scDiffEq: drift-diffusion modeling of single-cell dynamics with neural stochastic differential equations

Tentative order: Michael E. Vinyard[1-4], Anders W. Rasmussen[2], Ruitong Li[2,5], Gad Getz[2-4,6,**], & Luca Pinello[2-6. **]

1. Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA., USA
2. Broad Institute of MIT and Harvard, Cambridge, MA., USA
3. Center for Cancer Research, Massachusetts General Hospital, Boston, MA., USA
4. Department of Pathology, Massachusetts General Hospital, Boston, MA., USA
5. Department of Biomedical Informatics, Harvard Medical School, Cambridge, MA., USA
6. Department of Pathology, Harvard Medical School, Boston, MA., USA

** co-supervised the work

Abstract

Single-cell sequencing measurements facilitate the reconstruction of dynamic biology by capturing snapshots of the molecular profiles of individual cells. Cell fate decisions in development and disease are orchestrated through an intricate balance of deterministic and stochastic regulatory events. Drift-diffusion equations are effective in modeling single-cell dynamics from high-dimensional single-cell measurements. While existing solutions describe the deterministic dynamics associated with the drift term of these equations at the level of cell state, the diffusion is modeled as a constant across cell states. To fully understand the dynamic regulatory logic in development and disease, models explicitly attuned to the balance between deterministic and stochastic biology are required. Addressing these limitations, we introduce scDiffEq, a generative framework for learning neural stochastic differential equations that approximate the deterministic and stochastic dynamics in biology. Using lineage-traced single-cell data, we demonstrate that scDiffEq offers improved reconstruction of held-out cell states and prediction of cell fate from multipotent progenitors during hematopoiesis. By imparting *in silico* perturbations to multipotent progenitor cells, we find that scDiffEq accurately recapitulates the dynamics of CRISPR-perturbed hematopoiesis. Using scDiffEq, we simulate high-resolution developmental cell trajectories, modeling their drift and diffusion, enabling us to study their time-dependent gene-level dynamics.

Introduction

Dynamical systems underpin fundamental processes in biology and disease, including developmental differentiation and cancer. Gene expression is a common molecular proxy used to characterize cell types and states. Single-cell measurements such as single-cell RNA-sequencing (scRNA-seq) can capture snapshots of both stable cell states as well as transient states occupied by cell subpopulations undergoing transition between more stably observed cell states. While individual cells are destroyed upon measurement, scRNA-seq facilitates rapid profiling of thousands of cells, which has enabled the development of computational strategies to infer the relationship between an observed cell state and its past and future states. These approaches facilitate the study of relationships between cell states as well as between cell states and cell fates in cell trajectories and enable new insights into the regulatory dynamics underlying developmental processes and disease.

The evolution of tools to study dynamics from single-cell molecular data have grown increasingly sophisticated, leveraging emerging techniques from machine learning and domain knowledge of underlying biology [1–6]. Trajectory inference methods have offered effective approaches for pseudotemporal ordering in low-dimensional representations of cell state though remain limited to correlative analyses of genes with pseudotime, restricting their ability to provide insights on the underlying mechanisms that give rise to these trajectories. RNA velocity leverages reasonable biophysical assumptions regarding nascent, mature and degradation of RNA transcripts to infer future cell states on short timescales. Methods, including Dynamo and CellRank, use RNA velocity to infer cell trajectories and fate on extended timescales [7,8]. However, methods for estimation of RNA velocity from single-cell expression are sensitive to preprocessing and smoothing operations and struggle to accurately model multi-fated trajectories[9]. Thus, methods that use RNA velocity as an input feature depend on the validity of the assumptions made about transcriptional kinetics during velocity estimation.

Drift-diffusion equations have been used to model cellular dynamics from snapshot data (**Fig**. **1a**). In high dimensions, such as those encountered in single-cell gene expression analyses, finding analytical solutions to partial differential equations is computationally intractable. Thus, to gain traction in modeling single-cell dynamics, initially-proposed frameworks necessitated strong assumptions. One of the first solutions to this framework, population balance analysis (PBA), proposed leveraging properties of spectral graph theory to model cell states at steady state via a weighted random walk through a cell neighbor graph [10].

PRESCIENT is a generative recurrent neural network that learns a drift field from coarse, time-resolved cell data using an optimal transport-based loss function. This drift field is regularized according to the assumption that cells exist in a gradient potential landscape – i.e, each forward step taken by the model is the negative gradient of the model output (potential), representing a Waddington Landscape of cell development [11,12].

Dynamo learns a smoothed vector field from noisy velocity estimates that serves as a drift term in the drift-diffusion framework [7]. PBA, PRESCIENT, and Dynamo propose models using a drift-diffusion framework though fix diffusion to a uniform Gaussian noise term, treating its magnitude as a tunable hyper-parameter. This treatment of diffusion thus assumes the stochastic dynamics of individual cells are cell state-independent, preventing further study of the stochastic nature of gene expression as a function of cell state.

At the molecular level, stochasticity is required to facilitate the development of diverse cell types that originate from a common progenitor [13]. This stochasticity functions in tandem with more deterministic evolved regulatory mechanisms to give rise to cellular diversity observed during dynamic developmental processes. Understanding the interplay between stochastic and deterministic gene expression is essential to build interpretable models of the complex processes underlying cell decision making. We sought to better understand when in time and where in gene expression space cells lean on stochasticity to make decisions and how this translates to coordinated changes in transcriptional states. Historically, differential equations have served as a workhorse of biological modeling. However, modeling complex, biological systems, even in low dimensions, typically requires assumptions built on decades of empirical observations. Fortunately, recent advances in deep learning, mainly neural differential equations, have provided a solution to numerically approximate dynamics governed by complex differential equations [14–16]. Neural stochastic differential equations (neural SDEs) offer a direct framework to parameterize both the drift and diffusion terms of a drift-diffusion equation, each with a deep neural network (**Fig**. **1b**) [15].

In this work, we build on existing models of cell dynamics, taking advantage of neural differential equations, to present scDiffEq, a deep learning framework that learns neural SDEs from embeddings of cell states to model and study their dynamics (**Fig**. **1c**). We benchmark scDiffEq against existing methods in approximating cell dynamics, using multi-time point lineage-traced scRNA-seq data (**Fig**. **1d**-**f**). We note distinct improvements in scDiffEq's prediction of cell fate from multipotent progenitor cells. We next observe scDiffEq's enhanced ability to interpolate distributions of held-out cell populations. We showcase scDiffEq's ability to identify genes correlated with cell diffusion, distinctly separate from those associated with cell drift. This highlights the critical importance of modeling diffusion in our understanding of cellular dynamics and offers a framework for its biological interpretation. Specifically, we investigate this through a detailed study of neutrophil/monocyte decision-making from a multipotent progenitor cell wherein we attempt to pinpoint cell fate decision-making in both time and gene expression space. We demonstrate, in continuous time resolution, the recovery of key gene regulatory relationships and study how these processes relate to both cell drift and diffusion properties.

Results

Learning neural differential equations with scDiffEq. scDiffEq is based on neural Stochastic Differential Equations (SDEs) and is designed to accept cell input of any dimension. Contemporary methods including PRESCIENT use principal component analysis (PCA) as a preprocessing step. For straightforward comparison, we use the first 50 principal components (PCs) of a z-scored gene expression matrix. scDiffEq requires the annotation of an initial position from which it solves an initial value problem (IVP), to begin fitting the neural SDE describing the dynamics of the observed cell manifold. When discreetly-labeled time points are provided, scDiffEq computes the Wasserstein Distance of cells sampled from the observed cell population against those it predicts (**Fig**. **1c**).

To illustrate the scDiffEq framework, in our experiments, we used a lineage-traced scRNA-seq dataset profiling mouse hematopoiesis based on the Lineage And RNA Recovery (LARRY) barcoding system, measured over three time points (days 2, 4, 6, post-barcode transduction). In total, the LARRY dataset comprises 130,887 scRNA-seq cell profiles. 49,302 (37.7%) of these measured cell states were successfully transduced with one of 5,864 lineage barcodes. 28,249 cells were profiled on day 2, 48,498 cells on day 4, and 54,140 on day 6. 4,638/28,249 day 2 cells (16.4%), 14,985/48,498 day 4 cells (30.9%), and 29,679/54,140 day 6 cells (54.8%) were lineage barcoded. On day 2, the 4,638 barcoded cells spanned 2,672 unique barcodes. The 14,985 barcoded cells observed on day 4 spanned 4,101 unique barcodes and the 29,679 barcoded cells observed on day 6 spanned 3,956 unique barcodes. At each time point, the most abundant barcodes occupied 8 cells, 56 cells, and 157 cells representing 0.17%, 0.38%, and 0.53% of barcoded cells, respectively.

Using this dataset, we demonstrate how scDiffEq samples cells from the day 2 population and approximates a small, discrete step, “forward”, advancing at a specified interval (dt). At each annotated, observed time point (such as days 4 and 6), it computes the Wasserstein Distance between the predicted cell population at that time point with a randomly sampled subset of the observed cells at the same time; this distance is approximated using the Sinkhorn Divergence (**Methods**). scDiffEq is then iteratively optimized to minimize the Sinkhorn Divergence between the predicted and observed cell manifolds, summed over each real time point (day 4 and day 6, using the LARRY dataset). The stochastic predictions of scDiffEq produce synthetic cells that, while similar to those observed in the original data, are unique. A nearest neighbor graph trained on the original data provides a means of mapping the predicted cells to a distribution of similar, real cells. To quantify how well the observed data might be recapitulated via simulation, we sampled an increasing number of simulated trajectories from the model and used a nearest neighbor graph to map the predicted cells to the observed cell manifold, revealing that we are able to recapitulate 59.5% of the LARRY dataset manifold from just 10,000 initial cells (**Suppl**. **Fig**. **1**). Once the model has converged, cell trajectories simulated in conjunction with the IVP-solver may then accurately represent a cell traversing through the learned latent time and space, enabling prediction of cell trajectories from clonal lineage families.

