scDVF: Single-cell Transcriptomic Deep Velocity Field Learning with Neural Ordinary Differential Equations

Authors

Zhanlin Chen[1], William C. King[2], Mark Gerstein[1,3,4,*], Jing Zhang[5,*]

Affiliations

1. Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA.
2. Healthcare and Life Sciences, Microsoft, Redmond, WA 98052, USA.
3. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA.
4. Department of Computer Science, Yale University, New Haven, CT 06520, USA.
5. Department of Computer Science, Yale University, New Haven, CT 06520, USA.

*Corresponding authors. Email: pi@gersteinlab.org, jing.zhang@uci.edu

Abstract

Recent advances in single-cell RNA sequencing technology provided unprecedented opportunities to simultaneously measure the gene expression profile and the transcriptional velocity of individual cells, enabling us to sample gene regulatory network dynamics along developmental trajectories. However, traditional methods have been challenged in offering a fundamental and quantitative explanation of the dynamics as differential equations due to the high dimensionality, sparsity, and complex gene interactions. Here, we present scDVF, a neural-network-based ordinary differential equation that can learn to model non-linear, high-dimensional single-cell transcriptome dynamics and describe gene expression changes of individual cells across time. We applied scDVF on multiple published datasets from different technical platforms and demonstrate its utility to 1) formulate transcriptome dynamics of different timescales; 2) measure the instability of individual cell states; and 3) identify developmental driver genes upstream of the signaling cascade. Benchmarking with state-of-the-art methods shows that scDVF can improve velocity field representation accuracy by at least 50% in out-of-sample cells. Further, our perturbation studies revealed that single-cell dynamical systems may exhibit properties similar to chaotic systems. In summary, scDVF allows for the data-driven discovery of differential equations that delineate single-cell transcriptome dynamics.

Teaser

Using neural networks to derive the ordinary differential equations behind single-cell transcriptome dynamics.

MAIN TEXT

Introduction

Single-cell RNA-sequencing (scRNA-seq) captures a transcriptomic snapshot of a dynamic biological process. However, many current analysis methods view scRNA-seq as a static dataset. For example, Monocle constructs minimum spanning trees in the cellular manifold as bifurcation trajectories (*1*). Palantir uses Markov transition matrices to model neighboring cell transitions (*2*). More generally, diffusion pseudotime simulates diffusion to create pseudo-temporal ordering of cells in the data manifold (*3*). Although these computational methods have been effective in highlighting the dynamics behind single-cell transcriptomes, a fundamental question remains: can we derive quantitative equations that

accurately explain the gene expression dynamics of transitioning single cells? Discovering these equations as a function of time could answer questions about the cell fates and the driving forces behind developmental trajectories.

Recovering the dynamics from sparse and noisy scRNA-seq data is a difficult task because the cells are destroyed during data collection. With the development of RNA velocity, we can compute the time derivative of the expression state using the ratio of unspliced versus spliced transcripts (*4*). However, RNA velocity only predicts the future state of cells on the timescale of hours. We reasoned that it might be possible to extrapolate farther into the future by piecing together information from cells at different developmental times. Nevertheless, it is challenging to explicitly derive differential equations that model all gene interactions. Further, evaluating the generalizability of differential equations is still an open question. Previous approaches have relied on time-resolved scRNA-seq and linear ordinary differential equations (ODEs) to model the dynamics of regulatory networks (*5*, *6*). However, linear systems may fail to capture the non-linearity of single-cell dynamics. For example, heterodimeric cooperation in transcription factors is a known non-linear relationship, where two different transcription factors must bind to activate another gene (*7*). Moreover, single-cell dynamical systems have a high degrees-of-freedom due to the high dimensionality of the data, which could lead to errors in any dimension (*8*).

Inspired by recent developments in neural ODEs and data-driven dynamical systems (*9*, *10*), we present a computational framework called scDVF that learns to formulate the dynamics underlying scRNA-seq experiments by modeling the gene expression changes of single cells across time. With a deep-learning architecture, our approach can model non-linear, high-dimensional gene interactions in single-cell dynamical systems. Further, we can perform *in silico* studies to explore the behavior of biological processes over time. In this regard, scDVF differs substantially from most single-cell methods, in that the objective of our framework is to derive neural-network-based differential equations describing single-cell gene expression dynamics. To illustrate the robustness and general validity of our approach, we performed analyses on developmental mouse pancreatic endocrinogenesis, dentate gyrus, and developing neocortex, representing scRNA-seq experiments from different tissues, technical platforms, and developmental time scales (*11*, *12*). With two additional data sources (mouse gastrulation, developing human forebrain), we demonstrate the ability for scDVF to deconvolve gene co-expression networks and benchmarked our method against the state-of-the-art vector-field learning approach *SparseVCF* (*4*, *13*, *14*).

Results

Neural ODEs for Modeling Single-cell Transcriptome Dynamics

In a gene regulatory network, the expression of certain genes can increase or decrease the expression of other genes. In a broader biological context, a cell transitioning along its developmental trajectory can signal a cascade of gene expression changes. These gene-to-gene interactions can be formulated as a function of time using differential equations. More specifically, each cell represents an instance of the dynamics sampled from the single-cell dynamical system. If the gene expression state of a cell is the vector $\vec{x}$, then the increase or decrease in the gene expression with respect to time is the RNA velocity vector $\frac{\partial \vec{x}}{\partial t}$. Rather than deriving a system of linear ODEs $\frac{\partial \vec{x}}{\partial t} = \boldsymbol{A}\vec{x}$ with matrix $\boldsymbol{A}$, we train a variational autoencoder (VAE) $\boldsymbol{\alpha}(\vec{x})$ to learn the mapping from the gene expression state $\vec{x}$ to the RNA

