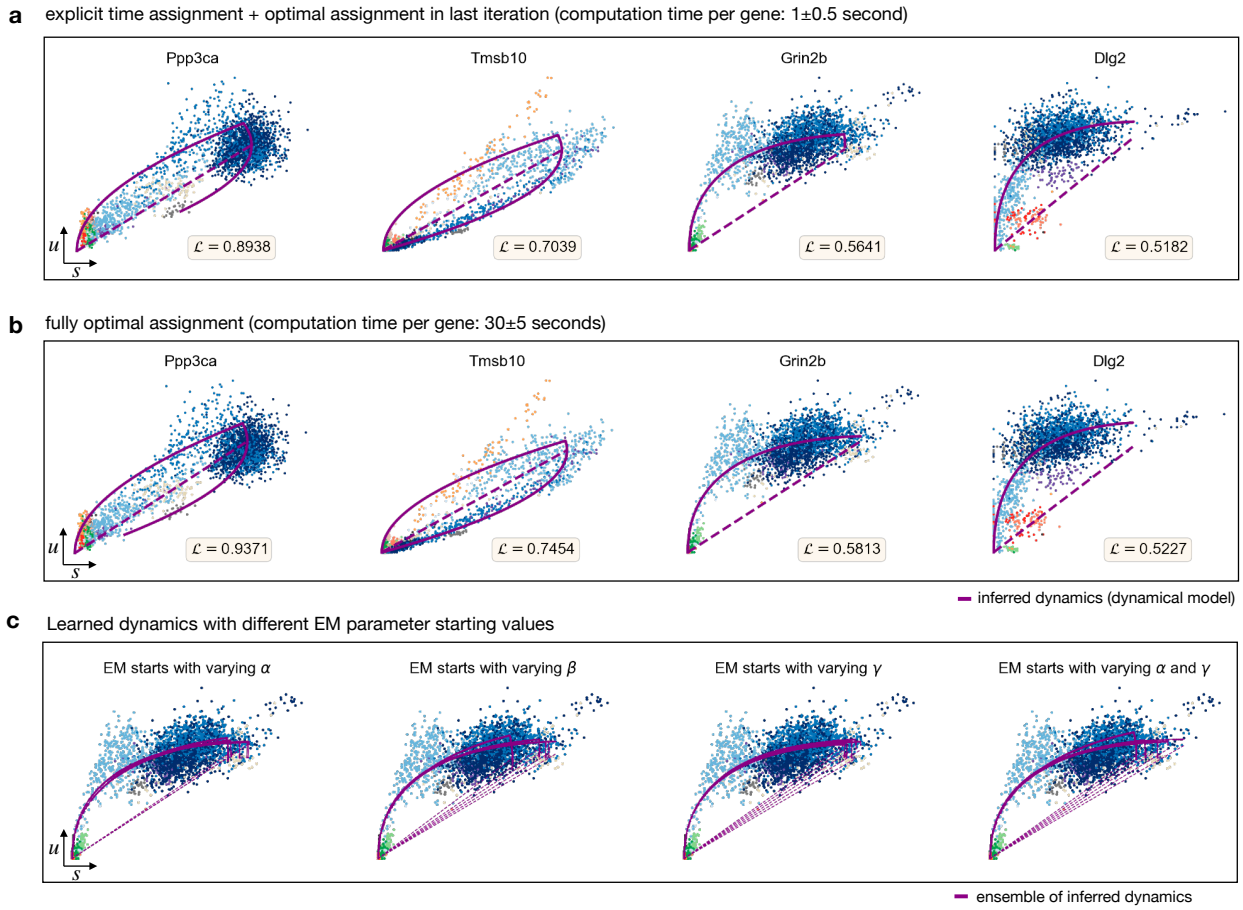
Supplementary Figure 4 | Identified driver genes with pronounced dynamic behavior.
a. Phase portraits of top likelihood-ranked genes for hippocampal neurogenesis. Most of the these have been reported to play a crucial role in neurogenesis (e.g *Grin2b, Map1b, Dlg2*)[24,25]. *Ppp3ca* has the highest likelihood and largest contribution to the velocity vector field. Its vital role has been demonstrated by associating a reduction of *Ppp3ca* activity with tauopathy in Alzheimer's disease[27].
