Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(8): 8505–8536.
DOI: 10.3934/mbe.2022395
Received: 21 February 2022
Revised: 21 April 2022
Accepted: 19 May 2022
Published: 10 June 2022

Research article

Comparison of cell state models derived from single-cell RNA sequencing data: graph versus multi-dimensional space

Heyrim Cho[1,4,*], **Ya-Huei Kuo**[2] **and Russell C. Rockne**[3,4,*]

[1] Department of Mathematics, University of California Riverside, Riverside, CA, USA

[2] Department of Hematologic Malignancies Translational Science, City of Hope, Duarte, CA, USA

[3] Department of Computational and Quantitative Medicine, Division of Mathematical Oncology, City of Hope, Duarte, CA, USA

[4] Interdisciplinary Center for Quantitative Modeling in Biology, University of California Riverside, Riverside, CA, USA

* **Correspondence:** Email: heyrim.cho@ucr.edu, rrockne@coh.org.

Abstract: Single-cell sequencing technologies have revolutionized molecular and cellular biology and stimulated the development of computational tools to analyze the data generated from these technology platforms. However, despite the recent explosion of computational analysis tools, relatively few mathematical models have been developed to utilize these data. Here we compare and contrast two cell state geometries for building mathematical models of cell state-transitions with single-cell RNA-sequencing data with hematopoeisis as a model system; (i) by using partial differential equations on a graph representing intermediate cell states between known cell types, and (ii) by using the equations on a multi-dimensional continuous cell state-space. As an application of our approach, we demonstrate how the calibrated models may be used to mathematically perturb normal hematopoeisis to simulate, predict, and study the emergence of novel cell states during the pathogenesis of acute myeloid leukemia. We particularly focus on comparing the strength and weakness of the graph model and multi-dimensional model.

Keywords: next generation sequencing data; cell state evolution; phenotype structured models; partial differential equation; hematopoeisis

1. Introduction

The ability to apply genome sequencing methods to single-cells has revolutionized biology [1]. Technologies enabling single-cell sequencing are advancing rapidly, with datasets as large as hundreds of thousands of cells are common [2]. RNA-sequencing is currently the most prevalent form of single-cell genomic analysis [1]. The sequencing of RNA at the cellular level enables the interrogation of gene transcription, which is used as a high-dimensional fingerprint which characterizes the identity of the cell. For this reason, single-cell RNA-sequencing data (scRNA-seq) has been used as a tool to study cell identity and state-transitions at the cellular level.

The most frequently studied cell state-transition is cellular differentiation; the process of a cell and its progeny to perform specialized tasks through transformation from a less differentiated stem-like state to a more differentiated state. ScRNA-seq is used to identify cells in various states of differentiation primarily through one or both of two primary methods: 1) clustering of cells with similar features [3], or 2) though trajectory inference (TI) [4]. Clustering analysis relies on the definition of a similarity metric, and may rely on a pre-defined number of clusters (e.g., k-means), or may use optimization methods to identify clusters (e.g., Leiden). There are a wide variety of clustering methods and similarity metrics to choose from, which may give drastically different results [5]. Similarly, trajectory inference methods may use pre-defined relationships between cells or may use optimization methods to identify these relationships to construct graphs which are then used to infer paths, or trajectories, between cell states. In addition, various approaches aim to characterize the cell fate landscape, for instance, by a parameterized landscape based on bifurcation analysis [6, 7], by using a measure of entropy of cell states: SCENT [8] and scEpath [9], or by mapping cells to a landscape on optimized parameters: HopLand [10] and Topslam [11].

A significant limitation of these approaches is if the graph structure and underlying relationships between the cells is unknown. As shown in a comprehensive review of trajectory inference methods by Saelens et al. (2019) [4], most TI algorithms have difficulty inferring even simple graphs which may include cycles or disconnected subgraphs. Because of the limitations of clustering and trajectory inference in analysing these data, we suggest that a hypothesis-driven and mathematical approach to the analysis of scRNA-seq data to study cell state transitions is warranted.

Moreover, single-cell genomic sequencing suffers from a number of challenges in analysis. Beyond the several choices to be made for even simple analyses such as clustering or visualization, the data may be sparse and incomplete. Gene "drop outs" and background signal (noise) can complicate differential expression and clustering analyses. For this reason, analysis of these data have remained fairly superficial despite the wealth of information contained in these high-dimensional datasets. Moreover, results obtained from analysis of single-cell sequencing datasets may be very sensitive to choices in the method of analysis and algorithm parameters. To date, single-cell sequencing data have not been effectively leveraged as inputs into mathematical models.

Here we compare two cell state geometries of cell state-transitions modeling with scRNA-seq data. Building on our prior work [12], we model cell differentiation as a continuous process. To elaborate this concept, when cell type-A becomes cell type-B, the cell states during the transition process are often classified into more steps as type-A½ or types-A¼, A²⁄₄, A¾. The continuous cell states can be considered as a limit of these states. We develop phenotype structured cell state models assuming continuous cell states using reaction-diffusion-advection partial differential equations (PDE) solved on: (i) an abstracted graph and (ii) a multi-dimensional continuum space. We compare and contrast these two cell state geometries with hematopoeisis as a model system. This manuscript is structured as follows: first we present the PDE model on a graph and in continuous space, then we apply the model to two datasets, see [13,14]. We examine the impact of various graph construction and trajectory inference methods on the geometry of the cell state space, and solve the model on these geometries. We then use the model the study the effects of perturbing 1) the graph structure 2) expression of select subsets of genes 3) and cell state transition dynamics by perturbing flow on the graph or by modifying the dynamics in the continuous space. We predict novel dynamics of leukemia pathogenesis by perturbing normal hematopoeisis and conclude with a comparative analysis of our approach and description of future directions for mathematical modeling with single-cell genomic sequencing data. A summary of our workflow is shown in Figure 1.

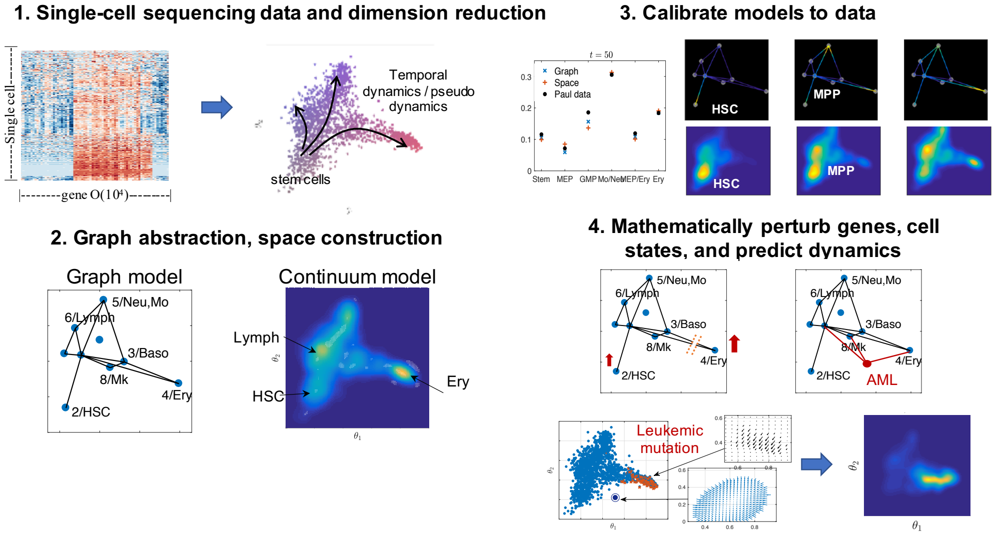
1. Single-cell sequencing data and dimension reduction
Single cell
gene O(10⁴)
Temporal dynamics / pseudo dynamics
stem cells
3. Calibrate models to data
t = 50
Graph
Space
Paul data
Stem MEP GMP Mo/Neu MEP/Ery Ery
HSC
MPP
2. Graph abstraction, space construction
Graph model
Continuum model
5/Neu,Mo
6/Lymph
3/Baso
8/Mk
4/Ery
2/HSC
Lymph
HSC
Ery
4. Mathematically perturb genes, cell states, and predict dynamics
AML
Leukemic mutation

Figure 1. Step-by-step illustration of our modeling process. 1. Processed single-cell sequencing expression matrices are represented in a reduced dimension space through one of many dimension reduction techniques such as diffusion mapping, PCA, t-SNE, or UMAP. 2. Cell clusters are inferred to construct the cell state geometry either the graph or multi-dimensional continuum of cell states. 3. From these representations, models are calibrated to the transport of cell distribution along the graph or in the cell state space. 4. The models can then be utilized to perturb genes and cell states. The calibrated models predict cell state-transitions and the emergence of novel cell states.

2. Materials and method

2.1. Modeling cell state-transitions in a continuous cell state-space

In this section we develop PDE models of cell dynamics in the continuous phenotype space identified by dimension reduction techniques. For a given single-cell genomic sequencing dataset

$$\{\boldsymbol{g}^i\}_{i=1}^N, \qquad \boldsymbol{g}^i = (g_1^i, g_2^i, ..., g_G^i) \in \mathbb{R}^G,$$

where N is the number of cells and $\boldsymbol{g}^i$ is a G-dimensional vector of gene expression of the i-th cell, the dimension reduction method can be written as an operator $\mathcal{P} : \mathbb{R}^G \to \Gamma \subset \mathbb{R}^n$ where the reduced dimensional space is truncated at the n-th dimension and $n \ll G$. We denote the reduced space variable as

$$\theta = \mathcal{P}(g) \in \Gamma \subset \mathbb{R}^n, \qquad \theta = (\theta_1, \theta_2, ..., \theta_n), \tag{2.1}$$

and the i-th single-cell data can be transformed into the reduced space as $\mathcal{P}(g^i) = \theta^i = (\theta^i_1, \theta^i_2, ..., \theta^i_n)$. Various dimension reduction techniques exist to construct a mapping $\mathcal{P}$, including principal component analysis (PCA), diffusion mapping, and t-distributed stochastic neighbor embedding (t-SNE). While different techniques provide different shapes and differentiation spaces, we choose diffusion mapping due to its ability to capture non-linear structure of high-dimensional data, and to well reproduce global trajectory of data [15]. We comment that if the reduced dimensional space is not clear to truncate at a low-dimension, one can consider low-dimensional marker genes according to the cell state of interest, and semi-supervised learning approach can be applied to obtain the low-dimensional reduced space. We also comment that it is common to remove the effect of cell cycles from the gene expression data to eliminate the cell state regarding their location along the cell cycle [16].

2.2. PDE model of cell state-transitions on a multi-dimensional reduced component space

We first develop a cell state model that describes the dynamics of cell distribution $u(t, \theta)$ on the reduced component space Γ, where $\theta \in \Gamma$ is the variable that represents continuous cell state. Three highly distinctive dynamic regimes of the cell states are considered, namely, directed cell transition, birth-death process, and random phenotypic instability. Such model can be written as an advection-reaction-diffusion PDE that governs the cell distribution $u(t, \theta)$ as

$$\partial_t u(t, \theta) = -\nabla \cdot (V(t, \theta)u(t, \theta)) + R(\theta, u(t, \theta)) + \nabla \cdot (D(\theta)\nabla u(t, \theta)), \tag{2.2}$$

with zero Dirichlet boundary condition. The three terms in our equation that involve parameterized functions V, R, and D, represent cell differentiation, population growth, and phenotypic instability, respectively.

Let us first describe the advection term $V \in \mathbb{R}^n$ that represents directed cell differentiation, where we propose two candidates for modeling V, denoted as $\mathbf{v}_1$ and $\mathbf{v}_2$. The first candidate $\mathbf{v}_1$ assumes an attractor cell states of homeostasis. Assuming that the magnitude of phenotypic instability is with a magnitude ν, that is, $D(\theta) = \nu$, one can compute the advection term $\mathbf{v}_1$ as

$$\mathbf{v}_1(\theta) = \nu\nabla_\theta U(\theta), \tag{2.3}$$

where $U(\theta)$ can be computed from the homeostasis distribution $u_s(\theta)$ that can be regarded as the cell landscape that the hematopoiesis system desires to maintain. As in the Boltzmann-like distribution from equilibrium statistical mechanics [17], we compute $U(\theta)$ as the exponent of $u_s(\theta)$ in the exponential form, in other words, $U(\theta) = -\ln(u_s(\theta))$. There are multiple methods to compute the cell landscape, so called quasi-potential, that focuses on relative stability of multiple attractors and models cell differentiation as transition between the attractor states [6–11]. Here, we compute the cell landscape empirically by assuming that the entire single-cell data is a representative subset of the entire hematopoiesis system, and by using density approximation methods. In particular, we use kernel density estimation [18] from the projected single-cell data $\theta^i \in \Gamma$, i.e., $u_s(\theta) = \frac{1}{N}\sum_{i=1}^{N} K_h(\theta - \theta^i)$ where we chose the standard normal density function as the kernel function K_h with bandwidth parameter $h > 0$.

The second candidate $\mathbf{v}_2$ models the dynamics of cell state transition. We model this term using a mechanistic approach that describes the symmetric and asymmetric cell division of stem cells to more differentiated cells. In particular, we consider the following form

$$\mathbf{v}_2(t, \theta) = \mathbf{c}(\theta)\,[2(1 - a(\theta))r(\theta)]\,s(t), \tag{2.4}$$

that is parameterized by the proliferation rate $r(\theta)$ and the self renewal rate $a(\theta)$ [19]. $\mathbf{c}(\theta) \in \mathbb{R}^n$ represents the direction and magnitude of differentiation on the phenotype space that we can estimate with either temporal data or pseudotime inference methods [4]. We note that the self renewal rate $a(\theta)$ is the proportion of cells that remains in cell state θ, while $1-a(\theta)$ cells further differentiate into matured states. This can be counted from symmetric and asymmetric stem cell division. In addition, we assume a signal parameter $s(t)$ that controls the active differentiation term, where $s(t) = 1/(1 + km(t))$ and $m(t)$ is the number of matured cells. Finally, we comment that the directed cell transition is simulated as $V = \mathbf{v}_1 + \mathbf{v}_2$, that is a sum of cell transition to attain homeostasis and active cell differentiation.