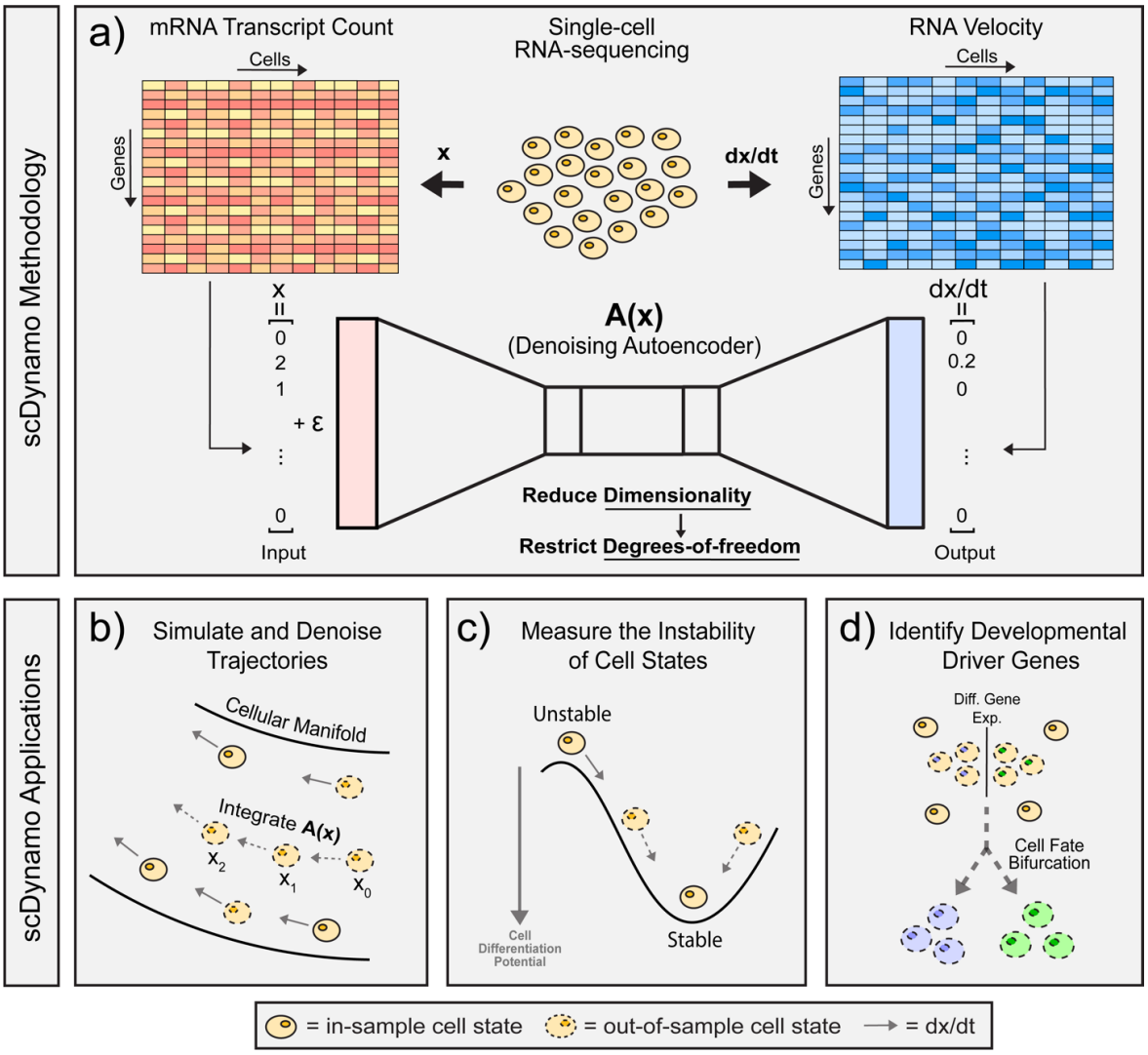

Fig. 1. Schematic for scDVF. a) Gene expression profiles and the corresponding transcriptional velocities can be derived from scRNA-seq data. After learning the mapping between gene expression and RNA velocity, the VAE is a neural differential equation that encapsulates the transcriptome dynamics. **b)** Given an initial condition and time, our framework can solve for the future gene expression state by integrating the VAE with any black-box ODE solver. **c)** Our approach can simulate trajectories to evaluate the instability of cell states in a dynamical system. **d)** scDVF can perform in silico perturbation studies to identify the developmental driver genes that determine the fate of cell bifurcations.

velocity $\frac{\partial \vec{x}}{\partial t}$ using data from each cell (Equation 1, Fig. 1a). Therefore, this VAE is a non-linear ODE and encapsulates the gene expression dynamics of individual cells in scRNA-seq. Then, given some initial gene expression state close to the data, we can numerically compute the future (or past) gene expression states with any black-box ODE solver. For example, given gene expression state $\overrightarrow{x_0}$ at time $t = 0$, we can use the Euler's method to find the gene expression state at $\overrightarrow{x_1}$, and iteratively for $\overrightarrow{x_2}, \ldots, \overrightarrow{x_n}$ (Equation 2, 3).

$$\frac{\partial \overrightarrow{x_t}}{\partial t} = \boldsymbol{\alpha}(\overrightarrow{x_t}) \qquad \text{(Equation 1)}$$

$$\overrightarrow{x_{t+1}} = \overrightarrow{x_t} + \frac{\partial \overrightarrow{x_t}}{\partial t} \qquad \text{(Equation 2)}$$

$$= \overrightarrow{x_t} + \boldsymbol{\alpha}(\overrightarrow{x_t}) \quad \text{(Equation 3)}$$

By sequentially computing the next gene expression state, scDVF can outline the developmental trajectory of single cells through time. Further, with different initial conditions $\overrightarrow{x_0}$, our framework can derive detailed insights into the future (or past) of different cell states. Here, we explored three applications of scDVF. First, we simulated and denoised developmental trajectories by extrapolating the dynamics to out-of-sample cells (Fig. 1b). Second, we evaluated the instability of cell states by tracking gene expression changes along simulated trajectories (Fig. 1c). Third, we performed *in silico* perturbation studies to investigate how initial gene expression conditions impact the fate of cell bifurcations (Fig. 1d).

Deriving the Neural Equations Underlying Pancreatic Endocrinogenesis

To evaluate whether scDVF can uncover the dynamics from sparse and noisy scRNA-seq experiments, we considered a dataset of mouse pancreatic endocrinogenesis with transcriptomes profiled at E15.5 with the Chromium Single-cell 3' Library from 10x Genomics (Fig. 2a) (*13*). Here, we show that summarizing the dynamics as neural ODEs can derive new insights from the data.

First, we examined a hypothetical trajectory simulated from scDVF. When simulating hypothetical trajectories, future state predictions rely on previous out-of-sample predictions. Hence, we evaluated scDVF with an out-of-sample initial condition. After training the VAE on neocortex cell states and velocities, we generated an out-of-sample cell as the initial condition. The out-of-sample cell is simulated by adding noise to the gene expression state of an existing cell, thereby representing a cell state that did not previously exist in the data. Then, we incrementally solved for the future gene expression states of the out-of-sample cell using scDVF. The simulated developmental path shows that our predicted gene expression states moved along existing trajectories in the data manifold (Fig. 2b). In pancreatic endocrinogenesis, the out-of-sample cell started as an endocrine progenitor, developed into a pre-endocrine cell, and ultimately became a beta cell. Further, when the VAE is solved with evenly distributed time increments, the distances between intermediate states reflect the magnitude of the RNA velocity vectors. Faster rates of change in gene expression generated more separated intermediate states. Conversely, slower rates of change produced a denser collection of intermediate points along the manifold.

When the VAE represents gene expression dynamics, we can visualize the latent layer embeddings to gain insights into the low-dimensional dynamic manifold. Similar to gene expression embeddings, the chronological and hierarchical order of developmental trajectories in the latent layer are properly encoded (Fig. 2b). In pancreatic endocrinogenesis, ductal cells represent a major starting state. Alpha and beta cells represent a major terminal state. The simulated cell migrates along existing trajectories in the low-dimensional dynamic manifold.

Characterizing Cell State Instability with the Cell Criticality Index

Next, we aimed to characterize the stable and unstable fixed points of this single-cell dynamical system. By looking forward in time, we can numerically approximate the instability of single-cell states, which we call the cell criticality index (CCI). For a cell, the CCI is defined as the cumulative information change, or the cumulative Kullback–Leibler (KL)-divergence, between gene expression distributions at each time step in the developmental trajectory. In order words, cell states that undergo large changes across time