Benchmarking models of single-cell dynamics with multi-time point lineage-traced scRNA-seq data. Benchmark datasets paired with standardized analysis tasks are required to validate and compare predictive models. However, single-cell data generation destroys the measured cell, impeding the observation of ground truth relationships between measured cells and their true past and future states. Despite this inherently obfuscated observation of cell-cell relationships, we and others have employed the LARRY dataset as an approximation of the ground truth real-time cell dynamics (**Fig**. **2a, b**). The LARRY dataset circumvents the destruction of cell trajectories by transducing multipotent progenitor (MPP) cells with lentiviral barcodes, which are heritably propagated to their daughter cells. Briefly, two days post-transduction, the MPP cells were sampled for scRNA-seq and divided into parallel wells. Each well of cells were measured via scRNA-seq four and six days, post-transduction. While not every cell sampled for scRNA-seq retained a heritable barcode, 5,864 lineage barcodes were recovered, spanning 49,302 of the 130,887 measured cells and thus enabling the coarse reconstruction of real-time cell development in hematopoiesis over three time points (**Fig**. **2c, Suppl**. **Fig**. **2**). This dataset is among a growing collection of lineage-traced single-cell datasets that pair heritable DNA barcodes with single-cell sequencing measurements, offering temporally dependent descriptions of cell states and thereby a real-time reconstruction of the temporal dynamics including state-state and state-fate relationships [17–24].

Here, we adapt and build on benchmark tasks that have previously been used in conjunction with this LARRY dataset [12,24] to compare the accuracy of models aimed at learning and predicting cellular dynamics on two tasks: fate bias prediction (Task 1) and prediction of intermediate cell states at unobserved time points (Task 2). For Task 1 we reasoned that for a given heritably barcoded cell observed at day 2 in the LARRY dataset, should a matching barcode be identified in another cell at one or more later time points in the dataset, we can infer that those cells are clonally related. Thus, in a multipotent cell system, the barcode of a progenitor cell enables a glimpse into how that progenitor cell may or may not be biased towards formation of a specific cell fate. Taking advantage of the LARRY dataset’s ability to highlight state-fate relationships, we first benchmarked scDiffEq against several methods, including methods specialized towards modeling single-cell transcriptomic data as well as more general classification algorithms such as nearest-neighbors and logistic regression. The goal of the fate prediction task is to accurately infer the final “fate bias” or relative proportion of cell fates formed, from a given progenitor cell, compared to the real values tabulated for each lineage observed in the LARRY dataset (**Fig**. **1d**). To prepare the LARRY dataset for each task, we followed the pre-processing procedure demonstrated

in the work describing PRESCIENT (**Methods**). Briefly, for Task 1, fate prediction, we first segmented the dataset into train and test sets wherein all day 2 cells and the day 4 and day 6 cells from Well 1 were used as the training set. Cells in Well 2 were reserved for the test set (**Fig**. **2b**, **Methods**). We then randomly initialized scDiffEq over five seeds, fitting the model as described above on the training set. For each cell lineage, we then sampled the model, simulating 2,000 trajectories to arrive at a representative approximation of the model's predicted outcome.

A discriminative classifier such as logistic regression predicts cell fates with greater accuracy (57.7%) than any of the single-cell focused methods that we benchmarked here. While a classifier method offers a reasonable prediction of cell fate from standardized input features, they provide relatively little information towards the underlying molecular processes and dynamics, beyond classification. We thus begin our benchmark with these baselines towards determining the learnable information content of scRNA-seq state descriptions that may be used to make predictions of future cell states, focusing on the relative gains in insight each method offers in the context of their predictive capabilities. Torch-PBA, our PyTorch implementation of the PBA framework (**Methods**), predicted cell fates with a mean accuracy (n=5) of 53.6%. Dynamo offers in-depth analyses of generatively sampled cell trajectories through only predicted cell fates with 26.3% accuracy. CellRank, which does not employ any drift-diffusion modeling assumptions though is widely-adopted, predicted cell fates with only 15.6% accuracy. Notably, Dynamo and CellRank, both of which rely on RNA-velocity estimates performed worse than all other methods, none of which use information derived from ratios of nascent and mature mRNA transcripts. PRESCIENT outperformed CellRank and Dynamo, predicting cell fate with an mean (n=5) accuracy of 39.4% (without KEGG weights) and 48.7% (with KEGG weights). "KEGG weights" refer to the growth-informed weights (via KEGG gene expression signature) used in the Wasserstein Distance loss calculation as performed in previous studies using a similar optimal transport framework. scDiffEq outperformed all methods tailored towards single-cell analysis in cell fate prediction with mean accuracies (n=5) of 52.5% (without KEGG weights) and 53.3% (with KEGG weights), representing a ~4.6% improvement in prediction accuracy of the existing state-of-the-art method, PRESCIENT (**Fig**. **2d**). In general, we find that modeling non-uniform diffusion (as compared to PRESCIENT) enables us to more accurately predict fates outside of Neutrophils and Monocytes (**Suppl**. **Fig**. **3**).

While fate prediction is informative with respect to state-fate relationships, discriminative models are not designed to also recapitulate a biological process. Both scDiffEq and PRESCIENT enable prediction of intermediate, unobserved cell states. In Task 2, interpolation of cells from a withheld time point, we compared scDiffEq to PRESCIENT, following the procedure outlined in Yeo, et. al., 2021. Briefly, models are fit to the LARRY dataset using only cells from day 2 and day 6, withholding cells at day 4. Day 4 cells then serve as the test set by which models are evaluated (**Fig**. **1e**). Successful reconstruction of the withheld time point was measured using the Wasserstein Distance loss function, approximated as the Sinkhorn Divergence (arbitrary units)[25]. Over five seeds, PRESCIENT was able to achieve a mean training distance of 14.94 and a test distance of 25.85. scDiffEq was able to minimize the training distance to 13.74 and the test distance to 24.56 (**Fig**. **2e**). As described above, while PRESCIENT is grounded in similar assumptions to scDiffEq, it is restricted to learning functions constrained to a potential gradient and does not explicitly parameterize the diffusion term in the drift-diffusion framework. scDiffEq models of comparable size to PRESCIENT, in terms of parameters (two fully-connected layers of 400 nodes) though unconstrained to a gradient of potential were unable to outperform PRESCIENT, we were eventually able to improve upon the performance achieved by PRESCIENT through increased model complexity: we composed the drift and diffusion networks of the scDiffEq model for this task using two fully-connected layers of 4000 nodes and two fully-connected layers of 800 nodes, respectively.

We note a distinct advantage in the approach taken by scDiffeq and PRESCIENT, to learn and subsequently simulate a latent time, from coarse, real time measurements of cells. Using a sample model prediction from scDiffEq, we demonstrate, over the course of a simulated population, cells beginning with zero error from their sampled d2 population, moving away from that population and eventually minimizing their distance to the observed d4 and d6 populations at increments of 0.1d (**Fig**. **2f**). While many methods are capable of assigning an arbitrary pseudotime to a biological process, relatively few are able to generate a latent time based in real time units.

To encourage transparency and forward progress within the field, we make our benchmark implementation available as an open-source Python package such that the community may readily apply and evaluate new models to this benchmarking framework (**Methods**).

Model predictions are corroborated using in silico perturbations. Inspired by the work presented in Yeo, et. al., 2021, we next asked whether scDiffeq could simulate expected changes to biological systems under perturbed conditions. We introduced *in silico* perturbations for an ensemble of transcription factors (TFs) known to be involved in granulopoiesis and neutrophil development: *Lmo4*, *Dach1*, *Klf4*, and *Cebpe*. Each perturbation was introduced to the z-scored gene expression matrix of 200 randomly selected undifferentiated cells from day 2. The unperturbed, zero-centered z-score values of target genes in selected cells are then directly set to a positive value for over-expression or a negative value for perturbations that represent knockdown or knockout. For each perturbed progenitor cell, the modified gene expression z-scores were transformed into the latent space using the original PCA model - these values were used as input to scDiffEq to simulate the resulting trajectories. We compared the simulated trajectory for perturbed cells to that of an unperturbed control simulation, qualitatively noting a shift in population density from monocytes towards neutrophils, when the neutrophil TF ensemble was perturbed to z=10 (**Fig**. **2g**, **h**). Next, we systematically profiled an array of expression perturbations spanning z-scores = 2.5, 5, and 10 for simulated overexpression and -2.5, -5, and -10 for simulated gene expression knockdown. Compared to the control with a mean neutrophil fate fraction of 0.129 and mean monocyte fate fraction of 0.577±0.024, for each overexpression experiment of z = 2.5, 5, and 10, a corresponding dose-dependent response, increasing the fraction of neutrophils predicted (0.171±0.021, 0.208±0.015, and 0.293±0.024, respectively, $p<0.05$ using Welch's independent two-sided *t*-test) while decreasing the fraction of monocytes predicted (0.507±0.019, 0.431±0.032, and 0.267±0.023, respectively, $p<0.05$) at the final time point, t = 6d (**Fig**. **2i**, **j**). Similarly, for simulated knockdowns, while z = -2.5 does not produce a significantly different fraction of neutrophils, z = -5 and z = -10 produce dose-dependent decreases to the fraction of neutrophils formed (0.101±0.010 and 0.075±0.010, respectively, $p < 0.05$) and z = -2.5, -5, and -10 all produced significant corresponding increases in the fraction of monocytes formed (0.628±0.022, 0.680±0.022, and 0.734±0.027, respectively, $p<0.05$) (**Fig**. **2i**, **j**).