The reaction term represents the growth rate that consists of proliferation and apoptosis. We comment that the calibration of this term requires additional data to scRNA-seq , particularly, the population level growth data, to uniquely calibrate the model. It has been studied that the cell dynamics cannot be uniquely determined without imposing the reaction term [20]. More recently, there has been efforts to estimate the proliferation rate directly from scRNA-seq data by cellular barcoding techniques [21]. In our simulations, we cluster the single-cell data into biologically well known cell types, for instance, in case of hematopoiesis, myeloid progenitors, lymphoid progenitors, macrophages, and obtain the proliferation rate and self renewal rate from the literature. We consider the logistic growth term as following

$$R(\theta, u) = r(\theta)(1 - d(\theta, u))u, \tag{2.5}$$

where $r(\theta)$ is the proliferation rate and $d(\theta)$ is the apoptosis term assuming a logistic growth as $d(\theta, u) = \min\{\frac{u}{u_s(\theta)}, \bar{d}\}$, where $\bar{d}$ models the maximum magnitude of apoptosis rate.

The second-order diffusion term represents the instability on the phenotypic landscape of the cells that should be taken into account when modeling the macroscopic cell density. We simply consider a constant term $D(\theta) = \nu$. Assuming that the cell state trajectory is subject to Gaussian white noise, the diffusion coefficient can be estimated as the variance of the cell trajectory $\theta(t)$ on the reduced space, $\nu = \mathrm{Var}(\theta(t))/4$. However, since we do not have the data of cell trajectories, one can estimate the value as a limit of random walk as $\nu = (\Delta x)^2/(4\Delta t)$, assuming that Δx is the step size of the phenotypic fluctuation in Δt time [22]. See Appendix A for the detail of the model.

2.3. PDE model of cell state-transitions solved on a graph

Although the continuum-based multi-dimensional model provides a framework to study cell states, it is not always straightforward to map back locations in the space to novel or otherwise unknown cell states. Moreover, a central feature of contemporary analysis of scRNA-seq data is clustering and inferring relationships between clusters of known cell types [4]. Therefore, we develop a model that can describe cell state-transition dynamics on a graph that represents relationships between known cell types identified with clusters, extended from our previous work in [12]. An immediate advantage of this cell state geometry is that it is convenient to employ biological insights from well-known classical discrete cell states.

The continuum of differentiation cell states is assumed to be on the graph obtained from the scRNA-seq data, for instance, using partition-based graph abstraction (PAGA) algorithm [23]. We project the graph on the reduced component space, and denote the nodes as $\{v_k\}_{k=1}^{n_v}$ and the edges as e_{ij} connecting in the direction from the i-th to the j-th node. For convenience of notation, the edges are also denoted as $\{e_k\}_{k=1}^{n_e}$ with the index mapping $I : \mathcal{J} \to \{1, ..., n_e\}$ on the set of nontrivial edges $(i, j) \in \mathcal{J}$. The end points in the direction of cell transition are $\{a_k\}_{k=1}^{n_e}$ and $\{b_k\}_{k=1}^{n_e}$, where $\bigcup_{k=1}^{n_e}\{a_k,\ b_k\} = \{v_k\}_{k=1}^{n_v}$.

The model follows the dynamics of the cell distribution on the graph, $u(x, t)$, where $x \in e_k$ is a one-dimensional variable that parameterizes the differentiation continuum space location along the edges. We annotate the cell distribution on each edge e_k as $u_k(x, t)$ such that $u(x, t) = \{u_k(x, t)\}_{k=1}^{n_e}$, and model the cell density by an advection-reaction-diffusion equation [24] as

$$\frac{\partial u_k}{\partial t} \;=\; -\frac{\partial}{\partial x}\left(V_k(x)u_k\right) + R_k(x)u_k + \frac{\partial}{\partial x}(D_k(x)\frac{\partial u_k}{\partial x}), \quad x \in e_k = \overline{a_k\, b_k}. \tag{2.6}$$

The three terms are similarly modeled as the multi-dimensional model in Eq (2.2), representing cell differentiation, population growth, and phenotypic instability. To summarize once more, the advection coefficient $V_k(x)$ models the cell differentiation and the transition between the nodes, that is, different cell types. We model the advection term in two parts as in the reduced component space model, $V_k(x) = \mathrm{v}_{k,1}(x) + \mathrm{v}_{k,2}(x)$, and compute them as

$$\mathrm{v}_{k,1}(x) = \nu\partial_x U_k(x), \quad \mathrm{v}_{k,2}(t, x) = [2(1 - a_k(x))r_k(x)]\, s(t). \tag{2.7}$$

Here, $u_{s,k}(x) = e^{-U_k(x)}$ is the homeostasis cell distribution on the k-th edge, ν is the magnitude of phenotypic instability, $r_k(x)$ is the proliferation rate, $a_k(x)$ is the self-renewal rate, and $s(t)$ is the signal parameter. Cell proliferation and apoptosis can be modeled by the reaction coefficient $R_k(x)$ as

$$R_k(x, u) = r_k(x)(1 - d_k(x, u))u. \tag{2.8}$$

Finally, the diffusion term that represents phenotypic fluctuation is taken as $D_k(x) = \nu$.

In addition to the governing equation on the edges, the boundary condition at the nodes are critical when describing the dynamics on the graph. The boundary condition on the cell fate PDE model can be classified into three types, the initial nodes that do not have inflow $\mathrm{N}_I \doteq \{v_k \notin \bigcup_{j=1}^{n_e}\{b_j\},\ {\scriptstyle k=1,...,n_v}\}$, e.g., stem cells, the final nodes without outflow $\mathrm{N}_F \doteq \{v_k \notin \bigcup_{j=1}^{n_e}\{a_j\},\ {\scriptstyle k=1,...,n_v}\}$, e.g., the most differentiated cells, and the intermediate nodes, $\mathrm{N}_T \doteq \{\bigcup_{j=1}^{n_e}\{a_j\}\} \bigcap \{\bigcup_{j=1}^{n_e}\{b_j\}\}$. On the intermediate nodes $v_n \in \mathrm{N}_T$, mixed boundary condition is imposed for continuity of the density and flow as following,

$$\begin{aligned}\sum_{(i,n)\in\mathcal{J}} \mathcal{B}_{I[i,n]}(u, b_{I[i,n]}) &= \sum_{(n,j)\in\mathcal{J}} \mathcal{B}_{I[n,j]}(u, a_{I[n,j]}),\\ u(b_{I[i,n]}) = u(a_{I[n,j]}), &\quad \text{for all } (i, n) \in \mathcal{J},\ (n, j) \in \mathcal{J},\end{aligned} \tag{2.9}$$

where $\mathcal{B}_{I[i,j]}(u, x) \doteq V_{I[i,j]}(x)u(x) - D_{I[i,j]}(x)\frac{\partial}{\partial x}u(x)\Big|_{x_{I[i,j]}}$, $b_{I[i,n]}$ is the right end point of the edge between nodes i and n, and $a_{I[n,j]}$ is the left end point of the edge between nodes n and j. The cell outflow boundary conditions on the final nodes, $v_n \in \mathrm{N}_F$, are imposed as reflecting boundary conditions

$$\sum_{(i,n)\in\mathcal{J}} \mathcal{B}_{I[i,n]}(u, b_{I[i,n]}) = 0,$$

and $u(b_{I[i,n]}) = u(b_{I[j,n]})$ for all (i, n) and (j, n) in $\mathcal{J}$, and similarly on the initial nodes $v_n \in \mathrm{N}_I$.

2.4. Quantification of cell state-transition dynamics

Let us define some useful quantities to interpret model predictions in the multi-dimensional cell state-space and on a graph. The total number of cells from the cell distribution on either a graph or a continuous manifold can be computed as

$$\rho(t) \doteq \sum_k \int_{e_k} u_k(t, x)dx, \qquad \rho(t) \doteq \int_\Gamma u(t, \theta)d\theta, \tag{2.10}$$

respectively. In addition, we compute the number of cells of a specific cell type by assigning a weight, w_k, that corresponds to cells in the k-th cluster as

$$\rho_k(t) \doteq \int_\Gamma u(t, \theta)w_k(\theta)d\theta, \tag{2.11}$$

with $\sum_k w_k(\theta) = 1$. We assign weights based on the relative cell density of each clusters estimated with kernel density estimation. In the graph model, we assign the cell states along the edge to be the cell type of the closest node, so that we can compute the number of the k-th node cell type as $\rho_k(t) \doteq \sum_{m=I(k,j)} \int_{a_m}^{\frac{a_m+b_m}{2}} u_m(t, x)dx + \sum_{m=I(i,k)} \int_{\frac{a_m+b_m}{2}}^{b_m} u_m(t, x)dx$.

Although we can understand the continuum of cell states by mapping cells in intermediate states back to known discrete cell types, we also desire to interpret the continuous cell states in their location without reference to the canonical cell identities. For such purpose, we characterize cell states by identifying genes that are strongly correlated to a location in the space or moving in a direction toward a cell state. This extends finding the genes that are correlated to specific reduced space components to analyze the reduced cell state space [25]. First, to characterize the cell state θ^* in the reduced space, we use a function $f_{\theta^*}(\theta)$ centered at θ^* as $f_{\theta^*}(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\|\theta - \theta^*\|^2/2\sigma^2\right]$, and compute the correlation between the function values and the j-th gene expression levels as

$$r_{f,j} \doteq \mathrm{corr}(\boldsymbol{f}, \boldsymbol{g}_j), \tag{2.12}$$

where $\boldsymbol{f}$ represents the vector of function evaluation at each single-cell data point θ^i, that is, $\boldsymbol{f} = \{f_{\theta^*}(\theta^i)\}_{i=1}^N$, and $\boldsymbol{g}_j = \{g_j^i\}_{i=1}^N$. An alternate quantity to examine is the genes that are related to a certain direction in the reduced component space. For instance, the correlation between the j-th gene and the k-th reduced component $\theta_k = \{\theta_k^i\}_{i=1}^N$ and to a certain vector $\boldsymbol{v} = \{v_k\}_{k=1}^n$ can be computed as

$$r_{k,j} \doteq \mathrm{corr}(\theta_k, \boldsymbol{g}_j), \qquad r_{\boldsymbol{v},j} \doteq \sum_k v_k\, r_{k,j}, \tag{2.13}$$

respectively. Regarding Eq (2.13) as global quantities, we can also compute the local correlation on the subdomain of the reduced space Ω_d by collecting the cell indices that lie within the subdomain $\Gamma_d = \{i\,|\,\theta^i \in \Omega_d\}$, that is, $r_{k,j}|_{\Gamma_d} \doteq \mathrm{corr}((\theta_k, \boldsymbol{g}_j)|_{i\in\Gamma_d})$ and $r_{\boldsymbol{v},j}|_{\Gamma_d} \doteq \sum_k v_k\, \rho_{k,j}|_{\Gamma_d}$. Although these metrics may provide candidates of potential genes that are related to the cell state to be analyzed, we emphasize that these need to be verified experimentally by observing the cell state change after perturbing the genes. See Section 4.2 for the limitations of this approach.

3. Simulation of continuous cell state models on multi-dimensional space versus graph

In this section, we employ the framework developed in the previous section to the mouse hematopoiesis cell data from Nestorowa et al. (2016) [13] and Paul et al. (2015) [14]. We obtain the graph and multi-dimensional space models of hematopoiesis cell state and focus on comparing the strengths and weaknesses of the two models.

The hematopoiesis single-cell data from [13, 14] projected on the first two diffusion component space are shown in Figure 2A, where distinct cell types, including lymphoid-primed multipotent progenitors (Lymph); common myeloid progenitors (CMP); megakaryocyte-erythroid progenitors (MEP); granulocyte-macrophage progenitors (GMP); erythrocytes (Ery); neutrophils (Neu); monocytes (Mo); megakaryocytes (Mk); basophils (Baso), classified in the original papers are illustrated with different colors. We truncate the diffusion component at two since the reduced two-dimensional space describes the dynamics of our interest, that is, from strong to weak stemness. The first two diffusion components θ_1 and θ_2 represent cell maturation in both data sets. In Nestorowa data, the first diffusion component separates the stem cells to myeloid lineages, particularly MEP cells and the second diffusion component describes GMP cells and the lymphoid lineages. In Paul data, the first and second reduced component represents MEP and GMP lineage, respectively. We remark that the cells that are the most stem-like in Paul data are CMPs, that is more matured than the ones in Nestorowa data, that includes the long-term and short-term HSCs. In addition to the single-cell data, the Figure 2B presents the abstracted graphs obtained using PAGA [23]. Further refinements of the graph will eventually become the full single-cell data, where each single-cell being counted as distinct cell states, and it depicts the hierarchy of cell states (see Figure A5).

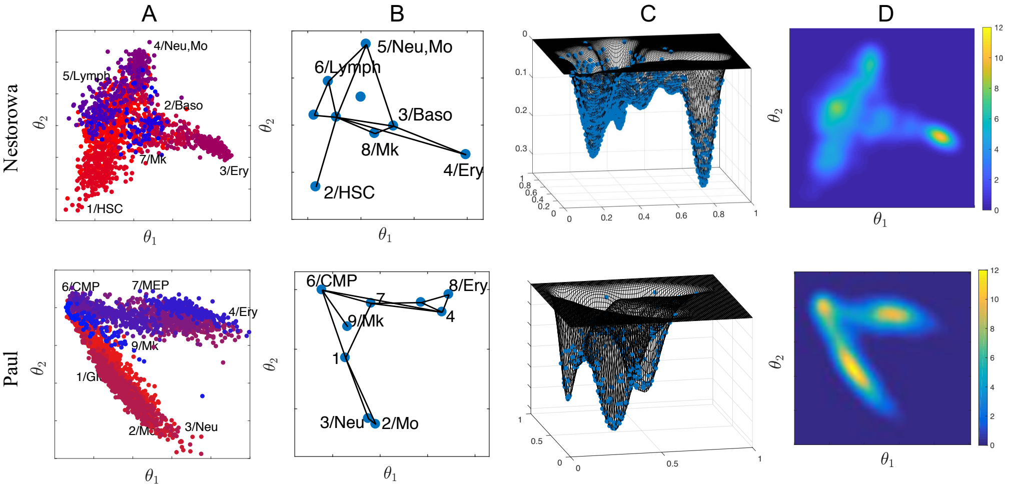

Figure 2. From discrete to continuum cell states. A) Single-cell data from Nestorowa et al. (2016) and Paul et al. (2015) projected on the first two diffusion component space. B) Graph obtained by PAGA algorithm projected on the diffusion component space. Distinct cell types classified in the original paper are either illustrated with different colors (A) or annotated on the graph nodes (B). C, D) Multi-dimensional continuum cell state distribution on the diffusion component space computed by kernel density estimation. They are used as homeostasis distribution.