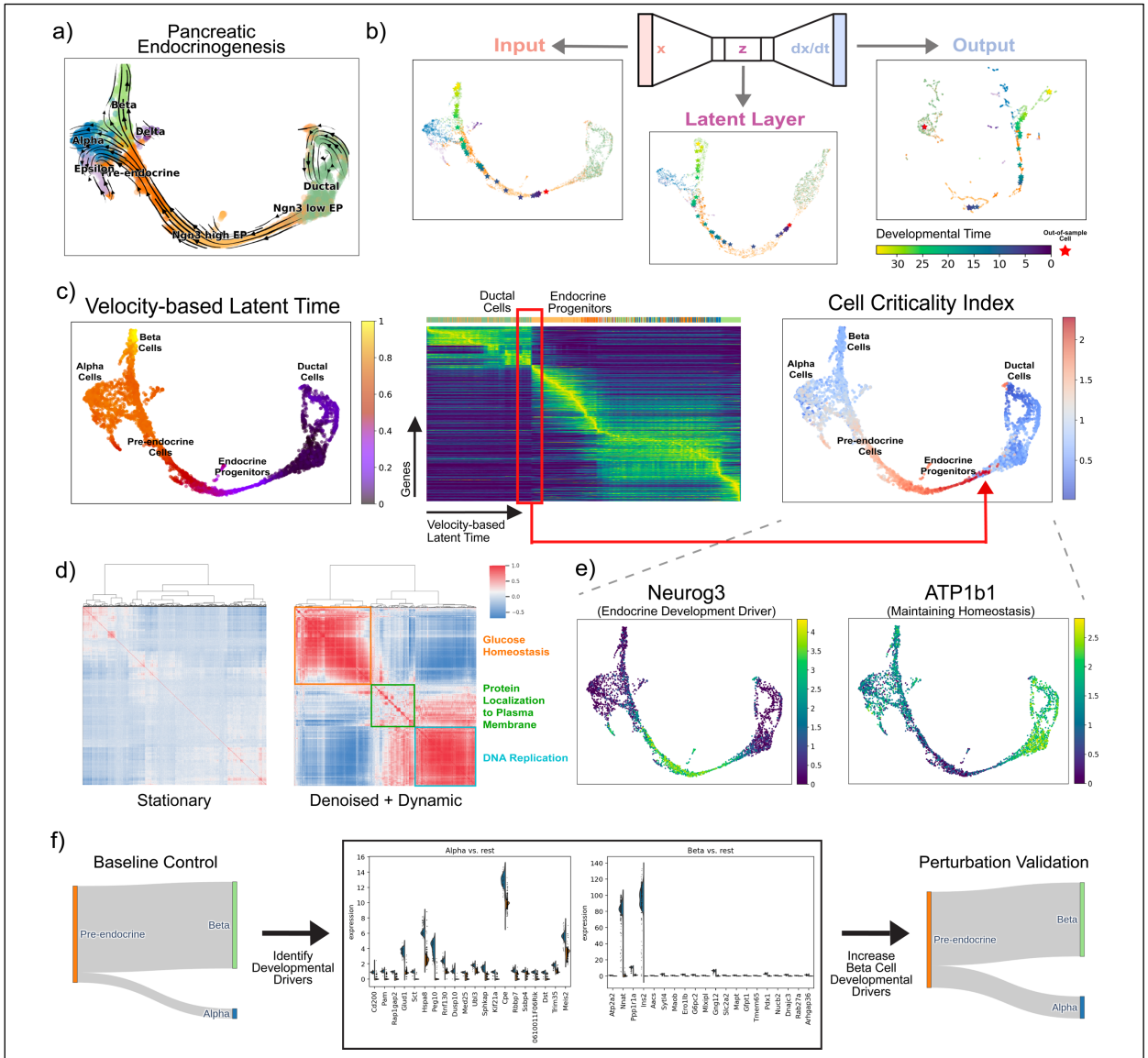

Fig. 2. Pancreatic Endocrinogenesis as a Dynamical System. a) Pancreatic endocrinogenesis phase portraits projected in a low-dimensional embedding. Here, each cell is represented by a gene expression state vector and an RNA velocity vector. **b)** Simulating the trajectory (in viridis) of an out-of-sample cell (in red) forward in developmental time, visualized in UMAP embedding of the gene expression state (x), autoencoder latent layer (z), and velocity (dx/dt). **c)** The CCI reveals unstable fixed points indicative of cell fate commitment. **d)** Simulating dynamic trajectories can effectively denoise gene co-expression networks in sparse and noisy scRNA-seq experiments, especially when compared to the static baseline. **e)** Genes that highly correlate with the CCI reveal driving forces behind endocrine progenitor dynamics. In particular, *Neurog3*, which positively correlates with the CCI, is a known pre-endocrine master regulator, and *ATP1b1*, which negatively correlates with the CCI, has an important function in maintaining homeostasis in stable and stationary cell types. These genes support the CCI as a metric for the stability of single-cell states. **f)** Differential gene expression analysis on the simulated pre-endocrine cells reveals key putative genes that correlate with the fate of transforming into an alpha versus a beta cell. Early perturbation of the top differentially expressed genes associated with an alpha cell fate resulted in a higher proportion of alpha cells from the perturbed pre-endocrine cells, suggesting a causal relationship through *in silico* studies.

will have a high CCI, whereas cell states that only go through small changes will have a low

CCI.

For each cell, we used scDVF to compute a developmental path such that the cell arrived at a steady terminal state. Then, we calculated the CCI along each path (Fig. 2c). The resulting developmental topology is similar to the classical Waddington landscape (*15*). In particular, the CCI can reveal unique topological information in the developmental landscape not directly observed in latent or pseudo time. For example, the endocrine progenitor states exhibit a higher criticality, whereas the ductal cell and differentiated endocrine cell states experience a lower criticality. When ductal cells are transforming into islet cell types, the heightened criticality in endocrine progenitors represents fate commitment or a point of no return during development. In dynamical systems, this suggests that the cell states with low criticality are located at a stable fixed point, and the cell identity would remain stable even with small gene expression perturbations. More interestingly, the endocrine progenitors are located at an unstable fixed point with properties similar to a chaotic system in which a small perturbation may result in large downstream changes. The instability of cell states can be substantiated by examining the genes that best correlate with the CCI (Fig. 2e). For example, previous experiments have shown that *Neurog3*, which positively correlates with the CCI, is a known driver for endocrine commitment, and *ATP1b1*, which negatively correlates with the CCI, is important for maintaining homeostasis in stable and stationary cell types (*16*, *17*). The expression of these genes supports the CCI as a metric for evaluating the instability of single-cell states.

Conducting *in Silico* Perturbation Studies with scDVF

Lastly, we investigated the behavior of this dynamical system with similar perturbation studies pioneered by (*18*). The goal of *in silico* perturbation studies is to computationally identify which initial gene expression conditions impact the fate of cell bifurcations. In short, we randomly sampled pre-endocrine cells ($n = 1{,}000$) as the initial conditions. By allowing these simulated pre-endocrine cells to naturally evolve according to the dynamics learned by scDVF, we observed a baseline 9:1 ratio of terminal beta versus alpha cell states. The ratio of terminal cell states indicates that the beta cell state is a stronger attractive terminal state than the alpha cell state, which corroborates with previous conclusions (*19*). Then, we performed differential gene expression between initial conditions of different fates. The results suggest that early expression perturbations in key upstream genes correlate with the fate of developmental bifurcations (Fig. 2f).

Further, we formulated a way to perform hypothesis testing and to infer causal relationships at developmental branching points (*20*). To investigate which developmental driver genes cause progenitor cells to prefer one trajectory over another, we strategically increased the expression of alpha-cell-related developmental driver genes in another set of simulated pre-endocrine cells. We hypothesized that this perturbation would lead to a larger proportion of alpha cells as terminal cell states. Indeed, we observed a statistically significant increase in the proportion of astrocytes (23%) compared with the baseline (10.3%; binomial test $p < 10^{-5}$) under the dynamics learned by scDVF. Thus, *in silico* perturbation studies can be used to efficiently and comprehensively identify developmental driver genes upstream of the signaling cascade. More interestingly, simulation results suggest that pancreatic endocrinogenesis development exhibits properties similar to chaotic systems, where small perturbations in key upstream genes determine the fate of cell bifurcations. In other words, small variations in the initial conditions of a cell may result in large downstream changes.

Exploring the Neural Equations Behind the Developing Mouse Dentate Gyrus

Further, we evaluated whether scDVF can uncover the dynamics of a dataset from a different tissue, developmental timescale, and technical platform. We considered an scRNA-seq experiment of the developing mouse dentate gyrus with transcriptomes profiled