b. Phase portraits of top likelihood-ranked genes for pancreatic endocrinogenesis. These genes have been associated with hormone processing (e.g. *Cpe and Pcsk2*) and secretion (Abcc8)[33–35]. They play a crucial role for insulin regulation and normal β-cell function. Pro-insulin to insulin in β-cells as well as Glucagon in α-cells is processed by the endocrine signature genes *Pcsk2* and *Cpe*. *Top2a* is a cell cycle gene, *Gnao1* and the transcriptional co-repressor *Cbfa2t2* are involved endocrine development[52,53].



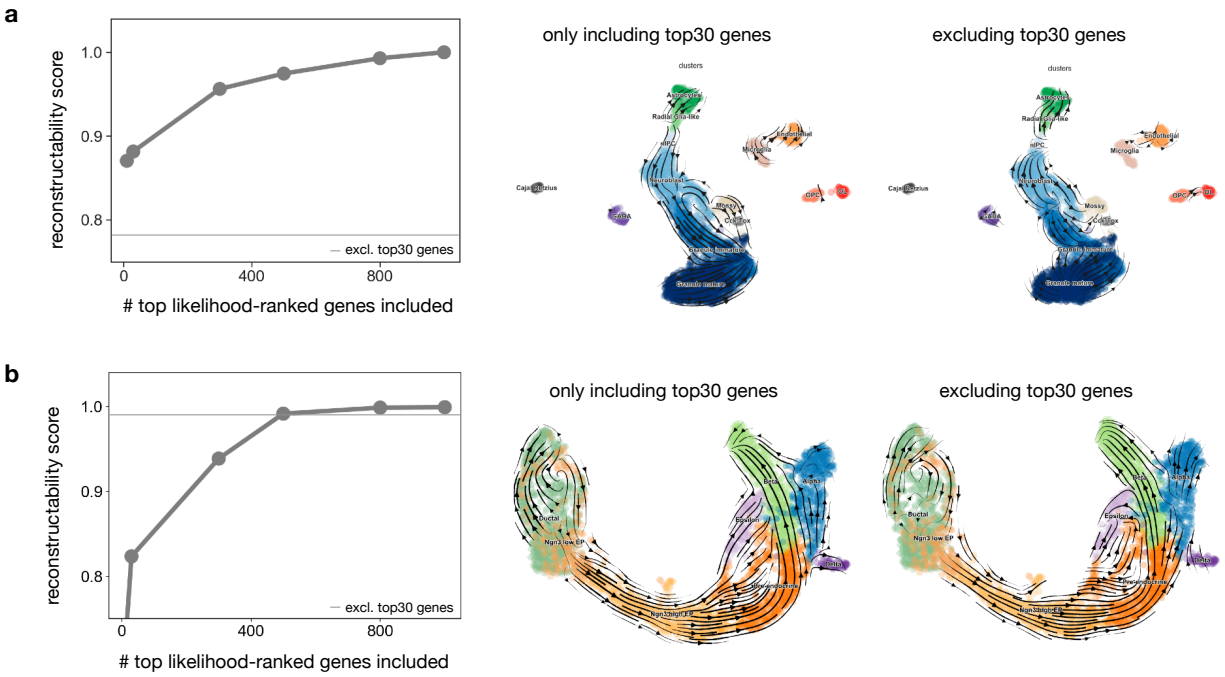
Supplementary Figure 5 | Convergence illustrated by log-likelihood contours.
a. The objective log-likelihood contours for *Tmsb10* from dentate gyrus neurogenesis at varying rates of transcription α, splicing β and degradation γ (optimal value in black). The dynamical model converges to an optimum (purple circle). On the right, the convexity of the landscape around the optimum and its convergence is shown to hold for top likelihood-ranked genes in hippocampal neurogenesis, illustratively for transcription and splicing rates.
b. Objective log-likelihood contours for *Cpe* and top likelihood-ranked genes in pancreatic endocrinogenesis demonstrate the convergence of the dynamical model to an optimal parameter set of kinetic rates.