The homeostatic cell distribution u_s on the reduced dimensional space is computed by kernel density approximation [18, 26]. The computed cell landscapes viewed in different angles are shown in Figure 2C,D. The cell distribution on the graph can be similarly obtained after reallocating the cells to the node, that is, the center of each cluster. The cell distribution on the continuum space provides an intuitive method to compare the relative concentration of different cell lineages, including the intermediate cell states. We observe high concentration of MEP and Ery cells that are localized at the far right (large θ_1) in both data. The Nestorowa data has more diverse cells including the common lymphoid progenitors that are visible on the left (small θ_1, and intermediate θ_2), while the Paul data has evenly distributed cell states among the most stem-like cells (CMP) and the two different lineages.

Let us summarize the properties of the graph and multi-dimensional space models before we present simulation results. The graph model has its strength that the cell lineages between the known cell states can be more easily identified as compared to the multi-dimensional space model. The cell concentration moving toward different edges can be clearly distinguished as the cell lineages to different cell states. Accordingly, counting the number of cells in each discrete cell state is more straightforward, for instance, by integrating the cell distribution along the edges half way. Although the multi-dimensional space model has ambiguity on classifying the cells into discrete cell types, the number of cells in each discrete cell type can be computed by assigning weights to integrate as in Eq (2.11). Moreover, we emphasize that the advantage of clear cell states in the graph model is also its limitation at the same time, since it restricts the model to only study the known cell types and lineages. The advantage of the multi-dimensional space model is its potential of exploring novel cell states that deviates from known cell types. While the graph model cannot explore the cell states that are not already included in the graph structure, the multi-dimensional space model can immediately study the abnormal trajectories and emergence of cells at any space location. We will show later in our simulation that the hypothesis of genetic alterations can be studied directly in the multi-dimensional space model, without projecting it on the graph structure. Moreover, the multi-dimensional space model is more sensitive to genetic variations than the graph model, although when the variation is large and a considerably distinct cell state arises, the graph model can append another cluster node. See Table 1 for a summary.

Table 1. Comparison of the cell state model on graph versus multi-dimensional space in n dimensions. The computational cost is estimated by denoting M as the number of discretized grid points in one-dimension. We comment that the computational cost of a PDE solver can be up to a third power of the degree of freedom.

	Graph model	Multi-dimensional model
Cell state Interpretation	Comprehensible as intermediate cell states	Difficult to interpret
Cell state Exploration	Limited to graph structure	Freedom to explore novel and unconventional cell states
Computational cost (Degree of freedom)	O(M)	O(M^n)

In the following sections, we consider mainly two application problems, namely, normal hematopoiesis and abnormal hematopoiesis differentiation, resulting in myeloid leukemia as application examples of our modeling approach.

3.1. Calibrating the mathematical models to normal hematopoiesis

We demonstrate that normal hematopoiesis can be visualized by both models on the graph and on the space of two-dimensional diffusion components, (see Figures A1 and A2 for the advection and reaction terms used in the multi-dimensional space model). Figure 3 shows a cluster of stem cells differentiating into the entire cell states on the graph and reduced space using Nestorowa data [13] and Paul data [14]. The initial condition is imposed as approximately 10% of cell capacity in normal condition mostly composed with stem cells. On both graph and multi-dimensional space models, nontrivial amount of most matured cell states, particularly, Ery and Neu/Mo cells arise around pesudotime $t = 30$, and further recovers the full landscape after approximately $t = 50$. In particular, the observation that the matured cells quickly proliferate to fill in the space agrees in both simulations from Nestorowa and Paul, while the effect is more significant in Paul's data due to shorter distances of the matured cell states from the initially administered cells.

The advantage of the graph model is apparent that we can observe distinct cell states as a mass of cells shifting from a node to distinct edges toward different cell states. For instance, the cells differentiating from the MPP cluster to either Neu/Mo lineage and Ery lineage can be clearly separated in the graph models, while those can be ambiguous in the two-dimensional distribution. Still, we can compute the number of cells in each cell types in both models as shown in Figure 3B. We observe that the number of cells reaches the maximal capacity at later times around $t = 100$, with the intended ratio of cell numbers in each discrete cell type approximating the given data [13] in Figure 3C. The recovery is more rapid for larger values of v and larger number of initial stem cells $\rho(0)$ (see Figure A6).We remark that the continuous cell states of hematopoeisis is also depicted in conventional flow cytometry which is typically used to identify distinct cell populations based on expression of cell surface markers. We performed Fluorescence Activated Cell Sorting (FACS) analysis of bone marrow cells isolated from normal C57Bl/6 mice (age 6-8 weeks). Distinct myeloid progenitor types (CMP, GMP and MEP) are typically differentiated by the expression of CD16/32 and CD34 markers within the myeloid lineage progenitor cell compartment. Figure 3D shows representative FACS data with respect to CD16/32 and CD34 expression that is used to identify the CMP, GMP, and MEP cell types within the normal myeloid progenitor compartment. Although the FACS data is conventionally clustered and gated into three cell types, continuity of CD16/32 and CD34 expression can be observed that agrees with our graph abstraction and multi-dimensional cell state geometries.

3.2. Using the model framework to simulate acute myeloid leukemia (AML) pathogenesis and progression

In this section, we once more compare the graph and multi-dimensional space models with an application to abnormal differentiation under leukemia pathogenesis and progression. We first consider AML model in the context similar to the previous section that involves known progenitor cells that exemplifies the advantage of graph models. However, we will show how aberrant differentiation and phenotypic plasticity of leukemia pathogenesis motivates the spatial model.

AML results from aberrant differentiation and proliferation of transformed leukemia-initiating

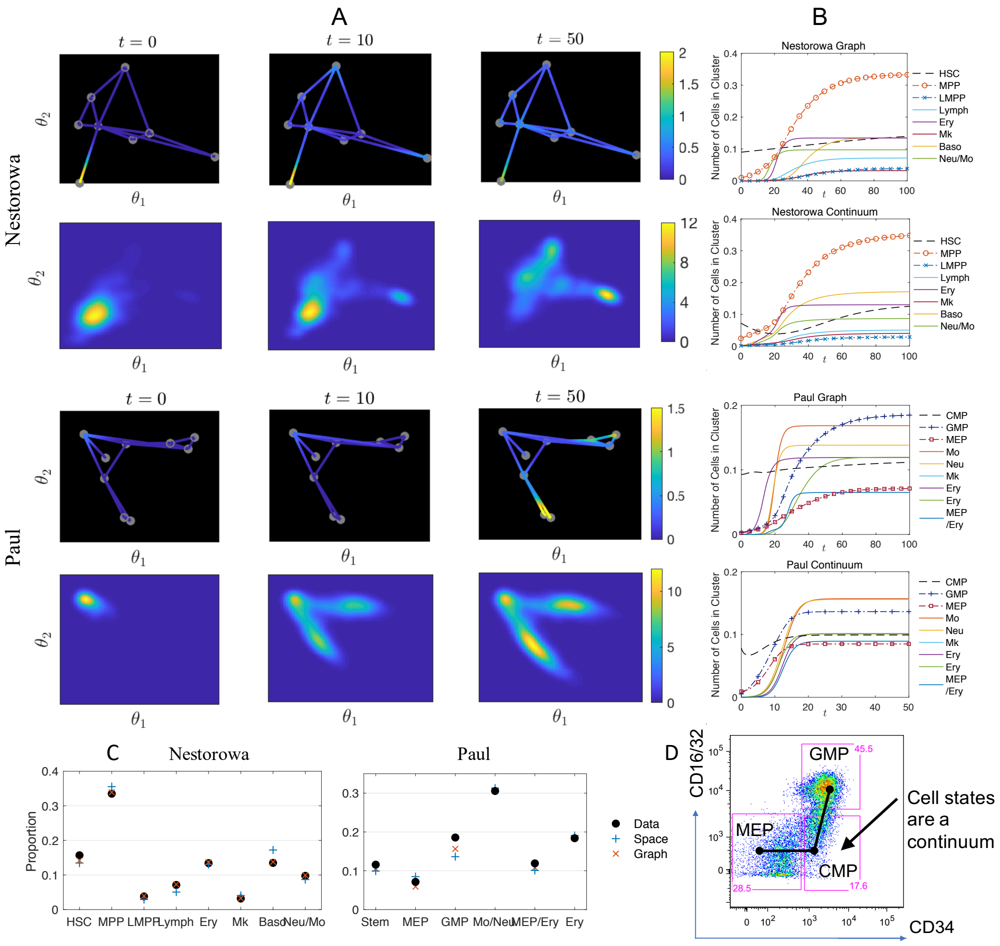

Figure 3. Dynamics of cell distribution during normal hematopoeisis. A) Evolution of cell state densities $u(t, x)$ on the graph with 8 to 9 nodes, and $u(t, \theta)$ on the diffusion component space during normal hematopoeisis using single-cell data from Nestorowa et al. (2016) and Paul et al. (2015). The shown dynamics is in pseudotime t. B) The pseudotime dynamics of the number of cells in each cell cluster, where the number is normalized so that the total cell number in equilibrium state is one. The initial stem cells differentiate to progenitors and more matured cell states and recover the entire cell landscape. C) Numbers of cells in each type/cluster using the multi-dimensional space model and the graph model are successfully calibrated to the observed data so that at $t = 100$ each model predicts the correct cell ratios to within ±5%. D) The continuous cell states of hematopoeisis is also depicted in the FACS data set collected from the normal mouse bone marrow. Bone marrow cells were gated for myeloid progenitor cell markers (lineage-negative, Sca1-negative, cKit-positive). Conventionally, the expression levels of CD16/32 and CD34 are used to distinguish CMP, GMP, and MEP cell types within the myeloid progenitor compartment, however, the continuity of marker expression agrees with our graph abstraction and multi-dimensional cell state geometries.

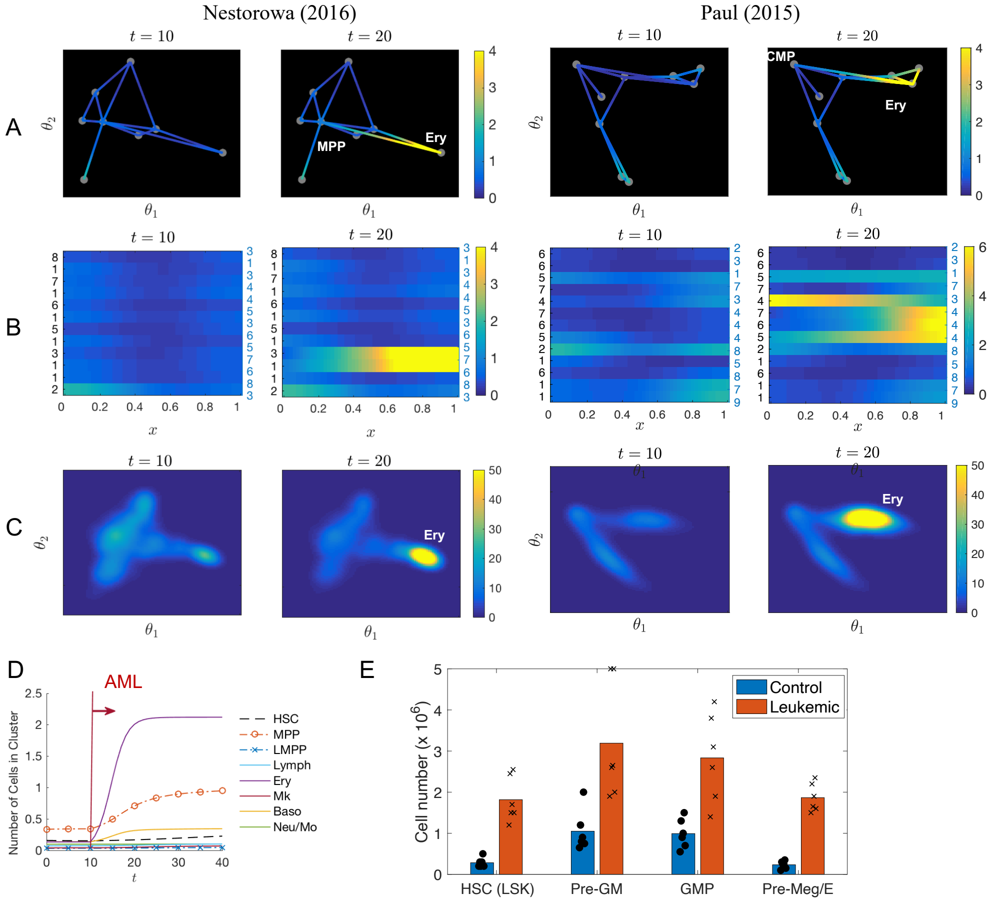

Figure 4. Predicting abnormal cell differentiation during leukemia progression. A,B,C) Cell distributions during leukemia pathogenesis and progression are shown on the graph model (A, B) and the multi-dimensional space model (C). Plot (B) shows an alternative way to plot the graph based model solution by stacking the cell distribution on each edge horizontally. The number on the left and right shows the node numbers shown in Figure 2B. They show the effect of over-proliferation and differentiation block in the myeloid lineages. In particular, we observe a rapid expansion of cell states near MEP and Ery in both Nestorowa and Paul data, after the initiation of AML at $t = 10$. D) The number of leukemic Ery cells show an increase within ten days in both graph and multi-dimensional space models. E) Experimental result reproduced from [27] that shows a rapid expansion of pre-Meg/E (MEP) population in leukemic mouse compared to normal mouse (Control).

cells and abnormal progenitor cells. We model AML pathogenesis based on known behavior of a genetic Cbfb-MYH11 (CM) knock-in mouse model that recapitulates somatic acquisition of a chromosomal rearrangement, inv(16)(p13q22) [28, 29], commonly found in approximately 12 percent of AML cases. Inv(16) rearrangement results in expression of a leukemogenic fusion protein CBFβ-SMMHC, which impairs differentiation of multiple hematopoietic lineages at various stages [30–32]. Most notable in such leukemia pathogenesis and progression is the increased in abnormal myeloid progenitors, which has an MEP-like immunophenotype and a CMP-like differentiation potential [31]. Experimental studies [27, 33] show that such MEP attains a predominant increase in pre-megakaryocyte/erythroid (Pre-Meg/Ery) population (ranging from 5 to 12 fold) accompanied by impaired erythroid lineage differentiation as shown in Figure 4E. This refined phenotypic Pre-Meg/Ery population consists partly of the CMP and MEP populations which are identified using conventional markers [13, 34].