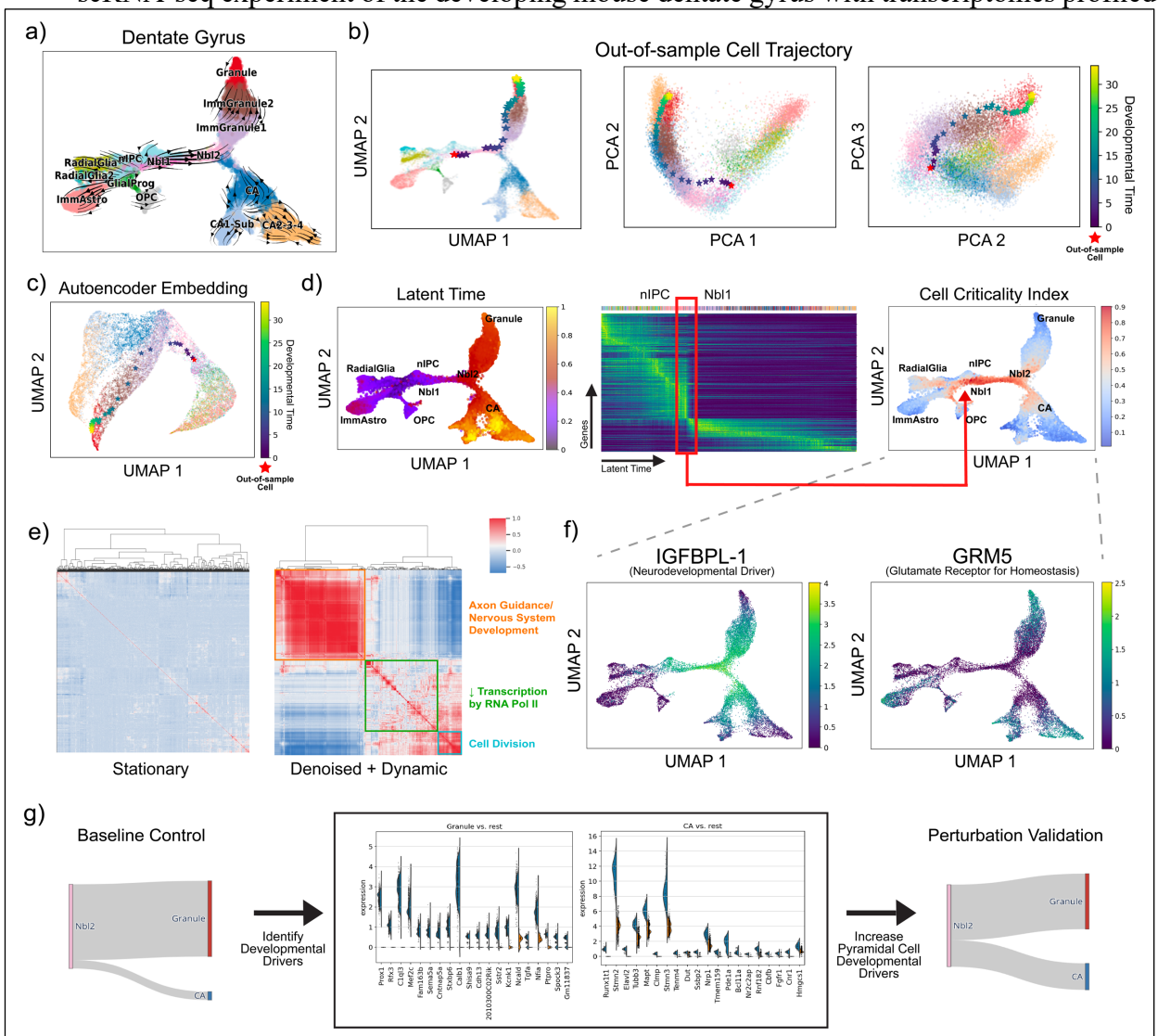

Fig. 3. Developing Mouse Dentate Gyrus as a Dynamical System. a) Dentate gyrus phase portraits projected in a low-dimensional embedding. Here, each cell is represented by a gene expression state vector and an RNA velocity vector. **b)** Simulating the trajectory (in viridis) of an out-of-sample cell (in red) forward in developmental time, visualized in UMAP and PCA embeddings. **c)** Visualizing the low-dimensional manifold of the high-degrees-of-freedom single-cell dynamical system, with the simulated out-of-sample cell trajectory (in red and viridis). **d)** The CCI reveals unstable fixed points indicative of cell fate commitment. **e)** Simulating dynamic trajectories can effectively denoise gene co-expression networks in sparse and noisy scRNA-seq experiments, especially when compared to the static baseline. **f)** Genes that highly correlate with the CCI reveal driving forces behind dentate gyrus dynamics. In particular, *IGFBPL-1*, which most positively correlates with the CCI, regulates neurodevelopment, and *GRM5*, which most negatively correlates with the CCI, encodes glutamate receptors in stable and stationary neurons. These genes further substantiate the CCI as a metric for the stability of single-cell states. **g)** Differential gene expression analysis on the simulated pre-endocrine cells reveals key putative genes that correlate with the fate of transforming into a pyramidal versus granule cells. Early perturbation of the top differentially expressed genes associated with a pyramidal cell fate resulted in a higher proportion of pyramidal cells from the perturbed Nbl2 cells, suggesting a causal relationship through *in silico* studies.

using droplet-based scRNA-seq (Fig. 3a) (*11*). After obtaining a neural network representation of the dentate gyrus dynamics, an out-of-sample cell was simulated by perturbing the gene expression state of an *Nbl2* cell. With the out-of-sample cell as the initial condition, we used scDVF to simulate an out-of-sample cell trajectory, which moved along the existing granule cell trajectory in the data (Fig. 3b). Further, the VAE embeddings properly encoded the developmental hierarchy of cell types in the low-dimensional dynamic manifold (Fig. 3c).

When examining critical cell states in the dentate gyrus, we observed an abrupt gene expression shift in the developmental manifold, which can be visualized when ordering cells in latent time (Fig. 3d). Specifically, the change in gene expression marks the transition from *nIPC* to *Nbl1* cells and suggests fate commitment during the transition. After calculating the CCI, we found that cells experiencing this abrupt change also have a high criticality, which substantiates the CCI as a metric for quantifying the instability of cell states. The robustness of the CCI as an instability measure is also highlighted by its most strongly correlated genes in the dentate gyrus. For example, *IGFBPL-1*, which most positively correlates with the CCI, drives neuron differentiation in progenitor cells, and *GRM5*, which most negatively correlates with the CCI, encodes glutamate receptors in stable and differentiated neurons (Fig. 3f) (*21*, *22*).

Lastly, we conducted *in silico* perturbation studies to determine the genetic drivers behind dentate gyrus cell fate decisions. We randomly sampled upstream *Nbl2* cells ($n = 1{,}000$) as the initial conditions and allowed the simulated *Nbl2* cells to naturally evolve according to the dynamics captured by scDVF, which resulted in either terminal granule or pyramidal cell states. Then, we performed differential expression analysis on the initial conditions (i.e., the simulated *Nbl2* cell states) of different fates (Fig. 3f). The top differentially expressed gene associated with a granule cell fate was *Prox1*. This gene has also been previous identified by RNA velocity and experimentally validated as being necessary for granule cell formation; moreover, the deletion of *Prox1* leads to the adoption of the pyramidal neuron fate (*21*). In addition, scDVF identified the top pyramidal neuron developmental driver gene as *Runx1t1*, which was recently shown to induce pyramidal neuron formation, with its deletion resulting in reduced neuron differentiation *in vitro* (*23*). As further validation, we increased the expression of pyramidal neuron developmental driver genes in simulated *Nbl2* cells and observed an elevated proportion of pyramidal neurons as terminal states (from 10% to 30%; binomial test $p < 10^{-7}$) under the dynamics captured by scDVF. In summary, *in silico* perturbation studies can be a low-cost alternative for identifying developmental driver genes. Further, the results show that scDVF is robust on scRNA-seq from different tissues, developmental timescales, and technical platforms.

Learning the Neural Equations Underlying the Developing Mouse Neocortex

Moreover, we evaluated whether scDVF can learn the dynamics underlying a more complex developmental process. We examined a scRNA-seq experiment of developing mouse neocortex with transcriptomes profiled at E14, E15, E16, and E17 (*12*). After batch correction and velocity-field learning, an out-of-sample intermediate progenitor (IP) was generated as the initial condition. The simulated developmental trajectory suggests that the hypothetical IP transitioned into a migrating neuron (MN), then subsequently became a CorticoFugal Neuron (CFN). The distance between each time step was larger when the simulated cell was an IP compared to a MN, indicating the IPs are transitioning at a faster rate compared to the MNs (Fig. 4a). The latent layer embeddings also properly encoded the progression from neural progenitors to differentiated neurons (Fig. 4b).

In addition to the latent time ordering of cells, we also observed high criticality in IPs and MNs, but low criticality in apical progenitors (AP) with the CCI. This suggests that APs play a more important role in self-renewal. Further, it indicates that IPs are pre-programmed for neuronal differentiation through a broader shift in gene expression patterns. Additionally, as MNs branch into IntraTelencephalic Neurons (ITN) and CFNs, scDVF can effectively disentangle the underlying bifurcation dynamics. When we allowed a set of simulated MNs ($n = 1000$) to evolve according to the dynamics learned by scDVF, it resulted in a baseline 3:7 ratio of ITNs vs. CFNs. After perturbing the initial differentially expressed genes associated with an ITN-fate in another set of simulated MNs ($n = 1000$), we observed a 10% increase in ITN proportion compared to baseline. These observations indicate that scDVF can accurately formulate an overarching ordinary differential equation of mouse neocortex development spanning several embryonic days. Further, the results suggest that our framework can identify driver genes that play a crucial role in the developmental dynamics with a continuous velocity field learned by scDVF.