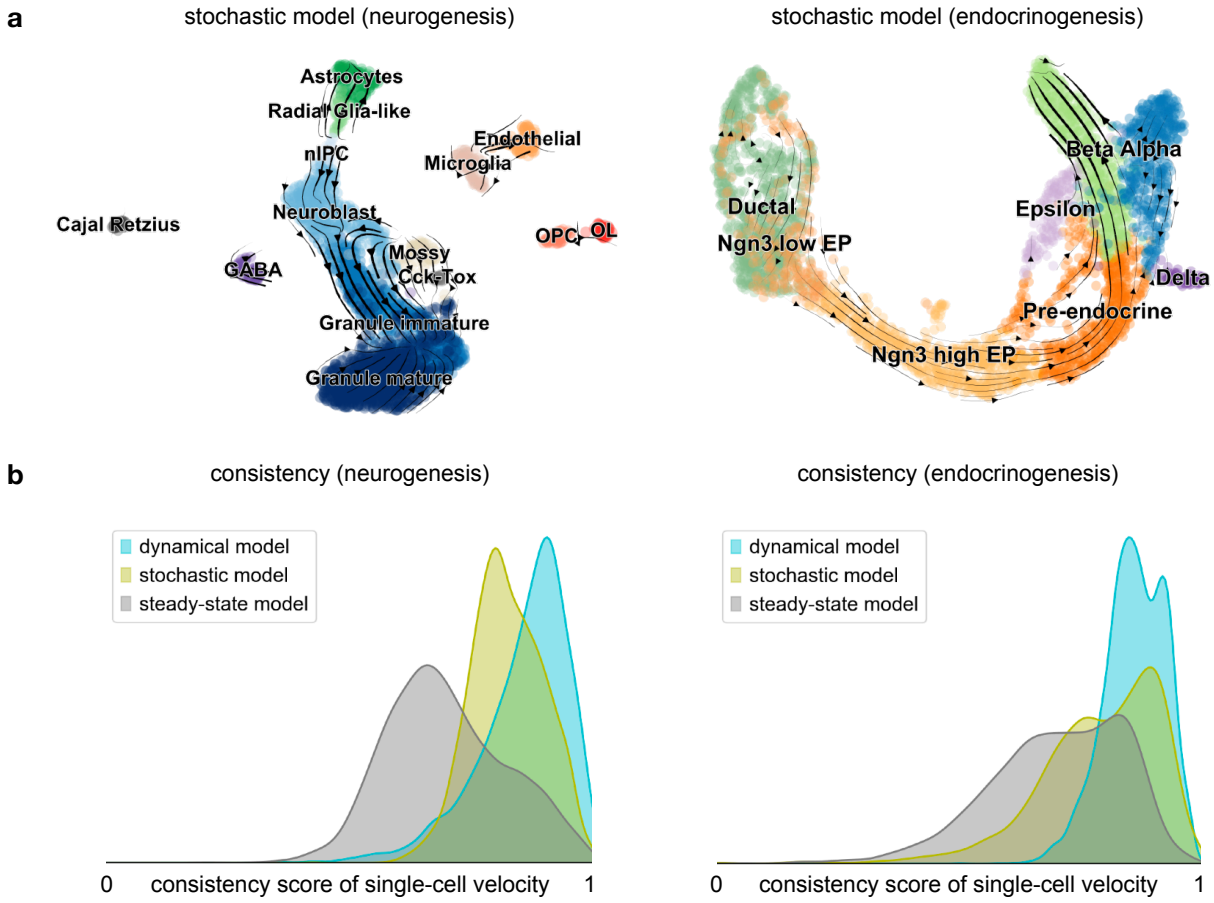
Supplementary Figure 6 | Experimental time as validation and velocity-inferred connectivities.
a. Experimental timepoints (postnatal day 12 and 35) affirm the inferred directionality obtained from RNA velocity, in particular the granule, astrocyte and GABA lineage.
b. Connectivity graph (left) and velocity graph (right). The connectivities between data points are obtained from euclidean distances in PCA space and show multiple connections between Mossy, Neuroblast and Granule cells, which explains why the UMAP-based embedding positions Mossy cells next to the Neuroblast population. Contrarily and in concordance with biological knowledge, the cell-to-cell transition probabilities obtained from velocities show that Mossy cells are only connected within the population and do not have any connections to other cell types such as Neuroblasts.