In our model, abnormal leukemic progenitors are regarded as intermediate cell states along the edges connecting CMP (or MPP) and MEP (and Ery) in the graph model, and the corresponding locations in the multi-dimensional space model. We assume a 10-fold increase on average in those population by lowering $d(\theta)$ and $d_k(x)$ in Eqs (2.5) and (2.8) that controls the local cell capacity. In addition to over-proliferation, another aspect of the leukemia pathogenesis of our interest is the impaired differentiation of erythroid lineage differentiation, where it can be modeled by blocking the cell differentiation $V(\theta)$ in Eq (2.6) and $V_k(x)$ in Eq (2.2) toward Ery.

The corresponding results are shown in Figure 4, where we modify the model to leukemia progression at $t = 10$. The cell distribution changes from the normal cell landscape at $t = 10$ to an increased population of Ery (MEP) and nearby cells at $t = 20$ in both graph and continuum models. We observe a 10-fold increase in the Ery (MEP) population, which includes the abnormal myeloid progenitors, in both graph and multi-dimensional space model across the data sets as shown in Figure 4D,E. The rapid emergence of AML occurs within two week period, corresponding to the expansion of the leukemic cell phenotype. We observe a rise in the MPP or CMP cluster as shown in the results from Nestorowa data, that is similar in Paul data as well. The total proportion of leukemic cells comprise 50–60% of the total population.

In the leukemia pathogenesis simulation in this section, we focus on studying the leukemic cells as a variation of cell states classified using conventional markers. In this case, the graph model can interpret and include the dynamics of such cells, as well as the multi-dimensional space model. While the simulation outcome between the graph and multi-dimensional space model is similar, the graph model is computationally more efficient due to the fewer number of unknowns as compared to the two-dimensional space model. However, to study the abnormal cell states that may appear far away from the known or existing landscape, we will show in the following section that the multi-dimensional space model has more freedom to include those new cell states and disrupted trajectories. We will study the impact of perturbation of genes in the graph and multi-dimensional space model, particularly focusing on alterations of genes known to be involved in leukemia pathogenesis.

3.3. *In silico experiments of gene expression perturbation*

In this section we investigate the sensitivity of altering specific genes in a prescribed manner and the impact of this perturbation in the graph and multi-dimensional space models. We keep our focus on leukemia pathogenesis and progression and consider alterations of 38 genes that are reported to be related to leukemia stem cells [35, 36], although we emphasize that these genes serve simply as examples, and are not intended to model the precise biological process of AML pathogenesis. The j-th gene expression level of i cell, g^i_j, is modified as

$$\widetilde{g}^i_j = 2^{\gamma_j} g^i_j, \qquad 0 \le \log_2(\widetilde{g}^i_j + 1) \le 16, \tag{3.1}$$

where γ_j is the $\log_2$-fold change compared to normal cells. The full list of altered genes and magnitudes γ_j from [35, 36] are shown in A1. Details of the model equation and parameters, and the $\log_2$-fold change is in the range of $\gamma_j \in [-3.5, 2.7]$. In addition, we consider the extreme case of gene alteration as the maximum level $\log_2(\widetilde{g}^i_j + 1) = 16$ for up-regulated genes and $\log_2(\widetilde{g}^i_j + 1) = 0$ for down-regulated genes. Figure 5 shows examples of the gene expression levels in log scale that we modify including the up-regulated genes, GPR56, GATA2, and MZB1, and the down-regulated genes, LGALS3, LY86, and ANXA5. The given single-cell data in normal condition is plotted together with the hypothetically altered levels of gene expressions in regular leukemia pathogenesis and extreme levels of alteration. Although the case of extreme alteration is unrealistic, we consider such case to illustrate an example where the graph abstraction and dimension reduction algorithm clearly distinguishes the leukemic cells.

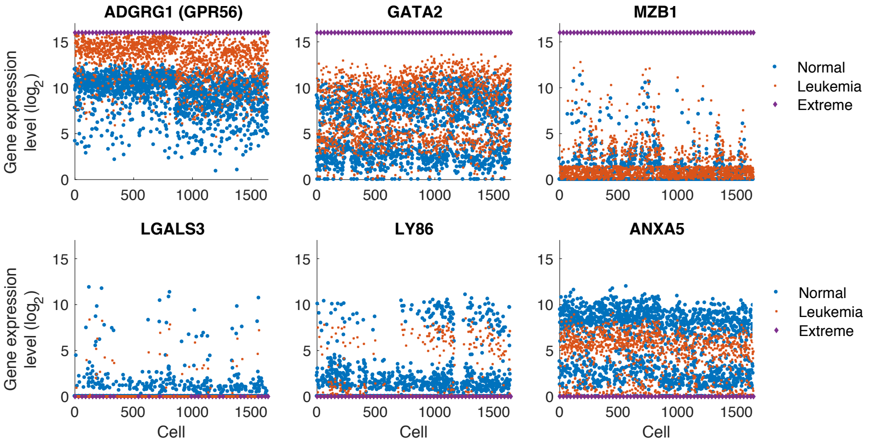

Figure 5. Perturbing genes associated with leukemia stem cells. Examples of expression levels of genes in $\log_2$ scale, that are associated with leukemia stem cells and pathogenesis, including up-regulated GPR56, GATA2, and MZB1, and down-regulated LGALS3, LY86, and ANXA5. We show these subset of genes simply to illustrate the process. The normal single-cell data $\log_2(g^i_j + 1)$ (blue circle) and modified gene expression $\log_2(\widetilde{g}^i_j + 1)$ (red square) computed as Eq (3.1) are shown together, with the case of extreme levels of either 16 or 0 (purple diamond).

3.4. *Effects of gene perturbation on graph abstraction and multi-dimensional reduced component space*

We first study the sensitivity of the reduced component space using diffusion mapping [15]. Figure 6A,B compares the altered leukemic single-cell data $\widetilde{g}^i$ projected on the normal reduced space (θ_1, θ_2). The left-most column shows the projected leukemic single-cell data $\mathcal{P}(\widetilde{g}^i)$ in the normal reduced space, where the leukemic cells are located toward the left-bottom compared to the normal cells in Nestorowa data, and upwards in Paul data. The effect of gene modification is shown more clearly in the presented vector field $\mathcal{P}(\widetilde{g}^i) - \mathcal{P}(g^i)$.

Similarly, we study the impact of leukemia-associated gene perturbation in graph abstraction using PAGA [23]. Figure 6C,D shows the clustered cell types and the corresponding graph using perturbed leukemic scRNA-seq data. The presented results are computed with Nestorowa data. The clustered cell types and leukemic cells are annotated to depict the cluster properties. In Figure 6C, which is the

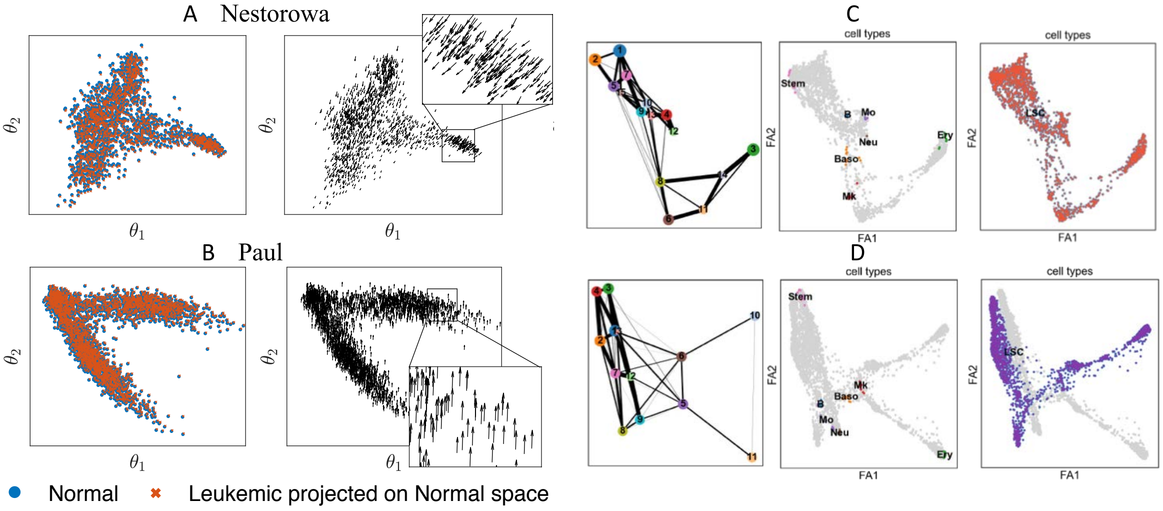

Figure 6. Effects of perturbing genes on cell state-space. A,B) Projection of perturbed leukemic single-cell data on the normal diffusion component space, and the directional vectors $\mathcal{P}(\widetilde{g^i}) - \mathcal{P}(g^i)$ representing the altered cell state by leukemic perturbation. The top figures are computed with Nestorowa data [13] (A) and the bottom figures with Paul data [14] (B). C,D) Graph computed from perturbed leukemic single-cell data and their cluster information. The annotation shows that the graph abstraction algorithm does not distinguish the perturbed leukemic cells in regular magnitude to the normal cells, so that the perturbed information is lost (C). However, when single-cell data is modified to the extreme values of gene expression level, the algorithm distinguishes the leukemic cells, although the data is unrealistic (D).

case of leukemia progression with single-cell data altered in regular magnitude, we observe that there is no cluster that separates the leukemic cells. Thus, the information of leukemic gene alteration is lost within the clustering algorithm, and the model on such abstracted graph is not capable of studying the perturbed cells. On the other hand, when the gene levels are modified to their extreme levels, the perturbed leukemic single-cell data are clustered into separate nodes as shown in Figure 6D. In this case, although the graph model is able to study the perturbed cells as separate nodes, we comment that this level of perturbation is an unrealistic scenario due to the extreme levels of gene expression.

A strength of the multi-dimensional cell state model is its capability of interpreting the perturbation of gene expression levels or new incoming cell data regardless of its relation to the primary data (Figure 6). As shown in our results, the leukemic alteration is successfully projected in the reduced space, while the abstracted graph lost the information. Although the projected directions in the reduced space can be once more projected on the graph, it does not make sense to do so when the direction is orthogonal to the edges. The multi-dimensional space model has its advantage especially in this case, where the projected direction of cell states can be directly implemented.

3.5. *Simulating AML pathogenesis by perturbing known leukemia-associated genes*

In this section, we incorporate the perturbed leukemia-associated gene data in the AML simulation using the multi-dimensional space model. In particular, we are interested in studying the impact of leukemia-associated gene alteration on the cell distribution during AML progression. We compute the abnormal differentiation of leukemic cells by projecting the altered single-cell data of MEP and Ery cells to the normal diffusion component space $\mathcal{P}(\widetilde{g}^i)$ as it is shown in Figure 7. The aberrant differentiation vector $\mathbf{v}^1_{aml} = \mathcal{P}(\widetilde{g}^i) - \mathcal{P}(g^i)$ shows the shifts of cells toward the location where no cell data occupies. Therefore, in addition to modifying the advection term according to the altered gene data, we assume an emergence of new abnormal cell state. In particular, we take the cell state location at $\theta^* = (0.610,\ 0.215)$ in Nestorowa data, and at $\theta^* = (0.6,\ 1.0)$ in Paul data, and use Gaussian functions centered at θ^* to obtain $\mathbf{v}^2_{aml}$. The corresponding vector fields are also shown in Figure 7.

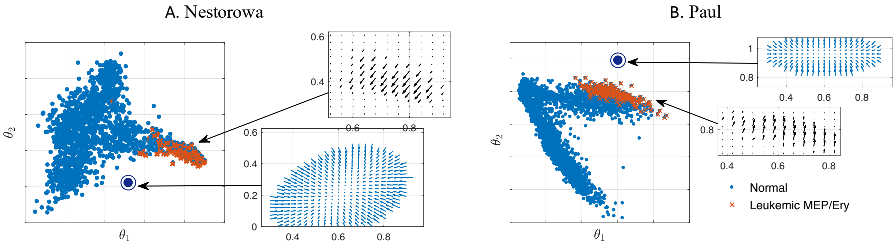

Figure 7. Modeling AML pathogenesis and progression by perturbing cell states directly in the cell state-space. The direction of abnormal cell differentiation $\mathbf{v}^1_{aml}$ (black) is computed from the projection of altered leukemic MEP and Ery cells (×) to the normal diffusion component space as $\mathcal{P}(\widetilde{g}) - \mathcal{P}(g)$. Alternatively, we assume a source of abnormal cell state ($\odot$) at $\theta^* = (0.610, 0.215)$ in Nestorowa data (A) and at $\theta^* = (0.6,\ 1)$ in Paul data (B) to model $\mathbf{v}^2_{aml}$ (blue).