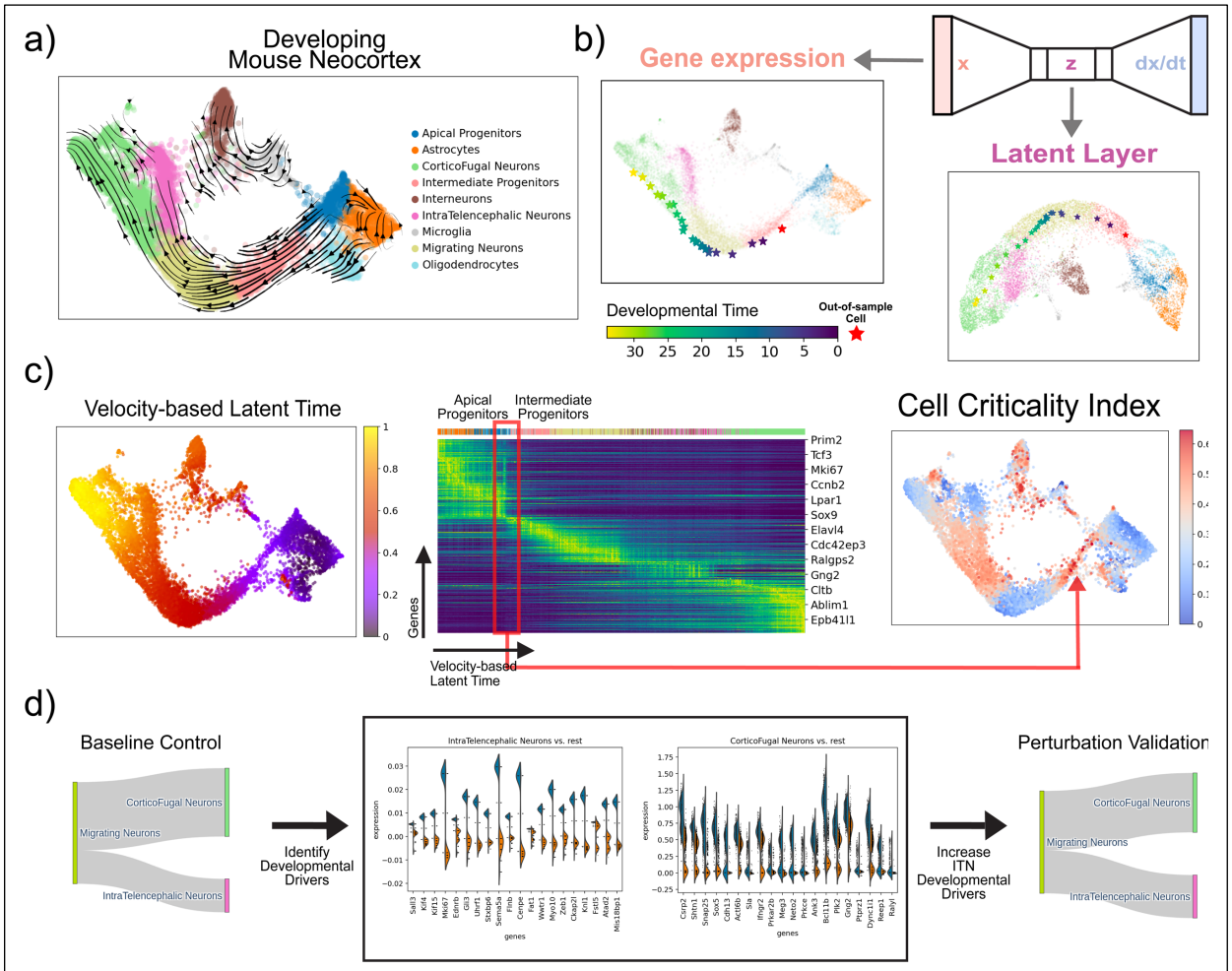

Fig. 4. Developing Mouse Neocortex Over Multiple Embryonic Days. a) Mouse neocortex phase portraits projected in a low-dimensional embedding. **b)** Simulating the trajectory (in viridis) of an out-of-sample cell (in red) forward in developmental time, visualized in gene expression embeddings and VAE latent layer embeddings. **c)** The CCI reveals unstable fixed points indicative of cell fate commitment. **d)** Differential gene expression analysis on the simulated MNs reveals key putative genes that correlate with the fate of transforming into a ITN versus CFN. Early perturbation of the top differentially expressed genes associated with a pyramidal cell fate resulted in a 10% higher proportion of ITNs from the perturbed MNs.

Comparing scDVF with Existing Methods

RNA velocity predicts gene expression change of individual cells in the timescale of hours. Previous simulations in this study used hypothetical progenitor cells as the initial conditions and computed trajectories into the future resulting in differentiated cells as terminal states. Conversely, we can use differentiated (or terminal) cells as the initial conditions and rewind time with scDVF. Then, the retrograde developmental trajectory represents the gene dynamics that would have resulted in the terminal cell types.

Due to sparse and noisy measurements, it is often challenging to detect strong correlation between genes in scRNA-seq, thereby making it difficult to find coherent functional modules in gene co-expression networks (*24–26*). However, denoising VAEs in scDVF can reduce the variability along a developmental trajectory due to the sparsity and noise associated with scRNA-seq (Fig. 5a). We hypothesize that cells in denoised trajectories simulated from scDVF (with a representative initial condition) could amplify the correlations within functional gene modules (Fig. 5b). Indeed, the gene co-expression network of cells in retrograde trajectories has more significant gene correlations compared to co-expression networks from static cells (Fig. 2e, Fig. 3e). Further, we biclustered the co-expression matrix into gene clusters. By benchmarking our approach on four datasets, we demonstrate that the gene clusters discovered from our method are more coherent by comparing the gene ontology (GO) enrichments. The benchmarks show that functional gene modules found from denoised and dynamic cells in retrograde trajectories have at least two orders of magnitude higher enrichment for cell-type-specific GO terms compared to static cell clusters (Fig. 5c). Therefore, the retrograde trajectories computed by scDVF can effectively disentangle trajectory-specific gene regulatory networks and serve as a computational solution for boosting signal-to-noise ratios in single-cell gene co-expression networks.

Further, scDVF qualitatively differs from existing ODE-based regulatory networks (*27*). First, explicitly deriving differential equations for biological processes is only feasible for examining small-scale systems (*28–31*). In contrast, scDVF can capture high-dimensional interactions and can scale to a large number of variables. Second, scDVF uses a neural network to learn potentially non-linear gene interactions, which is more suitable for modeling complex biological processes compared to linear ODEs and other kernel-based sparse approximation methods (*32*, *33*). In particular, we compared scDVF with state-of-the-art vector field learning approach SparseVFC (*6*). Benchmarking results show that our method has at least 50% reduction in out-of-sample velocity prediction loss across all datasets, indicating that scDVF can learn a more accurate representation of the velocity vector fields and compute future cell states with better numerical precision (Fig. 5d). Lastly, many previous ODE-based methods used pseudo-time as a substitute for time. In comparison, scDVF uses RNA velocity, which reflects developmental time (*5*).

Discussion

Although many effective tools have been developed to illuminate the dynamics of single-cell data, existing methods have mostly viewed single-cell datasets as a static manifold (e.g., minimum spanning trees, Markov matrices, diffusion etc.). In reality, many underlying biological processes captured by single-cell sequencing are dynamical systems, where individual cells are transitioning from one state to another. Hence, deriving accurate differential equations that quantify gene expression dynamics of single cells can answer many questions about the cell fates and the genetic drivers behind developmental trajectories.

Explicitly deriving differential equations for all gene interactions is a challenging task. Therefore, we tackled the problem with a data-driven approach. We considered each cell in