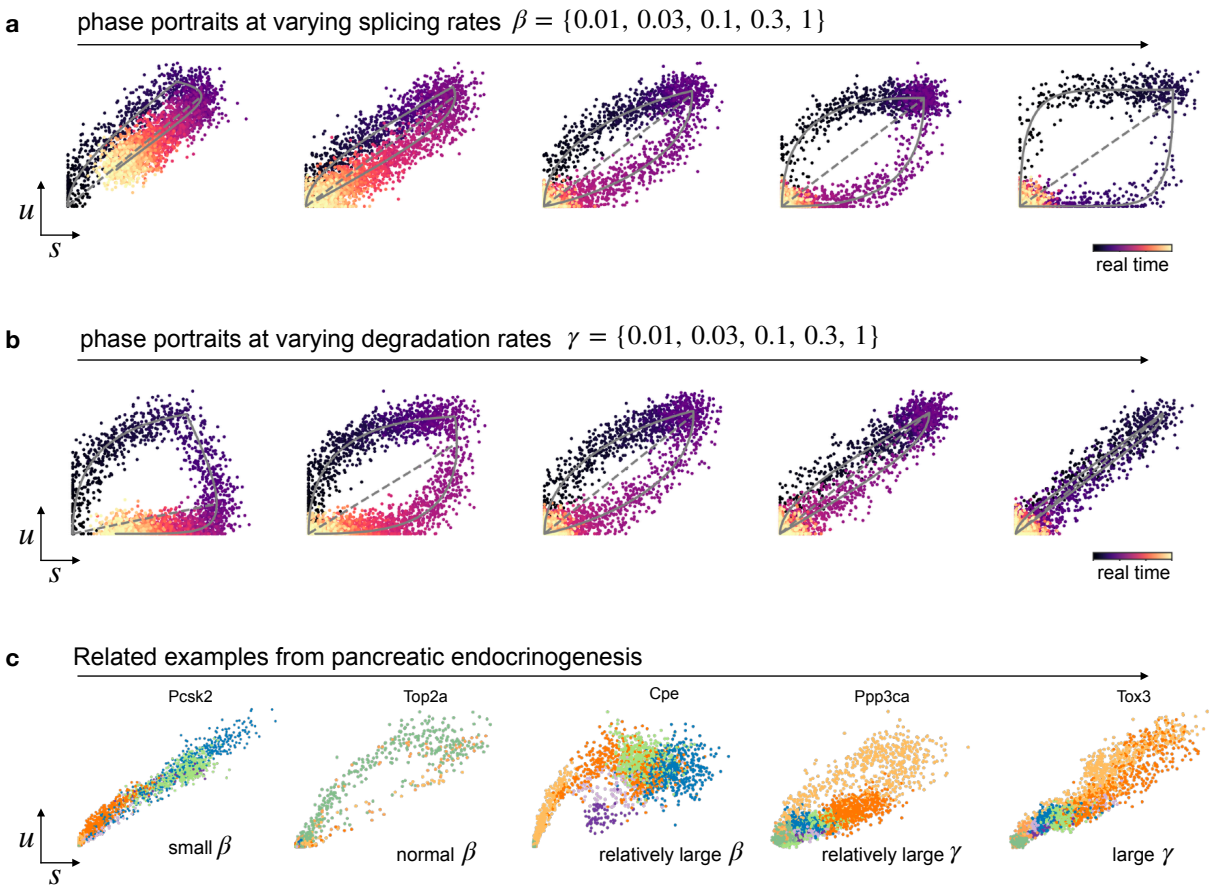
Supplementary Figure 7 | Impact of explicit time assignments and different EM starting values.
a. The optimization procedure starts with fitting the phase trajectory with explicit time assignments, which serves as an approximation to the optimal assignments. Optimal assignments are done only in the very last iteration. Fitting takes about one second per gene.
b. Optionally, time assignments can also be conducted optimally likelihood-based, which yields marginal improvements in the resulting likelihoods of up to 4%. It fundamentally yields the same results and comes at the computational expense of about 30 seconds per gene.
c. The ensemble of trajectories obtained with multiple initialized EM starting values of kinetic rates demonstrates the robustness of the model as differences in the resulting trajectories are marginal.