Decomposing the learned dynamics described by drift and diffusion from scDiffEq. Proceeding with a model checkpoint optimized for fate prediction, without further refitting, we next reevaluated each cell in the original dataset using both the independent model components: the neural networks for drift and diffusion to obtain the corresponding 50-dimension drift and diffusion vectors for each cell state. Each 50-dimension vector describes the instantaneous drift or diffusion "velocity" of each evaluated state. We summarized the magnitude of the drift and diffusion forces for each cell as a scalar value by computing the L2Norm of the aforementioned 50-dimension vectors. We used a nearest neighbor graph to smooth these values and visualize using UMAP (**Fig**. **2b**, **Suppl**. **Fig**. **4**), noting, qualitatively, cell- and group-specific fluctuations in both drift and diffusion.

Next, we simulated 2,000 trajectories from each of the 2,081 cell states observed in d2 that are also heritably observed in later time points. We show three representative examples spanning the reconstruction of single-fate lineages (**Fig**. **3a**), bi-fated lineages (**Fig**. **3b**) and lineages with three or more fates (**Fig**. **3c**). As before, for each simulated trajectory we reevaluate every simulated cell state against the drift and diffusion components of the model from which they were generated. For each of the 2,081 simulations, we computed the mean drift and diffusion at each 0.1d interval, grouping each trajectory according to its fate multiplicity or, the number of fates reached from a given progenitor state (**Suppl**. **Fig**. **5**). We focus on mono- and bi-fated simulations fated predominantly towards neutrophil or monocyte. We note mean maximum drift values of 13.40±1.97 and 15.08±1.94 for mono- and bi-fated simulations fated towards neutrophil while simulations fated towards monocyte reach maximum drift values of 18.55±2.57 and 16.80± 2.49 for mono- and bi-fated simulations, respectively. Mono- and bi-fated neutrophil-fated simulations reach a mean maximum diffusion of 6.25±0.90 and 5.46±0.86 while monocyte-fated simulations reach maximum diffusion values of 0.86±1.22 and 4.76±1.05, respectively (**Fig**. **3d**, **e**). Summarizing these observations: for monocyte- and neutrophil-fated trajectories, the temporal drift appears to remain relatively uniform, regardless of which fate is reached or if only one or both fates are reached from an initial state. In contrast, of trajectories that reach only a single fate, those biased towards neutrophils demonstrate distinct temporal diffusion from those biased towards monocytes; these differences being reduced for bi-fated trajectories.

The bi-fated simulation shown in **Fig**. **3b** highlights a multipotent progenitor cell that is empirically observed, using the lineage tracing data to produce both monocytes and neutrophils. Using this simulation as an example to explore the learned model drift and diffusion terms, we decompose the cell state-specific drift and diffusion along the trajectory as in **Fig**. **2d**, and visualize each individual state using UMAP (**Fig**. **3f**, **g**).

Within a simulation, attributes of each sampled trajectory may be annotated at each incremental time step (dt = 0.1d), allowing us to establish connections between observations in one state and those in other states or outcomes. Specifically, we can categorize trajectories based on their final states (fates) in simulations where we observe the formation of multiple fates, enabling us to condition attributes on these fates and make comparisons of attributes between conditions. Further, this categorization enables us to analyze the points of divergence among these trajectories. For instance, we can identify unique features of trajectories that originate from the same initial cell though ultimately reach separate fates (e.g., monocyte and neutrophil). Applying this strategy to the bi-fated simulation shown in **Fig**. **3b**, we compute the mean drift and diffusion values (plotted with standard deviation), at each time step, for both monocyte-fated trajectories as well as neutrophil-fated trajectories (**Fig**. **3h**, **i**). Since PCA was used to embed cells into the input latent space, we linearly inverted the scDiffEq-predicted 50-dimension cell states to estimate z-scored gene expression. Subsequently, we can arrive at non-negative estimates of gene expression by converting the gene z-scores back to log-normalized, non-negative expression values. This procedure enables estimation of gene expression for every scDiffEq-predicted cell state. We then compute the correlation of reconstruction expression for each gene, conditioned on fate, to drift and diffusion, conditioned on fate, within the simulation. Normalized expression values of the top 25 correlated genes, for drift and diffusion in neutrophil and monocyte-fated trajectories are displayed in **Fig**. **3h**, **i**.

Elane (**Fig**. **3j**) and *Mpo* (**Fig**. **3k**) are markers of neutrophil development [24,26]. Predicted expression of both Elane and Mpo were moderately anti-correlated with Neutrophil drift (rho=-0.66, p=3.0e-06; rho=-0.53, p=3.8e-4, respectively using Student's two-sided *t*-test)) though moderately and highly correlated neutrophil diffusion, respectively (rho=0.47, p=1.7e-3; rho=0.71, p=2.2e-07 using Student's two-sided *t*-test). The predicted temporal expression profiles of *Elane* and *Mpo* were not correlated to monocyte drift. However, predicted expression profiles of both *Elane* and *Mpo* were observed to exhibit strong positive correlation with monocyte diffusion (rho=0.66, p=3.0e-06; rho=0.83, p=4.2e-11, respectively using Student's two-sided *t*-test).

Granulocyte colony-stimulating factor receptor (G-CSF-R, also known as CD114) is encoded by *Csf3r*. G-CSF binding to G-CSF-R is an essential stimulation event during neutrophil development and proliferation [27]. Temporal expression of *Csf3r* (**Fig**. **3l**) was observed as moderately correlated to drift (rho=0.33, p=3.6e-02, using Student's two-sided *t*-test) and moderately anti-correlated to diffusion (rho=-0.54, p=2.4e-04, using Student's two-sided *t*-test) in neutrophil-fated trajectories. *Csf3r* is also highly anti-correlated to monocyte diffusion (rho=-0.69, p=5.5e-07, using Student's two-sided *t*-test).

CCAAT/enhancer-binding protein alpha (Cebpa) is a crucial TF in the differentiation of immature granulocytes [28,29]. A recent, detailed CRISPR screening investigation of mouse hematopoiesis demonstrates that loss of *Cebpa* results in decreased neutrophils and monocytes from multipotent progenitors. However, in the same study, knockout of *Irf8* in MPPs was shown to selectively decrease development of monocytes without the same impact on the development of neutrophils. Predicted temporal expression of *Cebpa* (**Fig**. **3m**) was observed to be in the top 2% of genes positively correlated with diffusion (rho=0.97, pval=1.3e-25, using Student's two-sided *t*-test) though not significantly correlated with drift in monocyte-fated trajectories. In neutrophil-fated trajectories, *Cebpa* was observed to be in the top 15% of genes positively correlated with diffusion (rho=0.87, p=7.5e-14, using Student's two-sided *t*-test) though not significantly correlated with drift. Predicted temporal expression of *Irf8* (**Fig**. **3n**) was found to be in the top 2% of genes positively anti-correlated with drift (rho=-0.93, p=2.7e-18, using Student's two-sided *t*-test) and highly anti-correlated with diffusion (rho=-0.76, p=9.5e-09, using Student's two-sided *t*-test) of monocyte-fated trajectories. *Irf8* was also found to be in the top 2% of genes positively anti-correlated with drift (rho=-0.83, p=2.7e-11, using Student's two-sided *t*-test) of neutrophil-fated trajectories though only moderately correlated with Neutrophil diffusion (rho=0.42, p=5.7e-03, using Student's two-sided *t*-test). In agreement with the Giladi et al., 2018 CRISPR perturbation results, *Irf8* is more tightly correlated with both the drift and diffusion of Monocyte-fated trajectories as compared that of neutrophil-fated trajectories.

Granulocyte-monocyte progenitor cells (GMPs) are the immediate precursor to granulocytes like neutrophils and agranulocytes like monocytes. *Flt3* plays an important role in the development of GMPs towards monocyte lineages [30–32]. Predicted temporal expression of *Flt3* (**Fig**. **3o**) was observed to be moderately correlated with drift in both monocyte- and neutrophil-fated trajectories (rho=0.39, p=1.1e-02; rho=0.35, p=2.5e-02, respectively using Student's two-sided *t*-test). Strongly correlated (top 15%) with diffusion in both monocytes (rho=0.93, p=3.3e-18, using Student's two-sided *t*-test) and neutrophils (rho=0.87 , p=8.9e-14, using Student's two-sided *t*-test).

Finally, we use *Spi1* and *Gfi1* as a brief vignette to contextualize our findings. *Spi1* encodes the pioneer TF protein, PU.1 [33]. The GFI1 protein, encoded by *Gfi1* is a transcriptional repressor in hematopoiesis. *Gfi1* and *Spi1* function antagonistically through a protein-protein interaction in regulating mouse hematopoiesis. Here we observe this relationship to be captured by the model, potentially reflected by their relative correlations to the drift and diffusion forces acting on cells during myeloid development. *Spi1* is strongly anti-correlated to monocyte diffusion (rho=-0.96, p=1.6e-22, using Student's two-sided *t*-test) and highly anti-correlated to neutrophil diffusion (rho=-0.71, p=1.7e-07, using Student's two-sided *t*-test). In contrast, *Gfi1* is highly correlated with monocyte diffusion (rho=0.76, p=9.3e-09, using Student's two-sided *t*-test) and moderately correlated with neutrophil diffusion (rho=0.47, p=2.1e-03, using Student's two-sided *t*-test). In terms of drift, while *Spi1* is highly anti-correlated to drift in the monocyte trajectory (rho=-0.60, p=3.6e-05, using Student's two-sided *t*-test), *Gfi1* is not correlated. Conversely, while *Gfi1* is highly anti-correlated to drift in neutrophil trajectory (rho=-0.67, p=1.5e-06, using Student's two-sided *t*-test), *Spi1* is not correlated to drift in the neutrophil trajectory (**Fig**. **3p**, **q**). Further work will be required to demonstrate this observation over the initiation of several models followed by significance testing.