For AML progression, the advection term is modeled with the prescribed vector field as $V = \mathbf{v}_1 + c_{aml}\mathbf{v}^1_{aml}$ or $V = \mathbf{v}_1 + c_{aml}\mathbf{v}^2_{aml}$, where c_{aml} parameterizes the perturbation magnitude. We further perturb the model by increasing the proliferation of the new leukemic cells at θ^* by appending $r_{aml} f_{\theta^*}(\theta)$ to R, where r_{aml} parameterizes the over-proliferation. The cell distribution $u(t, \theta)$ with $\mathbf{v}^1_{aml}$ and $\mathbf{v}^2_{aml}$ for different values of $c_{aml} = 1, 2, 10$ are presented in Figure 8 and Abnormal cell state transitions during leukemia pathogenesis and progression. The distribution of cell states $u(t, \theta)$ show abnormal cell states emerging during leukemic progression after $t = 10$ modeled in the advection term as $V = c_{aml}\mathbf{v}^1_{aml}$ (top) and $V = \mathbf{v}_1 + c_{aml}\mathbf{v}^2_{aml}$ (bottom), with various levels of $c_{aml} = 1, 2, 10$. Larger magnitude of c_{aml} results in more disrupted cell landscape. . In the cell landscape with $\mathbf{v}^1_{aml}$, we observe increased MEP cells and abnormal progenitors arising in the direction of left-bottom, especially for large values of c_{aml}. The model with $\mathbf{v}^2_{aml}$, a new cell state further down in the cell space emerges and dominates the population. With the model $V = \mathbf{v}_1 + 2\mathbf{v}^2_{aml}$, new abnormal cells appear around $t = 10$ and dominate the population at $t = 30$. The total number of cells is plotted in Figure 8B, where the effects of the parameters, c_{aml} and r_{aml}, are shown more clearly. The total number of cells increases more than 10 times the initial size after $t = 30$ when $c_{aml} = 10$ and $r_{aml} = 0$. When the over-proliferation term is appended as $r_{aml} = 1$, the total number of cells increases more rapidly, for example, up to 100 times the initial size and the number of cells in most of the myeloid lineage increases. Our simulation results agree with the experimental data, where unconventional cell states emerge during leukemia progression and eventually overtakes the entire progenitor population as observed by FACS analysis of bone marrow progenitor cells isolated from CM knock-in preleukemic and leukemic mice (Figure 8C). The predominant population observed in leukemic bone marrow does not fall within the typical gates in conventional cell clustering based on data from normal control mice (Figure 4E). Although this novel population would had been classified as MEP, pre-Meg/E, Pre-GM, and GMP cells in the graph model (Figure 4A,B), we emphasize that they are distinct population and the multi-dimensional model is capable of incorporating novel cell states. Although we comment that, in the graph model, a new cell type can be included by adding a new node to the original graph.

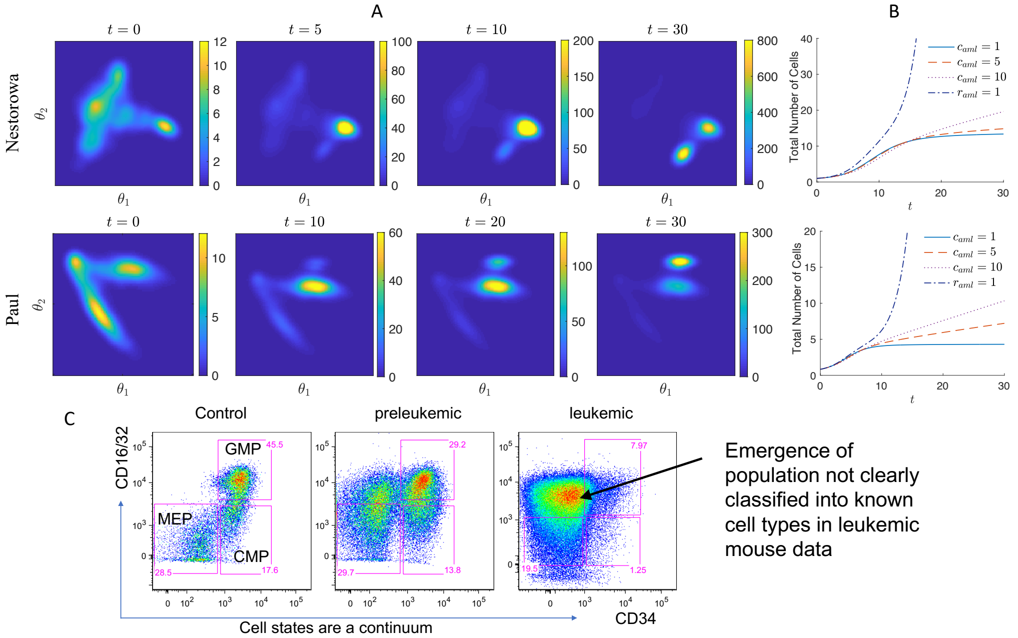

Figure 8. Cell state-transition dynamics during leukemia pathogenesis and progression. A) The evolution cell state distribution $u(t, \theta)$ with $c_{aml} = 2$. B) The total number of cells in AML condition is computed using model $V = \mathbf{v}_1 + c_{aml}\mathbf{v}_{aml}^2$. More rapid progression of AML in terms of cell number is observed for larger values of c_{aml} and r_{aml}. C) FACS analysis for CD34 and CD16/32 expression in myeloid progenitor compartment of control (left), preleukemic (center) and leukemic (right) CM knock-in mouse shows emergence of unconventional cell states during leukemic progression that eventually dominate the entire progenitor population. Our multi-dimensional cell state model is capable of incorporating those novel cell states.

3.6. *Interpretation of new cell states in the multi-dimensional model*

The remaining question is how to interpret the new cell states in the multi-dimensional space model that may arise far away from the cell states identified by conventional markers. Hence, we propose some measures in Eqs (2.12),(2.13) to guide the interpretation. Figure 9 shows an example of the rescaled correlation quantities $r_{f,j}$ and $r_{v,j}$ computed with Nestorowa data. The first row show results of the correlation $r_{v,j}$ to the average leukemic directional vector $\boldsymbol{v} = (-0.068, -0.206)$. The gene expression levels of genes that have large values of $r_{v,j}$ are depicted in the figure, namely, PLAC8 and CAR2. We remark that those genes have strong local correlation $r_{v,j}|_{\Gamma_d}$ on $\Gamma_d = \{0.3 \geq \theta_1 \geq 0.9, 0.3 \geq \theta_2 \geq 0.5\}$ as well. Figure 9 shows the correlation of all 3991 genes, where the red bars highlight the leukemia related genes we modify (A1. Details of the model equation and parameters) and we observe large magnitudes in some of the genes. The second row shows the correlation $r_{f,j}$ to a specific cell state at the reduced space, where we choose $\theta^* = (0.5, 0.35)$, which is approximately an intermediate location between MEP and CMP cells, and $f_{\theta^*}(\theta) = \frac{1}{2\pi 0.05} \exp\left[-\|\theta - \theta^*\|^2/0.1\right]$. APOE and CLEC12a genes show the largest magnitude of $r_{f,j}$, and similarly, we can identify the leukemia related genes that show strong correlation to cell state θ^*. Although more careful and rigorous approach should be developed to characterize the new arising cell states, $r_{f,j}$ and $r_{v,j}$ defined in Eqs (2.12),(2.13) provides an efficient method of initial screening of possible related genes.

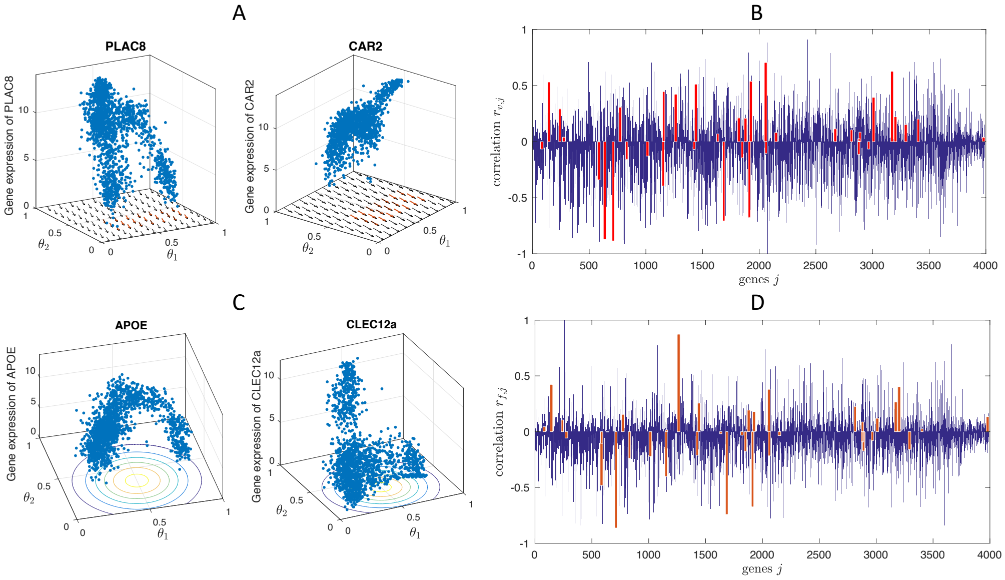

Figure 9. Interpretation and mapping of model-predicted novel cell states. In order to identify novel cell states predicted by the mathematical model, gene expression levels $\log_2(g_j^i + 1)$ that are strongly correlated to the direction of leukemic alteration $\boldsymbol{v} = (-0.068, -0.206)$ (A), and to the reduced space location $\theta^* = (0.5, 0.35)$ (C). The rescaled correlation $r_{f,j}$ (B) and $r_{v,j}$ (D) computed for all the genes in Nestorowa data are shown, and the leukemia related genes are marked in red bars.

4. Discussion

We have shown how to construct mathematical models of cell state-transitions using scRNA-seq data. We compare two cell state geometries: solving equations on graphs and solving equations on a multi-dimensional cell state-space. Each cell state geometry has its strengths and limitations. Selecting a model for a given application or dataset will depend on the type of biological data and the nature of the scientific question.

When the modeling application and quantity of interest includes well-known cell lineages and relation between the conventional cell states, the graph model is more appropriate due to its ability of distinguishing distinct cell lineages more clearly compared to the multi-dimensional space model. Dynamics of cell numbers in specific cell states, alteration of proliferation and apoptosis in particular cell state, differentiation block, and emergence of intermediate cell states can be quantified and studied in a straightforward manner. However, to explore cell states beyond known cell lineages, the continuum space model is more advantageous since it includes all intermediate and pathological cell states, rather than confining the model into presumed cell lineages. Moreover, the continuum model can incorporate a relatively small genetic and epigenetic alteration that the graph abstraction may not recognize, and study abnormal trajectories that yield unconventional cell states.

We selected and perturbed genes to simulate AML based on genes known to be associated with leukemia pathogenesis. We do not intend for this to be an accurate model of the biological process, rather, as an illustration of how one may select sets of genes and perturb them in a prescribed fashion in order to study the effect on cell state-transition dynamics. This approach assumes that AML pathogenesis originates from changes in gene expression in specific cell subsets, which is limited by our identification of these genes based on published literature. We acknowledge this is a limitation of the modeling approach, although we also note that our model predictions are consistent with known features of AML progression.

4.1. Comparison to other approaches

Although at the time of this work there are relatively few mathematical models published which utilize single-cell sequencing data, there are a few notable exceptions. Of particular note are works which use modeling and simulation to generate synthetic *in silico* gene expression datasets [37]. These important approaches to mechanism-based mathematical modeling may also be used to study and predict the effects of perturbations on cell state distributions. They may also be used as computational controls to benchmark analysis tools and potentially to benchmark and compare mathematical models, although using a model to benchmark other models can lead to consistent but incorrect circular reasoning and caution is warranted. Another example is Ferrall-Fairbanks and Papalexi et al, who use mathematical analysis to generate novel quantifications of cell heterogeneity in cancer or immune cell subsets respectively [38, 39]. These methods may be used to map and interpret novel cell states predicted by mathematical models or similarly as a method to interpret model-predicted changes in cell heterogeneity following a perturbation.

Schiebinger et al compute and predict differentiation trajectories in cell development using optimal transport (OT) [40,41]. This approach considers the optimal transport of cells as a mass flowing along differentiation trajectories, and is conceptually the most similar to our approach. As presented, Schiebinger et al do not use the OT framework to examine perturbations of cell states or genes along the differentiation trajectory, although this is possible with an OT model. Setty et al present a method to compute cell fate probabilities [42], which may also be achieved by inferring cell state-transition dynamics with lineage trees [43]. Fischer et al have demonstrated a method for inferring population dynamics from single-cell sequencing data [44], where the model equation is identical to our graph based model developed in [12]. Jiang et al develops a dynamic inference approach to derive a Fokker-Planck type PDE on a graph considering an energy landscape and optimal transport [45]. Sharma et al use longitudinal sequencing to study drug-induced infidelity in the stem cell hierarchy [46], and Karaayvaz et al show how to use single-cell sequencing to examine drug resistance in breast cancer [47]. These approaches and analysis methods may be used to inform and potentially calibrate mathematical models of cell population dynamics or response to treatment-induced perturbations.

Recently, vector fields derived from RNA velocity [48] have been used to infer potential energy or fitness landscapes for cell state-transitions. These approaches may be used to inform the computational domain for mathematical models that we present here, however, an important limitation of the RNA-velocity approach is extrapolation of the vector field outside of the data range. This underscores the need for hypothesis-based and model-guided approaches to inform the shape of these fields. This limitation also applies to the rapidly growing field of deep learning approaches [49] to analyze single-cell sequencing data, namely, whether the learning algorithm can effectively make predictions to datasets which are not sufficiently similar to those upon which it has been trained. We believe that the future likely involves a merger of mathematical modeling with machine learning, in which mathematical models are used to inform learning approaches and impute sparse data as has been recently shown by Gaw and Rockne et al [50, 51]. Among the recent works that align with this direction, PRESCIENT algorithm aims to learn the underlying differentiation landscape from time-series scRNA-seq data [21]. Moreover, dynamo framework improves RNA velocity using kinetic models to reconstruct continuous vector fields that predict cell fates [52].