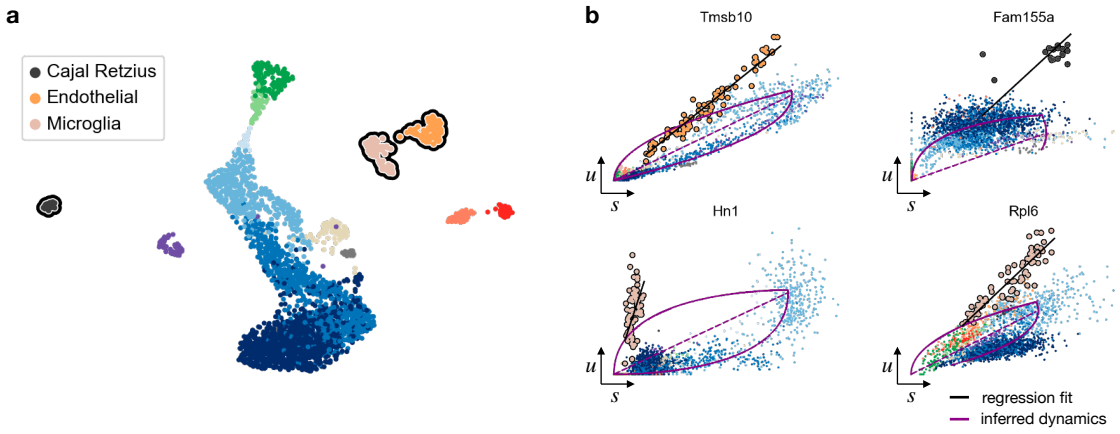
Supplementary Figure 8 | Computational validation of top likelihood-ranked genes.
a. To validate the contribution of the top likelihood-ranked genes to the overall inferred dynamics, we defined a reconstructability score. It is given by the median correlation of the velocity graph that results from (i) including all genes and (ii) only including a subset of likelihood-ranked genes. For dentate gyrus, the top 30 genes almost fully explain the inferred dynamics. If we include all other genes and only exclude the top 30 genes, the dynamics cannot be fully reconstructed (score displayed by horizontal line) which, for instance, yields blackflows in the nIPCs. That confirms that the systematically identified genes are transcriptomically highly relevant, making them candidates for important drivers of the underlying developmental process.
b. In the pancreatic endocrinogenesis dataset, the reconstructability score similarly confirms that the top likelihood-ranked genes explain most of the dynamics. However, as opposed to the dentate gyrus example, a much larger set of genes contributes to explaining the dynamics. That is indicated by the reconstructability curve that starts lower. Also the perfect reconstruction when excluding the top 30 genes (as indicated by the horizontal line and the velocity streamplot) suggest that many other genes have sufficiently evident kinetics to describe the underlying dynamics.