Discussion

scDiffEq is a drift-diffusion framework that leverages neural stochastic differential equations (SDEs) to capture the deterministic and stochastic dynamics of single-cell. Using the LARRY dataset, we built on a previously-implemented set of benchmarking tasks that aim to ascertain our models' ability to learn cell dynamics. Through these benchmarking tasks, we demonstrate an improved reconstruction of cell dynamics as measured by cell fate prediction (Task 1) and reconstruction of unseen cell populations (Task 2) when compared with existing methods. In developing this benchmark, we have built on existing literature and expanded accessibility, transparency, and utility of these benchmarks for future studies.

We explore the relationship between learned drift and diffusion fields, as they relate to cell fate decision-making. As a generative model, scDiffEq facilitates simulation of cellular differentiation from hematopoietic progenitors to mature granulocytes. These simulations offer insights into the deterministic and stochastic changes in gene expression across trajectories, and how they differ by terminal fate. At the gene-level, we demonstrate that our simulated trajectories recapitulate experimentally observed patterns of expression in neutrophil and monocyte lineages. We identify several genes with temporal expression profiles that demonstrate a strong correlation to fate-specific drift or diffusion, including *Irf8*, *Gfi1*, *Spi1*, and *Cebpa,* all of which are known to regulate myelogenesis and mediate other key gene regulatory relationships.

scDiffEq presents a flexible drift-diffusion framework for modeling the dynamics from single-cell data. While we benchmark and demonstrate this method using scRNA-seq data, we anticipate that this framework is likely amenable to modeling data from other single-cell modalities, including spatial transcriptomics or measurements multiplexed with perturbations. As a generalizable method for training neural SDEs, scDiffEq establishes a foundation on which new methods for biologically-informed regularization may be implemented. We anticipate this method to serve both as an instrumental tool for studying single-cell trajectories and cell-fate decisions, as well as an important methodological step towards developing the next generation of generative models for studying biological dynamics from single-cell data.

References

1. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
2. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
3. Chen, H. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).
4. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
5. Wang, S.-W., Herriges, M. J., Hurley, K., Kotton, D. N. & Klein, A. M. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat. Biotechnol.* **40**, 1066–1074 (2022).
6. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 1517 (2019).
7. Qiu, X. *et al.* Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690–711.e45 (2022).
8. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
9. Gorin, G., Fang, M., Chari, T. & Pachter, L. RNA velocity unraveled. *PLoS Comput. Biol.* **18**, e1010492 (2022).
10. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2467–E2476 (2018).
11. Hashimoto, T., Gifford, D. & Jaakkola, T. Learning Population-Level Diffusions with Generative RNNs. in *Proceedings of The 33rd International Conference on Machine Learning* (eds. Balcan, M. F. & Weinberger, K. Q.) vol. 48 2417–2426 (PMLR, 20--22 Jun 2016).
12. Yeo, G. H. T., Saksena, S. D. & Gifford, D. K. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nat. Commun.* **12**, 3222 (2021).
13. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
14. Chen, R. T. Q. & Rubanova, Y. Neural ordinary differential equations. *Adv. Neural Inf. Process. Syst.* (2018).

15. Kidger, P., Foster, J., Li, X. & Lyons, T. J. Neural SDEs as Infinite-Dimensional GANs. in *Proceedings of the 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) vol. 139 5453–5463 (PMLR, 18--24 Jul 2021).
16. Kidger, P. On Neural Differential Equations. *arXiv [cs.LG]* (2022).
17. VanHorn, S. & Morris, S. A. Next-Generation Lineage Tracing and Fate Mapping to Interrogate Development. *Dev. Cell* **56**, 7–21 (2021).
18. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
19. Biddy, B. A. *et al.* Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).
20. Bowling, S. *et al.* An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell* **181**, 1693–1694 (2020).
21. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
22. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
23. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
24. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, (2020).
25. Feydy, J. *et al.* Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (eds. Chaudhuri, K. & Sugiyama, M.) vol. 89 2681–2690 (PMLR, 16--18 Apr 2019).
26. Metzler, K. D. *et al.* Myeloperoxidase is required for neutrophil extracellular trap formation: implications for innate immunity. *Blood* **117**, 953–959 (2011).
27. Lieschke, G. J. *et al.* Mice lacking granulocyte colony-stimulating factor have chronic neutropenia, granulocyte and macrophage progenitor cell deficiency, and impaired neutrophil mobilization. *Blood* **84**, 1737–1746 (1994).

28. Porse, B. T. *et al.* E2F Repression by C/EBPα Is Required for Adipogenesis and Granulopoiesis In Vivo. *Cell* **107**, 247–258 (2001).

29. Johansen, L. M. *et al.* c-Myc is a critical target for c/EBPalpha in granulopoiesis. *Mol. Cell. Biol.* **21**, 3789–3806 (2001).

30. Buza-Vidas, N. *et al.* FLT3 expression initiates in fully multipotent mouse hematopoietic progenitor cells. *Blood* **118**, 1544–1548 (2011).

31. Böiers, C. *et al.* Expression and role of FLT3 in regulation of the earliest stage of normal granulocyte-monocyte progenitor development. *Blood* **115**, 5061–5068 (2010).

32. Kim, S.-W. *et al.* Flt3 ligand induces monocyte proliferation and enhances the function of monocyte-derived dendritic cells in vitro. *J. Cell. Physiol.* **230**, 1740–1749 (2015).

33. Dahl, R., Iyer, S. R., Owens, K. S., Cuylear, D. D. & Simon, M. C. The transcriptional repressor GFI-1 antagonizes PU.1 activity through protein-protein interaction. *J. Biol. Chem.* **282**, 6473–6483 (2007).

Main figures

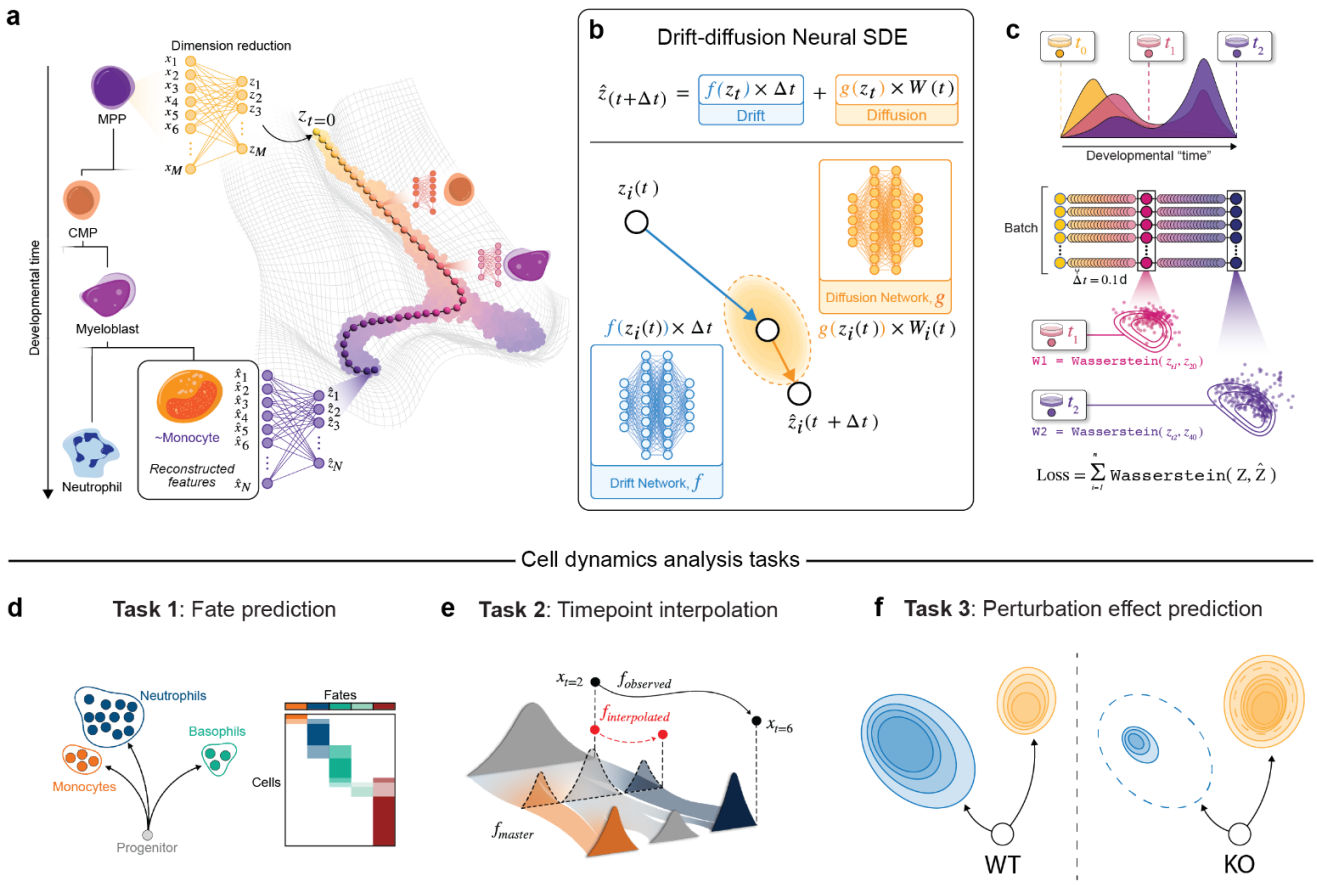


Figure 1. scDiffEq algorithm overview **a**. Conceptual overview of modeling a dynamical cell system such as hematopoietic development. **b**. Modeling cell drift and diffusion with neural differential equations in scDiffEq. **c**. Schematic diagram of scDiffEq training. **d-f**. Graphical summary of applications and analyses enabled by scDiffEq.