4.2. Opportunities and limitations of modeling with single-cell sequencing data

There are pros and cons, opportunities and limitations to mathematical modeling with single-cell sequencing data. The advantages and potential opportunities include: a wealth of available data, richness and complexity of each data set, a focus on the cell level, opportunity to study dynamics in hierarchically structured state-based relationships between cells, and an ability to perturb individual cells and/or genes within cells to predict dynamics of state-change at cellular level. The most significant strength of mathematical modeling is the ability to use and generate hypotheses that may not be directly evident from the data; for example, extrapolation of RNA velocity fields beyond the dataset boundaries or to interpret and predict novel cell states which may not otherwise be clearly identified with known canonical cell state markers. Another advantage of our approach is the use of pseudo-time analysis of data collected at a single timepoint to calibrate the models, however, the models can also be calibrated directly to time-sequential single-cell datasets, which we expect to become more commonly available as single-cell sequencing continues to be used as a tool to study cell dynamics.

The disadvantages and limitations include: the potential for misleading or incorrect inference due to poor data quality including drop-outs, small non-representative samples of large heterogeneous populations, batch effects, no physical or micro-environmental context, no direct or physical interactions between cells, and the possibility of model predictions to be sensitive to methods of dimension reduction, graph abstraction, state-space construction, and potentially sequencing platform. Sensitivity of the modeling to experimental and computational methods may be directly studied and mitigated as we have shown in this work, however this remains potentially a significant source of uncertainty and variability in the modeling calibration and predictions. Studying the sensitivity of our modeling framework regarding different noise scenarios and applying noise reduction methods is our future work [53].

In terms of computational cost, the graph model is more efficient since it is a multiple of one-dimensional cost, while the cost of implementing the space model increases exponentially as the dimension of reduced space increases. In our simulation, the computational cost to simulate up to time $t = 50$ with step size $\Delta t = 10^{-3}$ and $O(100^2)$ degrees of freedom in one-dimension is around 25 seconds in the graph model with 8 nodes, while it takes around 230 seconds in the continuum model with two dimensions. In short, the continuum model runs approximately 10 times longer than the graph model with 8 nodes in our example. Therefore, the multi-dimensional cell state geometry will be reasonable only when the reduced component can be truncated at two- to three-dimension, unless the numerical method is carefully implemented, and we emphasize that the graph model will be more advantageous in terms of computational cost than the continuum model especially when higher dimensional reduced space is necessary.

4.3. Future work and applications

Future applications of this approach is to explore hypothesis in the resolution of single-cell genomics and study altered and novel cell states with genetic and epigenetic alterations in various biological systems and pathogenesis. We look forward to compare the model prediction to sampling/sequencing of perturbed biological system, for instance, to examine scRNA-seq data from leukemic progenitor cells. Moreover, we anticipate to incorporate effects of external perturbations such as therapy in future studies.

There are opportunities for further enhancements in our model in improving the model of cell landscape dynamics to accurately estimate cell transition pathways in the reduced component space, for instance, minimum action paths [6] and bifurcation [7, 54]. The model can be improved by obtaining parameter functions or mappings of biological quantities directly from single-cell sequencing data, for example, more precisely infer the proliferation rate function. Also, developing methodologies to obtain reduced component space that captures desired characteristic of cell states [55] will help us explore our approach for other biological settings where cell states are less clearly characterized. Moreover, we propose to develop quantities, such as index of critical state transitions [54, 56], in the phenotype space that could be used to predict forthcoming major alterations in development and diseases. We also expect to be able to infer the potential landscape directly from the RNA velocity vector field [48, 52].

5. Conclusions

In summary, despite the explosion of computational tools to analyze single-cell sequencing data, there have been relatively few mathematical models developed which utilize this data. Here we begin to explore the possibilities—and limitations—of dynamical modeling with single-cell RNA-seq data. We hope this work paves the way for development of mathematical models to guide the interpretation of these complicated datasets as they begin to be collected after biological perturbations (eg., cancer, treatment, altered developmental processes), sequentially over time, or sampled spatially within biological tissues.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers U01CA250046 (Y.H.K., R.C.R.), R01CA178387 (Y.H.K.), R01CA205247 (Y.H.K.), P30CA033572 and included work performed in the Analytical Cytometry and Biostatistics and Mathematical Oncology Cores. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest

The authors declare that there is no conflict of interest.

Code availability

The simulation codes are available from https://github.com/heyrim/Mathematical-modeling-with-single-cell-sequencing-data

References

1. V. Svensson, R. Vento-Tormo, S. A. Teichmann, Exponential scaling of single-cell RNA-seq in the past decade, *Nat. Protoc.*, **13** (2018), 599–604. https://doi.org/10.1038/nprot.2017.149
2. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, et al., Massively parallel digital transcriptional profiling of single cells, *Nat. Commun.*, **8** (2017), 14049. http://dx.doi.org/10.1038/ncomms14049
3. T. Stuart, R. Satija, Integrative single-cell analysis, *Nat. Rev. Genet.*, **20** (2019), 257–272. http://www.nature.com/articles/s41576-019-0093-7
4. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods: towards more accurate and robust tools, *Nat. Biotechn.*, **37** (2019), 547–554. https://doi.org/10.1038/s41587-019-0071-9
5. V. Y. Kiselev, T. S. Andrews, M. Hemberg, Challenges in unsupervised clustering of single-cell RNA-seq data, *Nat. Rev. Genet.*, **20** (2019), 273–282. https://doi.org/10.1038/s41576-018-0088-9
6. R. D. Brackston, E. Lakatos, M. P. Stumpf, Transition state characteristics during cell differentiation, *PLoS Comput. Biol.*, **14** (2018), e1006405. https://doi.org/10.1371/journal.pcbi.1006405
7. E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, et al., Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape, *Proc. Nat. Academy Sci.*, **111** (2014), E5643–E5650. https://doi.org/10.1073/pnas.1408993111
8. A. E. Teschendorff, T. Enver, Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome, *Nat. Commun.*, **8** (2017), 15599, http://dx.doi.org/10.1038/ncomms15599 .
9. S. Jin, A. L. Maclean, T. Peng, Q. Nie, ScEpath: Energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data, *Bioinformatics*, **34** (2018), 2077–2086. https://doi.org/10.1093/bioinformatics/bty058
10. J. Guo, J. Zheng, HopLand: Single-cell pseudotime recovery using continuous Hopfield network-based modeling of Waddington's epigenetic landscape, *Bioinformatics*, **33** (2017), i102–i109. https://doi.org/10.1093/bioinformatics/btx232
11. M. Zwiessele, N. D. Lawrence, Topslam: Waddington Landscape Recovery for Single Cell Experiments, preprint, BioRxiv, 2017: 057778. https://doi.org/10.1101/057778

12. H. Cho, K. Ayers, L. DePillis, Y. h. Kuo, J. Park, A. Radunskaya, et al., Modeling acute myeloid leukemia in a continuum of differentiation states, *Lett. Biomath.*, **5** (2018), S69–S98.

13. S. Nestorowa, F. K. Hamey, B. Pijuan Sala, E. Diamanti, M. Shepherd, E. Laurenti, et al., A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation, *Blood*, **128** (2016), 20–32. https://doi.org/10.1182/blood-2016-05-716480

14. F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, et al., Transcriptional heterogeneity and lineage commitment in myeloid progenitors, *Cell*, **163** (2015), 1663–1677. https://doi.org/10.1016/j.cell.2015.11.013

15. L. Haghverdi, F. Buettner, F. Theis, Diffusion maps for high-dimensional single-cell analysis of differentiation data, *Bioinformatics*, **31** (2015), 2989–2998. https://doi.org/10.1093/bioinformatics/btv325

16. M. Barron, J. Li, Identifying and removing the cell-cycle effect from single-cell rna-sequencing data, *Sci. Rep.*, **6** (2016), 33892. https://doi.org/10.1038/srep33892

17. J. Wang, L. Xu, E. Wang, Potential landscape and flux framework of nonequilibrium networks: robustness, dissipation, and coherence of biochemical oscillations, *Proc. Nat. Acad. Sci.*, **105** (2008), 12271–12276. https://doi.org/10.1073/pnas.0800579105

18. Z. I. Botev, D. P. Kroese, The generalized cross entropy method, with applications to probability density estimation, *Methodol. Comput. Appl. Probab.*, **13** (2011), 1–27. https://doi.org/10.1007/s11009-009-9133-7

19. M. Doumic, A. Marciniak-Czochra, B. Perthame, J. P. Zubelli, A structured population model of cell differentiation, *SIAM J. Appl. Math.*, **71** (2011), 1918–1940. https://doi.org/10.1137/100816584

20. C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, A. M. Klein, Fundamental limits on dynamic inference from single-cell snapshots, *Proc. Nat. Acad. Sci.*, **115** (2018), E2467–E2476. https://doi.org/10.1073/pnas.1714723115

21. G. H. T. Yeo, S. D. Saksena, D. K. Gifford, Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions, *Nat. Commun.*, **12** (2021), 3222. https://doi.org/10.1038/s41467-021-23518-w

22. L. C. Evans, *An Introduction to Stochastic Differential Equations*, American Mathematical Society, 2014.

23. F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, et al., PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells, *Genome Biol.*, **20** (2019), 59. https://doi.org/10.1186/s13059-019-1663-x

24. L. C. Evans, *Partial Differential Equations*, 2nd edition, American Mathematical Society, 2010.

25. R. C. Rockne, S. Branciamore, J. Qi, D. E. Frankhouser, D. O'Meally, W. K. Hua, et al., State-transition analysis of time-sequential gene expression identifies critical points that predict development of acute myeloid leukemia, *Cancer Res.*, **80** (2020), 3157–3169. https://doi.org/10.1158/0008-5472.CAN-20-0354

26. A. W. Bowman, A. Azzalini, *Applied Smoothing Techniques for Data Analysis*, Oxford University Press Inc., New York, 1997.

27. Q. Cai, R. Jeannet, W. K. K. Hua, G. J. Cook, B. Zhang, J. Qi, et al., Cbfβ-smmhc creates aberrant megakaryocyte-erythroid progenitors prone to leukemia initiation in mice, *Blood*, **128** (2016), 1503–1515. https://doi.org/10.1182/blood-2016-01-693119

28. P. Liu, S. A. Tarlé, A. Hajra, D. F. Claxton, P. Marlton, M. Freedman, et al., Fusion between transcription factor cbf beta/pebp2 beta and a myosin heavy chain in acute myeloid leukemia, *Science*, **261** (1993), 1041–1044. https://doi.org/10.1126/science.8351518

29. P. P. Liu, C. Wijmenga, A. Hajra, T. B. Blake, C. A. Kelley, R. S. Adelstein, et al., Identification of the chimeric protein product of the cbfb-myh11 fusion gene in inv(16) leukemia cells, *Genes Chromosomes Cancer*, **16** (1996), 77–87. https://doi.org/10.1002/(SICI)1098-2264(199606)16:2< 77 :: $AID - GCC1$ >3.0.CO;2-%23

30. L. H. Castilla, L. Garrett, N. Adya, D. Orlic, A. Dutra, S. Anderson, et al., The fusion gene cbfb-myh11 blocks myeloid differentiation and predisposes mice to acute myelomonocytic leukaemia, *Nat. Genet.*, **23** (1999), 144–146. https://doi.org/10.1038/13776

31. Y. H. H. Kuo, S. F. Landrette, S. A. Heilman, P. N. Perrat, L. Garrett, P. P. Liu, et al., Cbfβ-smmhc induces distinct abnormal myeloid progenitors able to develop acute myeloid leukemia, *Cancer Cell*, **9** (2006), 57–68. https://doi.org/10.1016/j.ccr.2005.12.014

32. Y. H. H. Kuo, R. M. Gerstein, L. H. Castilla, Cbfβ-smmhc impairs differentiation of common lymphoid progenitors and reveals an essential role for runx in early b-cell development, *Blood*, **111** (2008), 1543–1551. https://doi.org/10.1182/blood-2007-07-104422

33. C. J. H. Pronk, D. J. Rossi, R. Mansson, J. L. Attema, G. L. Norddahl, C. K. F. Chan, et al., Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy, *Cell Stem Cell*, **1** (2007), 428–442. https://doi.org/10.1016/j.stem.2007.07.005

34. K. Akashi, D. Traver, T. Miyamoto, I. L. Weissman, A clonogenic common myeloid progenitor that gives rise to all myeloid lineages, *Nature*, **404** (2000), 193–197. https://doi.org/10.1038/35004599

35. S. Ng, A. Mitchell, J. A. Kennedy, W. C. Chen, J. Mcleod, N. Ibrahimova, et al., A 17-gene stemness score for rapid determination of risk in acute leukaemia, *Nature*, **540** (2016), 433–437, http://dx.doi.org/10.1038/nature20598

36. C. Pabst, A. Bergeron, V. P. Lavall, J. Yeh, P. Gendron, G. L. Norddahl, et al., GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo, *Blood*, **127** (2017), 2018–2027. https://doi.org/10.1182/blood-2015-11-683649

37. T. D. Sherman, L. T. Kagohara, R. Cao, R. Cheng, M. Satriano, M. Considine, et al., CancerInSilico: An R/Bioconductor package for combining mathematical and statistical modeling to simulate time course bulk and single cell gene expression data in cancer, *PLOS Comput. Biol.*, **14** (2019), e1006935. https://doi.org/10.1371/journal.pcbi.1006935

38. M. C. Ferrall-Fairbanks, M. Ball, E. Padron, P. M. Altrock, Leveraging single cell RNA sequencing experiments to model intra-tumor heterogeneity, *Clin. Cancer Inf.*, **3** (2019), 1–10. http://doi.org/10.1200/CCI.18.00074

39. E. Papalexi, R. Satija, Single-cell RNA sequencing to explore immune cell heterogeneity, *Nat. Rev. Immunol.*, **18** (2018), 35–45. https://doi.org/10.1038/nri.2017.76 .

40. G. Schiebinger, J. Shu, R. Jaenisch, A. Regev, E. S. Lander, Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming, *Cell*, **176** (2019), 928–943. https://doi.org/10.1016/j.cell.2019.01.006 .