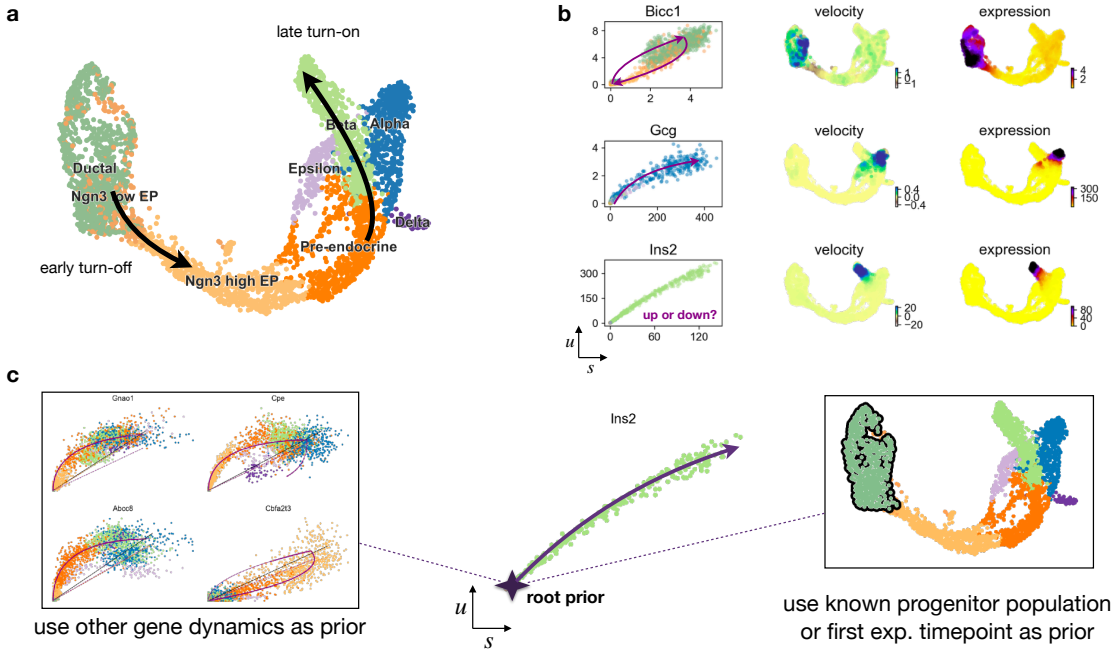
Supplementary Figure 9 | Consistency of single-cell velocities inferred by the stochastic model. **a.** The stochastic model captures the dynamics inferred by the dynamical model to a large extent for dentate gyrus neurogenesis (left) and pancreatic endocrinogenesis (right). It shows great resemblance in dentate gyrus, indicates cycling endocrine progenitors and resolves the endocrine lineage. However, like the steady-state model, the stochastic model also induces a backflow in the α-cells.
b. The consistency score is defined for each cell as the correlation of its velocity with the velocities of neighboring cells. As expected, both in dentate gyrus and pancreas, the stochastic model yields consistencies of single-cell velocities higher than for the steady-state model and lower than for the dynamical model.



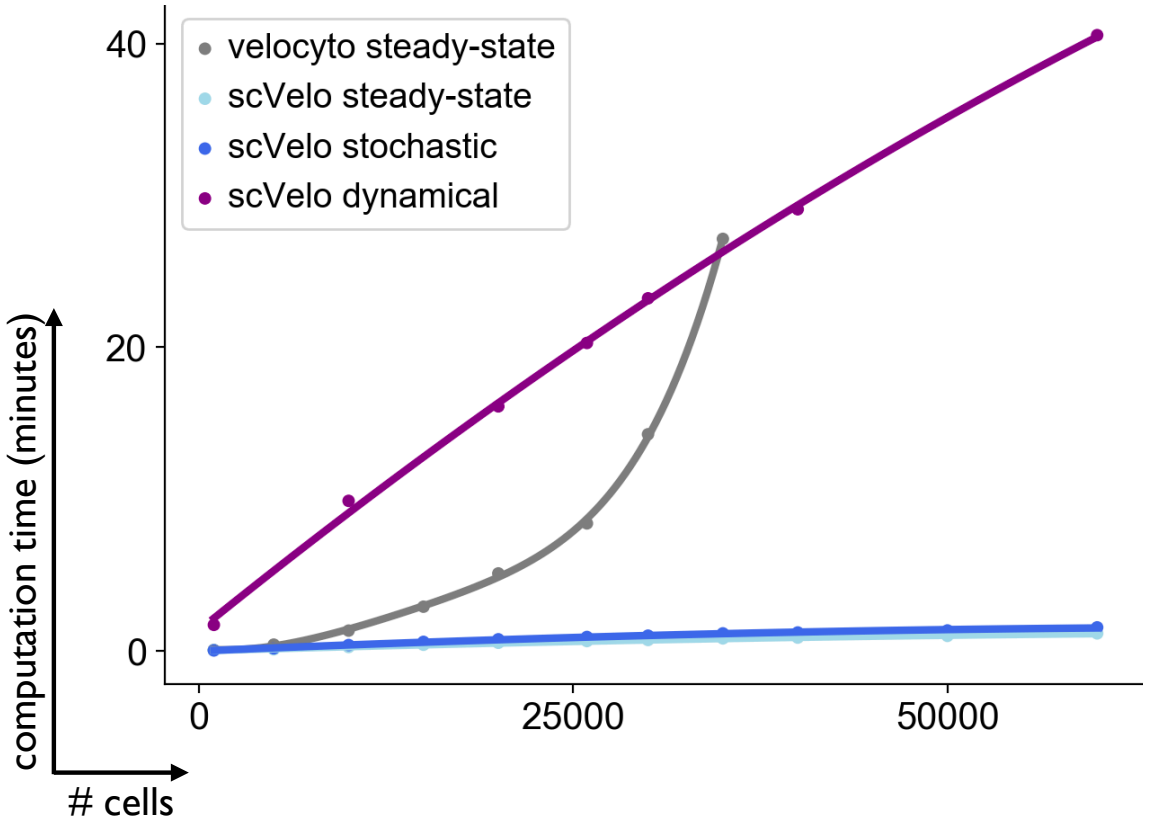
Supplementary Figure 10 | Splicing kinetics at varying reaction rates.
a. Gene unspliced/spliced phase portraits for simulated kinetics at varying splicing rates show that the curvature increases and that the kinetics becomes more pronounced with increasing splicing rate.
b. The curvature increases with decreasing degradation rates.
c. Phase portraits of selected top likelihood-ranked genes in pancreatic endocrinogenesis show how the genes from real data mirror the simulated kinetics, which can be explained by the inferred parameters of splicing and degradation rates.



Supplementary Figure 11 | Identifying kinetic regimes with a differential kinetic test.
a. UMAP-based embedding of dentate gyrus neurogenesis. Cajal Retzius (CR) cells, Endothelials and Microglias are detected to fall into kinetic regimes that are different from the main granule lineage.
b. Clusters and lineages that display different kinetic regimes are detected with a likelihood-ratio test. Cajal Retzius (CR) cells, Endothelials and Microglias display kinetic behavior in phase diagrams that cannot be sufficiently explained by a single model for the overall dynamics. While CR cells are terminal, the other cell types seem to form their own sub-lineages as displayed in phase space.