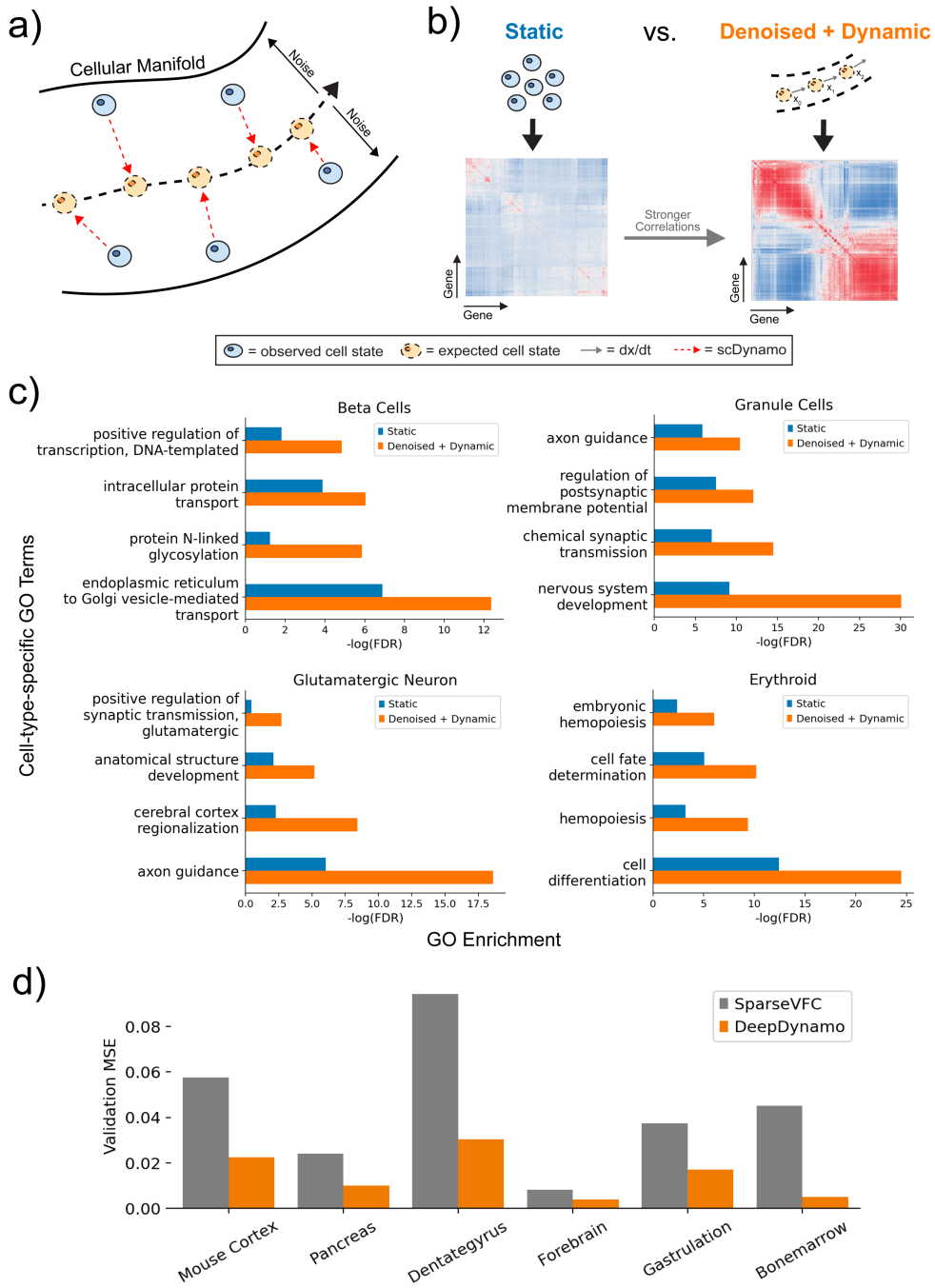

Fig. 5. Disentangling trajectory-specific gene co-expression networks. a) Schematic for reducing variability along a developmental trajectory due to sparsity and noise in scRNA-seq experiments with denoising VAEs in scDVF. **b)** We used a representative initial condition (e.g., the median expression profile for a cluster of cells) to simulate denoised cell trajectories according to the dynamics learned by scDVF. Compared to static cell clusters, the dynamic cells in denoised trajectories have stronger correlation between genes, which leads to better coherence between functional gene modules. **c)** GO enrichment on cell-type-specific terms from the most significant functional gene module. Our method improves upon existing gene co-expression network approaches on cell-type-specific GO term enrichment by at least two orders of magnitude. **d)** Benchmarking with state-of-the-art vector field learning method SparseVFC shows that scDVF can improve out-of-sample velocity vector prediction accuracy by at least 50%.

scRNA-seq as an instance sampled from a dynamic system, composed of a state vector (gene expression, $\vec{x}$) and a velocity vector (RNA velocity, $\frac{\partial \vec{x}}{\partial t}$). Then, we trained a neural network $\boldsymbol{\alpha}(\vec{x})$ to learn potentially non-linear mappings from the state $\vec{x}$ to the velocity $\frac{\partial \vec{x}}{\partial t}$ of each cell. With a trained VAE $A(\vec{x})$ that takes part in the differential equation $\frac{\partial \vec{x}}{\partial t} = \boldsymbol{\alpha}(\vec{x})$, we can integrate the VAE with any black box ODE solver to compute the future (or past) gene expression states.

Overall, our scDVF framework learns a continuous vector-field representation of single-cell velocities and allows hypothetical cells to evolve according to the dynamics learned from existing cells in the data. Using the ability to simulate future gene expression trajectories, we devised a metric to quantify the instability of individual cells called the CCI. Through perturbing cell states with high criticality, *in silico* gene perturbation studies can computationally identify key upstream driver genes that determine the fate of cell bifurcations. Lastly, by rewinding the developmental time of differentiated cells, retrograde trajectories can deconvolute trajectory-specific gene co-expression networks and discover more coherent cell-type-specific gene modules.

Previous approaches utilize pseudo-time to construct a temporal-ordering of cells and pluripotency metrics to measure the differentiation potential of a cell, similar to quantifying the "potential energy" of Waddington landscapes. However, these "potential energy" metrics are limited in describing dynamical systems. Theoretically, the potential energy is converted into conservative forces, where the total work done by a cell becomes independent of the developmental path taken. In order to more accurately capture the expression changes along specific developmental trajectories, we designed a new metric called the CCI, analogous to the "kinetic energy" of Waddington landscapes. In our analysis, we demonstrated that this metric could highlight fixed points in single-cell dynamical systems. Moreover, previous studies have formulated cell fate decisions as high-dimensional critical state transitions (*34*, *35*). Therefore, we further bring awareness to the dynamical perspective of single-cell data and advocate for new metrics that quantify the kinetics of single-cell experiments.

More interestingly, single-cell processes have long been hypothesized to exhibit properties similar to chaotic systems (*8*, *36*, *37*). By recovering single-cell gene expression dynamics with scDVF, we observed chaotic behaviors in the *in silico* gene perturbation studies, where a small change in the initial gene expression state may result in a large difference in the future states, also known as the butterfly effect. Specifically, small perturbations in developmental driver genes of progenitor cells can alter the cell fate at developmental branching points both *in vitro* and *in silico*. If single-cell dynamics exhibit chaotic properties, under the right biological conditions, the chaos can spontaneously evolve into lockstep patterns according to the Kuramoto model of synchronization (*38*). Hence, synchronization models could be a possible explanation for emergent tissue-level behaviors from single cells. These effects could be explored by incorporating the gene interaction dynamics between cells. Currently, scDVF only models the gene expression dynamics within a single cell. A future direction could expand the state space of scDVF and incorporate gene interactions between spatially neighboring cells with spatial transcriptomics (*39*). Another future direction includes incorporating concurrently resolved protein and chromatin accessibilities and their velocities into the dynamical model as a multi-modal representation of the cell state (*40*, *41*).

Materials and Methods

Data Collection and Preprocessing

scRNA-seq data (pancreatic endocrinogenesis, dentate gyrus, mouse gastrulation, and human forebrain) were downloaded. After computing the gene expression count matrix. The top 3,000 variable genes and cells with a minimum of 20 transcripts were selected. Velocity genes were found using log-transformed data, and the moments were estimated using the top 30 principal components and the top 30 nearest neighbors. Dynamical velocity vectors were computed using the raw counts.

Variational Autoencoder Architecture

High-dimensional single-cell dynamical systems are difficult to model due to high degrees of freedom. For example, the number of features can sometimes be larger than the number of data points. Consequently, gene expression would only vary in a small portion of dimensions. Therefore, modeling the gene expression dynamics of a low-dimensional manifold embedded in high-dimensional data is a challenging task. Fortunately, autoencoders can reduce the dimensionality of the data by introducing an information bottleneck. Accordingly, when used to represent dynamical systems, autoencoders can restrict cell transitions to only movements along the low-dimensional manifold.

A variational autoencoder with four dense layers (size 64 as the intermediate layer and size 16 as the latent layer) was constructed using the Tensorflow and Keras packages (*42*, *43*). The VAE takes the gene expression state as input, and outputs the RNA velocity. In the VAE, the encoder layers with weights W_e and biases b_e produces the hidden layer $h(x)$, which parametrizes the location and scale of i gaussian distributions. Then, a sample from each reparametrized gaussian distribution z_i is used as input for the decoder layer with weights W_d and biases b_d. The architecture can be expressed as:

$$EncoderLayer(x) = h(x) \quad \text{(Equation 4)}$$

$$= Relu(b_e + W_e * x) \quad \text{(Equation 5)}$$

$$\mu_i(x) = EncoderLayer(h(x)) \quad \text{(Equation 6)}$$

$$\sigma^2{}_i(x) = EncoderLayer(h(x)) \quad \text{(Equation 7)}$$

$$\epsilon_i \sim N(0, I) \quad \text{(Equation 8)}$$

$$z_i \sim \mu_i(\frac{\partial x}{\partial t}) + \epsilon_i * \sigma^2{}_i\,(\frac{\partial x}{\partial t}) \quad \text{(Equation 9)}$$

$$DecoderLayer(z_i) = Relu(b_d + W_d * z_i) \quad \text{(Equation 10)}$$

where the $Relu$(z) activation function is:

$$Relu(z) = max(0, z) \quad \text{(Equation 11)}$$

We used the mean squared error reconstruction loss with the Adam optimizer. To prevent overfitting and encourage a sparse representation of latent embeddings, L1 regularization was added to all layers. The evidence lower bound loss function with L1 regularization where $\lambda = 1 \times 10^{-6}$, $q(z|\frac{\partial x}{\partial t}) = N(\mu_i(\frac{\partial x}{\partial t}), diag(\sigma^2{}_i(\frac{\partial x}{\partial t})))$ $p(z) = N(0, I)$, can be described as:

$$L\left(\frac{\partial x}{\partial t}, \widehat{\frac{\partial x}{\partial t}}\right) = KL - Divergence + Reconstruction\ Loss + Regularization$$

(Equation 12)

$$= KL(q\left(z\middle|\frac{\partial x}{\partial t}\right) \,||\, p(z)) - \sum_{i=1}^{D}\left(\frac{\partial x}{\partial t} - \frac{\widehat{\partial x}}{\partial t}\right)^2 + \lambda\sum_{i=1}^{D}|W_e| + \lambda\sum_{i=1}^{D}|W_d|$$

(Equation 13)

Because the input and output vectors are sparse, a small learning rate of 0.00001 was used. Early stopping was added once the validation loss did not improve for three consecutive epochs.