41. G. Schiebinger, Reconstructing developmental landscapes and trajectories from single-cell data, *Curr. Opin. Syst. Biol.*, **27** (2021), 100351. https://doi.org/10.1016/j.coisb.2021.06.002

42. 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), 451–460, http://dx.doi.org/10.1038/s41587-019-0068-4

43. S. Hormoz, Z. S. Singer, J. M. Linton, Y. E. Antebi, B. I. Shraiman, M. B. Elowitz, Inferring cell-state transition dynamics from lineage trees and endpoint single-cell measurements, *Cell Syst.*, **3** (2016), 419–433. https://doi.org/10.1016/j.cels.2016.10.015

44. D. S. Fischer, A. K. Fiedler, E. M. Kernfeld, R. M. J. Genga, A. Bastidas-ponce, M. Bakhti, et al., Inferring population dynamics from single-cell RNA-sequencing time series data, *Nat. Biotechnol.*, **37** (2019), 461–468. https://doi.org/10.1038/s41587-019-0088-0 .

45. Q. Jiang, S. Zhang, L. Wan, Dynamic inference of cell developmental complex energy landscape from time series single-cell transcriptomic data, *PLOS Comput. Biol.*, **18** (2022), e1009821. https://doi.org/10.1371/journal.pcbi.1009821

46. A. Sharma, E. Y. Cao, V. Kumar, X. Zhang, H. S. Leong, A. M. L. Wong, et al., Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy, *Nat. Commun.*, https://doi.org/10.1038/s41467-018-07261-3 .

47. M. Karaayvaz, S. Cristea, S. M. Gillespie, A. P. Patel, R. Mylvaganam, C. C. Luo, et al., Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq, *Nat. Commun.*, **9** (2018), 3588. https://doi.org/10.1038/s41467-018-06052-0 .

48. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, et al., RNA velocity of single cells, *Nature*, **560** (2018), 494–498. https://doi.org/10.1038/s41586-018-0414-6

49. G. Eraslan, Ž. Avsec, J. Gagneur, F. J. Theis, Deep learning : new computational modelling techniques for genomics, *Nat. Rev. Genet.*, **20** (2019). https://doi.org/10.1038/s41576-019-0122-6

50. N. Gaw, A. Hawkins-Daarud, L. S. Hu, H. Yoon, L. Wang, Y. Xu, et al., Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric MRI, *Sci. Rep.*, **9** (2019), 10063. https://doi.org/10.1038/s41598-019-46296-4

51. R. C. Rockne, A. Hawkins-Daarud, K. R. Swanson, J. P. Sluka, J. A. Glazier, P. Macklin, et al., The 2019 mathematical oncology roadmap, *Phys. Biol.*, **16** (2019), 4. https://doi.org/10.1088/1478-3975/ab1a09

52. X. Qiu, Y. Zhang, J. D. Martin-Rufino, C. Weng, S. Hosseinzadeh, D. Yang, et al., Mapping transcriptomic vector fields of single cells, *Cell*, **185** (2022), 690–711. https://doi.org/10.1016/j.cell.2021.12.045

53. S. K. Chu, S. Zhao, Y. Shyr, Q. Liu, Comprehensive evaluation of noise reduction methods for single-cell rna sequencing data, *Briefings Bioinf.*, **23** (2022), bbab565. https://doi.org/10.1093/bib/bbab565

54. M. Mojtahedi, A. Skupin, J. Zhou, I. G. Castaño, R. Y. Leong-Quong, H. Chang, et al., Cell fate decision as high-dimensional critical state transition, *PLoS Biol.*, **14** (2016), 1–28. https://doi.org/10.1371/journal.pbio.2000640

55. C. Li, L. Zhang, Q. Nie, Landscape reveals critical network structures for sharpening gene expression boundaries, *BMC Syst. Biol.*, **12** (2018), 67. https://doi.org/10.1186/s12918-018-0595-5

56. J. I. Joo, J. X. Zhou, S. Huang, K. H. Cho, Determining relative dynamic stability of cell states using boolean network model, *Sci. Rep.*, **8** (2018), 12077. https://doi.org/10.1038/s41598-018-30544-0

57. B. E. Shepherd, P. Guttorp, P. M. Lansdorp, J. L. Abkowitz, Estimating human hematopoietic stem cell kinetics using granulocyte telomere lengths, *Exp. Hematol.*, **32** (2004), 1040–1050. https://doi.org/10.1016/j.exphem.2004.07.023

58. E. P. Cronkite, Kinetics of granulopoiesis, *Clin. Haematol.*, **8** (1979), 351–370.

59. S. Hao, C. Chen, T. Cheng, Cell cycle regulation of hematopoietic stem or progenitor cells, *Int. J. Hematol.*, **103** (2016), 487–497. https://doi.org/10.1007/s12185-016-1984-4

60. E. M. Pietras, M. R. Warr, E. Passegué, Cell cycle regulation in hematopoietic stem cells, *J. Cell Biol.*, **195** (2011), 709–720. https://doi.org/10.1083/jcb.201102131

61. T. Stiehl, A. D. Ho, A. Marciniak-Czochra, The impact of CD34+ cell dose on engraftment after SCTs: personalized estimates based on mathematical modeling, *Bone Marrow Transp.*, **49** (2014), 30–37. https://doi.org/10.1038/bmt.2013.138

62. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, et al., Geometric diffusions as a tool for harmonic analysis and structure definition of data : Diffusion maps, *Proc. Natl. Acad. Sci.*, **102** (2005), 7426–7431. https://doi.org/10.1073/pnas.0500334102

63. L. Haghverdi, M. Büttner, F. Wolf, F. Buettner, F. Theis, Diffusion pseudotime robustly reconstructs lineage branching, *Nat. Methods*, **13** (2016), 845–848. https://doi.org/10.1038/nmeth.3971

64. M. Jacomy, T. Venturini, S. Heymann, M. Bastian, Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software, *PLOS One*, **9** (2014), e98679. https://doi.org/10.1371/journal.pone.0098679

Appendix

A1. Details of the model equation and parameters

The model terms require interpolation of single-cell data to the continuum cell state space. We use clustering to identify cell types and their cell type properties to assign those to each single-cell. The following are the values we take for cluster properties.

By denoting $\bar{r}_i$ as the assigned proliferation rate of the i-th cluster, we compute the intermediate level of proliferation in the graph model by linear interpolation as

$$r_k(x) = r_{I(i,j)}(x) = \bar{r}_i + (\bar{r}_j - \bar{r}_i)x, \quad x \in [0, 1], \tag{A.1}$$

Table A1. Summary of the required parameters. The following table summarizes the parameters in our model terms, V, R, D, V_k, R_k, and D_k, and their biological meaning with the range. The ranges are found from the literature [19,57–61] experimentally measuring the cell cycle and self-renewal rate of the well known hematopoeisis cell types.

Parameters	Biological meaning	Range	
$r(\theta)$, $r_k(x)$	proliferation rate	[0.00215, 1]	[57–60]
$a(\theta)$, $a_k(x)$	self-renewal rate	$[0.1, 0.8]$	[19,61]
$\boldsymbol{c}(\theta)$	differentiation vector	$[0, 1]$	estimated
ν	phenotypic fluctuation	$[0, 0.0027]$	estimated
$\bar{d}$	apoptosis rate	0.6925	[61]

Table A2. Parameter values of proliferation and self-renewal rate. The following values are taken for each single-cell, $\bar{r}^i$ and $\bar{a}^i$, based on their clustered cell types [19,57–61], and then used for computing $r(\theta)$, $r_k(x)$, $a(\theta)$, and $a_k(x)$.

	hematopoietic stem cells (HSC) ↔ progenitor cells (HPC)			
cell type	HSC	MPP, LMPP, CMP	MEP, GMP	Neu/Mo, Ery
proliferation	0.01125 (8.8 weeks)	0.05658 (12.25 days)	0.1612 (4.3 days)	0.6931 (1 day)
self-renewal	0.77	0.7689	0.7359	0.66

assuming that the overall proliferation of intermediate cell states change gradually. In the multi-dimensional model, we compute the interpolation based on local means as

$$r(\theta) = \frac{1}{|I_\theta|} \sum_{i \in I_\theta} \bar{r}^i, \quad I_\theta = \{i \,|\, \|\theta^i - \theta\| < \bar{\theta}\}, \tag{A.2}$$

where we take $\bar{\theta} = 0.04$. The self-renewal rate functions $a_k(x)$ and $a(\theta)$ are computed similarly. See Cell proliferation rate $r(\theta)$ and self-renewal rate $a(\theta)$ computed from the single-cell data. The black dots are the rates of data. for $r(\theta)$ and $a(\theta)$ computed for Nestorowa data.

To compute the multi-dimensional function on the continuum space from the single-cell data, we employ the *kernel density method* [18,26], that is a non-parametric way to estimate the density function based on a finite data sample. Using the single-cell samples in the reduced component space, $\{\theta^i\}_{i=1}^N$, the method approximates the density function as

$$u_s(\theta) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{(\theta - \theta^i)}{h}\right),$$

where K is the kernel smoothing function that we take it as a Gaussian function and h is the bandwidth. The optimal bandwidth to estimate normal densities can be computed by $(4\hat{\sigma}^5/3N)^{1/5}$, where $\hat{\sigma}$ is the standard deviation and N is the sample size, and the optimal bandwidth for our data is computed as 0.0383 to 0.0456, however in our simulation, we choose a slightly smaller value, $h = 0.03$, to reveal more features of multiple modes. Figure 2(C,D) shows the results computing $u_s(\theta)$ using Nestorowa and Paul data, and A2(a) shows the corresponding $\mathbf{v}_1(\theta)$.

In the diffusion term, we explore the parameter ν so that the phenotypic instability does not dominate the cell maturation. We compute the parameter in the range of $\nu \leq (L/T_d)^2/4$, where the distance in the diffusion space is $L = 1$ and the time that HSC differentiates to the progenitors is $T_d = 5 \sim 30$ (day), that is, $\nu \leq 0.0027 \sim 0.01$, and we consider $\nu = 0.001$. Quantifying the local phenotypic instability in the reduced component space, and justifying this term is our future work.

To compute the reduced component space using dimension reduction approaches, we employ *diffusion mapping*. See [15, 62] for the detail of the algorithm. We take the cosine distance, $k(x^i, x^j) = 1 - corr(x^i, x^j)$ for the Nestorowa data and the gaussian distance $k(x^i, x^j) = \exp\left(-\frac{\|x^i - x^j\|^2}{2\sigma^2}\right)$ for Paul data with $\sigma = 50$. From $L(i, j) = k(x^i, x^j)$, the diffusion mapping use parameter α to tune the influence of density of the data points as

$$L^{(\alpha)} = D^{-\alpha} L D^{-\alpha}, \quad M = (D^{(\alpha)})^{-1} L^{(\alpha)},$$

where $D^{(\alpha)}(i, i) = \sum_j L^{(\alpha)}(i, j)$, and we choose $\alpha = 0.5$. From the eigen-decomposition of $M\phi = \lambda\phi$ and ordered eigenvalues $1 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$, the corresponding right eigenvectors, $\phi_1, \phi_2, \cdots$ are the diffusion components. We truncate the reduced space at the second diffusion component, where the eigenvalues are $\lambda_1 = 0.1039$, $\lambda_2 = 0.0326$, $\lambda_3 = 0.0167$, $\lambda_4 = 0.0135$ for Nestorowa data, and $\lambda_1 =$ 2.4653e-03, $\lambda_2 =$ 5.8338e-04, $\lambda_3 =$ 9.7792e-05, $\lambda_4 =$ 7.0364e-05 for Paul data. For a comparison of diffusion mapping to two-dimensional reduced component space using other dimension reduction algorithms, see Comparison of dimension reduction methods. Dimension reduction algorithms that focus on preserving local structure are not appropriate to infer global trajectory. Compare the following figures to diffusion component space. They are computed with Nestorowa (top) and Paul (bottom) data and projected on the reduced component space of ForceAtlas2 (FA) [64] and t-stochastic neighbor embedding (tSNE). .

For the pseudotime inference, we use the algorithm developed in [63]. The diffusion distance between two cells are computed as

$$D_t^2(x^i, x^j) = \sum_{k=1}^{n} \lambda_k^{2t} (\phi_k^i - \phi_k^j)^2,$$

and the pseudotime distances are computed based on this distance. We choose three extreme points in each of the three clusters, stem cell, Ery, and Neu cell types, that are the furtherest in the diffusion component space, and infer the lineage between the extreme cells. After computing the pesudotime of each single-cell we compute the local average direction to the neighborhood cells that are in later pesudotime similar as in Eq (A.2). The computed results are shown in Pseudotime dynamics. The homeostasis cell differentiation vector $\mathbf{v}_1$ (a), and the direction of active cell differentiation obtained from diffusion pseudotime analysis (b), and that interpolated at the grid points $\mathbf{v}_2$ (c) are presented. We remark that, $\mathbf{v}_2$ corresponds to the cell differentiation along the edges in the graph model. (b) with the interpolated vector at the grid points, Pseudotime dynamics. The homeostasis cell differentiation vector $\mathbf{v}_1$ (a), and the direction of active cell differentiation obtained from diffusion pseudotime analysis (b), and that interpolated at the grid points $\mathbf{v}_2$ (c) are presented. We remark that, $\mathbf{v}_2$ corresponds to the cell differentiation along the edges in the graph model. (c).

Table A3. Gene alterations in Leukemic stem cells. From the genes that are reported in [35, 36], we find all the gene that are in Nestorowa and Paul data. The following table is the genes and their altered magnitude. See [35] Extended Data Table 1 for the 17 genes and [36] Supplemental Table S4 for approximately 80 genes.