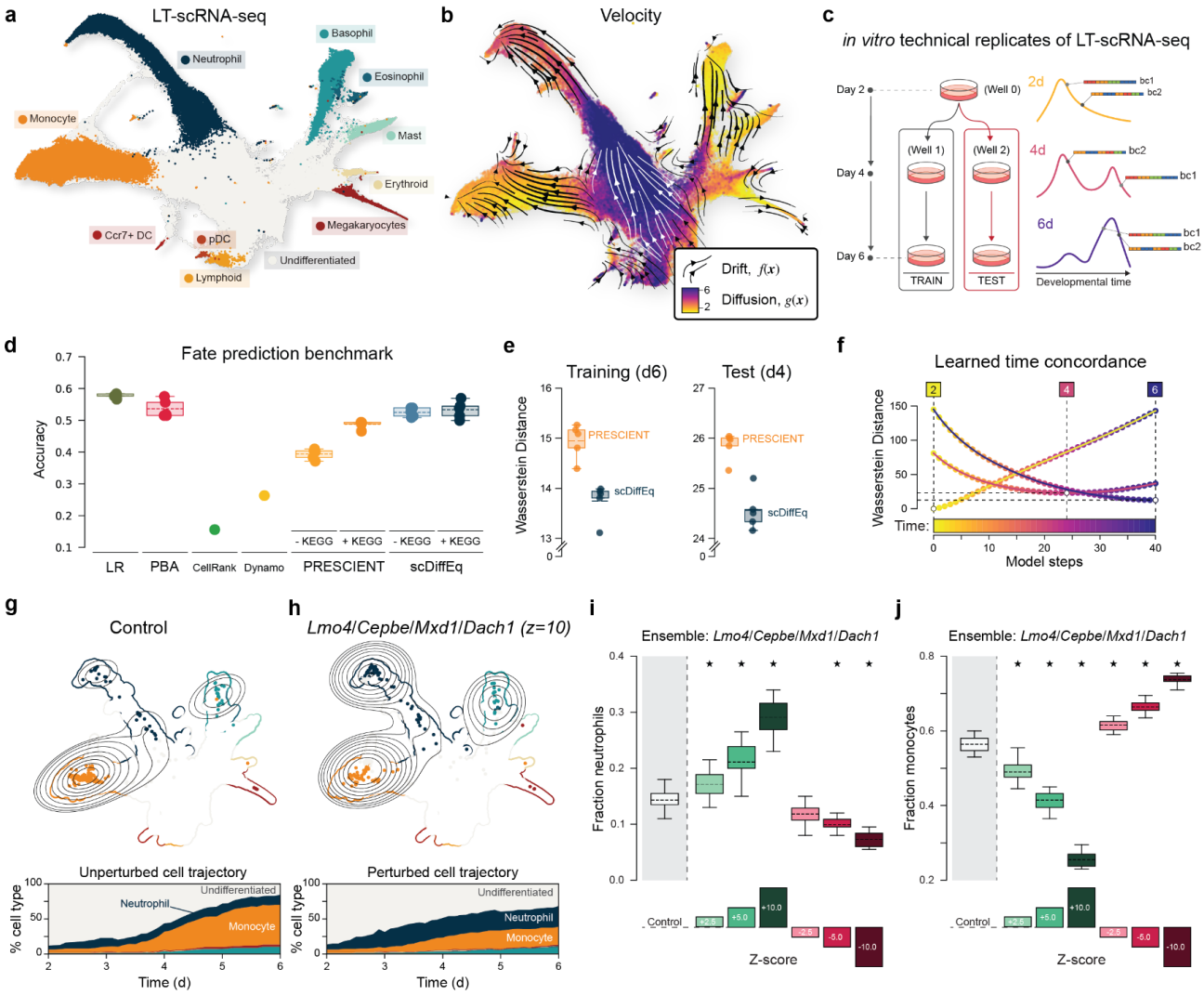


Figure 2. Benchmarking models of cell dynamics using the LARRY dataset. **a**. UMAP of the in vitro LARRY scRNA-seq dataset colored by cell type labels. **b**. UMAP stream plot illustrating scDiffEq-learned cell velocity decomposed into drift (vector field) and diffusion (L2Norm plotted using the colormap). **c**. Schematic overview of the lineage-tracing strategy and experimental setup of the in vitro LARRY dataset. **d**. Fate prediction accuracy for each tested method (n=5, except for CellRank and Dynamo, for which we were only able to recover a single deterministic result. **e**. Task 2 performance, comparing the relative ability of scDiffEq and PRESCIENT to minimize the Sinkhorn Divergence (y-axis) for the training (d6) and test (d4) sets. **f**. Sinkhorn Divergence distance (y-axis) of the predicted distribution of cells, at each discretized model time step (0.1d) against the true cell population. White points indicate distance minima from the true distributions of d4, d6 cells. **g**. UMAP (top) highlighting the predicted distribution of cells at t = d6 under control (unperturbed) conditions and the corresponding percentages of cell state labels at each time point in the simulated (unperturbed) trajectory. **h**. UMAP (top) highlighting the predicted distribution of cells at t = d6 under perturbation conditions and the corresponding percentages of cell state labels at each time point in the simulated (perturbed) trajectory. **i**. Fraction of neutrophils predicted at t = 6d over varying Z-score magnitudes of the ensembled perturbation. **j**. Fraction of monocytes predicted at t = 6d over varying Z-score magnitudes of the ensembled perturbation. In both (**i**) and (**j**), Stars indicate p-value < 0.05, when compared to the unperturbed control.

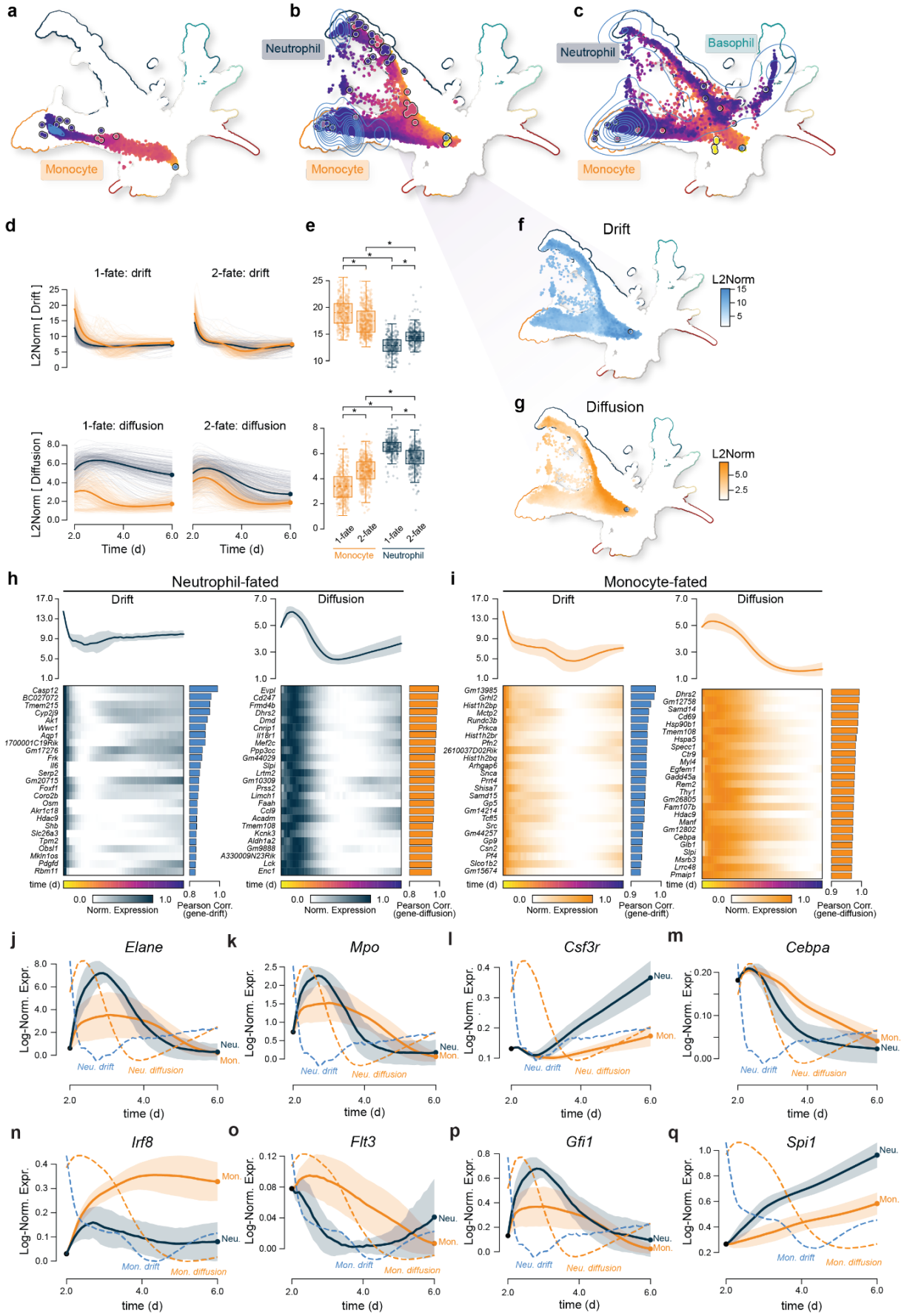


Figure 3. Decoupling cell-specific drift and diffusion. **a-c**. UMAP of an scDiffEq model simulation of a relatively