Initial Value Problems and ODE Solvers for Integration

Our framework can be used to predict gene expression profiles across time. Given t_0 and $\vec{x}(0) = x_0$, this is an initial value problem with the goal of solving $\vec{x}(t) = \overrightarrow{x_t}$ for any t:

$$\frac{\partial \vec{x}(t)}{\partial t} = f\big(t, \vec{x}(t)\big) \qquad \text{(Equation 14)}$$

Here, f is only a function of the state $\overrightarrow{x_t}$ such that $f = A(\overrightarrow{x_t})$. Then the equation becomes:

$$\frac{\partial \overrightarrow{x_t}}{\partial t} = \boldsymbol{\alpha}(\overrightarrow{x_t}) \qquad \text{(Equation 15)}$$

The first-order Euler's method for finding the state $\overrightarrow{x_{t+1}}$ is:

$$\overrightarrow{x_{t+1}} = \overrightarrow{x_t} + \frac{\partial \overrightarrow{x_t}}{\partial t} \qquad \text{(Equation 16)}$$

$$= \overrightarrow{x_t} + \boldsymbol{\alpha}(\overrightarrow{x_t}) \qquad \text{(Equation 17)}$$

However, we can utilize higher-order ODE solvers from the SciPy package to find a more accurate solution (*44*). The explicit Runge-Kutta method of order 8 (DOP853) was used to obtain the most accurate solutions, but it has a slow runtime. Explicit Runge-Kutta method of order 5 (RK53) can be used to trade off accuracy for a faster runtime. In practice, cells in this study were integrated to a maximum of 35 discrete steps (each with 5 intermediate steps) forward in time, which should be experimentally derived for each dataset.

Addressing Drift Effects

In control theory, using only the previous state and the velocity vectors to predict the next state can result in a phenomenon called "dead reckoning," where the errors accumulate after each step (*45*). To mitigate this effect, we employed two strategies:

1. Instead of a traditional VAE, we trained a denoising VAE to reduce the variance of predicted RNA velocity. By adding a small Gaussian noise ϵ to the gene expression input during training, we could increase the generalizability of the input space and improve extrapolations to out-of-sample cells.

$$\frac{\partial \vec{x}}{\partial t} = \boldsymbol{\alpha}(\vec{x} + \epsilon) \qquad \text{(Equation 18)}$$

2. As we integrated the VAE over time, we found reference cells in the data manifold every few steps and continued integration from the reference cell, as a form of high-gain Kalman filter. We designated the intermediate step size as a hyperparameter relative to the step size. For example, after integrating for five intermediate steps, we projected the predicted (or extrapolated) gene expression state to the original dataset

using the top 30 principal components. Then, we identified the K-nearest neighbors ($K = 30$) within the PCA embeddings. The reference cell is defined as the median expression profile among those K-nearest neighbor cells from the dataset, and ODE integration continued from this reference cell. This allowed our prediction to adhere closely to the data manifold and reduced the degree-of-freedom due to numerical errors. Further, finding reference cells in the data also constructed boundary conditions when integrating a dynamical system. For example, once the extrapolated state went beyond the cellular manifold, there were no cells in the data to serve as a reference, but the nearest neighboring cells from the dataset could still construct a reference cell from where integration could continue.

Measuring Instability with the Cell Criticality Index

By solving for the developmental path of a single cell, we can measure the amount of gene expression change along a trajectory, rather than comparing only the difference between the start and end states. Previously, StemID used the entropy of the gene expression distribution to heuristically identify stems cells in single-cell transcriptome data, where pluripotent cells tend to have a more uniform gene expression distribution with a higher entropy and differentiated cells tend to have a lower entropy (*46*). If $\overrightarrow{x^g}$ denotes the expression state of the genes g, then the StemID of the gene expression state is defined as:

$$\text{StemID}(\vec{x}) = -\sum_{i \in g} x^i log\left(\frac{x^i}{x^i}\right) \qquad \text{(Equation 19)}$$

We reasoned that a change in the gene expression distribution (e.g., from high to low entropy) can be captured using the relative entropy (or the KL-divergence). Based on this idea, we devised a measure to quantify the capacity for any cell to undergo gene expression change in the dynamical system. The CCI is calculated as the cumulative information change, or the cumulative KL-divergence, between gene expression distributions at each step in the developmental trajectory. Different from StemID, the CCI can quantify the gene expression change of a cell regardless of the pluripotency. As an analogy, StemID measures the "potential energy" of a cell's ability to differentiate, whereas the CCI measures the "kinetic energy" of a cell's ability to change. If $\overrightarrow{x_t^g}$ denotes the expression state of the genes g at time t, then the cumulative KL-divergence for $T = 35$ steps can be defined as:

$$CCI(\vec{x}) = \sum_{t=0}^{T} KL(\overrightarrow{x_{t+1}^g} \,||\, \overrightarrow{x_t^g}) \qquad \text{(Equation 20)}$$

$$= \sum_{t=0}^{T} \sum_{i \in g} x_{t+1}^i log\left(\frac{x_{t+1}^i}{x_t^i}\right) \qquad \text{(Equation 21)}$$

***In Silico* Perturbation Studies**

We divide an *in silico* perturbation study into three steps:

1. A sample of initial gene expression states ($n = 1{,}000$) was randomly generated. First, we solved the random initial gene expression states over time to establish a developmental baseline. Specifically, we aimed to observe the natural proportion of terminal cell types that could arise from the dynamical system without any intervention.
2. Then, we identified differentially expressed genes in the initial gene expression states that correlate with development into a particular terminal cell type later in time. Differential gene expression was performed using the *scanpy* package with Wilcoxon test and Bonferroni corrections (*47*).

3. Lastly, we perturbed only the differentially expressed genes in another set of randomly sampled initial gene expression states to test whether the perturbation increases the proportion of cells for the terminal cell type.

To sample initial gene expression states, we computed the median expression profile of progenitor cells and added Laplace distributed noise using the variance of those genes in progenitor cells to randomly increase or decrease gene expression. For the perturbation, exponentially distributed noise was added only to the top 100 differentially expressed genes for the randomly sampled cells to specifically increase the expression of the top differentially expressed genes. Terminal cell identity was determined by projecting the data onto the top 30 principal components and by using K-nearest neighbor classification (with $K = 30$). With the scVelo package, the dynamical mode estimates a variance for each gene over all cells, whereas the stochastic mode estimates a variance for each cell. Note that to model stochasticity in the stochastic mode, our framework could be easily adapted to also learn the variance of the velocity vectors (as neural stochastic ODEs). All initial gene expression states were integrated for 35 timesteps each with 5 intermediate steps.

Retrograde Trajectory Simulation

Similar to *in silico* perturbation studies, we computed the median expression profile of a terminal cell type (beta cells, granule cells, glutamatergic neurons, and erythroid) in each scRNA-seq experiment (mouse pancreatic endocrinogenesis, dentate gyrus, human forebrain, and mouse gastrulation) as the representative initial condition. A set of cells ($n = 50$) were sampled from each representative initial condition by adding Laplace distributed noise using the variance of the terminal cell type gene expression. The retrograde trajectory for each cell was simulated by subtracting the predicted RNA velocities from the gene expression state during integration:

$$\overrightarrow{x_{t-1}} = \overrightarrow{x_t} - \frac{\partial \overrightarrow{x_t}}{\partial t} \qquad \text{(Equation 22)}$$

$$= \overrightarrow{x_t} - \boldsymbol{\alpha}(\overrightarrow{x_t}) \qquad \text{(Equation 23)}$$

After integrating for 15 discrete steps each with 5 intermediate steps, a gene correlation matrix of the cells in retrograde trajectories was calculated.

Gene Ontology Enrichment Analysis

Hierarchical biclustering was performed on the co-expression matrices, and three gene clusters were identified from each co-expression matrix, representing three functional modules. We performed GO enrichment analysis on each functional module using GOATOOLS with Fisher's exact test (*48*). Further, we calculated the Benjamin-Hochberg false discovery rates to correct for multiple testing. To compare between two co-expression matrices, we considered the most significant enrichment out of the three clusters for each GO term. In Fig. 3a and Fig. 5a, the most significantly enriched GO terms associated with biological processes are listed next to each gene cluster.