Supplementary Figure 12 | Accounting for insufficiently observed kinetics with a root prior.
a. A gene's activity may be observed in a small time window towards the very ends of a process, thereby disclosing only a small fraction of the kinetics.
b. If a gene is up-regulated only at the very end (late turn-on, e.g. *Ins2*) or down-regulated at the very beginning (early turn-off, e.g. *Bicc1*) of a developmental process, it manifests as a straight line rather than a curve in the unspliced-to-spliced phase diagram. The lack of observed curvature challenges the dynamical model to determine whether an up- or down-regulation should be fit, as illustrated by *Ins2*.
c. We addressed 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 Figure 13 | Runtime comparison.
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 comparison, we run the pipelines of velocyto's steady-state model and all scVelo models on an Intel Core i7 CPU with 3.7GHz and 64GB RAM. We used 1k highly variable genes and varied the cell numbers from 1k to 60k cells. We further run the pipelines for 300k cells to test for scalability. The cells are sampled from the pancreas development dataset containing 25,919 transcriptome profiles. In less than a minute, scVelo manages to run the full velocity pipeline for the steady-state as well as stochastic model on 35k cells. The full splicing dynamics including kinetic rate parameters, latent time and velocities, is inferred in a longer but practicable runtime of 20 minutes. As it scales linearly with the number of cells and genes, its runtime is exceeded by velocyto's quadratic runtime on large cell numbers of 35k and higher. For large cell numbers, also memory efficiency becomes a critical aspect. Velocyto cannot process more than 40k cells as it runs out of memory, while scVelo scales to more than 300k cells.