Up-regulated Gene	log2-fold change	Down-regulated Gene	log2-fold change
CD34	2.1500	LGALS3	−3.4901
LAPTM4B	1.8000	CYBB	−2.9546
MMRN1	1.3600	CD36	−2.7661
SOCS2	1.2400	ANXA5	−2.6349
CDK6	1.2300	LY86	−2.5564
CPXM1	1.2000	IRF8	−2.4982
EMP1	1.0100	SAMHD1	−2.4580
GPR56	2.7004	GRN	−2.3659
GATA2	1.8875	RNASE6	−2.3585
LPIN1	1.6323	FCER1G	−2.2934
MZB1	1.4854	S100A9	−2.2447
ZSCAN18	1.3219	TLR4	−2.1078
GUCY1A3	1.2630	FCGRT	−2.1016
SPNS2	1.2016	S100A8	−2.0116
PTK7	1.2016	CLEC12A	−1.8730
ABCC1	1.1375	MNDA	−1.8417
SYTL1	1.0704	IL13RA1	−1.7515
MAGED1	1.0704	SGK1	−1.7418
ARHGAP25	1.0704		
SLA2	1.0000		

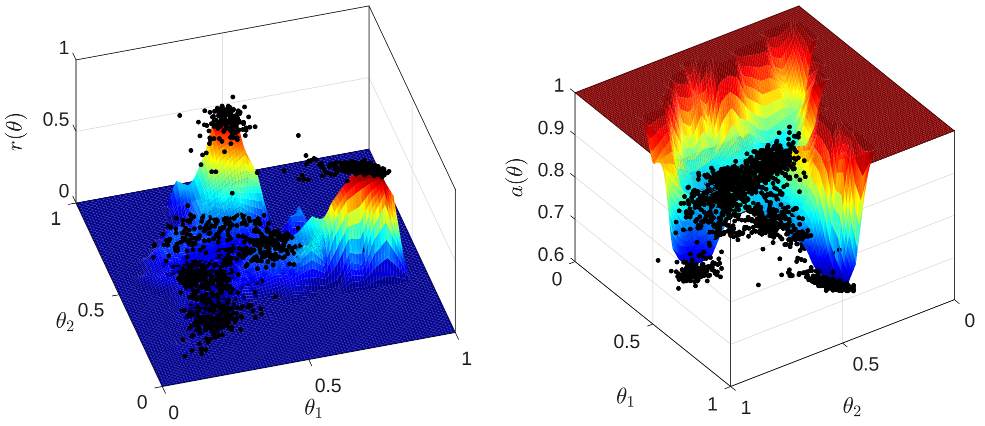

Figure A1. Cell proliferation rate $r(\theta)$ and self-renewal rate $a(\theta)$ computed from the single-cell data. The black dots are the rates of data.

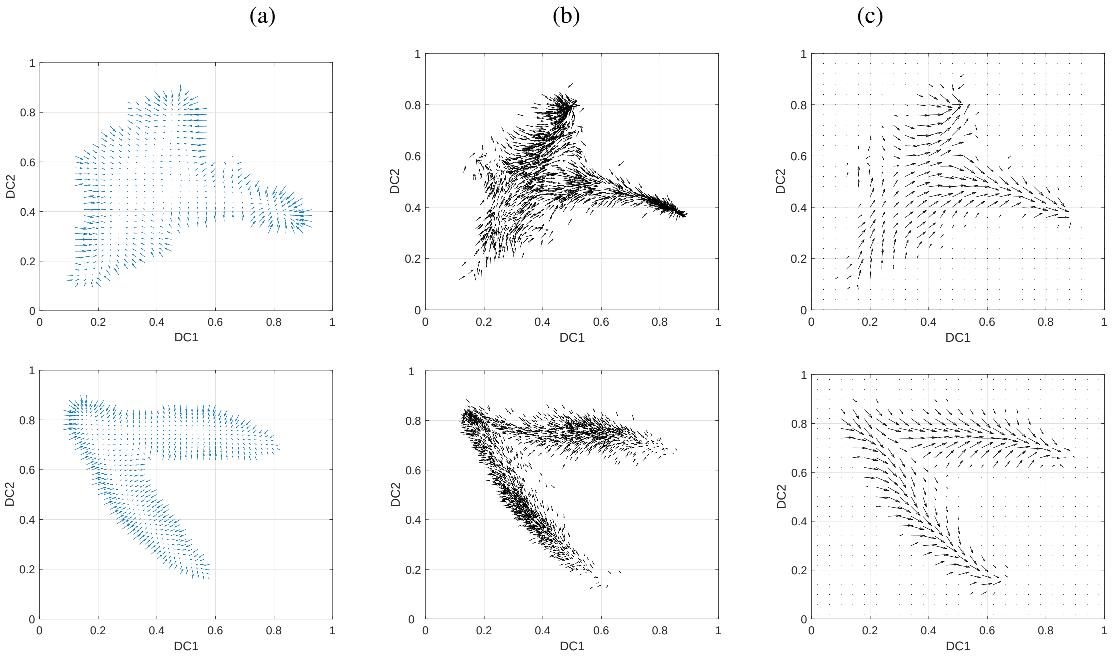

Figure A2. Pseudotime dynamics. The homeostasis cell differentiation vector $\mathbf{v}_1$ (a), and the direction of active cell differentiation obtained from diffusion pseudotime analysis (b), and that interpolated at the grid points $\mathbf{v}_2$ (c) are presented. We remark that, $\mathbf{v}_2$ corresponds to the cell differentiation along the edges in the graph model.

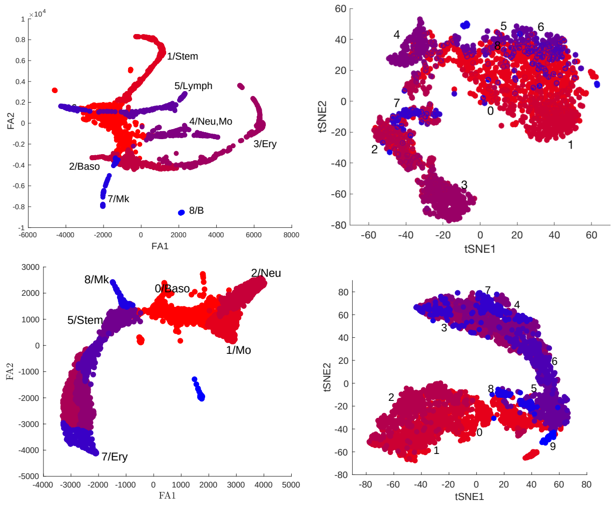

Figure A3. Comparison of dimension reduction methods. Dimension reduction algorithms that focus on preserving local structure are not appropriate to infer global trajectory. Compare the following figures to diffusion component space. They are computed with Nestorowa (top) and Paul (bottom) data and projected on the reduced component space of ForceAtlas2 (FA) [64] and t-stochastic neighbor embedding (tSNE).

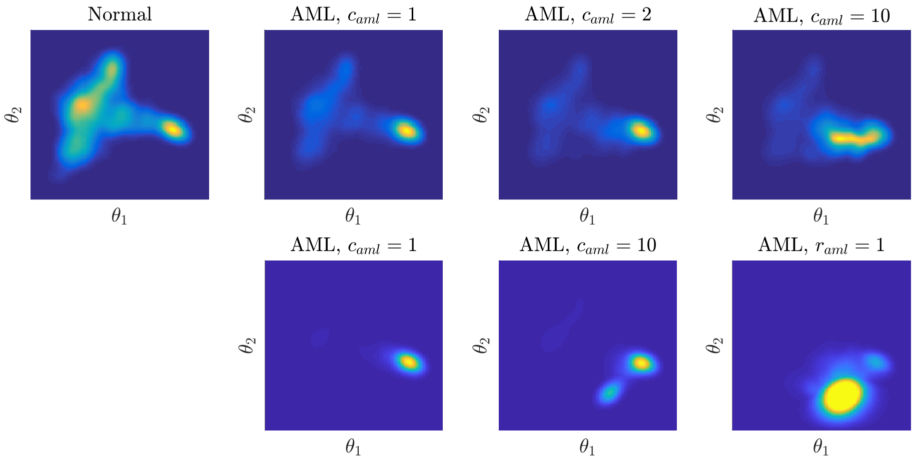

Figure A4. Abnormal cell state transitions during leukemia pathogenesis and progression. The distribution of cell states $u(t, \theta)$ show abnormal cell states emerging during leukemic progression after $t = 10$ modeled in the advection term as $V = c_{aml}\mathbf{v}^1_{aml}$ (top) and $V = \mathbf{v}_1 + c_{aml}\mathbf{v}^2_{aml}$ (bottom), with various levels of $c_{aml} = 1, 2, 10$. Larger magnitude of c_{aml} results in more disrupted cell landscape.

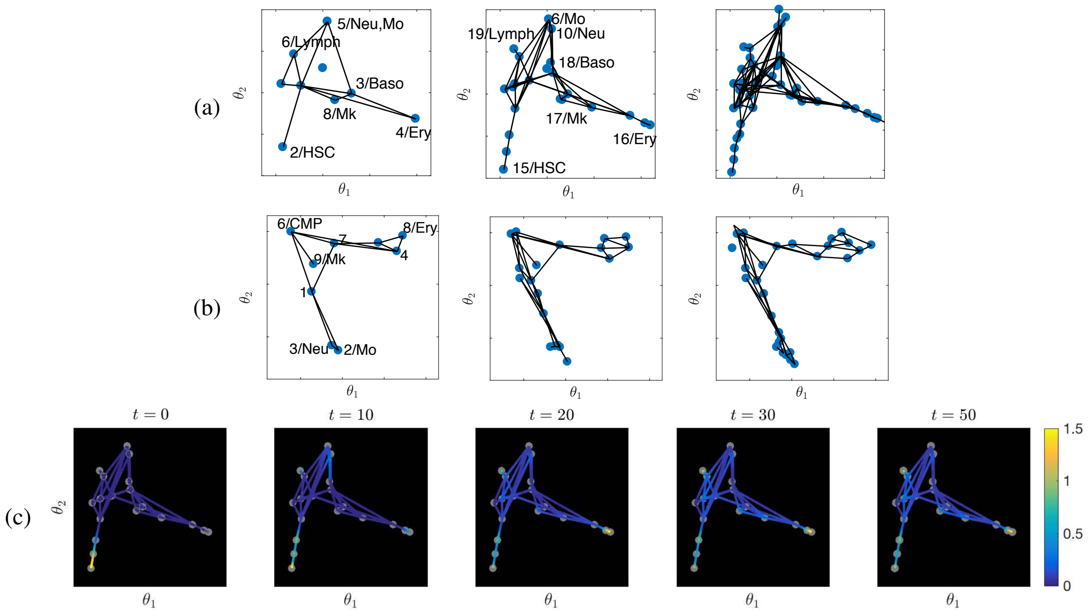

Figure A5. From discrete to continuum cell states. The hierarchy of graphs using partition-based graph abstraction [23] and single cell data from Nestorowa et al. (2016) (a) and Paul et al. (2015) (b). The single-cell data can be regarded as the most refined graph. The simulation of normal hematopoiesis on graph with 19 nodes (c), that is comparable to Figure 3A, illustrates the hierarchy of cell distribution toward the entire reduce space.

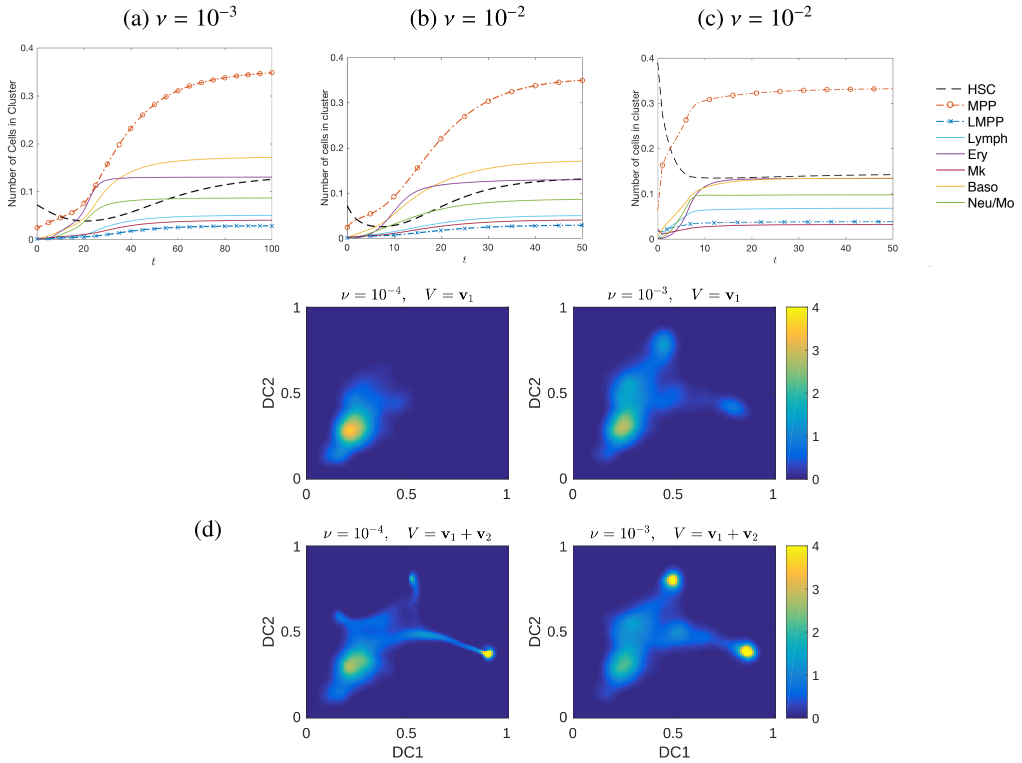

Figure A6. Model sensitivity to parameters. Using single cell data from Nestorowa et al. (2016), number of cells and its dynamics in each cluster up to $t = 50$ for different values of ν and initial stem cell numbers $\rho(0)$ are shown in (a–c). The dynamics of cells in each cluster for $\nu = 10^{-3}$ with $\rho(0) = 0.1$ (a), $\nu = 10^{-2}$ with $\rho(0) = 0.1$ (b), and $\nu = 10^{-2}$ with larger initial number of cells, $\rho(0) = 0.5$ (c) shows that the recovery is more rapid for larger values of ν and larger number of initial stem cells $\rho(0)$. Cell distribution $u(\theta, t)$ at intermediate time $t = 14$ for advection terms $\mathbf{v}_1$ and $\mathbf{v}_1 + \mathbf{v}_2$, and $\nu = 10^{-4}$ or $\nu = 10^{-3}$ are shown in (d). The distributions are distinct, where larger values of ν increases overall rate of differentiation, while adding $\mathbf{v}_2$ prioritizes recovery of the most matured cells.