mono-fated (**a**), bi-fated (**b**), and multi(3+)-fated (**c**) d2 progenitor cell from the in vitro LARRY scRNA-seq dataset. The simulation is colored according to time and plotted against the observed cell manifold. Blue kernel density maps indicate the distribution of simulated cells at d6. **d.** Mean L2Norm of cell drift (top) and diffusion (bottom) at each time point, for every trajectory predominantly biased towards the neutrophil or monocyte fates, fate multiplicity. Each simulation (N=2,000 Individual trajectories) is represented by the lighter-weight lines and colored according to fate, while the bold line represents the mean of all trajectories, conditioned on fate. The left two subplots show simulations that reach only a single fate, while the right two subplots show those that are bi-fated. **e**. Mean maximum value of drift (top) and diffusion (bottom), for each group of simulations shown in (d). * indicates $p < 0.05$. **f**, **g**. UMAP of the scDiffEq simulation from (**b**), colored according to the smoothed L2Norm of cell (**f**) drift and (**g**) diffusion. **h**, **i**. Temporal L2Norm of cell drift (top, left) and diffusion (top, right) for (**h**) neutrophil-fated and (**i**) monocyte-fated trajectories sampled from the model simulation shown in (**b**). Below each are the top 25 rank-ordered drift- and diffusion-correlated gene expression predictions with respective Pearson correlation shown to the right. **j**-**q**. Temporal expression of the critical regulatory genes in neutrophil and monocyte development: *Elane* (**j**), *Mpo* (**k**), *Csf3r* (**l**), *Cebpa* (**m**), *Irf8* (**n**), *Flt3* (**o**), *Gfi1* (**p**), and *Spi1* (**q**). Overlaid the mean and standard deviation of gene expression are dotted lines indicating the relevant drift or diffusion from the simulation shown in (**h**), and (**i**).

Methods

Data preprocessing.

Following the procedure from PRESCIENT, we filtered non-highly-variable genes, regressed out genes associated with a list of cell cycle genes. We then scaled the log-normalized expression counts and performed PCA using sci-kit learn. As described above, we have assembled a package to reproducibly fetch, preprocess, and format the LARRY dataset and interface that dataset with models to recapitulate the benchmarking efforts we and others have used with this data. First, the data matrices and associated cell- and gene-level metadata are fetched from: https://github.com/AllonKleinLab/paper-data/tree/master/Lineage_tracing_on_transcriptional_landscapes_links_state_to_fate_during_differentiation (commit: af842ce).

Drift-diffusion model of cell dynamics.

Allow $z \in R^G$ to be a cell's position in a low dimensional representation of transcriptional space. We assume the dynamics of a cell's state change can be represented as a drift-diffusion equation of the following form:

$$dz(t) = f(z(t))dt + g(z(t))dW(t)$$

Where $f(z)$ and $g(z)$ are drift and diffusion vector fields defined on R^G. t is an unobserved latent time, and $W(t)$ is Brownian motion. To constrain $f(z)$, adopt Yeo et al.'s approach of defining the drift field as the negative gradient of some potential function $\psi(z)$, a scalar field defined on R^G. As such the equation can now be written as

$$dz(t) = -\nabla\psi(z(t))dt + g(z(t))dW(t)$$

By integrating forward in time, we can compute the future state of a cell, i.e.

$$z(t_1) = z(t_0) + \int_{t_0}^{t_1} -\nabla\psi(z(t))dt + \int_{t_0}^{t_1} g(z(t))dW(t)$$

Sinkhorn Divergence.

Fitting a model that successfully reconstructs cellular populations from progenitors requires a metric to compare simulated cellular populations to those observed. For this we utilize the metric of Sinkhorn divergence, an unbiased entropically regularized Wasserstein distance [25].

Consider two discrete sets of cells, $Z_{obs} = \{z_1^{obs}, z_2^{obs}, \ldots z_n^{obs}\}$ and $Z_{sim} = \{z_1^{sim}, z_2^{sim}, \ldots z_m^{sim}\}$ where each z_i^{obs}, z_j^{sim} is in R^G. We define two discrete probability measures on R^G, $\mu_{obs} = (Z_{obs}, w_{obs})$ and $\upsilon_{sim} = (Z_{sim}, w_{sim})$ where w_{obs}, w_{sim} are non-negative weight vectors on the standard simplex Δ^G, i.e. for the observed cells weights, $\sum_{i=1}^{n} w_i^{obs} = 1$ and $w_i^{obs} \geq 0 \; \forall \; i \in \{0, 1, \ldots n\}$.

The entropically regularized Wasserstein distance between μ_{obs} and υ_{sim}, with a squared Euclidean cost can be written,

$$W_\lambda(\mu_{obs}, \upsilon_{sim}) = \sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \frac{1}{2}\|z_i^{obs} - z_j^{sim}\|_2^2 - \frac{1}{\lambda}(\log log\left(P_{ij}\right) - 1)$$

Where $\|z_i^{obs} - z_j^{sim}\|_2^2$ is the squared Euclidean distance between a simulated and observed cell in R^G and λ is the strength of entropic regularization. P is a transport plan where each element P_{ij} represents mass transported from point z_i^{obs} to z_j^{sim}.

The marginals of the transport plan must be equal to the measures μ_{obs} and υ_{sim}, i.e. the matrix P is subject to the row sum and column sum constraints imposed by our weights w_{obs} and w_{sim}, i.e.

$$\sum_{j=1}^{m} P_{ij} = w_i^{obs} \; \forall\, i \in \{1,\ 2,\ \ldots n\} \text{ and } \sum_{i=1}^{n} P_{ij} = w_j^{obs} \; \forall\, j \in \{1,\ 2,\ \ldots m\}$$

For some regularization strength λ, the Sinkhorn divergence between μ_{obs} and υ_{sim} can be written,

$$S_\lambda\left(\mu_{obs},\ \upsilon_{sim}\right) = W_\lambda\left(\mu_{obs},\ \upsilon_{sim}\right) - \frac{1}{2} W_\lambda\left(\mu_{obs},\ \mu_{obs}\right) - \frac{1}{2} W_\lambda\left(\upsilon_{sim},\ \upsilon_{sim}\right)$$

Where W_λ is the entropically regularized Wasserstein distance. The inclusion of self-transport terms forces the Sinkhorn divergence to be equal to zero when $\mu_{obs} = \upsilon_{sim}$.

Model fitting.

The goal of our model training is to learn a parameterization of $\psi(z)$ and $g(z)$ that captures dynamics that simulate an observed dataset Z_{obs} from a set of progenitor cells $Z_0 \in Z_{obs}$. We parameterize the scalar potential field $\psi(z)$ and the diffusion vector field $g(z)$ with deep neural networks (discuss more next section).

Simulation of the process in (3) via Euler's method can be done via the first order discretization:

$$z_{t+\Delta t} \approx z_t - \nabla\psi\left(z_t\right)\Delta t + g\left(z_t\right) W(\Delta t)$$

Where $\Delta t = \frac{t_{final}}{N_{steps}}$ given user-specified parameters t_{final}, N_{steps}. Allow $Z_0 \in Z_{obs}$ to be a user-defined subset of cells that are assumed to have been measured at $t = 0$, i.e. the initial population. For each training iteration, we randomly select N_{cells} from this population and forward integrate each cell's position to times $\{\Delta t,\ 2\Delta t \ldots t_{final}\}$ via the Euler discretization in (4). Allow $Z_{fwd,\,T}$ to be the resulting set of all N_{cells} forward integrated positions at time T from the initial population. Similarly, $Z_{obs,\,T}$ is the set of transcriptional profiles for the cells observed at time point T.

We utilize the same growth weights that are computed in Yeo et al. to arrive at our weights $w_{obs,\,T}$ and $w_{sim,\,T}$. We now form the probability measures $\mu_{obs,T} = (Z_{obs,T},\ w_{obs,T})$ and $\upsilon_{sim} = (Z_{sim,T},\ w_{sim,T})$. For multiple time points $\{1,\ 2,\ \ldots T_{final}\}$, we calculate our overall cross-sectional loss as

$$loss = \sum_{T=1}^{T_{final}} S_\lambda\left(\mu_{obs,T},\ \upsilon_{sim,\,T}\right)$$

We aim to find parameters for $\psi(z)$ and $g(z)$ that minimize this loss function. We do so by implementing our forward integration approach with a Sinkhorn divergence loss in PyTorch optimizer. We train the model by optimizing weights of $\psi(z)$ and $g(z)$ via the Adam optimizer.

Benchmark task one: fate prediction.

scDiffEq. We fit scDiffEq over 250 epochs, exposing the model to cells from Well 0 (day 2) and Well 1 (day 4 and day 6), using 90% of that data for training and 10% for validation, reshuffling that split at every epoch. . Evaluation was performed by predicting the clonal fate bias of cells in Well 2 (day 4 and day 6) from the corresponding clonal progenitors in Well 0 (day 2).

PRESCIENT. We fit PRESCIENT, using their recommended parameters, as described in their Yeo et al., 2021. Briefly, using cells from Well 0 (day 2) and Well 1 (day 4 and day 6), PRESCIENT was trained over 2500 epochs. Evaluation was

performed by predicting the clonal fate bias of cells in Well 2 (day 4 and day 6) from the corresponding clonal progenitors in Well 0 (day 2). Predicted cell states at day 6 were labeled using the pre-fit nearest neighbor classifier.

Population balance analysis (PBA). To predict fate bias with PBA, we re-implemented the original PBA algorithm in PyTorch to provide a significant boost to the speed of downstream analyses. Following the procedure described in Weinreb, et al., we independently sampled 20,000 cells from the training set over five seeds. PBA requires designating "source" and "sink" points in the cell manifold. Undifferentiated cells were designated as the "source". To designate the "sink" points, we first pre-computed the potential values for each cell in the training set. For each cell fate, the cell with the minimum value of potential and its 20 nearest neighbors were designated as "sink" points. S = 10 was used as the corresponding "S" parameter for each sink point cell. Undifferentiated cells were assigned R = 0.2, while fated "sink" cells were assigned R = -0.2. Cells not designated as source or sink were assigned S = 0 and R = -1.0e-03. PBA next computes fate bias directly, producing a cell x fate bias matrix. We set the stochasticity, D = 1.0. Each cell in the training set was assigned a fate bias. To evaluate progenitor cells in the test set, we used a nearest neighbor graph built from the progenitor cells in the training set, subsequently mapping (using k = 20 neighbors) the fate biases generated for the cells in the training set to those in the test set.