References

1. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
2. M. Setty, V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis, D. Pe'er, Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37** (2019), pp. 451–460.

3. L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods*. **13**, 845–848 (2016).
4. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, others, RNA velocity of single cells. *Nature*. **560**, 494–498 (2018).
5. H. Matsumoto, H. Kiryu, C. Furusawa, M. S. Ko, S. B. Ko, N. Gouda, T. Hayashi, I. Nikaido, SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*. **33**, 2314–2321 (2017).
6. X. Qiu, Y. Zhang, J. D. Martin-Rufino, C. Weng, S. Hosseinzadeh, D. Yang, A. N. Pogson, M. Y. Hein, K. H. J. Min, L. Wang, others, Mapping transcriptomic vector fields of single cells. *Cell* (2022).
7. M. Bouhaddou, M. R. Birtwistle, Dimerization-based control of cooperativity. *Mol. Biosyst.* **10**, 1824–1832 (2014).
8. A. Uthamacumaran, A review of dynamical systems approaches for the detection of chaotic attractors in cancer networks. *Patterns*. **2**, 100226 (2021).
9. R. T. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, Neural ordinary differential equations. *ArXiv Prepr. ArXiv180607366* (2018).
10. S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci.* **113**, 3932–3937 (2016).
11. H. Hochgerner, A. Zeisel, P. Lönnerberg, S. Linnarsson, Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.* **21**, 290–299 (2018).
12. D. J. Di Bella, E. Habibi, R. R. Stickels, G. Scalia, J. Brown, P. Yadollahpour, S. M. Yang, C. Abbate, T. Biancalani, E. Z. Macosko, others, Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature*. **595**, 554–559 (2021).
13. A. Bastidas-Ponce, S. Tritschler, L. Dony, K. Scheibner, M. Tarquis-Medina, C. Salinno, S. Schirge, I. Burtscher, A. Böttcher, F. J. Theis, others, Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*. **146**, dev173849 (2019).
14. B. Pijuan-Sala, J. A. Griffiths, C. Guibentif, T. W. Hiscock, W. Jawaid, F. J. Calero-Nieto, C. Mulas, X. Ibarra-Soria, R. C. Tyser, D. L. L. Ho, others, A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*. **566**, 490–495 (2019).
15. E. Ferrell James, Bistability. *Bifurc. Waddingtons Epigenetic Landsc. Curr. Biol.* **22**, R458–R466 (2012).
16. M. E. Bechard, E. D. Bankaitis, S. B. Hipkens, A. Ustione, D. W. Piston, Y.-P. Yang, M. A. Magnuson, C. V. Wright, 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).
17. Z. Li, S. A. Langhans, Transcriptional regulators of Na, K-ATPase subunits. *Front. Cell Dev. Biol.* **3**, 66 (2015).
18. E. N. Lorenz, Deterministic nonperiodic flow. *J. Atmospheric Sci.* **20**, 130–141 (1963).
19. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
20. D. E. Wagner, A. M. Klein, Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21** (2020), pp. 410–427.
21. Y. Gonda, H. Sakurai, Y. Hirata, H. Tabata, I. Ajioka, K. Nakajima, Expression profiles of Insulin-like growth factor binding protein-like 1 in the developing mouse forebrain. *Gene Expr. Patterns*. **7**, 431–440 (2007).

22. E. M. Batista, J. G. Doria, T. H. Ferreira-Vieira, J. Alves-Silva, S. S. Ferguson, F. A. Moreira, F. M. Ribeiro, Orchestrated activation of mGluR5 and CB 1 promotes neuroprotection. *Mol. Brain*. **9**, 1–17 (2016).
23. L. Zou, H. Li, X. Han, J. Qin, G. Song, Runx1t1 promotes the neuronal differentiation in rat hippocampus. *Stem Cell Res. Ther.* **11**, 1–10 (2020).
24. S. G. Galfrè, F. Morandin, M. Pietrosanto, F. Cremisi, M. Helmer-Citterich, COTAN: scRNA-seq data analysis based on gene co-expression. *NAR Genomics Bioinforma.* **3**, lqab072 (2021).
25. D. Mercatelli, F. Ray, F. M. Giorgi, Pan-Cancer and Single-Cell Modeling of Genomic Alterations Through Gene Expression. *Front. Genet.* **10**, 671 (2019).
26. B. D. Harris, M. Crow, S. Fischer, J. Gillis, Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain. *Cell Syst.* (2021).
27. G. Karlebach, R. Shamir, Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 (2008).
28. S. Li, P. Brazhnik, B. Sobral, J. J. Tyson, A quantitative study of the division cycle of Caulobacter crescentus stalked cells. *PLoS Comput. Biol.* **4**, e9 (2008).
29. H. H. McAdams, L. Shapiro, A bacterial cell-cycle regulatory network operating in time and space. *Science*. **301**, 1874–1877 (2003).
30. J. Holtzendorff, D. Hung, P. Brende, A. Reisenauer, P. H. Viollier, H. H. McAdams, L. Shapiro, Oscillating global regulators control the genetic circuit driving a bacterial cell cycle. *Science*. **304**, 983–987 (2004).
31. E. G. Biondi, S. J. Reisinger, J. M. Skerker, M. Arif, B. S. Perchuk, K. R. Ryan, M. T. Laub, Regulation of the bacterial cell cycle by an integrated genetic circuit. *Nature*. **444**, 899–904 (2006).
32. M. S. Yeung, J. Tegnér, J. J. Collins, Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci.* **99**, 6163–6168 (2002).
33. D. C. Weaver, C. T. Workman, G. D. Stormo, in *Biocomputing'99* (World Scientific, 1999), pp. 112–123.
34. M. Mojtahedi, A. Skupin, J. Zhou, I. G. Castaño, R. Y. Leong-Quong, H. Chang, K. Trachana, A. Giuliani, S. Huang, Cell fate decision as high-dimensional critical state transition. *PLoS Biol.* **14**, e2000640 (2016).
35. A. E. Teschendorff, A. P. Feinberg, Statistical mechanics meets single-cell biology. *Nat. Rev. Genet.* **22** (2021), pp. 459–476.
36. C. Furusawa, K. Kaneko, Chaotic expression dynamics implies pluripotency: when theory and experiment meet. *Biol Direct*. **4** (2009), p. 17.
37. M. L. Heltberg, S. Krishna, M. H. Jensen, On chaotic dynamics in transcription factors and the associated effects in differential gene regulation. *Nat. Commun.* **10**, 1–10 (2019).
38. Y. Kuramoto, *Chemical oscillations, waves, and turbulence* (Courier Corporation, 2003).
39. P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, others, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. **353**, 78–82 (2016).
40. G. Gorin, V. Svensson, L. Pachter, Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 1–6 (2020).
41. M. Tedesco, F. Giannese, D. Lazarević, V. Giansanti, D. Rosano, S. Monzani, I. Catalano, E. Grassi, E. R. Zanella, O. A. Botrugno, others, Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nat. Biotechnol.*, 1–10 (2021).
42. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek

Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015), (available at https://www.tensorflow.org/).

43. F. Chollet, Building autoencoders in keras. *Keras Blog*. **14** (2016).

44. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, others, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*. **17**, 261–272 (2020).

45. S. Beauregard, H. Haas, in *Proceedings of the 3rd Workshop on Positioning, Navigation and Communication* (2006), pp. 27–35.

46. D. Grün, M. J. Muraro, J.-C. Boisset, K. Wiebrands, A. Lyubimova, G. Dharmadhikari, M. van den Born, J. Van Es, E. Jansen, H. Clevers, others, De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*. **19**, 266–277 (2016).

47. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).

48. D. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. W. Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, others, GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 1–17 (2018).

Acknowledgments

Funding: Research reported in this publication was supported by the National Institutes of Health under award numbers [insert here].

Author contributions:

Conceptualization: ZC
Methodology: WK, ZC
Investigation: ZC, WK, MB, JZ
Visualization: ZC
Supervision: MB, JZ
Writing—original draft: ZC
Writing—review & editing: MB, JZ

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The mouse pancreatic endocrinogenesis, dentate gyrus, neocortex, gastrulation, and human forebrain datasets used for this study can be found in the NCBI Gene Expression Omnibus (GEO) repository with accession numbers GSE132188, GSE95753, GSE153164, GSE87038, and in Sequence Read Archive (SRA) under accession code SRP129388, respectively. All source code to reproduce this study can be found on Github at https://github.com/gersteinlab/scDVF.