Dynamo. We adapted the author-published tutorial for scRNA-seq. Briefly, this approach uses PCA-transformed embeddings as input and relies on the "stochastic" model for computing expression dynamics. Importantly, the function, `dyn.tl.gene_wise_confidence` was used to identify and filter genes whose velocity contradicts the "direction" of cell "movement" based on user-provided initial and final states.

CellRank. We adapted the author-published tutorial for running CellRank using the RNA velocity kernel, which relies on pre-computing RNA velocity values using the scVelo dynamical model. Per the tutorial, the "combined kernel" was created from a velocity kernel and connectivity kernel, weighted 0.80 and 0.20, respectively. *n* = 10 macrostates were used to compute the absorption probabilities towards terminal states / fates.

Labeling cells using approximate nearest neighbors. Following the procedure described by Yeo, et al., we used a nearest neighbors classifier to label simulated cells, based on their relative position with respect to the observed manifold. We used the annoy classifier from Spotify, fitting the model using all observed cells in the 50-dimension PCA space, using 10 trees, 20 neighbors, and euclidean distance.

Logistic regression. We used the scikit- learn implementation of logistic regression, fitting the model to the PCA representation of the training set cells and their tabulated fate biases. We then predict the fate biases of the PCA representation of cells from the test set. We then compute accuracy scores using sci-kit learn, for all evaluated cells.

Benchmark task two: recovery of a withheld time point.

Following the protocol outlined in Yeo, et,. al., 2021, briefly, we fit both scDiffEq and PRESCIENT to the subset of data in days 2 and 6 for which lineage information was recorded. scDiffEq was trained for 100 epochs. PRESCIENT models were trained for 2500 epochs. Evaluation of each model was performed by sampling 10,000 cells from day 2, with replacement, weighted by the empirically derived rate of cell proliferation and simulating one trajectory per cell. We then computed the Wasserstein distance via Sinkhorn divergence between the simulated cell populations at both day 4 and day 6 against the observed cell populations at these respective time points. PRESCIENT was evaluated at every 100 epochs. scDiffEq was evaluated at every epoch. For both models, we report the test error for the epoch with the minimum training error.

Analysis task three: *In silico* gene perturbations.

Following the protocol outlined in Yeo, et,. al., 2021, briefly, we introduced perturbations to 200 cells sampled from Day 2 of the LARRY dataset. For each gene perturbed, we set the z-scored expression to the indicated target value and re-transformed the scaled expression matrix using the pre-fit PCA model (to the original data). These perturbed cells were used as input to the scDiffEq model for predicting future states. We then label the model-predicted latent states using a nearest neighbor classifier fit to the original dataset. We used Welch's independent two-sided t-test to assess significance

in comparing the fractions of neutrophil and monocyte fates across conditions (perturbed vs. unperturbed). Each presented comparison was performed over 10 seeds.

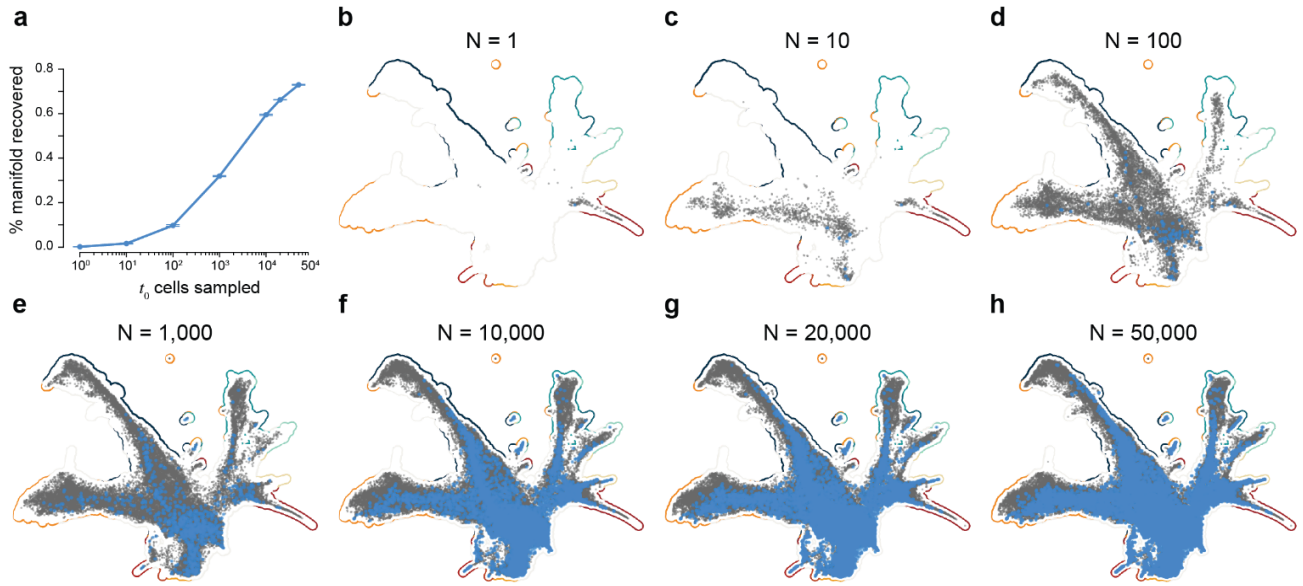
Author contributions

All authors contributed to the conceptualization of the methodology, experiments, and analyses. M.E.V. performed the M.E.V., R.L. and A.R. performed the fate-prediction benchmarking experiments. M.E.V. performed the time point interpolation experiments. M.E.V. performed the gene perturbation experiments. M.E.V. and R.L. performed the investigation of model attributes. M.E.V. wrote the software package. All authors wrote the manuscript. G.G. and L.P. provided supervision and guidance in designing the experimental strategy and funded the research.

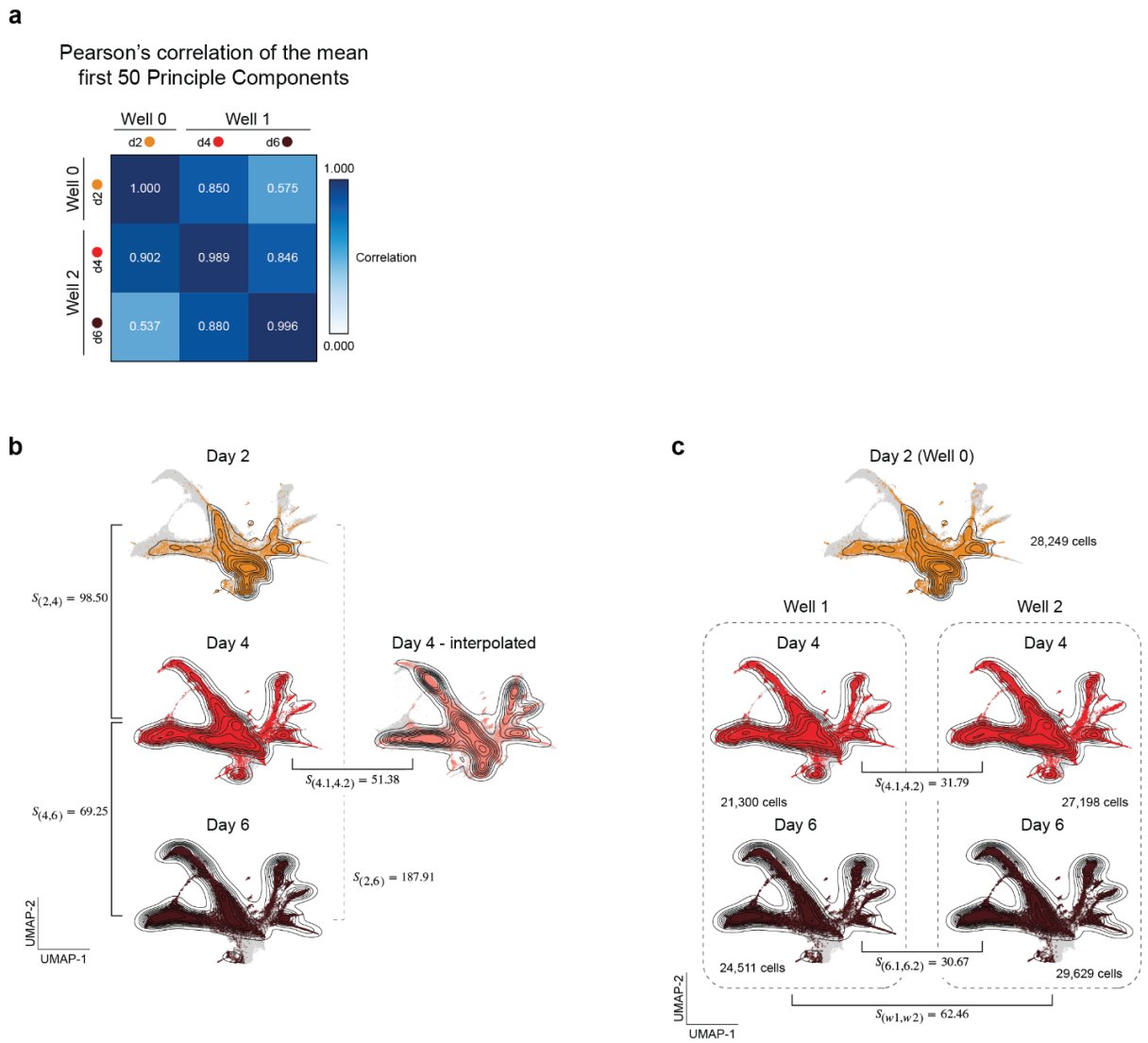
Acknowledgements

We graciously thank Caleb Weinreb and Allon Klein for their assistance in using the dataset produced in *Weinreb*, et al. 2020. We would like to thank Sachit Saksena, Grace Hui Ting Yeo, and David K. Gifford for their assistance in creating a benchmark comparison to PRESCIENT (Yeo et al., 2021). We would like to thank the entire Pinello and Getz Labs for thoughtful feedback and discussion throughout the preparation of this manuscript. We thank Shankara Anand for pivotal discussions that guided us towards the useful implementation of neural differential equations. We thank Dongya Jia, Joseph Zhou, and Herbert Levine for useful conversations and assistance in simulating synthetic datasets for proof-of-concept experiments.

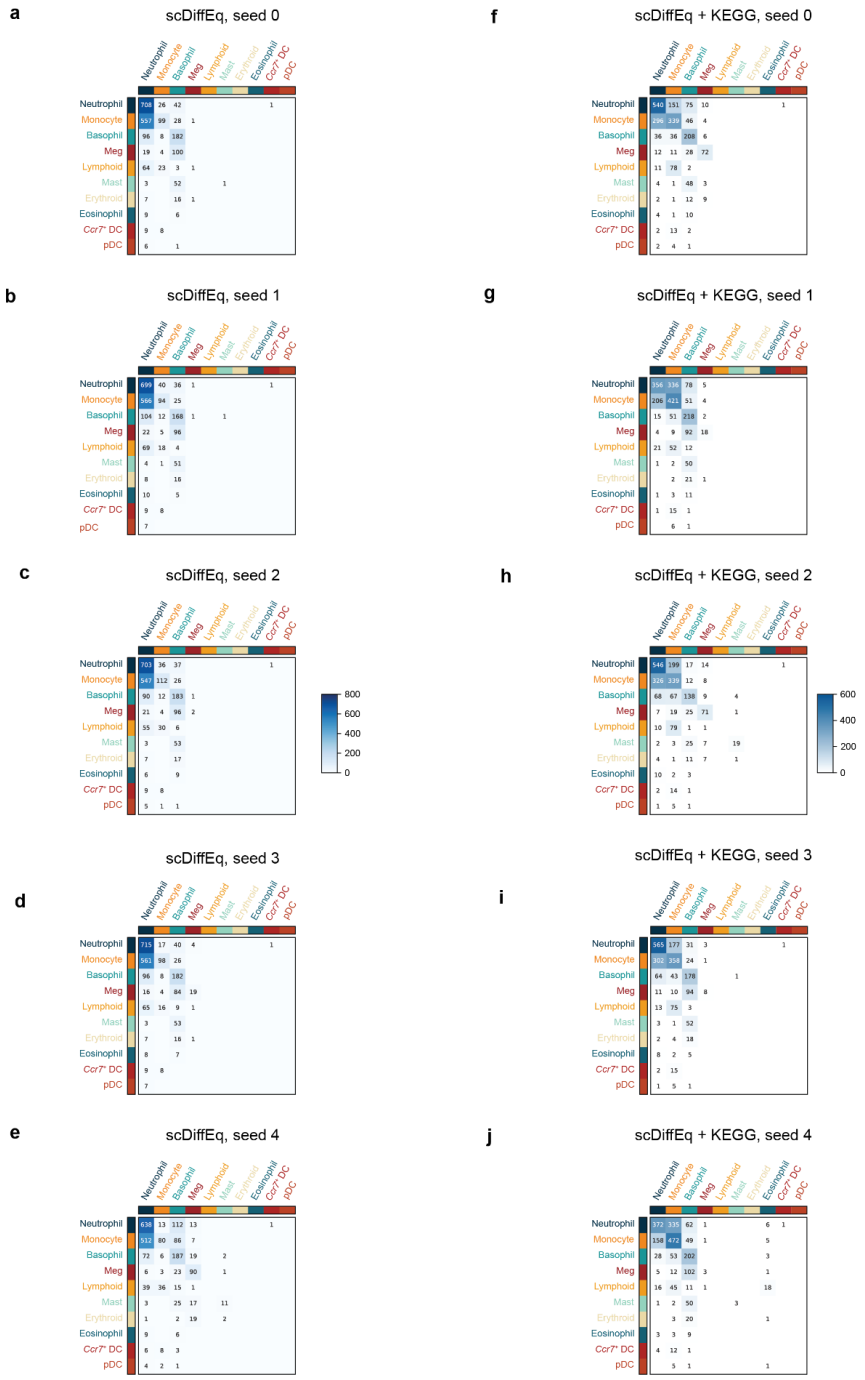
Supplementary Figures



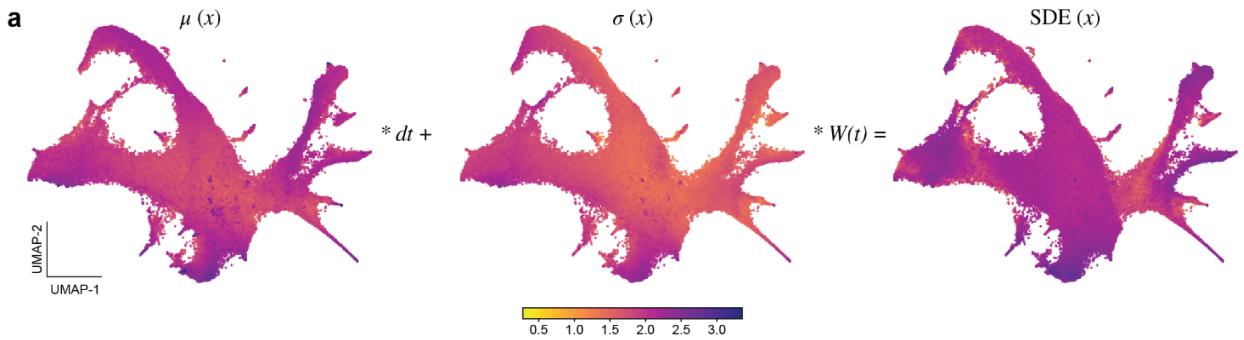
Supplementary Figure 1. Reconstruction of the observed manifold. **a**. Percent manifold recovery, using nearest neighbor (k=20) mapping. **b**-**h**. UMAP plots showing reconstruction of the original LARRY dataset manifold (background) using model simulations (gray) from 1, 10, 100, 1000, 10,000, 20,000, and 50,000. randomly sampled initial cells, highlighted in blue.



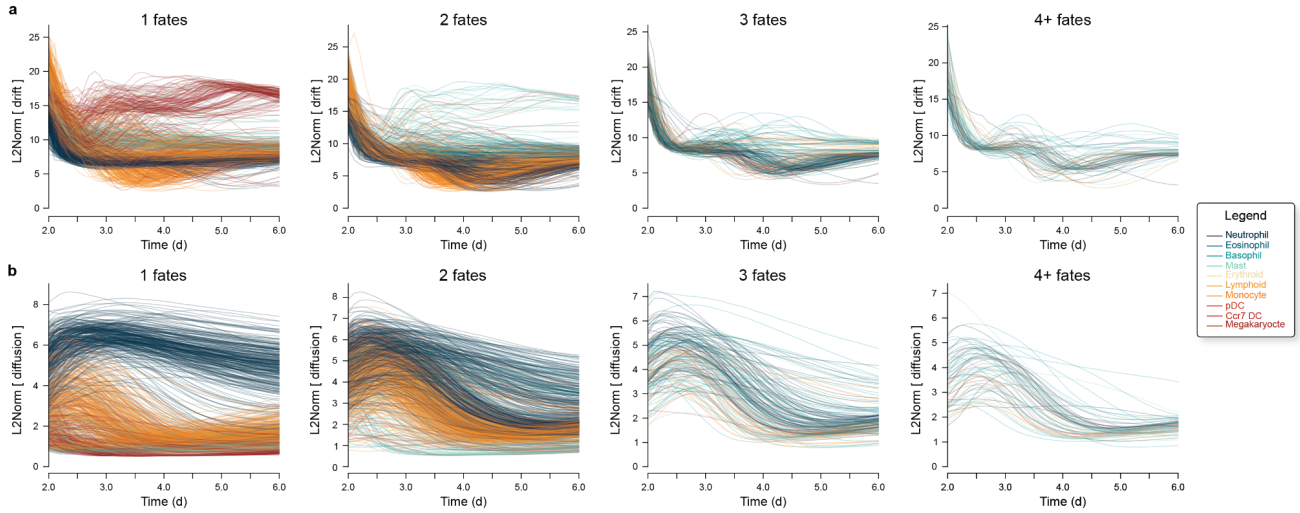
Supplementary Figure 2. Fate prediction (task 1) benchmark baselines. **a**. Pearson's correlation coefficient of the mean of the first 50 principal components, by well. **b**. UMAP plots highlighting time-specific sub-populations of the LARRY scRNA-seq dataset annotated with the Sinkhorn divergence between time points as well as a linear interpolation of the d4 distribution. **c**. Sinkhorn divergence between wells at each time point.



Supplementary Figure 3. scDiffEq fate prediction confusion matrices. **a**-**e**. scDiffEq without KEGG weights. **F**-**j**. scDiffEq with KEGG weights.



Supplementary Figure 4. Decomposed drift and diffusion from the learned neural SDE, applied to all cells of the LARRY dataset, superimposed on the structure of the learned neural SDE.



Supplementary Figure 5. Mean temporal L2Norm of model-predicted (**a**) drift and diffusion (**b**) diffusion, per simulated trajectory, grouped by fate multiplicity.